VNiVERSiDAD D SALAMANCA

CiC CENTRO DE INVESTIGACIÓN DEL CÁNCER

Desarrollo de algoritmos bioinformáticos para estudios de genómica funcional: aplicaciones en cáncer

TESIS DOCTORAL

Celia Fontanillo Fontanillo

Director:

Dr. Javier De Las Rivas Sanz

Salamanca, Febrero de 2013

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 "***Desarrollo de algoritmos bioinformáticos para estudios de genómica funcional: aplicaciones en cáncer***", presentada por Dña. **Celia Fontanillo Fontanillo** alumna 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 13 de febrero de 2013

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 general

Índice de figuras

Índice de tablas

Introducción general: Bioinformática y genómica funcional

Podemos definir la bioinformática como una ciencia interdisciplinar que consiste en la aplicación de técnicas computacionales, matemáticas y estadísticas a la clasificación y el análisis de información biológica. Desde una perspectiva a más alto nivel, la bioinfomática trata de analizar grandes cantidades de datos para lograr entender problemas tanto de biología fundamental como de aspectos biológicos relacionados con enfermedades.

Los avances en bioinformática han estado siempre ligados a los avances técnicos y tecnológicos que posibilitan la adquisición de grandes cantidades de datos biológicos. El primer gran esfuerzo tecnológico fue el *International Human Genome Project* (HGP) (Watson, 1990) que posibilitó la introducción de instrumentos de secuenciación automatizada de DNA a mediados de los 80, aunque los resultados del primer borrador del genoma humano no vieron la luz hasta el 2001 (Lander et al., 2001; Venter et al., 2001). Este gran paso en la automatización ha hecho posible el desarrollo de tecnologías capaces de obtener cantidades masivas de información dando lugar a las denominadas ciencias *ómicas*: genómica, transcriptómica, metabolómica, proteómica, etc. Se ha pasado así de estudiar entidades biológicas de manera independiente a un estudio holístico o global de todos los genes, proteínas y biomoléculas.

La bioinformática integra diferentes áreas de conocimiento como la estadística, la minería de datos y en general, la teoría de la información para desarrollar técnicas capaces de almacenar, analizar y extraer información útil de las grandes cantidades de datos generadas por las tecnologías *ómicas*. Esta ciencia integrativa puede subdividirse a su vez en diferentes áreas dependiendo del problema biológico al que buscan aportar soluciones. Así podemos definir subáreas como la **biología estructural**, centrada en el análisis y comparación de secuencias, predicción de estructura de proteínas y clasificación de estructuras 3D; la **biología de sistemas** que comprende el modelado de sistemas y procesos biológicos y el análisis de redes biomoleculares complejas; la **genómica funcional** que incluye la búsqueda y anotación de genes en genomas, los análisis de expresión génica, estudios de regulación de genes, análisis de mutaciones, etc; el **análisis de imágenes biomédicas** centrado en la automatización del procesamiento, cuantificación y análisis de imágenes clínicas especialmente orientado al diagnóstico; la **biología evolutiva y genómica comparativa** que estudia el origen y los descendientes de los sistemas vivos así como su evolución en el tiempo tomando como base la correspondencia entre genes y otras características genómicas en diferentes organismos; y la **minería de datos biomédicos** que utiliza técnicas de lingüística computacional y estadística para organizar de manera automática el conocimiento biomédico generado.

El trabajo de investigación desarrollado en esta memoria de Tesis Doctoral se encuadra fundamentalmente en la subárea de la genómica funcional. El objetivo de la genómica funcional es entender las relaciones que existen entre el genoma de un organismo, incluyendo sus genes, proteínas, funciones e interacciones, y el fenotipo de dicho organismo, es decir, la manifestación externa de dicho genotipo. Pretende expandir y sintetizar el conocimiento genómico y proteómico para proporcionar un mejor comprensión de las propiedades de un organismo a nivel celular o sistémico. Una característica fundamental de la genómica funcional es su aproximación global, que normalmente involucra técnicas de masivas de alto rendimiento en lugar de la aproximación tradicional de análisis gen a gen.

La genómica funcional estudia la variación en la abundancia de los genes y sus productos génicos, como mRNAs y proteínas, en una muestra biológica. Esta variación puede ser estudiada a lo largo del tiempo, como por ejemplo durante el desarrollo de un organismo, en diferentes lugares, como las diferencias entre distintas partes del cuerpo o tipos celulares, o en diferentes estados biológicos o patológicos que afectan a genes, cromosomas, RNA o proteínas.

En la presente memoria se ha profundizado en diferentes aspectos de la genómica funcional orientados a la mejor comprensión de características y mecanismos de desarrollo tumoral aplicados a diferentes tipos de cáncer. El **primer capítulo** se centra en el análisis de los cambios en la abundancia de los genes en múltiples estados, de manera que sea posible la diferenciación y clasificación en base a la variación de dichos genes. Un aspecto importante en el análisis de estas variaciones es la búsqueda de relaciones y asociaciones entre los genes que ayudarán en la identificación de los procesos o funciones desreguladas en cada uno de los tipos patológicos estudiados.

La posibilidad de entender cómo una mutación u otro tipo de alteración genómica conduce a la expresión de un determinado fenotipo tiene implicaciones muy importantes a la hora de entender enfermedades complejas en las que la inestabilidad cromosómica es elevada, como es el caso de la mayoría de los procesos tumorales. En el **segundo capítulo** de la presente Tesis Doctoral se estudiarán las alteraciones somáticas en el DNA, es decir, cambios en el DNA adquiridos durante la vida de un organismo que no provienen su la línea germinal. En el caso de cáncer, el análisis de estas alteraciones proporcionará información sobre el desarrollo y progresión tumoral, lo que puede ayudar a un diseño más eficaz de marcadores moleculares trasladables a la clínica.

En la mayoría de tumores, el número de alteraciones genómicas es muy elevado. Sin embargo, no todas las mutaciones tienen el mismo tipo de penetrancia ni afectan por igual en la manifestación de un determinado fenotipo. Una de las herramientas de las que dispone la genómica funcional para la discriminación de las diferentes mutaciones es la integración de datos. El análisis de las alteraciones cromosómicas con sus efectos sobre los niveles de expresión de los genes y sobre la abundancia de las proteínas codificadas permite identificar aquellas alteraciones con efectos más significativos sobre un sistema biológico estudiado. Es por ello que la integración de diferentes capas de información provenientes de estudios genómicos, transcriptómicos y proteómicos resulta mucho más útil que el estudio independiente de cada una de ellas. El **tercer capítulo** plantea la integración de datos genómicos y transcriptómicos en la identificación de los elementos desencadenantes de los procesos tumorales. Conocer cuáles son las causas de una determinada patología tumoral ayudará al desarrollo de tratamientos y terapias que mejoren la supervivencia de los pacientes que la padecen.

Por último, otro de los aspectos fundamentales de la genómica funcional es la identificación de las funciones o procesos biológicos asociados a los estados analizados, ya sean los procesos activos en cada etapa del desarrollo de un organismo, en diferentes lugares o tipos celulares o procesos desregulados en diferentes estados patológicos. En el **cuarto capítulo** se abordará la identificación y asignación de funciones y procesos biológicos a estados estudiados con técnicas genómicas.

Objetivos

El objetivo general de esta Tesis Doctoral es el desarrollo y aplicación de **algoritmos y métodos bioinformáticos** para el análisis de datos biológicos procedentes de diversas plataformas genómicas de tipo microarrays de alta densidad, así como su integración e interpretación para obtener una visión global de los genes y procesos biológicos alterados. Estos métodos se aplicarán a varios estudios experimentales concretos de **cáncer** sobre muestras humanas de pacientes.

De modo más específico, se proponen los siguientes cuatro objetivos:

1. Diseño y desarrollo de un **clasificador multiclase** para diferenciar varios tipos y subtipos patológicos, basado en análisis de microarrays de expresión génica derivados de muestras clínicas de pacientes. Integración en un algoritmo que explora los datos de expresión y los parámetros de clasificación para construir redes de genes marcadores de cada tipo o subtipo biológico estudiado. Para todo ello se utilizarán datos procedentes de microarrays de oligos alta densidad para muestras de mRNA, que miden simultáneamente la expresión de la mayoría de los genes del genoma humano.

2. Desarrollo de un método y un flujo de trabajo optimizado para el análisis cuantitativo de **alteraciones genómicas del número de copias** de DNA (*Copy Number Alterations*, CNA) así como para la detección de **puntos de ruptura** (*breakpoints*) en el genoma. Para esto se utilizarán datos procedentes de microarrays de oligos alta densidad de DNA, es decir, microarrays genómicos que detectan variaciones y polimorfismos en el genoma humano.

3. Desarrollo de un método de **análisis integrado** de datos de microarrays de expresión (mRNA) y datos de microarrays genómicos (DNA); y estudio de la correlación entre las alteraciones genómicas en **número de copias** (CN) y las alteraciones transcriptómicas de la **expresión génica** (GE).

4. Desarrollo de un algoritmo robusto para **análisis biológico funcional** basado en asociación recíproca múltiple de **genes y términos** derivados de diferentes espacios de anotación biológica. Dicho algoritmo se centrará sobre todo en la eliminación de redundancias y en la simplificación de los resultados que se obtienen por las técnicas clásicas de análisis de enriquecimiento funcional (*Functional Enrichment Analysis*).

En todos los casos, estos métodos se desarrollarán trabajando con **muestras humanas de pacientes** correspondientes a varios tipos de **cáncer** (leucemias, cáncer de colon, glioblastomas), que provienen de series experimentales publicadas y de trabajos concretos realizados en colaboración con distintos grupos clínicos y experimentales del CiC-IBMCC/USAL/HUS.

Capítulo

1

Diseño y construcción de un clasificador multiclase para subtipos patológicos basado en expresión génica: *geNetClassifier*

1.1 Introducción: Transcriptómica y perfiles de expresión génica

El transcriptoma es el conjunto total de tránscritos (RNA) de una determinada especie derivados de la transcripción de su DNA genómico. La transcripción principal corresponde a los *locus* génicos que codifican para proteínas, de los cuales se derivan los RNAs mensajeros codificantes (denominados clásicamente como mRNAs ó *protein coding RNAs*). También existen otros tipos de RNAs que no son codificantes para proteínas (denominados en general ncRNAs), cuyas funciones son muy variadas y se están caracterizando mejor en los últimos años: tRNAs, rRNAs, snRNAs, miRNAs, lncRNAs, lincRNAs, etc. En todos los organismos vivos existe una asociación directa entre la información genómica codificada en el DNA, la información transcrita a RNA y el fenotipo o características de los individuos. Las herramientas clásicas para la cuantificación y estudio de los mRNAs codificantes han sido y siguen siendo ampliamente utilizadas en los laboratorios de biología molecular, como por ejemplo las técnicas de *Northern blot*, RT-PCR, ESTs, SAGE, etc. Estas técnicas analizan el mRNA de manera individual o para un conjunto pequeño de genes; sin embargo, hasta la aparición de técnicas genómicas y transcriptómicas globales –como los microarrays– no fue posible el estudio de todo el transcriptoma a gran escala (Schena et al., 1995). Más recientemente las técnicas de secuenciación masiva del transcriptoma (principalmente RNA-seq) junto con otras técnicas globales están permitido una mejor caracterización de distintos transcritos derivados de cada *locus* génico. Esta caracterización incluye las estructuras detalladas de exones/intrones, las isoformas activas, así como la múltiple variabilidad (i.e. polimorfismos) que existe en los genomas y transcriptomas de organismos complejos como el humano.

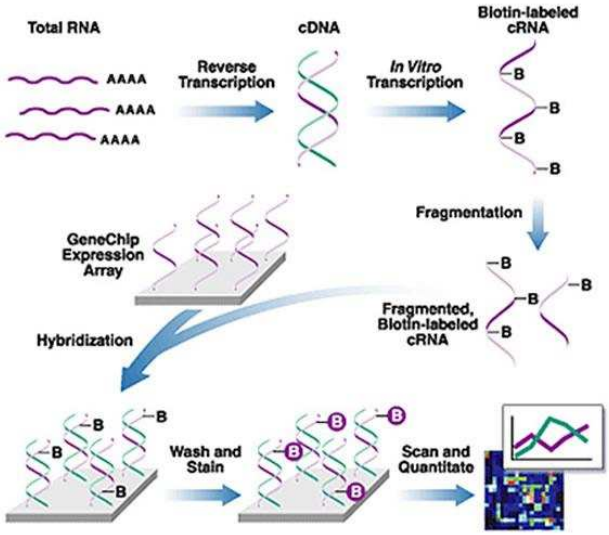


Figura 1.1: Esquema del proceso de hibridación de un microarray - El proceso de hibridación de un microarray comprende 6 fases **(i)** extracción de los mRNA con colas poli-A; **(ii)** la transcripción reversa *in vitro* para dar las cadenas de cDNA complementarias; **(iii)** el marcaje con moléculas fluorescentes; **(iv)** la fragmentación de las secuencias; **(v)** la hibridación en el microarray y lavado de restos no unidos; y **(vi)** la lectura de la fluorescencia con un *scanner* de alta resolución. (Fuente: www.affymetrix.com)

Los estudios de expresión global (llamados a menudo estudios de genómica funcional) se centran primeramente en el análisis de los genes codificantes y sirven para construir perfiles de expresión de los genes conocidos del genoma (i.e. "*gene expression profiles*") en las muestras biológicas en las condiciones estudiadas. En este capítulo nos centraremos principalmente en la construcción y análisis de este tipo de perfiles en muestras clínicas obtenidas de pacientes con distintas patologías.

Desde el punto de vista tecnológico, en la caracterización de los perfiles de expresión en diferentes tipos celulares, tejidos o fenotipos patológicos los dispositivos microarrays y las técnicas de ultrasecuenciación –i.e. nuevas técnicas de secuenciación masiva, NGS– han demostrado obtener resultados coherentes (Malone and Oliver, 2011). En comparación con las técnicas de ultrasecuenciación, los microarrays de oligos –especialmente los de la plataforma *Affymetrix, Inc.*– han sido usados en numerosos estudios biomédicos publicados. El mayor grado de madurez de los microarrays de oligos, de los que se conocen y han sido estudiadas las posibles fuentes de ruido y desviación para las que se han desarrollado toda una serie de algoritmos y técnicas estadísticas robustas de determinación de señal y correción de ruido, hacen que estos microarrays sigan siendo una fuente de datos muy utilizada. También, hoy por hoy, la diferencia de precios entre las dos tecnologías –microarrays *versus* ultrasecuenciación– hace que para el análisis de series largas, con más de 20 o 30 muestras, los arrays sigan siendo una metodología más asequible a la hora de obtener perfiles de expresión génica.

1.1.1 Microarrays para medir la expresión génica global

Los microarrays de oligonucleótidos de DNA son dispositivos que incluyen colecciones de miles de moléculas de oligonucleótidos de DNA de un organismo concreto inmovilizadas sobre un soporte sólido formando una micromatriz en la que la localización de cada oligo es conocida. Estos oligos con una secuencia específica de DNA son conocidos como sondas. El primer paso para la utilización de los microarrays de expresión es la extracción del mRNA de una muestra. Este mRNA es fragmentado y marcado con una molécula fluorescente antes de ser depositado sobre la superficie del array. Una vez depositados los fragmentos marcados, éstos hibridarán con las sondas complementarias inmovilizadas. De esta manera, cuando el array es iluminado las moléculas fluorescentes de los fragmentos que han hibridado en determinadas regiones o puntos de la micromatriz emiten luz que puede ser escaneada y cuantificada (Figura 1.1).

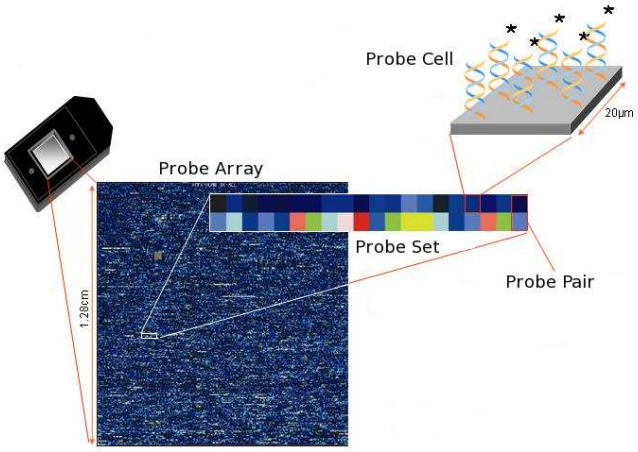


Figura 1.2: Diseño de un microarray *GeneChip* de *Affymetrix* - Esquema que representa la estructura del array Human Genome U133A de Affymetrix. Este array incluye conjuntos de 11 pares de sondas agrupados en un *probeset* que representa un mRNA. Cada par de sondas está formado por un PM y un MM cada uno con miles de copias en una única celda. (Fuente: www.affymetrix.com)

El soporte en el que se inmovilizan las sondas depende de la compañía que fabrica los arrays, siendo cristal o silicio para *Affymetrix* (http://www.affymetrix.com) o *Agilent* (http://www.agilent.com) y microesferas para *Illumina* (http://www.illumina.com). Los más populares y más usados entre los arrays de oligonucleótidos son los *GeneChip* desarrollados por *Affymetrix* (Irizarry et al., 2003a). Cada uno de estos microarrays contiene entre 40 y 60000 conjuntos de sondas (llamados *probesets*) con secuencias de todo el transcriptoma de la especie estudiada, representando en el caso de humano unos 25000 genes. Cada conjunto de sondas (*probeset*) está a su vez constituído por entre 11 y 16 sondas de oligonucleótidos distintas que corresponden a distintas regiones codificantes del gen que representan. Finalmente, para cada sonda de secuencia específica existe un oligo denominado *perfect match* (PM, i.e. oligo de 25 nucleótidos que se corresponde exactamente con una sección de la molécula de mRNA de interés) junto a un oligo llamado *mismatch* (MM, i.e. oligo de 25 nucleótidos que se construye cambiando el nucleótido central de la secuencia del PM). El propósito de estas sondas MM de control es permitir la cuantificación del ruido de fondo y la estimación de hibridaciones inespecíficas.

Un ejemplo de este tipo de arrays es el *Human Genome U133A* de *Affymetrix*, cuyo diseño puede verse en la figura 1.2. En las nuevas versiones de los arrays este diseño con MM y PM ha sido sustituido por sondas PM únicamente.

Entre los principales objetivos para los que se utilizan este tipo de microarrays destacan:

1. La búsqueda de genes que sufren cambios significativos en sus niveles de expresión en dos estados diferentes. Es decir, análisis de la expresión diferencial entre dos estados.
2. La clasificación de muestras basada en los perfiles de expresión de sus genes. Es decir, construcción de perfiles de expresión normalizados para cada muestra a lo largo de todos los genes y búsqueda de grupos de muestras similares de acuerdo a esos perfiles.
3. La búsqueda de patrones o grupos de genes relacionados que cambian su expresión de manera conjunta en una serie de muestras o condiciones. Es decir, construcción de perfiles de expresión normalizados para cada gen a lo largo de las muestras y búsqueda de pares de genes que co-expresan de acuerdo a esos perfiles.

1.1.2 Análisis de perfiles de expresión génica para clasificación de muestras

Como se ha indicado en el apartado anterior, los microarrays que miden la expresión génica global pueden ser utilizados para la clasificación de muestras de estados o subestados biológicos concretos. Para ello se debe proceder a la construcción de perfiles de expresión normalizados para cada muestra a lo largo de todos los genes y, basados en esos perfiles, buscar grupos de muestras similares.

En este capítulo presentaremos el diseño y construcción de un clasificador para múltiples clases que permita diferenciar tipos y subtipos biológicos o patológicos basado en datos de expresión génica. El algoritmo se denomina *geNetClassifier* y está diseñado para construir clasificadores transparentes y para obtener las redes de genes asociadas a cada clase a partir de datos procedentes de microarrays de expresión global. El método se basa en técnicas de aprendizaje automático (*machine learning*) que permiten extraer conocimiento de un conjunto de muestras y extrapolar tal conocimiento a la clasificación de nuevas muestras o individuos problema. Esta clasificación se construye de modo que no sea una caja negra de toma de decisiones, sino que proporcione información acerca de la influencia o valor de cada gen en la clasificación. De este modo, los genes seleccionados contituirán el punto de partida para la identificación de marcadores moleculares de los estados estudiados.

En las siguientes secciones se introducen brevemente conceptos clave en el desarrollo del algoritmo; se comentan las diferentes técnicas de aprendizaje automático alternativas para la construcción de clasificadores y se describe someramente la técnica empleada en el algoritmo (Máquinas de Vector Soporte, SVM), así como el método Bayesiano Empírico Paramétrico (BEP) utilizado en la selección de genes. Finalmente, se detalla también cada uno de los pasos que constituyen el algoritmo *geNetClassifier* y se presenta como ejemplo de aplicación la clasificación y caracterización de 4 tipos de leucemia diferentes.

1.2 Métodos

1.2.1 Métodos de aprendizaje automático: SVM

La disciplina del aprendizaje automático o *machine learning* incluye técnicas y metodologías destinadas a la búsqueda y establecimiento de patrones que permitan identificar las características de un estado o un evento a partir de un conjunto estructurado de datos empíricos medidos para dicho evento. El establecimiento preciso de estos patrones permitirá luego la realización de predicciones automáticas sobre el estado o evento en cuestión.

Los problemas de predicción y búsqueda de patrones son frecuentes en el análisis de datos de alta dimensionalidad. Éste es el caso de los datos procedentes de microarrays en los que se mide la expresión de miles de genes simultáneamente en un conjunto de muestras. Diversos algoritmos han sido aplicados con éxito a la búsqueda de genes marcadores de estado y a la consecuente clasificación de estados biológicos y patológicos. Algunos de los algoritmos de clasificación más utilizados en estudios que analizan datos genómicos o transcriptómicos son: **(i)** Análisis lineal discriminante (*Linear Discriminant Analysis*, LDA); **(ii)** K-vecinos más próximos (*K-Nearest Neighbours*, k-NN); **(iii)** Redes neuronales (*Neural Networks*, NN); y **(iv)** Máquinas de Vector Soporte (*Support Vector Machines*, SVM). Una amplia revisión que incluye los fundamentos y descripción de estos métodos así como su aplicabilidad a estudios genómicos se puede encontrar en (Gentleman et al., 2005).

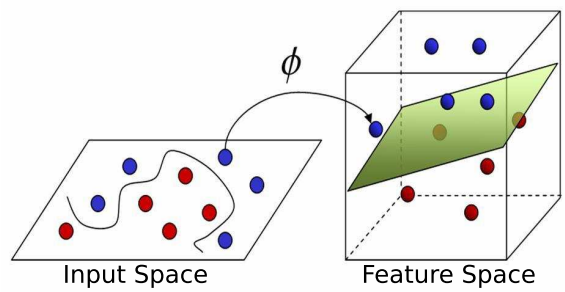


Figura 1.3: Transformación del espacio de entrada de SVM - La aplicación de una función de transformación ϕ del espacio de entrada en un espacio de características de más dimensiones permite una separación lineal de las clases. (Modificado de (Keyvanrad and HomayounpourM.M, 2011))

La mayoría de los métodos de aprendizaje automático citados presentan problemas con datos de alta dimensionalidad. Sin embargo, SVM permite trabajar sin problemas con este tipo datos. Esta característica unida a los rendimientos demostrados en el análisis de datos procedentes de microarrays de expresión (Brown et al., 2000; Furey et al., 2000) hacen de SVM una metodología adecuada para el tratamiento de los perfiles de expresión génica.

Las máquinas de vector soporte son un conjunto de algoritmos de aprendizaje automático supervisado destinados a la clasificación. El objetivo de estas técnicas es obtener el hiperplano óptimo capaz de separar dos clases. Muchos problemas no son separables mediante un hiperplano de manera sencilla por lo que SVM realiza una transformación previa de los vectores de entrada (n-dimensionales) en vectores de dimensión más alta en los que el problema de separación pueda resolverse linealmente. Es decir, los objetos del espacio de entrada son mapeados a un nuevo espacio de características que tiene mayor dimensionalidad utilizando un conjunto de funciones (*kernels*), de tal forma que en lugar de buscar una curva compleja que los separe puedan ser separados por medio de un hiperplano 1.3.

De todos los hiperplanos posibles SVM elige aquel que maximiza la separación entre las dos clases. Este hiperplano es definido por medio de una serie de instancias de entrenamiento que actúan como límites, denominadas vectores soporte (*support vectors*).Las nuevas instancias se clasifican de acuerdo al lado del hiperplano en el que se encuentran. Una representación del hiperplano de separación con los vectores soporte se puede ver en la figura 1.4.

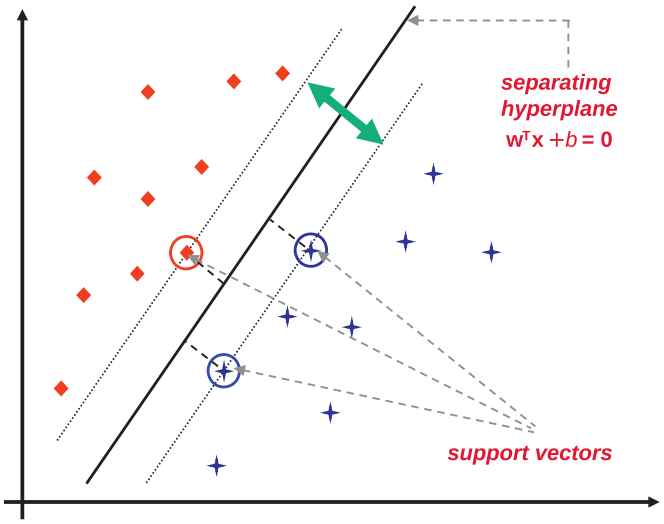


Figura 1.4: Vectores soporte para SVM - El método SVM binario selecciona el hiperplano (línea en negrita) que maximiza el margen entre las dos clases. Este hiperplano es determinado por ciertas instancias utilizadas en el entrenamiento que sirven de frontera (vectores soporte marcados con círculos).

Las SVM fueron inicialmente diseñadas para la clasificación binaria, aunque han sido extendidas a la clasificación de múltiples clases utilizando dos aproximaciones fundamentales (Statnikov et al., 2005):

1. *One-Versus-Rest* (OVR): Es conceptualmente la SVM multiclase más sencilla. Se basa en la construcción de k clasificadores binarios: clase 1 frente al resto, clase 2 frente al resto, ... clase k frente al resto. La función de decisión global elige la clase que se corresponde con el valor máximo de las k funciones de decisión binarias (*winner-takes-all*).
2. *One-Versus-One* (OVO): Este método implica la construcción de SVM binarios para cada par de clases, en total
$$\binom{k}{2} = \frac{k(k-1)}{2}$$
La función de decisión global asigna cada instancia a la clase que tiene el mayor número de votos según una estrategia de mayorías (*max wins strategy*).

Se ha demostrado que si el problema multiclase es no-separable y algunos de los subproblemas binarios son separables el rendimiento de clasificación OVO es mejor que el rendimiento de clasificación OVR (Kressel, 1999). En la figura 1.5 se muestra un ejemplo de ambas aproximaciones.

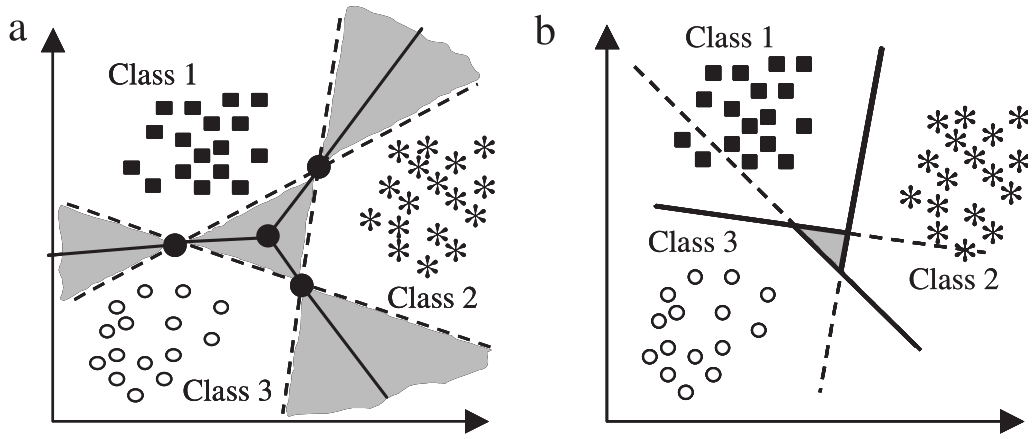


Figura 1.5: Esquemas alternativos para SVM multiclase - SVM multiclase aplicado a la separación de 3 clases. **(a)** OVR construye 3 clasificadores que separan cada clase individualmente del resto de las clases. **(b)** OVO construye también 3 clasificadores, pero separa clase 1 frente a clase 2, clase 1 frente a clase 3 y clase 2 frente a clase 3. (Modificado de (Statnikov et al., 2005))

1.2.2 Selección de variables para la clasificación

Los análisis de expresión procedentes de datos de microarrays parten típicamente de un conjunto de decenas o cientos de muestras en las que se han medido varios miles de genes (en torno a 20000 genes). Debido a la alta dimensionalidad de este tipo de datos, en los algoritmos de aprendizaje supervisado, y más concretamente en los de clasificación, es necesario hacer una pre-selección de los genes como variables para evitar el problema del sobreajuste o sobreentrenamiento (*over-training*) (Guyon et al., 2002). Para evitar el sobreajuste, la reducción del número de variables (*Feature Subset Selection*) (Bell and Wang, 2000) se ha convertido en un paso requerido en el entrenamiento de clasificadores con datos de muy alta dimensión.

Aunque no hay un método específico que se haya demostrado el mejor para realizar esta tarea, las aproximaciones son fundamentalmente de dos tipos:

1. Indirecta o *filter*: hace uso de heurísticos para determinar el subconjunto de variables óptimo, es decir, establece una regla matemática que es capaz de guiar el proceso de búsqueda hacia una solución optimizada.
2. Directa o *wrapper*: cada posible subconjunto de variables candidato es evaluado directamente en el modelo de clasificación de modo exhaustivo.

La exploración de todas las posibles combinaciones de variables es computacionalmente muy costosa y no mejora en exceso la exactitud o precisión del clasificador (Inza et al., 2004). Esto ha propiciado que la aproximación de filtrado (i.e. el método indirecto) esté mucho más extendida para la selección de genes en la clasificación de datos de microarrays. Dentro de las posibles estrategias de filtrado para ayudar a la selección la más común es ordenar los genes en base a análisis de expresión diferencial entre clases, lo cual permite ordenar los genes por los valores de cambio entre clases o por la significación estadística de tales cambios. A continuación se describe el método que usamos en este trabajo para generar ese orden de genes (*gene rankig*).

1.2.2.1 **Método bayesiano empírico paramétrico (PEB) para generar un orden de genes**

El ordenamiento o *ranking* de los genes que constituye el primer paso de la aproximación de filtrado para la selección de variables se realiza en base a la expresión diferencial de cada uno de los genes en cada clase analizada. La aproximación estándar para la búsqueda de genes diferencialmente expresados consiste en probar una hipótesis para cada gen utilizando variantes de los estádisticos *t* o *F*. Dada la muy elevada cantidad de variables testadas, los p-valores proporcionados por estos estadísticos deben ser corregidos con alguno de los métodos de corrección para contrastes múltiples (*multiple testing*) (Tusher et al., 2001).

Por otro lado, el número de muestras o de réplicas biológicas disponibles es habitualmente limitado y muchas veces no es suficiente a la hora de estimar las diferencias en los niveles de expresión. Por ello, si en lugar de tratar cada gen independientemente se analizan los niveles de expresión del conjunto de genes simultáneamente es posible extraer información acerca de la variabilidad de estos genes. Los métodos bayesianos empíricos (*Empirical Bayes*, EB) utilizan la variabilidad del conjunto de genes y por tanto se adaptan fácilmente a problemas en los que el conjunto de variables es más alto que el número de muestras o individuos. En general los métodos paramétricos clásicos se centran en el control de la proporción esperada de falsos positivos (*False Discovery Rate*, FDR) para dar un valor de significación a cada variable gen, mientras que los métodos EB logran el ordenamiento o *ranking* de las variables genes en base a sus patrones de expresión utilizando probabilidades posteriores.

El primer método bayesiano utilizado con datos de expresión génica fue un método bayesiano empírico paramétrico basado en modelos jerárquicos que fue empleado para detectar cambios en la expresión génica para un único array de cDNA (Newton et al., 2001). Este método fue extendido para múltiples condiciones o clases (Kendziorski et al., 2003) y, a continuación refinado para evitar la asumpción de un coeficiente de variación constante para todos los genes (Lo and Gottardo, 2007). Esta última formulación incluye una estimación de los parámetros del modelo con un algoritmo de esperanza-maximización (*Expectation-Maximization*, EM). El paquete de Bioconductor (www.bioconductor.org) *EBarrays* proporciona una implementación de estos métodos en R.

El método bayesiano empírico paramétrico (*Parametric Empirical Bayesian method*, PEB) calcula las probabilidades bayesianas o probabilidades posteriores de patrones de expresión diferencial para múltiples condiciones o clases. Este método busca inicialmente un modelo que caracterice la distribución de probabilidad de las medidas de expresión de un gen j en un conjunto de muestras $x_j = (x_{j1}, x_{j2}, \ldots, x_{jI})$. Con la hipótesis de partida de que las muestras son intercambiables

podemos plantear las medidas x_{ij} como desviaciones aleatorias de una distribución observada $f_{obs}(\cdot \mid \mu_j)$.

Cuando se comparan los valores de expresión en dos estados, el conjunto de muestras $1, 2, \ldots, I$ se divide en dos subconjuntos s_1 y s_2. Se pueden plantear entonces dos hipótesis: la hipótesis nula inicial en que la media de los valores de expresión para cada grupo sea la misma ($\mu_{j1} = \mu_{j2}$), es decir, expresión equivalente EE_j, o que las medias sean distintas ($\mu_{j1} \neq \mu_{j2}$), es decir, exista expresión diferencial DE_j.

Un gen EE tendrá valores $x_j = (x_{j1}, x_{j2}, \ldots, x_{jI})$ de acuerdo a una distribución

$$f_0(x_j) = \int \left(\prod_{i=1}^{I} f_{obs}(x_{ij} \mid \mu) \right) \pi(\mu) d\mu \tag{1.1}$$

Y para un gen DE los valores $x_j = (x_{j1}, x_{j2})$ siguen una distribución

$$f_1(x_j) = f_0(x_{j1}) f_0(x_{j2}) \tag{1.2}$$

Si p es el porcentaje de genes DE, entonces $1 - p$ es el porcentaje de genes EE y la distribución marginal de los datos sería:

$$p f_1(x_j) + (1 - p) f_0(x_j) \tag{1.3}$$

Aplicando el teorema de Bayes la probabilidad posterior de expresión diferencial del gen j se obtiene mediante la ecuación:

$$\frac{p f_1(x_j)}{p f_1(x_j) + (1 - p) f_0(x_j)} \tag{1.4}$$

Para extender la aplicación a más de dos condiciones basta con generalizar 1.3 para m condiciones

$$\sum_{k=0}^{m} p_k f_k(x_j) \tag{1.5}$$

1.2.3 **Normalización y obtención de la señal sumarizada para cada gen**

El preprocesamiento de los microarrays de expresión constituye un paso fundamental que repercute directamente sobre todos los análisis posteriores. Es por ello necesaria una exploración de los métodos de normalización y obtención de la señal sumarizada para cada gen del microarray. Es importante también tener en cuenta cuáles son las entidades biológicas que van a constituir el punto de partida de los algoritmos, y, si es necesario, redefinirlas para adecuarlas a los objetivos buscados.

La cuantificación de la cantidad de mRNA en una muestra es calculada a partir de la señal fluorescente de hibridación obtenida del microarray. Esta señal está afectada por la fluorescencia inespecífica del propio array, que será necesario corregir para eliminar el ruido de fondo. Existen también otras fuentes de variación que pueden influir en las mediciones de los niveles de expresión como pueden ser: diferencias en la cantidad de mRNA total hibridado en distintas muestras, eficiencia de la hibridación, propiedades ópticas del *scanner*, etc.

A la hora de comparar microarrays es necesario tener en cuenta estas fuentes de variación en cada uno de ellos. La normalización es el término para describir la eliminación de esa variación. Es

decir, se trata de eliminar el efecto de la tecnología o de artefactos en las mediciones de fluorescencia. También se debe tener en cuenta que para calcular la intensidad del conjunto de sondas (*probeset*) que miden cada gen en el microarray es necesario integrar o sumarizar el conjunto de medidas diferentes que corresponden al mismo mRNA.

Para la ejecución de estos 3 pasos (eliminación del ruido de fondo, normalización y sumarización o integración de las sondas del mismo *probeset*/gen) exiten diferentes métodos. Algunos de los algoritmos más utilizados son:

1. MAS5 (Liu et al., 2003): Es el método original propuesto por *Affymetrix*. Es un método paramétrico que evalúa cada array de manera independiente y utiliza tanto las sondas PM (*Perfect Match*) como las MM (*MisMatch*).
2. RMA (Irizarry et al., 2003b): Utiliza la información del conjunto de arrays en lugar de evaluar array por array e incluye una normalización no paramétrica basada en cuantiles. Utiliza únicamente las sondas PM.
3. dChip (Li and Wong, 2001): También multiarray aunque optimizado para réplicas técnicas. Hay disponibles dos variantes, una que utiliza únicamente PM y otra que utiliza tanto PM como MM.
4. PLIER (Affymetrix, 2005): Desarrollado por *Affymetrix* para mejorar el anterior, MAS5. Está basado en la generación de modelos para el análisis de múltiples muestras simultáneamente. Permite la utilización de modelos para PM y MM y para PM únicamente.

La diferencia en los valores de sumarización obtenidos con estos métodos para las mismas muestras unido a la complejidad de la evaluación de los algoritmos ha motivado una gran cantidad de trabajos orientados a la comparación de las diferentes alternativas, (Irizarry et al., 2003a, 2006; Millenaar et al., 2006; Qin et al., 2006) entre otros. De estos estudios se puede extraer que, en general, los algoritmos que utilizan únicamente PM tienen una menor varianza y por tanto mayor precisión. Este hecho unido a la buena correlación de los datos normalizados con RMA con datos de RT-PCR ha motivado la decisión de utilizar RMA para el preprocesamiento de datos de microarrays de expresión (Millenaar et al., 2006).

1.2.4 **Redefinición de las sondas de medida de los microarrays a genes**

Además de las fuentes de ruido comentadas anteriormente otro aspecto que afecta al análisis de los microarrays de expresión se deriva de la definición arbitraria de conjuntos de sondas (*probesets*) para la medida de los genes. La mayoría de los análisis de datos procedentes de microarrays identifican como genes un conjunto predeterminado de sondas de oligonucleótidos definidas por el fabricante. La arquitectura conocida de muchos genes humanos está en constante evolución y va cambiando con el tiempo al aumentar el conocimiento, con lo que la referencia para los genes y transcritos utilizada en la definición de los conjuntos de sondas del microarray no se corresponde con la versión más actualizada y refinada del genoma. Esta situación se ve agravada con la existencia de ciertos conjuntos sondas (*probesets*) definidos como distintos pero que mapean a un mismo gen, provocando redundancia. Todo ello aumenta bastante la imprecisión que la asignación de los valores de expresión a los genes.

Para evitar esta fuente de ruido adicional se han redefinido los genes re-mapeando cada una de las sondas del microarray sobre una versión actualizada del transcriptoma humano obtenida de *Ensembl* (http://www.ensembl.org). De esta manera cada gen queda definido por todo el conjunto de sondas del array que mapean de manera exclusiva en los límites del locus génico conocido correspondiente. Esta redefinición se ha llevado a cabo utilizando los ficheros proporcionados en

la sección *Probe Mapping* de la herramienta bioinformática *GATExplorer* (Risueno et al., 2010) (http://bioinfow.dep.usal.es/xgate/mapping/mapping.php).

1.3 Desarrollo del algoritmo de clasificación múltiple: *geNetClassifier*

El análisis de los perfiles de expresión de genes en distintos tipos de tejidos, bajo diferentes condiciones o en distintos estados patológicos está fundamentalmente orientado a la búsqueda de marcadores biológicos para dichas situaciones. Estos marcadores biológicos serán aquellos genes que permitan distinguir, bien solos o bien en combinación con otros, los diferentes estados o clases comparadas.

En esta sección se presenta *geNetClassifier*, un algoritmo diseñado para la búsqueda de genes marcadores que permitan diferenciar estados patológicos. El algoritmo diseñado se centra en la búsqueda de estas firmas moleculares, sin embargo analiza también las relaciones existentes entre los genes característicos identificados. La figura 1.6 presenta un esquema con los pasos fundamentales del algoritmo junto con los diferentes tipos de información o resultados que se obtendrán a partir de los perfiles de expresión.

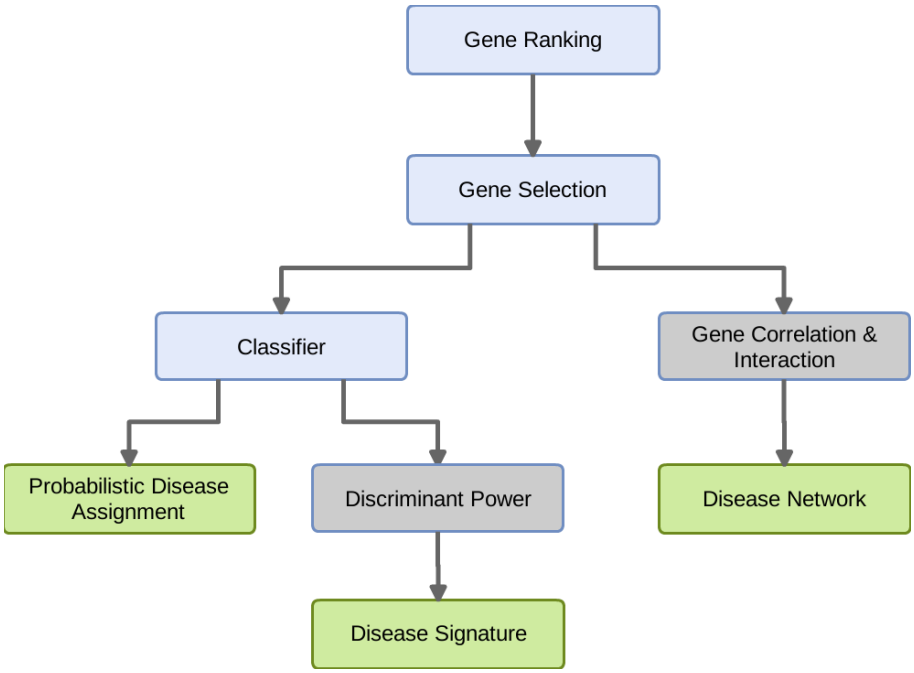


Figura 1.6: Partes del algoritmo *geNetClassifier* - Representación esquemática de los pasos fundamentales del algoritmo. En verde aparecen resaltados los principales resultados buscados y en gris las métricas o características en las que se apoyan.

El primer paso de *geNetClassifier* consiste en el establecimiento de un *ranking* de los genes basado en las diferencias de expresión entre los estados. Este *ranking* de genes corresponde, como se ha indicado en la sección anterior, a la ordenación de variables previa a la exploración y selección que realizará el clasificador. De este modo, el *ranking* sirve como punto de partida para la selección del número de genes óptimo que permita distinguir un estado del resto de los estudiados. Esta selección de genes propios de cada estado se utilizará fundamentalmente en la obtención de tres tipos de resultados:

1. *Probabilistic Disease Assigment* o asignación de nuevos individuos a cada una de las clases o estados estudiados.
2. *Disease Signature* o firma molecular característica de cada uno de los estados patológicos. El proceso de clasificación no será únicamente una caja negra de toma de decisiones sino que proporcionará información acerca de la importancia en la clasificación de cada una de las variables exploradas, es decir, su poder discriminante (*Discriminant Power*). Así, las variables más importantes, es decir, los genes con mayor influencia, constituirán el punto de partida para el análisis de marcadores de los estados estudiados.
3. *Disease Networks* o redes de asociación entre los genes candidatos a ser marcadores para cada uno de los estados patológicos. La selección de variables proporcionará el punto de partida para el estudio de las relaciones entre las variables o genes que se realizará mediante dos tipos de análisis:
   (a) análisis de la relación entre genes mediante medida de la coexpresión (*gene coexpression*).
   (b) análisis de la interacción entre genes mediante medida de la información mutua entre ellos (*mutual information*).

En las siguientes subsecciones se detallan y comentan cada uno de los pasos del algoritmo desarrollado.

1.3.1 *Ranking* de genes utilizando PEB

Como se ha comentado en la sección 1.2.2, el elevado número de variables o características comparado con el número de muestras en los datos procedentes de microarrays de expresión hace necesaria una reducción de las mismas para evitar el sobre-entrenamiento cuando se utilizan algoritmos de aprendizaje supervisado. Para ello se ha optado una aproximación de filtrado. En la implementación desarrollada a cada una de las variables o genes se le asigna un coeficiente de relevancia que establece un *ranking* entre ellas. Así las variables predictivas quedan ordenadas respecto a las clases seleccionándose las k primeras para inducir con ellas. En las primeras posiciones del *ranking* aparecerán las variables que despejan una mayor cantidad de incertidumbre en el problema, mientras que en las zonas finales estarán aquellos atributos sin aparente relación con el problema abordado.

Para establecer el *ranking* de los genes se utiliza el método Bayesiano Empírico Paramétrico (*Parametric Empirical Bayes method*, PEB) (Lo and Gottardo, 2007) basado en el análisis de la señal de expresión. Los fundamentos teóricos de este método han sido presentados en la sección 1.2.2.1. PEB busca los genes que presentan una expresión diferencial significativa cuando se comparan las muestras en un estado frente a los demás (*One Versus Rest*, OVR) y devuelve las probabilidades posteriores de expresión diferencial para cada uno de los genes en cada estado. Además el algoritmo implementado para *geNetClassifier* calcula la diferencia de la media de expresión de cada gen en cada estado respecto al resto. Utilizando estos dos estadísticos, la probabilidad posterior y la diferencia de medias, los genes son ordenados en base a su relevancia.

Para lograr una mayor especificidad el *ranking* se construye con la condición de que cada gen no puede estar presente en más de una lista siendo asignado únicamente a la lista del estado en el que es más relevante. Con ello se consigue que no haya solapamiento entre los genes de cada estado.

1.3.2 **Selección de genes como variables para la clasificación**

El objetivo de este paso del algoritmo es la selección del número de genes mínimo que permita construir el clasificador óptimo. El problema se centra en el establecimiento de un punto de corte que seleccione un número concreto de los genes con mejores posiciones en el *ranking* para ser utilizados como variables en la construcción del clasificador y lograr la menor tasa de error posible.

El algoritmo diseñado para establecer el número mínimo de genes que proporcionan el menor error de clasificación consiste en la construcción de clasificadores con un número creciente de genes, comenzando con un gen para cada clase y añadiendo un gen más en cada paso tomado por orden en el *ranking*. Para cada uno de los clasificadores construidos se evalúa la tasa de errores utilizando validación cruzada. El número de genes seleccionado se corresponderá con aquel que alcance la tasa de error mínima. Este algoritmo aparece representado en la figura 1.7.

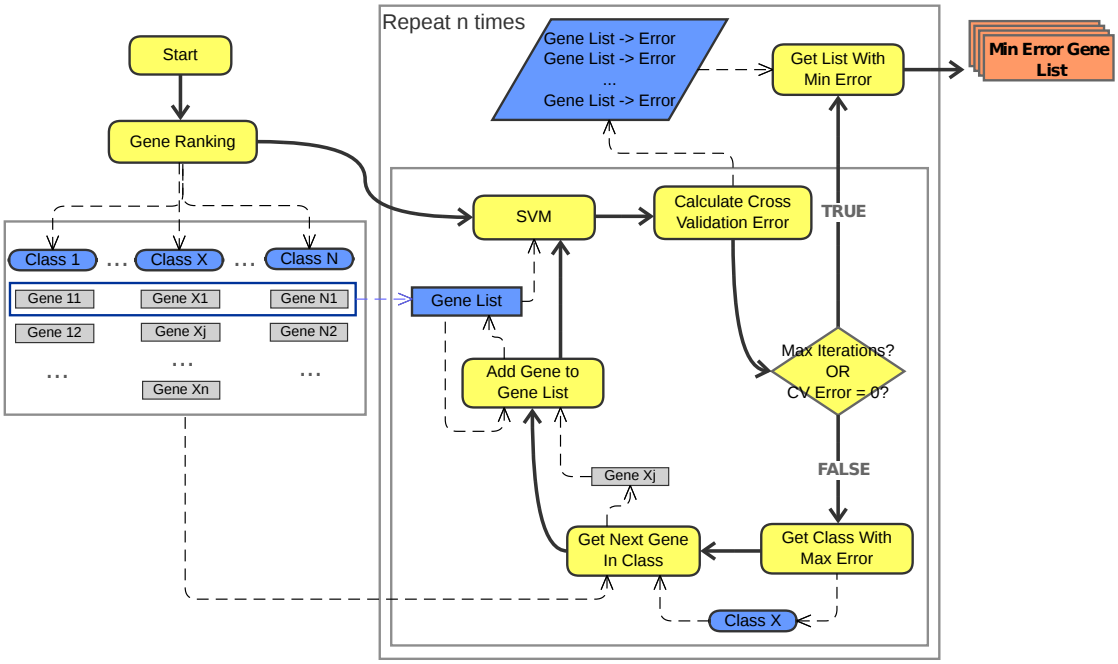


Figura 1.7: Esquema con el proceso de selección de genes - Las flechas continuas representan el flujo de ejecución y las flechas discontinuas representan el acceso o la escritura de datos.

El número de genes seleccionado varía ligeramente en diferentes ejecuciones debido a la aleatorización de las muestras en la validación cruzada. Para lograr una mayor estabilidad todo el proceso se repite un número n de veces ($n = 6$ por defecto). Finalmente el número de genes óptimo elegido será el mayor número de genes seleccionado en cada una de las iteraciones, excluyendo aquellos considerados atípicos (*outliers*). El hecho de tomar este número de genes en lugar de la media o el número mínimo aumenta la sensibilidad del predictor y permite la segregación de muestras que, con un número menor de genes, no habrían podido ser correctamente clasificadas.

Un ejemplo del número de genes seleccionado en cada iteración para cada clase se puede ver en la figura 1.8 que corresponde a la clasificación de muestras en cinco clases: cuatro tipos de leucemias y una quinta clase normal no leucemia. Las barras sólidas representan el número de genes finalmente elegido para cada una de las clases: 4 para ALL, 11 para AML, 1 para CLL, 8 para CML y 9 para NoL. Este número de genes serán tomados por orden del *ranking* establecido y serán las variables utilizadas en la construcción del clasificador final.

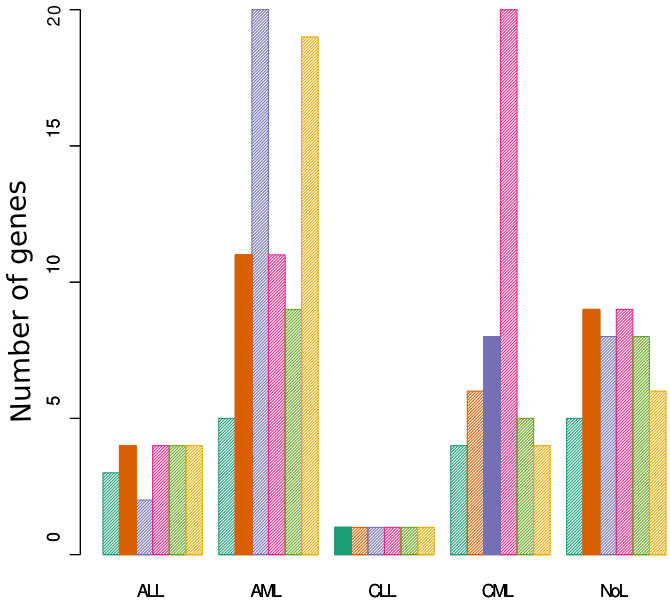


Figura 1.8: Número de genes con tasas de error mínimas - Las barras representan para cada clase el número de genes que han proporcionado las tasas de error mínimas en 6 iteraciones del algoritmo. Cada color representa una iteración y las barras sólidas el número de genes final seleccionado.

1.3.3 Construcción del clasificador y búsqueda de genes marcadores

El método de aprendizaje automático seleccionado para la clasificación ha sido SVM. Las principales características que han llevado a esta elección han sido comentadas en la sección 1.2.1. Se ha utilizando la implementación para multiclase *One-Versus-One* (mcSVM-OvO) con *kernel* lineal proporcionada en el paquete de R *e1071*.

El método SVM es transparente, es decir la utilización de este tipo de clasificador nos permite obtener información acerca del papel de cada gen (cada variable) en la clasificación y la importancia del mismo en el establecimiento de las fronteras entre las diferentes clases. De esta manera podemos ahondar en la biología subyacente a la clasificación.

Al utilizar una aproximación OvO, la SVM define un conjunto de vectores soporte para la separación de cada par de clases. Estos vectores soporte incluyen los coeficientes de *Lagrange* para cada coordenada o cada gen. Utilizando una combinación de estos coeficientes podemos obtener una medida del poder discriminatorio de cada gen respecto a cada par de clases. Una representación de este poder o valor discriminante para el gen VPREB1 puede verse en la figura 1.9 donde cada barra representa los coeficientes de *Lagrange* apilados en cada clase. De este modo, se puede definir el parámetro ”poder discriminante”(*discriminant power*) de cada gen como la diferencia entre las clases con los valores máximos de la suma sus coeficientes de Lagrange. En el gráfico aparace representado como la diferencia entre la mayor barra (que indica la clase que mejor marca) y la siguiente barra más cercana. Esta distancia se corresponde a la separación entre las dos líneas rojas marcadas en la figura 1.9.

Este *poder discriminante* es un nuevo parámetro clave que nos permite identificar los genes que mejor distinguen las clases de entre todos los utilizados en la clasificación y, de este modo, aporta información adicional respecto a los genes seleccionados.

1.3.3.1 Aplicación del clasificador para la asignación a clase

Una vez seleccionados los genes que se utilizarán como variables en el clasificador, éste es entrenado con las muestras disponibles para la identificación del estado o fenotipo de un nuevo individuo

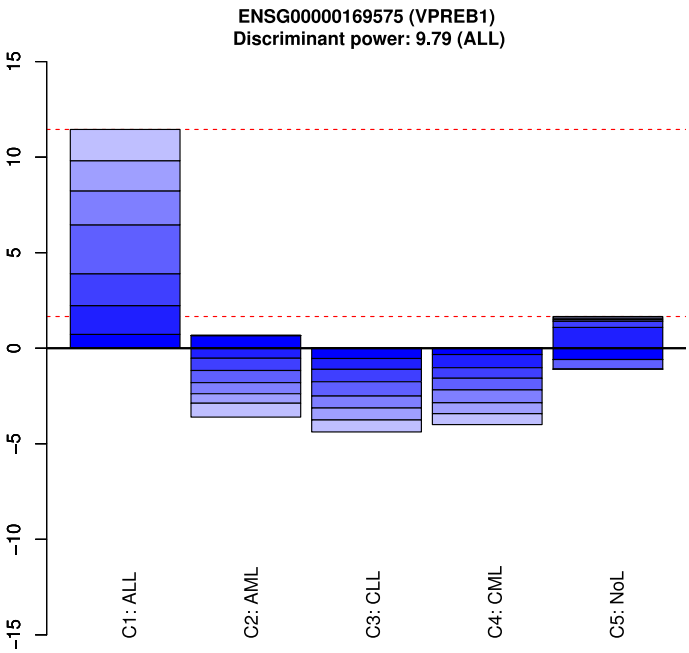


Figura 1.9: Poder discriminante - Representación del poder discriminante para el gen VPREB1 como la diferencia entre los dos valores mayores de la suma máxima de los coeficientes de Lagrange de los vectores soporte para cada clase.

o muestra. El método SVM utilizado proporciona las probabilidades de asignación de una muestra problema a cada una de las clases, sin embargo, muchas veces resulta difícil distinguir a cuál de las clases es posible asignar la muestra. Por ello se ha desarrollado dentro del algoritmo de clasificación un paso que tiene en cuenta estas probabilidades de asignación, pero también el número de clases a distinguir y la probabilidad de confusión con el resto.

Para asignar una muestra a una clase determinada se deben cumplir dos condiciones:

1. La probabilidad de asignación proporcionada por el clasificador deber ser al menos el doble que la probabilidad al azar.
2. La diferencia de probabilidad con la siguiente clase más probable debe ser también mayor que 0.8 veces la probabilidad al azar.

Por ejemplo, para asignar una muestra a una determinada clase en un clasificador que trata de discernir entre 5 clases, la probabilidad de asignación debe ser al menos del 40 % (2∗20 %) y la mínima probabilidad de confusión con otras clases tiene que ser del 16 % (0.8 ∗ 20 %). Es decir, si la probabilidad de asignación a 2 de las clases es del 50 % y del 40 % respectivamente, esa muestra no será asignada. Esta situación es equivalente a la acción de un experto cuando no emite un juicio debido a que no está seguro.

Utilizando esta estratégia no todas las muestras serán identificadas o incluidas dentro de alguna de las clases, sin embargo se logra que el número de falsos positivos o error de tipo I (*False Positives*, FP) disminuya. El objetivo buscado es un equilibrio entre los FP y el *call rate* o porcentaje de asignación de las muestras:

$$\text{Call rate} = \frac{\text{Asignados}}{\text{Asignados} + \text{No Asignados}} \tag{1.6}$$

La figura 1.10 muestra una representación de la probabilidad de asignación a la clase más probable frente a la diferencia de las probabilidades entre las dos clases más probables. Esta representación permite ver los efectos de la variación de los umbrales de asignación establecidos. Con los umbrales establecidos por defecto 5 muestras, de un total de 50, quedarían sin asignar a ninguna clase

(*not-assigned*), disminuyendo el *call rate*, pero sin incrementar el número de falsos positivos ya que la muestra en rojo, erróneamente clasificada no es asignada.

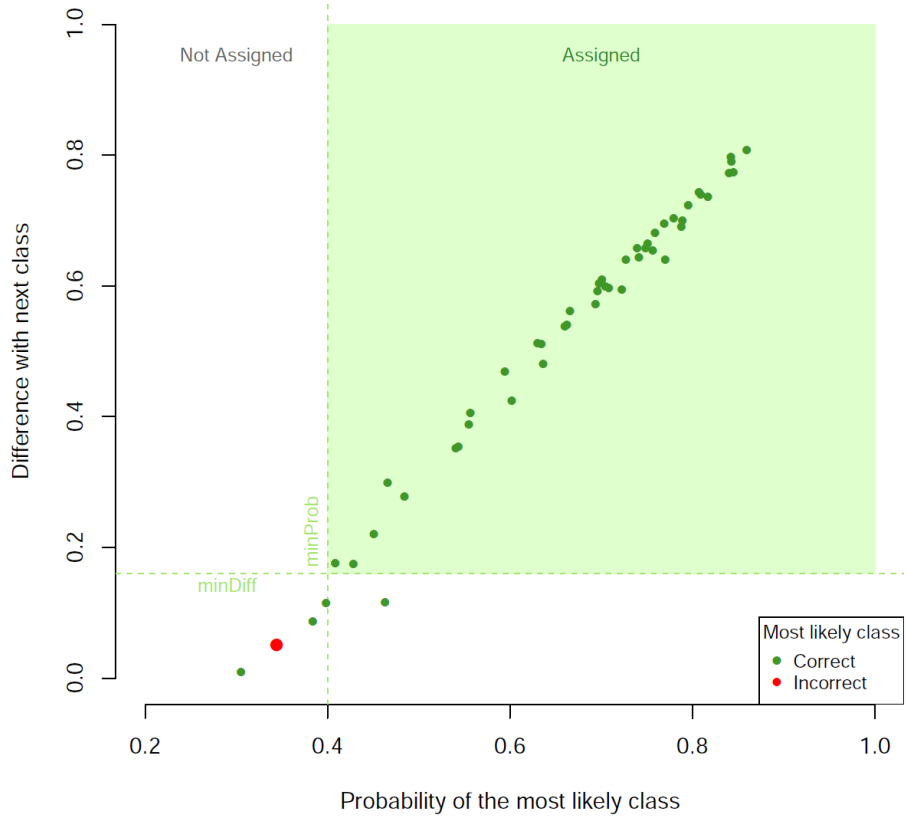


Figura 1.10: Estrategia de asignación - Probabilidades de asignación de un conjunto de muestras a la clase más probable frente a la diferencia entre las dos clases más probables. Las líneas verticales representan los umbrales utilizados en el algoritmo para el caso de 5 clases. Se presenta la asignación de 50 muestras.

1.3.3.2 Estimación del error de generalización

La estimación del error de generalización de un clasificador mediante validación cruzada (*Cross Validation*, CV) una vez que han sido seleccionados los parámetros y variables óptimas en un determinado estudio es un estimador sesgado el error real del clasificador puesto que deja fuera de dicha estimación todo el proceso de selección y ajuste de parámetros y variables. Para lograr un estimador independiente que tenga en cuenta todo el proceso se han desarrollado estrategias más robustas de estimación del error de generalización basadas en la aplicación de métodos de validación cruzada durante la construcción de los clasificadores. En concreto, la validación cruzada anidada o doble (*Double-Nested Cross Validation*, nCV) reduce considerablemente el sesgo y proporciona una estimación del error del método de clasificación mucho más ajustada a la que se correspondería con un conjunto de datos independiente (Varma and Simon, 2006).

En la nCV se simula una validación independiente añadiendo un segundo bucle de CV sobre el inicial utilizado para la estimación del número óptimo de genes. Un ejemplo simplificado para entender el funcionamiento es el mostrado en la figura 1.11. El conjunto de datos total se divide en r grupos mutuamente excluyentes (r*-Fold Cross Validation*). El error es estimado en el bucle externo (*outer loop*) y el bucle interno (*inner loop*) determina el valor óptimo de los parámetros/variables,

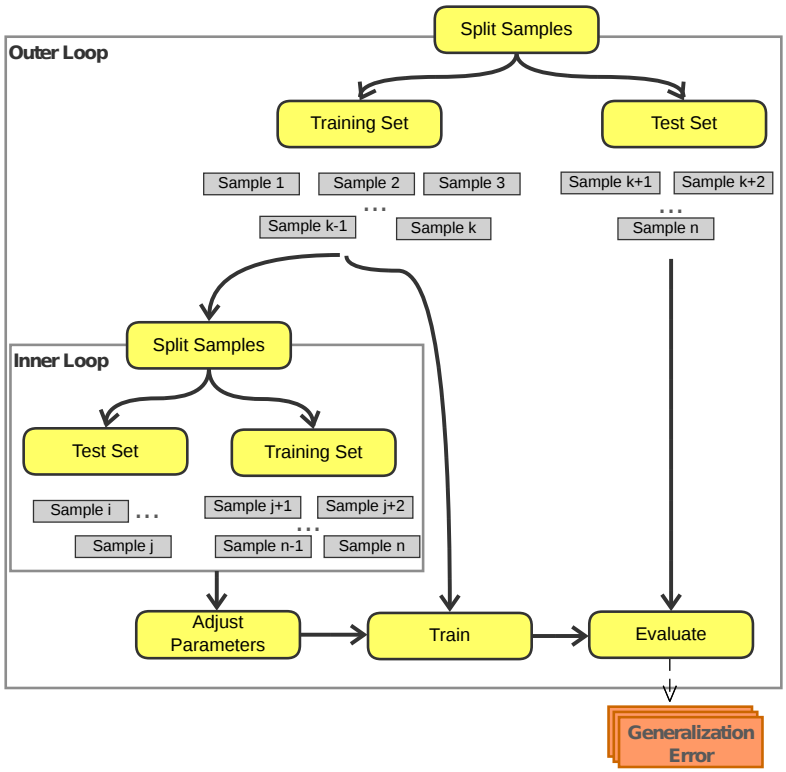


Figura 1.11: Validación cruzada doble/anidada (nCV) - Simplificación del esquema de nCV implementado. Comprende dos bucles, un bucle interno para la estimación del número de variables a utilizar en la clasificación y otro externo para la evaluación y estimación del error de generalización.

en nuestro caso el número de genes. Es importante resaltar que para obtener una estimación válida del error de generalización todos los pasos del algoritmo, incluyendo la selección de genes, tienen que estar integrados dentro de la estrategia de CV.

La estimación del error de generalización se basa en el cálculo de tres parámetros:

1. *Sensibilidad* o tasa de verdaderos positivos (*True Positive Rate*, TPR)

$$sensibilidad(TPR) = \frac{TP}{TP + FN} \tag{1.7}$$

2. *Especificidad* o tasa de verdaderos negativos (*True Negative Rate*, TNR)

$$especificidad(TNR) = \frac{TN}{TN + FP} \tag{1.8}$$

3. *Coeficiente de correlación de Matthews* que mide de modo equilibrado el balance entre verdaderos y falsos positivos y negativos (*Matthews Correlation Coefficient*, MCC).

$$MCC = \frac{TP * TN - TP * FN}{\sqrt{(TP + FP)(FP + FN)(TN + FP)(TN + FN)}} \tag{1.9}$$

1.3.4 Asociación entre genes marcadores en cada clase

Entre los genes situados en las cabeceras del *ranking* de expresión diferencial aparecen algunos con perfiles de expresión similares. Estos genes asociados o relacionados podrían considerarse variables redundantes y poco útiles en la clasificación.De hecho la eliminación de variables dependientes para reducir el conjunto de variables de entrada óptimo para la clasificación sin que

se reduzca, *a priori*, el rendimiento del clasificador constituye un tema muy estudiado (Ding and Peng, 2005; Liu et al., 2009, 2011).

Sin embargo, estas asociaciones entre genes marcadores pueden resultar muy útiles desde el punto de vista biológico. La identificación de conjuntos de genes asociados desregulados de la misma manera en un estado patológico puede ayudar en la identificación de los procesos o funciones en los que es posible que estén cooperando.

El estudio de estas asociaciones se aborda desde dos perspectivas complementarias. Por un lado se analiza la co-modulación de los genes mediante un análisis de coexpresión. Por otro lado se analiza la cantidad de información que un gen puede aportar sobre la modulación de otro mediante el análisis de la información mutua entre cada par de genes.

El análisis de coexpresión está basado en el cálculo de la correlación de Pearson entre cada par de genes:

$$r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{(\sum_{i=1}^{n} X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}} \tag{1.10}$$

La interacción o dependencia entre cada par de genes ha sido estimada mediante el cálculo de la información mutua:

$$I(X;Y) = \sum_{y \in Y}\sum_{x \in X} P(x,y)\log\frac{P(x,y)}{P(x)P(x)} \tag{1.11}$$

Las correlaciones e interacciones calculadas entre cada par de genes permiten la construcción de redes de genes basadas en la asociación gen a gen. La identificación de módulos con alta densidad de conexiones permite descubrir grupos de genes que pueden estar siendo modulados conjuntamente o bien implicados en el mismo proceso biológico.

En el ejemplo de la figura 1.12 se muestra la red construida para un tipo de leucemia. En esta red se aprecia un módulo o grupo de genes sobre-expresados (en rojo) y otro conjunto más pequeño de genes reprimidos o infra-expresados (en verde). La mayoría de las asociaciones entre genes corresponden en este caso a coexpresión (lineas azules).

1.4 Aplicación a datos de leucemia

Las leucemias constituyen un conjunto de enfermedades caracterizadas por una proliferación hematopoyética anómala. Ciertos tipos de células sanguíneas se multiplican a un ritmo superior al normal y no se diferencian de modo adecuado, dando lugar a masas neoplásicas que impiden el desarrollo de los demás tipos celulares hematológicos. Dependiendo de la evolución de la enfermedad y de la línea celular a la que afectan las leucemias se pueden clasificar en agudas (A) o crónicas (C), mieloides (ML) o linfoides (LL). Las leucemias agudas se caracterizan por un aumento muy rápido de las células inmaduras, lo que impide que la médula ósea pueda producir las sanas correctamente. La crónica sin embargo, se distingue porque aunque puede producir las células maduras, éstas siguen siendo de alguna forma más numerosas y además defectuosas. Su progresión puede llevar meses o incluso años y suele darse principalmente en ancianos (i.e. personas mayores de 65 años). Las leucemias linfoides, se denominan así porque las células afectadas son aquellas que dan lugar a los linfocitos. Las mieloides, afectan a los mielocitos, que posteriormente se desarrollan

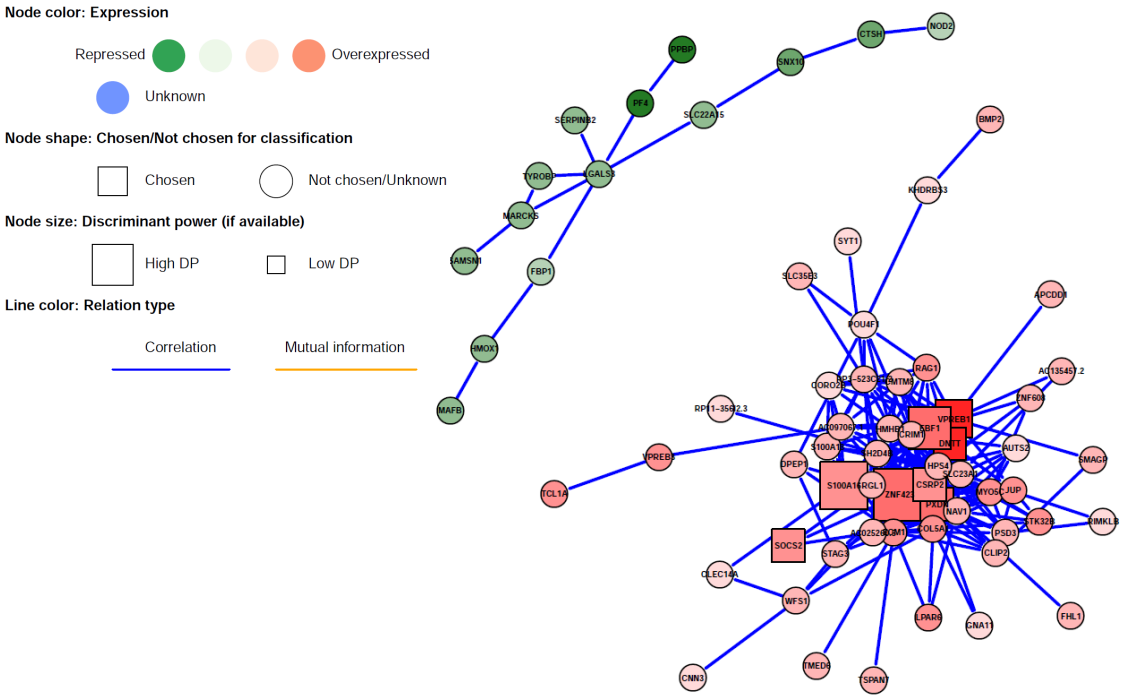


Figura 1.12: Red de interacción para Leucemias Agudas Linfoblásticas - Red de genes obtenida calculando correlaciones e información mutua. En verde genes reprimidos en ALL y en rojo sobre-expresados. Aparecen representados con cuadrados los genes utilizados para la construcción del clasificador con tamaño en función del poder discriminante.

en glóbulos rojos, blancos o plaquetas. Un esquema del proceso de diferenciación de las células sanguíneas puede verse en la figura 1.13.

Según lo descrito, existen cuatro grandes clases de leucemias que son las siguientes:

1. Leucemia linfoblástica aguda (*Acute Lymphoblastic Leukemia*, ALL)
2. Leucemia mieloide aguda (*Acute Myeloid Leukemia*, AML)
3. Leucemia linfocítica crónica (*Chronic Lymphocytic Leukemia*, CLL)
4. Leucemia mieloide crónica (*Chronic Myeloid Leukemia*, CML)

En esta sección se aplicarán los algoritmos desarrollados sobre un conjunto de 50 muestras humanas de pacientes con leucemia analizados con microarrays de expresión de *Affymetrix* modelo *HG U133 plus 2.0*. Este conjunto consiste en 10 muestras de cada uno de los tipos principales descritos: ALL, AML, CLL y CML; más un conjunto de 10 muestras de individuos sanos (NoL). Con la aplicación de los algoritmos desarrollados se pretende encontrar la firma molecular génica propia de cada uno de estos tipos de leucemia y construir un clasificador que permita diferenciar los diferentes subtipos. El clasificador proporcionará datos acerca de los genes que constituyen dicha firma molecular candidatos a ser biomarcadores en cada una de las clases.

1.4.1 *Ranking* de genes asociados a cada subtipo de leucemia

El análisis de las probabilidades posteriores de los genes calculadas con PEB en el primer paso del algoritmo permite comparar las diferentes clases o estados patológicos estudiados proporcionando un tamaño aproximado de las firmas moleculares para cada uno de ellos. Observando el número

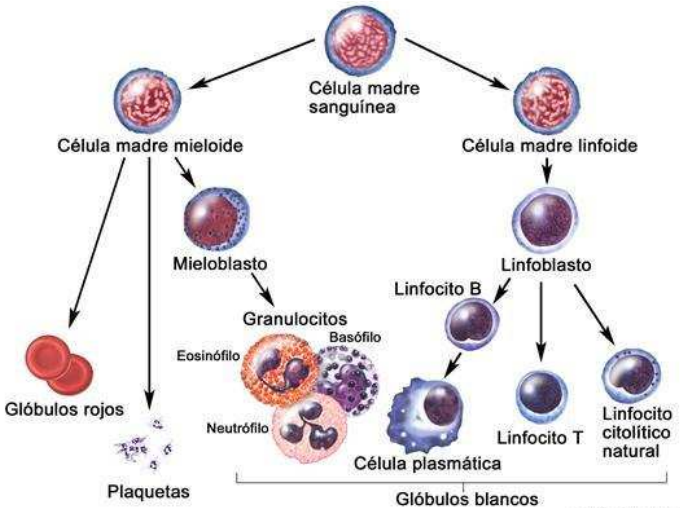


Figura 1.13: Esquema de la hematopoyesis - Evolución y diferenciación de las células sanguíneas en las dos líneas principales mieloide y linfoide a partir de las células madre hematopoyéticas (HSCs) que se encuentran en la médula ósea. (Fuente: www.genome.gov).

de genes con probabilidades posteriores por encima de un umbral común (por ejemplo 0.75), podremos comparar los estados patológicos y distinguir cuáles están asociados con la desregulación de una mayor cantidad de genes y por tanto de procesos.

En el *ranking* de genes para los 4 subtipos de leucemia se puede observar que el número de genes asociados a cada uno ellos es muy diferente. La figura 1.14 muestra la distribución de genes ordenados en base a las probabilidades posteriores de expresión diferencial para cada subtipo de leucemia. Tan sólo 308 genes están asociados a AML con una probabilidad posterior mayor que 0.75, mientras que para la misma significación estadística CLL incluye más de dos mil. Esto sugiere que en AML están afectados procesos muy concretos, mientras que CLL sería una enfermedad más sistémica, con más genes afectados y procesos desregulados.

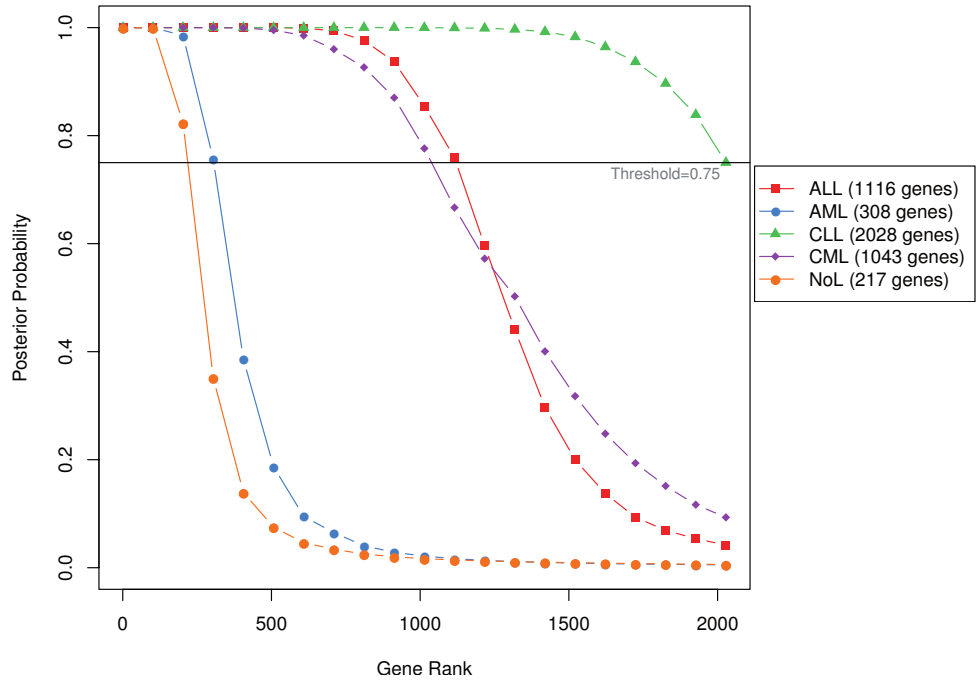


Figura 1.14: Distribución de probabilidades posteriores de expresión diferencial en 4 subtipos de leucemia - Genes ordenados en base a su probabilidad posterior para cada uno de los subtipos de leucemia. En la leyenda se indica el número de genes con probabilidad posterior > 0.75.

1.4.2 Genes seleccionados para cada subtipo de leucemia

Como se ha descrito, el algoritmo de clasificación utiliza el orden de los genes en el *ranking* para guiar la selección de variables que servirán para distinguir cada uno de los subtipos patológicos o clases. El clasificador que obtuvo una mayor precisión para la diferenciación de los 4 subtipos de leucemias ha sido construido con 34 genes distribuidos entre las clases como indica la tabla 1.1.

ALL	AML	CLL	CML	NoL
8	5	4	4	13

Tabla 1.1: Número de genes seleccionados para diferenciar cada tipo de leucemia

La figura 1.15 presenta la evolución de las tasas de error de clasificación que se han obtenido con un número creciente de genes añadidos en el proceso de selección de variables. Cada uno de los colores representa una iteración de este proceso de selección. Como se observa los errores son estables en cada una de las iteraciones con tasas de error similares para el mismo número de genes.

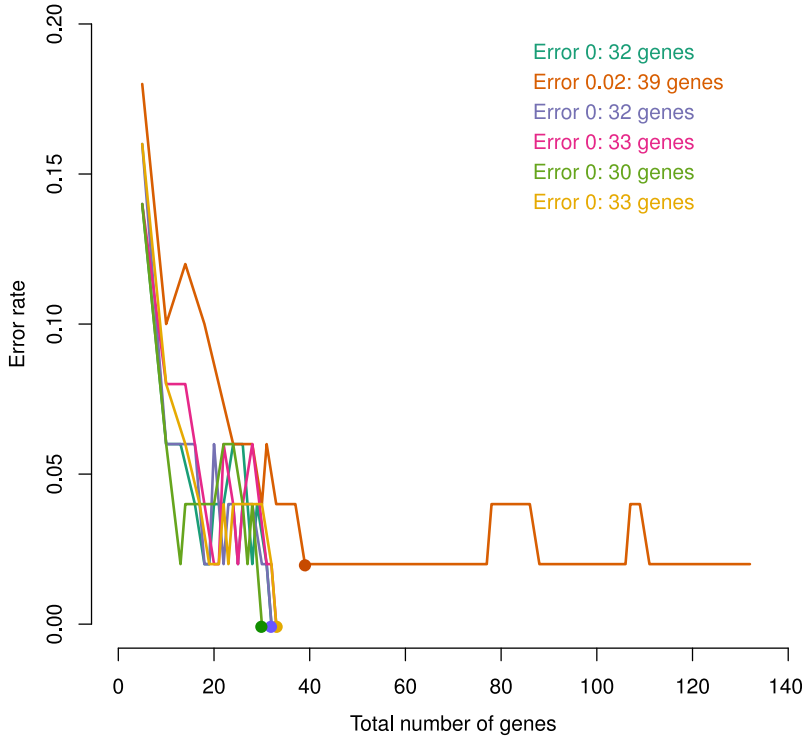


Figura 1.15: Tasas de error para distintos números de genes seleccionados - Tasas de error observadas en el proceso de validación de los clasificadores de leucemias construidos con un número creciente de genes. Cada color representa una de las n iteraciones (n=6).

La tabla 1.4.2 contiene la lista con los genes marcadores de cada subtipo de leucemia seleccionados por el clasificador. Además del nombre del gen la tabla incluye: **(i)** la posición que ocupa en el *ranking* en la clase a la que ha sido asociado; **(ii)** el valor de la probabilidad posterior calculada con PEB; **(iii)** la diferencia entre el nivel de expresión medio del gen en las muestras de la clase que discrimina y el nivel medio de expresión en el resto de muestras; y **(iv)** el poder discriminante de cada gen derivado del clasificador. El valor de la probabilidad posterior de expresión diferencial aparece siempre como 1 en la tabla debido a los límites de representación numérica en la aproximación a 1 (0.9999...)

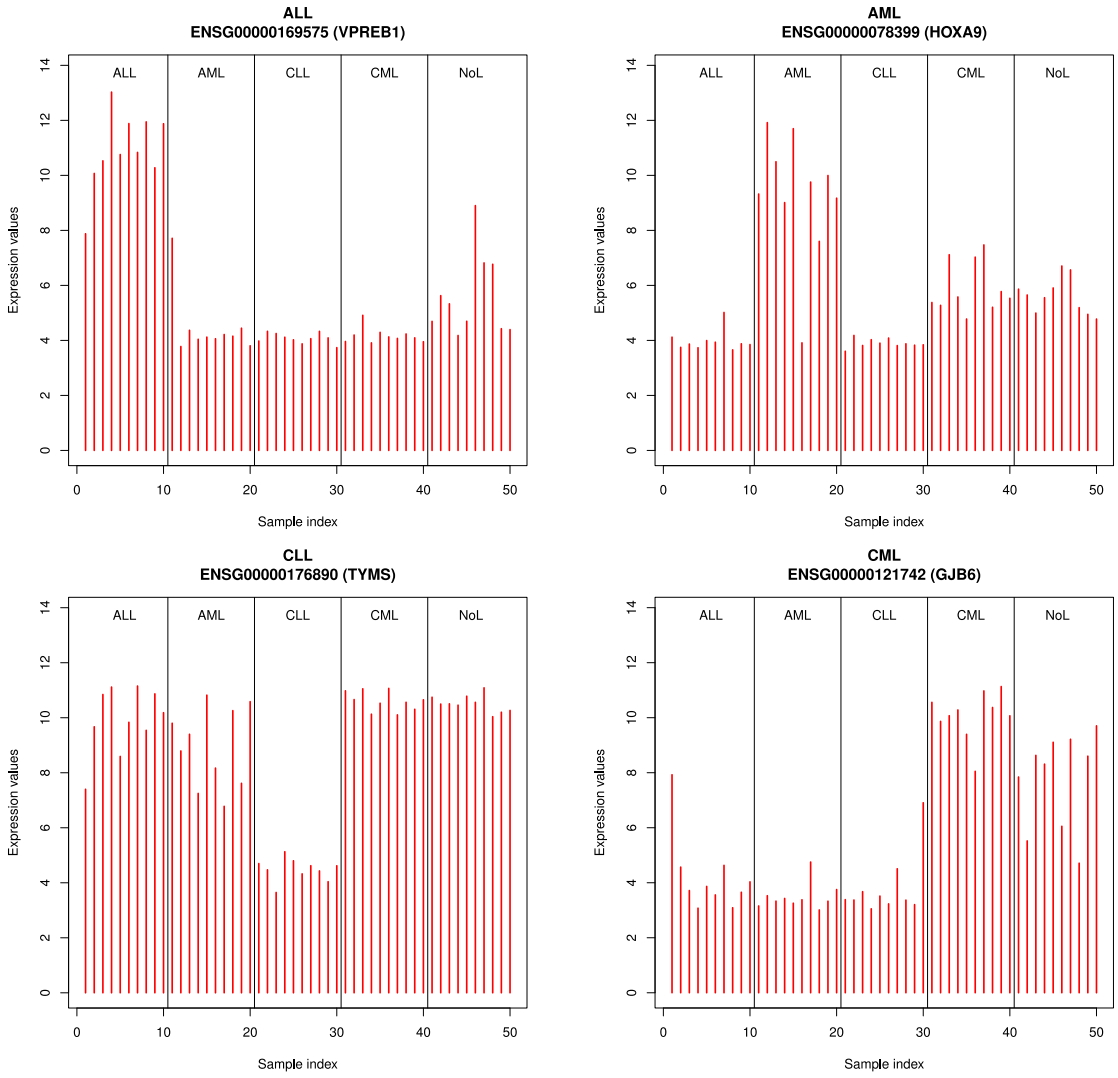


Figura 1.16: Perfiles de expresión de genes asociados a leucemias - Niveles de expresión de los genes con mejor probabilidad posterior de expresión diferencial en cada subtipo de leucemia (VPREB1, HOXA9, TYMS y GJB6). Cada barra representa el nivel de expresión de un gen en una muestra. Las muestras están agrupadas de 10 en 10 en cada subtipo de leucemia.

Gene	Ranking position	Class	Posterior probability	Mean expression difference	Discriminant power
VPREB1	1	ALL	1	6.33	9.74
ZNF423	2	ALL	1	5.10	12.77
DNTT	3	ALL	1	6.89	8.58
EBF1	4	ALL	1	5.42	10.50
PXDN	5	ALL	1	5.04	8.31
S100A16	6	ALL	1	4.34	11.12
CSRP2	7	ALL	1	4.05	8.98
SOCS2	8	ALL	1	4.54	7.77
HOXA9	1	AML	1	4.44	7.78
MEIS1	2	AML	1	3.28	10.84
CD24L4	3	AML	1	-4.49	-2.99
ANGPT1	4	AML	1	2.74	10.03
CCNA1	5	AML	1	2.56	10.15
AC079767.3	1	CLL	1	6.51	11.83
TYMS	2	CLL	1	-5.52	-10.49
FCER2	3	CLL	1	4.59	7.65
NUCB2	4	CLL	1	-5.61	-10.95
GJB6	1	CML	1	5.25	6.40
PRG3	2	CML	1	4.98	5.95
LY86	3	CML	1	-2.20	-4.22
AC091062.1	4	CML	1	2.43	3.68
IGHV3-23	1	NoL	1	3.49	4.85
IGLV3-19	2	NoL	1	2.74	5.81
IGKV4-1	3	NoL	1	2.66	3.10
IGLV1-47	4	NoL	1	2.35	3.76
FGF13	5	NoL	1	2.69	-1.50
IGLV3-25	6	NoL	1	2.47	-2.49
IGHV3-9	7	NoL	1	2.11	4.94
NMU	8	NoL	1	1.97	4.42
SMPDL3A	9	NoL	1	1.95	4.74
KLRB1	10	NoL	1	2.23	4.71
RNF182	11	NoL	1	1.84	2.36
RFESD	12	NoL	1	2.37	5.61
SLC25A21	13	NoL	1	1.49	8.24

Tabla 1.2: Genes seleccionados para diferenciar cada tipo de leucemia

Los perfiles de expresión del primer gen de la tabla para cada clase (VPREB1, HOXA9, TYMS y GJB6) se pueden ver en la figura 1.16, donde las líneas verticales representan el nivel de expresión del gen en cada una de las muestras.

Cada uno de estos genes tiene una probabilidad posterior de expresión diferencial calculada con el método PEB. Además, el clasificador proporciona el poder discriminante, comentado en la sección 1.3.3. El orden de los genes establecido en base a la probabilidad de expresión diferencial no tiene porqué coincidir exactamente con el orden basado en el poder discriminante. Sin embargo ambas medidas pueden utilizarse conjuntamente para analizar los genes de interés en cada subtipo de leucemia analizado. Así por ejemplo la probabilidad posterior de expresión diferencial del gen VPREB1 en ALL es muy alta, sin embargo su poder discriminante en el clasificador no lo es tanto (de hecho es menor que el poder discriminante de otros genes que están más abajo en el *ranking*).

La figura 1.17 muestra el poder discriminante de los genes que ocupan las 4 primeras posiciones del *ranking* para AML, ordenados de izquierda a derecha. En este gráfico también se aprecia que MEIS1, homólogo humano de un gen murino conocido por su papel en el desarrollo de leucemia mieloide (Moskow et al., 1995; Smith et al., 1997) es el gen que mejor diferencia AML del resto, aunque sea HOXA9 el que presenta una mayor probabilidad posterior. La interacción entre ambos

genes ha sido además reportada en células mieloides indicando que sobre-expresión de ambos genes es sufiente para la inducción de leucemia mieloide en ratones (Shen et al., 1999).

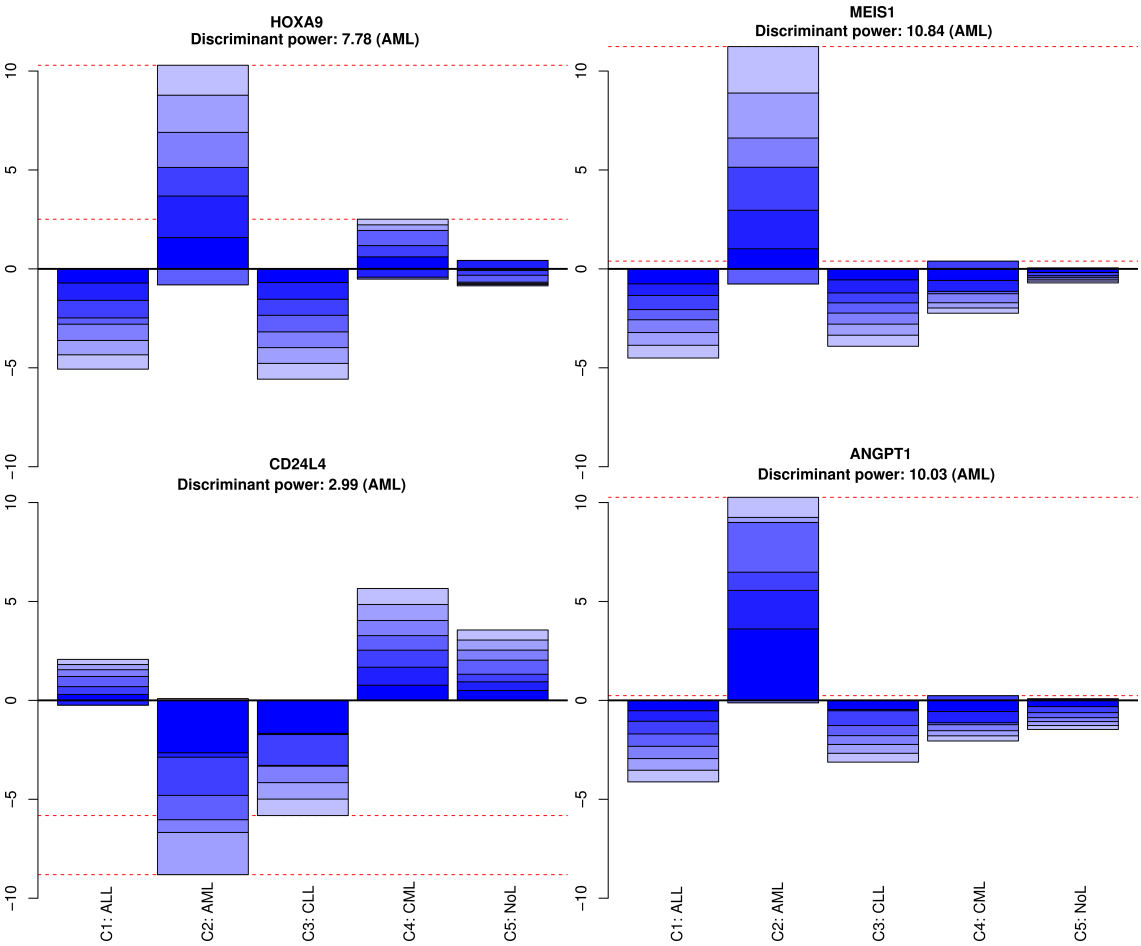


Figura 1.17: Poder discriminante - Representación que muestra el parámetro definido como *poder discriminante* para los genes que ocupan las 4 primeras posiciones del *ranking* de expresión diferencial para AML.

1.4.3 Estimación del error de generalización para el clasificador de leucemias

En la figura 1.15 se observa que el error cometido al clasificar los 4 tipos de leucemias siempre tiende a 0 o está muy cerca de 0 cuando se llega a un número óptimo suficiente de genes. Esta estimación del error con validación cruzada es una medida aproximada que, muchas veces, puede ser optimista respecto a los valores reales de clasificación con muestras independientes. Una estimación más robusta es la validación cruzada anidada (nCV) descrita en 1.3.3.2. En la tabla 1.4.3 se proporcionan los valores de sensibilidad, especificidad y el coeficiente de correlación de Mathew (MCC) para cada una de las clases, así como el *call rate* o porcentaje de muestras clasificadas utilizando nCV.

	Sensibilidad	Especificidad	MCC	Call Rate
ALL	100	100	100	90
AML	100	100	100	80
CLL	100	100	100	90
CML	100	100	100	100
NoL	100	100	100	90

Tabla 1.3: Parámetros de estimación del error de clasificación

1.5 Discusión

Los perfiles de expresión génica derivados de datos de microarrays han sido ampliamente utilizados para construir clasificadores o predictores utilizando diferentes técnicas de aprendizaje automático (*machine learning*, ML). Una de las aplicaciones principales de esta tecnología es la clasificación de tipos y subtipos de enfermedades. El problema fundamental que surge en este escenario es la identificación de las mejores características para la clasificación, es decir aquellas que no solamente reduzcan al mínimo los errores cometidos sino que tengan un sentido biológico coherente con las enfermedades analizadas. De hecho la mayoría de los métodos funcionan como una caja negra respecto a las entidades biológicas tras la clasificación, y por tanto no resultan útiles en la identificación de los procesos biológicos subyacentes.

Por otro lado, los métodos de clasificación más frecuentemente usados son binarios –es decir, para asignación entre dos clases– y no suelen permitir la opción de *no-asignación* que se daría cuando hay duda en el sistema experto, como sucede frecuentemente en las asignaciones dadas por expertos humanos.

En este capítulo se ha propuesto un algoritmo diseñado para la construcción de clasificadores multiclase que permitan identificar qué genes están asociados a cada una de las clases y en qué grado. El método incluye la posibilidad de *no-asignación* y, además, proporciona para cada clase una red de genes derivada que facilita la interpretación de los procesos biológicos asociados. Estas redes pueden ser estudiadas mediante el análisis de los módulos o conjuntos de genes más conectados o relacionados.

La definición de las entidades que van a ser utilizadas como variables es un aspecto que normalmente no se tiene en cuenta en los trabajos centrados en la clasificación de estados patológicos a partir de datos genómicos. La mayoría de estos trabajos toma los niveles de expresión de los *probesets* como variables en lugar de los genes en su definición biológica más exacta. Si el punto de partida del método se corresponde a entidades definidas arbitrariamente la dificultad de encontrar el procesos coherentes en los resultados obtenidos se ve incrementada. De este modo, una redefinición y mapeo de las sondas directamente sobre los loci génicos reduce parte del ruido inherente a la tecnología genómica utilizada, solucionando inconsistencias como el hecho de que los valores de los *probesets* asociados a un mismo gen presente niveles de expresión diferentes.

El análisis de los genes asociados a cada una de las clases proporciona una idea general acerca de la complejidad de las enfermedades estudiadas. Aquellas con un número alto de genes significativamente asociados es probable que afecten a procesos o funciones más generales o a un mayor número de ellos. Además se proporciona información del poder discrimiante de cada uno de los genes de manera complementaria a la significación encontrada mediante el análisis de expresión diferencial, lo que da una idea independiente de la relevancia de cada gen en la clasificación.

Una de las técnicas más utilizadas para la selección de variables es la identificación y eliminación de las asociaciones o dependencias entre las mismas para minimizar las redundancias, criterio MRMR (*Minimum Redundancy-Maximum Relevance*) (Ding and Peng, 2005). En la teoría general

de aprendizaje computacional (ML) se plantea que eliminando estas redundancias se mantiene el poder predictivo y se disminuye el número de variables necesarias para la clasificación. El método propuesto permite la identificación y la eliminación de estas asociaciones, sin embargo la información de las relaciones entre los genes es mantenida y utilizada en la construcción de redes de interacción como un paso más hacia la interpretabilidad biológica de los resultados.

En conclusión el algoritmo desarrollado *geNetClassifier* proporciona un método de clasificación robusto asegurado mediante validación cruzada anidada y centrado en el acceso transparente a las entidades biológicas que clasifican. El algoritmo multiclase propuesto constituye una ventaja frente a las aproximaciones tradicionales focalizadas en la minimización de los errores de clasificación sin tener en cuenta su interpretabilidad.

Capítulo

2

Análisis de alteración de número de copias de DNA en cáncer

2.1 Introducción: Alteración del número de copias de DNA

El presente capítulo se centra en el análisis del DNA genómico humano, concretamente en el análisis de las alteraciones de número de copias de DNA que sufren las células en determinadas patologías o situaciones de transformación. En inglés estas alteraciones se denominan habitualmente *Copy Number Alterations* y se citan con el acrónimo CNA.

En esta sección se definen los conceptos básicos de CNA y las plataformas genómicas experimentales de tipo *microarrays* que permiten la cuantificación de este tipo de datos de una manera global. Las siguientes secciones de este capítulo describen el flujo de trabajo y metodologías desarrolladas para el análisis de CNA. En los últimos años han sido desarrollados numerosos métodos para la normalización y segmentación como CRMAv2 (Bengtsson et al., 2009) o CBS (Venkatraman and Olshen, 2007). Sin embargo, el consenso de metodologías para lograr un análisis robusto de series con múltiples muestras para datos de número de copias no es tan firme como las metodologías actualmente usadas para datos de expresión. Por este motivo el capítulo se divide en dos secciones diferenciadas: **(i)** primero sobre el preprocesamiento y normalización de muestras individuales; **(ii)** segundo sobre el análisis unificado de conjuntos de muestras.

Además, dentro del análisis unificado se han desarrollado algoritmos para la detección de las regiones mínimas de alteración recurrente (*Minimal Common Regions*, MCR) así como para la detección de regiones con puntos de ruptura recurrentes (*Recurrent Breakpoint Regions*, RBR).

Por último, el trabajo incluye los resultados de la aplicación de los algoritmos diseñados a la identificación de regiones alteradas en muestras humanas de cáncer colorectal (*Colo-Rectal Cancer*, CRC). Este análisis permite identificar de forma consistente las regiones más significativamente alteradas en CRC.

2.1.1 Definición de alteración del número de copias de DNA

La gran mayoría de los organismos eucariotas metazoos –incluyendo todos los mamíferos– son organismos diploides. Esto significa que sus células tienen dos copias del genoma completo, es decir, dos copias de cada una de las moléculas de DNA que constituyen los cromosomas. En el caso del ser humano las células somáticas normales cuentan con 46 cromosomas, 23 pares de cromosomas, de los cuales 22 pares son iguales (llamados autosomas) y un par son los cromosomas sexuales (llamados heterocromosomas X e Y) que son iguales en la mujer (XX) y diferentes en el hombre (XY).

La diploidía celular se mantiene bien controlada en las células somáticas durante su replicación. De este modo, en la división de la célula en dos células hijas –que es el proceso conocido como mitosis–, cada molécula de DNA es replicada y cada nueva célula hija recibe una copia de ese DNA. Algunas veces durante este proceso se producen errores que conllevan cambios en las moléculas de DNA y por tanto en el genoma. Hay diferentes tipos de errores que pueden ir desde modificaciones de secuencia en una única base nucleotídica (*Single Nucleotide Polymorphism*, SNP) a alteraciones que pueden afectar a todo un cromosoma. Algunos de estas modificaciones como mutaciones puntuales, translocaciones o inversiones no afectan al número de copias de DNA, sin embargo otras modificaciones sí pueden afectarlo. En este capítulo nos centraremos únicamente en aquellas alteraciones que pueden modificarlo.

Una alteración del número de copia de DNA (*Copy Number Alteration*, CNA) es el incremento o disminución patológica de una parte del genoma que puede abarcar desde un cromosoma entero a un segmento de pocos cientos o miles de pares de bases. Las CNAs son de dos tipos principales:

- **Ganancias** o **amplificaciones** en las que un segmento de DNA se replica más de una vez, con lo que el número de copia total de DNA correspondiente a esa región es mayor que 2 (siendo, com se ha indicado, el estado normal en los autosomas diploide, 2 copias).
- **Pérdida** o **deleción** en la que un segmento de DNA se pierde, con lo que el número de copias totales de DNA en esa región pasa a ser 1 ó 0 (según se haya perdido una o ambas copias) .

Las CNAs están involucradas en el desarrollo y progresión de diferentes tipos de enfermedades complejas, especialmente en cáncer. La amplificación de una región genómica que codifica un oncogén (*Oncogene*) o la deleción de una región que codifica un gen supresor tumoral (*Tumor Suppressor Gene*, TSG) puede contribuir a la transformación de una célula en tumoral. De este modo, mediante el estudio de estas CNAs es posible identificar aquellas regiones cuya alteración juega un papel importante en el desarrollo y progresión de la enfermedad. El fin último de la detección del número de copias de DNA es establecer las regiones cromosómicas asociadas al estado patológico y, en particular en cáncer, asociadas a la progresión tumoral, a la supervivencia de los enfermos o las posibilidades de éxito de determinados tratamientos (Kallioniemi, 2008).

2.1.2 Cuantificación del número de copias de DNA

Existen diferentes plataformas experimentales que permiten la cuantificación de las CNAs midiendo los valores de número de copias de un genoma en posiciones concretas. Por un lado están las técnicas genómicas de gran escala y análisis masivo, como los microarrays de número de copias de DNA, y por otro las técnicas específicas de pequeña escala que permiten estudiar un número reducido de regiones concretas, como la hibridación y marcaje con sondas fluorescentes *in situ* (*Fluorescence In Situ Hybridization*, FISH). En cuanto a los microarrays, hay dos tipos principales que permiten la cuantificación del número de copias: los arrays de CGH (*Comparative Genomic*

Hybridization, aCGH) y los arrays de SNPs (*oligonucleotide-based Single Nucleotide Polimorfism* arrays).

2.1.2.1 **Arrays de CGH**

En los arrays de CGH cadenas largas de DNA con localizaciones genómicas conocidas están inmovilizadas en cada *spot* del array. El DNA de la muestra problema (*test sample*) -muestra del tumor en el caso de estudios de cáncer- junto con el de una muestra de referencia normal (*reference sample*), marcados con diferentes fluoróforos, se hibridan sobre el chip. Normalmente estos arrays son de tipo "dos-colores.º "dos-canales"(*two-channel arrays*), mezclándose la muestra problema y la referencia tras el marcaje respectivo de cada una con un tipo de fluoróforo (por ejemplo los *cyanine*: verde Cy3 y rojo Cy5). Las intensidades de fluorescencia de ambas muestras se miden y se comparan calculando la relación entre ellas, es decir, el ratio normalmente transformado a escala logarítmica (*log-ratio*). Este ratio es un valor que permite estimar la diferencia en el número de copias entre las dos muestras hibridadas para una determinada localización en el genoma. La figura 2.1 muestra una representación esquemática del proceso de hibridación en este tipo de arrays.

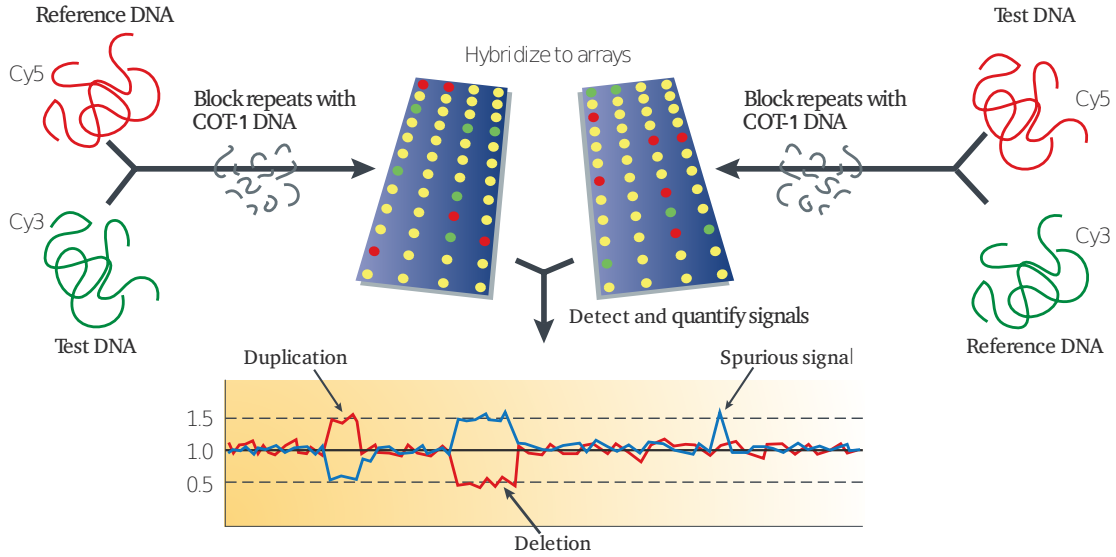


Figura 2.1: Arrays de CGH. - Representación esquemática de la hibridación de microarrays de CGH de tipo "dos-colores"(Cy3 y Cy5). Modificado de (Feuk et al., 2006).

Si, por ejemplo, el número de copias en un determinado locus es mayor en el tumor que en la referencia una mayor cantidad de moléculas de DNA tumoral hibridarán en el *spot* del micorray que representa ese *locus*, mientras que, comparativamente, la cantidad de moléculas de la referencia será menor. Las localizaciones de estos *loci* se corresponderán con genes que probablemente estén asociados con dicho tumor.

El análisis de este tipo de datos siempre se realiza de una manera comparativa tomando los ratios (test / referencia) y nunca cuantificando los valores absolutos de hibridación de una única muestra.

2.1.2.2 **Arrays de SNPs**

Los arrays de SNPs fueron inicialmente diseñados para detectar polimorfismos en el DNA genómico en una población. Los polimorfismos de un único nucleótido (*Single Nucleotide Polimorfism*,

llamados SNPs, son las variaciones más frecuentes en el genoma y se definen como variaciones de una única base en la secuencia de DNA en al menos un 1 % de la población.

Los arrays de SNPs disponibles en el mercado son microarrays de oligonucleótidos de alta densidad. Estos arrays son de tipo ün-color.º ün-canal"(*single-channel arrays*). En su diseño, descrito de modo básico, constan de una superficie sólida de material inerte que hace de soporte sobre el que se inmovilizan de modo ordenado (en pequeñas áreas bien definidas) cadenas cortas de DNA de hebra simple (oligonucleótidos) con secuencias específicas del DNA o del genoma que se quiere testar. Sobre estas micro áreas o celdas, que son de miles a cientos de miles en un array, se hibrida una muestra del DNA problema preparado adecuadamente con pasos de digestión, amplificación, fragmentación y marcaje. La fluorescencia que marca la muestras permite una lectura y cuantificación precisa, realizada con aparatos diseñados para ello. La figura 2.2 tomada de *Affymetrix Genome-Wide Human SNP Array 6.0 data sheet* muestra de manera esquemática el proceso de hibridación de un array de SNPs. Este tipo de arrays de SNPs han demostrado un mejor rendimiento que aCGH en la detección de ganancias o pérdidas de una sola copia de diferentes tamaños genómicos (Hehir-Kwa et al., 2007).

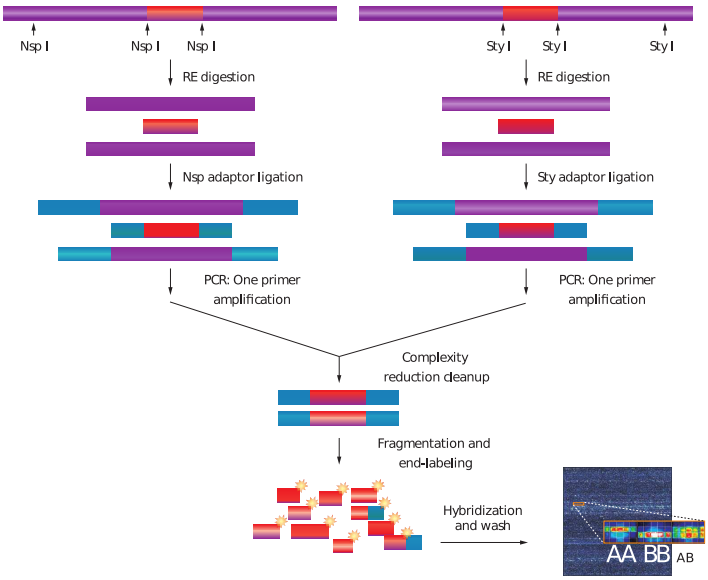


Figura 2.2: Arrays de SNPs. - Representación esquemática de la hibridación de microarrays de SNPs de tipo ün-color". (Fuente: www.affymetrix.com)

Para cada uno de los SNPs, el microarray contiene sondas para interrogar cada uno de los alelos (A y B) pudiendo corresponder a 3 formas: monocigótica AA, heterocigótica AB ó monocigótica BB. La cuantificación de cada alelo independientemente se debe a que estos microarrays fueron diseñados para la detección del "genotipo"de cada SNP; es decir, el genotipado de miles de polimorfismos simples a la vez. El análisis del genotipo permite estudiar la susceptibilidad de cada individuo a determinadas enfermedades cuando éstas están ligadas a ciertos genes (*expression Quantitative Trait Loci*, eQTL). También estos microarrays pueden ser utilizados para buscar la combinación de diferentes SNPs como marcadores genéticos en estudios llamados de .ªsociación genómica"(*Genome Wide Association Studies*, GWAS). En todos estos tipos de estudios el problema principal es que se necesitan siempre cientos o miles de muestras, y por ello de microarrays, para lograr un análisis estadístico robusto. Sin embargo, en los últimos años los arrays de SNPs han sido utilizados de modo muy eficiente en la determinación del número de copias de DNA.

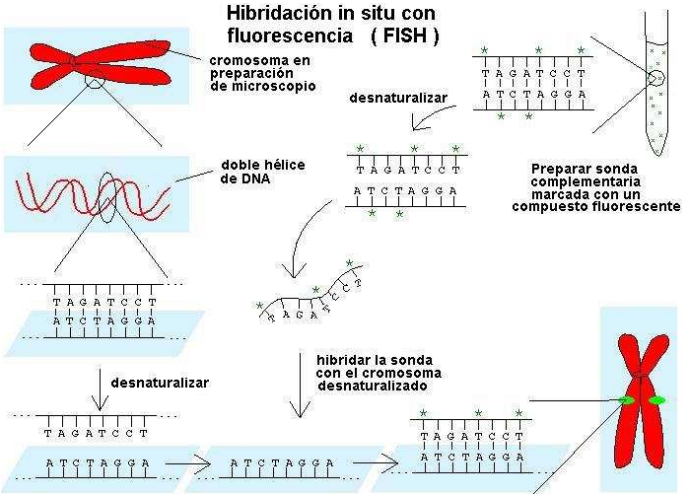


Figura 2.3: Técnica FISH - Esquema del método de hibridación *in situ* con fluorescencia. (Fuente: *National Human Genome Research Institute*, NHGRI, www.genome.gov).

Otra ventaja que presenta la utilización de los arrays de SNPs aplicados al cálculo del número de copias es la posibilidad de detección de la pérdida de heterocigosidad (*Loss of Heterozygosity*, LOH). La LOH ocurre cuando de los dos alelos presentes de un gen uno está mutado y el alelo normal se pierde. Por ejemplo, si se produce una mutación maligna en un gen supresor tumoral que hace que este gen ya no sea funcional y la segunda copia funcional del gen desaparece, es posible que se desencadene un proceso tumoral. Si el gen que presenta esta LOH tiene un número de copias normal diploide (*copy-neutral LOH*) es posible detectarlo mediante arrays de SNPs, pero no mediante arrays de CGH.

Existen diferentes plataformas y compañías que proporcionan arrays de SNPs con características similares como *Affymetrix*, *NimbleGen*, *Illumina* o *Agilent*. El flujo de trabajo desarrollado para este tipo de datos ha sido diseñado y probado con arrays de SNPs de *Affymetrix*. Sin embargo, una vez realizada la normalización en el preprocesamiento de los datos, que es dependiente de la plataforma y del tipo de arrays, el resto de los algoritmos desarrollados pueden ser aplicados independientemente del tipo específico de array y plataforma utilizados.

2.1.2.3 Hibridación *in situ* con fluorescencia, FISH

La hibridación *in situ* con fluorescencia (*Fluorescence In Situ Hybridization*, FISH) es una técnica de marcaje de cromosomas mediante la cual los cromosomas son hibridados con sondas que emiten fluorescencia y que, gracias a ello, permiten la visualización y distinción de los cromosomas así como de las anomalías que puedan presentar. La técnica FISH se basa en la utilización de fragmentos de DNA de secuencia específica, denominados sondas, que corresponden a regiones concretas del genoma y que son etiquetados con una sustancia fluorescente. Estas sondas marcadas se unen por hibridación específica al DNA y permiten identificar los cromosomas y regiones cromosómicas correspondientes utilizando microscopía fluorescente. Un esquema de FISH puede verse en la figura 2.3.

Esta técnica se emplea sobre todo para la detección de grandes deleciones, duplicaciones o translocaciones en células tumorales ya que se realiza sonda a sonda y no a escala genómica. La técnica de FISH tiene una resolución mucho menor que cualquiera de los arrays comentados anteriormente de cara al mapeo de alteraciones en todo el genoma; sin embargo, ha sido una técnica muy utilizada en estudios citogenéticos y actualmente es todavía muy usada como técnica experimental de validación de series de pacientes para pequeños conjuntos de sondas específicas seleccionadas.

2.2 Preprocesamiento: Análisis de muestras individuales

El preprocesamiento de los datos tiene un efecto significativo en el aumento de la resolución en la detección de alteraciones de número de copias de DNA alcanzable con las diferentes plataformas de microarrays de oligonucleótidos de alta densidad (Hehir-Kwa et al., 2007). El preprocesamiento incluye las operaciones preliminares sobre los datos de fluorescencia hasta llegar a obtener un valor que cuantifique el número de copias de DNA para cada región. Se lleva a cabo de manera independiente para cada una de las muestras y sin establecer asociaciones con la enfermedad estudiada o con ningún tipo de variable. Comprende tres pasos principales: **(i)** cálculo de la señal cruda normalizada, **(ii)** segmentación y **(iii)** discretización.

2.2.1 Cálculo de la señal cruda normalizada

El primer paso en el análisis de número de copias de DNA es la obtención del número de copia estimado de cada sonda a partir de los datos de fluorescencia del dispositivo microarray usado. Este paso lo denominamos cálculo de la señal cruda normalizada (*normalized raw signals*), y se puede a su vez subdividir en varios pasos que son muy dependientes del tipo de dispositivo y tecnología genómica que se use en el estudio. Incluye las correcciones respecto al ruido y a las distintas desviaciones posibles en la señal. Una vez minimizadas estas desviaciones mediante métodos de corrección y normalización, la señal de las diferentes sondas hibridadas será más fácilmente comparable. Como se ha indicado, los métodos concretos aplicados en este paso son muy dependientes de la tecnología de cuantificación de DNA utilizada y del tipo de dispositivos empleados (fabricante, modelo, etc). Sin embargo, se pueden establecer, de manera general, un conjunto de procesamientos diferentes:

1. Eliminación de la señal de fondo o *background*: estima y sustrae la hibridación residual inespecífica que puede presentar cada array o cada una de las zonas del array.
2. Normalización interna del array: intenta corregir las diferencias entre la señal de las diversas sondas para hacerlas comparables.
3. Sumarización: suma de modo robusto las señales múltiples para asignar un único valor a cada región genómica teniendo en cuenta los valores de las sondas que mapean en la misma.

Existen varios algoritmos que implementan variaciones sobre estos pasos entre los que cabe destacar dChip (Li and Wong, 2001), CNAG (Nannya et al., 2005), CRMA (Bengtsson et al., 2008) o CRMAv2 (Bengtsson et al., 2009) como los más utilizados. En este trabajo se ha utilizado CRMAv2, diseñado específicamente para los arrays de SNPs de *Affymetrix* e implementado en el paquete de R *aroma.affymetrix* (http://www.aroma-project.org). Este método de cálculo de la señal cruda normalizada incluye concretamente los siguientes pasos (Bengtsson et al., 2009):

1. Estimación y corrección de la hibridación cruzada entre la señal de los pares de sondas de los distintos alelos para un mismo SNP, que son medidos por el valor *Perfect Match* del alelo A (PMA) y el *Perfect Match* del alelo B (PMB).
2. Estimación de la afinidad de cada sonda basada en su secuencia de aminoácidos mediante el ajuste de un modelo de afinidad para cada array usando sólo un subconjunto de sondas con valores neutrales de número de copia. Los valores de intensidad son corregidos en base a la suma de los efectos individuales de cada aminoácido en cada posición de la sonda.
3. Sumarización de la señal de las k sondas réplica que existen para cada SNP j en cada muestra i en cada uno de los alelos (A y B): θ_{ij} . Esta sumarización puede realizarse mediante un modelo aditivo si las sondas para un SNP son réplicas idénticas:
   $\theta_{ijA} = median_k(PM_{ijk})$

$\theta_{ijB} = median_k(PM_{ijk})$
$\theta_{ij} = \theta_{ijA} + \theta_{ijB}$
O bien mediante un modelo multiplicativo en el que se tiene en cuenta la afinidad de cada sonda, si dichas sondas no son réplicas exactas sino que la posición del SNP está desplazada un número de nucleótidos dentro de la sonda:
$PM_{ik} = \phi_k * \theta_i + \xi_{ik}$

4. Corrección de los efectos debidos a la diferente longitud de los fragmentos de PCR. Este efecto surge al utilizar dos enzimas de restricción distintos en la fragmentación de las muestras.
5. Cálculo del número de copia crudo para cada SNP relativo a la referencia no alterada: $CN_{ij} = log_2(\theta_{ij}/\theta_{Rj})$, donde θ_{ij} y θ_{Rj} son los valores de la señal sumarizadas de las sondas para cada SNP en las muestras problema y en las referencias, respectivamente.

Como resultado de este proceso de cálculo de la señal normalizada obtenemos para cada SNP el valor estimado de su número de copia CN_{ij}, al que nos referiremos como *log2ratio* crudo.

2.2.2 **Segmentación**

Los métodos de segmentación constituyen una familia de algoritmos que han sido aplicados recientemente a datos genómicos. Un algoritmo de segmentación divide un conjunto ordenado de datos en regiones de elementos adyacentes con valores similares. A cada uno de los segmentos identificados se les asigna un único valor que representará a todos los elementos en la región. Cada perfil genómico de número de copias puede verse entonces como una sucesión de segmentos que representan regiones homogéneas en el genoma.

La aplicación de este paso en el preprocesamiento es posible debido a dos características principales de los datos de número de copias de DNA. En primer lugar, por la naturaleza de los datos de número de copias de DNA, ya que existe una relación entre las sondas próximas en el genoma. En segundo lugar, porque el proceso biológico subyacente es discreto con valores de 1 copia de la región, 2 copias, 3 copias ... (deleción o amplificación de las regiones cromosómicas de DNA) mientras que la señal que lo mide es contínua. Aprovechando estas características es posible reducir parte del ruido inherente a la tecnología y a la heterogeneidad de las células en una misma muestra mediante el agrupamiento de sondas adyacentes en los estados de número de copia posibles.

Teniendo en cuenta esta asociación los algoritmos de segmentación parten de la suposición de que dos regiones del genoma adyacentes tienen el mismo número de copias a no ser que se haya producido alguna alteración. La idea intuitiva de cambio de estados ha llevado al desarrollo de varios métodos basados en modelos de Markov ocultos (*Hidden Markov Models*, HMM) (Fridlyand et al., 2004), sin embargo existen también otras aproximaciones metodológicas para las búsquedas de estos cambios de estado. Entre los algoritmos más utilizados para la segmentación se encuentran *Circular Binary Segmentation* (CBS) (Olshen et al., 2004) (Venkatraman and Olshen, 2007), PennCNV (Wang et al., 2007), CGHseq (Picard et al., 2005) y GLAD (Hupé et al., 2004) entre otros. Como método de segmentación utilizado para el preprocesamiento de los datos previo a la aplicación de los algoritmos desarrollados se ha utilizado CBS, evaluado en estudios independientes que estiman su precisión frente a otros algoritmos (Willenbrock and Fridlyand, 2005; Lai et al., 2005).

La idea fundamental de CBS es considerar cada cromosoma como un anillo cerrado con los dos extremos del cromosoma unidos. Cada uno de estos anillos se divide en dos partes comparando el

valor del número de copia de cada una con un *t-test*:

$$Z_{ij} = \frac{(S_j - S_i)/(j-i) - (S_n - S_j + S_i)/(n-j+i)}{\sqrt{1/(j-i) + 1/(n-j-i)}} \tag{2.1}$$

para cada par de posiciones i, j. S_k es la suma de los *log2ratio* crudos desde el primer SNP hasta el *k-ésimo*. Si el valor máximo de Z_{ij} está por encima de un umbral calculado mediante *bootstrap*, entonces i y j delimitan un segmento.

El método es utilizado de manera recursiva hasta que no se identifica ningún otro segmento. Se ha utilizado la versión implementada en el paquete de R *DNAcopy* que realiza modificaciones en el *bootstrapping* optimizando el tiempo de procesamiento. Cada uno de los segmentos identificados queda descrito con la mediana de los *log2ratios* para los SNPs incluidos en el mismo. En este trabajo nos referiremos a estos valores resultado de la segmentación como *log2ratio* segmentados o valores del número de copias segmentado (sCN).

2.2.3 **Discretización**

Uno de los pasos clave del pre-procesamiento para el análisis del número de copia consiste en la discretización de los estados de las regiones cromosómicas siendo, para muestras diploides al menos de tres estados básicos: deleción o pérdida (*loss*) ($<$2 copias), no cambio o normal (*neutral*) ($=$2 copias) y ganancia (*gain*) ($>$2 copias). A estos estados se les puede añadir otro cuarto estado de amplificación (*amplification*) ($>$3 copias) para distinguir ganancias más drásticas con un número de copias superior a tres.

La asignación a estos estados constituye un tema abierto en los análisis de la alteración del número de copias de DNA y no existe aún un método que haya sido ampliamente admitido o que presente una clara ventaja frente al resto.

La mayoría de los estudios de CNAs en cáncer se basan en la dispersión de los datos respecto a un valor de centralidad, que se espera que sea un valor de no cambio de número de copia o normal. Así establecen umbrales basados en las desviaciones estándar sobre la media, como es el caso de (Aguirre et al., 2004) o (Tonon et al., 2005) que definen las ganancias y pérdidas como *log2ratios* $>$ 4 desviaciones estándar sobre el cuantil 50 % de los datos.

Esta asignación puede realizarse directamente sobre los *log2ratios* crudos como en el caso anterior o aplicarse sobre datos segmentados. Los investigadores Willenbrock and Fridlyand (Willenbrock and Fridlyand, 2005) determinaron que la discretización sobre valores previamente segmentados o suavizados es preferible frente a la asignación directa sobre los *log2ratios* crudos. En el trabajo mencionado estiman la variabilidad experimental con un estadístico más robusto, la desviación absoluta de la mediana (MAD, *Median Absolute Deviation*). Calculan este estadístico para las diferencias entre los valores de *log2ratios* segmentados (sCN) y crudos (CN). Así, para cada SNP i ($i = 1, \ldots, n$) tenemos $X_i = |sCN_i - CN_i|$, cuya desviación absoluta de la mediana sería:

$$MAD = median_i(|X_i - median(X)|) \tag{2.2}$$

Calculados los valores de MAD, dichos autores establecen entonces un umbral de 3 veces la MAD (aunque utilizan 2.5 en el conjunto de datos real) para la asignación de ganancias, pérdidas o no cambio. Otros trabajos que analizan alteraciones genómicas en cáncer, como (Fridlyand et al., 2006), utilizan también este valor MAD para la discretización.

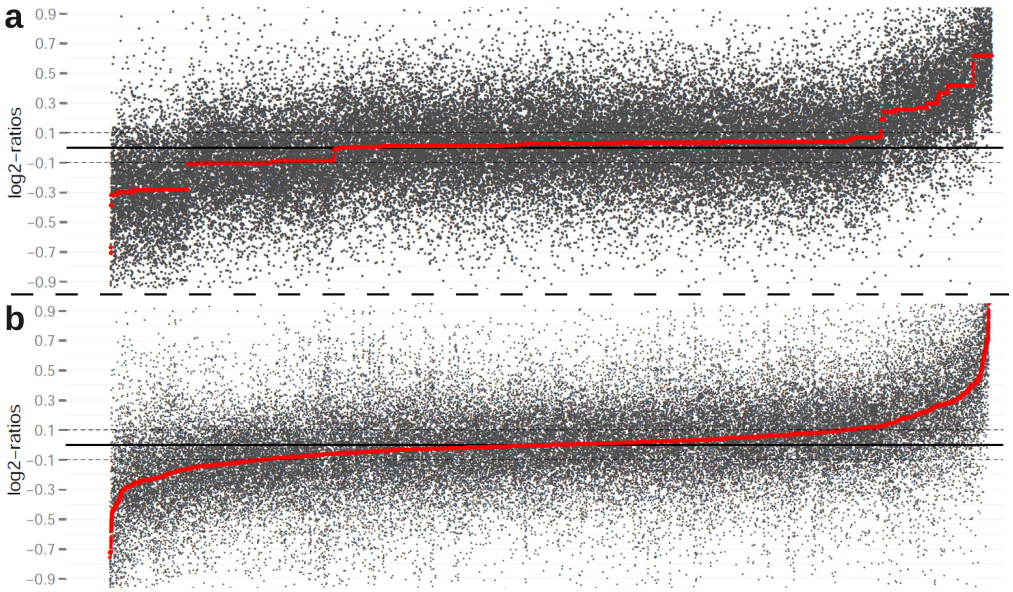


Figura 2.4: Representación de los segmentos de todo el genoma ordenados por valor creciente de *log2ratio* - Los puntos rojos representan los valores de *log2ratio* de los datos segmentados (sCN). Los puntos grises representan los valores de *log2ratio* crudos (CN) para cada sonda en cada segmento. **(a)** Segmentos en una única muestra. **(b)** Segmentos para un conjunto de muestras tumorales.

Atendiendo únicamente a los valores segmentados (sCN) es posible postular que aquellas regiones en las que no se ha producido ninguna alteración el valor medio del segmento se corresponderá con un *log2ratio* en torno a 0, que indicaría el no cambio. Los segmentos con valores de *log2ratio* > 0 se corresponderán con regiones ganadas y los segmentos con valores < 0 se corresponderán a su vez con regiones perdidas en la muestra problema frente a la referencia. Como se puede apreciar en la parte **(a)** figura 2.4 si ordenamos los segmentos de acuerdo a sus *log2ratios* de modo creciente, de menor a mayor, se observa una función con saltos. Es razonable asumir que estos saltos se corresponden con un número de copias particular (Olshen et al., 2004). Sin embargo, al ordenar los segmentos no sólo de una muestra sino de todo el conjunto de muestras los escalones se suavizan (figura 2.4 **(b)**). Este suavizado, en el que sucede una pérdida de escalones y umbrales claros, puede ser más evidente cuando los datos provienen de muestras clínicas debido principalmente a la variabilidad de los individuos, a la heterogeneidad de los tumores analizados o a la diferencia en los porcentajes de infiltración tumoral (es decir, a la existencia de diferentes porcentajes de células tumorales en cada una de las muestras). Cuando se produce este efecto de suavizado, sin que se pueda determinar una causa concreta conocida, la identificación de los umbrales se puede realizar mediante a una búsqueda analítica de los puntos de inflexión de la curva.

Junto a la aproximación descrita para la búsqueda de umbrales, la discretización se puede acometer de otro modo planteándola como un problema de agrupamiento (*clustering*) de los segmentos en los tres estados posibles predefinidos de número de copias de DNA: ganancia, normalidad y pérdida de la región genómica.

Para lograr esta discretización basada en agrupamiento, proponemos la aplicación de un algoritmo de agrupamiento no supervisado con un número de estados predefinidos como es *k-means* (Lloyd, 1982). Este algoritmo se puede considerar una variante del algoritmo de Esperanza-Maximización (*Expectation-Maximization*, EM) en el que para un conjunto de observaciones (valor medio de los segmentos) $(x_1, x_2, ..., x_n)$ y dado un conjunto inicial de semillas $m_1^{(1)}, ..., m_k^{(1)}$ procede en dos

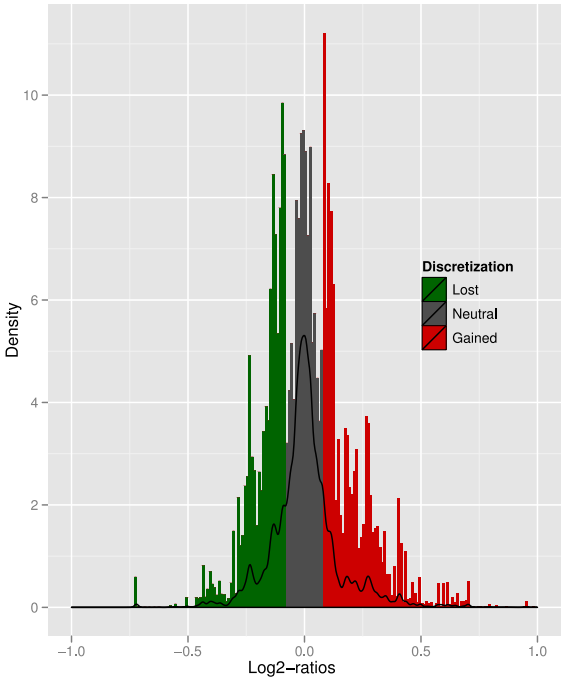


Figura 2.5: Distribución de densidad de los valores de *log2ratios* segmentados de un conjunto de muestras de cáncer colorectal - Histograma y distribución de densidad de *log2ratios* segmentados. Marcados en verde, gris y rojo los intervalos del histograma catalogados como pérdidas, normalidad y ganancias respectivamente.

pasos independientes:

1. **Paso E** (o paso de asignación), en el que cada segmento es asignado a la media más cercana:

$$S_i^{(t)} = \{x_p : \|x_p - m_i^{(t)}\| \leq \|x_p - m_j^{(t)}\| \forall 1 \leq j \leq k\} \tag{2.3}$$

2. **Paso M** (o paso de actualización), en el que se recalcula la media que será el centroide de cada clúster:

$$m_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} \sum_{x_j \in S_i^{(t)}} x_j \tag{2.4}$$

De esta manera los segmentos se agrupan en k clústers, donde k es el número de estados de número de copias de DNA a considerar.

La figura 2.5 muestra la distribución de densidad de los valores de *log2ratios* segmentados en el que se diferencian con distintos colores las regiones que incluyen los estados de ganancia (*gained*, rojo), normalidad (*neutral*, gris) y pérdida (*lost*, verde) de copias de DNA. El número de clústers o grupos buscado puede variar, incluyendo por ejemplo estados para deleciones (pérdida total de la región cromosómica) y amplificaciones (ganancia de más de 1 copia de DNA para una región) además de los 3 estados definidos anteriormente.

Conviene mencionar que si se establecen únicamente 3 grupos, debido a la naturaleza del algoritmo de *k-means* que busca el centroide de cada clase, es necesario tener en cuenta la existencia de regiones con un número muy alto de copias (más de 3 ó 4). Estas regiones pueden modificar en el centroide del estado de ganancia provocando el desplazamiento del grupo y la modificación del umbral esperado. Es necesario descartar las regiones extremas o atípicas para que no afecten en exceso al establecimiento de umbrales. La determinación de estas regiones o *outliers* se lleva a cabo utilizando la definición de (Tukey, 1977) que considera *outliers* a aquellos valores comprendidos fuera del rango definido por:

$$[Q_1 - k(Q_3 - Q_1), Q_3 + k(Q_3 - Q_1)] \tag{2.5}$$

donde Q_1 y Q_3 son los cuartiles inferior y superior de las distribución de *log2ratios* y k es una constante definida como $k = 1{,}5$. Gráficamente estos valores extremos pueden verse representados en la figura 2.6.

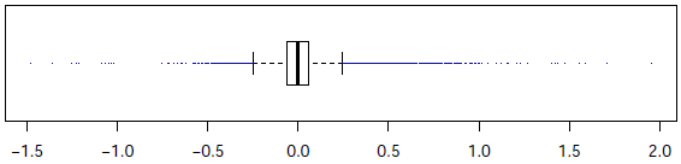


Figura 2.6: Boxplot de los valores de *log2ratios* segmentados para un conjunto de muestras de cáncer colorectal. - Representados en azul los puntos considerados *outliers* según la regla de *Tukey*.

2.3 Análisis unificado de conjuntos de muestras

La localización de alteraciones en el número de copias de DNA en muestras individuales como la realizada en los pasos anteriores constituye sólo el inicio en la determinación de regiones y genes críticos en una enfermedad. Para extraer conclusiones extrapolables sobre genes desregulados en una determinada enfermedad o en condiciones particulares la determinación de las alteraciones en muestra individual no es suficiente, y es necesario el análisis simultáneo de un conjunto de muestras.

Para este análisis unificado se han desarrollado algoritmos que detectan regiones alteradas recurrentemente (*Minimal Common Regions*, MCR), o regiones con puntos de ruptura recurrentes (*Recurrent Breakpoints Regions*, RBR) y algoritmos para el análisis diferencial de regiones alteradas.

2.3.1 Detección de regiones mínimas comunes (MCR) de alteración

Las regiones más fuertemente asociadas a un estado patológico son aquellas que están alteradas de manera recurrente o común en los individuos o muestras de dicho estado. Es probable que las alteraciones más frecuentes sean las que producen los cambios funcionales importantes para la progresión y el desarrollo de la enfermedad, mientras que las alteraciones que ocurren en un pequeño subconjunto de muestras se deban a efectos individuales no comunes sin relevancia para el proceso patológico. Bajo esta hipótesis el análisis de las CNA en una determinada enfermedad se centrará en la búsqueda de regiones comunes o recurrentes, más concretamente de las regiones mínimas alteradas de forma recurrente en el conjunto de muestras (*Minimal Common Regions*, MCR).

Para la búsqueda de las MCR se ha diseñado un método basado en la frecuencia de alteración de la regiones similar al propuesto en (Aguirre et al., 2004) y (Tonon et al., 2005). La selección de las MCR comprende los siguientes pasos:

1. Asignación de la frecuencia de alteración de cada región cromosómica, de modo independiente para las ganancias y para las pérdidas.
2. Selección de aquellas regiones con una frecuencia de alteración por encima de un umbral (*z-score* $> 2{,}1$) como regiones significativas candidatas a ser MCR. Serán descartados aquellos cromosomas que no contengan ninguna región significativa.

3. Búsqueda de los picos o sub-regiones con mayor recurrencia en las regiones significativas como regiones candidatas.
4. Se identificarán como MCR las regiones candidatas con una frecuencia de al menos el 80 % de la frecuencia de alteración máxima en el cromosoma y que comprendan al menos 5 sondas.

La caracterización de estas regiones comunes comprende también la anotación de la amplitud de la alteración utilizando para ello la mediana del valor de *log2ratio* de la región en las muestras que presentan la alteración (es decir, aquellas cuyos *log2ratios* están por encima o por debajo del umbral de discretización) considerando también de modo independiente las ganancias/amplificaciones y las pérdidas/deleciones.

2.3.2 Detección de regiones con puntos de ruptura (*breakpoints*) recurrentes

Las CNAs no siempre tienen como objetivo la desregulación de un gen incluido en la región alterada, muchas veces el objetivo es la desregulación del gen o locus génico en la frontera de dicha. Un ejemplo es el gen de fusión TMPRSS2-ERG en cáncer de próstata (Tomlins et al., 2005). Este gen de fusión es el resultado de una deleción en el cromosoma 21 en la que los dos genes implicados se sitúan en los extremos de la región. Los límites de las regiones con cambios en el número de copias determinan localizaciones en las que se ha producido la ruptura del cromosoma, pero existen también otros reordenamientos como traslocaciones e inversiones en los que no se modifica "aparentemente" en el número de copias. Sin embargo, incluso en el caso de este tipo de reordenamientos balanceados se producen pequeñas deleciones en las regiones adyacentes (Kolomietz et al., 2001; Watson et al., 2007). Esto hace posible el análisis de los puntos de ruptura a partir de los datos de arrays de SNPs.

Definimos entonces punto de ruptura o *breakpoint* como la transición de un estado de número de copia (ganancia, no cambio o pérdida) a otro. A partir de esta definición y de manera análoga a las MCR establecemos las regiones con puntos de ruptura recurrentes como aquellas regiones cromosómicas que de manera recurrente o frecuente presentan transiciones de un estado a otro en las diferentes muestras.

El estudio y determinación de estas regiones con puntos de ruptura recurrentes es importante a la hora de identificar genes que posiblemente estén alterados en la enfermedad. Para ello hemos diseñado un algoritmo que analiza la densidad de puntos de ruptura e identifica regiones cromosómicas con alta densidad en base al número de muestras diferentes que presentan puntos de ruptura en la misma región.

La determinación de las regiones con puntos de ruptura recurrentes se realiza para cada cromosoma independientemente. La figura 2.7 esquematiza el proceso de definición de regiones candidatas con puntos de ruptura recurrentes en el cromosoma 17 para un conjunto de muestras de cáncer. El proceso consiste en los siguientes pasos:

1. Cálculo del número total de puntos de ruptura N que corresponden a cualquier transición de estado (ganancia/neutral/perdida) en todas las muestras.
2. División del cromosoma en intervalos del mismo tamaño. El número de intervalos K se determina mediante la regla de *Sturges*
3. Cálculo de la frecuencia de puntos de ruptura en cada intervalo. En el cálculo de estas frecuencias se asignarán pesos diferentes si una misma muestra presenta varios puntos de ruptura dentro del mismo intervalo.

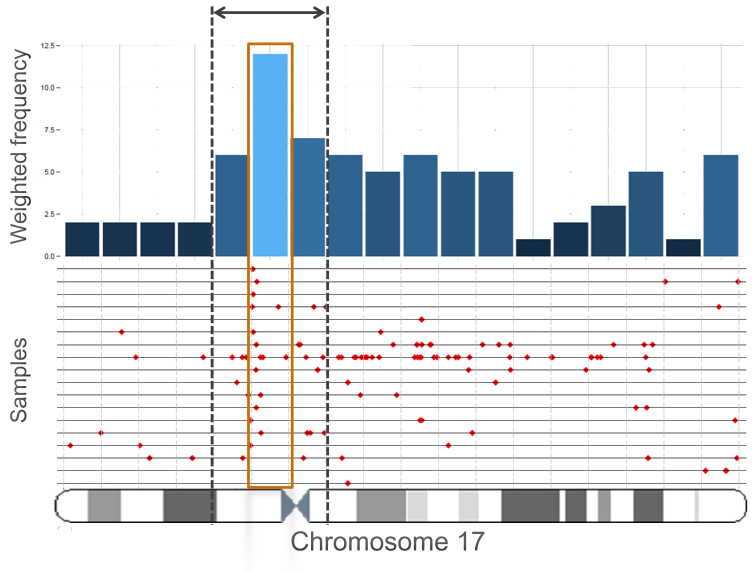


Figura 2.7: Esquema del algoritmo para la detección de puntos de ruptura recurrentes - Puntos de ruptura en el cromosoma 17 para muestras de cáncer colorectal representadas en sus localizaciones cromosómicas (puntos rojos). Las barras azules representan las frecuencias de puntos de ruptura en el intervalo cromosómico ponderadas teniendo en cuenta que una muestra puede presentar varios puntos de ruptura. El recuadro rojo señala el intervalo con una probabilidad de puntos de ruptura significativamente diferente al resto de intervalos y las líneas verticales discontinuas representan la región extendida a los intervalos adyacentes del intervalo seleccionado.

4. Determinación de los intervalos con una frecuencia significativamente diferente del resto de intervalos. Estos intervalos significativos se expandirán a los intervalos adyacentes para constituir las regiones candidatas a contener los puntos de ruptura recurrentes.
5. Refinamiento de las regiones candidatas mediante la aplicación recursiva del método. Estas regiones candidatas serán divididas en intervalos del mismo tamaño recalculándose las frecuencias ponderadas de puntos de ruptura en cada uno de ellos. El algoritmo procederá hasta que no existan diferencias significativas en las frecuencias de los intervalos. La región candidata será entonces marcada como RBR.

El algorimo desarrollado se presenta en mayor detalle en el siguiente pseudocódigo, que recibe como parámetros de entrada S, P y B definidos como:

- $S = \{s_1, \ldots, s_m\}$, Un conjunto de m muestras definidas por las localizaciones de sus puntos de ruptura respectivos $s_i = \{b_1, ..., b_r\}$.
- $P = [p_1, p_2]$, Las posiciones de inicio y fin de la región genómica que se va a considerar. Al comienzo del algoritmo corresponde a las posiciones de inicio y fin del cromosoma de estudio.
- $B = \bigcup_{i=1}^{m} s_i = \{b_1, \ldots, b_n\}$, El conjunto de las localizaciones genómicas (*loci*) de todos los puntos de ruptura n en la región delimitada por P. Al comienzo del algoritmo este conjunto corresponde a todos los puntos de ruptura del cromosoma de estudio.

Algorithm 1 Algoritmo para la detección de regiones con puntos de ruptura recurrentes

procedure DIVIDEINTERVAL(P,n) ▷ Divide the interval according to the Sturges rule
  $K \leftarrow 1 + 3{,}22 \log |B|$
  $R \leftarrow (p_2 - p_1)/K$ ▷ Range of the intervals
  $P'_1 \leftarrow [p_1, p_1 + R)$
  for $i = 2$ to $K - 1$ **do**
    $P'_i \leftarrow [p_1 + (i-1)R, p_1 + iR)$
  end for
  $P'_K \leftarrow [p_1 + (k-1)R, p_2]$
  return $P' = \{P'_1, \ldots, P'_K\}$
end procedure

procedure RECURRENTBREAKS(S, P, B)
  $P' \leftarrow$ DIVIDEINTERVAL(P,$|B|$) ▷ $|B| = n$: total number of breakpoints
  $m \leftarrow |S|$ ▷ Total number of samples
  $K \leftarrow |P'|$ ▷ K: number of intervals
  for $i = 1$ to K **do**
    $B'_i \leftarrow \{b \in B / b \subset P'_i\}$
    for $j = 1$ to m **do**
      $x_{ij} \leftarrow count(\{b \in s_j / b \subset P'_i\})$
    end for
    d-score$_i \leftarrow \sum_{j=1}^{m} \log(x_{ij} + 1)$ ▷ Weighted frequency
  end for
  for $i = 1$ to K **do**
    $z_i \leftarrow$ (d-score$_i - mean$(d-score))$/sd$(d-score)
  end for
  if $notAny(z > 2)$ **then** ▷ Recursivity end condition
    return P
  else
    $breakRegions \leftarrow \emptyset$
    for $i = 1$ to K **do**
      if $z_i > 2$ **then**
        $regionStart \leftarrow i$
        while $z_i > 2$ **do**
          $i \leftarrow i + 1$
        end while
        $regionEnd \leftarrow i$
        $breakRegions \leftarrow breakRegions \cup [regionStart, regionEnd]$
      end if
    end for
    for all $breakRegions$ **do**
      $P \leftarrow P'_{regionStart-1} \cup P'_{regionEnd+1}$ ▷ Extend candidate region to adjacent intervals
      $B \leftarrow B'_{regionStart-1} \cup B'_{regionEnd+1}$
      return RECURRENTBREAKS(S, P, B)
    end for
  end if
end procedure

2.4 Aplicación a datos de cáncer colorectal (CRC)

El cáncer colorectal o cáncer de colon (*Colo-Rectal Cancer*, CRC) es el segundo tipo de cáncer más frecuente en los países desarrollados. Este tipo de cáncer se suele desarrollar lentamente durante muchos años. La mayoría comienzan como un pólipo o abultamiento en la mucosa del intestino grueso y crecen hacia el centro del colon o el recto transformándose un adenocarcinoma o tumor maligno. El desarrollo comprende cuatro etapas diferenciadas que se caracterizan por el tamaño del tumor, cuánto ha penetrado en la mucosa, si ha invadido órganos adyacentes y cuántos ganglios linfáticos ha afectado. Los órganos adyacentes más frecuentemente afectados por los procesos metastásicos del cáncer de colon son el hígado y los pulmones.

La estratégia de análisis y los métodos desarrollados comentados en las secciones 2.2 y 2.3 han sido aplicados a un conjunto de datos de cáncer colorectal con metástasis en el hígado. Este conjunto de datos experimentales consiste en un total de 23 muestras de tumores primarios metastásicos además de otras 23 muestras de las metástasis respectivas en hígado que fueron hibridadas en microarrays de SNPs (*250K Affymetrix SNP Mapping NspI, StyI arrays*). Como controles se hibridaron muestras pareadas de la sangre periférica de los mismos pacientes. Más información sobre los datos y detalles sobre su obtención se encuentran en (Sayagués et al., 2010) y (Muñoz Bellvis et al., 2012).

2.4.1 Evaluación de los métodos de discretización aplicados a CRC

El establecimiento de los umbrales para la discretización es un paso crítico en el análisis de CNA, puesto que los estados definidos constituyen el punto de partida en los análisis de recurrencia posteriores. En nuestro estudio sobre tumores primarios metastásicos de CRC, la evaluación y determinación de los umbrales de ganancia, pérdida y normalidad se ha realizado estableciendo como *Gold Standard* datos del número de copia de diversas regiones cromosómicas obtenidas experimentalmente mediante FISH (2.1.2.3). Los datos de FISH se correponden al uso de 45 sondas hibridadas para cada muestra, distribuidas a lo largo de los 22 autosomas humanos. De estas 45 sondas se han seleccionado 24 que mapean en al menos 10 SNPs para obtener unos resultados comparables a los datos de los arrays de SNPs. Para esta comparación se ha tomado la mediana de los SNPs que mapean en cada una de las sondas para cada array, discretizándolo en los 3 estados definidos: ganancia, pérdida o no cambio.

Para establecer el valor óptimo de los umbrales de determinación de ganancias y pérdidas se han construido curvas de sensibilidad y especificidad incrementando progresivamente los umbrales. Estas curvas pueden verse en la figura 2.8. El punto en el que se cruzan las curvas de sensibilidad y especificidad representa el umbral c cuyo valor maximiza conjuntamente las dos funciones. Este umbral c puede estimarse como el valor que minimiza la diferencia entre la sensibilidad y la especificidad, es decir, el umbral para el cual la diferencia entre ambos parámetros se hace cero (*Minimized Difference Threshold*, MDT) (Jimenez-Valverde and Lobo, 2007).

Los valores obtenidos para el parámetro MDT en las muestras de CRC analizadas se corresponde con un umbral en torno a 0.06 para la determinación de ganancias (*Gain threshold*) y en torno a -0.07 para la determinación de pérdidas (*Loss threshold*). Como se aprecia en la tabla 2.4.1 los umbrales óptimos derivados de los datos de FISH no difieren demasiado de los umbrales establecidos por los métodos comentados en la sección 2.2.3 que son: **(i)** cálculo de la desviación absoluta de la mediana (MAD); **(ii)** agrupamiento en 3 estados (pérdida, no cambio y ganancia de número de copias) sin eliminar *outliers*; **(iii)** agrupando en 3 estados con eliminación de *outliers* y **(iv)** agrupamiento en 4 estados (incluyendo amplificación). Sin embargo, si se consideran únicamente

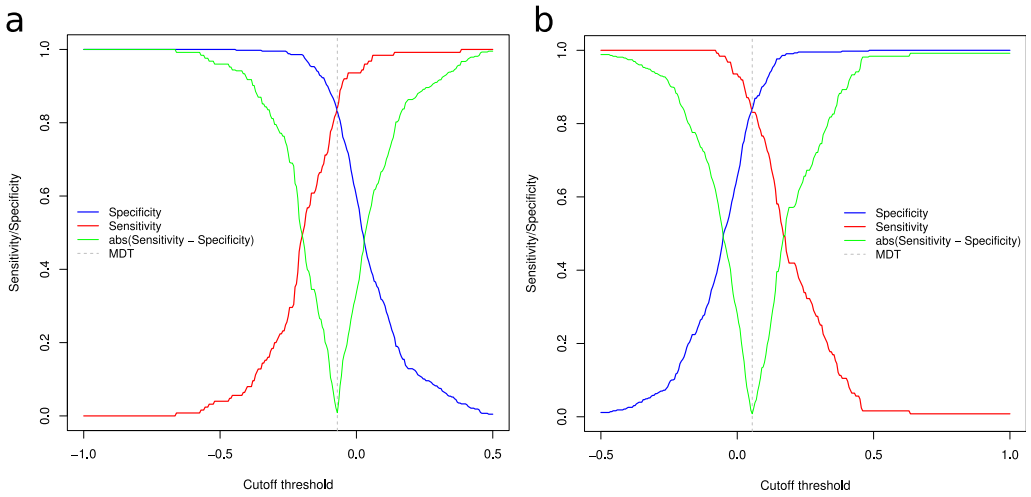


Figura 2.8: Curvas de sensibilidad y especificidad para distintos umbrales de discretización - Curvas de especificidad, sensibilidad y valor absoluto de la diferencia de ambas marcando el punto en el que se minimiza esta diferencia (*Minimized Difference Threshold*, MDT): **(a)** umbrales para la determinación de pérdidas y **(b)** umbrales para la determinación de ganancias.

3 estados sin eliminar los *outliers*, el umbral de ganancias se ve afectado por las amplificaciones y se sobreestima (valor 0.16) produciéndose una desviación mayor respecto al valor óptimo observado experimentalmente.

	Loss threshold	Gain threshold	Loss Sensitivity /Specificity	Gain Sensitivity /Specificity
Minimized Difference Threshold (MDT)	-0.07	0.06	0.84 / 0.83	0.83 / 0.85
(i) Median Absolute Deviation (MAD)	-0.10	0.10	0.74 / 0.89	0.76 / 0.91
(ii) Three states with outliers	-0.08	0.16	0.82 / 0.86	0.53 / 0.98
(iii) Three states without outliers	-0.07	0.06	0.84 / 0.83	0.83 / 0.85
(iv) Four states	-0.10	0.08	0.74 / 0.89	0.79 / 0.87

Tabla 2.1: Umbrales de discretización obtenidos con diferentes métodos.

Con el estudio descrito tenemos una clara estimación del grado de especificidad y sensibilidad esperados para unos umbrales determinados de ganancia y pérdida. La elección de unos umbrales que optimicen a la vez tanto la especificidad y como la sensibilidad puede no ser siempre la mejor opción, ya que en muchos casos interesa tener una especificidad no menor del 95 % (es decir, no admitir más de un 5 % de falsos positivos). Por ello, en el análisis concreto del conjunto de muestras de CRC se ha optado por una aproximación más conservadora estableciendo el umbral en el valor que obtiene una especificidad del 95 % (ver línea azul en la figura 2.8). De este modo, se determinó usar como umbrales 0.09 y -0.09 para ganancias y pérdidas, respectivamente.

2.4.2 Identificación de regiones de alteración recurrente en CRC

Como se ha mencionado previamente, la probabilidad de que una región cromosómica contenga genes críticos para una enfermedad está relacionada con la recurrencia de la alteración en los individuos con dicha enfermedad. El análisis de la frecuencia de alteración de las regiones es muy importante a la hora de encontrar regiones que contengan genes ”conductores” que dirijan el desarrollo y progresión de la enfermedad.

Los análisis de recurrencia de alteración se realizaron con el conjunto de muestras de CRC citado, incluyendo en este caso datos de 20 pacientes. La figura 2.9 **a** presenta un *heatmap* con los valores de *log2ratio* segmentados (sCN) para cada muestra. En el *heatmap* se aprecia que hay ciertas regiones que aparecen frecuentemente alteradas en la mayoría de muestras. Esto se aprecia mejor en el mapa con las frecuencias de alteración para cada región mostrado en la figura 2.9 **b**. Son particularmente interesantes los picos con las frecuencias más altas, calculados con el algoritmo de búsqueda de MCR (*Minimal Common Regions*) comentado en la sección 2.3.1. Las MCR con ganancia de número de copias o amplificación se corresponden, para los tumores de CRC, a regiones localizadas en los cromosomas: 7, 8q, 13q y 20q, (con un incremento del número de copias en más del 60 % de los tumores). Por su parte, los cromosomas 1p, 8p, 17p y 18 presentan pérdidas muy recurrentes (presentes en más del 70 % de los tumores para 1p, 17p y 18 y más del 65 % para 8p).

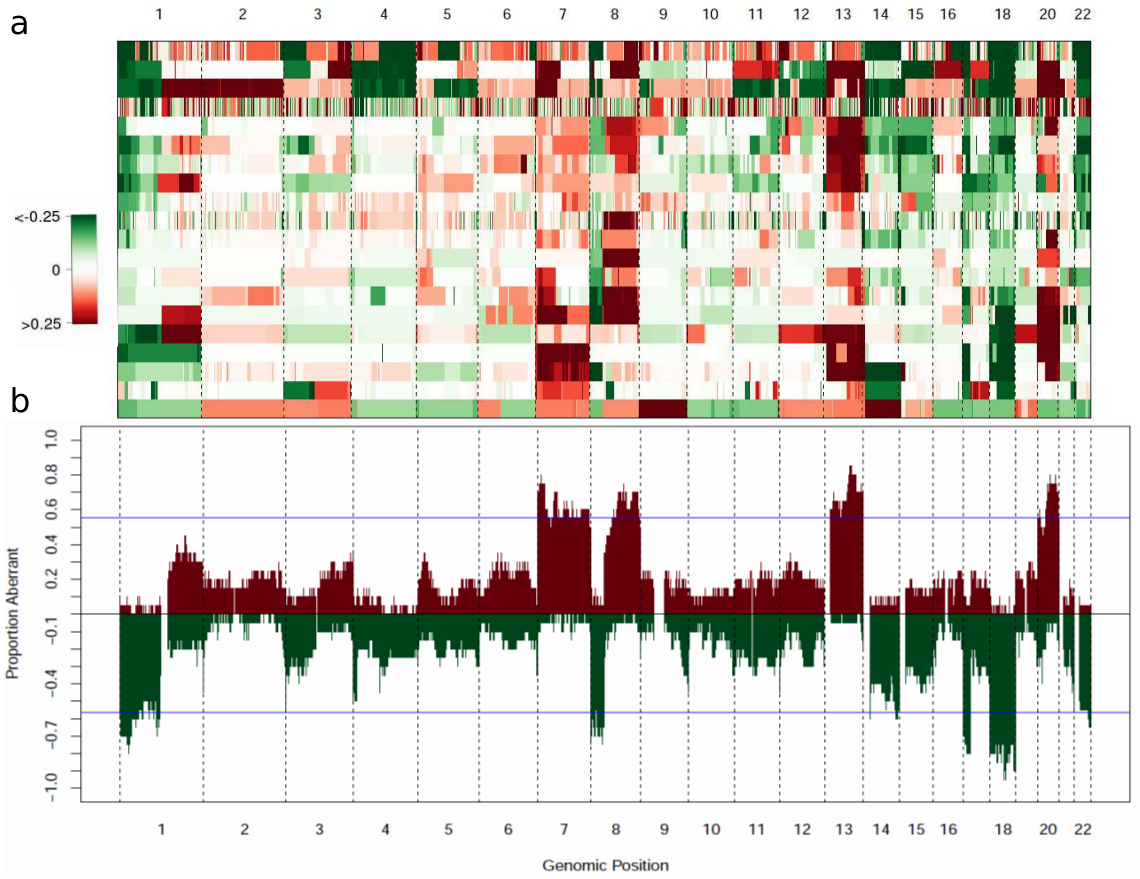


Figura 2.9: Heatmap y frecuencias de alteración de muestras de CRC - Resumen con las alteraciones más frecuentes encontradas en CRC. El eje x representa las localizaciones genómicas en cada uno de los cromosomas situados en orden creciente uno a continuación de otro. **(a)** Heatmap con los valores de número de copia (*log2ratio* segmentados) en cada región para cada muestra de CRC. **(b)** En rojo (por encima de 0), frecuencias de ganancias de número de copias de DNA; en verde (por debajo de 0), las frecuencias de pérdida de número de copias de DNA.

La mayoría de estas MCR se corresponden con regiones pequeñas ($<$ 1Mb) que contienen al menos un gen candidato a ser explorado como gen de interés. Bastantes de estos genes han sido previamente asociados con CRC o con metástasis tumoral. Las tablas 2.4.2 y 2.4.2 contienen estas regiones $<$ 1Mb para las ganancias y las pérdidas de número de copia respectivamente. En estas tablas se muestra información más detallada sobre la localización de la región, su tamaño y el

número de SNPs que incluyen, la recurrencia (mostrada como porcentaje de muestras que presentan alteración), la mediana del *log2ratio* segmentado en las muestras que presentan la alteración y los símbolos identificativos de los genes incluidos en la región.

Region	Region length (Kb)	Number of SNPs	Recurrence (% samples)	Median log2ratio	Genes
Chr1: 7334675 - 7347651	13	7	70	-0.21	-
Chr1: 26131131 - 26191419	60	16	74	-0.18	PAFAH2
Chr1: 26194668 - 26738275	544	70	70	-0.19	PAFAH2, SCARNA17, SCARNA18, EXTL1, TRIM63, PDIK1L, GRRP1, ZNF593, CNKSR1, CATSPER4, CCDC21, UBXN11, CD52, AIM1L, ZNF683, LIN28, DHDDS, HMGN2
Chr1: 29528243 - 29633233	105	17	70	-0.20	-
Chr1: 29894172 - 29922153	28	17	70	-0.21	-
Chr1: 30778256 - 30843796	66	28	70	-0.18	-
Chr8: 198834 - 392556	194	46	70	-0.26	FAM87A, FBXO25
Chr8: 400640 - 539716	139	29	70	-0.26	FBXO25, C8orf42
Chr8: 11285446 - 11566690	281	85	65	-0.17	C8orf12, C8orf13, BLK, C8orf14
Chr8: 23264737 - 23277681	13	8	70	-0.16	-
Chr8: 23295574 - 23519768	224	68	65	-0.17	LOXL2, ENTPD4, SLC25A37
Chr8: 29372156 - 29426943	55	15	65	-0.17	-
Chr8: 31294142 - 31566697	273	41	65	-0.17	-
Chr8: 32105734 - 32675812	570	196	70	-0.15	-
Chr8: 32692415 - 33064331	372	67	65	-0.15	NRG1
Chr17: 5965727 - 6261243	296	73	74	-0.21	WSCD1
Chr17: 6634969 - 6666473	32	12	74	-0.22	TEKT1
Chr17: 10693238 - 11021844	329	89	78	-0.22	-
Chr17: 14234746 - 14967525	733	214	78	-0.23	-
Chr17: 14984724 - 15082587	98	18	78	-0.23	PMP22
Chr17: 19622919 - 20156497	534	66	74	-0.23	ULK2, AKAP10, CYTSB
Chr17: 20792902 - 20844368	51	8	74	-0.23	USP22
Chr18: 41130655 - 41494986	364	134	91	-0.20	SLC14A2
Chr18: 44872826 - 45204796	332	32	87	-0.21	-
Chr18: 45410728 - 45497910	87	29	91	-0.21	-
Chr18: 45654114 - 46036475	382	144	91	-0.21	MYO5B, CCDC11
Chr18: 46252199 - 46288353	36	12	91	-0.22	-
Chr18: 46291648 - 46467497	176	53	87	-0.22	MAPK4
Chr18: 69742176 - 69938053	196	38	87	-0.23	FBXO15
Chr18: 74952886 - 75029618	77	18	87	-0.23	-

Tabla 2.2: MCR con delecciones recurrentes en CRC

El mismo algoritmo de detección de regiones mínimas comunes ha sido aplicado a las muestras de las metástasis en hígado de los mismos pacientes. Se han identificado MCR con ganancia en el número de copias en regiones localizadas en los cromosomas 7p, 8q24, 13q y 20q con incremento del número de copias de DNA en más del 70 % de las muestras. A su vez se han identificado también regiones con pérdidas recurrentes en los cromosomas 1p (70 %), 8p23 (70 %), 17p (90 %), 18q22 (95 %) y 22q13 (70 %). En estas regiones se localizan oncogenes y genes supresores tumorales previamente asociados con procesos metastáticos. Las tablas con la información a cerca de las regiones identificadas y sus genes no se han añadido en la memoria por simplicidad, pero se muestran detalladamente en las tablas 1 y 2 en (Muñoz Bellvis et al., 2012): *"Unique genetic profile of sporadic colorectal cancer live metastasis versus primary tumors as defined by high-density single-nucleotide polymorphism arrays"* que ha sido añadido como apéndice en la presente memoria.

En el análisis de MCR se detectan también regiones más extensas ($>$ 1Mb) recurrentemente alteradas que comprenden decenas de genes en los cromosomas 8q, 17p y 22q. Estas regiones están descritas en más detalle en (Sayagués et al., 2010). En estas regiones de mayor tamaño no es posible identificar aquellos genes que puedan ser candidatos a dirigir el proceso tumoral utilizando únicamente la información del número de copias de DNA. Para esta discriminación e identificación

Region	Region length (Kb)	Number of SNPs	Recurrence (% samples)	Median log2ratio	Genes
Chr7: 6633026 - 6709622	77	11	70	0.18	ZNF316, ZNF12
Chr7: 7957012 - 7981482	24	10	70	0.18	GLCCI1
Chr7: 8109452 - 8142826	33	11	70	0.20	ICA1
Chr7: 8255230 - 8280496	25	10	74	0.18	ICA1
Chr7: 9676276 - 9690241	14	8	70	0.20	-
Chr7: 10461770 - 10486412	25	8	74	0.18	-
Chr7: 12514442 - 12576898	62	9	74	0.18	SCIN
Chr7: 12579777 - 12725149	145	20	70	0.20	SCIN, ARL4A, SNORA64
Chr7: 20303440 - 20340777	37	14	70	0.20	ITGB8
Chr7: 20660167 - 20868295	208	37	70	0.18	ABCB5, SP8
Chr8: 86214670 - 86946337	732	52	65	0.20	LRRCC1, E2F5, CA13, CA1, CA3, CA2
Chr8: 87377186 - 87789535	412	65	65	0.20	WWP1, FAM82B, CPNE3, CNGB3
Chr8: 88872540 - 89066702	194	24	65	0.20	WDR21C
Chr8: 91686333 - 91735940	50	10	65	0.22	TMEM64
Chr8: 94759374 - 95077320	318	44	65	0.20	RBM12B, C8orf39, TMEM67, PPM2C
Chr8: 95294349 - 95435061	141	28	65	0.19	GEM
Chr8: 95593385 - 95776644	183	36	65	0.20	KIAA1429, RBM35A
Chr8: 101919388 - 102577101	658	69	61	0.23	YWHAZ, GRHL2
Chr8: 122649759 - 122760879	111	21	61	0.22	HAS2
Chr8: 125380146 - 125811489	431	87	61	0.21	TMEM65, TRMT12, RNF139, TATDN1, NDUFB9, MTSS1
Chr8: 128638191 - 128724583	86	25	65	0.23	-
Chr8: 129180096 - 129268067	88	43	65	0.21	mir-1208
Chr8: 130906244 - 131222249	316	35	65	0.23	FAM49B, SNORA25, ASAP1
Chr8: 133845345 - 133868639	23	9	65	0.23	PHF20L1
Chr8: 133882656 - 133900665	18	6	65	0.23	-
Chr8: 133913690 - 133953202	40	10	61	0.22	PHF20L1, TG
Chr8: 134275461 - 134665414	390	113	61	0.23	WISP1, NDRG1, ST3GAL1
Chr8: 135527585 - 135836235	309	97	65	0.20	ZFAT
Chr8: 136498075 - 136866133	368	74	65	0.20	KHDRBS3
Chr8: 137055200 - 137091177	36	12	65	0.21	-
Chr13: 71257370 - 71353336	96	17	74	0.25	DACH1
Chr13: 73603130 - 73627939	25	10	78	0.25	KLF12
Chr13: 74972248 - 75117835	146	26	78	0.25	COMMD6, UCHL3, LMO7
Chr13: 76352482 - 76366765	14	11	78	0.25	KCTD12
Chr13: 78098212 - 78143588	45	7	78	0.25	C13orf7
Chr13: 78805700 - 79077299	272	46	78	0.21	RBM26, NDFIP2
Chr13: 79621013 - 79845948	225	40	78	0.21	SPRY2
Chr20: 29309964 - 29460709	151	22	78	0.25	DEFB115, DEFB116, DEFB117, DEFB118, SNORA40, DEFB119
Chr20: 33776127 - 33954944	179	21	78	0.28	RBM39, PHF20, COX7BP2
Chr20: 37766095 - 38339016	573	131	83	0.24	HSPEP1
Chr20: 43550795 - 43804853	254	82	78	0.25	SPINT3, SPINLW1, WFDC8, WFDC9, WFDC10, WFDC11, WFDC13, SPINT4
Chr20: 44503157 - 44574807	72	20	78	0.28	ZNF663, ZNF840, ZNF334
Chr20: 46929063 - 47265311	336	55	78	0.26	ARFGEF2, SNAP23P, CSE1L, STAU1
Chr20: 51015479 - 51116032	101	25	78	0.25	TSHZ2
Chr20: 54234467 - 54365833	131	40	78	0.26	MC3R

Tabla 2.3: MCR con ganancias o amplificaciones recurrentes en CRC.

resulta necesario incluir información adicional. En los últimos años se han realizado numerosos esfuerzos en integración de datos de diferente procedencia, como por ejemplo la integración de información genómica de número de copias de DNA con información transcriptómica de expresión génica (mRNAs). La integración de esta información heterogénea puede ayudar en la separación de los genes cuya CNA afecta fuertemente a su nivel de expresión y así poder identificar y discriminar únicamente subconjuntos de genes mejor definidos dentro de regiones alteradas extensas.

2.4.3 **Búsqueda de puntos de ruptura frecuentes en CRC**

Las localizaciones genómicas en las que se producen cambios en el número de copias de DNA (puntos de ruptura) de manera recurrente en el conjunto de muestras de una patología son *a priori* regiones importantes especialmente afectadas por las alteraciones de CN. De este modo, la identificación de regiones de cambio o ruptura frecuente en las muestras de CRC puede permitir la identificación de genes clave en la progresión de esta enfermedad.

Para identificar dichas regiones con puntos de ruptura frecuente sobre el conjunto de tumores de CRC se ha aplicado el algoritmo diseñado y explicado anteriormente en la sección 2.3.2. Las regiones identificadas se muestran en la tabla 2.4.3. La tabla contiene el número de SNPs que comprende cada región identificada junto con el número total de puntos de ruptura y, entre paréntesis, el número de muestras diferentes que presentan dichos puntos de ruptura en esa región. Se muestra también el número total de SNPs en el cromosoma así como el número total de puntos de ruptura de las muestras en ese cromosoma. Finalmente, utilizando una distribución hipergeométrica se calcula p-valor de enriquecimiento significativo un puntos de ruptura para cada una de estas regiones respecto a cada cromosoma total. En la figura 2.10 se representan las 4 regiones con puntos de ruptura recurrentes más significativas.

Chromosome band	Chromosomal position(Kb)	SNPs in region	Breakpoints (samples)	Breakpoints in chromosome	SNPs in chromosome	p-value
1p12	116620 - 123939	716	12 (9)	254	40083	1.51E-03
4p16.1	7984 - 10259	513	13 (7)	155	32252	1.11E-06
7p22.3	141 - 837	49	11 (6)	201	25734	1.10E-13
8p11.23	35498 - 38405	386	10 (8)	173	27421	1.44E-04
10p15.3	102 - 695	103	7 (6)	178	28431	3.71E-06
17p11.2	20156 - 22976	182	18 (15)	133	11233	3.64E-12
20p12.1	13875 - 16089	550	16 (9)	145	12378	4.71E-04

Tabla 2.4: Regiones con puntos de ruptura recurrentes detectados en muestras de CRC

De las regiones identificadas en CRC, aquella que presenta un mayor número de puntos de ruptura con una mayor significación estadística es la región que comprende el cromosoma 17p11.2: 18 puntos de ruptura en 15 muestras diferentes que incluyen al menos un punto de ruptura. Este punto de ruptura ha sido estudiado en profundidad en (González-González et al., 2012) donde se valida con otro conjunto más amplio de muestras de CRC utilizando FISH y con otro conjunto independiente de 119 muestras de CRC metastático (Poulogiannis et al., 2010) obtenidas de la base de datos pública GEO (GSE12520) (http://www.ncbi.nlm.nih.gov/geo). En este trabajo se demuestra el valor pronóstico del punto de ruptura detectado y la sonda diseñada para su detección. La figura 2.11 tomada de (González-González et al., 2012) muestra el diseño de esta sonda utilizada para detectar la ruptura recurrente observada en el cromosoma 17 (región 17p11.2).

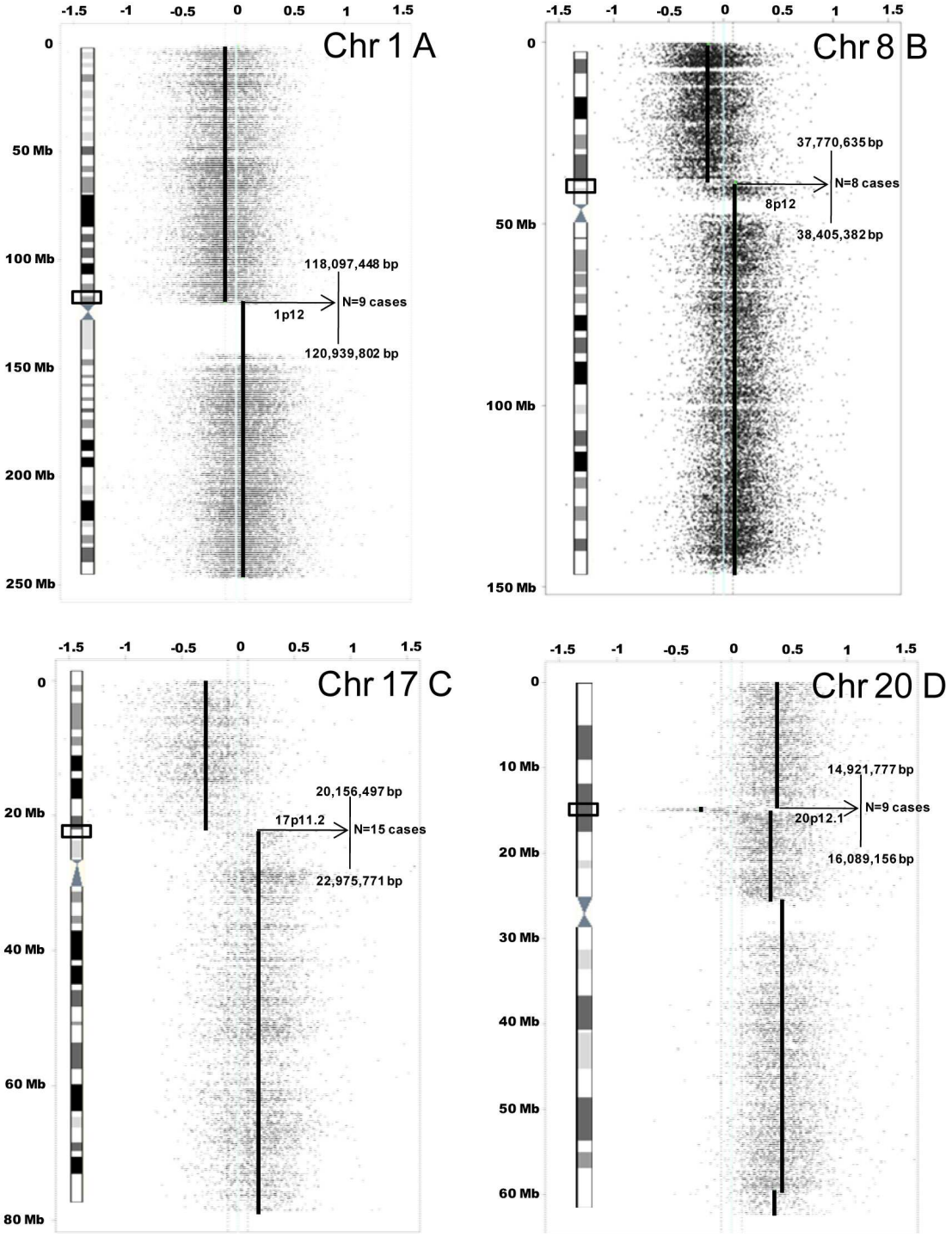


Figura 2.10: Regiones con puntos de ruptura recurrentes en CRC - RBR más significativas en CRC: (A) región 1p12, (B) 8p12, (C) 17p11.2 y (D) 20p12.1.

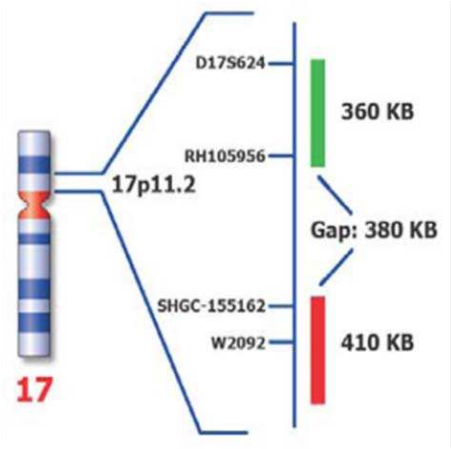


Figura 2.11: Punto de ruptura en el cromosoma 17p11.2 - Representación esquemática de la sonda diseñada para la detección del punto de ruptura del cromosoma 17p11.2 mediante FISH.

2.5 Discusión

El desarrollo de plataformas genómicas experimentales que permiten la cuantificación del número de copias de DNA a lo largo de genomas completos ha hecho posible el análisis sistemático de las alteraciones en número de copias (CNA) en diferentes tipos de cáncer (Beroukhim et al., 2010). De hecho, una característica común a un gran número de oncopatologías malignas y procesos tumorales es la alta inestabilidad genómica que se observa comparada con células normales, que provoca desde alteraciones de brazos completos de cromosomas a alteraciones focales que comprenden unos pocos genes.

En este capítulo se ha diseñado y desarrollado un flujo de trabajo completo para el análisis de las alteraciones del número de copias de DNA, es decir, para análisis de CNA. En este flujo de trabajo se han utilizado metodologías y algoritmos ampliamente validados y aceptados como CRMAv2 (Bengtsson et al., 2009) y CBS (Venkatraman and Olshen, 2007), asi como nuevos algoritmos diseñados *ad hoc*. En concreto se han explorado diferentes técnicas para la discretización de los datos en estados de ganancia (*gain*), no alteración (*NA*) y pérdida (*loss*) y se ha implementado un algoritmo basado en el agrupamiento con *kmeans* y la eliminación de los *outliers* como elementos atípicos que distorsionan la discretización.

En la detección de regiones recurrentes de alteración se han desarrollado dos tipos de algoritmos. Por un lado, se ha desarrollado una estrategia para la detección de regiones mínimas comunes (MCR), definidas como las regiones cromosómicas más pequeñas que contienen una determinada alteración (ganancia o pérdida) en un porcentaje significativo de las muestras de una enfermedad o estado. Esta estrategia está basada en métodos implementados en diferentes trabajos, especialmente (Aguirre et al., 2004). Por otro lado, se ha desarrollado también un algoritmo para la detección de regiones con puntos de ruptura recurrentes (RBR), es decir, regiones cromosómicas en las que es más probable que se produzcan fallos en la replicación, con rupturas y recombinaciones, asociadas a las cuales se originen alteraciones en el número de copias de DNA. Estas rupturas recurrentes suelen estar asociadas con enfermedades concretas y, en cáncer, con subtipos patológicos definidos.

Los algoritmos diseñados, integrados en el marco de trabajo descrito, han sido aplicados con éxito a un conjunto de muestras humanas de cáncer colorectal (CRC) metastásico. En este conjunto de datos se han identificado regiones con un aumento del número de copias de DNA en más del 60 % de las muestras en los cromosomas 7, 8q, 13q y 20q (regiones ganadas significativamente), así como regiones en las que se han perdido una o varias copias de DNA con una recurrencia de más del 65 % en los cromosomas 1p, 8p, 17p y 18 (regiones perdidas significativamente). La mayoría

de estas regiones han sido previamente detectadas en otros estudios similares de CRC metastásico (De Angelis et al., 1999), (Diep et al., 2003), (Höglund et al., 2002), (Diep et al., 2006), confirmando la validez de los análisis. Además, se han detectado nuevas regiones con alteraciones recurrentes que contienen genes asociados con cáncer, muchos de los cuales han sido descritos con relación a la patogénesis del CRC y con procesos metastásicos. Estas regiones están comentadas de forma más detallada en la discusión de trabajo (Sayagués et al., 2010) titulado *Mapping of genetic abnormalities of primary tumours from metastatic CRC by high-resolution SNP arrays* y que se adjunta en esta memoria.

El algoritmo de detección de MCR ha sido aplicado también con éxito al conjunto de muestras de metástasis en hígado, identificando regiones comunes alteradas de la misma manera en los tumores primarios y sus respectivas metástasis. Estas regiones comunes incluyen ganancias de los cromosomas 7, 8q, 13q y 20 y pérdidas en 1p, 8p, 17p, 18 y 22q que comprenden prácticamente todas las regiones identificadas alteradas en los tumores primarios. Sin embargo, también se han encontrado diferencias entre ambos conjuntos de datos que obedecen por un lado a un aumento de la frecuencia de alteración respecto a la frecuencia detectada en los tumores primarios y por otro lado a la adquisición de nuevas alteraciones como deleciones en los cromosomas 4 y 10q y amplificaciones en los cromosomas 5p y 6p. Estas nuevas alteraciones incluyen 11 genes asociados previamente con el proceso metastásico de CRC (Muñoz Bellvis et al., 2012) y podrían estar asociadas bien con el proceso metastásico en sí o con la adaptación de las células metastásicas al microentorno en el hígado.

Finalmente, el algoritmo diseñado para la detección de puntos de ruptura ha posibilitado la determinación de las regiones cromosómicas más inestables asociadas al CRC. Entre las regiones detectadas destaca la región del cromosoma 17 que comprende la banda 17p11.2. Este punto de ruptura ha sido validado con un conjunto de muestras de CRC más amplio e independiente en (González-González et al., 2012). Esta región detectada está caracterizada por una arquitectura compleja con repeticiones en bajo número de copias denominadas LCRs (*Low Copy Repeats*) que, aparentemente, influyen en la inestabilidad genómica y facilitan los reordenamientos genómicos (Sharp et al., 2006; Carvalho and Lupski, 2008). Esta inestabilidad provoca que muchas veces, durante la división del centrómero en la mitosis y meiosis se produzcan alteraciones y cruzamientos, llevando a que la división se realice en el plano transversal en lugar de en el vertical. Esto provoca que uno de los brazos del cromosoma original se pierda y el otro se duplique dando lugar a un isocromosoma con dos brazos genéticamente idénticos entre sí pero en sentido inverso (Barbouti et al., 2004). El isocromosoma del brazo largo del cromosoma 17 ha sido asociado con diferentes tipos de tumores como el meduloblastoma, donde es la alteración más frecuente (50 %, (Biegel, 1997)), linfoma no-Hodgkin (NHL), leucemia mieloide aguda (AML), leucemia linfoide crónica (CLL) y síndromes mielodisplásicos (MDS) (Babicka et al., 2007; Scheurlen et al., 1999). La amplificación del cromosoma 17q (el punto de ruptura 17p11.2 conlleva normalmente una pérdida del brazo p del cromosoma y una ganancia del brazo q del mismo) es un factor de mal pronóstico en la gran mayoría de los casos reportados. Sin embargo, hasta ahora la presencia de este isocromosoma no había sido asociada con la supervivencia en cáncer colorectal metastásico.

Capítulo

3

Análisis combinado de perfiles de expresión génica y de número de copias de DNA

3.1 Introducción: Integración de datos *ómicos*

Las tecnologías genómicas y proteómicas de gran escala (como los microarrays de alta densidad) han posibilitado la cuantificación de características celulares globales a diferentes niveles biomoleculares, como pueden ser los niveles de expresión génica (Schena et al., 1995), el número de copias de DNA (Pollack et al., 1999), la expresión de proteínas (Haab et al., 2001), la metilación del DNA (Yan et al., 2000), etc. Cada uno de estos tipos de datos proporciona una visión de los procesos celulares desde un punto de vista diferente aunque complementario. La integración de varias de estas capas de datos proporciona una visión global más completa y detalla que la que puedan ofrecer cada unas de las capas por separado.

El presente capítulo está enfocado a la integración de dos de estas capas de datos ”ómicos”: el número de copias de DNA y los perfiles de expresión génica. Esta integración tiene como objetivo discernir las alteraciones germinales que influyen en la aparición y progresión de numerosas patologías malignas, es decir, que están frecuentemente asociadas a las transformaciones que suceden en los procesos tumorales.

Como la introducción a los datos de expresión y de número de copias de DNA, así como a las técnicas de cuantificación y medida de los mismos han sido ya presentadas en los capítulos 1 y 2 respectivamente; en este capítulo se presentan directamente las ventajas derivadas de dicha integración junto con los problemas que han motivado el desarrollo del algoritmo implementado.

A continuación se detalla la motivación de este trabajo y se describen cada uno de los pasos llevados a cabo para la integración, así como los métodos empleados para el análisis conjunto de los datos. Para finalizar se muestran los resultados obtenidos con el algoritmo diseñado aplicado a un conjunto de datos de Glioblastoma Multiforme.

3.2 Motivación: Número de copias de DNA (CN) y expresión génica (GE)

El análisis de datos de expresión ha sido utilizado con éxito en la identificación de rutas moleculares (*pathways*) y en la identificación de subgrupos o subtipos diferentes de cáncer (Tusher et al., 2001; Subramanian et al., 2005). Sin embargo, en la caracterización de genes conductores el consenso entre diferentes firmas moleculares publicadas en estudios independientes es muy escaso (Ein-Dor et al., 2005; Fan et al., 2006; Sims, 2009). Por el contrario, los análisis publicados sobre las alteraciones en número de copias del DNA (CNAs) para una misma enfermedad suelen ser bastante consistentes, identificando regiones similares aunque varíen en tamaño o estén ligeramente desplazadas (Beroukhim et al., 2010; Bignell et al., 2010). Las alteraciones genómicas pueden ser por tanto predictores más fiables y estables en la localización de genes conductores, ya que normalmente implican la amplificación o delección de oncogenes y genes supresores tumorales con un papel importante en la carcinogénesis.

La principal desventaja del análisis de CNA es que normalmente da como resultado regiones genómicas de gran amplitud en las que pueden llegar a incluirse un gran número de *loci* génicos. Aparecen así dificultades para separar aquellos genes de la región cuya desregulación causa y promueve el desarrollo del tumor, que denominaremos genes conductores (*driver genes*), de aquellos otros genes que acompañan al estado transformado patológico o que incluso simplemente están accidentalmente situados en la región con CNA pero que no tienen un efecto causal sobre la enfermedad, que denominaremos genes pasajeros (*passenger genes*). La clave de los estudios de genómica funcional que buscan biomarcadores para distintas enfermedades o estados biológicos patológicos es encontrar los genes que son verdaderamente causales.

En ciertos casos genes situados en regiones con modificación en el número de copias de DNA no presentan cambios en sus niveles de expresión, o viceversa, genes que presentan alteraciones en sus niveles de expresión no modifican su número de copias de DNA (*gene dosage*). Akavia y colaboradores postulan que los genes conductores deberían presentar ambas características: estar amplificados y sobre-expresados o delecionados e infra-expresados (Akavia et al., 2010). Otros autores demostraron que los genes que están consistentemente sobre-expresados en regiones genómicas amplificadas son necesarios para el desarrollo tumoral de las células, de modo que amplificación y sobre-expresión recurrentes marcan genes que pueden ser utilizados como potenciales dianas terapéuticas (Bernard-Pierrot et al., 2008; Natrajan et al., 2009). De este modo, una estrategia eficaz para disminuir el número de genes pasajeros y ayudar a diferenciarlos de los genes conductores consiste en la integración de información del número de copias de DNA junto con datos de expresión génica.

En los últimos años han surgido numerosas aproximaciones a este tipo de integración (Pollack et al., 2002; Kotliarov et al., 2009; Akavia et al., 2010; Turner et al., 2010; Kim et al., 2011). Estas aproximaciones han demostrado la utilidad de dicha integración a la hora de identificar genes conductores basándose en la hipótesis de que los cambios de expresión en los genes asociados a una determinada enfermedad están frecuentemente inducidos por alteraciones genómicas. La idea fundamental que subyace en este tipo de análisis es la identificación de alteraciones recurrentes y genes cuyos perfiles de expresión estén asociados (correlacionados normalmente) con dichas alteraciones. Sin embargo, la mayoría de estos trabajos buscan las relaciones entre los genes de una manera individual; es decir, utilizan los valores de expresión basándose en una estrategia gen a gen sin tener en cuenta las localizaciones cromosómicas de estos genes.

La aproximación que se propone en este capítulo está basada sin embargo en una exploración global de la relación entre el número de copias de DNA y los niveles de expresión génica. Nues-

tra aproximación se apoya en la hipótesis de que debería existir un comportamiento común a los genes que están bajo la influencia de CNAs (Ortiz-Estevez et al., 2011). Para analizar este comportamiento común se propone la aplicación de un algoritmo de segmentación a los dos tipos de datos (i.e. número de copias CN y expresión génica GE) de manera independiente. Del mismo modo que los algoritmos de segmentación identifican conjuntos de SNPs contiguos en el genoma con el mismo número de copias de DNA; la aplicación del algoritmo de segmentación a los valores de expresión génica permitirá identificar conjuntos de genes con expresión similar en regiones genómicas y reducir posibles efectos de regulación génica que no estén relacionados con la localización cromosómica.

En (Ortiz-Estevez et al., 2011) se analiza la correlación entre los dos tipos de datos segmentados (CN y GE) y se demuestra que dicha segmentación mejora las correlaciones encontradas utilizando los genes de modo independiente. Sin embargo, no se propone una metodología para el análisis y la discriminación de las regiones candidatas y de los genes conductores identificados en las mismas.

3.3 Desarrollo metodológico: Integración de datos de expresión génica y de datos de número de copias de DNA

El presente capítulo describe el desarrollo de un método para la integración de datos procedentes de microarrays de expresión y de microarrays de DNA basado en las localizaciones cromosómicas de ambos tipos de datos publicado en (Fontanillo et al., 2012). Para una mayor facilidad de lectura abreviaremos los datos de expresión génica como GE y los datos de número de copias de DNA como CN.

El flujo de análisis propuesto se encuentra representado de manera esquemática en la figura 3.1 que incluye el preprocesamiento propuesto en (Ortiz-Estevez et al., 2011). En el flujo de trabajo se muestra que ambos tipos de datos se preprocesan de manera independiente y se aplica un algoritmo de segmentación de manera también independiente a CN y GE. Una vez obtenidos los valores de número de copia de DNA segmentados y los valores de expresión segmentados se realiza la integración de ambos tipos de datos. El análisis de la asociación entre ambos permitirá seleccionar las regiones candidatas junto con los genes alterados en cada una de ellas.

3.3.1 Normalización y sumarización

En los capítulos previos se ha comentado la importancia del preprocesamiento en el análisis de datos de GE procedentes de microarrays (1) y datos de CN procedentes de microarrays de SNPs (2). En el flujo de análisis conjunto de ambos tipos de datos el preprocesamiento se realiza de forma independiente para ambos.

Por un lado, los datos de GE son preprocesados con RMA (Irizarry et al., 2003a) y, por otro, los datos de CN son preprocesados con CRMA (Bengtsson et al., 2009). Para ambos algoritmos es necesario un fichero de descripción del array (*Chip Definition File*, CDF) que contiene información sobre el mapeo de las sondas del array. Las sondas de los arrays de expresión serán agrupadas para cada *locus* génico utilizando el mapeo proporcionado por *GATExplorer* (http://bioinfow.dep.usal.es/xgate) (Risueno et al., 2010) y las sondas de los arrays de SNPs serán agrupadas para cada uno de los SNPs mapeados segun las asignaciones proporcionadas por *Affymetrix* usando la plataforma *aroma.affymetrix* (http://www.aroma-project.org).

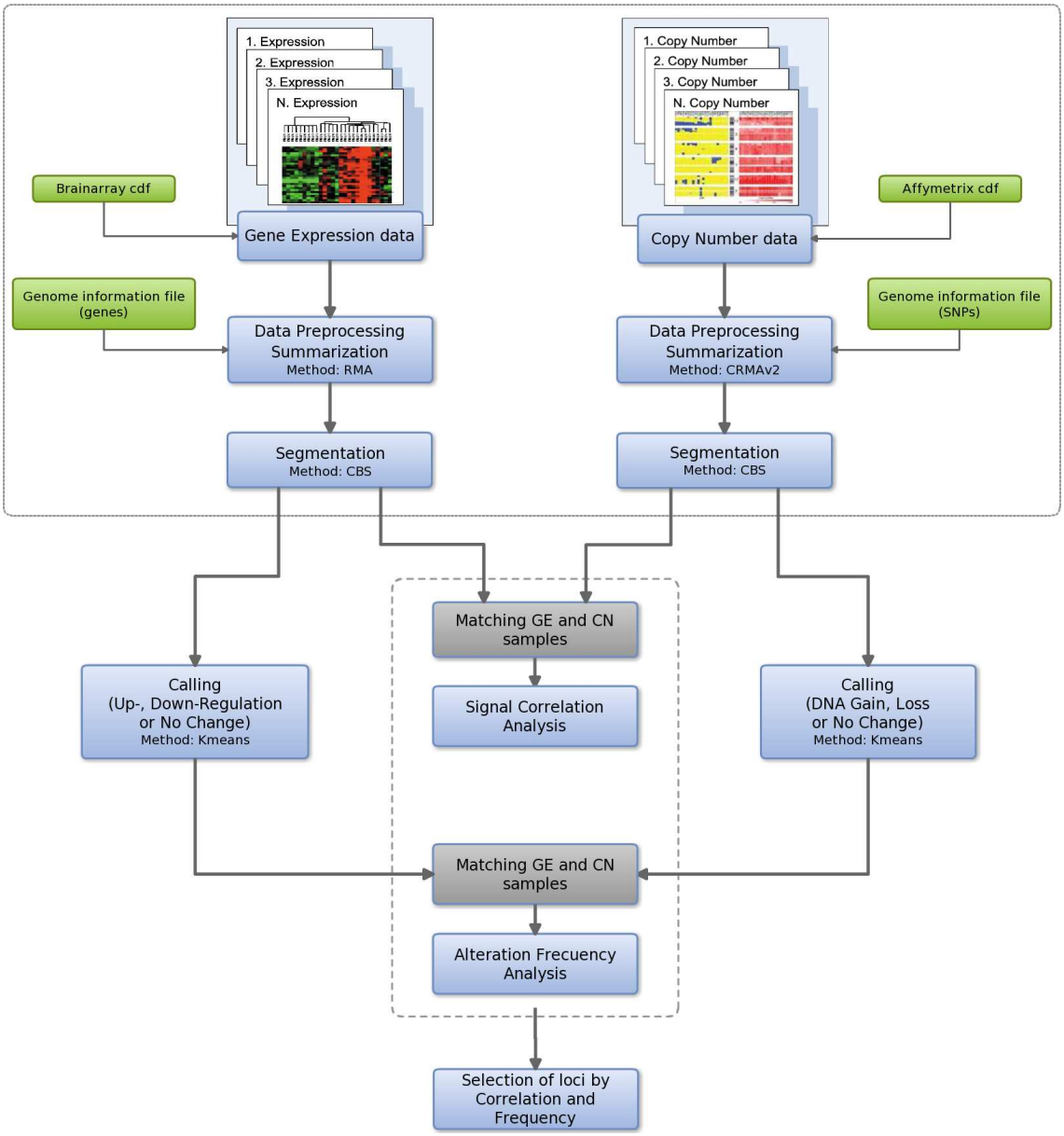


Figura 3.1: Esquema del flujo del análisis integrado de datos de expresión y número de copias de DNA - Los pasos iniciales de preprocesamiento y segmentación incluidos en el recuadro gris han sido adaptados de (Ortiz-Estevez et al., 2011). El resto de pasos incluyen el análisis de la correlación entre la señal segmentada de expresión y de número de copias y las frecuencias de alteración conjunta.

3.3.2 **Segmentación**

Llamamos segmentación al proceso de dividir un conjunto de datos ordenado en regiones de elementos adyacentes que tienen valores similares. A cada una de estas regiones se le asigna un valor que representa a todos los datos que pertenecen a la misma.

Entre los diferentes algoritmos de segmentación comentados en 2.2.2, se ha utilizado CBS (*Circular Binary Segmentation*) (Venkatraman and Olshen, 2007). Dicho algoritmo ha sido aplicado a los datos de GE y de CN de manera independiente.

Al igual que en el preprocesamiento de los datos de CN descrito en 2.2.1, el número de copias de cada región viene determinado por la comparación con una muestra sana de referencia (*log2ratio* de señal de la muestra tumoral alterada *versus* la señal de la muestra de referencia normal), en los datos de GE se han utilizado también los valores de expresión relativos a muestras sanas pareadas (*log2ratio* de los niveles de expresión de los genes en la muestra tumoral *versus* los niveles de expresión en la muestra de referencia normal). Esto permite discernir aquellos genes en los que se produce un cambio en los niveles de expresión en el tumor, de los genes cuyos valores de expresión son altos en cualquier célula independientemente de si es tumoral.

Para datos de CN los algoritmos de segmentación proporcionan una buena estimación del número real de copias de DNA, ya que se centran en la búsqueda de los puntos de ruptura donde se producen estos cambios. En contraposición, la segmentación de GE no es el mejor estimador para el nivel de expresión propio de un gen, sino una medida indirecta de los efectos sobre la expresión de mecanismos de regulación relacionados con la localización genómica. Al suavizar los valores de expresión de cada gen teniendo en cuenta la expresión de los genes adyacentes se están considerando sobre todo los efectos asociados a la localización de los genes en el genoma, como pueden ser la alteración del número de copias de la región u otros mecanismos epigenéticos como la metilación del DNA o la descondensación de la cromatina. De esta manera, se suavizan o enmascaran los mecanismos de control transcripcional específicos de un gen no asociados a su *locus* genómico.

La segmentación de datos de expresión plantea un problema añadido que no está presente en los datos que provienen de arrays de SNPs. Un SNP es un único punto con una localización concreta en el genoma, sin embargo un gen para el que se tiene un único valor de expresión puede ocupar desde unos pocos cientos hasta decenas de miles de pares de bases. Para los algoritmos de segmentación cada uno de estos genes tiene que ser considerado de manera puntual, por lo que para la segmentación se ha utilizado como localización concreta el punto medio del *locus* génico independientemente del número de pares de bases en las que se extienda el gen.

3.3.3 **Emparejamiento de los datos de expresión y número de copias de DNA**

El proceso para emparejar los datos de GE y CN no es trivial puesto que las sondas e identificadores utilizados en ambos tipos de datos no son los mismos ni tienen la misma distribución a lo largo del genoma.

La aproximación mayoritaria es la tomada en (Turner et al., 2010) que asigna a cada gen presente en el microarray de expresión el valor de CN correspondiente a un valor de centralidad, media o mediana, de los SNPs localizados en el locus de dicho gen. Sin embargo, ocurre que, dependiendo del tipo de array de SNPs utilizado, únicamente un porcentaje de los genes asociados contienen algún SNP.

En (Kotliarov et al., 2009) extienden esta definición y toman para cada *probeset* definido en el microarray de expresión el valor de la media de los SNPs localizados en una ventana de 1 Mbp alrededor del centro de dicho *probeset* logrando así una mayor cobertura. Esta aproximación presenta el inconveniente de que los SNPs tomados para cada uno de los *probesets* pueden solaparse. Además sigue siendo necesaria también la realización de una equivalencia posterior entre *probeset* y gen.

La segmentación tanto de los datos de expresión como de número de copias de DNA hace posible una nueva aproximación. El algoritmo implementado aprovecha la ventaja de los datos segmentados en los que ya no se dispone de información puntual de cada sonda en posiciones concretas, sino que las regiones o segmentos abarcan todo el genoma completo. Cada una de los segmentos calculados para cada tipo de datos está definido por la media del conjunto de SNPs o del conjunto de genes incluidos en cada uno de ellos. De esta manera, es posible asignar a cada gen un valor de expresión y de número de copias de DNA correspondiente al valor del segmento en el que se encuentra. Si en un *locus* hay uno o más puntos de ruptura que dan lugar a varios segmentos con valores diferentes para un mismo gen, se asignará a dicho gen el valor del segmento que incluya una mayor proporción del *locus* génico.

Utilizando los datos segmentados se construyen dos matrices con los valores de CN segmentados (sCN) o con los valores de GE segmentados (sGE), respectivamente. De esta manera sCN_{ij} se corresponde con el valor del número de copias para el gen i en la muestra j y sGE_{ij} con el valor de expresión del gen i en la muestra j.

3.3.4 Correlación entre niveles de expresión y número de copias de DNA

En la integración de los datos de expresión y de número de copias de DNA la determinación de la influencia de las CNAs sobre los niveles de expresión de los genes incluidos en las regiones alteradas adquiere un papel importante. Como se ha indicado al principio de este capítulo, varios estudios han demostrado que las CNAs pueden estar relacionadas con modificaciones similares en los niveles de expresión de algunos genes específicos (Pollack et al., 2002; Kotliarov et al., 2006; Bungaro et al., 2009).

En el análisis integrado desarrollado, la cuantificación de la influencia o asociación entre CN y GE se realiza mediante el cálculo de los coeficientes de correlación de *Pearson* entre los valores segmentados obtenidos: sCN y sGE. Así, para cada gen i el coeficiente de correlación se define como:

$$r_i = \frac{\sum_{j=1}^{n} (sCN_{ij} \cdot sGE_{ij}) - n \cdot \overline{sCN_i} \cdot \overline{sGE_i}}{n \cdot \sigma_{sCN_i} \cdot \sigma_{sGE_i}} \tag{3.1}$$

La utilización únicamente del coeficiente de correlación basado en datos segmentados permite la detección de regiones donde existe una fuerte influencia de las CNAs sobre los valores de expresión. Sin embargo, la segmentación sobre GE reduce la sensibilidad y no permite identificar fácilmente aquellas CNAs que afecta a un número muy reducido de genes o a genes cuyos *loci* están muy distantes. En el cálculo de sCN las sondas de los arrays de SNPs están homogéneamente distribuidas en el genoma, con la sola excepción de algunas regiones bien localizadas (como los centrómeros donde la densidad de sondas es mucho menor que en el resto). En el cálculo de sGE esto no sucede así, existen regiones genómicas en las que la densidad de genes es muy baja y la variabilidad de su expresión muy alta. Es por ello que los segmentos obtenidos para los datos de expresión son más extensos y consecuentemente, existe una mayor dificultad en la detección

de puntos de ruptura. Este hecho reduce la sensibilidad en la detección de regiones muy pequeñas afectadas por las CNAs y plantea un problema en la búsqueda de correlaciones para aquellos genes en regiones de densidad génica baja y con mucha variabilidad en la expresión de genes cercanos.

Para aunmentar la sensibilidad y evitar la pérdida de estas regiones en el análisis de las correlaciones, el algoritmo realiza una búsqueda de genes cuyos valores de expresión difieren significativamente de la media del segmento en el que se encuentran. Se seleccionan aquellos genes que son considerados *outliers* en su segmento en un porcentaje de las muestras (por defecto un tercio de las mismas). Una vez identificados estos genes se recalculan las correlaciones entre los valores de sCN y sus valores de expresión sin segmentar. Se recuperan así ciertos genes aislados pero significativos que se comportan de modo independiente a sus genes más cercanos, que de otro modo habrían quedado ocultos en la segmentación.

3.3.5 **Alteraciones consistentes y recurrentes en los niveles de CN y GE**

Según la hipótesis de la presión selectiva las alteraciones que confieren a la célula tumoral una ventaja sobre el resto son mantenidas, con lo que las alteraciones comunes observadas en un elevado número de muestras serán probablemente más importantes en el desarrollo de la enfermedad que aquellas que se producen en un número muy pequeño de muestras, aunque sean de una intensidad mayor.

A la hora de identificar regiones clave buscaremos por tanto aquellas que se producen de manera recurrente y consistente. Esta alteración recurrente"significa alterada en una alta proporción de los individuos, y çonsistente"significa alterada de la misma forma en ambos tipos de datos (es decir, una ganancia en CN que conlleve una sobre-expresión de los genes en la región o una pérdida en CN que conlleve una infra-expresión de los genes en la región). De esta manera se discriminará aquellos genes pasajeros que o bien no ven modificada su expresión o bien esta expresión está regulada por otros factores independientes del número de copias de DNA.

Para acometer la búsqueda de estas alteraciones consistentes y recurrentes se realiza un análisis basado en la estratificación en categorías de los segmentos en ambos tipos de datos. De este modo, las regiones genómicas se clasifican en diferentes categorías dependiendo de los valores de CN y de GE.

En los datos de CN las categorías en las que se discretizan los segmentos también son 3:

1. Ganancia en el número de copias de DNA (3 copias o más) (*Gained*, G)
2. Pérdida en el número de copias de DNA (1 copia o menos) (*Lost*, L)
3. Sin cambios en el número de copias de DNA (2 copias) (*No Changed*, N)

Los segmentos de GE son discretizados en 3 categorías:

1. Regulados sobre-expresados (*Up-regulated*, U)
2. Regulados infra-expresados o reprimidos (*Down-regulated*, D)
3. Sin cambios en la expresión (*No Changed*, N)

La combinación de los dos tipos de estratificaciones permite la discretización de las regiones genómicas en 9 categorías representadas en la tabla de contingencia presentada en la figura 3.2.

La discretización de los datos en estos 9 estados se realiza a partir de los valores de sGE y de sCN para cada gen en cada una de las muestras obtenidos del emparejamiento de los dos tipos de datos explicado en la sección 3.3.3. La determinación de los estados o categorías establecidos es independiente para cada tipo de datos y se realiza de forma análoga a la descrita en el capítulo 2

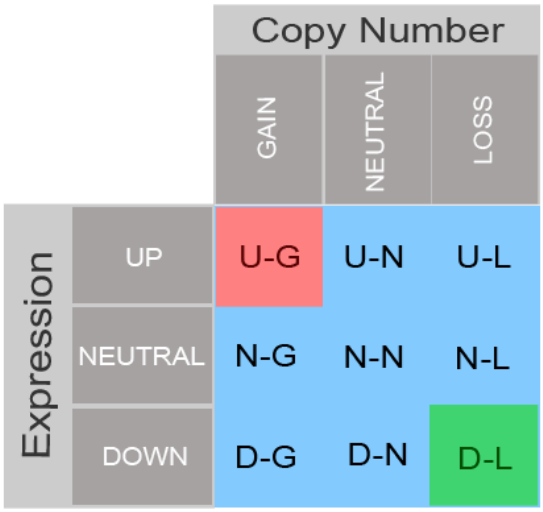


Figura 3.2: Posibles estados para cada región basados en la categorización de los segmentos de CN y de GE - De acuerdo a las categorías establecidas para los datos de número de copias de DNA (Ganancia, Pérdida y No Cambio) y de expresión (Sobre-expresión, Infra-expresión y No Cambio) las regiones pueden clasificarse en 9 estados diferentes mostrados en la tabla de contingencia.

en la sección 2.2.3. Esto es, tomando las distribuciones de los valores de expresión y de número de copias para el conjunto de genes mapeados en ambos tipos de datos se identifican los valores atípicos u *outliers* de estas distribuciones, que se eliminarán para el cálculo de los umbrales. A continuación se utiliza un algoritmo de agrupamiento (*K-means*) en 3 clases o clústers para cada tipo de datos de manera independiente. Se establecen como umbrales de ganancia y pérdida y sobre-expresión e infra-expresión los límites o fronteras entre las clases determinadas por el algoritmo de agrupamiento.

Utilizando estos umbrales para cada muestra se clasifican los genes en las 9 categorías descritas. De estas 9 categorías resultan interesantes aquellas que implican la alteración consistente en ambos tipos de datos, es decir, cuando la sobre-expresión (*up* U) está asociada con un aumento del número de copias de DNA (*gain* G) y cuando la infra-expresión génica (*down* D) está asociada con la pérdida (*loss* L) de alguna de las copias de DNA. Estas categorías (U-G y D-L) están marcadas en rojo y verde respectivamente en la tabla de contingencia de la figura 3.2. El algoritmo buscará aquellos genes que de manera recurrente se encuentran en estas categorías en la mayoría de las muestras. Es decir, se buscarán patrones comunes de modulación de los genes mediante el análisis de todas las muestras.

La determinación de los patrones recurrentes o comunes se realiza mediante el análisis de las distribuciones de frecuencias empíricas para las categorías U-G y D-L mencionadas anteriormente. En estas distribuciones se identifican aquellos genes con una frecuencia de alteración por encima del cuantil 90 (C_{90}).

A partir de los genes identificados se establecen las regiones genómicas recurrentemente alteradas de manera que cada una de estas regiones estará formada por los genes en localizaciones cromosómicas contiguas con una frecuencia de alteración U-G y D-L por encima del umbral establecido. Cuando existan dos de estas regiones recurrentes separadas por 3 o menos genes serán combinadas en una única región con los límites de comienzo y fin marcados por el inicio de la primera región y el fin de la segunda respectivamente.

3.3.6 Identificación de regiones genómicas clave en la alteración

El método desarrollado identifica por un lado regiones candidatas con una alta correlación entre CN y GE y por otro lado las regiones recurrentemente alteradas de forma consistente entre las

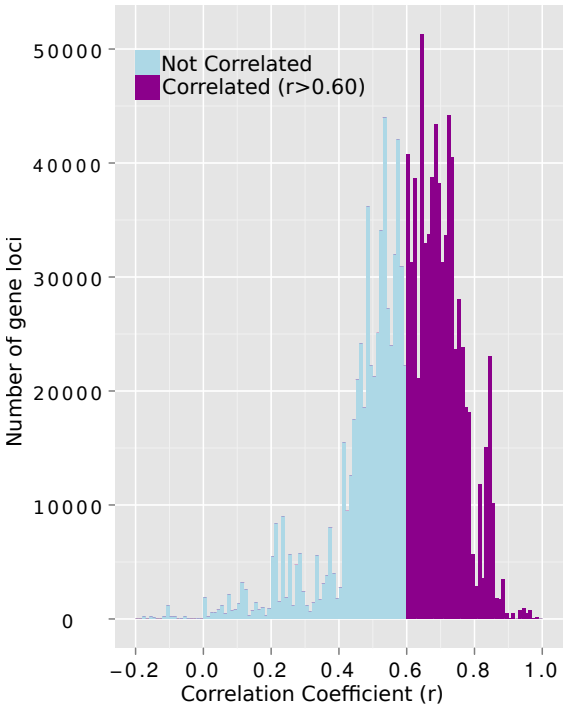


Figura 3.3: Distribución de densidad de los coeficientes de correlación para los datos de GBM - En morado se representa el número de genes totales, considerando todas las muestras, con correlaciones significativas ($r \geq 0.60$) entre los niveles de expresión génica y el número de copias de DNA para los datos de GBM. En azul se representa el número de genes con correlaciones no significativas ($r \leq 0.60$).

muestras testadas para ambos tipos de datos. Una región genómica será definida como clave en el desarrollo y progresión tumoral si y sólo si cumple ambas condiciones. Es decir, la frecuencia de alteración de la región es elevada y además la CNA determina la desregulación de los genes asociados a la misma. Estas regiones clave son regiones candidatas a contener los genes conductores causales asociados con la enfermedad o el tipo de cáncer analizado.

3.4 Aplicación a un conjunto de muestras de Glioblastoma Multiforme (GBM)

El método diseñado ha sido aplicado a un conjunto de datos sobre muestras de Glioblastoma Multiforme (GBM), que es el tumor más común y de peor pronóstico de los tumores del sistema nervioso central con una mediana de supervivencia de aproximadamente 14 meses (Furnari et al., 2007). Los datos para el análisis con el algoritmo desarrollado han sido tomados del estudio (Kotliarov et al., 2006) que consta de 64 muestras de tumores de pacientes de GBM de los que se han hibridado tanto microarrays de DNA como microarrays de expresión. En concreto, para DNA se ha utilizado el modelo *Genechip Human Mapping 100K* de *Affymetrix* y para el análisis de los niveles de expresión génica el *Genechip Human Genome U133 Plus 2.0*. A parte de las muestras tumorales y para tener una referencia con la que poder comparar se han hibridado también arrays de ambos tipos con 21 muestras no tumorales de cariotipo normal, tomadas de resecciones del lóbulo temporal de paciences epilépticos.

3.4.1 Correlación entre CN y GE en muestras de GBM

Aproximadamente el 55 % de los 21281 genes humanos mapeados (versión h37 del *Genome Reference Consortium*, GRC) en ambos tipos de datos están relacionados con un coeficiente de correlación de *Pearson* $r > 0.60$, correspondiente con un p-valor ajustado con el método de *Bonferroni*

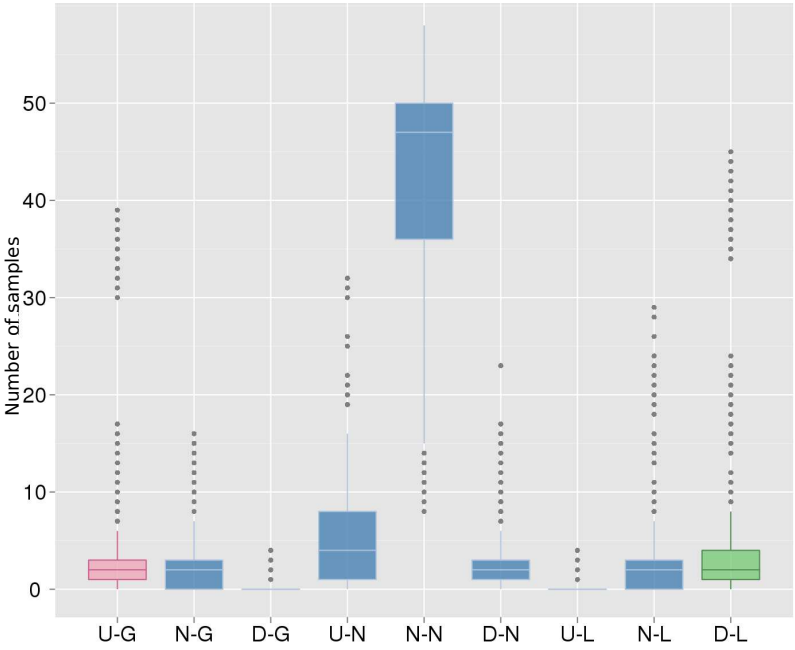


Figura 3.4: *Boxplots* con las frecuencias de alteración conjunta para CN y GE en GBM - Distribuciones de densidad del número de muestras de GBM asignadas a cada una de las 9 categorías (U-G, N-G, D-G, U-N, N-N, D-N, U-L N-L D-L) para todos los *loci* génicos humanos.

$<$0.005. El elevado porcentaje de genes correlacionados muestra la influencia del número de copias de DNA sobre los valores de expresión de los genes. Siendo más astringente en la selección del umbral mínimo de correlación la cobertura se ve reducida al 26 % de los genes para un umbral de correlación r $>$ 0.70 y a un 6 % de los genes para un r $>$ 0.80. La figura 3.3 muestra el número de *loci* génicos para los diferentes valores de correlación, en la que se aprecia que la mayoría presenta correlaciones significativas entre los datos sCN y sGE. Se consideran regiones con valores de correlación significativa aquellas regiones formadas por al menos dos *loci* génicos contiguos en los que el coeficiente de correlación es mayor de 0.60 (r $>$ 0.60) correspondiente con un p-valor $<$ 0.005 (corregido mediante el método de *Bonferroni*).

3.4.2 Frecuencia combinada de alteración de CN y GE en muestras de GBM

Partiendo de los datos emparejados de sCN y sGE se han estratificado los genes en las 9 categorías representadas en la tabla de contingencia 3.2. La figura 3.4 presenta las distribuciones del número de muestras en cada categoría para todos los genes.

Como cabría esperar, la categoría de no cambio (N-N) es la más frecuente. La mayoría de los genes no presentan ninguna alteración ni en el número de copias de DNA ni en sus niveles de expresión, mientras que en categorías como D-G o U-L (correspondientes a tendencias opuestas entre GE y CN) no existen apenas genes (tan sólo 4) y además presentan esta tendencia en un número muy pequeño de muestras (siempre en menos de 5 muestras). Por otro lado, en las categorías de interés, es decir, regiones donde se produce una sobre-expresión colocalizada con una ganancia de CN (U-G) o una infra-expresión colocalizada con una pérdida de CN (D-L), se observa que un número reducido de genes aparece alterado en ambos tipos de datos en un porcentaje elevado de las muestras (más de 30 muestras de las 64 analizadas).

El análisis de las distribuciones de frecuencias empíricas en estas categorías permite establecer un umbral correspondiente al cuantil del 10 % de frecuencias superiores. La determinación de los *loci* más frecuentemente asignados a estas categorías se corresponde con los genes que aparecen alterados en el 20 % y el 17 % de las muestras para las categorías U-G y D-L, respectivamente.

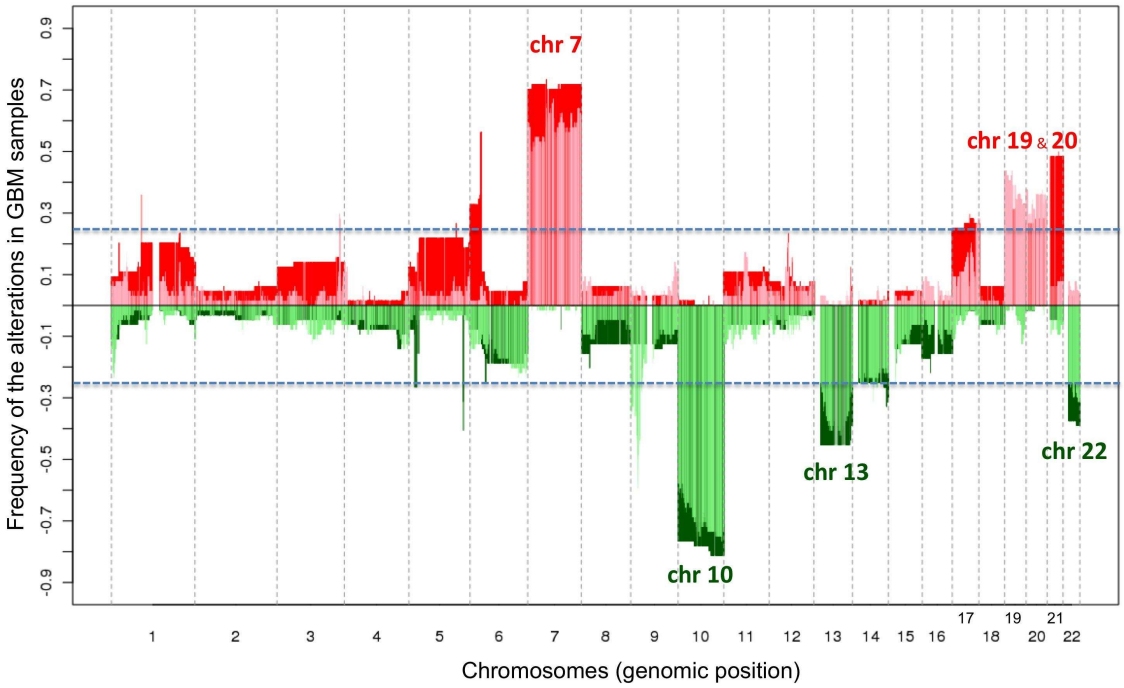


Figura 3.5: Frecuencias de alteración de GE y CN para datos de GBM - Porcentaje de muestras alteradas en cada localización genómica para cada cromosoma ordenados del 1 al 22, uno a continuación de otro. En rojo y rosa el porcentaje de muestras con sobre-expresión y aumento CN respectivamente. El porcentaje de muestras con infra-expresión está marcado en verde y con una disminución del número de copias de DNA en verde claro superpuesto. Las líneas azules marcan una frecuencia de alteración en el 25 % de las muestras.

Se establecen como regiones candidatas aquellas formadas por uno o más *loci* contiguos con frecuencias de alteración por encima de los umbrales establecidos. Estas regiones candidatas se corresponden con regiones en los cromosomas 7, 19 y 20 para U-G y regiones en los cromosomas 10, 13 y 22 para D-L. En la representación esquemática de estas alteraciones en la figura 3.5 se observan como las mayores frecuencias se localizan en dichos cromosomas. En la figura también se puede ver que la mayoría de las CNAs y los cambios de expresión se superponen en el genoma, lo que indica una fuerte asociación entre ambos tipos de datos, reportada también en (Ortiz-Estevez et al., 2011). Existen sin embargo algunas regiones (como el brazo p del cromosoma 6 o el cromosoma 21) en las que se observa que la sobre-expresión de los genes situados en estos *loci* no se corresponde con un aumento de CN. Se puede suponer que esta no correspondencia debe obedecer otros mecanismos de regulación transcripcional distintos en los que no se produce esa asociación entre CN y GE.

3.4.3 **Identificación de genes conductores en regiones candidatas para GBM**

Como se ha descrito anteriormente, el método desarrollado identifica como regiones candidatas aquellas que presentan una alta correlación entre los valores de CN y GE y que, además, están frecuentemente alteradas en el mismo sentido de manera consistente. La figura 3.6 muestra una visión global de estas dos características para cada región en los 22 autosomas humanos. En ella se representan el porcentaje de muestras con alteraciones (el área rosa representa el porcentaje de muestras U-G y el área verde el porcentaje de muestras D-L) y la correlación (r) entre los valores de CN y GE (en azul correlaciones no significativas, $r \leq 0.60$, y en rojo correlaciones significativas, $r > 0.60$) para las muestras de GBM.

Imponiendo esta doble condición se obtienen las regiones alteradas de manera recurrente y consistente que se corresponden con los cromosomas 7, 10, 13q, 14q, 20 y 22q. Las regiones candidatas obtenidas son consistentes con las alteraciones previamente descritas. Por ejemplo, en el trabajo (de Tayrac et al., 2009), donde se reportan ganancias en los cromosomas 7 (73 %) y 20 (16 %) y pérdidas en 10 (58 %), 13q (31 %) y 22q(21 %); y en el trabajo de (Ruano et al., 2006) en el que también se identifican como regiones clave los cromosomas 7, 10, 13 y 20.

Los regiones más significativas identificadas se muestran de manera detallada en las tablas 3.4.3 y 3.4.3, tanto para las regiones que muestran ganancia y sobre-expresión como para las regiones que muestran pérdida e infra-expresión, respectivamente. Estas tablas contienen información acerca del tamaño de la región (posiciones de comienzo y fin en Megabases (Mb)), frecuencia media de alteración de la región (porcentaje de muestras) y valor medio de la correlación (r) entre los valores de sCN y sGE. Las correlaciones calculadas con los valores de expresión sin segmentar para los genes con valores de expresión muy diferentes del resto de su segmento aparecen marcadas con *. Además se indica el número de genes totales que comprende la región y se señalan los genes previamente asociados con algún tipo de cáncer (según el censo de genes de cáncer del Sanger Center: *Cancer Gene Census List*, http://www.sanger.ac.uk/genetics/CGP).

Las alteraciones más importantes y frecuentes en Glioblastoma son la amplificación del cromosoma 7 y la delección del cromosoma 10 (Ohgaki et al., 2004), hecho que corrobora las frecuencias de alteración encontradas en los datos analizados. Los resultados para estos cromosomas obtenidos con el método propuesto han sido ampliados en la figura 3.7. Las alteraciones abarcan prácticamente los cromosomas completos en más del 50 % de las muestras, aunque también se observa que existen pequeñas regiones en las que la influencia de las alteraciones en el número de copias no se corresponde con la desregulación de la expresión génica (son pequeñas zonas con correlaciones no significativas indicadas como puntos azules).

Figura 3.6: Esquema de las regiones candidatas en GBM - Gráfico de los 22 autosomas. El eje X representa las localizaciones genómicas en cada cromosoma. Los puntos azules y rojos diferencian las correlaciones no significativas ($r \leq 0.60$) y significativas ($r > 0.60$). Las áreas rosa y verde señalan el porcentaje de muestras en las categorías U-G y D-L respectivamente.

Chr	Cytobands	Start (Mb)	End (Mb)	Frequency U-G (%)	Correlation (average r)	Number of genes	Cancer genes
7	p22.3-p21.3	14	12407	54	0.75	70	CARD11, PMS2
7	p21.2,p21.1	13981	18582	48	0.83	15	ETV1
7	p21.1	20736	20825	50	0.81	2	
7	p15.3,p15.2	22277	26373	50	0.85	21	HNRNPA2B1
7	p15.2	26682	27830	50	0.68	15	HOXA11, HOXA13, HOXA9
7	p14.3	29901	33407	50	0.78	25	
7	p14.3,p14.2	34693	36658	48	0.72	8	
7	p14.1-p12.1	37857	50759	50	0.72	63	IKZF1
7	p11.2	54820	54827	59	0.77 *	1	SEC61G
7	p11.2	55087	55275	61	0.77 *	1	EGFR
7	p11.2	55572	56044	59	0.70	5	(VOPP1)
7	p11.2-q11.21	57270	66582	57	0.66	18	SBDS
7	q11.22-q21.2	69661	91852	58	0.68	104	ELN, HIP1, AKAP9
7	q21.2,q21.3	92738	97976	57	0.72	32	
7	q22.1-q34	98456	141707	56	0.71	278	MET, SMO, BRAF, CREB3L2, PIK3CG
7	q34-q36.3	141955	158879	56	0.75	121	EZH2, MLL3, PIP
20	p13-q13.33	73	62897	26	0.83	464	

Tabla 3.1: Regiones candidatas sobre-expresadas y ganadas (U-G).

Chr	Cytobands	Start (Mb)	End (Mb)	Frequency D-L (%)	Correlation (average r)	Number of genes	Genes
10	p13-p12.2	16746	24410	60	0.64	28	MLLT10
10	p12.1-p11.22	25190	47922	60	0.63	79	KIF5B, RET
10	p11.2	38239	38265	61	0.60 *	1	ZNF25
10	q11.22,q11.23	49204	53405	63	0.68	42	NCOA4
10	q21.1-q23.1	59989	82352	64	0.65	130	CCDC6, FAM22B, MYST4, PRF1
10	q23.1,	84191	87743	66	0.66	7	
10	q23.31-q23.33	91498	97024	67	0.62	36	LGI1
10	q24.1	97391	97763	66	0.62	4	
10	q24.1-q26.3	98081	134474	69	0.67	196	FGFR2, NFKB2, SUFU, TCF7L2, TLX1, MXI1, VTI1A
13	q12.13,q12.2	27693	28017	28	0.60	5	
13	q12.2,q12.3	28367	30382	30	0.61	12	CDX2, FLT3
13	q13.3-q21.2	35882	61059	35	0.73	103	FOXO1, LCP1, LHFP, RB1
13	q22.2-q31.1	76452	80913	37	0.63	13	
13	q32.1-q33.2	95813	106131	33	0.64	34	
13	q33.3	108171	108931	28	0.95	4	
13	q34	110423	111549	23	0.89	8	
14	q11.2-q24.2	19686	70583	18	0.75	315	BCL2L2, CCNB1IP1, NIN, NKX2-1
14	q24.2,q24.3	73223	76275	17	0.69	44	
14	q24.3,q31.1	77826	80673	17	0.75	9	
22	q11.1-q13.33	16158	51225	25	0.73	398	BCR, CHECK2, CLTCL1, EWSR1, MN1, MYH9, NF2, PDGFB, SMARCB1

Tabla 3.2: Regiones candidatas infra-expresadas y perdidas (D-L).

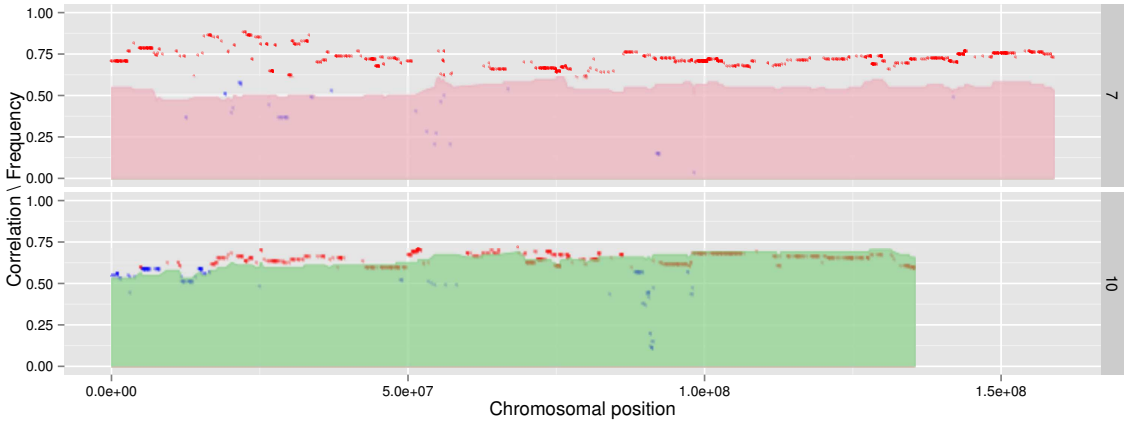


Figura 3.7: Esquema de las alteraciones de los cromosomas 7 y 10 en GBM - Vista detallada de los cromosomas 7 y 10. El eje X representa las localizaciones genómicas en cada cromosoma. Los puntos azul y rojo representan las correlaciones no significativas ($r \leq 0.60$) y significativas ($r > 0.60$), respectivamente. Las áreas rosa y verde señalan el porcentaje de muestras en las categorías U-G y D-L.

La alteración de estos dos cromosomas está fundamentalmente asociada con la sobre-expresión del gen EGFR y la infra-expresión del gen PTEN, respectivamente. El gen EGFR se encuentra sobre-expresado y su *locus* con un incremento en el número de copias de DNA en más del 60 % de las muestras analizadas y constituye la alteración más frecuente en GBM. Esta alteración de EGFR se asocia con genes en *loci* cercanos que regulan su función. Así por ejemplo, VOPP1 (conocido como ECOP: (*EGFR coamplified and overexpresed protein*) aparece también alterado muy significativamente y está en la región adyacente a EGFR (ver tabla 3.4.3), aunque no es un gen anotado en el censo de genes de cáncer (por ello se incluye entre paréntesis). La alteración de PTEN en GBM es más discutida. Revisiones como (Reifenberger and Collins, 2004) señalan que la alteración de este gen aparece en aproximadamente la mitad de los GBM *de novo*, pero en sólo un 10 % de los GBM desarrollados a partir de gliomas de menor grado. La variabilidad en las frecuencias de alteración publicadas para PTEN en diferentes subtipos de GBM concuerda con la variabilidad en perfiles de expresión y de número de copias observados para PTEN en las muestras de GBM analizadas. Además, el nivel de expresión de este gen no se ve afectado en gran medida por las CNAs en su *locus*: la correlación observada entre GE y CN es tan sólo de $r = 0.38$. Estos hechos señalan que PTEN quizás no sea el mejor macador genómico que caracteriza GBM y también por ello no lo detecten nuestros análisis de modo consistente. Por otro lado, la alteración de este gen ha sido asociada también a la pérdida de genes supresores tumorales adicionales en el cromosoma 10, como LGI1 (Chernova et al., 1998) y MXI1 (Wechsler et al., 1997), que si se observan alterados y están incluidos en la tabla.

3.5 Discusión

La integración de datos genómicos de número de copias de DNA junto con datos transcriptómicos de expresión génica facilita la identificación de genes conductores implicados en el desarrollo de enfermedades complejas en las que la inestabilidad genómica juega un papel fundamental, como es el caso del cáncer. En este capítulo se ha presentado un método para la identificación de alteraciones en CN asociadas a cambios diferenciales en GE comunes en la aparición o desarrollo de un estado patológico. El método se propone también como herramienta útil para la búsqueda de genes conductores causales de un proceso tumoral bajo la hipótesis de que los cambios de ex-

presión asociados a un determinado tipo de cáncer son inducidos frecuentemente por alteraciones genómicas.

La utilización de algoritmos de segmentación sobre lo datos de expresión reduce los efectos sobre los niveles de expresión génica no relacionados con sus localizaciones cromosómicas. Mediante la segmentación las señales de sobre-expresión o infra-expresión de genes en *loci* muy cercanos se suavizan y por ello si los efectos de la regulación no están vinculados a las posiciones genómicas tienden a cancelarse. La relación entre CN y GE que se mantiene tras la segmentación será por tanto aquella principalmente asociada a la localización genómica. Tras estos análisis, las discrepancias que puedan seguir existiendo entre los cambios en CN y los cambios de GE serán debidas a otros tipos de regulación, como: regulación epigenética, metilación de DNA o metilación, acetilación y fosforilación de histonas (Wilson et al., 2006; Jones and Baylin, 2007), etc. El análisis de las regiones donde existen estas discrepancias constituye un punto de partida para descubrir otro tipo de regulaciones, abriendo la puerta a la integración por ejemplo de datos epigenéticos como la realizada en (Stransky et al., 2006). En todo caso, el análisis previo de CN y GE utilizando segmentación tiene la ventaja de reducir los efectos de regulación no asociados a la localización genómica.

Otra ventaja del método desarrollado derivada de la utilización de algoritmos de segmentación es la posibilidad de integrar datos provenientes de arrays de diferente resolución sin tener que recurrir al mapeo de sondas de un array a otro y al promediado de sondas con la consiguiente pérdida de información y cobertura.

El método de integración de datos y análisis ha sido aplicado a datos de cáncer GBM identificándose regiones con alteraciones recurrentes de CN en las que se enmarcan genes cuya alteración es clave en el desarrollo de la enfermedad. Los resultados obtenidos son consistentes con alteraciones previamente publicadas en otros trabajos (de Tayrac et al., 2009; Ruano et al., 2006). La presencia de oncogenes y genes supresores tumorales previamente conocidos y asociados con la aparición y progresión de GBM en los resultados obtenidos avala el método desarrollado. El resto de regiones candidatas y genes identificados pueden constituir un punto de partida para la validación de nuevos genes conductores de GBM y estudios posteriores de la progresión de la enfermedad.

Aunque la utilización de datos de GE reduce el tamaño y el número de regiones candidatas respecto a las que se obtendrían únicamente utilizando datos de CN, en muchas de las regiones resultantes en el análisis de GBM la mayoría de los genes ven alterada su expresión y no es difícil elegir de entre ellos un gen conductor concreto. Por un lado, es posible que estas regiones contengan varios de genes conductores, cada uno de ellos contribuyendo en parte de manera individual al desarrollo tumoral, pero que en conjunto tengan un efecto oncogénico aditivo. Por otro lado, también es posible que algunos de los genes con cambios significativos en sus perfiles de expresión sean simplemente genes pasajeros cuya sobre-expresión o infra-expresión no confiera una ventaja para el origen del tumor. La discriminación o determinación de genes conductores frente a genes pasajeros únicamente podría hacerse mediante estudios funcionales examinando los efectos de la combinación de alteraciones entre ellos.

Capítulo

4

Algoritmo de análisis biológico funcional: *GeneTerm Linker*

4.1 Introducción: Análisis biológico funcional

El presente capítulo está centrado en la búsqueda y el análisis de las funciones y procesos en los que están implicados conjuntos de genes derivados generalmente de experimentos de alto rendimiento a gran escala. Cuando se estudian las características de una enfermedad o las consecuencias de la aplicación de un estímulo sobre una célula u organismo, este estudio no concluye generalmente con una lista de los genes diferencialmente expresados o alterados junto con sus p-valores. A menudo los investigadores recorren esa lista escogiendo manualmente aquellos genes que les resultan más interesantes para el proceso estudiado y dirigen la interpretación de los resultados hacia procesos conocidos. Para evitar el sesgo y proporcionar una aproximación más exhaustiva y dirigida por los datos es necesario un análisis de los *pathways* o funciones (Khatri et al., 2012). Este análisis se denomina análisis de enriquecimiento funcional (*Functional Enrichment Analysis*). Sin embargo, muchas veces este tipo de análisis no es aplicado debido a que los resultados obtenidos con la mayoría de las herramientas disponibles son incluso más difíciles de interpretar y resumir que la lista de genes inicial. Es común que tras realizar un análisis de enriquecimiento funcional sobre una lista de varios cientos de genes se obtengan varias listas con cientos de funciones enriquecidas, tras lo cual el investigador termine eligiendo de nuevo de esas listas aquellas funciones que le resultan más interesantes.

En este capítulo se hace una breve introducción de conceptos necesarios para entender el análisis de enriquecimiento funcional planteando sus principales desventajas y necesidades de mejora. Se describe a continuación el método desarrollado para la simplificación de los resultados obtenidos con las herramientas de enriquecimiento funcional, *GeneTerm Linker*, que facilita la interpretación de las funciones biológicas sobre-representadas. Se muestran también los resultados de la validación del método con conjuntos de genes seleccionados así como su aplicación a datos de experimentales.

4.1.1 Principales espacios de anotación biológica

El análisis de enriquecimiento funcional está basado en la utilización de bases de datos de anotaciones biológicas que incluyen grupos de genes asociados a funciones biológicas específicas como vías de señalización celular, vías metabólicas, procesos celulares, etc. Esta información característica de los genes es necesario que pueda ser consultada de una manera sencilla y automática. Entre las bases de datos más utilizadas en este tipo de análisis se encuentran *Gene Ontology* (GO), *Kyoto Encyclopedia of Genes and Genomes* (KEGG) y *UniProt*.

4.1.1.1 Ontología de genes: *Gene Ontology* (GO)

Gene Ontology o GO (Ashburner et al., 2000; Consortium, 2010) es un repositorio de atributos de genes y sus productos con un vocabulario controlado que constituye, hoy por hoy, una de las principales bases de datos de conocimiento biológico.

Debido a que una parte importante de los genes que especifican las funciones biológicas básicas están compartidos por todos los organismos eucariotas, el conocimiento del papel de una proteína en un organismo puede ser extrapolado a otros organismos. Basándose en esta premisa, el trabajo del *GO Consortium* consistió en producir un vocabulario controlado y estructurado que describiese el papel de los genes y sus productos dentro de cualquier organismo. Para ello se crearon tres ontologías independientes:

(i) Procesos biológicos (*Biological Process*, GO-BP), se refiere a un objetivo biológico en el que contribuye el gen. Un proceso está compuesto por un conjunto de funciones moleculares que, a menudo, implican una transformación química o física. Ejemplos de proceso biológico se consideran 'crecimiento celular y mantenimiento' o 'transducción de señales'.
(ii) Funciones moleculares (*Molecular Functions*, GO-MF), se define como una actividad bioquímica de un producto génico. Describe la actividad realizada, pero no dónde ni cuándo sucede. Ejemplos de funciones moleculares pueden ser 'enzima' o 'kinasa'.
(iii) Componentes celulares (*Cellular Components*, GO-CC), se refiere al lugar de la célula donde está activo un determinado producto génico. Ejemplos pueden ser 'retículo endoplásmico' o 'nucleosoma'.

Las relaciones entre los genes o productos génicos con los términos de la ontología no son unívocas sino de tipo uno a muchos, permaneciendo así fiel a la realidad biológica en la que un mismo gen o proteína puede verse involucrado en más de un proceso simultáneamente. Un ejemplo de términos y relaciones en las tres ontologías puede verse en la figura 4.1

La ventaja de usar ontologías es la capacidad de representar no sólo las entidades, sino también las relaciones existentes entre ellas. El desarrollo y uso de un vocabulario biológico controlado y estructurado en GO representa el conocimiento biológico actual y a la vez permite organizar los nuevos conocimientos que sean añadidos con posterioridad. Además permite el acceso de forma más sencilla a la información, tanto para las personas como para las herramientas computacionales construidas para su manejo.

4.1.1.2 Vías metabólicas y de señalización: *Kyoto Encyclopedia of Genes and Genomes* (KEGG)

La enciclopedia KEGG, *Kyoto Encyclopedia of Genes and Genomes* es una base de datos que contiene información de las funciones a alto nivel de un sistema biológico como la célula, el organismo y el ecosistema, integrando información que va desde un nivel molecular hasta genómico

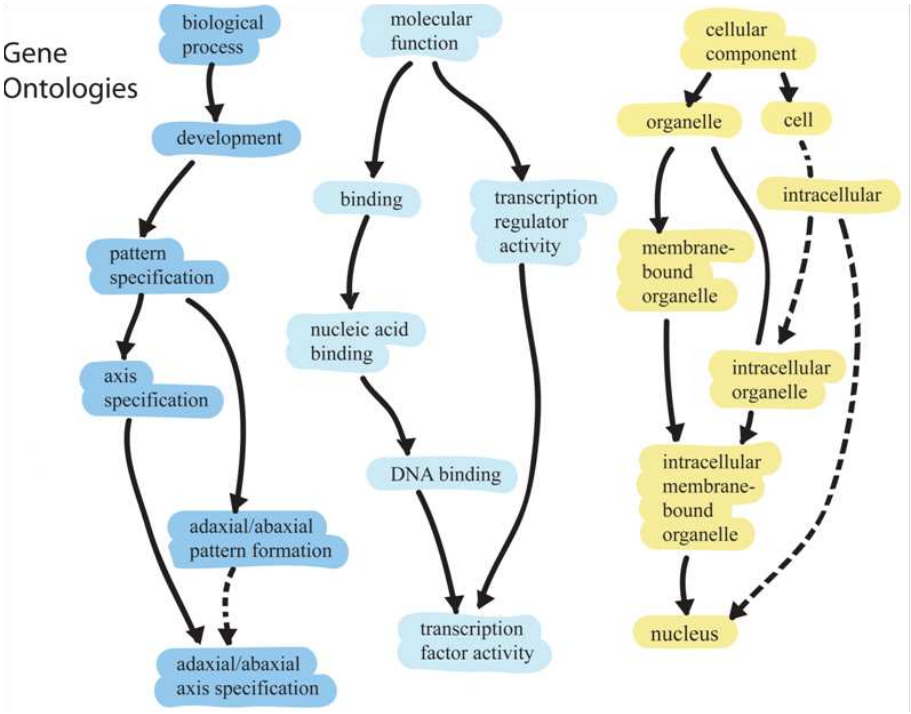


Figura 4.1: Ontologías génicas(GO) - Ejemplo de representación de los 3 espacios de categorías ontológicas que incluye *Gene Ontology* (GO).

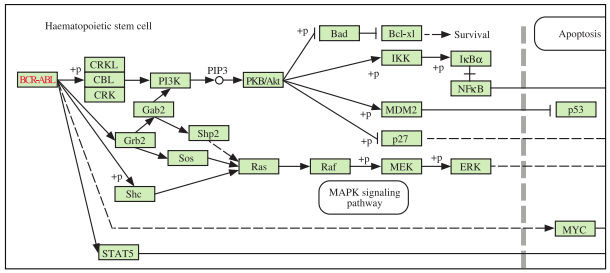
(Kanehisa and Goto, 2000). La parte más completa de KEGG corresponde a las vías y rutas metabólicas y de señalización (*pathways*). Ésta es la parte que se utilizará en el presente capítulo como espacio de anotación.

El diseño global actual de KEGG es modular, integra hasta 16 bases de datos diferentes que se incluyen dentro de 3 categorías más generales (Kanehisa et al., 2010) (Kanehisa et al., 2012):

(i) Información genómica (*Genomic Information*), que consiste en bloques de construcción moleculares formados por genes y proteínas.
(ii) Información química (*Chemical Information*), que incluye biomoléculas pequeñas (metabolitos y productos/sustratos celulares), reacciones entre ellas y otras estructuras químicas derivadas de estas reacciones.
(iii) Información de sistemas (*Systems Information*), que integra los bloques de información anteriores en diagramas de interacciones moleculares, reacciones y redes relacionales. Incluye los mapas de vías y rutas moleculares, jerarquías funcionales, enfermedades (entendidas como alteraciones de las rutas) y drogas (entendidos como alteradores de las rutas).

Como se puede ver, la mayor parte de la información contenida en KEGG es modular y gradual, esto es, unos paquetes de información son construidos a partir de otros hasta llegar a contener información de las funciones sistémicas de la célula o el organismo. Todo el conocimiento experimental de esas funciones sistémicas es añadido y organizado de tres formas diferentes: como un mapa de las rutas metabólicas y de señalización (*Pathway map*), como listas de componentes biomoleculares (*Simple list*), o como datos estructurados de una manera jerárquica (*Hierarchical list*). Un ejemplo asociado a la leucemia mieloide crónica puede verse en la figura 4.2 tomada de (Kanehisa et al., 2010).

(**a**) Pathway map



(**b**) Simple list

Disease gene
- BCR-ABL (translocation)
- EVI1 (overexpression)
- AML1 (translocation)
- p16/INK4A (mutation)
- p53 (mutation)
- RB1 (mutation)

Carcinogen
- 1,3-Butadiene

Marker
- BCR-ABL (translocation)
- WT1

Drug
- Imatinib mesylate (Gleevec)
- Hydroxyurea
- Interferon-alpha

(**c**) Hierarchical list

- **Cancers**
  - **Cancers of the Nervous System**
  - **Cancers of the Digestive System**
  - **Cancers of Haematopoietic and Lymphoid Tissues**
    - H00003 Acute myeloid leukemia (AML) [PATH:hsa05221]
    - H00001 Acute lymphoblastic leukemia (ALL) (Precursor B
    - H00002 Acute lymphoblastic leukemia (ALL) (Precursor T
    - H00004 Chronic myeloid leukemia (CML) [PATH:hsa05220]
    - H00005 Chronic lymphocytic leukemia (CLL)
    - H00007 Hodgkin lymphoma
    - H00006 Hairy-cell leukemia
    - H00008 Burkitt lymphoma
    - H00009 Adult T-cell leukemia
    - H00010 Multiple myeloma
    - H00011 Lymphoplasmacytic lymphoma
    - H00012 Polycythemia vera
  - **Cancers of the Breast and Female Genital Organs**
  - **Cancers of Soft Tissues and Bone**
  - **Skin Cancers**
  - **Cancers of the Urinary System and Male Genital Organs**
  - **Cancers of Endocrine Organs**
  - **Head and Neck Cancers**
  - **Cancers of the Lung and Pleura**
- **Immune System Diseases**
- **Nervous System Diseases**
- **Circulatory System Diseases**

Figura 4.2: Tipos de representación de información en KEGG - El ejemplo muestra (a) el pathway para leucemia mieloide crónica (hsa05220), (b) los elementos asociados con leucemia mieloide crónica en KEGG DISEASE (H00004) y (c) parte de la estructura jerárquica para la clasificación de enfermedades humanas (br08402). Adaptado de (Kanehisa et al., 2010).

4.1.1.3 **Estructura y función de proteínas: *Integrated repository of protein families, domains and functional sites* (InterPro)**

La función de cada proteína está intrínsecamente ligada a su estructura. La secuencia primaria de aminoácidos que la integra, los plegamientos secundario, terciario y cuaternario, así como su interacción con otras proteínas y otras biomoléculas para formar macroestructuras más complejas (complejos proteicos) definen sus características y funciones. En este sentido es importante encontrar patrones en su estructura puesto que pueden estar ligados a determinados comportamientos o funciones celulares. Una fuente de información para obtener este tipo de datos es InterPro (Hunter et al., 2012). InterPro (www.ebi.ac.uk/interpro/) es una base de datos de familias de proteínas establecidas en base a modelos y patrones que integran información de alineamientos de secuencias e identificación de dominios y de sitios funcionales. Está estructurado también de manera jerárquica, como se puede apreciar en la figura 4.3 en la que se muestra la estructura de la familia de proteínas tirosina quinasa.

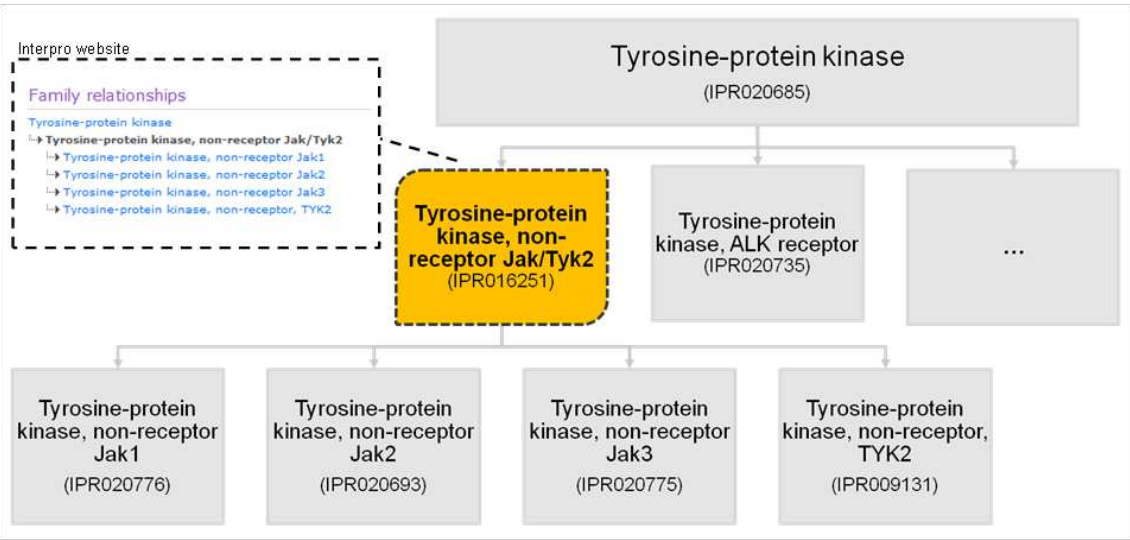


Figura 4.3: Estructura de una familia de proteínas en InterPro - El ejemplo muestra la estructura jerárquica de la familia de proteínas tirosina quinasa en InterPro. (Fuente: www.ebi.ac.uk/interpro).

InterPro está formado por un consorcio de bases de datos independientes que son: PROSITE (Hulo et al., 2008), PRINTS (Attwood et al., 2003), Pfam (Finn et al., 2010), ProDom (Bru et al., 2005), SMART (Letunic et al., 2012), TIGRFAMs (Haft et al., 2003), PIRSF (Nikolskaya et al., 2006), SUPERFAMILY (Wilson et al., 2009), PANTHER (Mi et al., 2010) y Gene3D (Lees et al., 2012). Cada una de estas bases de datos utiliza metodologías y algoritmos diferentes, así como diferentes tipos de información biológica de proteínas para inferir patrones. Todas estas bases de datos tienen un tamaño similar, sin embargo difieren en contenido y pueden considerarse complementarias. La integración de estas bases de datos se realiza de forma manual por expertos evitando duplicidades y redundancias. Así, esta integración y unificación proporciona una mayor cobertura además de unos resultados más fiables y permite, en algunos casos, buscar las relaciones biológicas existentes entre los diferentes patrones constituyentes.

4.2 Motivación: Problemas del análisis biológico funcional

El análisis de enriquecimiento funcional, normalmente citado como *Functional Enrichment Analysis* (EA) (Huang et al., 2009a), facilita la tarea de inferir implicaciones funcionales en conjuntos de genes o conjuntos de proteínas que cooperan. El problema de las técnicas de enriquecimiento subyace en la redundancia de sus resultados, incluyendo, en muchos casos, información trivial que puede contribuir a enmascarar otras realidades biológicas más relevantes presentes en el estudio realizado. El análisis de enriquecimiento funcional clásico no se centra en la solución de los problemas derivados de la repetitividad y el solapamiento de los diferentes espacios de anotación. Así por ejemplo podemos encontrar frecuentemente problemas como:

(i) **Redundancia**: Existen numerosos términos redundantes, repetidos en los diferentes espacios de anotación (por ejemplo GO:0007049: *cell cycle* y KEGG hsa04110: *cell cycle*) o términos equivalentes con el mismo significado biológico (por ejemplo GO:0007049: *cell cycle* y GO:0022402: *cell cycle process*).

(ii) **Imprecisión**: Sesgo causado por la anotación demasiado frecuente de términos genéricos, inespecíficos, poco precisos en los espacios de anotación (por ejemplo GO:0050789: *protein binding* que incluye alrededor del 40 % de los genes de *Homo Sapiens* anotados en GO-MF).

(iii) **Falta de anotación**: Carencias en la anotación de genes ampliamente conocidos, como es el caso de NRAS, que no está anotado al término GO:0043410: *positive regulation of MAPKKK cascade*, y, sin embargo, es parálogo al gen HRAS que cumple un papel importante en la vía de señalización de MAPK kinasas.

En este marco, se plantea la necesidad de solventar los problemas descritos y otras dificultades que surgen a la hora de interpretar los resultados derivados de análisis biológicos funcionales complejos. Lograr una simplificación de estos resultados es fundamental, pero para ello es necesario realizar un análisis exhaustivo de los espacios de anotación.

De este modo, como paso previo antes de acometer los problemas detectados en la mayoría de las herramientas de enriquecimiento, se han analizado y comparado las distribuciones de frecuencias de los términos biológicos en las dos bases de datos más utilizadas para anotación: GO en sus tres espacios de anotación y KEGG (ambas comentadas previamente en 4.1.1). En estas bases de datos las funciones están anotadas con términos específicos que describen los roles biológicos utilizando vocabularios controlados y estructurados. Estos vocabularios no son estáticos en el tiempo sino que evolucionan de forma natural modulando su significado, tal y como lo hace el lenguaje cotidiano. Un ejemplo de esta modulación puede ser la pérdida de significado que experimentan algunos términos al convertirse en términos populares utilizados en gran variedad de contextos.

El análisis del número de genes asignados a cada término revela que las distribuciones no son homogéneas. La figura 4.4 presenta las distribuciones de densidad del número de genes por término. En el caso de GO más del 50 % de los términos tienen asignados menos de 4 genes, algo que no sucede en KEGG, con una distribución más homogénea semejante a una campana de *Gauss*. Centrando la atención en la diferencia de longitudes de las colas de las distribuciones se observa la presencia de términos con miles de genes anotados en GO que aparecen como términos atípicos o *outliers* de las distribuciones (representados como círculos en los boxplots de la figura 4.5.

En la tabla 4.2 se resumen los tres términos con un mayor número de genes para GO-BP, GO-MF, GO-CC junto con el número de genes anotados para *Homo sapiens* en cada uno. El término más utilizado en GO-BP está asociado con aproximadamente el 6 % de los genes humanos (término GO:0007165 *Signal transduction*: 1872 genes asignados). Estos términos, como es el caso de GO:0005515: *Protein binding*, son los elementos atípicos en la distribución, que están sobrerepresentados y pueden ser considerados como términos promiscuos y poco informativos, de-

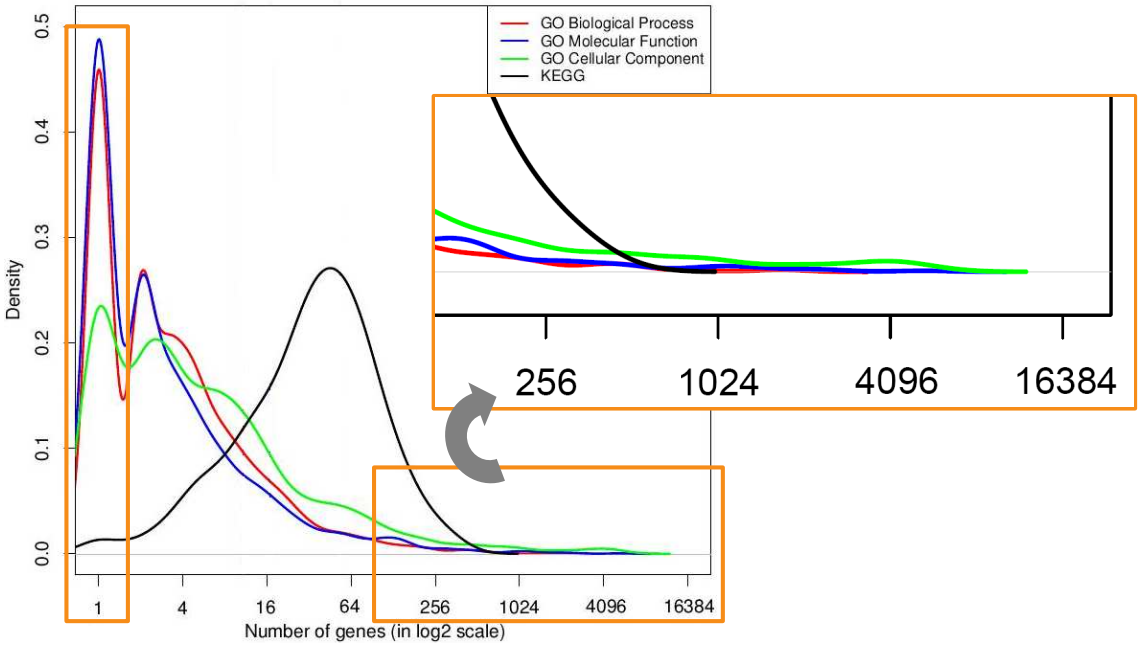


Figura 4.4: Distribuciones del número de genes anotados a cada término - Cada color representa un espacio de anotación diferente mostrado en la leyenda. El número de genes por término está representado en escala logarítmica. La zona ampliada corresponde a términos que se asignan a un gran número de genes.

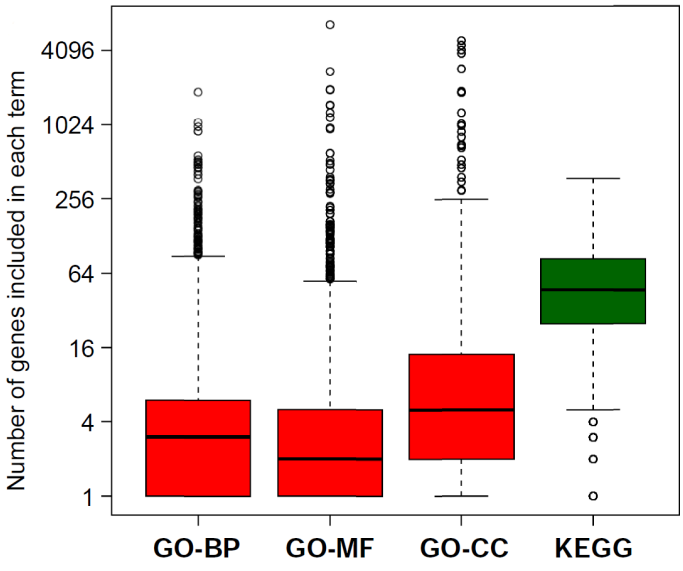


Figura 4.5: Distribuciones del número de genes anotados a cada término - Se presentan los *boxplots* de las distribuciones de número de genes anotados a términos GO-BP, GO-MF, GO-CC (en color rojo) y KEGG (en color verde) para *Homo sapiens*.

masiado generales para proporcionar información clara y útil por ellos mismos en unos resultados de análisis funcional.

Categoría	Términos	# Genes
GO-BP		
GO:0007165	Signal transduction	1872
GO:0006355	Regulation of transcription, DNA dep.	1063
GO:0045449	Regulationof transcription	987
GO-MF		
GO:0005515	Protein binding	6618
GO:0046872	Metal ion binding	2759
GO:0008270	Zinc ion binding	1956
GO-CC		
GO:0005634	Nucleus	4877
GO:0005737	Cytoplasm	4457
GO:0016021	Integral to membrane	4109

Tabla 4.1: Términos sobre-representados en los tres espacios de anotación de GO.

4.3 Desarrollo metodológico del algoritmo

De forma general, los métodos de enriquecimiento funcional buscan patrones frecuentes de asociación entre los genes y las descripciones o funciones anotadas en bases de datos biológicas. Estos patrones pueden ser analizados como conjuntos de elementos (*itemsets*) frequentes. Una definición formal de un conjunto de elementos frecuentes, *frequent itemset*, podría ser:

Dado un conjunto de ítems $I = \{i_1, ..., i_n\}$ y una base de datos de transacciones $T = \{t_1, ..., t_m\}$ donde cada transacción es un subconjunto de I, $F \subseteq I$ es un conjunto de ítems frecuente si está incluido en un número de transacciones mayor que un umbral dado, ε.

En el contexto biológico de enriquecimiento funcional se pueden establecer equivalencias e identificar los términos en los espacios de anotación como ítems y los genes como las transacciones que los soportan. De esta manera los resultados de las herramientas de enriquecimiento funcional, entendidos como *frequent itemsets* pueden simplificarse como un conjunto de términos asociados a un conjunto de genes con un *score* o un p-valor que es derivado del análisis de enriquecimiento e indica la fortaleza de la asociación.

Formalmente definimos cada uno de estos conjuntos genes-términos o *GeneTerm sets* como una tupla: $E_i =< G_i, A_i, p_i >$ donde E_i es el *GeneTerm set* i-ésimo de los resultados, G_i es un conjunto de genes $\{g_1, g_2, \ldots, g_n\}$, A_i es un conjunto de términos o anotaciones biológicas $\{a_1, a_2, \ldots, a_n\}$ y p_i es el p-valor.

En términos de enriquecimiento se puede decir que A_i es el conjunto de términos sobre-representados en una lista de genes y G_i es el conjunto de genes que soporta esa sobre-representación con un p-valor p_i.

La mayoría de las herramientas de análisis de enriquecimiento funcional, bien de tipo GSEA, *Gene Set Enrichment Analysis*) (Subramanian et al., 2005) o análisis de enriquecimiento tradicionales, proporcionan largas de listas de *GeneTerm sets* derivadas del análisis de diferentes espacios de anotación. Estas listas contienen términos y funciones redundantes y muchas veces demasiado genéricas que enmascaran parte de las funciones más específicas y no permiten extraer conclusiones biológicas significativas de una manera sencilla.

En este contexto proponemos un método basado en la búsqueda y construcción de “metagrupos” de genes y términos. Estos “metagrupos” estarán compuestos por *GeneTerm sets* seleccionados y relacionados, minimizando la influencia de términos y anotaciones redundantes y poco informativas, y de optimizando la significación biológica de los resultados de enriquecimiento.

En los siguientes apartados se describen los cincos pasos secuenciales de los que consta el método propuesto.

4.3.1 **Paso 1: Filtrado de términos poco informativos**

Tal y como se muestra en la sección 4.2, aquellos términos con un número de genes asociado mucho mayor que el que cabría esperar en un determinado espacio de anotación incorporan información poco útil por sí solos, mientras que pueden estar enmascarando otros patrones interesantes menos frecuentes, pero aún así significativos.

Una vez analizado cada espacio de anotación disponible para cada organismo e identificados los términos poco informativos, el primer paso consiste en eliminar aquellos *Geneterm sets* que incluyen únicamente estos términos poco informativos. Es decir, dado un elemento E_i del conjunto de resultados de enriquecimiento será eliminado si y sólo si el conjunto de términos A_i está compuesto únicamente por los denominados “términos poco informativos”.

Sea $R = \{a_1, \ldots, a_r\}$ el conjunto de términos poco informativos, $E_i =< G_i, A_i, p_i >$ será filtrado $\Leftrightarrow A_i \subseteq R$.

De este modo, si el *Geneterm set* contiene una combinación de este tipo de términos junto con otros que no pertenecen a esta categoría no será eliminado puesto que los términos poco informativos pueden resultar útiles para matizar la información proporcionada. En este paso se produce una reducción considerable en el número de *Geneterm sets* puesto que filtra elementos no descriptivos por sí solos aunque los mantiene siempre y cuando maticen la información. La figura 4.6 contiene una representación esquemática del proceso para dos términos poco informativos, C y K.

Genes	Terms	P-value
gen1, gen2, gen3, gen4	A	0.001
gen8, gen7	M, L	0.003
gen1, gen2	A, B, C	0.007
gen1, gen3	K	0.002
gen8, gen9	J, K, L	0.005
gen1, gen2, gen3	A, C	0.025
gen5, gen6	K, C	0.035
...	...	...

Genes	Terms	P-value
gen1, gen2, gen3, gen4	A	0.001
gen8, gen7	M, L	0.003
gen1, gen2	A, B, C	0.007
gen8, gen9	J, K, L	0.005
gen1, gen2, gen3	A, C	0.025
...	...	...

Figura 4.6: Esquema del filtrado de términos poco informativos - Los términos C y K han sido identificados como poco informativos en sus respectivos espacios de anotación y por ello los *GeneTerm sets* en los que aparecen solos son eliminados.

4.3.2 **Paso 2: Generación de módulos funcionales**

El segundo paso del método se centra en la agrupación de *Geneterm sets* similares para formar módulos funcionales que faciliten la interpretación de los resultados de enriquecimiento.

Es posible encontrar clústers coherentes de *itemsets* relacionados de manera que el conjunto total sea más fácilmente entendible (Toivonen et al., 1995). Por un lado se han desarrollado algoritmos basados en el agrupamiento de ítems (Lent et al., 1997; Liu et al., 1999). Sin embargo, la aplicación del coeficiente de similitud de *Jaccard* utilizado para medir la distancias entre los conjuntos de transacciones que soportan los *itemsets* obtiene mejores resultados que las medidas de similitud calculadas en base a los ítems (Gupta et al., 1999).

Apoyados en estos estudios, para el agrupamiento de *itemsets* frecuentes en este paso del algoritmo se utilizan las medidas de similitud basadas en transacciones (es decir, genes en el contexto de *Geneterm sets*). Estas medidas capturan mejor las interacciones entre los conjuntos de *items* y son robustas e independientes del tamaño de los datos. Además en el algoritmo desarrollado se ha tenido en cuenta el p-valor, la fortaleza de la asociación entre las transacciones y los ítems (es decir, entre los genes y los términos).

Para ello, para cada *Geneterm set* E_i se crea un vector v_i como sigue:

$$v_i = (\delta(g_1, G_i), \delta(g_2, G_i), \ldots, \delta(g_M, G_i), M \times p_i)$$

donde:

$$\delta(g_k, G_i) = \begin{cases} 1 & g_k \in G_i \\ 0 & g_k \notin G_i \end{cases} \tag{4.1}$$

y M es el número total de genes en el conjunto de *Geneterm sets*.

De esta manera el vector v_i contiene 1 en la posición k-ésima si el gen k está incluido en el conjunto de genes del *Geneterm set* E_i y 0 en caso contrario. Además se añade el p-valor ponderado por el número total de genes M en la coordenada $M + 1$ del vector.

Las distancias entre cada par de vectores v_i son calculadas usando la distancia Coseno, generalización para vectores no binarios derivada de la distancia de *Jaccard* empleada en (Gupta et al., 1999) y (Plasse et al., 2007):

$$D(E_i, E_j) = 1 - \cos(\vec{v_i}, \vec{v_j}) = 1 - \frac{v_i \cdot v_j}{\|v_i\|\|v_j\|} \tag{4.2}$$

Una vez que las distancias han sido calculadas para cada par de *Geneterm sets* E_i y E_j es necesario utilizar un algoritmo de agrupamiento que nos permita definir metagrupos preliminares. Se ha utilizado el método *Ward* (Ward, 1963) de agrupamiento jerárquico puesto que el conjunto de *Geneterm sets* es homogéneo, sin valores extremos que puedan alterar el agrupamiento. Como se desprende de los resultados obtenidos en la comparación de diferentes métodos de aglomerativos realizada en (Plasse et al., 2007), los grupos obtenidos con el método *Ward* son equilibrados con un número de elementos muy similar. La figura 4.7 muestra los árboles jerárquicos obtenidos con diferentes algoritmos de agrupamiento. Cada uno de los extremos u hojas del árbol se corresponde con un *Geneterm set*. En los diferentes árboles se observa un agrupamiento similar, sin embargo el método *Ward* proprorciona grupos más compactos y definidos.

A la hora de definir los grupos se considera un punto de corte heurístico basado en la profundidad del árbol. Inicialmente se establece como umbral el 20 % de la altura del árbol generado, pero

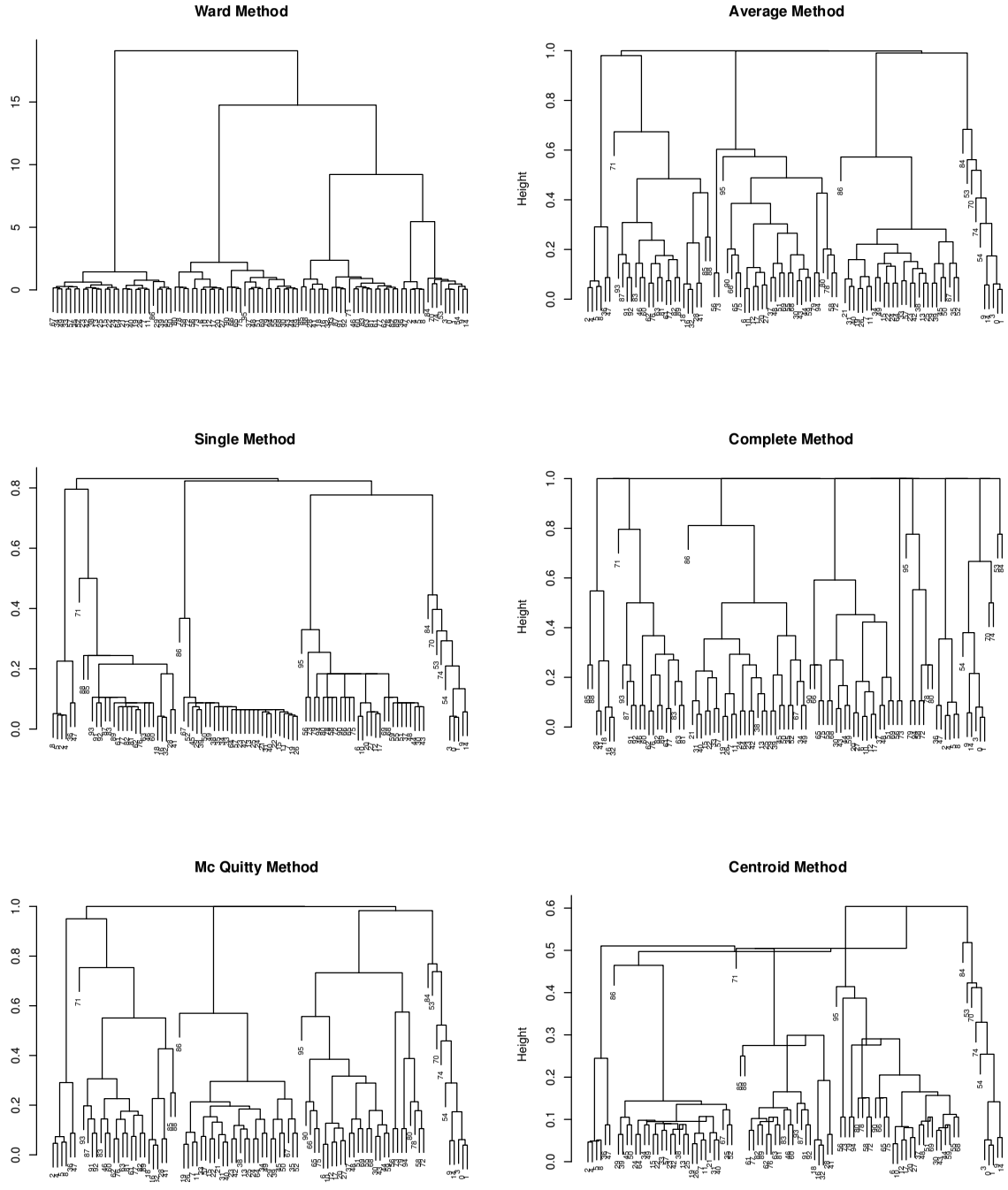


Figura 4.7: Resultado de 6 métodos de agrupamiento jerárquico no supervisado - Visualización del agrupamiento de *Geneterm sets* con seis métodos de agrupamiento no supervisado.

si éste no es suficiente para lograr al menos una división de los datos, se incrementa el umbral un 10 % en la altura iterativamente hasta lograr la agrupación de al menos 2 *Geneterm sets* en el clúster. De este modo, no se fuerza el establecimiento de un número de grupos predefinido, sino que se agrupan solamente aquellos *Geneterm sets* que representen información verdaderamente relacionada, módulos funcionales de genes y términos. Conviene notar que el modelo implementado se basa en la utilización agrupamiento jerárquico no supervisado con el método de *Ward*, sin embargo es independiente del algoritmo concreto utilizado pudiéndose implementar otras versiones cambiando el método de agrupamiento.

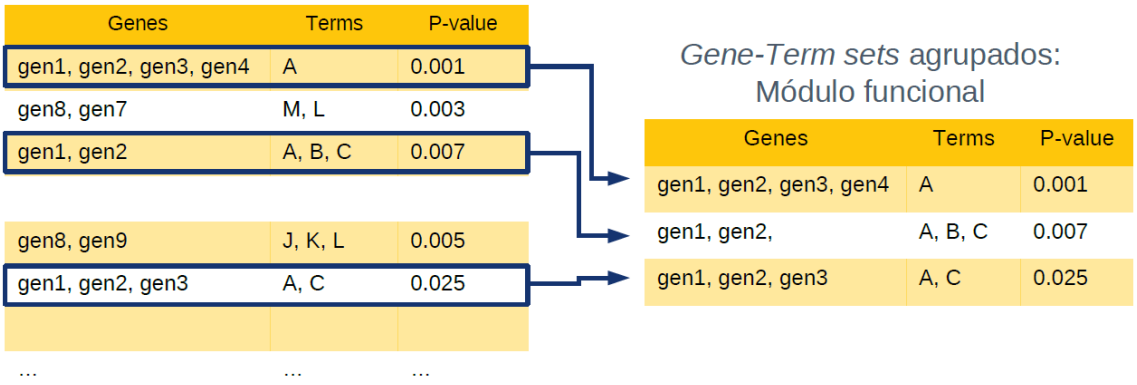


Figura 4.8: Agrupamiento de *Gene-Term sets* en módulos funcionales - Representación esquemática del proceso del algoritmo en el que se logra el agrupamiento de *Geneterm sets* en módulos funcionales o metagrupos preliminares.

El proceso correspondiente a este paso del algoritmo está representado esquemáticamente en la figura 4.8.

4.3.3 Paso 3: Convergencia de términos

Todo el proceso de agrupamiento está dirigido por las transacciones, es decir, los genes de los *Geneterm sets*. Así cada módulo funcional contiene anotaciones biológicas con información de los procesos en los que participan los genes que lo forman. Sin embargo, es frecuente encontrar diferentes grupos de genes implicados en los mismos procesos biológicos o funciones. Uno de los objetivos del método propuesto es establecer una correspondencia recíproca entre los genes y los términos biológicos, por lo que es necesario tener en cuenta también dichos términos a la hora de definir módulos funcionales más generales.

Para ello, una vez realizado el agrupamiento preliminar basado en los genes, el método combina los clústers que están involucrados en funciones similares agrupando recursivamente los que comparten los mismos términos biológicos. Se construyen así módulos funcionales donde la convergencia de genes y términos está maximizada. El proceso se representa esquemáticamente en la figura 4.9.

4.3.4 Paso 4: Eliminación de redundancias

Una vez generados los módulos funcionales es todavía posible compactar y reducir más su tamaño sin disminuir la información proporcionada eliminando aquellos *Geneterm sets* redundantes en cada uno.

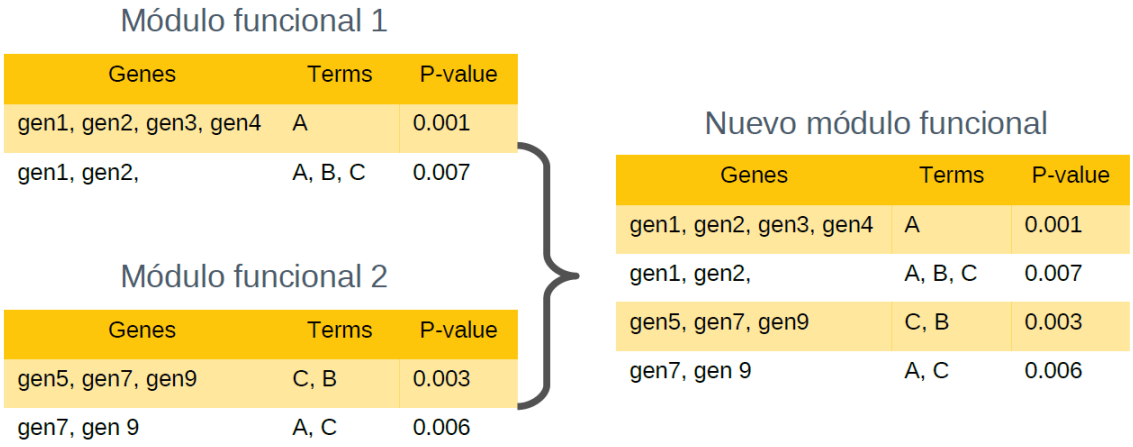


Figura 4.9: Convergencia de módulos funcionales en base a los términos - Representación esquemática de la convergencia de módulos funcionales (i.e. los metagrupos preliminares derivados del agrupamiento) que se realiza teniendo en cuenta los términos biológicos que los componen.

Toivonen y colaboradores (Toivonen et al., 1995) proponen el concepto de *cover* de un conjunto de reglas de asociación como el mínimo subconjunto que contiene todas las relaciones presentes en el conjunto original. Es posible extender este concepto a un módulo funcional de *Geneterm sets* con la idea de mantener la completitud de los datos eliminando redundancias sin perder información. En este contexto no sólo es necesario mantener las anotaciones o combinaciones de anotaciones enriquecidas, sino también conservar todos los genes que soportan dicho enriquecimiento.

Cada módulo funcional contiene un conjunto de *Geneterm sets*, que a su vez están formados por un conjunto de términos y un conjunto de genes. De esta manera podemos decir que el módulo queda definido por la unión de los genes y términos de los *Geneterm sets* que lo constituyen. Sin embargo, no todos los *Geneterm sets* añaden información. Aquellos cuyo conjunto de genes y anotaciones está incluido en otros *Geneterm sets* del mismo módulo resultan redundantes. El *cover* de un módulo será entonces un subconjunto de sus *Geneterm sets* que garantice su completitud, es decir, que mantenga la descripción del módulo intacta. Formalmente, dado un módulo $\Gamma = \{E_1, E_2 \dots E_n\}$ y un subconjunto $\Delta \subseteq \Gamma$, decimos que:

$$\Delta \textit{ es cover de } \Gamma \iff \Big(\bigcup_{E_k \in \Delta} \gamma(E_k) = \bigcup_{E_k \in \Gamma} \gamma(E_k)\Big) \wedge \Big(\bigcup_{E_k \in \Delta} \alpha(E_k) = \bigcup_{E_k \in \Gamma} \alpha(E_k)\Big) \tag{4.3}$$

donde: $\gamma(E_i) = G_i$ y $\alpha(E_i) = A_i$

En la búsqueda del *cover* es importante tener en cuenta la significación o fortaleza de la asociación entre los genes y los términos. Para ello los *Geneterm sets* del metagrupo son ordenados por su p-valor eliminando consistentemente aquellos con peor p-valor y que no añaden ningún gen o término nuevo al módulo. La figura 4.10 muestra esquemáticamente este proceso, al final del cual se consiguen los módulos funcionales finales o metagrupos.

4.3.5 Paso 5: Significación y coherencia de los metagrupos finales

Una vez obtenidos los módulos funcionales o metagrupos finales y establecido su *cover* es posible estimar su relevancia, calidad y coherencia. Para conocer la relevancia o significación estadística

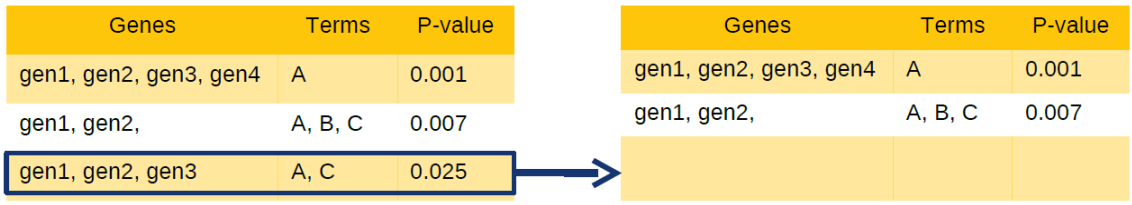


Figura 4.10: Eliminación de *Geneterm sets* redundantes - Representación esquemática del proceso del algoritmo en el que se eliminan los *Geneterm sets* redundantes en cada uno de los módulos funcionales construidos en el paso anterior.

de los metagrupos finales se aplica un test hipergeométrico en el que se tienen en cuenta los genes que los caracterizan y los genes anotados a cada uno de los *Geneterm sets* que lo forman.

A la hora de estimar la calidad de los metagrupos funcionales es importante evaluar su compactación, es decir, la homogeneidad dentro del grupo y la proximidad o separación de los diferentes grupos entre sí. El coeficiente Silueta (*Silhouette coefficient*) es una medida de validación de resultados de agrupamiento que tiene en cuenta ambos parámetros. Este coeficiente varía entre -1 y 1 de manera que valores cercanos a 1 nos indican que el módulo está bien definido, es homogéneo o compacto y está suficientemente diferenciado del resto. De la misma manera un valor de Silueta negativo o cercano a 0 indicaría que el módulo es difuso y algunos de sus *Geneterm sets* podrían estar agrupados igualmente en otros módulos.

Una vez obtenidas estas medidas definimos formalmente metagrupo como la tupla $M_i =< G_i, T_i, p_i, s_i >$ donde G_i es la unión de todos los genes de los *Geneterm sets* que forman el módulo $G_i = \{g_1, g_2, \ldots, g_m\}$, T_i es la unión de los términos de los *Geneterm sets* del modulo $T_i = \{t1, t2, \ldots, t_n\}$, p_i es el p-valor obtenido con el test hipergeométrico, y s_i el coeficiente Silueta para dicho módulo.

Adicionalmente se han calculado otras medidas de caracterización del metagrupo como el diámetro, distancia máxima entre los *Geneterm sets* del módulo, o el coeficiente de similitud, que es la media de las similitudes entre cada par de *Geneterm sets*.

4.4 Aplicación y validación del algoritmo *GeneTerm Linker*

En este apartado se presentan los resultados de validación del método descrito que hemos denominado: ***GeneTerm Linker***. Para validar la capacidad del método propuesto para encontrar metagrupos de genes y anotaciones funcionalmente relacionados se han utilizado varios conjuntos de datos.

En primer lugar, se ha utilizado un conjunto de 59 proteínas nucleares de levadura (*Saccharomyces cerevisiae*) que constituyen 5 complejos bien definidos mostrados en la tabla 4.4. Estos 5 grupos de genes/proteínas han sido bien caracterizados mediante métodos experimentales de detección de interacciones proteína-proteína y constituyen un conjunto que ha sido también usado para encontrar regiones densamente conectadas en redes de interacción de proteínas (Bader and Hogue, 2003). La figura 4.11 muestra la red de interacción de proteínas generada para este conjunto de datos, construida utilizando la base de datos APID y la herramienta APID2NET (Prieto and De Las Rivas, 2006) y (Hernandez-Toro et al., 2007) respectivamente. APID se utilizó como fuente de información sobre los datos experimentales conocidos para las interacciones y APID2NET como herramienta bioinformática de visualización integrada en *Cytoscape* (www.cytoscape.org).

Complex (Member proteins)	Number of proteins
a. mRNA cleavage and polyadenylation specificity factor complex (CFT1, CFT2, FIP1, GLC7, MPE1, PAP1, PFS2, PTA1, PTI1, REF2, RNA14, YSH1, YTH1)	13
b. anaphase-promoting complex (APC1, APC2, APC4, APC5, APC9, APC11, CDC16, CDC23, CDC26, CDC27, DOC1)	11
c. proteasome, 19/22S regulator complex (RPN1, RPN2, RPN3, RPN5, RPN6, RPN8, RPN10, RPN11, RPN13, RPT1, RPT3, RPT6)	12
d. U6 snRNP complex (DCP1, KEM1, LSM1, LSM2, LSM3, LSM4, LSM5, LSM6, LSM7, LSM8, PAT1)	11
e. SAGA complex (ADA2, GCN5, HFI1, NGG1, SGF11, SGF29, SGF73, SPT3, SPT7, SPT8, SPT20 TRA1)	12

Tabla 4.2: Proteínas de levadura seleccionadas que constituyen 5 complejos bien definidos

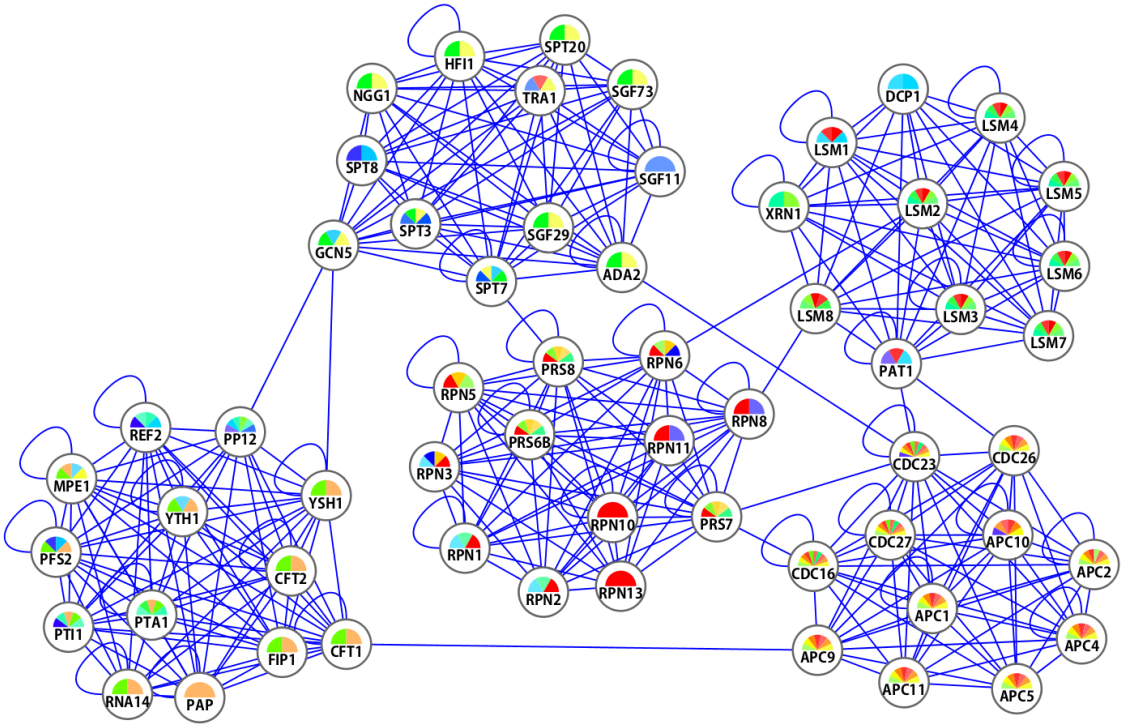


Figura 4.11: Red de 59 proteínas de levadura obtenida mediante datos experimentales de interacción - Los nodos representan las 59 proteínas y los enlaces se corresponden con las interacciones experimentales obtenidas de APID. En cada nodo están marcados con diferentes colores los términos biológicos asignados en GO-BP e InterPro a cada proteína.

Bajo la premisa de que las proteínas de un complejo deberían estar anotadas a las mismas funciones biológicas, el método de análisis recíproco de genes y términos desarrollado, ***GeneTerm Linker***, debería ser capaz de reconstruir 5 módulos utilizando únicamente anotaciones sobre la estructura y funciones de los 59 genes individualmente.

Para probar esta hipótesis se realizó un análisis de enriquecimiento funcional de los 59 genes utilizando *GeneCodis* (Nogales-Cadenas et al., 2009) donde se seleccionaron las bases de datos de anotación *Gene Ontology* (*Biological Process*, *Molecular Function y Celular Component*), InterPro y KEGG, con un soporte mínimo de 4 genes y utilizando un test hipergeométrico con corrección por FDR para calcular el enriquecimiento. Los resultados de GeneCodis fueron 127 *Geneterm sets*, de los cuales 31 contenían únicamente anotaciones poco informativas descritas en la sección 4.3.1.

La aplicación de ***GeneTerm Linker*** simplificó y organizó los resultados en 5 metagrupos con un total de 49 *Geneterm sets*. Los resultados obtenidos pueden verse en detalle en el material suplementario de Fontanillo et al. (2011).

El resumen del análisis de enriquecimiento funcional en 5 metagrupos facilita la visualización y su interpretación. La tabla 4.4 muestra los genes que forman cada metagrupo, sus p-valores, los valores Silueta y los términos o anotaciones asociados a cada uno de ellos. Los p-valores obtenidos con el test hipergeométrico son muy significativos y los valores de Silueta altos ($> 0{,}5$) reflejando que los metagrupos encontrados son compactos y constituyen unidades diferenciadas. Como se puede observar en esta tabla los metagrupos se corresponden con los complejos analizados e incluyen de modo casi exacto los genes descritos para cada uno (ver 4.4). Las principales funciones y roles biológicos de cada complejo están representadas en el conjunto de anotaciones respectivo, donde algunos de los términos son sinónimos como en el metagrupo 3 *GO:0000502:Proteasome complex* y *KEGG:03050:Proteasome* y otros complementarios como *GO:0046540:U4/U6 tri-snRNP complex* e *IPR:001163:Like-Sm ribonucleoprotein (LSM) domain*.

Met.	Genes	g(t)	G(u)	P-valor	Silueta	Términos
1	GLC7, REF2, YTH1, FIP1, PAP1, PFS2, CFT1, RNA14, PTI1, PTA1, MPE1, CFT2, YSH1	13(59)	15(7103)	2.26E-26	0.529	mRNA cleavage and polyadenylation specificity factor complex (CC), mRNA polyadenylation (BP), mRNA cleavage (BP), RNA 3'-end processing (BP), transcription termination (BP), mRNA cleavage factor complex (CC), metal ion binding (MF), termination of RNA polymerase II transcription, poly(A)-coupled (BP), termination of RNA polymerase II transcription, poly(A)-independent (BP)
2	CDC23, APC5, CDC16, APC2, APC1, DOC1, APC9, APC11, APC4, CDC27, CDC26, GLC7, TRA1	11(59)	11(7103)	4.85E-24	0.823	Cell cycle-yeast (KEGG), modification-dependent protein catabolic process (BP), ubiquitin-protein ligase activity (MF), sister chromatid segregation (BP), protein ubiquitination (BP), spindle elongation (BP), anaphase-promoting complex (CC), cyclin catabolic process (BP), anaphase-promoting complex-dependent (BP), metaphase/anaphase transition (BP), Ubiquitin mediated proteolysis (KEGG)
3	RPN13, RPN8, RPN1, RPT1, RPN3, RPN10, RPN11, RPN2, RPN5, RPT3, RPT6, RPN6, APC2, DOC1	14(59)	90(7103)	8.11E-15	0.609	Ubiquitin-dependent protein catabolic process (BP), proteasome complex (CC), Proteasome (KEGG), proteasome storage granule (CC), proteasome regulatory particle, lid subcomplex (CC), proteasome regulatory particle, base subcomplex (CC), endopeptidase activity (MF), enzyme regulator activity (MF)
4	LSM5, DCP1, PAP1, LSM3, LSM8, LSM6, LSM1, LSM4, LSM7, LSM2, PTA1, KEM1, PAT1, YSH1	14(59)	93(7103)	1.31E-14	0.750	RNA degradation (KEGG), tRNA processing (BP), U4/U6 x U5 tri-snRNP complex (CC), ribonucleoprotein complex (CC), nuclear mRNA splicing (BP), Like-Sm ribonucleoprotein (LSM) domain (IPR), Like-Sm ribonucleoprotein (LSM)-related domain (IPR), nuclear-transcribed mRNA catabolic process (BP), U6 snRNP (CC), Spliceosome (KEGG), small nucleolar ribonucleoprotein complex (CC), RNA splicing (BP), cytoplasmic mRNA processing body (CC), RNA catabolic process (BP)
5	NGG1, HFI1, TRA1, SPT20, SGF29, SPT7, SGF73, SPT8, SGF11, GCN5, SPT3, ADA2, RPN6, CFT2	14(59)	142(7103)	5.26E-12	0.570	SAGA complex (CC), chromatin modification (BP), histone acetylation (BP), SLIK (SAGA-like) complex (CC), RNA polymerase II transcription factor activity (MF), transcription cofactor activity (MF), positive regulation of transcription (BP), histone acetyltransferase activity (MF), Ada2/Gcn5/Ada3 transcription activator complex (CC), transcription factor TFIID complex (CC), DNA-directed RNA polymerase II (CC), transcription coactivator activity (MF), ER-nuclear signaling pathway (BP), transcription from RNA polymerase II promoter (BP)

Tabla 4.3: Resultado del análisis independiente con *GeneTerm Linker* de 59 proteínas nucleares de levadura

De forma más visual la figura 4.12 muestra una red funcional que ha sido construida utilizando sólo los resultados de ***GeneTerm Linker***. En esta red funcional los nodos corresponden a los genes y los enlaces representan que dos genes están en el mismo *Geneterm set* (enlace negro) o en el mismo metagrupo pero no en el mismo *Geneterm set* (enlace gris). En esta figura se aprecia mejor cada uno de los metagrupos obtenidos por el algoritmo, y las relaciones funcionales entre los genes que los constituyen.

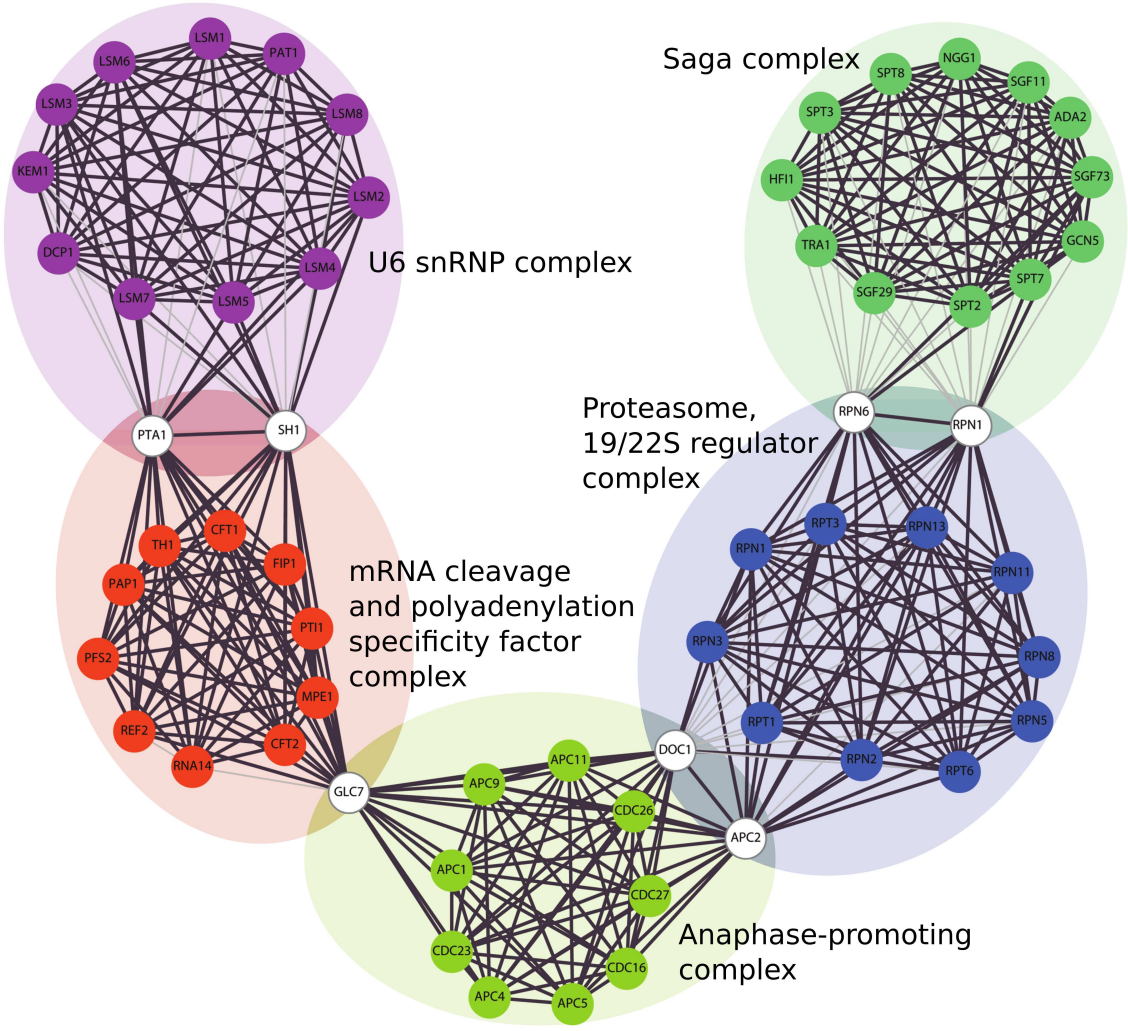


Figura 4.12: Red funcional derivada de datos de *GeneTerm Linker* para 59 proteínas de levadura - La figura muestra la red funcional obtenida a partir de los metagrupos encontrados por *GeneTerm Linker* a partir de 59 proteínas de levadura. Los metagrupos o módulos funcionales quedan bien definidos y se descubren 7 proteínas que hacen de puente entre varios de los módulos.

4.4.1 Comparación del método con otras aproximaciones de anotación funcional

Existen numerosas aproximaciones bioinformáticas y métodos para el análisis de enriquecimiento funcional (Huang et al., 2009a), (Khatri et al., 2012), sin embargo la búsqueda de módulos funcionales derivados de análisis de enriquecimiento ha sido poco explorada. La herramienta más utilizada para este tipo de análisis modular es *Functional Annotation Clustering*, que pertenece a la plataforma bioinformática DAVID (*DAVID Bioinformatics Resources*) (Huang et al., 2009b),

denominada DAVID FAC en adelante. Esta herramienta realiza el agrupamiento de las anotaciones enriquecidas en una lista de genes basándose en el grado de co-asociación entre los genes de la lista inicial de partida.

Hemos comparado el método ***GeneTerm Linker*** desarrollado en este trabajo con la herramienta **DAVID FAC** utilizando el mismo conjunto de 59 proteínas que constituyen los 5 complejos de levadura analizados en el apartado 4.4. Con ambas herramientas se seleccionaron los mismos espacios de anotación: KEGG, los tres espacios de GO e InterPro; con soporte mínimo de 4 genes. El resto de parámetros en la herramienta de DAVID FAC se mantuvieron con sus valores por defecto en una de las pruebas y en otra ejecución se ajustaron hasta obtener 5 módulos. Aunque esta última opción no se corresponde con un análisis estándar puesto que se conocen a priori el número de módulos a definir, su realización se consideró oportuna para obtener una comparación en el mejor de los supuestos posibles para DAVID FAC.

Para comparar los resultados obtenidos con ambas herramientas se construyó una matriz de co-ocurrencias de los genes/proteínas como se describe en (Halkidi et al., 2001). De esta manera para cada par de genes **g1** y **g2** existen cuatro posibilidades:

(a) Si **g1** y **g2** pertenecen al mismo complejo y son asociados en el mismo metagrupo o módulo obtenido se consideran verdaderos positivos (TP)
(b) Si **g1** y **g2** pertenecen al mismo complejo y no son incluidos en el mismo metagrupo o módulo obtenido se consideran falsos negativos (FN)
(c) Si **g1** y **g2** no pertenecen al mismo complejo y son incluidos en el mismo metagrupo o módulo obtenido se consideran falsos positivos (FP)
(d) Si **g1** y **g2** no pertenecen al mismo complejo ni están en el mismo metagrupo: se consideran verdaderos negativos (TN)

A partir de estos conceptos básicos que definen el grado de acuerdo entre lo esperado y lo observado, podemos definir los siguientes parámetros estadísticos para evaluar los métodos:

Exactitud o estadístico *Rand*:

$$Rand = (TP + TN)/(TP + TN + FP + FN) \tag{4.4}$$

Coeficiente de *Jaccard*:

$$J = TP/(TP + FP + FN) \tag{4.5}$$

En la tabla 4.4.1 se muestran los resultados de los análisis realizados con ***GeneTerm Linker*** y DAVID FAC. Como se puede observar *GeneTerm Linker* recupera mejor los módulos buscados consiguiendo medidas de exactitud y coeficiente de *Jaccard* mejores que DAVID FAC. Incluso ajustando los valores de los parámetros en DAVID FAC para obtener el número de grupos deseado se obtienen menos TP y más FP.

4.4.2 **Validación con conjuntos de datos más amplios y evaluación de la tolerancia al ruido**

Las técnicas ómicas de alto rendimiento generan gran cantidad de datos que permiten el análisis de estadios biológicos en situaciones experimentales específicas. Estas técnicas permiten la aproximación al problema analizado desde un punto de vista ómico o global, sin embargo sacrifican a cambio parte de su precisión. Debido a esto, las listas de genes obtenidas tras estos análisis a gran escala incluyen genes falsos positivos que aparentemente no guardan una relación funcional con

	GeneTerm Linker	**DAVID FAC (por defecto)**	**DAVID FAC (5 grupos)**
Grupos buscados	5	5	5
Grupos encontrados	5	15	5
Combinaciones de genes	1711	1711	1711
TP	320	320	254
FN	82	1179	132
FP	0	0	66
TN	1309	212	1259
Coeficiente de *Jaccard*	**0.769**	**0.213**	**0.562**
Exactitud *Rand*	**0.952**	**0.311**	**0.884**

Tabla 4.4: Análisis comparativo de resultados obtenidos con DAVID FAC y *GeneTerm Linker* utilizando el set de datos de 59 proteínas de levadura correspondientes a 5 complejos.

el resto. La existencia de estos falsos positivos ha motivado la evaluación de ***GeneTerm Linker*** para enfrentarse a estas situaciones de ruido.

Para demostrar estas capacidades hemos validado el método frente a tres series de datos biomoleculares sobre genes/proteínas que se han recolectado procedentes de 3 repositorios de diferente naturaleza biológica. Además estas series contienen datos de varias especies:

(a) **Complejos CORUM**: Complejos de proteínas identificados en mamíferos tomados de la base de datos CORUM (http://mips.helmholtz-muenchen.de/genre/proj/corum) (Ruepp et al., 2010).
(b) **Vías SGD**:Vías de señalización y vías metabólicas de levadura tomadas de la base de datos sobre rutas biomoleculares incluida en SGD (http://www.yeastgenome.org) (Engel et al., 2010).
(c) **Enfermedades OMIM**: Genes implicados en enfermedades humanas tomados de la base de datos de OMIM (http://www.ncbi.nlm.nih.gov/omim) (Hamosh et al., 2005).

De estas bases de datos se seleccionaron complejos/vías/enfermedades con al menos 8 genes anotados tomando de entre ellos 10 grupos aleatorios de cada repositorio, es decir, un total de 30 conjuntos de referencia.

Cada conjunto seleccionado fue analizado utilizando ***GeneTerm Linker*** evaluando cuántos de los genes asociados a cada complejo/vía/enfermedad están incluidos en el metagrupo más significativo recuperado. Para evaluar la tolerancia al ruido cada conjunto es analizado no sólo con los genes que lo componen sino añadiendo genes seleccionados al azar entre el conjunto total de cada repositorio. De esta manera se han introducido dos niveles de ruido: 20 % y 60 %; de manera que si por ejemplo un complejo analizado está formado por 10 genes, a la hora de introducir un 20 % de ruido se añaden otros 2 genes más tomados al azar de la base de datos de todos los complejos.

Los resultados utilizando ***GeneTerm Linker*** sobre los conjuntos de referencia se muestran en la tabla 4.4.2, que presenta en cada fila los datos del metagrupo más significativo encontrado por el algoritmo además de su coincidencia con el correspondiente conjunto de referencia. Como ejemplo, en el caso del primer complejo "*C complex spliceosome*" está compuesto por 80 genes a los que se ha añadido un 20 % de ruido extra conteniendo en total 96 para ser analizados. El metagrupo más significativo encontrado por GeneTerm Linker contiene 68 genes de los cuales los 68 pertenecen al complejo (es decir son verdaderos positivos, TP).

Siguiendo los mismos pasos se calcularon los resultados para los 30 conjuntos de referencia. Como los genes que pertenecen a cada conjunto de referencia son conocidos es posible calcular los

índices de precisión y exhaustividad (*precision and recall*) del método. En un escenario de recuperación de información se define precisión como el número de elementos relevantes obtenido respecto al total de resultados recuperados; es una medida de la exactitud y fidelidad. Por otro lado, el índice de exhaustividad (*recall*) se define como el número de elementos relevantes recuperados respecto al total de elementos relevantes (los que deberían haber sido recuperados); es por tanto un indicador de la completitud. Una medida más general se puede obtener combinando las dos anteriores con una media armónica de ambas que se denomina medida F (*F-measure* o $F_1 score$). Estadísticamente estos parámetros están relacionados con los errores de tipo I y tipo II y se definen del siguiente modo:

$$Precision = TP/(TP + FP) \tag{4.6}$$

$$Recall = TP/(TP + FN) \tag{4.7}$$

$$F_1 score = 2 * (Precision * Recall)/Precision + Recall) \tag{4.8}$$

Los resultados del análisis, mostrados en la tabla 4.4.2, indican que la precisión media tanto para complejos como vías (*pathways*) o enfermedades con un nivel de ruido del 20 % es cercana al 100 % (100 %, 97.80 %, 99.74 % respectivamente). El método es por tanto preciso y robusto frente a las perturbaciones, recuperando en cada metagrupo la mayoría de la señal funcional incluida en los conjuntos originales probados. Se observa también en esta tabla que el número de genes de los conjuntos de referencia no afecta a las tasas de error del método ya que la precisión y el *recall* no se ven asociados con dicho tamaño.

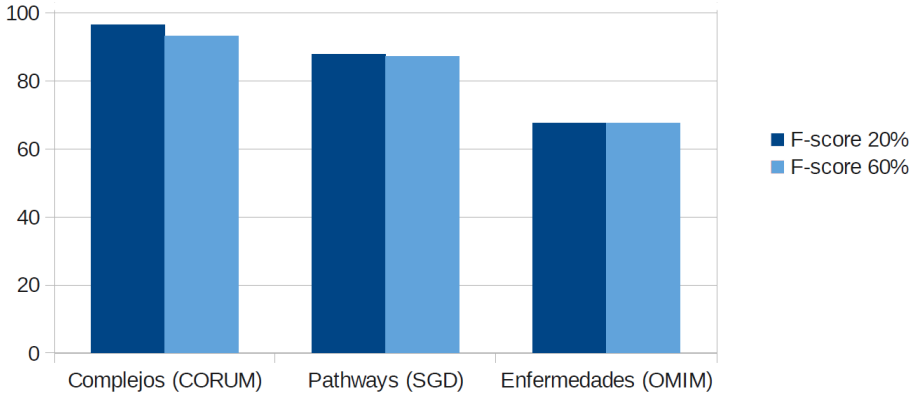


Figura 4.13: Comparación de F1scores - F1score medios para los 10 conjuntos analizados en las bases de datos CORUM, OMIM y SGD introduciendo un 20 % y un 60 % de ruido

En la figura 4.13 se muestra el $F_1 score$ medio para los 10 conjuntos de cada repositorio introduciendo porcentajes de ruido del 20 y 60 %. Como se puede observar, el hecho de añadir genes no relacionados con cada conjunto en el análisis no tiene una especial influencia a la hora de recuperar los grupos funcionales. Sin embargo, sí que se observan diferencias entre las bases de datos utilizadas pasando de valores de *recall* medio por encima del 90 % para los complejos (CORUM) a valores entorno al 65 % en el caso de las enfermedades (OMIM). Esta observación es coherente con las expectativas, ya que desde el punto de vista biológico molecular se ha de esperar un descenso de la coherencia y compactación funcional de las unidades cuando pasamos de CORUM a SGD y a OMIM. Resulta claro que la cohesión y la similitud funcional de genes asociados en un complejo multiproteína con relaciones estructurales y funcionales de tipo molecular es mucho mayor que la que cabría esperar en una vía metabólica en la que existe asociación funcional y no estructural. Y ésta a su vez es aún mayor que la cohesión funcional de los genes de una enfermedad donde la asociación entre los genes es muchas veces heurística, observacional o fenomenológica y no realmente asociada a una causa molecular conocida.

	Genes iniciales	Genes analizados	Genes metagrupo	TP	Precision (%)	Recall (%)	F1score (%)
Complejos (CORUM)							
C complex spliceosome	80	96	68	68	**100.00**	**85.00**	**91.89**
Mediator (transcriptional coactivator)	32	39	28	28	**100.00**	**87.50**	**93.33**
Proteasome (20S/26S)	22	27	22	22	**100.00**	**100.00**	**100.00**
RNA polymerase II (RNAPII)	26	32	24	24	**100.00**	**92.31**	**96.00**
F1F0-ATP synthase, mitochondrial	16	20	14	14	**100.00**	**87.50**	**93.33**
DAB, transcription preinitiation complex	16	20	16	16	**100.00**	**100.00**	**100.00**
Exosome	11	14	11	11	**100.00**	**100.00**	**100.00**
eIF3 complex	13	16	11	11	**100.00**	**84.62**	**91.67**
Nup 107-160 nuclear pore subcomplex	9	11	9	9	**100.00**	**100.00**	**100.00**
CENP-A NAC-CAD kinetochore complex	13	16	13	13	**100.00**	**100.00**	**100.00**
Vías/Pathways (SGD)							
Gluconeogenesis	22	27	22	22	**100.00**	**100.00**	**100.00**
TCA cycle, aerobic respiration	22	27	23	22	**95.65**	**100.00**	**97.78**
Sphingolipid metabolism	19	23	16	16	**100.00**	**84.21**	**91.43**
Biosynthesis of purine nucleotides	35	42	21	21	**100.00**	**60.00**	**75.00**
Lipid-linked oligosaccharide biosynthesis	12	15	12	12	**100.00**	**100.00**	**100.00**
Ergosterol biosynthesis	11	14	11	11	**100.00**	**100.00**	**100.00**
Superpathway of glucose fermentation	14	17	13	13	**100.00**	**92.86**	**96.30**
Fatty acid biosynthesis, initial steps	17	21	13	12	**92.31**	**70.59**	**80.00**
Inositol phosphate biosynthesis	19	23	11	11	**100.00**	**57.89**	**73.33**
Folate biosynthesis	18	22	10	9	**90.00**	**50.00**	**64.29**
Enfermedades (OMIM)							
Retinitis pigmentosa	51	62	39	38	**97.44**	**74.51**	**84.44**
Deafness	84	101	31	31	**100.00**	**36.90**	**53.91**
Cardiomyopathy	44	53	19	19	**100.00**	**43.18**	**60.32**
Epidermolysis bullosa	11	14	11	11	**100.00**	**100.00**	**100.00**
Congenital disorder of glycosylation	23	28	11	11	**100.00**	**47.83**	**64.71**
Muscular dystrophy	25	30	11	11	**100.00**	**44.00**	**61.11**
Glycogen storage disease	19	23	9	9	**100.00**	**47.37**	**64.29**
Leigh syndrome	8	10	7	7	**100.00**	**87.50**	**93.33**
Acute Leukemia	37	45	9	9	**100.00**	**24.32**	**39.13**
Diabetes mellitus	13	16	5	5	**100.00**	**38.46**	**55.56**

Tabla 4.5: Efectos de la introduccion del 20 % de ruido sobre la precision y el *recall*

4.4.3 Aplicación del método a conjuntos de datos experimentales

La validación de ***GeneTerm Linker*** en las secciones previas se ha realizado con conjuntos de datos seleccionados con resultados conocidos *a priori* que permiten evaluar el método y estimar errores y precisión. Sin embargo, para aceptar la utilidad mas práctica y general es necesaria su validación en escenarios típicos de análisis de enriquecimiento como pueden ser análisis de genes diferencialmente expresados o análisis de listas de genes derivadas de otros tipos de técnicas genómicas. Para demostrar la utilidad del método en ejemplos de datos experimentales reales se han seleccionado las firmas moleculares obtenidas en dos estudios de diferente naturaleza:

1. **Estudio de Alzheimer**: Análisis del desarrollo de ovillos neurofibrilares marcadores primarios de la enfermedad de Alzheimer (Dunckley et al., 2006). En este estudio se identifican 225 *Affymetrix probesets* como posibles genes implicados en el desarrollo neurofibrilar.
2. **Estudio de Cáncer**: Progresión del cáncer de mama a través del análisis de los genes que correlacionan con los diferentes estadíos patológicos de la enfermedad (Ma et al., 2003). En este estudio se identifican 174 genes marcadores para el pronóstico de la enfermedad.

Estas dos firmas moleculares fueron analizadas con GeneTerm Linker para encontrar los procesos biológicos subyacentes en cada estudio. Los resultados se muestran en las tablas 4.4.3 y 4.4.3 para los datos de Alzheimer y cáncer de mama respectivamente. En el caso del estudio de Alzheimer las funciones que aparecen más claramente están relacionadas con el transporte de iones de sodio y potasio, las vías de señalización de calcio, actividad post-sináptica, enfermedad de alzheimer y vías de señalización de Wnt y GnRH. Las funciones más sobre-representadas en la firma molecular de la progresión del cáncer de mama son la ubiquitinación y degradación de proteínas en el proteasoma, regulación de apoptosis, mitosis y meiosis con presencia de genes en los microtúbulos y cinetocoro y vías de señalización de EGF. Ambas firmas funcionales son coherentes con las características patológicas de cada enfermedad, es decir, con los procesos neurodegenerativos o con los procesos tumorales, demostrando que ***GeneTerm Linker*** es capaz de encontrar genes y funciones asociados de modo relevante y preciso.

Genes	Silueta	P-valor	Términos
CACNA1C, CACNA1D, HCN1, HCN3, KCNB1, KCNJ9, KCNMA1, KCNV1, KCTD2, SCN2A, SCN3A, SCN3B	0.55	4.50E-10	Ion transport (IPR), voltage-gated ion channel activity (MF), sodium ion transport (BP), potassium ion transport (BP), voltage-gated potassium channel activity (MF), voltage-gated potassium channel complex (CC)
ADD3, ATP2B1, CACNA1C, CACNA1D, CALM3, CAMK2D, CHRM1, GAP43, P2RX5, PLCB1, SLC3A2, SNTA1	0.42	1.01E-07	Calmodulin binding (MF), Calcium signaling pathway (KEGG), calcium ion transport (BP)
ANO3, CHRM1, GABBR2, GABRB3, GRIA3, MCHR1, P2RX5, TSPO	0.41	1.55E-03	Postsynaptic membrane (CC), Neuroactive ligand-receptor interaction KEGG), ion channel activity (MF)
APOE, CACNA1C, CACNA1D, CALM3, CAMK2D, FZD3, KCNMA1, MAPK8, MAPT, PLCB1, PPP1R14A, PRKG1, SORT1	0.30	4.33E-07	Calcium signaling pathway (KEGG), Long-term potentiation (KEGG), GnRH signaling pathway (KEGG), Alzheimer's disease (KEGG), Vascular smooth muscle contraction (KEGG), Melanogenesis (KEGG), Neurotrophin signaling pathway (KEGG), Wnt signaling pathway (KEGG)
CACNA1C, CHRM1, GABBR2, GABRB3, GAP43, GRIA3, HOMER1, KCNMA1, LZTS1, NRXN3, SEPT3, SNTA1, SYNGR1	0.26	2.34E-10	Postsynaptic membrane (CC), postsynaptic density (CC), synapse (CC), synaptic transmission (BP)
ACTN1, APOE, CACNA1C, CACNA1D, CALM3, CHRM1, CNTN4, CNTNAP1, CUX2, CXCL12, DDN, ECHDC1, FAT3, GAP43, HCN1, HHIP, ICAM2, KCNMA1, LRRFIP1, LZTS1, MAPK8, MAPT, MTSS1, MYO9A, NLGN4X, NPTXR, NRCAM, NRXN3, NTNG1, PAK7, PDCD6, PLCB1, PRKCI, RASGRP1, ROBO2, RUNX1T1, S100A16, SCN2A, SCN3A, SELPLG, SLC3A2, SORT1, STRC, SYNGAP1, THY1, TRPS1, VWF, WASF2	-0.05	3.61E-20	Cell adhesion molecules (KEGG), integrin binding (MF), response to calcium ion (BP), cell projection (CC), cell surface (CC), Concanavalin A-like lectin/glucanase, subgroup (IPR), Concanavalin A-like lectin/glucanase (IPR), apical plasma membrane (CC), external side of plasma membrane (CC), IQ calmodulin-binding region (IPR), MAPK signaling pathway (KEGG), SH3 domain binding (MF), enzyme binding (MF), dendrite (CC), axon (CC), catalytic activity (MF), Axon guidance (KEGG), Focal adhesion (KEGG), transcription repressor activity (MF), Regulation of actin cytoskeleton (KEGG)

Tabla 4.6: Resultados del análisis con *GeneTerm Linker* del set de datos experimental de Alzheimer

Genes	Silueta	P-valor	Términos
PLK1, PSMA7, PSMB2, PSMB3, PSMD12, UBE2C	0.73	2.04E-06	Regulation of ubiquitin-protein ligase activity during mitotic cell cycle (BP), anaphase-promoting complex-dependent proteasomal ubiquitin-dependent protein catabolic process (BP), proteasome complex (CC), Proteasome (KEGG)
IL2RA, ING4, IP6K2, MCM6, POLL, RRM2, TOP2A	0.60	4.81E-04	DNA replication (BP), positive regulation of apoptosis (BP)
AURKA, BUB1, KIF11, PBK, PLK1, SCYL2, UBE2C	0.52	7.75E-06	Mitosis (BP), protein kinase activity (MF)
BIRC5, CDCA8, CENPA, ESPL1, HJURP, NDC80, SKA2, SMC1A, TOP2A	0.32	1.80E-08	Chromosome segregation (BP), chromosome, centromeric region (CC), chromosome (CC)
ANLN, AURKA, BIRC5, CDC25B, CDCA8, CEP55, ESPL1, GABARAP, KIF11, NUSAP1, RACGAP1, SKA2	0.32	9.49E-12	Spindle microtubule (CC), mitosis (BP), cytokinesis (BP), spindle (CC), microtubule binding (MF)
AURKA, BUB1, ESPL1, HJURP, NDC80, NUSAP1, PGR, PLK1, SMC1A	0.25	1.52E-10	Oocyte meiosis (KEGG), condensed chromosome kinetochore (CC), mitotic sister chromatid segregation (BP), kinetochore (CC), Cell cycle (KEGG)
BUB1, CCNA2, CDC25B, ESPL1, MCM6, PGR, PLK1, SMC1A, UBE2C	0.22	5.79E-08	Mitosis (BP), Cell cycle (KEGG), Progesterone-mediated oocyte maturation (KEGG)
ANLN, AURKA, BIRC5, BUB1, CCNA2, CDC25B, CDCA3, CEP55, CKS2, KIF11, NDC80, PBK, PLK1, SKA2, TOP2A, UBE2C	0.08	2.15E-13	Mitosis (BP), spindle organization (BP), phosphoinositide-mediated signaling (BP)
BRE, CBX7, CENPA, DNALI1, HDAC7, ING4, KIF11, KIF5C, L3MBTL, RAD51, RAG2, SATB1, SFPQ, SMC1A, TOP2A, WDR5	-0.02	1.52E-09	DNA repair (BP), chromatin binding (MF), chromatin modification (BP), microtubule motor activity (MF)
ANLN, ARPC1A, B3GNT3, BCL2, CDKN3, CHST10, EFEMP2, ESPL1, FBLN2, FOXO4, FZD5, GABARAP, ING4, LRRC59, LTBP3, MARCKS, MASP2, NDFIP2, NUDT21, PLAGL1, PLK1, SLC27A1, SLC35B1, SPCS3, ZMPSTE24	-0.03	1.97E-12	Calcium ion binding (MF), EGF-like calcium-binding (IPR), EGF-like calcium-binding (IPR), EGF-type aspartate/asparagine hydroxylation site (IPR), EGF calcium-binding (IPR), microsome (CC), cell cycle arrest (BP), actin cytoskeleton (CC), centrosome (CC), Golgi membrane (CC)

Tabla 4.7: Resultados del análisis con *GeneTerm Linker* del set de datos experimental de cáncer de mama

El hecho de simplificar y agrupar los genes y los términos en metagrupos facilita en gran manera la interpretación de los resultados. En la tabla 4.4.3, que resume los resultados de estos análisis, se puede observar cómo el número de *Geneterm sets* obtenidos con la herramienta de enriquecimiento antes de aplicar ***GeneTerm Linker*** es demasiado alto para una interpretación inmediata de los resultados. El uso de ***GeneTerm Linker*** reduce significativamente (de 101 a 59 y de 103 a 68) el número de *Geneterm sets*, manteniendo todas las funciones encontradas inicialmente y estableciendo asociaciones entre los términos que, de otra manera, quedarían ocultas.

	Alzheimer Dunckley et al. (2006)	**Cáncer de mama Ma et al. (2003)**
Firma génica (publicada)	225 Affymetrix *probesets*	200 genes
Genes identificados	176 genes	174 genes
Grupos DAVID FAC	58	68
Metagrupos GeneTerm Linker	6	10
***GeneTerm sets* iniciales**	101	103
***GeneTerm sets* redundantes**	42	35
***GeneTerm sets* finales**	59	68

Tabla 4.8: Resumen de los resultados de *GeneTerm Linker* para dos conjuntos de datos experimentales

4.5 Implementación de *GeneTerm Linker* en un servidor web

GeneTerm Linker está diponible como herramienta web que permite a cualquier usuario acceder al método y utilizarlo de modo sencillo con sus listados de genes o proteínas problema. La aplicación bioinformática está disponible en **http://gtlinker.cnb.csic.es** y **http://cicblade.dep.usal.es:8000**.

En la figura 4.14 se presenta una captura de pantalla de la página de entrada a la herramienta. Internamente la aplicación está implementada en los lenguajes de programación R (http://www.r-project.org) y Ruby (http://ruby-lang.org) y está asociada a la aplicación de análisis de enriquecimiento funcional *GeneCodis* (http://genecodis.cnb.csic.es) a través de su *Web Service*. La asociación de *GeneTerm Linker* a una herramienta de enriquecimiento previa es necesaria porque, como se ha explicado en el desarrollo, el método realiza un análisis post-enriquecimiento de las relaciones genes-términos encontradas. En este sentido *GeneTerm Linker* podría acoplarse a cualquier salida de una herramienta de análisis de enriquecimiento. La asociación con *GeneCodis* se ha implementado para facilitar la utilización práctica de *GeneTerm Linker* y para permitir que los investigadores usuarios puedan partir de la lista cruda inicial de genes/proteínas problema sin necesidad de pasar por el paso intermedio de enriquecimiento. Finalmente, indicar que en el sitio *web* de *GeneTerm Linker* se incluye una ayuda para su uso con detalles respecto al método y el modo de ver e interpretar los resultados.

4.6 Discusión

La anotación funcional se ha considerado el cuello de botella de los estudios biomédicos desde la aparición de las técnicas masivas de producción de datos biomoleculares (Medrano-Soto et al., 2008; Llewellyn and Eisenberg, 2008). Para muchos de los genes estudiados mediante estas técnicas no se conocen o no se han anotado funciones específicas y otros han sido anotados únicamente en base a homología, asignando funciones conocidas a secuencias similares. Esta falta de conocimiento preciso, los errores introducidos muchas veces en las anotaciones por homología y la

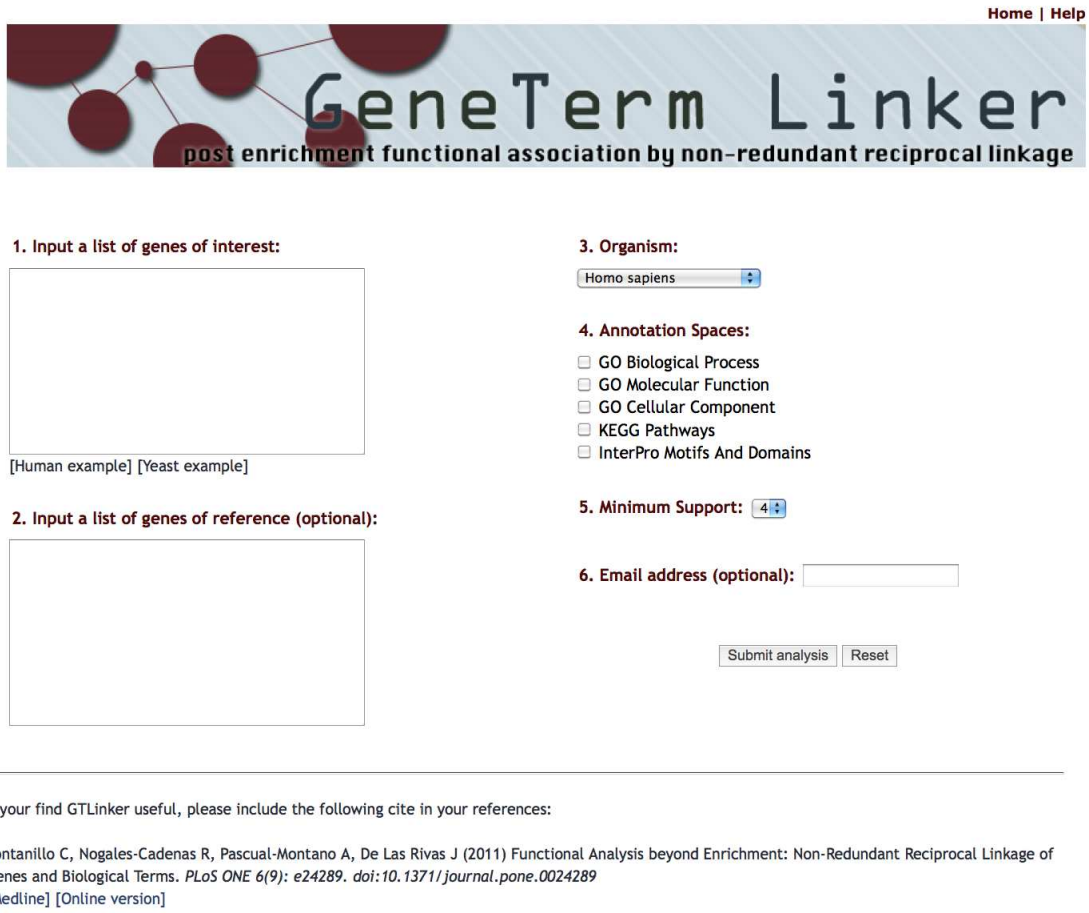


Figura 4.14: Aplicación web del método *GeneTerm Linker* - Captura de pantalla de la página de inicio de *GeneTerm Linker* en la que se introducen los datos básicos esenciales: listado de genes, lista de referencia (opcional), organismo, espacios de anotación que se quieren utilizar y soporte mínimo para construir los *Geneterm sets*.

utilización frecuente de términos “de moda” poco informativos son dificultades añadidas a las que las herramientas de enriquecimiento deben enfrentarse.

En este capítulo se ha propuesto un nuevo método bioinformático de análisis funcional denominado ***GeneTerm Linker*** que pretende solventar o minimizar los efectos descritos. El método ha sido desarrollado especialmente para la combinación de múltiples espacios de anotación con el objetivo de eliminar redundancias y reducir la complejidad de los resultados de anotación funcional automática.

La caracterización funcional de listas de genes/proteínas derivadas de técnicas ómicas debería proporcionar idealmente menos conjuntos de genes/proteínas anotados que el número de elementos que contiene la lista inicial (Merico et al., 2010). Sin embargo, las herramientas actuales no proporcionan resultados sencillos ni fácilmente interpretables. El método propuesto en este capítulo intenta minimizar este problema proporcionando un resultado simplificado y preciso, donde los genes y términos están agrupados en metagrupos evaluados por su significación, coherencia funcional y similitud.

Además la herramienta puede facilitar la inferencia de relaciones funcionales entre genes que no han sido anotadas específicamente en ninguna base de datos, pero que, sin embargo, aparecen implícitas cuando se analizan los genes que pertenecen a un mismo metagrupo.

En la figura 4.15 se muestra un esquema con los fundamentos que guiaron el desarrollo de la herramienta. La potencia del método radica en su capacidad para combinar múltiples fuentes de anotación sin depender de su estructura interna a la hora de proporcionar resultados simplificados. Las aproximaciones de la simplificación de los resultados de enriquecimiento mediante la búsqueda de asociaciones basadas en la estructura jerárquica de las bases de datos de anotación (sobre todo aplicadas en *Gene Ontology*) consiguen una mejora de los resultados, sin embargo no son extrapolables a otros espacios de anotación con diferente estructura y organización. ***GeneTerm Linker*** hace posible la integración de múltiples espacios incluyendo tanto GO como KEGG, InterPro, etc.

Una contribución complementaria de este estudio es el análisis comparativo de los espacios de anotación utilizados. En la sección 4.2 se muestra que KEGG es más estable y contiene menos términos atípicos. Estas diferencias son debidas fundamentalmente al “recurado” exhaustivo por expertos de KEGG y al hecho de que GO muchas veces anote términos muy generales para su aplicación en organismos muy diversos. Esta falta de especificidad y la utilización masiva de términos que se hacen demasiado populares (como *Signal transduction* o *Regulation of transcription*, influye en gran medida en la calidad de los resultados de enriquecimiento.

Como conclusión, es posible afirmar que se ha construido una herramienta innovadora cuyo valor radica en proporcionar una solución coherente y simplificada al problema de la anotación funcional, pero que al mismo tiempo permite ahondar en la naturaleza de los resultados para descubrir nuevas relaciones. Para lograr esto la herramienta construida va un paso más lejos que las herramientas de enriquecimiento actuales permitiendo la integración bases de datos heterogéneas y eliminando los problemas derivados de la utilización de términos redundantes o poco informativos. Para más información se puede consultar el artículo publicado sobre GeneTerm Linker titulado: ”*Functional analysis beyond enrichment: non-redundant reciprocal linkage of genes and biological terms*”, que se adjunta en esta memoria (Fontanillo et al., 2011).

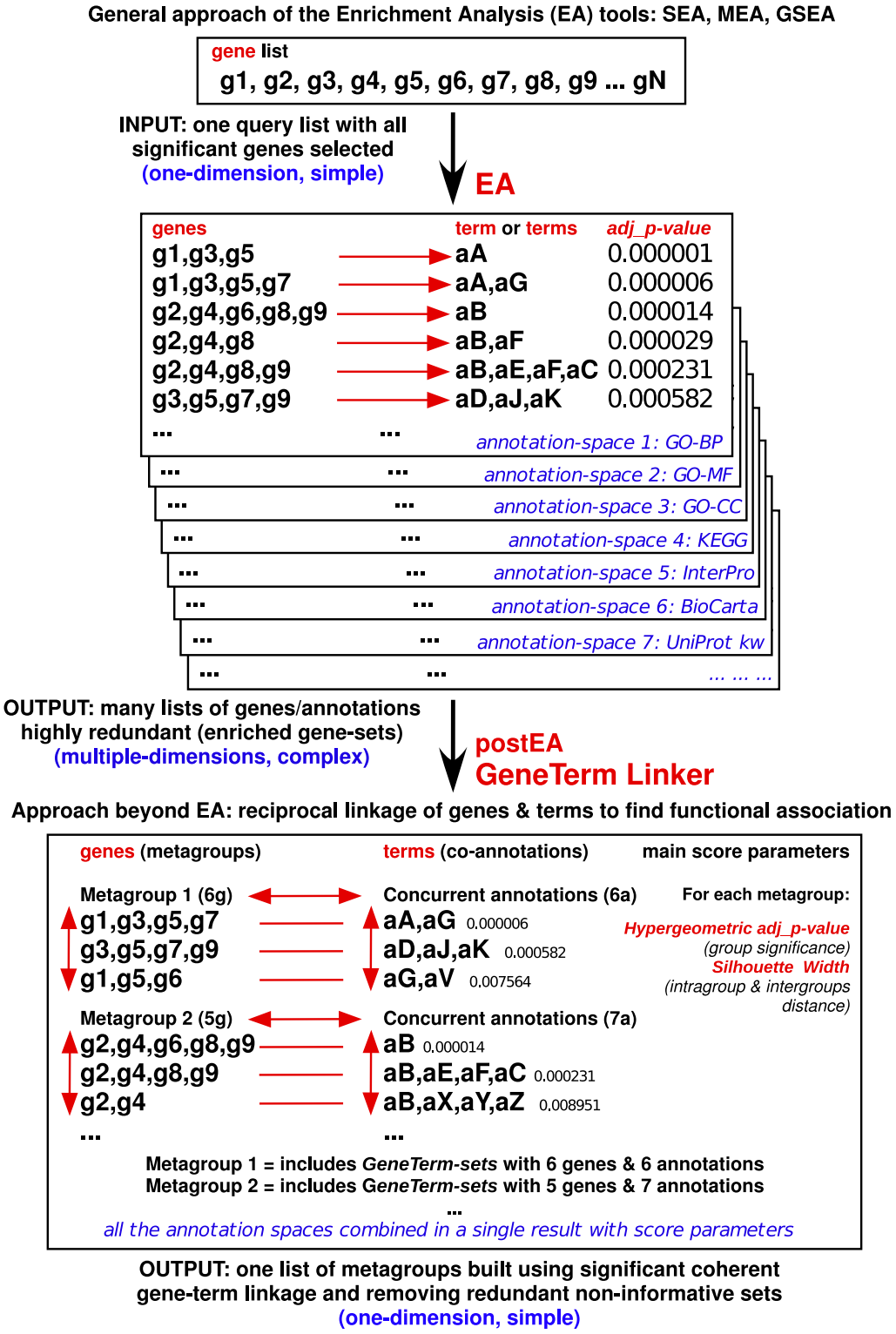


Figura 4.15: Esquema descriptivo del método de GeneTerm Linker - Esquema que describe el problema y el objetivo del método de GeneTerm Linker ilustrándolo como herramienta de post-enriquecimiento e indicando los parámetros en los que se apoya para construir los metagrupos de genes y co-anotaciones de múltiples términos biológicos.

Conclusiones generales

El objetivo fundamental de la genómica funcional es entender cómo funciona el genoma en su conjunto mediante el análisis de la expresión de todos y cada uno de sus genes y de los múltiples factores que regulan o influyen la expresión de los mismos. La recolección sistemática de información y datos procedentes de tecnologías genómicas experimentales globales a gran escala proporciona un punto de partida para desvelar la actividad del genoma y el comportamiento de los sistemas vivos asociado a su genoma. Se trata de expandir el alcance de la investigación biológica desde el estudio de genes individuales al estudio de todos los genes de una célula en un momento determinado. Desde esta perspectiva en la presente memoria de Tesis Doctoral se han abordado el desarrollo de distintos métodos y algoritmos bioinformáticos y su aplicación a series de datos de estudios en cáncer obtenidos por diversas técnicas genómicas. A continuación se hace un resumen de las cuatro partes del trabajo desarrollado y se proponen una conclusiones finales a modo de sumario.

En el **capítulo 1** se han utilizado perfiles de expresión derivados de datos de microarrays que permiten diferenciar tipos o subtipos de enfermedades así como identificar una firma molecular propia de cada uno de los estados. Se ha desarrollado el algoritmo *geNetClassifier* que, mediante el análisis de estos perfiles de expresión, proporciona un método de clasificación multiclase robusto, evaluado mediante validación cruzada anidada y centrado en el acceso transparente a las entidades biológicas. Este clasificador realiza una asignación probabilística de nuevas muestras a cada uno de los estados, de manera que permite reducir el número de falsos positivos y lograr una mayor semejanza con el proceso de decisión que llevaría a cabo un experto. Además de identificar los genes marcadores para cada enfermedad, el algoritmo analiza las relaciones entre dichos genes de manera que se facilita la creación de redes de genes asociados a cada subtipo patológico y la identificación de los procesos biológicos desregulados en cada estado en los que cooperan dichos genes marcadores.

Los procesos tumorales están dirigidos habitualmente por la acumulación de alteraciones en el DNA, normalmente con incrementos y disminuciones en el número de copias de DNA. En el **capítulo 2** se ha implementado un flujo de trabajo completo para el análisis de estas alteraciones. Los algoritmos diseñados se basan en la discretización de los valores cuantitativos continuos del número de copias de DNA obtenidos mediante arrays de SNPs en 3 estados: amplificación (o ganancia), deleción (o pérdida) y no cambio. A partir de esta discretización se han desarrollado dos algoritmos que buscan alteraciones recurrentes propias de determinadas patologías. El primer algoritmo identifica regiones mínimas con alteraciones comunes en un conjunto de muestras, que se corresponderán con las alteraciones germinales o más importantes implicadas en el desarrollo de las patologías, normalmente procesos tumorales. El segundo algoritmo identifica las regiones con puntos de ruptura, es decir, en las se producen frecuentemente los cambios en el estado de número de copias, normalmente asociadas con el desarrollo tumoral. Utilizando estos algoritmos

se han identificado regiones alteradas candidatas a marcadores moleculares de cáncer colorectal metastático. Además se ha descubierto un punto de ruptura en el cromosoma 17p11.2 relacionado con la supervivencia de los pacientes con este tipo de tumor metastático.

El análisis de las alteraciones genómicas del número de copias de DNA proporciona una información valiosa para comprender el origen y desarrollo tumoral, sin embargo, no todas las alteraciones definidas producen los mismos efectos sobre la expresión génica. La integración de estos datos junto con los perfiles de expresión génica facilita la identificación de genes conductores claves en el desarrollo y progresión tumoral. En el **capítulo 3** se ha presentado un método para la identificación de alteraciones en el número de copias de DNA asociados a cambios en la expresión génica. El método propuesto permite la integración de datos procedentes de microarrays de RNA y de DNA que tienen diferente resolución y reduce, gracias a la segmentación, los efectos de regulación no asociados a la localización genómica. De este modo, se consiguen identificar genes conductores en regiones genómicas candidatas que refinan los resultados obtenidos analizando por separado ambas capas de información.

Por último, en el **capítulo 4** se ha propuesto un método bioinformático de análisis de enriquecimiento funcional que permite la combinación de múltiples espacios de anotación con el objetivo de eliminar redundancias y reducir la complejidad de los resultados de anotación funcional automática. Se ha desarrollado una herramienta web con el método propuesto que facilita la interpretación de los resultados de enriquecimiento mediante el filtrado de términos generales identificados en el análisis de diferentes espacios de anotación y posibilita la inferencia de relaciones funcionales entre genes pobremente anotados.

De modo global, el trabajo descrito en esta memoria proporciona un conjunto de herramientas y algoritmos que permiten estudiar la asociación entre genotipo y fenotipos patológicos. Estudiando las alteraciones genómicas y los cambios en la expresión génica es posible comprender mejor las funciones y procesos que están teniendo lugar en las células y que, de alguna manera, están impulsando el desarrollo tumoral o patológico. El análisis de datos procedentes de técnicas genómicas como los microarrays de expresión y de SNPs permite identificar genes marcadores o causales y ahondar en los mecanismos que rodean la aparición y la progresión de enfermedades complejas como el cáncer. La profundización en el conocimiento de las enfermedades puede traducirse en tratamientos mejor dirigidos y más específicos, así como en un diagnóstico precoz que posibilite una mejor calidad de vida y una mejora de la supervivencia de los pacientes. Como aporte adicional las herramientas y algoritmos desarrollados en este trabajo son independientes de la tecnología utilizada para la cuantificación de las señales génicas y genómicas y, por ello, pueden ser fácilmente adaptables a otras técnicas experimentales en auge como las nuevas técnicas de secuenciación masiva simplemente adaptando el preprocesamiento de los datos.

Finalmente, como **CONCLUSIONES FINALES** resumidas del trabajo se puede decir:

1. La aplicación de métodos de aprendizaje automático (*machine learning*) transparentes basados en datos de expresión génica global permite construir sistemas expertos (i.e. clasificadores) capaces de diferenciar subtipos de enfermedades e identificar las entidades biomoleculares que los definen. Además, el análisis de los parámetros internos de estos sistemas expertos permite derivar nuevas características de las entidades biomoleculares, como es el poder discriminante de los genes.

2. La búsqueda sistemática a nivel genómico de alteraciones recurrentes en el número de copias de DNA en muestras de pacientes con tipos de cáncer específicos permite identificar regiones del genoma que incluyen genes cuya alteración es clave para el desarrollo de esos

procesos tumorales. Esta estrategia se ha demostrado eficaz un estudio de cáncer colorectal metastásico.

3. El desarrollo de métodos bioinformáticos integrativos de varios tipos de señales genómicas, como son los cambios de expresión génica y la alteración del numero de copias, se ha demostrado eficaz cuando se siguen estrategias coherentes y paralelas de procesado de ambos tipos datos basadas en la co-localización genómica de las señales.

4. La anotación biológica y análisis de enriquecimiento funcional de listas de genes marcadores derivadas de datos genómicos se ve optimizada cuando se exploran simultáneamente múltiples espacios de anotación, se evitan redundancias y se filtran términos poco informativos. Además, estas estrategias facilitan encontrar módulos funcionales con grupos de genes y términos asociados.

Bibliografía

Affymetrix, I. (2005). Guide to probe logarithmic intensity error (PLIER) estimation. Technical report. 13

Aguirre, A. J., Brennan, C., Bailey, G., Sinha, R., Feng, B., Leo, C., Zhang, Y., Zhang, J., Gans, J. D., Bardeesy, N., Cauwels, C., Cordon-Cardo, C., Redston, M. S., DePinho, R. A., and Chin, L. (2004). High-resolution characterization of the pancreatic adenocarcinoma genome. *Proc Natl Acad Sci U S A*, 101(24):9067–72. 38, 41, 52

Akavia, U. D., Litvin, O., Kim, J., Sanchez-Garcia, F., Kotliar, D., Causton, H. C., Pochanard, P., Mozes, E., Garraway, L. a., and Peér, D. (2010). An integrated approach to uncover drivers of cancer. *Cell*, 143(6):1005–17. 56

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, 25(1):25–9. 72

Attwood, T. K., Bradley, P., Flower, D. R., Gaulton, A., Maudling, N., Mitchell, A. L., Moulton, G., Nordle, A., Paine, K., Taylor, P., Uddin, A., and Zygouri, C. (2003). PRINTS and its automatic supplement, prePRINTS. *Nucleic Acids Res*, 31(1):400–2. 75

Babicka, L., Ransdorfova, S., Brezinova, J., Zemanova, Z., Sindelarova, L., Siskova, M., Maaloufova, J., Cermak, J., and Michalova, K. (2007). Analysis of complex chromosomal rearrangements in adult patients with MDS and AML by multicolor FISH. *Leuk Res*, 31(1):39–47. 53

Bader, G. D. and Hogue, C. W. (2003). An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*, 4(1):2. 84

Barbouti, A., Stankiewicz, P., Nusbaum, C., Cuomo, C., Cook, A., Höglund, M., Johansson, B., Hagemeijer, A., Park, S.-S., Mitelman, F., Lupski, J. R., and Fioretos, T. (2004). The breakpoint region of the most common isochromosome, i(17q), in human neoplasia is characterized by a complex genomic architecture with large, palindromic, low-copy repeats. *Am J Hum Genet*, 74(1):1–10. 53

Bell, D. A. and Wang, H. (2000). A Formalism for Relevance and Its Application in Feature Subset Selection. *Machine Learning*, 41(2):175–195. 10

Bengtsson, H., Irizarry, R., Carvalho, B., and Speed, T. P. (2008). Estimation and assessment of raw copy numbers at the single locus level. *Bioinformatics*, 24(6):759–67. 36

Bengtsson, H., Wirapati, P., and Speed, T. P. (2009). A single-array preprocessing method for estimating full-resolution raw copy numbers from all Affymetrix genotyping arrays including GenomeWideSNP 5 & 6. *Bioinformatics*, 25(17):2149–56. 31, 36, 52, 57

Bernard-Pierrot, I., Gruel, N., Stransky, N., Vincent-Salomon, A., Reyal, F., Raynal, V., Vallot, C., Pierron, G., Radvanyi, F., and Delattre, O. (2008). Characterization of the recurrent 8p11-12 amplicon identifies PPAPDC1B, a phosphatase protein, as a new therapeutic target in breast cancer. *Cancer Res*, 68(17):7165–75. 56

Beroukhim, R., Mermel, C. H., Porter, D., Wei, G., Raychaudhuri, S., Donovan, J., Barretina, J., Boehm, J. S., Dobson, J., Urashima, M., Mc Henry, K. T., Pinchback, R. M., Ligon, A. H., Cho, Y.-J., Haery, L., Greulich, H., Reich, M., Winckler, W., Lawrence, M. S., Weir, B. A., Tanaka, K. E., Chiang, D. Y., Bass, A. J., Loo, A., Hoffman, C., Prensner, J., Liefeld, T., Gao, Q., Yecies, D., Signoretti, S., Maher, E., Kaye, F. J., Sasaki, H., Tepper, J. E., Fletcher, J. A., Tabernero, J., Baselga, J., Tsao, M.-S., Demichelis, F., Rubin, M. A., Janne, P. A., Daly, M. J., Nucera, C., Levine, R. L., Ebert, B. L., Gabriel, S., Rustgi, A. K., Antonescu, C. R., Ladanyi, M., Letai, A., Garraway, L. A., Loda, M., Beer, D. G., True, L. D., Okamoto, A., Pomeroy, S. L., Singer, S., Golub, T. R., Lander, E. S., Getz, G., Sellers, W. R., and Meyerson, M. (2010). The landscape of somatic copy-number alteration across human cancers. *Nature*, 463(7283):899–905. 52, 56

Biegel, J. A. (1997). Genetics of pediatric central nervous system tumors. *J Pediatr Hematol Oncol*, 19(6):492–501. 53

Bignell, G. R., Greenman, C. D., Davies, H., Butler, A. P., Edkins, S., Andrews, J. M., Buck, G., Chen, L., Beare, D., Latimer, C., Widaa, S., Hinton, J., Fahey, C., Fu, B., Swamy, S., Dalgliesh, G. L., Teh, B. T., Deloukas, P., Yang, F., Campbell, P. J., Futreal, P. A., and Stratton, M. R. (2010). Signatures of mutation and selection in the cancer genome. *Nature*, 463(7283):893–8. 56

Brown, M. P., Grundy, W.Ñ., Lin, D., Cristianini, N., Sugnet, C. W., Furey, T. S., Ares, M., and Haussler, D. (2000). Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc Natl Acad Sci U S A*, 97(1):262–7. 9

Bru, C., Courcelle, E., Carrère, S., Beausse, Y., Dalmar, S., and Kahn, D. (2005). The ProDom database of protein domain families: more emphasis on 3D. *Nucleic Acids Res*, 33(Database issue):D212–5. 75

Bungaro, S., DellÓrto, M. C., Zangrando, A., Basso, D., Gorletta, T., Lo Nigro, L., Leszl, A., Young, B. D., Basso, G., Bicciato, S., Biondi, A., te Kronnie, G., and Cazzaniga, G. (2009). Integration of genomic and gene expression data of childhood ALL without known aberrations identifies subgroups with specific genetic hallmarks. *Genes Chromosomes Cancer*, 48(1):22–38. 60

Carvalho, C. M. B. and Lupski, J. R. (2008). Copy number variation at the breakpoint region of isochromosome 17q. *Genome Res*, 18(11):1724–32. 53

Chernova, O. B., Somerville, R. P., and Cowell, J. K. (1998). A novel gene, LGI1, from 10q24 is rearranged and downregulated in malignant brain tumors. *Oncogene*, 17(22):2873–81. 69

Consortium, G. O. (2010). The Gene Ontology in 2010: extensions and refinements. *Nucleic Acids Res*, 38(Database issue):D331–5. 72

De Angelis, P. M., Clausen, O. P., Schjø lberg, A., and Stokke, T. (1999). Chromosomal gains and losses in primary colorectal carcinomas detected by CGH and their associations with tumour DNA ploidy, genotypes and phenotypes. *Br J Cancer*, 80(3-4):526–35. 53

de Tayrac, M., Etcheverry, A., Aubry, M., Saïkali, S., Hamlat, A., Quillien, V., Le Treut, A., Galibert, M.-D., and Mosser, J. (2009). Integrative genome-wide analysis reveals a robust genomic glioblastoma signature associated with copy number driving changes in gene expression. *Genes Chromosomes Cancer*, 48(1):55–68. 66, 70

Diep, C. B., Kleivi, K., Ribeiro, F. R., Teixeira, M. R., Lindgjaerde, O. C., and Lothe, R. A. (2006). The order of genetic events associated with colorectal cancer progression inferred from meta-analysis of copy number changes. *Genes Chromosomes Cancer*, 45(1):31–41. 53

Diep, C. B., Parada, L. A., Teixeira, M. R., Eknaes, M., Nesland, J. M., Johansson, B., and Lothe, R. A. (2003). Genetic profiling of colorectal cancer liver metastases by combined comparative genomic hybridization and G-banding analysis. *Genes Chromosomes Cancer*, 36(2):189–97. 53

Ding, C. and Peng, H. (2005). Minimum redundancy feature selection from microarray gene expression data. *J Bioinform Comput Biol*, 3(2):185–205. 21, 28

Dunckley, T., Beach, T. G., Ramsey, K. E., Grover, A., Mastroeni, D., Walker, D. G., LaFleur, B. J., Coon, K. D., Brown, K. M., Caselli, R., Kukull, W., Higdon, R., McKeel, D., Morris, J. C., Hulette, C., Schmechel, D., Reiman, E. M., Rogers, J., and Stephan, D. a. (2006). Gene expression correlates of neurofibrillary tangles in Alzheimer's disease. *Neurobiol Aging*, 27(10):1359–71. 93

Ein-Dor, L., Kela, I., Getz, G., Givol, D., and Domany, E. (2005). Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics*, 21(2):171–8. 56

Engel, S. R., Balakrishnan, R., Binkley, G., Christie, K. R., Costanzo, M. C., Dwight, S. S., Fisk, D. G., Hirschman, J. E., Hitz, B. C., Hong, E. L., Krieger, C. J., Livstone, M. S., Miyasato, S. R., Nash, R., Oughtred, R., Park, J., Skrzypek, M. S., Weng, S., Wong, E. D., Dolinski, K., Botstein, D., and Cherry, J. M. (2010). Saccharomyces Genome Database provides mutant phenotype data. *Nucleic Acids Res*, 38(Database issue):D433–6. 90

Fan, C., Oh, D. S., Wessels, L., Weigelt, B., Nuyten, D. S. A., Nobel, A. B., van't Veer, L. J., and Perou, C. M. (2006). Concordance among gene-expression-based predictors for breast cancer. *New Engl J Med*, 355(6):560–9. 56

Feuk, L., Carson, A. R., and Scherer, S. W. (2006). Structural variation in the human genome. *Nat Rev Genet*, 7(2):85–97. 33

Finn, R. D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J. E., Gavin, O. L., Gunasekaran, P., Ceric, G., Forslund, K., Holm, L., Sonnhammer, E. L. L., Eddy, S. R., and Bateman, A. (2010). The Pfam protein families database. *Nucleic Acids Res*, 38(Database issue):D211–22. 75

Fontanillo, C., Aibar, S., Sanchez-Santos, J. M., and De Las Rivas, J. (2012). Combined analysis of genome-wide expression and copy number profiles to identify key altered genomic regions in cancer. *BMC Genomics*, 13(Suppl 5):S5. 57

Fontanillo, C., Nogales-Cadenas, R., Pascual-Montano, A., and De Las Rivas, J. (2011). Functional Analysis beyond Enrichment: Non-Redundant Reciprocal Linkage of Genes and Biological Terms. *PLoS One*, 6(9):e24289. 86, 98

Fridlyand, J., Snijders, A. M., Pinkel, D., Albertson, D. G., and Jain, A.Ñ. (2004). Hidden Markov models approach to the analysis of array CGH data. *J Multivar Anal*, 90(1):132–153. 37

Fridlyand, J., Snijders, A. M., Ylstra, B., Li, H., Olshen, A., Segraves, R., Dairkee, S., Tokuyasu, T., Ljung, B. M., Jain, A.Ñ., McLennan, J., Ziegler, J., Chin, K., Devries, S., Feiler, H., Gray, J. W., Waldman, F., Pinkel, D., and Albertson, D. G. (2006). Breast tumor copy number aberration phenotypes and genomic instability. *BMC Cancer*, 6:96. 38

Furey, T. S., Cristianini, N., Duffy, N., Bednarski, D. W., Schummer, M., and Haussler, D. (2000). Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, 16(10):906–14. 9

Furnari, F. B., Fenton, T., Bachoo, R. M., Mukasa, A., Stommel, J. M., Stegh, A., Hahn, W. C., Ligon, K. L., Louis, D.Ñ., Brennan, C., Chin, L., DePinho, R. A., and Cavenee, W. K. (2007). Malignant astrocytic glioma: genetics, biology, and paths to treatment. *Genes Dev*, 21(21):2683–710. 63

Gentleman, R., Carey, V., Huber, W., Irizarry, R., and Dudoit, S. (2005). *Bioinformatics and Computational Biology Solutions Using R and Bioconductor (Statistics for Biology and Health)*. Springer-Verlag, 1 edition. 8

González-González, M., Muñoz Bellvis, L., Mackintosh, C., Fontanillo, C., Gutiérrez, M. L., Abad, M. M., Bengoechea, O., Teodosio, C., Fonseca, E., Fuentes, M., De Las Rivas, J., Orfao, A., and Sayagués, J. M. (2012). Prognostic Impact of del(17p) and del(22q) as Assessed by Interphase FISH in Sporadic Colorectal Carcinomas. *PLoS One*, 7(8):e42683. 50, 53

Gupta, G. K., Strehl, A., and Ghosh, J. (1999). Distance based clustering of association rules. In *Intelligent Engineering Systems Through Artificial Neural Networks (ANNIE)*, pages 759–764. ASME Press. 80

Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). Gene Selection for Cancer Classification using Support Vector Machines. *Machine Learning*, 46(1):389–422. 10

Haab, B. B., Dunham, M. J., and Brown, P. O. (2001). Protein microarrays for highly parallel detection and quantitation of specific proteins and antibodies in complex solutions. *Genome Biol*, 2(2). 55

Haft, D. H., Selengut, J. D., and White, O. (2003). The TIGRFAMs database of protein families. *Nucleic Acids Res*, 31(1):371–3. 75

Halkidi, M., Batistakis, Y., and Michalis, V. (2001). On Clustering Validation Techniques. *J Intell Inf Syst*, 17(2):107–145. 89

Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A., and McKusick, V. A. (2005). Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res*, 33(Database issue):D514–7. 90

Hehir-Kwa, J. Y., Egmont-Petersen, M., Janssen, I. M., Smeets, D., van Kessel, A. G., and Veltman, J. A. (2007). Genome-wide copy number profiling on high-density bacterial artificial chromosomes, single-nucleotide polymorphisms, and oligonucleotide microarrays: a platform comparison based on statistical power analysis. *DNA Res*, 14(1):1–11. 34, 36

Hernandez-Toro, J., Prieto, C., and De las Rivas, J. (2007). APID2NET: unified interactome graphic analyzer. *Bioinformatics*, 23(18):2495–7. 84

Höglund, M., Gisselsson, D., Hansen, G. B., Säll, T., Mitelman, F., and Nilbert, M. (2002). Dissecting karyotypic patterns in colorectal tumors: two distinct but overlapping pathways in the adenoma-carcinoma transition. *Cancer Res*, 62(20):5939–46. 53

Huang, D. W., Sherman, B. T., and Lempicki, R. a. (2009a). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res*, 37(1):1–13. 76, 88

Huang, D. W., Sherman, B. T., and Lempicki, R. a. (2009b). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc*, 4(1):44–57. 88

Hulo, N., Bairoch, A., Bulliard, V., Cerutti, L., Cuche, B. A., de Castro, E., Lachaize, C., Langendijk-Genevaux, P. S., and Sigrist, C. J. A. (2008). The 20 years of PROSITE. *Nucleic Acids Res*, 36(Database issue):D245–9. 75

Hunter, S., Jones, P., Mitchell, A., Apweiler, R., Attwood, T. K., Bateman, A., Bernard, T., Binns, D., Bork, P., Burge, S., de Castro, E., Coggill, P., Corbett, M., Das, U., Daugherty, L., Duquenne, L., Finn, R. D., Fraser, M., Gough, J., Haft, D., Hulo, N., Kahn, D., Kelly, E., Letunic, I., Lonsdale, D., Lopez, R., Madera, M., Maslen, J., McAnulla, C., McDowall, J., McMenamin, C., Mi, H., Mutowo-Muellenet, P., Mulder, N., Natale, D., Orengo, C., Pesseat, S., Punta, M., Quinn, A. F., Rivoire, C., Sangrador-Vegas, A., Selengut, J. D., Sigrist, C. J. A., Scheremetjew, M., Tate, J., Thimmajanarthanan, M., Thomas, P. D., Wu, C. H., Yeats, C., and Yong, S.-Y. (2012). InterPro in 2011: new developments in the family and domain prediction database. *Nucleic Acids Res*, 40(Database issue):D306–12. 75

Hupé, P., Stransky, N., Thiery, J.-P., Radvanyi, F., and Barillot, E. (2004). Analysis of array CGH data: from signal ratio to gain and loss of DNA regions. *Bioinformatics*, 20(18):3413–22. 37

Inza, I.ñ., Larrañaga, P., Blanco, R., and Cerrolaza, A. J. (2004). Filter versus wrapper gene selection approaches in DNA microarray domains. *Artif Intell Med*, 31(2):91–103. 11

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(4):e15. 7, 13, 57

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(2):249–64. 13

Irizarry, R. A., Wu, Z., and Jaffee, H. A. (2006). Comparison of Affymetrix GeneChip expression measures. *Bioinformatics*, 22(7):789–94. 13

Jimenez-Valverde, A. and Lobo, J. M. (2007). Threshold criteria for conversion of probability of species presence to either–or presence–absence. *Acta Oecologica*, 31:361–369. 45

Jones, P. A. and Baylin, S. B. (2007). The epigenomics of cancer. *Cell*, 128(4):683–92. 70

Kallioniemi, A. (2008). CGH microarrays and cancer. *Curr Opin Biotechnol*, 19(1):36–40. 32

Kanehisa, M. and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*, 28(1):27–30. 73

Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M., and Hirakawa, M. (2010). KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res*, 38(Database issue):D355–60. 73, 74

Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., and Tanabe, M. (2012). KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res*, 40(Database issue):D109–14. 73

Kendziorski, C. M., Newton, M. a., Lan, H., and Gould, M.Ñ. (2003). On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles. *Stat Med*, 22(24):3899–914. 11

Keyvanrad, M. and HomayounpourM.M (2011). Automatic Gender Identification Using Fusion of Generative and Discriminative Classifiers and Clustering of Spekaers from the Same Gender. *Computer Science and Engineering*, 1(1):22–25. 9

Khatri, P., Sirota, M., and Butte, A. J. (2012). Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges. *PLoS Comput Biol*, 8(2):e1002375. 71, 88

Kim, Y.-A., Wuchty, S., and Przytycka, T. M. (2011). Identifying causal genes and dysregulated pathways in complex diseases. *PLoS Comput Biol*, 7(3):e1001095. 56

Kolomietz, E., Al-Maghrabi, J., Brennan, S., Karaskova, J., Minkin, S., Lipton, J., and Squire, J. A. (2001). Primary chromosomal rearrangements of leukemia are frequently accompanied by extensive submicroscopic deletions and may lead to altered prognosis. *Blood*, 97(11):3581–8. 42

Kotliarov, Y., Kotliarova, S., Charong, N., Li, A., Walling, J., Aquilanti, E., Ahn, S., Steed, M. E., Su, Q., Center, A., Zenklusen, J. C., and Fine, H. a. (2009). Correlation analysis between single-nucleotide polymorphism and expression arrays in gliomas identifies potentially relevant target genes. *Cancer Res*, 69(4):1596–603. 56, 60

Kotliarov, Y., Steed, M. E., Christopher, N., Walling, J., Su, Q., Center, A., Heiss, J., Rosenblum, M., Mikkelsen, T., Zenklusen, J. C., and Fine, H. a. (2006). High-resolution global genomic survey of 178 gliomas reveals novel regions of copy number alteration and allelic imbalances. *Cancer Res*, 66(19):9428–36. 60, 63

Kressel, U. H.-G. (1999). Advances in kernel methods. chapter Pairwise c, pages 255–268. MIT Press, Cambridge, MA, USA. 10

Lai, W. R., Johnson, M. D., Kucherlapati, R., and Park, P. J. (2005). Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. *Bioinformatics*, 21(19):3763–70. 37

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton,

L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J. F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R. A., Muzny, D. M., Scherer, S. E., Bouck, J. B., Sodergren, E. J., Worley, K. C., Rives, C. M., Gorrell, J. H., Metzker, M. L., Naylor, S. L., Kucherlapati, R. S., Nelson, D. L., Weinstock, G. M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Smith, D. R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H. M., Dubois, J., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R. W., Federspiel, N. A., Abola, A. P., Proctor, M. J., Myers, R. M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D. R., Olson, M. V., Kaul, R., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G. A., Athanasiou, M., Schultz, R., Roe, B. A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W. R., de la Bastide, M., Dedhia, N., Blöcker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J. A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D. G., Burge, C. B., Cerutti, L., Chen, H. C., Church, D., Clamp, M., Copley, R. R., Doerks, T., Eddy, S. R., Eichler, E. E., Furey, T. S., Galagan, J., Gilbert, J. G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L. S., Jones, T. A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W. J., Kitts, P., Koonin, E. V., Korf, I., Kulp, D., Lancet, D., Lowe, T. M., McLysaght, A., Mikkelsen, T., Moran, J. V., Mulder, N., Pollara, V. J., Ponting, C. P., Schuler, G., Schultz, J., Slater, G., Smit, A. F., Stupka, E., Szustakowski, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y. I., Wolfe, K. H., Yang, S. P., Yeh, R. F., Collins, F., Guyer, M. S., Peterson, J., Felsenfeld, A., Wetterstrand, K. A., Patrinos, A., Morgan, M. J., de Jong, P., Catanese, J. J., Osoegawa, K., Shizuya, H., Choi, S., Chen, Y. J., and Szustakowki, J. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921. 1

Lees, J., Yeats, C., Perkins, J., Sillitoe, I., Rentzsch, R., Dessailly, B. H., and Orengo, C. (2012). Gene3D: a domain-based resource for comparative genomics, functional annotation and protein network analysis. *Nucleic Acids Res*, 40(Database issue):D465–71. 75

Lent, B., Swami, A., and Widom, J. (1997). Clustering Association Rules. In *International Conference on Data Engineering*, pages 1–25, Birmingham U.K. 80

Letunic, I., Doerks, T., and Bork, P. (2012). SMART 7: recent updates to the protein domain annotation resource. *Nucleic Acids Res*, 40(Database issue):D302–5. 75

Li, C. and Wong, W. H. (2001). Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc Natl Acad Sci U S A*, 98(1):31–6. 13, 36

Liu, B., Hsu, W., and Ma, Y. (1999). Pruning and Summarizing the Discovered Associations. In *Proceedings of the International Conference on Knowledge Discovery and Data Mining*, San Diego, California, USA. ACM. 80

Liu, L., Hawkins, D. M., Ghosh, S., and Young, S. S. (2003). Robust singular value decomposition analysis of microarray data. *Proc Natl Acad Sci U S A*, 100(23):13167–72. 13

Liu, Q., Sung, A. H., Chen, Z., Liu, J., Chen, L., Qiao, M., Wang, Z., Huang, X., and Deng, Y. (2011). Gene selection and classification for cancer microarray data based on machine learning and similarity measures. *BMC Genomics*, 12 Suppl 5:S1. 21

Liu, Q., Sung, A. H., Chen, Z., Liu, J., Huang, X., and Deng, Y. (2009). Feature selection and classification of MAQC-II breast cancer and multiple myeloma microarray gene expression data. *PloS One*, 4(12):e8250. 21

Llewellyn, R. and Eisenberg, D. S. (2008). Annotating proteins with generalized functional linkages. *Proc Natl Acad Sci U S A*, 105(46):17700–5. 96

Lloyd, S. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137. 39

Lo, K. and Gottardo, R. (2007). Flexible empirical Bayes models for differential gene expression. *Bioinformatics*, 23(3):328–35. 11, 15

Ma, X.-J., Salunga, R., Tuggle, J. T., Gaudet, J., Enright, E., McQuary, P., Payette, T., Pistone, M., Stecker, K., Zhang, B. M., Zhou, Y.-X., Varnholt, H., Smith, B., Gadd, M., Chatfield, E., Kessler, J., Baer, T. M., Erlander, M. G., and Sgroi, D. C. (2003). Gene expression profiles of human breast cancer progression. *Proc Natl Acad Sci U S A*, 100(10):5974–9. 93

Malone, J. H. and Oliver, B. (2011). Microarrays, deep sequencing and the true measure of the transcriptome. *BMC Biol*, 9:34. 6

Medrano-Soto, A., Pal, D., and Eisenberg, D. (2008). Inferring molecular function: contributions from functional linkages. *Trends Genet*, 24(12):587–90. 96

Merico, D., Isserlin, R., Stueker, O., Emili, A., and Bader, G. D. (2010). Enrichment map: a network-based method for gene-set enrichment visualization and interpretation. *PloS One*, 5(11):e13984. 98

Mi, H., Dong, Q., Muruganujan, A., Gaudet, P., Lewis, S., and Thomas, P. D. (2010). PANTHER version 7: improved phylogenetic trees, orthologs and collaboration with the Gene Ontology Consortium. *Nucleic Acids Res*, 38(Database issue):D204–10. 75

Millenaar, F. F., Okyere, J., May, S. T., van Zanten, M., Voesenek, L. A. C. J., and Peeters, A. J. M. (2006). How to decide? Different methods of calculating gene expression from short oligonucleotide array data will give different results. *BMC Bioinformatics*, 7:137. 13

Moskow, J. J., Bullrich, F., Huebner, K., Daar, I. O., and Buchberg, A. M. (1995). Meis1, a PBX1-related homeobox gene involved in myeloid leukemia in BXH-2 mice. *Mol Cell Biol*, 15(10):5434–43. 26

Muñoz Bellvis, L., Fontanillo, C., González-González, M., Garcia, E., Iglesias, M., Esteban, C., Gutierrez, M. L., Abad, M. M., Bengoechea, O., De Las Rivas, J., Orfao, A., and Sayagués, J. M. (2012). Unique genetic profile of sporadic colorectal cancer liver metastasis versus primary tumors as defined by high-density single-nucleotide polymorphism arrays. *Mod Pathol*, 25(4):590–601. 45, 48, 53

Nannya, Y., Sanada, M., Nakazaki, K., Hosoya, N., Wang, L., Hangaishi, A., Kurokawa, M., Chiba, S., Bailey, D. K., Kennedy, G. C., and Ogawa, S. (2005). A robust algorithm for copy number detection using high-density oligonucleotide single nucleotide polymorphism genotyping arrays. *Cancer Res*, 65(14):6071–9. 36

Natrajan, R., Lambros, M. B., Rodríguez-Pinilla, S. M., Moreno-Bueno, G., Tan, D. S. P., Marchió, C., Vatcheva, R., Rayter, S., Mahler-Araujo, B., Fulford, L. G., Hungermann, D., Mackay, A., Grigoriadis, A., Fenwick, K., Tamber, N., Hardisson, D., Tutt, A., Palacios, J., Lord, C. J.,

Buerger, H., Ashworth, A., and Reis-Filho, J. S. (2009). Tiling path genomic profiling of grade 3 invasive ductal breast cancers. *Clin Cancer Res*, 15(8):2711–22. 56

Newton, M. A., Kendziorski, C. M., Richmond, C. S., Blattner, F. R., and Tsui, K. W. (2001). On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *J Comput Biol*, 8(1):37–52. 11

Nikolskaya, A.Ñ., Arighi, C.Ñ., Huang, H., Barker, W. C., and Wu, C. H. (2006). PIRSF family classification system for protein functional and evolutionary analysis. *Evol Bioinform*, 2:197–209. 75

Nogales-Cadenas, R., Carmona-Saez, P., Vazquez, M., Vicente, C., Yang, X., Tirado, F., Carazo, J. M., and Pascual-Montano, A. (2009). GeneCodis: interpreting gene lists through enrichment analysis and integration of diverse biological information. *Nucleic Acids Res*, 37(Web Server issue):W317–22. 86

Ohgaki, H., Dessen, P., Jourde, B., Horstmann, S., Nishikawa, T., Di Patre, P.-L., Burkhard, C., Schüler, D., Probst-Hensch, N. M., Maiorka, P. C., Baeza, N., Pisani, P., Yonekawa, Y., Yasargil, M. G., Lütolf, U. M., and Kleihues, P. (2004). Genetic pathways to glioblastoma: a population-based study. *Cancer Res*, 64(19):6892–9. 66

Olshen, A. B., Venkatraman, E. S., Lucito, R., and Wigler, M. (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, 5(4):557–72. 37, 39

Ortiz-Estevez, M., De Las Rivas, J., Fontanillo, C., and Rubio, A. (2011). Segmentation of genomic and transcriptomic microarrays data reveals major correlation between DNA copy number aberrations and gene-loci expression. *Genomics*, 97(2):86–93. 57, 58, 66

Picard, F., Robin, S., Lavielle, M., Vaisse, C., and Daudin, J.-J. (2005). A statistical approach for array CGH data analysis. *BMC Bioinformatics*, 6:27. 37

Plasse, M., Niang, N., Saporta, G., Villeminot, A., and Leblond, L. (2007). Combined use of association rules mining and clustering methods to find relevant links between binary rare attributes in a large data set. *Comput Stat Data Anal*, 52(1):596–613. 80

Pollack, J. R., Perou, C. M., Alizadeh, A. A., Eisen, M. B., Pergamenschikov, A., Williams, C. F., Jeffrey, S. S., Botstein, D., and Brown, P. O. (1999). Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nat Genet*, 23(1):41–6. 55

Pollack, J. R., Sø rlie, T., Perou, C. M., Rees, C. A., Jeffrey, S. S., Lonning, P. E., Tibshirani, R., Botstein, D., Bø rresen Dale, A.-L., and Brown, P. O. (2002). Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proc Natl Acad Sci U S A*, 99(20):12963–8. 56, 60

Poulogiannis, G., Ichimura, K., Hamoudi, R. A., Luo, F., Leung, S. Y., Yuen, S. T., Harrison, D. J., Wyllie, A. H., and Arends, M. J. (2010). Prognostic relevance of DNA copy number changes in colorectal cancer. *J Pathol*, 220(3):338–47. 50

Prieto, C. and De Las Rivas, J. (2006). APID: Agile Protein Interaction DataAnalyzer. *Nucleic Acids Res*, 34(Web Server issue):W298–302. 84

Qin, L.-X., Beyer, R. P., Hudson, F.Ñ., Linford, N. J., Morris, D. E., and Kerr, K. F. (2006). Evaluation of methods for oligonucleotide array data via quantitative real-time PCR. *BMC Bioinformatics*, 7:23. 13

Reifenberger, G. and Collins, V. P. (2004). Pathology and molecular genetics of astrocytic gliomas. *J Mol Med (Berl)*, 82(10):656–70. 69

Risueno, 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. 14, 57

Ruano, Y., Mollejo, M., Ribalta, T., Fiaño, C., Camacho, F. I., Gómez, E., de Lope, A. R., Hernández-Moneo, J.-L., Martínez, P., and Meléndez, B. (2006). Identification of novel candidate target genes in amplicons of Glioblastoma multiforme tumors detected by expression and CGH microarray profiling. *Mol Cancer*, 5:39. 66, 70

Ruepp, A., Waegele, B., Lechner, M., Brauner, B., Dunger-Kaltenbach, I., Fobo, G., Frishman, G., Montrone, C., and Mewes, H.-W. (2010). CORUM: the comprehensive resource of mammalian protein complexes–2009. *Nucleic Acids Res*, 38(Database issue):D497–501. 90

Sayagués, J. M., Fontanillo, C., Abad, M. d. M., González-González, M., Sarasquete, M. E., Chillon, M. d. C., Garcia, E., Bengoechea, O., Fonseca, E., Gonzalez-Diaz, M., De las Rivas, J., Muñoz Bellvis, L., and Orfao, A. (2010). Mapping of genetic abnormalities of primary tumours from metastatic CRC by high-resolution SNP arrays. *PloS One*, 5(10):e13752. 45, 48, 53

Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270(5235):467–70. 5, 55

Scheurlen, W. G., Schwabe, G. C., Seranski, P., Joos, S., Harbott, J., Metzke, S., Döhner, H., Poustka, A., Wilgenbus, K., and Haas, O. A. (1999). Mapping of the breakpoints on the short arm of chromosome 17 in neoplasms with an i(17q). *Genes Chromosomes Cancer*, 25(3):230–40. 53

Sharp, A. J., Hansen, S., Selzer, R. R., Cheng, Z., Regan, R., Hurst, J. A., Stewart, H., Price, S. M., Blair, E., Hennekam, R. C., Fitzpatrick, C. A., Segraves, R., Richmond, T. A., Guiver, C., Albertson, D. G., Pinkel, D., Eis, P. S., Schwartz, S., Knight, S. J. L., and Eichler, E. E. (2006). Discovery of previously unidentified genomic disorders from the duplication architecture of the human genome. *Nat Genet*, 38(9):1038–42. 53

Shen, W. F., Rozenfeld, S., Kwong, A., Köm ves, L. G., Lawrence, H. J., and Largman, C. (1999). HOXA9 forms triple complexes with PBX2 and MEIS1 in myeloid cells. *Mol Cell Biol*, 19(4):3051–61. 27

Sims, A. H. (2009). Bioinformatics and breast cancer: what can high-throughput genomic approaches actually tell us? *J Clin Pathol*, 62(10):879–85. 56

Smith, J. E., Bollekens, J. A., Inghirami, G., and Takeshita, K. (1997). Cloning and mapping of the MEIS1 gene, the human homolog of a murine leukemogenic gene. *Genomics*, 43(1):99–103. 26

Statnikov, A., Aliferis, C. F., Tsamardinos, I., Hardin, D., and Levy, S. (2005). A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. *Bioinformatics*, 21(5):631–43. 10

Stransky, N., Vallot, C., Reyal, F., Bernard-Pierrot, I., de Medina, S. G. D., Segraves, R., de Rycke, Y., Elvin, P., Cassidy, A., Spraggon, C., Graham, A., Southgate, J., Asselain, B., Allory, Y., Abbou, C. C., Albertson, D. G., Thiery, J. P., Chopin, D. K., Pinkel, D., and Radvanyi, F. (2006). Regional copy number-independent deregulation of transcription in cancer. *Nat Genet*, 38(12):1386–96. 70

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. a., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., and Mesirov, J. P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*, 102(43):15545–50. 56, 78

Toivonen, H., Klemettinen, M., Ronkainen, P., Hatonen, K., and Mannila, H. (1995). Pruning and Grouping Discovered Association Rules. In *MLnet Workshop on Statistics, Machine Learning and Discovery in Databases*, Heraklion, Greece. 80, 83

Tomlins, S. A., Rhodes, D. R., Perner, S., Dhanasekaran, S. M., Mehra, R., Sun, X.-W., Varambally, S., Cao, X., Tchinda, J., Kuefer, R., Lee, C., Montie, J. E., Shah, R. B., Pienta, K. J., Rubin, M. A., and Chinnaiyan, A. M. (2005). Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science*, 310(5748):644–8. 42

Tonon, G., Wong, K.-K., Maulik, G., Brennan, C., Feng, B., Zhang, Y., Khatry, D. B., Protopopov, A., You, M. J., Aguirre, A. J., Martin, E. S., Yang, Z., Ji, H., Chin, L., and Depinho, R. A. (2005). High-resolution genomic profiles of human lung cancer. *Proc Natl Acad Sci U S A*, 102(27):9625–9630. 38, 41

Tukey, J. W. (1977). *Exploratory Data Analysis*. Addison-we edition. 40

Turner, N., Lambros, M. B., Horlings, H. M., Pearson, A., Sharpe, R., Natrajan, R., Geyer, F. C., van Kouwenhove, M., Kreike, B., Mackay, A., Ashworth, A., van de Vijver, M. J., and Reis-Filho, J. S. (2010). Integrative molecular profiling of triple negative breast cancers identifies amplicon drivers and potential therapeutic targets. *Oncogene*, 29(14):2013–23. 56, 59

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(9):5116–21. 11, 56

Varma, S. and Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, 7:91. 19

Venkatraman, E. S. and Olshen, A. B. (2007). A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics*, 23(6):657–63. 31, 37, 52, 59

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., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., Skupski, M., Subramanian, G., Thomas, P. D., Zhang, J., Gabor Miklos, G. L., Nelson, C., Broder, S., Clark, A. G., Nadeau, J., McKusick, V. A., Zinder, N., Levine, A. J., Roberts, R. J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Di Francesco, V., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A. E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T. J., Higgins, M. E., Ji, R. R., Ke, Z., Ketchum, K. A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G. V., Milshina, N., Moore, H. M., Naik, A. K., Narayan, V. A., Neelam, B., Nusskern, D., Rusch, D. B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z., Wang, A., Wang, X., Wang, J., Wei, M., Wides, R., Xiao, C., Yan, C., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Zhu, S., Zhao, S., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F., An, H., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Center,

A., Cheng, M. L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferriera, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y. H., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N.Ñ., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-Deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J. F., Guigó, R., Campbell, M. J., Sjolander, K. V., Karlak, B., Kejariwal, A., Mi, H., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yooseph, S., Allen, D., Basu, A., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y. H., Coyne, M., Dahlke, C., Mays, A., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D., Majoros, W., McDaniel, J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N., Nodell, M., Pan, S., Peck, J., Peterson, M., Rowe, W., Sanders, R., Scott, J., Simpson, M., Smith, T., Sprague, A., Stockwell, T., Turner, R., Venter, E., Wang, M., Wen, M., Wu, D., Wu, M., Xia, A., Zandieh, A., and Zhu, X. (2001). The sequence of the human genome. *Science*, 291(5507):1304–51. 1

Wang, K., Li, M., Hadley, D., Liu, R., Glessner, J., Grant, S. F. A., Hakonarson, H., and Bucan, M. (2007). PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res*, 17(11):1665–74. 37

Ward, J. H. (1963). Hierarchical Grouping to Optimize an Objective Function. *J Am Stat Assoc*, 58(301):236. 80

Watson, J. D. (1990). The human genome project: past, present, and future. *Science*, 248(4951):44–9. 1

Watson, S. K., DeLeeuw, R. J., Horsman, D. E., Squire, J. A., and Lam, W. L. (2007). Cytogenetically balanced translocations are associated with focal copy number alterations. *Hum Genet*, 120(6):795–805. 42

Wechsler, D. S., Shelly, C. A., Petroff, C. A., and Dang, C. V. (1997). MXI1, a putative tumor suppressor gene, suppresses growth of human glioblastoma cells. *Cancer Res*, 57(21):4905–12. 69

Willenbrock, H. and Fridlyand, J. (2005). A comparison study: applying segmentation to array CGH data for downstream analyses. *Bioinformatics*, 21(22):4084–91. 37, 38

Wilson, D., Pethica, R., Zhou, Y., Talbot, C., Vogel, C., Madera, M., Chothia, C., and Gough, J. (2009). SUPERFAMILY–sophisticated comparative genomics, data mining, visualization and phylogeny. *Nucleic Acids Res*, 37(Database issue):D380–6. 75

Wilson, I. M., Davies, J. J., Weber, M., Brown, C. J., Alvarez, C. E., MacAulay, C., Schübeler, D., and Lam, W. L. (2006). Epigenomics: mapping the methylome. *Cell Cycle*, 5(2):155–8. 70

Yan, P. S., Perry, M. R., Laux, D. E., Asare, A. L., Caldwell, C. W., and Huang, T. H. (2000). CpG island arrays: an application toward deciphering epigenetic signatures of breast cancer. *Clin Cancer Res*, 6(4):1432–8. 55

Apéndice: Publicaciones científicas realizadas durante la presente Tesis Doctoral

Publicaciones relacionadas con el Capítulo 1

Aibar, S., Fontanillo, C., Droste, C., and De Las Rivas, J. (2013). geNetClassifier: classify multiple diseases and build associated gene networks using gene expression profiles. En preparación.

Publicaciones relacionadas con el Capítulo 2

Sayagués, J. M., Fontanillo, C., Abad, M. M., González-González, M., Sarasqete, M. E., Chillon, M. C., García, E., Bengoechea, O., Fonseca, E., Gonzalez-Diaz, M., De Las Rivas, J., Muñoz-Bellvis, L., and Orfao, A. (2010). Mapping of genetic abnormalities of primary tumours from metastatic CRC by high-resolution SNP arrays. *PLoS One*,5(10):e13752.

Muñoz-Bellvis, L., Fontanillo, C., González-González, M., Garcia, E., Iglesias, M., Esteban, C., Gutierrez, M. L., Abad, M. M., Bengoechea, O., De Las Rivas, J., Orfao, A., and Sayagués, J. M. (2012). Unique genetic profile of sporadic colorectal cancer liver metastasis versus primary tumors as defined by high-density single-nucleotide polymorphism arrays. *Mod Pathol*, 25(4):590-601.

González-González, M., Muñoz-Bellvis, L., Mackintosh, C., Fontanillo, C., Gutiérrez, M. L., Abad, M. M., Bengoechea, O., Teodosio, C., Fonseca, E., Fuentes, M., De Las Rivas, J., Orfao, A., and Sayagués, J. M. (2012). Prognostic impact of del(17p) and del(22q) as assessed by interphase FISH in sporadic colorectal carcinomas. *PLoS One*, 7(8):e42683.

Publicaciones relacionadas con el Capítulo 3

Ortiz-Estevez, M., De Las Rivas, J., Fontanillo, C., and Rubio, A. (2011). Segmentation of genomic and transcriptomic microarrays data reveals major correlation between DNA copy number aberrations and gene-loci expression. *Genomics*, 97(2):86-93.

Fontanillo, C., Aibar, S., Sanchez-Santos, J. M., and De Las Rivas, J. (2012). Combined analysis of genome-wide expression and copy number profiles to identify key altered genomic regions in cancer. *BMC Genomics*, 13(Suppl 5):S5.

Publicaciones relacionadas con el Capítulo 4

Fontanillo, C., Nogales-Cadenas, R., Pascual-Montano, A., and De Las Rivas, J. (2011). Functional analysis beyond enrichment: non-redundant reciprocal linkage of genes and biological terms. *PLoS One*, 6(9):e24289.

Otras publicaciones

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(12):e3911.

Hernández, J. A., Rodríguez A. E., González, M., Benito, R., Fontanillo, C., Sandoval, V., Romero, M., Martín-Nuñez, G., de Coca, A. G., Fisac, R., Galende, J., Recio, I., Ortuño, F., García, J. L., De Las Rivas, J., Gutiérrez, N.C., San Miguel, J. F., and Hernández, J. M. (2009). A high number of losses in 13q14 chromosome band is associated with a worse outcome and biological differences in patients with B-cell chromic lymphoid leukemia. *Haematologica*, 94(3):364-71.

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.

De Las Rivas, J., and Fontanillo, C. (2010). Protein-protein interactions essentials: key concepts to building and analyzing interactome networks. *PLoS Comput Biol*, 6(6):e1000807.

De Las Rivas, J., and Fontanillo, C. (2012). Protein-protein interactions networks: unraveling the wiring of molecular machines within the cell. *Brief Funct Genomics*, 11(6):489-96.

.

OPEN ACCESS Freely available online

PLoS one

Mapping of Genetic Abnormalities of Primary Tumours from Metastatic CRC by High-Resolution SNP Arrays

José María Sayagués[1], Celia Fontanillo[2], María del Mar Abad[3], María González-González[1], María Eugenia Sarasquete[4], Maria del Carmen Chillon[4], Eva Garcia[5], Oscar Bengoechea[3], Emilio Fonseca[6], Marcos Gonzalez-Diaz[4], Javier De Las Rivas[2], Luís Muñoz-Bellvis[7☯], Alberto Orfao[1*☯]

1 Servicio General de Citometría, Departamento de Medicina and Centro de Investigación del Cáncer (IBMCC-CSIC/USAL), Universidad de Salamanca, Salamanca, Spain, **2** Grupo de Investigación en Bioinformática y Genómica Funcional, Centro de Investigación del Cáncer (IBMCC-CSIC/USAL), Universidad de Salamanca, Salamanca, Spain, **3** Departamento de Patología, Hospital Universitario de Salamanca, Salamanca, Spain, **4** Servicio de Hematología, Hospital Universitario, Centro de Investigación del Cáncer (IBMCC-CSIC/USAL), Salamanca, Spain, **5** Unidad de Genómica y Proteómica, Centro de Investigación del Cáncer (IBMCC-CSIC/USAL), Universidad de Salamanca, Salamanca, Spain, **6** Servicio de Oncología Médica, Departamento de Cirugía, Hospital Universitario de Salamanca, Salamanca, Spain, **7** Unidad de Cirugía Hepatobiliopancreática, Departamento de Cirugía, Hospital Universitario de Salamanca, Salamanca, Spain

Abstract

Background: For years, the genetics of metastatic colorectal cancer (CRC) have been studied using a variety of techniques. However, most of the approaches employed so far have a relatively limited resolution which hampers detailed characterization of the common recurrent chromosomal breakpoints as well as the identification of small regions carrying genetic changes and the genes involved in them.

Methodology/Principal Findings: Here we applied 500K SNP arrays to map the most common chromosomal lesions present at diagnosis in a series of 23 primary tumours from sporadic CRC patients who had developed liver metastasis. Overall our results confirm that the genetic profile of metastatic CRC is defined by imbalanced gains of chromosomes 7, 8q, 11q, 13q, 20q and X together with losses of the 1p, 8p, 17p and 18q chromosome regions. In addition, SNP-array studies allowed the identification of small (<1.3 Mb) and extensive/large (>1.5 Mb) altered DNA sequences, many of which contain cancer genes known to be involved in CRC and the metastatic process. Detailed characterization of the breakpoint regions for the altered chromosomes showed four recurrent breakpoints at chromosomes 1p12, 8p12, 17p11.2 and 20p12.1; interestingly, the most frequently observed recurrent chromosomal breakpoint was localized at 17p11.2 and systematically targeted the *FAM27L* gene, whose role in CRC deserves further investigations.

Conclusions/Significance: In summary, in the present study we provide a detailed map of the genetic abnormalities of primary tumours from metastatic CRC patients, which confirm and extend on previous observations as regards the identification of genes potentially involved in development of CRC and the metastatic process.

Citation: Sayagués JM, Fontanillo C, Abad MdM, González-González M, Sarasquete ME, et al. (2010) Mapping of Genetic Abnormalities of Primary Tumours from Metastatic CRC by High-Resolution SNP Arrays. PLoS ONE 5(10): e13752. doi:10.1371/journal.pone.0013752

Editor: Zhongjun Zhou, The University of Hong Kong, Hong Kong

Received July 7, 2010; **Accepted** October 6, 2010; **Published** October 29, 2010

Funding: This work has been partially supported by grants from the Consejeria de Sanidad, Junta de Castilla y Leon, Valladolid, Spain (SAN191/SA09/06 and SAN673/SA39/08), Fundacion Memoria de Don Samuel Solorzano Barruso, Salamanca, Spain, Caja de Burgos (Obra Social), Burgos, Spain, Grupo Excelencia de Castilla y Leon (GR37) and the RTICC from the Instituto de Salud Carlos III (ISCIII), Ministerio de Ciencia e Innovacion, Madrid, Spain (RD06/0020/0035-FEDER). JM Sayagués, M Gonzalez, ME Sarasquete and MC Chillon are supported by grants (CP05/00321, FI08/00721, CA08/00212 and CA/07/00077, respectively) from the ISCIII, Ministerio de Ciencia e Innovación, Madrid, Spain. 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: orfao@usal.es

☯ These authors contributed equally to this work.

Introduction

The development and progression of CRC is a multistep process leading to the accumulation of genomic alterations that occur at the single cell level over the lifetime of a tumour, from benign to invasive and metastatic states leading to patient death [1,2]. For many years, the genetics of metastatic CRC have been studied with an increasingly high variety of techniques from conventional cytogenetics [3] and fluorescence *in situ* hybridization (FISH) [4] to comparative genomic hybridization (CGH) [5] and array CGH (aCGH) [6]. Based on these techniques, many different recurrent genetic abnormalities have been identified in metastatic CRC which frequently include gains of chromosomes 8q, 13q and 20q [7,8] together with losses of the 1p, 8p, 17p and 18q chromosomal regions [9]. By contrast, detailed characterization of the common breakpoint regions as well as the identification of the specific genes targeted by such abnormalities has proven difficult with these approaches. This is partially due to the fact that these techniques have a relatively limited resolution which hampers identification of the specific cancer-associated genes recurrently targeted in such alterations. In fact, the highest resolution approaches applied so far to the study of CRC are based on aCGH (i.e. Camps *et al* who

applied a 185K oligonucleotide array with an estimated resolution of 16 kb, to the analysis of 32 primary CRC tumours) [10].

In recent years, the availability of high-density single nucleotide polymorphism (SNP) arrays has allowed identification of small regions of chromosomal gains and losses with a much higher resolution, down to 2.5 kb [11]. Thus, based on genome wide SNP arrays, fine mapping of chromosomal breakpoints and subsequent identification of the specific genes recurrently altered (deleted, gained or amplified) is achieved for individual samples. This allows for a more precise and detailed comparison of the breakpoint regions found in different tumours and their correlation with the clinical features of the disease.

In the present study we used 500K SNP mapping arrays with a mean distance between interrogated SNPs of 5.8 kb (median intermarker distance of 2.5 kb) to map genetic lesions present at diagnosis in primary tumours from a group of 23 sporadic CRC patients who developed liver metastasis. Our major goal was to define the most frequent recurrent breakpoint regions in metastatic CRC and the commonly gained and/or deleted genes in the altered chromosomes. In order to evaluate the reproducibility of the SNP-array results we performed parallel interphase FISH (iFISH) analyses of the same tumour samples using 24 probes directed against an identical number of regions from 20 different human chromosomes frequently altered in sporadic CRC.

Materials and Methods

Patients and samples

Tissue specimens were obtained from primary tumours from 23 patients (15 males and 8 females; median age of 68 years, ranging from 48 to 80 years) suffering from metastatic sporadic CRC. The study was approved by the local ethics committee of the University Hospital of Salamanca (Salamanca, Spain) and prior to entering to the study, informed consent was given by each individual.

In each case, the diagnosis and the classification of the tumours were performed according to the WHO criteria [12]. According to tumour grade, 13 cases corresponded to well-differentiated CRC, 8 to moderately- and 2 to poorly-differentiated tumours. Histopathological grade was confirmed in all cases in a second independent evaluation by an experienced pathologist.

From the 23 primary tumors, 16 were localized at the right (caecum, ascending or trasverse) or the left (descending and sigmoid) colon and 7 in the rectum. Mean size of primary tumors was of 5.2 ± 1.8 cm with the following distribution according to the TNM stage [13]: T3N0M1, 3 cases; T3N1M1, 9; T3N2M1, 3; T4N0M1, 5; T4N1M1, 1 and; T4N2M1, 2 patients. In all cases paired liver metastases were identified either at the time of colorectal surgery ($n = 14$) or during the first year after initial diagnosis ($n = 9$); the mean size of the largest liver metastases/patient was of 5.3 ± 2.8 cm (range: 2 to 10 cm).

After histopathological diagnosis was established, samples from representative areas of the primary tumours showing macroscopical infiltration, were used to prepare single cell suspensions to be stored (-20°C) in methanol/acetic (3/1; vol/vol) for further iFISH analyses [14]. The remaining tissue was either fixed in formalin and embedded in paraffin or frozen in liquid nitrogen, and stored at room temperature (RT) and at -80°C, respectively. From the paraffin-embedded tissue samples, sections were cut from three different areas representative of the tumoural tissue used to prepare single cell suspensions and placed over poly L-lysine coated slides. All tissues were evaluated after hematoxylin-eosin staining to confirm the presence of tumour cells and evaluate their quantity in samples to be studied by both iFISH and SNP-arrays. For SNP-array studies, tumour DNA was extracted from freshly-frozen tumour tissues mirror cut to those used for iFISH analyses which contained $\geq$65% epithelial tumour cells. In turn, normal DNA was extracted from matched peripheral blood (PB) leucocytes from the same patient. For both types of samples (tumour tissue and PB leucocytes), DNA was extracted using the QIAamp DNA mini kit (Qiagen, Hilden, Germany) following the manufacturer's instructions.

Analysis of single nucleotide polymorphism (SNP) arrays

Paired samples of purified tumoural DNA and normal PB DNA from individual patients were hybridized to two 250K Affymetrix SNP Mapping arrays (*Nsp*I and *Sty*I SNP arrays, Affymetrix, Santa Clara, CA) using a total of 250 ng of DNA per array, according to the instructions of the manufacturer. Fluorescence signals were detected using the GeneChip Scanner 3000 (Affymetrix). Average genotyping call rates of 94.4% and 97.3% were obtained for tumoral and paired normal PB DNA samples, respectively. Only those SNPs with a call rate $\geq$92.3% were used for further analyses.

In order to calculate genome-wide copy number (CN) changes in tumoural vs. normal samples, the *aroma.affymetrix* algorithm was used, following the CRMA v2 method, as described elsewhere (R-software package, Berkeley, CA) [15]. The following sequential steps were used for this purpose: i) calibration for crosstalks between pairs of allele probes; ii) normalization for probe nucleotide-sequence effects, and; iii) normalization for PCR fragment length- and probe localization-dependent effects. Then, data derived from both the 250K *Sty*I and the 250K *Nsp*I arrays was integrated into a single database and raw CN values calculated as transformed log2 values of the tumoural/normal ratio obtained for paired SNP fluorescence signals.

Log2 ratio values were then used to identify DNA regions which showed similar CN values, using the Circular Binary Segmentation (CBS) algorithm [16]. For the identification of altered (gained or lost) DNA regions, a threshold was established based on the changes observed in the log2 CN values (fluorescence intensity ratio) of sequential tumour DNA segments found for each individual. Therefore, log2 ratio >0.09 and <-0.09 were used as cut-off thresholds to define the presence of increased and decreased CN values, respectively. High-level gains (amplifications) were defined as regions with a mean log2 CN ratio $\geq$0.22 for $\geq$3 contiguous SNPs. The specific frequencies of both CN gains and losses per SNP were established and plotted along individual chromosomes for each individual case analyzed. Minimal common regions (MCR) of gain and loss were defined as the smallest group of contiguous SNPs ($\geq$3) with a high frequency of gains and losses (Z-score threshold $\geq$2.1) according to the overall distribution of CN values found in the entire tumour cell genome, respectively. Common recurrent breakpoint regions were defined as those chromosomal regions which recurrently showed transition from one CN state (gain, loss or no-change) to another for the whole set of individual samples analyzed, at a frequency of $\geq$35% of the cases ($n = 8/23$ samples).

Interphase fluorescence in situ hybridization (iFISH) studies

In all cases, iFISH studies were performed on an aliquot of the single cell suspension prepared from the tumour sample. A set of 24 locus-specific FISH probes directed against DNA sequences localized in 20 different human chromosomes, specific for those chromosomal regions more frequently gained or deleted in sporadic CRC [4,6,8,17,18] were systematically used to validate the results obtained with the SNP arrays (Table 1).

Table 1. A panel of 24 locus-specific FISH probes directed against 24 different regions localized in 20 different human chromosomes were used to validate the results obtained with the SNP arrays.

iFISH probe chromosome localization	iFISH probe length (kb)	Target gene	N. of SNPs inside the region identified by the iFISH probe
1p36	110	*P58*	120
1q25	620	*ABL2*	68
2p24	200	*NMYC*	38
3q26	839	*HTERC*	52
5p15.2	450	*D5S721*	118
6q23	740	*MYB*	88
7q31	200	*D7S486*	33
8p22	170	*LPL*	39
8q24	600	*CMYC*	159
9p21	190	*P16*	37
9q34	270	*ABL1*	33
10q23	370	*PTEN*	49
11q22	184	*ATM*	69
12p13	350	*TEL*	98
13q14	220	*RB1*	14
13q34	550	*LAMP1*	92
14q32	1500	*IGH*	82
15q22	540	*DAPK2*	38
17p13	145	*TP53*	12
18q21	750	*BCL2*	153
19q13	340	*CD37*	21
20q13.2	320	*ZNF217*	53
21q22	500	*AML1*	111
22q11.2	300	*BCR*	36

All probes were purchased from Vysis Inc (Chicago, IL, USA), except for the 3q26, 15p22 and 19q13 probes, which were obtained from QBIOgene Inc (Amsterdam, The Netherlands).

doi:10.1371/journal.pone.0013752.t001

The methods and procedures used for the iFISH studies have been previously described in detail [19]. Briefly, dried slides containing both the tumour cells' and the probes' DNA were denatured (1 min at 75°C) and hybridized overnight (37°C) in a Hybrite termocycler (Vysis Inc, Downers Grove, IL, USA). After this incubation, slides were sequentially washed (5 min at 46°C) in 50% formamide in a 2× saline sodium citrate buffer (SSC) and in 2XSSC. Finally, nuclei were counterstained with 35 μL of a mounting medium containing 75 ng/ml of 4,6-diamidino 2-phenylindole (DAPI; Sigma, St Louis, MO, USA); Vectashield (Vector Laboratories Inc, Burlingame, CA, USA) was used as antifading agent.

A BX60 fluorescence microscope (Olympus, Hamburg, Germany) equipped with a 100× oil objective was used to count the number of hybridization spots/nuclei for ≥200 cells/sample. Only those spots with a similar size, intensity and shape were counted in areas with <1% unhybridized cells; doublet signals were considered as single spots. A tumour was considered to carry a numerical abnormality for a given chromosomal region when the proportion of cells displaying an abnormal number of hybridization spots for the corresponding probe was at a percentage higher or lower than the mean value plus two standard deviations (SD) of the mean percentage obtained with the same probe in control samples (n = 10).

Quantitative Real-Time PCR

In order to validate the results obtained in the SNP-array studies, quantitative real-time polymerase chain reaction (RQ-PCR) was performed using the Step One Plus Real-Time PCR System (Applied Biosystems, Foster City, CA) in matched normal and tumoural samples in 18/23 cases. Expression of the *MAP2K4*, *MYC* and *BIRC7* genes was analyzed. We employed TaqMan® Gene Expression Assays designed by Applied Biosystems (Applied Biosystems, Foster City, CA) according to the manufacturers instructions, and the assays ID for the genes studied were as follows: Hs_00387426-m1 (*MAP2K4*), Hs_00153408-m1 (*MYC*) and Hs_00223384-m1 (*BIRC7*).

Each PCR was carried out in duplicate in a 10 uL volume using the TaqMan® Fast Universal Mastermix (Applied Biosystems) and the following cycling parameters: incubation at 95°C (20 sec), followed by 50 cycles at 95°C (1 sec) and an incubation at 60°C (20 sec). Analysis was made using StepOne software v2.0. The obtained data were normalized by using the internal housekeeping gene, *GAPDH*. Relative quantification was calculated using the equation $2^{-\Delta CT} = C_{TGENE}\text{-}C_{TGAPDH.}$ The final mRNA expression index in each sample was calculated as follows (arbitrary units; AU): mRNA expression index = *MYC* or *MAP2K4* or *BIRC7* mRNA value/ *GAPDH* mRNA value X 10,000 AU.

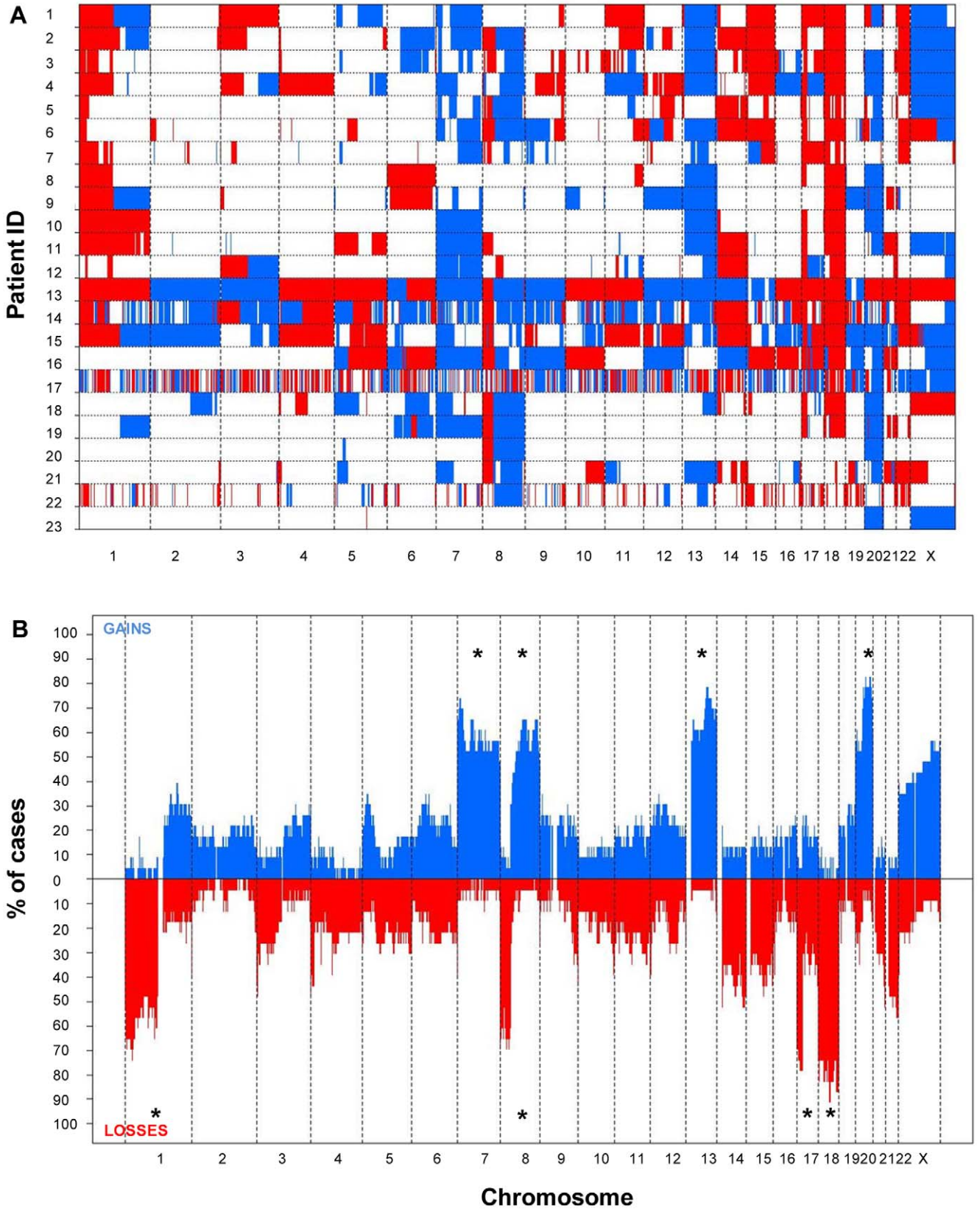


Figure 1. Metastatic colorectal cancer genome for the 23 CRC patients studied. In panel A an overall view of both the gained (blue areas) and lost (red areas) chromosome regions across the genome are shown for the 23 patients genotyped on the Affymetrix 500k SNP array platform. In panel B a summary plot showing the frequency of CN gains (plotted above zero values in the x-axis) and losses (plotted below zero values the x-axis) detected for each individual chromosome, is displayed. Those chromosome regions most frequently showing recurrent losses and gains by SNP

arrays were localized in chromosomes 1p, 8p, 17p and 18, and involved the whole chromosome 7 and the 8q, 13q and 20q chromosome regions, respectively.
doi:10.1371/journal.pone.0013752.g001

Statistical methods

For all continuous variables, mean values (and SD) and range were calculated using the SPSS software package (SPSS 12.0 Inc, Chicago, IL USA); for dichotomic variables, frequencies were reported. In order to evaluate the statistical significance of differences observed between groups, the Mann-Whitney U and X^2 tests were used for continuous and categorical variables, respectively (SPSS).

A multivariate stepwise regression analysis (regression, SPSS) was performed to determine the correlation between the structural and/or numerical abnormalities found for both iFISH, SNP-array techniques and their relationship with the expression of those genes analyzed by RQ-PCR. Only those iFISH probes with ≥12 SNPs localized in the iFISH mapped region (Table 1) were used for correlation studies with the CN status identified by the SNP array (gain vs. loss vs. no change) for those SNPs localized at each iFISH region. *P*-values <.01 were considered to be associated with statistical significance.

Results

Map of CN changes by SNP arrays

Overall CN changes for at least one chromosomal region were detected in all 23 tumors studied. The highest frequency of CN losses detected corresponded to chromosomes 1p (n = 17; 74%), 8p (n = 18; 78%), 14q (n = 15; 65%), 17p (n = 19; 83%), 18 (n = 21; 91%) and 22q (n = 17; 74%); in turn, CN gains more frequently involved chromosomes 1q (n = 10; 43%), 7 (n = 20; 87%), 8q (n = 17; 74%), 13q (n = 18; 78%), 20q (n = 20; 87%) and X (n = 13; 57%) (Figure 1); these (gained) chromosomes/chromosomal regions also revealed the highest level of genomic amplification (Table S1). In addition, gains and losses of many other chromosomal regions were identified at lower frequencies (Figure 1). An illustrating map of the most frequently gained/lost chromosome regions according to SNP-array studies, is shown in figure 2.

Of note, SNP arrays allowed the identification of 43 small DNA sequences (arbitrarily defined as regions of <1300 kb) which displayed recurrent CN changes (gains and losses). Interestingly, most of those regions which showed recurrent CN changes (n = 28/43) contained at least one known well-characterized gene, five contained known cancer-associated genes and one region held a microRNA gene (*MIR1208*), localized at chromosome 8q24.21 (Table 2). The exact number of small regions characterized by CN changes, as well as the relative proportion of CN gains vs. losses varied widely among the different chromosomes. The 43 small regions containing CN gains and losses were coded in those chromosomes more frequently affected by CN changes and their

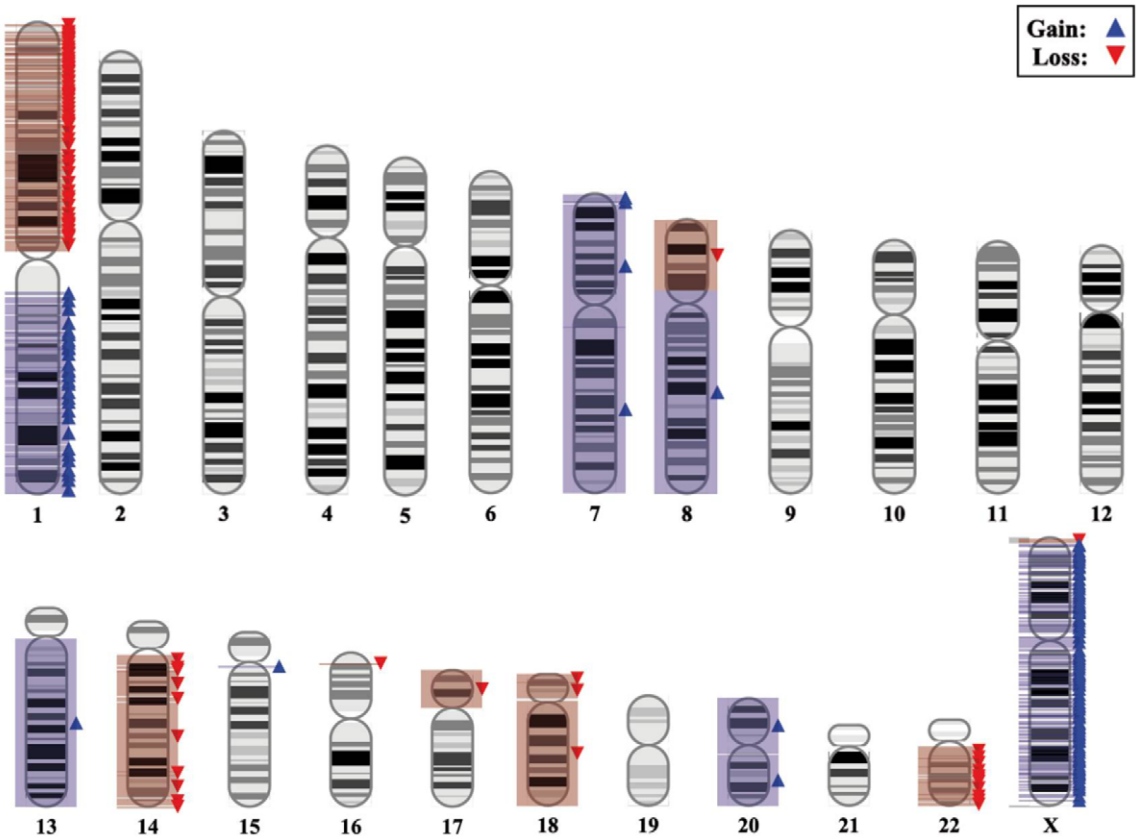


Figure 2. Representative karyotype of a primary metastatic colorectal tumor as determined by the Affymetrix 500K SNP array genotyping platform, showing summary results for those chromosome gains/losses more frequently detected in the colorectal tumor samples analyzed (n = 23).
doi:10.1371/journal.pone.0013752.g002

Table 2. Most frequently detected small regions (<1300 kb) of gain and loss in primary sporadic colorectal tumors genotyped on the Affymetrix 500K SNP array platform (n = 23).

Minimal common altered regions (bp)	Region length (bp)	N. of SNPs	Chromosome band	Event	% of altered cases	Gene list
Chr 1: 26,131,131-26,191,419	60,288	16	1p36.11	Deletion	74	*PAFAH2*
Chr 7: 8,255,230-8,280,496	25,266	10	7p21.3	Gain	74	*ICA1*
Chr 7: 10,461,770-10,486,412	24,642	8	7p21.3	Gain	74	–
Chr 7: 12,514,442-12,576,898	62,456	9	7p21.3	Gain	74	*SCIN*
Chr 8: 32,105,734-32,675,812	570,078	196	8p12	Deletion	70	–
Chr 8: 198,834-392,556	193,722	46	8p23.3	Deletion	70	*FAM87A, FBXO25*
Chr 8: 400,640-539,716	139,076	29	8p23.3	Deletion	70	*C8orf42*
Chr 8: 23,264,737-23,277,681	12,944	8	8p21.3	Deletion	70	–
Chr 8: 86,214,670-86,946,337	731,667	52	8q21.2	Gain	65	*LRRCC1,* ***E2F5, CA13, CA1****, CA3, CA2*
Chr 8: 87,377,186-87,789,535	412,349	65	8q21.3	Gain	65	***WWP1****, FAM82B, CPNE3, CNGB3*
Chr 8: 88,872,540-89,066,702	194,162	24	8q21.3	Gain	65	*WDR21C*
Chr 8: 91,462,487-91,474,759	12,272	2	8q21.3	Gain	65	–
Chr 8: 91,686,333-91,735,940	49,607	10	8q21.3	Gain	65	*TMEM64*
Chr 8: 94,759,374-95,077,320	317,946	44	8q22.1	Gain	65	*RBM12B, C8orf39, TMEM67, PPM2C*
Chr 8: 95,294,349-95,435,061	140,712	28	8q22.1	Gain	65	*GEM*
Chr 8: 95,593,385-95,776,644	183,259	36	8q22.1	Gain	65	*KIAA1429, RBM35A*
Chr 8: 128,638,191-128,724,583	86,392	25	8q24.21	Gain	65	–
Chr 8: 129,180,096-129,268,067	87,971	43	8q24.21	Gain	65	*MIR1208*
Chr 8: 130,906,244-131,222,249	316,005	35	8q24.21	Gain	65	*FAM49B*
Chr 8: 133,845,345-133,868,639	23,294	9	8q24.22	Gain	65	*PHF20L1*
Chr 8: 133,882,656-133,900,665	18,009	6	8q24.22	Gain	65	–
Chr 8: 135,527,585-135,836,235	308,650	97	8q24.22	Gain	65	*ZFAT*
Chr 8: 136,498,075-136,866,133	368,058	74	8q24.23	Gain	65	***KHDRBS3***
Chr 8: 137,055,200-137,091,177	35,977	12	8q24.23	Gain	65	–
Chr 13: 73,603,130-73,627,939	24,809	10	13q22.1	Gain	78	*KLF12*
Chr 13: 74,972,248-75,117,835	145,587	26	13q22.2	Gain	78	*COMMD6,* ***UCHL3, LMO7***
Chr 13: 75,689,304-75,689,865	561	2	13q22.2	Gain	78	–
Chr 13: 76,352,482-76,366,765	14,283	11	13q22.3	Gain	78	*KCTD12*
Chr 13: 78,098,212-78,143,588	45,376	7	13q31.1	Gain	78	*C13orf7*
Chr 13: 78,805,700-79,077,299	271,599	46	13q31.1	Gain	78	*RBM26, NDFIP2*
Chr 13: 79,621,013-79,845,948	224,935	40	13q31.1	Gain	78	***SPRY2***
Chr 17: 10,693,238-11,021,844	328,606	89	17p13.1	Deletion	78	–
Chr 17: 14,234,746-14,967,525	732,779	214	17p12	Deletion	78	–
Chr 17: 14,984,724-15,082,587	97,863	18	17p12	Deletion	78	*PMP22*
Chr 18: 41,130,655-41,494,986	364,331	134	18q12.3	Deletion	91	*SLC14A2*
Chr 18: 45,410,728-45,497,910	87,182	29	18q21.11	Deletion	91	–
Chr 18: 45,654,114-46,036,475	382,361	144	18q21.11	Deletion	91	*MYO5B, CCDC11*
Chr 18: 46,252,199-46,288,353	36,154	12	18q21.11	Deletion	91	–
Chr 20: 37,766,095-38,339,016	572,921	131	20q12	Gain	83	*HSPEP1*
Chr 20: 51,012,908-51,013,194	286	2	20q13.2	Gain	83	–
Chr 20: 52,991,500-54,234,439	1,242,939	325	20q13.2	Gain	83	*CBLN4*
Chr X: 134,159,698-134,160,254	556	2	Xq26.3	Gain	57	–
Chr X: 151,650,011-151,652,710	2699	2	Xq28	Gain	57	–

Genes which have been associated with cancer are shown in bold.
doi:10.1371/journal.pone.0013752.t002

distribution was as follows: chromosomes 1p, 1 region; 7p, 3; 8p, 4; 8q, 16; 13q, 7; 17p, 3; 18q, 4; 20q, 3, and; Xq, 2 region. In addition, other regions carrying recurrent large-scale CN gains and losses (arbitrarily defined as regions of >1500 kb) were identified at the 8q21.13, 17p12, 17p11.2, 22q13 and Xq25 chromosome segments (one in each chromosome). Interestingly,

each of these larger regions has been previously associated with malignancy and contained genes i) relevant to the metastatic process (i.e.: *TPD52*, *FABP5*, *MAP2K4*, *LLGL1*, *TOP3A*, *ALDH3A2*, *UPK3A*, *FBLN1*, *TYMP*), ii) associated with intracellular signaling processes (i.e.: *PAG1*, *ELAC2*, *RASD1* and *TNFRSF13B*) and iii) genes involved in the regulation of the cell cycle (i.e.: *FLCN*, *PEMT* and *XIAP*); in turn, three of these large CN regions showing CN losses and one with CN gains contained a total of 8 known microRNAs (Table 3).

Chromosomal regions showing high-level CN gains

The highest levels of genetic amplification were detected for the 7p15.2, 8q24.21, 13q12.13 and 20p12.3 chromosome bands with maximum fluorescence intensity $\log_2$ ratios of 0.99 (0.23±0.11), 1.45 (0.35±0.15), 1.47 (0.31±0.22) and 0.96 (0.28±0.11), respectively (Table 4). Several genes which are potentially involved in the pathogenesis of CRC are localized in these four chromosomal regions. Among others, these include the *CYCS* and *UPP1* genes on chromosome 7p, the *MYC* gene at chromosome 8q24.21, the *HSPH1* and *CDX2* genes at chromosome 13q and the *CDC25B*, *PLCB4*, *TNFRSF6B*, *OGFR*, *NTSR1*, *CDH4*, *CYP24A1* and *RGS19* genes in chromosome 20. The most commonly amplified single region (18/23 cases; 78%) corresponded to a region localized at chromosome 20q11.22 identified by the SNP_A-2220183 and the SNP_A-2039695 at the 33,776,127 bp and 33,954,944 bp positions, respectively (Table S1).

Interestingly, we recorded a statistically significant association between tumour grade and presence of gains/amplifications at the 20p13 chromosomal region localized between the 2,574,587 and 2,993,797 bp positions and assessed by 66 SNPs with a greater frequency of well- vs moderately-differentiated tumours- (11/13 (85%) vs 2/8 (25%); $p = 0.005$) among cases with this chromosomal alteration.

Recurrent chromosomal breakpoints identified by SNP-arrays

Based on the analysis of the distribution of chromosomal breakpoints defined by the SNP-arrays, four recurrent chromosomal breakpoints (arbitrarily defined as DNA segments showing CN changes in more than one third of the cases) were identified at chromosomes 1p12, 8p12, 17p11.2 and 20p12.1 (Figure S1). Chromosomes 1, 8 and 20 showed a high number (>145) of different breakpoint regions with a variable and heterogeneous distribution; in contrast, a highly prevalent breakpoint region was identified in the centromeric portion of chromosome 17p, between the genome coordinates 20,156,497 bp and 22,975,771 bp (15/19 patients with abnormalities for this chromosome), and a minimum size of 28.2 Mb for the recurrent breakpoint. In these 15 cases, the first gene affected on the retained telomeric side of the breakpoint region was the *CYTSB* gene and the first constantly deleted gene on the centromeric side was the *FAM27L* gene. Interestingly, in 13 of these 15 patients a preferential breakpoint occurred at the 21,769,828–22,975,771 genome coordinate where the *FAM27L* gene is coded.

Correlation between the chromosomal changes detected by SNP-arrays and both iFISH and RQ-PCR studies

In order to evaluate the consistency of the chromosomal changes identified by the SNP-arrays, iFISH analysis were performed in parallel for a total of 24 chromosome regions from 20 different chromosomes. Overall our results showed a high degree of correlation (mean r^2 of 0.73±02; range: 0.65 to 0.91) between both methods, including when such analysis was restricted to the most frequently altered regions ($r^2 \geq 0.67$) (Table 5).

In order to assess the impact of the information generated by SNP arrays, the expression of three genes (*MAP2K4*, *MYC* and

Table 3. Most frequently detected extensively altered chromosome regions with CN changes (>1500 kb) in primary sporadic colorectal tumors genotyped on the Affymetrix 500K SNP array platform (n = 23).

Extensively altered regions (bp)	Region length (bp)	Chromosome band	Event	% of altered cases	Gene list
Chr 8: 80,831,670-82,390,493	1,558,823	8q21.13	Gain	65	*HEY1, MRPS28,* ***TPD52****, ZBTB10, ZNF704,* ***PAG1****,* ***FABP5***
Chr 17: 11,135,229-14,009,355	2,874,126	17p12	Deletion	78	*DNAH9, ZNF18,* ***MAP2K4****, MIR744, MYOCD,* ***ELAC2****, HS3ST3A1, MIR548H3, COX10*
Chr 17: 16,270,540-19,616,367	3,345,827	17p11.2	Deletion	78	*TRPV2, C17orf45, C17orf76, ZNF287, ZNF624, CCDC144A,* ***TNFRSF13B****, C17orf84,* ***FLCN****, COPS3, NT5M, MED9,* ***RASD1****,* ***PEMT****, RAI1, SREBF1, MIR33B, TOM1L2, LRRC48, ATPAF2, C17orf39, DRG2, MYO15A, ALKBH5,* ***LLGL1****, FLII, SMCR7,* ***TOP3A****, SMCR8,* ***SHMT1****, NOS2B, TBC1D28, TRIM16L, FBXW10, FAM18B, PRPSAP2, SLC5A10, FAM83G, GRAP, EPN2, B9D1, MIR1180, MAPK7, MFAP4, ZNF179, SLC47A1,* ***ALDH3A2****, SLC47A2,* ***ALDH3A1****, ULK2*
Chr 22: 43,616,234-49,576,671	5,960,437	22q13	Deletion	57	*ARHGAP8, PHF21B, NUP50, C22orf9, MIR1249,* ***UPK3A****, FAM118A, SMC1B, RIBC2,* ***FBLN1****, ATXN10,* ***WNT7B****, C22orf26, MIRLET7A3, MIRLET7B, PPARA, PKDREJ, GTSE1, TRMU, CELSR1, GRAMD4, CERK, TBC1D22A, FAM19A5, C22orf34, BRD1, ZBED4, ALG12, CRELD2, PIM3, IL17REL, TTLL8, MLC1, MOV10L1, PANX2, TRABD, TUBGCP6, HDAC10, MAPK12, MAPK11, PLXNB2, FAM116B, SAPS2, SBF1, ADM2, MIOX, TMEM112B, NCAPH2, SCO2,* ***TYMP****, KLHDC7B, CPT1B, CHKB, MAPK8IP2,* ***ARSA****,SHANK3, ACR, RABL2B*
Chr X: 120,721,375-126,726,076	6,004,701	Xq25	Gain	57	*GRIA3, THOC2, MIR220A,* ***XIAP****, STAG2, SH2D1A, ODZ1, WDR40C, WDR40B, CXorf64*

Genes which have been associated with cancer are shown in bold.
doi:10.1371/journal.pone.0013752.t003

Table 4. Most frequently detected high-level amplified chromosome regions (average $\log_2$ copy number ratio $\geq$0.22) containing genes commonly associated with cancer in primary sporadic colorectal tumors genotyped on the Affymetrix 500K SNP array platform (n = 23).

Amplified chromosome regions (bp)	Chromosome band	Mean $\log_2$ Ratio	Maximum $\log_2$ Ratio	% of altered cases	Cancer associated genes
Chr 7: 21,060,948-21,773,238	7p15.3	0.22	0.51	57	*SP4*
Chr 7: 25,072,457-29,780,614	7p15.2	0.23	0.99	52	***CYCS***, *CHN2, JAZF1, HOXA1, HOXA4, HOXA5, HOXA7, HOXA9, HOXA10, HOXA11, HNRPA2B1*
Chr 7: 30,433,934-47,043,330	7p15.1	0.24	0.69	52	*SFRP4, AMPH, RALA, INHBA, PPIA, IGFBP3*
Chr 7: 47,249,414-48,538,115	7p12.3	0.23	0.51	57	***UPP1***
Chr 7: 50,305,027-50,512,587	7p12.2	0.24	0.51	61	*DDC, IKZF1*
Chr 8: 128,130,968-129,218,353	8q24.21	0.35	1.45	61	***MYC***
Chr 13: 22,371,210-23,251,245	13q12.12	0.29	0.81	57	*SACS*
Chr 13: 23,722,973-24,224,179	13q12.12	0.30	0.90	57	*ATP12A, PARP4*
Chr 13: 25,516,360-33,070,797	13q12.13	0.31	1.47	61	*BRCA2, RXFP2, HMGB1,* ***HSPH1****, SLC7A1, FLT1, FLT3,* ***CDX2****, PDX1, GTF3A*
Chr 20: 3,590,646-3,775,309	20p13	0.28	0.62	52	***CDC25B****, SIGLEC1, GFRA4*
Chr 20: 6,077,268-10,228,083	20p12.3	0.28	0.96	52	***PLCB4****, PLCB1*
Chr 20: 33,776,127-33,954,944	20q11.22	0.27	0.51	78	*RBM39, PHF20*
Chr 20: 47,898,202-49,082,996	20q13.13	0.27	0.55	74	*ADNP, BCAS4, PTPN1, CEBPB, SNAI1*
Chr 20: 52,203,846-52,261,791	20q13.2	0.27	0.55	74	***CYP24A1***
Chr 20: 59,237,873-59,740,719	20q13.33	0.27	0.59	74	***CDH4***
Chr 20: 59,926,031-62,297,793	20q13.33	0.28	0.82	74	*TAF4, SS18L1, LAMA5,* ***GATA5****, SLCO4A1,* ***NTSR1****,* ***OGFR****,* ***TCFL5****, DIDO1, BIRC7, EEF1A2, PTK6, STMN3,* ***NFRSF6B****, TPD52L2, SOX18,* ***RGS19****, OPRL1,*

Genes which have been commonly associated with colorectal cancer are shown in bold.
Only those regions with recurrently amplified DNA copy-number found in at least half of the cases, are listed.
doi:10.1371/journal.pone.0013752.t004

BIRC7) was further analyzed in detail using RQ-PCR. As expected from the SNP-array data, the *MYC* and *BIRC7* relative transcript levels were up-regulated in 15/18 (83%) and 14/18 (78%) tumours analyzed, respectively. Conversely, the *MAP2K4* gene was down-regulated in 16/18 (89%) tumours (Figure 3). Upon comparing the results obtained with the two methods, a significant ($p<0.001$) correlation was observed between the microarray data and the expression of the three genes evaluated by RQ-PCR techniques with correlation coefficients (r^2) of 0.88, 0.66 and 0.64 for *MAP2K4*, *MYC* and *BIRC7* genes, respectively.

Discussion

In this study we describe a comprehensive map of the genetic abnormalities present in primary tumors from metastatic CRC through the usage of high-resolution 500K SNP arrays. To our knowledge this is the most extensive study using high-resolution SNP-arrays to define the genetic alterations in this subgroup of CRC patients. Overall, our results confirm previous analyses using chromosome banding techniques [20], CGH [5], SKY [21], aCGH [6,10] and low-resolution 50k SNP-arrays [22].

Previous reports in which similar SNP-array tools have been applied to investigate the genetic profile of non-metastatic CRC [23] have shown in a subset of patients with advanced carcinomas in the absence of liver metastases (n = 18), a relatively low frequency of 1p, 8p, 9q, 14 and 17p losses and unique amplifications at chromosome 20q. Interestingly, among our series of metastatic CRC patients the frequency of losses at the same chromosomal regions was strikingly higher: 1p, 74% vs 11%; 8p, 78% vs 33%; 9q, 35% vs 6%; 14, 65% vs 39%; and; 17p, 83% vs 33%. In turn, we also detected additional amplifications at 7p, 8q and 13q, as well as at the 20q chromosomal region. In line with our observations, Al-Mulla *et al* [24] also found that, once compared to patients without metastatic disease (n = 30) CRC patients with liver metastases (n = 26) more frequently displayed losses of chromosomes 1p, 4, 5q, 8p, 9p, and 14q. Altogether, those results indicate that the genetic profile of metastatic CRC is defined by imbalanced gains/amplifications of chromosomes 7p, 8q, 13q and 20q together with losses of the 1p, 8p, 9p, 14q and 17p chromosomal regions [5,20,25–27]. In addition, here we describe new recurrently altered regions that contain cancer genes, many of which have been previously involved in the pathogenesis of CRC, at the same time, we provide detailed characterization of recurrent chromosomal breakpoints most frequently occurring in primary tumours from CRC patients who had developed liver metastases.

Interestingly, a relatively high degree of correlation was found between the cytogenetic alterations detected by SNP-arrays and iFISH studies. Despite this, slight differences were noted between both techniques. On one hand, these were due to the lower

Table 5. Primary colorectal cancer with liver metastasis (n = 23): correlation between the numerical changes detected by each individual iFISH probe used and the CN changes identified for the corresponding single nucleotide polymorphisms (SNPs) through SNP array studies.

Chromosomal region identified by the iFISH probe	R^2/*P*-value
1p36	0.75/<0.001
1q25	0.75/<0.001
2p24	0.65/0.001
3q26	0.81/<0.001
5p15.2	0.65/0.001
6q23	0.67/<0.001
7q31	0.67/<0.001
8p22	0.81/<0.001
8q24	0.79/<0.001
9p21	0.91/<0.001
9q34	0.77/<0.001
10q23	0.68/<0.001
11q22	0.82/<0.001
12p13	0.76/<0.001
13q14	0.74/<0.001
13q34	0.78/<0.001
14q32	0.82/<0.001
15q22	0.72/<0.001
17p13	0.80/<0.001
18q21	0.75/<0.001
19q13	0.65/<0.001
20q13.2	0.80/<0.001
21q22	0.74/<0.001
22q11.2	0.83/<0.001

R^2: Coefficient of correlation.

doi:10.1371/journal.pone.0013752.t005

sensitivity of the SNP-array vs. iFISH for the identification of chromosomal abnormalities present in only a small proportion of all cells in the sample (i.e. secondary genetic lesions absent in the ancestral tumour cell clones) [28]. On the other hand, they were attributable to the increased sensitivity of the SNP-array vs. iFISH studies as regards identification of small interstitial changes [11]. In this regard, our results show occurrence of a high number of CN changes involving minimal/small regions (<1.3 Mb) and to a less extent, also extensive/large (>1.5 Mb) regions which frequently went undetectable by iFISH. Interestingly, several of these small and large altered regions contain cancer-associated genes known to be involved in CRC and/or the metastatic process: i.e. the *TPD52* [29], *FABP5* [30], *MAP2K4* [31], *LLGL1* [32], *FBLN1* [33] and *TYMP* [34] genes.

Among all human chromosomes, chromosomes 17 and 18 were those more frequently found to be altered in our series, their abnormalities typically consisting on extensive deletions involving the *TP53* and *DCC* genes, respectively, in addition to other tumor suppressor genes, such as *MAP2K4* at 17p12. A potential role for chromosome 18q in the development of CRC with associated liver metastases has been previously reported [35]; in this regard, decreased expression of Smad4 in addition to *DCC*, has been pointed out as a potential target protein coded in chromosome 18q since it is associated with both liver and lymph node metastases [36]. In line with these findings we also identified loss of the *SMAD4* gene in the great majority (83%) of the metastatic cases analyzed. By contrast, the most frequently (78% of cases) amplified region was found in chromosome 20, at 20q11.22. This is a relatively small region of 178,817 bp which harbors 8 known genes, half of which have been associated with CRC: *TNFRSF6B* [37], *OGFR* [38], *NTSR1* [39] and *CDH4* [40]. Among these genes, overexpression of *TNFRSF6B* -a gene that belongs to the tumor necrosis factor receptor (*TNFR*) super-family- has been reported in advanced stages of CRC [37] and other tumors of the gastrointestinal tract [41], in association with an increased resistance to adjuvant chemotherapy [42]; in turn, increased *NTSR1* expression has been reported as an early event in colon tumorigenesis that contributes to tumor progression and an aggressive clinical behavior [39]. Similarly, we also identified amplification and overexpression of the *MYC* gene at 8q24 in the great majority of the primary tumors, which have both been previously suggested to be involved in disease progression to a metastatic tumour [28;43].

From the clinical point of view, gain/amplification of 20p13 was associated with a higher frequency of well vs. moderately-differentiated tumours. Noteworthy, this chromosomal region contains genes which have been previously associated with disease progression. Accordingly, Miyoshi N *et al* have recently suggested that overexpression of the *TGM2* gene in CRC patients is associated with a shorter overall survival [44] and expression of the *PTPRA* gene has been recurrently associated with progression of gastric cancer, including lymphovascular invasion and liver/peritoneal dissemination [45,46].

Apart from defining the most frequently altered genes in metastatic CRC, this study was also aimed at detailed characterization of the most frequent recurrent breakpoint regions associated with such genetic changes. The number of different breakpoints detected within individual chromosomes is usually considered as a surrogate marker for chromosomal instability in cancer. In the present study, we found 245 different breakpoints for chromosome 1. This frequency is significantly higher than that reported by others using aCGH analyses of CRC without distant metastases: 16 different chromosomes breakpoints found, in a group of 32 patients [10]. These results suggest that advanced-stage and metastatic CRC could be associated with a greater number of breakpoints and higher chromosomal instability. In line with this hypothesis, Knutsen *et al* [21] found 407 chromosomal breakpoints in 15 CRC cell lines, using spectral karyotyping with a high frequency of recurrent breakpoints in the centromeric (p11 to q11) or pericentromeric (p11.2 and q11.2) regions of chromosomes 12, 13, 14, 15, 17 18 and 20. Interestingly, in this latter study Knutsen *et al* [21] also found recurrent breakpoints at 17p11.2 in 6/15 cell lines.

In the present study, a high percentage of cases showed recurrent breakpoints for chromosomes 1, 8, 17 and 20. Most interestingly, breakpoints at chromosome 17p were preferentially localized at the genome coordinate 20,156,497–22,975,771 bp at 17p12 (15/23 cases); in most of these cases (12/15 cases), the breakpoint was restricted to the genome coordinate (21,769,828–22,975,771 bp) which maps for the *FAM27L* gene, a gene whose function remains to be elucidated. Whether, disruption of the *FAM27L* gene may also play a role in the malignant transformation and/or the metastatic process of CRC into the liver in addition to, inactivation of *TP53* and inhibition of apoptosis [47,48], remains to be elucidated. Nevertheless, it should be noted that Camps *et al* [10] have shown a higher frequency of 17p11.2 breakpoints in CRC patients with positive (8/16) vs. negative (4/

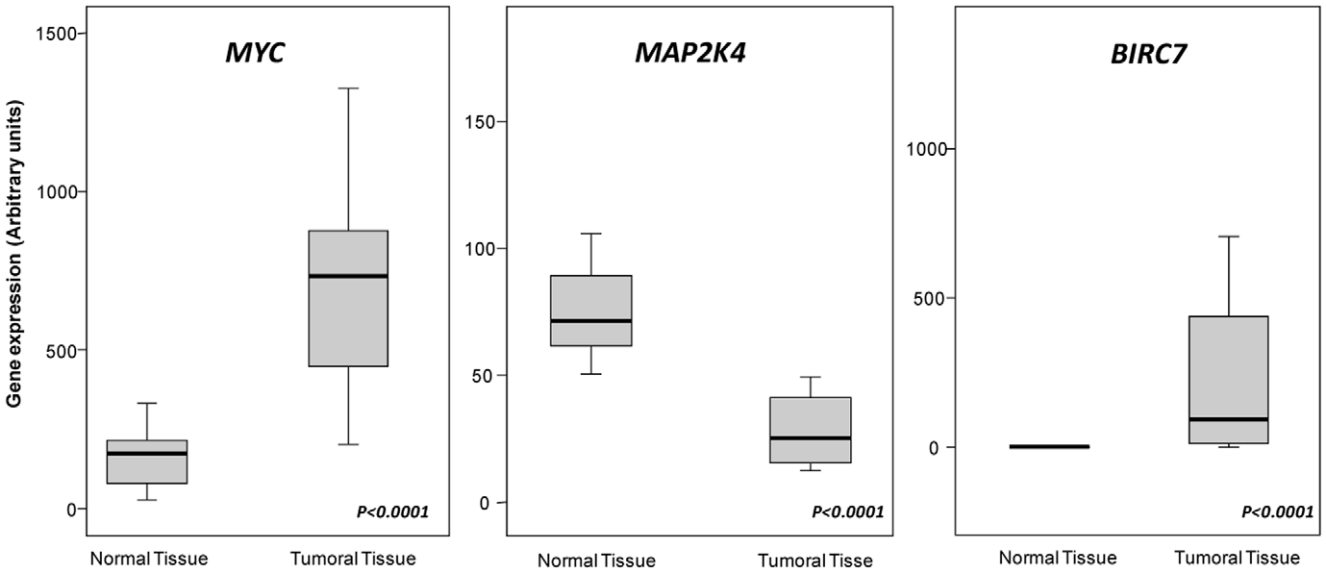


Figure 3. Expression levels of MYC, MAP4K and BIRC7 mRNA as assessed by RQ-PCR in metastatic CRC tumors and their corresponding paired normal tissue (n = 18). Note that MYC and BIRC7 mRNA levels from metastatic CRC tumours samples are significantly higher than in their paired normal tissues ($p<0.0001$). By contrast, MAP4K mRNA levels in metastatic CRC tumors are significantly lower than normal ($p<0.0001$).
doi:10.1371/journal.pone.0013752.g003

16) lymph nodes using aCGH. This breakpoint has been previously associated with an homogeneous genetic profile defined by a higher frequency of abnormalities of chromosomes 1p, 7, 8, 13q, 18q and 20q and an adverse clinical outcome [35,49–52]. Other recurrent chromosomal breakpoints found in our patients were localized in the 1p12, 8p12 and 20p12.1 chromosomal regions. Previous studies suggest that genes typically deregulated by these chromosome breaks included the *REG4* [53] and *NOTCH2* [54] genes at chromosome 1p12, *EIF4EBP1* [55] and *FGFR* [56] at chromosome 8p12, and the *FOXA2* [57] gene at chromosome 20p12; all these genes have been associated with the development and progression of CRC and the metastatic process in a variety of human cancers, including the development of liver metastases in CRC [53–57]. Additional GEP and functional studies as well as direct comparison of paired primary and metastatic tumours are required to validate our findings and to gain further insight into their role in metastatic CRC patients.

Supporting Information

Figure S1 Primary colorectal cancer with paired liver metastasis (n = 23): Identification of recurrent chromosomal breakpoint regions for the 1p12, 8p12, 17p11.2 and 20p12.1 chromosome regions as defined by the Affymetrix 500K SNP array genotyping platform. Breakpoints occurred in 9 cases (39%) at the 118097448-120939802 genome coordinate for chromosome 1 (panel A), in 8 cases (35%) at the 37770635-38405382 coordinate for chromosome 8 (panel B), in 15 cases (65%) at the 20156497-22975771 position for chromosome 17 (panel C) and in 9 cases (39%) at the 14921777- 16089156 genome coordinate for chromosome 20 (panel D).
Found at: doi:10.1371/journal.pone.0013752.s001 (4.70 MB TIF)

Table S1 Most frequently detected amplified regions (for >3 contiguous SNPs with average log2 copy number ratio >0.22) in primary colorectal tumours from metastatic CRC patients genotyped on the Affymetrix 500K SNP array platform (n = 23). Only recurrently amplified DNA copy-number regions found in at least half of the cases, are listed.
Found at: doi:10.1371/journal.pone.0013752.s002 (0.10 MB DOC)

Author Contributions

Conceived and designed the experiments: MGD AO. Performed the experiments: JMS MGG MES MdCC EG. Analyzed the data: JMS CF MGG MES MdCC JDLR LMB AO. Contributed reagents/materials/analysis tools: MdMA MES MdCC OB EF LMB. Wrote the paper: JMS AO.

References

1. Tsai MS, Su YH, Ho MC, Liang JT, Chen TP, et al. (2007) Clinicopathological features and prognosis in resectable synchronous and metachronous colorectal liver metastasis. Ann Surg Oncol 14: 786–94.
2. Macartney-Coxson DP, Hood KA, Shi HJ, Ward T, Wiles A, et al. (2008) Metastatic susceptibility locus, an 8p hot-spot for tumour progression disrupted in colorectal liver metastases: 13 candidate genes examined at the DNA, mRNA and protein level. BMC Cancer 8: 178–187.
3. Rigola MA, Casadevall C, Bernues M, Caballin MR, Fuster C, et al. (2002) Analysis of kidney tumors by comparative genomic hybridization and conventional cytogenetics. Cancer Genet Cytogenet 137: 49–53.
4. Garcia J, Duran A, Tabernero MD, Garcia PA, Flores CT, et al. (2003) Numerical abnormalities of chromosomes 17 and 18 in sporadic colorectal cancer: Incidence and correlation with clinical and biological findings and the prognosis of the disease. Cytometry B Clin Cytom 51: 14–20.
5. De Angelis PM, Clausen OP, Schjolberg A, Stokke T (1999) Chromosomal gains and losses in primary colorectal carcinomas detected by CGH and their associations with tumour DNA ploidy, genotypes and phenotypes. Br J Cancer 80: 526–35.
6. Lassmann S, Weis R, Makowiec F, Roth J, Danciu M, et al. (2007) 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 85: 293–304.
7. Hu XT, Chen W, Wang D, Shi QL, Zhang FB, et al. (2008) The proteasome subunit PSMA7 located on the 20q13 amplicon is overexpressed and associated with liver metastasis in colorectal cancer. Oncol Rep 19: 441–6.
8. Korn WM, Yasutake T, Kuo WL, Warren RS, Collins C, et al. (1999) Chromosome arm 20q gains and other genomic alterations in colorectal cancer metastatic to liver, as analyzed by comparative genomic hybridization

and fluorescence in situ hybridization. Genes Chromosomes Cancer 25: 82–90.
9. Tanaka T, Watanabe T, Kazama Y, Tanaka J, Kanazawa T, et al. (2006) Chromosome 18q deletion and Smad4 protein inactivation correlate with liver metastasis: A study matched for T- and N- classification. Br J Cancer 95: 1562–7.
10. Camps J, Grade M, Nguyen QT, Hormann P, Becker S, et al. (2008) Chromosomal breakpoints in primary colon cancer cluster at sites of structural variants in the genome. Cancer Res 68: 1284–95.
11. Walker BA, Morgan GJ (2006) Use of single nucleotide polymorphism-based mapping arrays to detect copy number changes and loss of heterozygosity in multiple myeloma. Clin Lymphoma Myeloma 7: 186–91.
12. World Health Organization. WHO International Histological Classification of Tumors, Vol 1-25. Geneva, 1967-1981; 2nd edn, Berlin: Springer-Verlag, 1988-.1992.
13. Greene FL (2007) Current TNM staging of colorectal cancer. Lancet Oncol 8: 572–3.
14. Vindelov LL, Christensen IJ, Nissen NI (1983) A detergent-trypsin method for the preparation of nuclei for flow cytometric DNA analysis. Cytometry 3: 323–327.
15. Bengtsson H, Irizarry R, Carvalho B, Speed TP (2008) Estimation and assessment of raw copy numbers at the single locus level. Bioinformatics 24: 759–67.
16. Venkatraman ES, Olshen AB (2007) A faster circular binary segmentation algorithm for the analysis of array CGH data. Bioinformatics 23: 657–663.
17. Habermann JK, Paulsen U, Roblick UJ, Upender MB, McShane LM, et al. (2007) Stage-specific alterations of the genome, transcriptome, and proteome during colorectal carcinogenesis. Genes Chromosomes Cancer 46: 10–26.
18. Ooi A, Huang CD, Mai M, Nakanishi I (1996) Numerical chromosome alterations in colorectal carcinomas detected by fluorescence in situ hybridization. Relationship to 17p and 18q allelic losses. Virchows Arch 428: 243–51.
19. Sayagues JM, Tabernero MD, Maillo A, Espinosa A, Rasillo A, et al. (2004) Intratumoral patterns of clonal evolution in meningiomas as defined by multicolor interphase fluorescence in situ hybridization (FISH): is there a relationship between histopathologically benign and atypical/anaplastic lesions? J Mol Diagn 6: 316–25.
20. Diep CB, Parada LA, Teixeira MR, Eknaes M, Nesland JM, et al. (2003) Genetic profiling of colorectal cancer liver metastases by combined comparative genomic hybridization and G-banding analysis. Genes Chromosomes Cancer 36: 189–97.
21. Knutsen T, Padilla-Nash HM, Wangsa D, Barenboim-Stapleton L, Camps J, et al. (2010) Definitive molecular cytogenetic characterization of 15 colorectal cancer cell lines. Genes Chromosomes Cancer 49: 204–23.
22. Sheffer M, Bacolod MD, Zuk O, Giardina SF, Pincas H, et al. (2009) Association of survival and disease progression with chromosomal instability: a genomic exploration of colorectal cancer. Proc Natl Acad Sci U S A 106: 7131–6.
23. Ghadimi BM, Grade M, Monkemeyer C, Kulle B, Gaedcke J, Gunawan B, et al. (2006) Distinct chromosomal profiles in metastasizing and non-metastasizing colorectal carcinomas. Cell Oncol 28: 273–81.
24. Al-Mulla F, AlFadhli S, Al-Hakim AH, Going JJ, Bitar MS (2006) Metastatic recurrence of early-stage colorectal cancer is linked to loss of heterozygosity on chromosomes 4 and 14q. J Clin Pathol 59: 624–30.
25. Paredes-Zaglul A, Kang JJ, Essig YP, Mao W, Irby R, et al. (1998) Analysis of colorectal cancer by comparative genomic hybridization: evidence for induction of the metastatic phenotype by loss of tumor suppressor genes. Clin Cancer Res 4: 879–86.
26. Hoglund M, Gisselsson D, Hansen GB, Sall T, Mitelman F, et al. (2002) Dissecting karyotypic patterns in colorectal tumors: two distinct but overlapping pathways in the adenoma-carcinoma transition. Cancer Res 62: 5939–46.
27. Diep CB, Kleivi K, Ribeiro FR, Teixeira MR, Lindgjaerde OC, et al. (2006) The order of genetic events associated with colorectal cancer progression inferred from meta-analysis of copy number changes. Genes Chromosomes Cancer 45: 31–41.
28. Sayagues JM, Abad MM, Barquero H, Gutierrez ML, Gónzalez-Gónzalez M, et al. (2010) Intratumoral cytogenetic heterogeneity of sporadic colorectal carcinomas suggests several pathways to liver metastasis. J Pathol 221: 308–319.
29. Payton LA, Lewis JD, Byrne JA, Bright RK (2008) Vaccination with metastasis-related tumor associated antigen TPD52 and CpG/ODN induces protective tumor immunity. Cancer Immunol Immunother 57: 799–811.
30. Pang J, Liu WP, Liu XP, Li LY, Fang YQ, et al. (2010) Profiling protein markers associated with lymph node metastasis in prostate cancer by DIGE-based proteomics analysis. J Proteome Res 9: 216–26.
31. Spillman MA, Lacy J, Murphy SK, Whitaker RS, Grace L, et al. (2007) Regulation of the metastasis suppressor gene MKK4 in ovarian cancer. Gynecol Oncol 105: 312–20.
32. Tsuruga T, Nakagawa S, Watanabe M, Takizawa S, Matsumoto Y, et al. (2007) Loss of Hugl-1 expression associates with lymph node metastasis in endometrial cancer. Oncol Res 16: 431–5.
33. Yang H, Rouse J, Lukes L, Lancaster M, Veenstra T, et al. (2004) Caffeine suppresses metastasis in a transgenic mouse model: a prototype molecule for prophylaxis of metastasis. Clin Exp Metastasis 21: 719–35.
34. Thean LF, Loi C, Ho KS, Koh PK, Eu KW, et al. (2010) Genome-wide scan identifies a copy number variable region at 3q26 that regulates PPM1L in APC mutation-negative familial colorectal cancer patients. Genes Chromosomes Cancer 49: 99–106.
35. Tanaka T, Watanabe T, Kitayama J, Kanazawa T, Kazama Y, et al. (2009) Chromosome 18q deletion as a novel molecular predictor for colorectal cancer with simultaneous hepatic metastasis. Diagn Mol Pathol 18: 219–25.
36. Tanaka T, Watanabe T, Kazama Y, Tanaka J, Kanazawa T, et al. (2008) Loss of Smad4 protein expression and 18q LOH as molecular markers indicating lymph node metastasis in colorectal cancer--a study matched for tumor depth and pathology. J Surg Oncol 97: 69–73.
37. Pitti RM, Marsters SA, Lawrence DA, Roy M, Kischkel FC, et al. (1998) Genomic amplification of a decoy receptor for Fas ligand in lung and colon cancer. Nature 396: 699–703.
38. Zagon IS, Donahue RN, McLaughlin PJ (2009) Opioid growth factor-opioid growth factor receptor axis is a physiological determinant of cell proliferation in diverse human cancers. Am J Physiol Regul Integr Comp Physiol 297: R1154–R1161.
39. Gui X, Guzman G, Dobner PR, Kadkol SS (2008) Increased neurotensin receptor-1 expression during progression of colonic adenocarcinoma. Peptides 29: 1609–15.
40. Miotto E, Sabbioni S, Veronese A, Calin GA, Gullini S, et al. (2004) Frequent aberrant methylation of the CDH4 gene promoter in human colorectal and gastric cancer. Cancer Res 64: 8156–9.
41. Bai C, Connolly B, Metzker ML, Hilliard CA, Liu X, et al. (2000) Overexpression of M68/DcR3 in human gastrointestinal tract tumors independent of gene amplification and its location in a four-gene cluster. Proc Natl Acad Sci U S A 97: 1230–5.
42. Mild G, Bachmann F, Boulay JL, Glatz K, Laffer U, et al. (2002) DCR3 locus is a predictive marker for 5-fluorouracil-based adjuvant chemotherapy in colorectal cancer. Int J Cancer 102: 254–7.
43. Camps J, Nguyen QT, Padilla-Nash HM, Knutsen T, McNeil NE, Wangsa D, et al. (2009) Integrative genomics reveals mechanisms of copy number alterations responsible for transcriptional deregulation in colorectal cancer. Genes Chromosomes Cancer 48: 1002–17.
44. Miyoshi N, Ishii H, Mimori K, Tanaka F, Hitora T, Tei M, et al. (2010) TGM2 is a novel marker for prognosis and therapeutic target in colorectal cancer. Ann Surg Oncol 17: 967–72.
45. Wu CW, Kao HL, Li AF, Chi CW, Lin WC (2006) Protein tyrosine-phosphatase expression profiling in gastric cancer tissues. Cancer Lett 242: 95–103.
46. Junnila S, Kokkola A, Karjalainen-Lindsberg ML, Puolakkainen P, Monni O (2010) Genome-wide gene copy number and expression analysis of primary gastric tumors and gastric cancer cell lines. BMC Cancer 10: 73.
47. Chen L, Jiang J, Cheng C, Yang A, He Q, et al. (2007) P53 dependent and independent apoptosis induced by lidamycin in human colorectal cancer cells. Cancer Biol Ther 6: 965–73.
48. Gemignani F, Moreno V, Landi S, Moullan N, Chabrier A, et al. (2004) A TP53 polymorphism is associated with increased risk of colorectal cancer and with reduced levels of TP53 mRNA. Oncogene 23: 1954–6.
49. Carvalho B, Postma C, Mongera S, Hopmans E, Diskin S, et al. (2009) Multiple putative oncogenes at the chromosome 20q amplicon contribute to colorectal adenoma to carcinoma progression. Gut 58: 79–89.
50. Ookawa K, Sakamoto M, Hirohashi S, Yoshida Y, Sugimura T, et al. (1993) Concordant p53 and DCC alterations and allelic losses on chromosomes 13q and 14q associated with liver metastases of colorectal carcinoma. Int J Cancer 53: 382–7.
51. Fijneman RJ, Carvalho B, Postma C, Mongera S, van Hinsbergh VW, et al. (2007) Loss of 1p36, gain of 8q24, and loss of 9q34 are associated with stroma percentage of colorectal cancer. Cancer Lett 258: 223–9.
52. Buffart TE, Coffa J, Hermsen MA, Carvalho B, van Dersijp IR, et al. (2005) DNA copy number changes at 8q11-24 in metastasized colorectal cancer. Cell Oncol 27: 57–65.
53. Oue N, Kuniyasu H, Noguchi T, Sentani K, Ito M, et al. (2007) Serum concentration of Reg IV in patients with colorectal cancer: overexpression and high serum levels of Reg IV are associated with liver metastasis. Oncology 72: 371–80.
54. Chu D, Zheng J, Wang W, Zhao Q, Li Y, et al. (2009) Notch2 expression is decreased in colorectal cancer and related to tumor differentiation status. Ann Surg Oncol 16: 3259–66.
55. Provenzani A, Fronza R, Loreni F, Pascale A, Amadio M, et al. (2006) Global alterations in mRNA polysomal recruitment in a cell model of colorectal cancer progression to metastasis. Carcinogenesis 27: 1323–33.
56. Sato T, Oshima T, Yoshihara K, Yamamoto N, Yamada R, et al. (2009) Overexpression of the fibroblast growth factor receptor-1 gene correlates with liver metastasis in colorectal cancer. Oncol Rep 21: 211–6.
57. Lehner F, Kulik U, Klempnauer J, Borlak J (2007) The hepatocyte nuclear factor 6 (HNF6) and FOXA2 are key regulators in colorectal liver metastases. FASEB J 21: 1445–62.

Modern Pathology (2012) 25, 590–601

Unique genetic profile of sporadic colorectal cancer liver metastasis *versus* primary tumors as defined by high-density single-nucleotide polymorphism arrays

Luís Muñoz-Bellvis[1], Celia Fontanillo[2], María González-González[3], Eva Garcia[4], Manuel Iglesias[1], Carmen Esteban[1], ML Gutierrez[3], MM Abad[5], Oscar Bengoechea[5], Javier De Las Rivas[2], Alberto Orfao[3,6] and JM Sayagués[3,6]

[1]*Unidad de Cirugía Hepatobiliopancreática, Departamento de Cirugía, Hospital Universitario de Salamanca, Salamanca, Spain;* [2]*Grupo de Investigación en Bioinformática y Genómica Funcional, Centro de Investigación del Cáncer (IBMCC-CSIC/USAL), Universidad de Salamanca, Salamanca, Spain;* [3]*Servicio General de Citometría, Departamento de Medicina and Centro de Investigación del Cáncer (IBMCC-CSIC/USAL), Universidad de Salamanca, Salamanca, Spain;* [4]*Unidad de Genómica y Proteómica, Centro de Investigación del Cáncer (IBMCC-CSIC/USAL), Universidad de Salamanca, Salamanca, Spain and* [5]*Departamento de Patología, Hospital Universitario de Salamanca, Salamanca, Spain*

Most genetic studies in colorectal carcinomas have focused on those abnormalities that are acquired by primary tumors, particularly in the transition from adenoma to carcinoma, whereas few studies have compared the genetic abnormalities of primary *versus* paired metastatic samples. In this study, we used high-density 500K single-nucleotide polymorphism arrays to map the overall genetic changes present in liver metastases ($n = 20$) from untreated colorectal carcinoma patients studied at diagnosis *versus* their paired primary tumors ($n = 20$). *MLH1, MSH2* and *MSH6* gene expression was measured in parallel by immunohistochemistry. Overall, metastatic tumors systematically contained those genetic abnormalities observed in the primary tumor sample from the same subject. However, liver metastases from many cases (up to 8 out of 20) showed acquisition of genetic aberrations that were not found in their paired primary tumors. These new metastatic aberrations mainly consisted of (1) an increased frequency of genetic lesions of chromosomes that have been associated with metastatic colorectal carcinoma (1p, 7p, 8q, 13q, 17p, 18q, 20q) and, more interestingly, (2) acquisition of new chromosomal abnormalities (eg, losses of chromosomes 4 and 10q and gains of chromosomes 5p and 6p). These genetic changes acquired by metastatic tumors may be associated with either the metastatic process and/or adaption of metastatic cells to the liver microenvironment. Further studies in larger series of patients are necessary to dissect the specific role of each of the altered genes and chromosomal regions in the metastatic spread of colorectal tumors.
Modern Pathology (2012) **25,** 590–601; doi:10.1038/modpathol.2011.195; published online 6 January 2012

Keywords: colorectal cancer; copy number change; FISH; liver metastases; SNP array

Correspondence: Professor A Orfao, MD, PhD, Centro de Investigación del Cancer, Paseo de la Universidad de Coimbra S/N, 37007 Salamanca, Spain.
E-mail: orfao@usal.es
[6]These authors contributed equally to this work and should be considered as senior last authors.
Received 3 August 2011; revised 17 October 2011; accepted 17 October 2011; published online 6 January 2012

Occurrence of distant metastasis in sporadic colorectal cancer (eg, liver metastasis) confers a poor prognosis. In fact, metastatic disease is the main cause of death in colorectal carcinoma patients, and the liver is the most common site for metastatic spread of colorectal carcinoma.[1,2] Current knowledge about the genetic pathways of clonal evolution

in colorectal carcinoma suggest that development of colorectal cancer could be triggered by the clonal expansion of cells that carry mutations, which most frequently involve the *APC, RAS, TP53* and/or *DCC* genes, and lead to a growth and/or survival advantage of tumor cells.[3] As metastatic cells derive from primary tumor cells, specific genomic alterations driving these ultimate steps of the metastatic cascade are expected to be acquired over the genomic profile of neoplastic cells from the primary tumor.[4] The genomic abnormalities, which are potentially characteristic of such advanced stages of the disease, are complex and so far, poorly described and partially understood. This relates to the fact that most genetic studies in colorectal cancer have focused on those abnormalities that are acquired in primary tumors, particularly, in the transition from adenoma to carcinoma, and few studies have compared these abnormalities with those observed in paired metastatic samples.[5–7] Despite this, multiple recurrent chromosomal abnormalities, which are found in primary tumors have been associated with metastatic colorectal carcinoma. Among others, these mainly include numerical changes such as gains of chromosomes 8q, 13q and 20q, and losses of the 1p, 8p, 17p and 18q chromosomal regions.[8–10] However, the molecular mechanisms underlying the association of such genetic profiles with metastatic colorectal carcinoma remain largely unknown.

Previous studies using conventional karyotyping,[5] comparative genomic hybridization (CGH),[5,7,11] fluorescence *in situ* hybridization (FISH)[9,11] or microsatellite markers to detect regions of loss of heterozygosity (LOH),[12] have largely failed in identifying recurrent chromosomal abnormalities acquired in metastatic *versus* primary colorectal tumors. This could be explained, at least in part, because of the relatively limited resolution of these techniques. More recently, the availability of high-density single-nucleotide polymorphism (SNP) arrays has facilitated the identification of small regions of chromosomal gains and losses because of its higher resolution (down to 2.5 kb),[13] and provides new opportunities in the identification of novel cancer genes involved in the metastatic process of colorectal cancer. However, previous reports in which high-density SNP arrays have been used to investigate the genetic profiles of colorectal carcinoma have specifically focused on primary tumor samples,[14] and to the best of our knowledge, no study has been reported so far in which high-density SNP arrays are employed to investigate the potential genetic differences between paired primary and metastatic tumors from colorectal carcinoma patients.

In the present study, we applied high-density (500 K) SNP mapping arrays—mean distance between the interrogated SNPs of 5.8 kb (median intermarker distance of 2.5 kb)—to map the overall genetic changes present in liver metastases from 20 untreated colorectal carcinoma patients studied at diagnosis *versus* their paired primary tumors ($n = 40$ samples). Our goal was to search for recurrent genetic differences between paired primary *versus* metastatic tumor samples that might contain candidate genes highly characteristic of metastatic liver disease.

Patients and methods

Patients and Samples

Tissue specimens from 20 sporadic colorectal adenocarcinomas and 20 paired liver metastases ($n = 40$ samples) were obtained from 20 patients (13 males and 7 females; median age of 70 years, ranging from 49 to 80 years) after informed consent had been given by each subject. It should be noted that only patients with metastatic lesions able to be resected were included in this cohort, which, therefore, is not representative of the whole colorectal cancer patient population. All patients underwent surgical resection of both tumor tissues at the Department of Surgery of the University Hospital of Salamanca (Salamanca, Spain). All tumors were diagnosed and classified according to the WHO criteria,[15] and they were all studied before any treatment was given. According to the tumor grade, 11 cases were classified as well-differentiated tumors, 8 as moderately- and one as poorly differentiated carcinomas. In all cases, histopathological grade was confirmed in a second independent evaluation by an experienced pathologist. Median follow-up at the moment of closing the study was of 37 months (range: 36–96 months). The study was approved by the local ethics committee of the University Hospital of Salamanca (Salamanca, Spain).

Seven primary tumors were localized in the rectum, and the other 13 were localized either in the right (cecum, ascending or transverse) or the left (descending and sigmoid) colon. The mean size of the primary tumors was of 5.3 ± 1.9 cm with the following distribution according to their TNM stage at diagnosis:[16] T3N0M0, two cases; T3N1M1, four cases; T3N1M0, four cases; T3N2M1, four cases; T4N0M1, one case; T4N0M0, three cases; T4N1M1, one case and; T4N2M1, one case. Liver metastases were identified either at the time of colorectal surgery ($n = 11$) or during the first year after initial diagnosis ($n = 9$); to date, patients have not shown any other metastasis. The mean size of the liver metastases was of 4.3 ± 2.2 cm.

After histopathological diagnosis was established, part of the primary tumor and its paired liver metastasis (both corresponding to a macroscopically tumoral region) were used to prepare single-cell suspensions. Once prepared, single-cell suspensions were resuspended in methanol/acetic acid (3/1; vol/vol) and stored at $-20\,^{\circ}C$ for further interphase FISH analyses, as recently described.[17] The remaining tissue was either fixed in formalin

and embedded in paraffin, or frozen in liquid nitrogen and stored at room temperature or at −80 °C, respectively. All tissues were evaluated after hematoxylin–eosin staining to confirm the presence of tumor cells and to evaluate their quantity in each individual sample. For SNP array studies, tumor DNA was extracted from representative areas of freshly frozen tumor tissues (primary tumors and liver metastases), which contained ≥65% epithelial tumor cells, localized mirror-cut to those used for iFISH analyses. In turn, normal DNA was extracted from peripheral blood leukocytes from the same patient. For the three types of samples (primary tumors, paired liver metastases and peripheral blood leukocytes), DNA was extracted using the QIAamp DNA mini kit (Qiagen, Hilden, Germany) following the instructions of the manufacturer.

SNP Array Studies

Each DNA sample derived from primary tumors and liver metastases and normal peripheral blood leukocytes was hybridized to two different 250 K Affymetrix SNP Mapping arrays (*Nsp*I and *Sty*I SNP arrays, Affymetrix, Santa Clara, CA, USA); for this purpose, 250 ng of DNA per array was used, according to the instructions of the manufacturer. Fluorescence signals were detected using the Affymetrix GeneChip Scanner 3000 (Affymetrix), and average genotyping call rates of 94.4, 91.5 and 97.3% were obtained for primary tumors, liver metastases and normal peripheral blood DNA samples, respectively.

To identify copy number changes throughout the whole tumor genome, the aroma.affymetrix algorithm was used, following the CRMA v2 method described elsewhere[18] (R-software package, http://www.aroma-project.org) and the following sequential steps: (i) calibration for crosstalks between pairs of allele probes; (ii) normalization for probe nucleotide-sequence effects; and (iii) normalization for PCR fragment-length and probe localization effects. Then, data from the 250K *Sty*I and 250K *Nsp*I arrays was integrated into a single database, and raw copy number values were calculated as transformed $\log_2$ values of the primary tumor/normal peripheral blood, liver metastasis/normal peripheral blood, liver metastases/primary tumor ratios calculated for each individual patient.

To identify DNA regions with similar copy number values, we used Circular Binary Segmentation as implemented in the DNAcopy Bioconductor package[19] with the default parameters; a *P*-value ≤0.01 for ≥5 markers per DNA segment was used to define points with changes. We used the smoothed value by assigning the median segment value to each probe. For the identification of altered (gained or lost) DNA regions, a threshold was established on the basis of the changes observed in the fluorescence intensity of sequential DNA segments for primary tumor *versus* peripheral blood, liver metastasis *versus* peripheral blood, and for liver metastases *versus* primary tumor samples, for each of the 20 patients studied. $\log_2$ ratio values >0.09 and <−0.09 were used as cut-off thresholds to define the presence of increased and decreased copy number values, respectively. High-level gains (DNA amplification) were defined as regions with a mean $\log_2$ copy number ratio ≥0.25. The specific frequencies of both copy number gains and losses per SNP were established and plotted along individual chromosomes for each tumor sample analyzed, for all individual cases studied. On the basis of the empirical frequency distribution of gains and losses among the 20 primary and the 20 metastatic tumor samples, respectively, we took the common altered regions grouping the contiguous SNPs with adjusted *P*-values <0.01 (false discovery rate correction, based on the Benjamini and Hochberg procedure).[20] Minimal common regions were defined as the smallest subset of SNPs in the altered regions with the highest frequency of gains and losses. At least five contiguous SNPs were required to define a region. Genes in these regions were identified using Ensembl release 53 (http://www.ensembl.org). The pattern of copy number changes of the primary tumors analyzed here has been previously reported in detail in a recent study.[21]

Interphase FISH Studies

To evaluate the reproducibility of the SNP array results and to assess background noise impact of this technique, interphase FISH analyses of the same tumor samples was performed in parallel, using 24 probes directed against an identical number of regions from 20 different human chromosomes, which are frequently altered in sporadic colorectal carcinomas. Overall, our results showed a high degree of correlation between both methods; this also holds true when such analysis was restricted to the most frequently altered regions, as previously described.[21]

Immunohistochemistry

One block of formalin-fixed paraffin wax-embedded adenocarcinoma tissue was selected in each case. In all cases, this block comprised an area of normal colonic mucosa adjacent to the tumor. Sections (4 μm) were affixed to Superfrost-plus slides (CML, Nemours, France) and dried overnight at 37 °C. Paraffin was removed and the tissue rehydrated using xylene and ethanol. Slides were subjected to microwave antigen retrieval in 10 mM citrate buffer (pH 6) at 85 °C for 35 min and cooled in phosphate-buffered saline, pH 7.4 (Sigma). Endogenous peroxidase activity was blocked with 2% hydrogen peroxide in methanol, and slides were washed with phosphate-buffered saline before overnight incubation with the appropriate antibody at a dilution of

1:100. Commercially available monoclonal antibodies against the nuclear proteins MLH1 (Clone G168-15; BD Biosciences, San Jose, CA, USA), MSH2 (Clone FE11; Biocare Medical, CA, USA) and MSH6 (Clone BC/44; Biocare Medical) were applied, followed by staining with Strept ABC complex/HRP Duet kit (DAKO, Copenhagen, Denmark) in conjunction with diamino benzedene 180 mg in 300 ml phosphate-buffered saline with 300 ml hydrogen peroxide. Sections were washed under running tap water and then lightly counterstained in Mayer's hematoxylin. Loss of expression was recorded when nuclear staining was absent from all malignant cells, but preserved in normal epithelial and stroma cells. Two observers assessed all cases independently.

Statistical Methods

For all continuous variables, mean values and their s.d. and range were calculated using the SPSS software package (SPSS 12.0. Chicago, IL USA); for dichotomic variables, frequencies were reported. To evaluate the statistical significance of differences observed between groups, the Mann–Whitney *U*-test and the χ^2-test were used for continuous and categorical variables, respectively (SPSS).

Results

Frequency and Chromosomal Localization of Copy Number Changes in Liver Metastasis from Colorectal Carcinoma

Overall, liver metastases from the 20 colorectal cancer patients analyzed systematically contained those chromosomal abnormalities that were identified in their paired primary colorectal carcinomas; (please note that the later have been previously described in detail for a larger series of patients).[21] Despite this, some aberrations were either newly acquired or more frequently found in liver metastases than in their paired primary tumors, which could reflect an increased genetic instability of neoplastic cells from metastatic *versus* primary tumor samples (Figure 1).

In detail, all liver metastases showed copy number changes in at least one chromosomal region. The highest frequency of copy number losses detected corresponded to chromosomes 1p ($n=16$; 80%), 17p ($n=18$; 90%) and 18q ($n=19$; 95%); in turn, copy number gains more frequently involved chromosomes 7p ($n=18$; 90%), 8q ($n=15$; 75%), 13q ($n=14$; 70%) and 20q ($n=15$; 75%) (see Supplementary Table S1). Interestingly, each of these regions has been previously found to contain

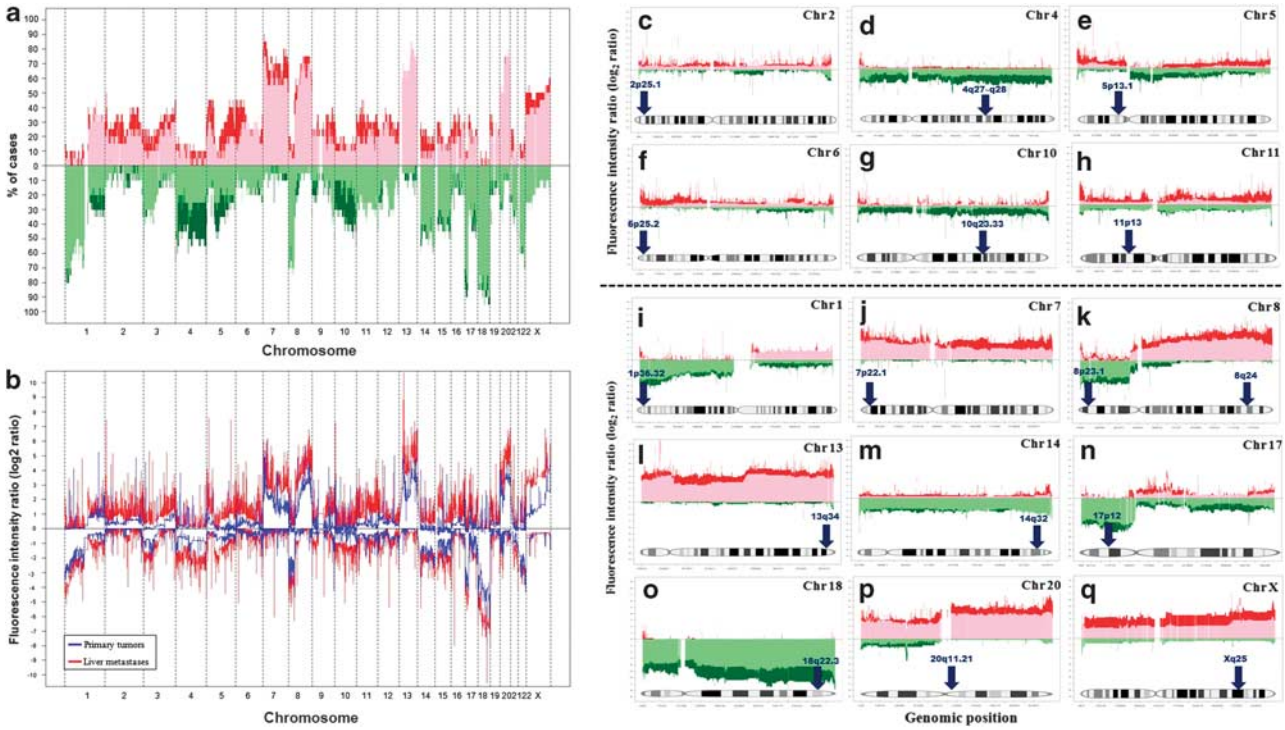


Figure 1 Metastatic colorectal cancer genome for the 20 colorectal carcinoma patients genotyped on the Affymetrix 500 K SNP array platform. A summary plot showing the frequency (**a** (in the left)) and fluorescence intensity $\log_2$ ratios (**b** (in the left)) of those copy number gains (plotted in red above zero values in the *x* axis) and losses (plotted in green below zero values in the *x* axis) identified in primary sporadic colorectal tumors (light colors), and their paired liver metastases (dark color) are displayed for the whole genome. The panels in the right show abnormalities identified in primary sporadic colorectal tumors (light colors) and their paired liver metastases (dark color) for individual chromosomes, which showed new abnormalities in metastatic *versus* primary tumor samples (**c–h**), or displayed an increased frequency of abnormalities in metastatic samples, which were already detected in primary tumors (**i–q**). Arrows point to regions of interest.

genes, which are altered/involved in colorectal cancer (eg, *ANGPT2, UBR5, KLF10, EIF3H, NOV, DCT, ABCC4, SLC15A1, EFNB2, IRS2, ING1, MAP2K4* (mitogen-activated protein kinase kinase-4)*, ID1, BCL2L1, MYLK2, CBFA2T2* and *E2F1*) and/or genes that are relevant to the metastatic process (eg, *ANGPT2, RRM2B, KLF10, RAD21, NOV, POU4F1, SPRY2, DCT, CLDN10, EFNB2, IRS2, COL4A2, ING1, MYH8, MAP2K4, ID1, BCL2L1, TPX2, MYLK2* and *E2F1*), in addition to genes associated with other malignancies (ie, *TRPS1, BTF3L1, DNAJC3, STK24, TM9SF2, LIG4, ARHGEF7, SCO1, MYOCD, GALR1, HCK* and *SMC1B*; Table 1). From them, the *ANGPT2, MAP2K4, E2F1, ID1* and *BCL2L1* genes have been reported to be involved in mechanisms that lead to increased cell proliferation and angiogenesis, and they have been found to be altered in both colorectal cancer and the metastatic events.

Frequency and Chromosomal Localization of High-Level Copy Number Gains in Liver Metastases from Colorectal Carcinomas

Chromosome 7 showed 17 regions, which displayed high level genetic amplification (Table 2) with mean $\log_2$ ratio fluorescence intensities of 0.28 (range: 0.25–0.36). These 17 regions were distributed along the whole chromosome 7 at the 7p22, 7p21, 7p15, 7p12, 7q22 and 7q36 chromosome bands, and they were all found to be altered (amplified) in $\geq$14/20 cases studied (70%; see Supplementary Table S2). These regions contain multiple genes, which have been recurrently associated with the pathogenesis of colorectal cancer and the metastatic process: *FSCN1, TWIST1, ITGB8, DFNA5, HOXA7, GRB10, EGFR, AZGP1, MCM7, EPHB4* and *MUC3A* (Table 2). In turn, for chromosome 8, only two regions of high-level genetic amplification (mean fluorescence intensities of 0.42 and 0.37, respectively) were detected; both regions were localized at the 8q24 chromosome band, and they involved the *MTSS1* and *ASAP1* genes (Table 2). Additionally, chromosomes 13 and 20 also displayed three regions (two in chromosome 13 and one in chromosome 20) with high level genetic amplification (mean fluorescence intensities of 0.45 and 0.43, respectively), containing genes potentially involved in the pathogenesis of colorectal cancer and the metastatic process, the *KLF5* and *IRS2* and the *MMP9* genes coded in the 13q22.1, 13q34 and 20q13.12 chromosomal regions, respectively (Table 2).

Acquired Chromosomal Abnormalities in Liver Metastases

In individual patients, primary tumors and their paired liver metastases frequently revealed the same chromosomal changes at both sites (Figure 1). However, liver metastases from 8 out of 20 cases showed acquisition of new genetic abnormalities that were not found in their paired primary tumors.

Table 1 Chromosomal regions, which most frequently displayed copy number alterations by SNP arrays, in colorectal liver metastases and that contain genes commonly associated with colorectal cancer and the metastatic process ($n = 20$)

Altered chromosomal regions (bp)	*Region length (bp)*	*Number of SNPs*	*Chromosome band*	*Event*	*Altered cases (%)*	*Cancer-associated genes*
Chr8: 6 319 564–6 393 980	74 416	41	8p23.1	Deletion	70	***ANGPT2***[a]
Chr8: 102 281 574–104 598 943	2 317 369	449	8q22.3	Gain	75	***RRM2B***, *UBR5*[a], ***KLF10***,
Chr8: 116 722 193–118 020 419	1 298 226	222	8q23.3	Gain	75	*TRPS1*, *EIF3H*[a], ***RAD21***
Chr8: 120 491 103–120 508 049	16 946	8	8q24.12	Gain	75	***NOV***[a]
Chr13: 75 649 333–83 553 225	7 903 892	1469	13q22.3	Gain	70	*BTF3L1, POU4F1, SPRY2*
Chr13: 93 770 996–94 062 605	291 609	72	13q31.3	Gain	70	***DCT***[a]
Chr13: 94 410 370–94 792 167	381 797	118	13q31.3	Gain	70	*ABCC4*[a]
Chr13: 94 803 129–95 581 085	777 956	80	13q31.3	Gain	70	***CLDN10,*** *DNAJC3*
Chr13: 96 635 106–99 063 992	2 428 886	530	13q32.1	Gain	70	*STK24*, *SLC15A1*[a], *TM9SF2*
Chr13: 105 576 623–106 900 640	1 324 017	327	13q33.2	Gain	70	***EFNB2***[a]
Chr13: 107 631 285–107 678 245	46 960	15	13q33.3	Gain	70	LIG4
Chr13: 109 092 251–110 808 910	1 716 659	406	13q34	Gain	70	***IRS2, COL4A2, ING1***[a], *ARHGEF7*
Chr17: 10 134 845–11 066 755	931 910	178	17p13.1	Deletion	90	***MYH8***, *SCO1*
Chr17: 11 124 244–12 787 020	1 662 776	312	17p12	Deletion	90	***MAP2K4***[a], *MYOCD*
Chr18: 71 151 155–73 487 286	2 336 131	645	18q22.3	Deletion	95	*GALR1*
Chr20: 29 314 247–30 386 296	1 072 049	103	20q11.21	Gain	75	***ID1***, ***BCL2L1***[a], ***TPX2***, *HCK, MYLK2*[a]
Chr20: 31 377 143–32 096 987	719 844	81	20q11.21	Gain	75	*CBFA2T2*[a], ***E2F1***[a]
Chr22: 44 114 817–44 172 947	58 130	16	22q13.2	Deletion	70	*SMC1B*

[a]Genes that have been described to be involved/altered in colorectal cancer, and genes that have been commonly associated with the metastatic process are shown in bold italics.

Table 2 Chromosomal regions, which most frequently displayed high-level genetic amplification by SNP arrays, and which contained genes commonly involved/altered in colorectal cancer and/or associated with the metastatic process ($n=20$)

Amplified chromosomal regions (bp)[a]	*Chromosomal band*	*Mean log_2 ratio*	*Maximum log_2 ratio*	*Altered cases (%)*	*Cancer-associated genes*
Chr7: 2 888 108–2 965 622	7p22.2	0.26	1.40	80	*CARD11*
Chr7: 4 624 574–5 634 592	7p22.1	0.26	1.13	85	***FSCN1***[b]
Chr7: 19 099 597–19 164 657	7p21.1	0.29	1.33	80	***TWIST1***[b]
Chr7: 20 315 252–20 348 199	7p15.3	0.26	0.93	80	***ITGB8***[b]
Chr7: 24 288 850–24 294 445	7p15.3	0.28	1.68	80	*NPY*[b]
Chr7: 24 761 544–24 764 289	7p15.3	0.27	1.45	80	***DFNA5***[b]
Chr7: 27 110 282–27 278 326	7p15.2	0.36	1.67	70	*HOXA5, HOXA7*[b], *HOXA9, HOXA11, HOXA13*
Chr7: 28 169 667–28 211 202	7p15.2	0.28	1.19	70	*JAZF1*
Chr7: 50 163 751–50 752 627	7p12.2	0.27	1.93	75	*IKZF1, DDC*
Chr7: 50 797 965–50 839 403	7p12.2	0.32	1.41	75	*SLC4A2, FASTK*
Chr7: 50 797 965–50 839 403	7p12.2	0.33	1.42	75	***GRB10***[b]
Chr7: 54 954 150–56 213 585	7p11.2	0.27	1.10	75	***EGFR***[b], *PSPH*
Chr7: 99 301 754–101 811 250	7q22.1	0.25	0.70	70	***AZGP1***[b], ***MCM7***[b], ***CUX1***, ***EPHB4***[b], ***MUC3A***[b], *MUC12*[b]
Chr7: 105 711 183–105 715 751	7q22.2	0.25	0.88	70	*PBEF1*
Chr7: 154 677 722–155 005 086	7q36.3	0.27	0.91	70	EN2
Chr7: 156 357 544–156 630 253	7q36.3	0.28	0.63	70	*MNX1, UBE3C*
Chr7: 156 893 472–158 147 850	7q36.3	0.29	1.34	70	*PTPRN2*
Chr8: 125 800 442–125 834 484	8q24.13	0.42	1.19	85	MTSS1
Chr8: 131 064 043–131 191 826	8q24.21	0.37	0.98	70	***ASAP1***[b]
Chr13: 72 497 695–72 659 497	13q22.1	0.41	0.73	70	***KLF5***[b]
Chr13: 79 810 102–79 825 947	13q31.1	0.46	1.39	70	SPRY2
Chr13: 90 792 026–90 811 945	13q31.3	0.47	1.50	70	*MIRHG1*
Chr13: 98 007 816–98 035 844	13q32.2	0.43	1.49	70	*STK24*
Chr13: 109 205 907–109 255 030	13q34	0.44	1.20	70	***IRS2***[b]
Chr13: 109 743 976–109 764 350	13q34	0.47	1.47	70	*COL4A1, COL4A2*
Chr13: 110 549 062–110 578 598	13q34	0.45	1.79	70	*ARHGEF7*
Chr20: 41 247 578–41 278 159	20q12	0.41	1.10	70	*PTPRT*
Chr20: 44 007 866–44 178 129	20q13.12	0.41	0.88	70	***MMP9***[b]
Chr20: 59 241 454–59 268 793	20q13.33	0.47	1.8	70	*CDH4*[b]

[a]Only those regions, which were recurrently amplified in at least 14 out of 20 cases analyzed (>70%) are listed.
[b]Genes that have been described to be involved/altered colorectal cancer, and genes that have been commonly associated with metastatic processes are shown in bold italics.
High-level genetic amplification was defined *versus* those with an average log_2 copy number ratio ≥ 0.25.

These new metastatic aberrations included copy number gains at chromosomes 2p, 5p, 6p, 7q and 11p, together with copy number losses of chromosomes 4, 5q and 10q (Table 3). The specific abnormalities, which were recurrently detected in 8 out of 20 colorectal carcinomas metastasis for those chromosomal regions that showed a normal diploid profile in their corresponding (paired) primary tumors, are shown in Figure 2. As illustrated, these metastatic abnormalities involved chromosomal regions which harbor i) tumor suppressor genes that have a key role in the metastatic process (eg, the *ANXA5*, *CCNA2*, *IL2* and *IL21* genes at chromosome 4q27; the *PLK4, IL15, GAB1, HHIP* and *SMAD1* genes coded at the 4q28.1 chromosome regions and the *PTEN* gene coded at the 10q23.33 chromosomal region) and; (ii) oncogenes (eg, the *PTGER4* and *PRKAA1* genes coded at chromosome 5p13.1, and both the *RIPK1* and *NQO2* genes coded at chromosome 6p25.2); copy number gains of the former two oncogenes have been associated with advanced colorectal carcinoma. Many other genetic aberrations were present in liver metastases from colorectal carcinoma analyzed, but at lower frequencies (Figure 1).

Correlation Between the Chromosomal Changes Detected by Interphase FISH and SNP Array Studies

Overall, the chromosomal abnormalities identified by interphase FISH in liver metastases showed profiles similar to those found by SNP array studies, also when such analysis was restricted to the most frequently altered regions. Thus, gains/amplification at 7q were detected in 60% of the cases by interphase FISH *versus* 70% by SNP array studies ($r^2=0.67$; $P<0.001$); similarly, gains/amplification of chromosomes 8q (found in 70% of cases by interphase FISH *vs* 75% by SNP array studies;

Table 3 Metastatic colorectal cancer genome for the 20 colorectal cancer patients genotyped on the Affymetrix 500 K SNP array platform: chromosomal abnormalities identified exclusively in liver metastases (and not in their paired primary tumors), which involved chromosomal regions that contain genes commonly associated with cancer and/or the metastatic process

Altered chromosomal regions (bp)	*Chromosome band*	*Event*	*Number of altered cases*[a]	*Mean log_2 ratio in liver metastases*	*Cancer-associated genes*
Chr2: 11 301 969–11 420 624	2p25.1	Gain	7	1.13	—
Chr4: 11 217 752–12 280 681	4p15.33	Deletion	7	−23.76	—
Chr4: 12 402 361–14 483 843	4p15.33	Deletion	7	−53.43	—-
Chr4: 14 537 247–16 743 438	4p15.33	Deletion	7	−48.17	***BST1***[b], ***FGFBP1***[b], ***PROM1***[b]
Chr4: 95 174 065–97 996 020	4q22.3	Deletion	7	−58.56	***BMPR1B***[b]
Chr4: 98 021 823–101 566 164	4q22.3	Deletion	7	−55.19	***EIF4E***[b], *ADH5*, ***MTTP***[b]
Chr4: 122 461 059–126 659 227	4q27	Deletion	8	−61.43	***ANXA5***[b], ***CCNA2***[b], ***IL2***[b], ***IL21***[b], *NUDT6*[b], *FAT4*[b]
Chr4: 126 668 154–147 205 681	4q28.1	Deletion	8	−324.84	***PLK4***[b], ***SLC7A11***, *NARG1*, *SETD7*, ***IL15***[b], *INPP4B*, ***GAB1***[b], ***SMARCA5***, ***HHIP***[b], ***SMAD1***[b]
Chr4: 147 218 521–148 654 896	4q31.22	Deletion	7	−26.13	*EDNRA, POU4F2*, *LSM6*
Chr4: 148 658 165–150 007 154	4q31.23	Deletion	7	−24.35	***EDNRA***, *ARHGAP10*, ***NR3C2***[b]
Chr4: 151 081 476–169 411 644	4q31.3	Deletion	7	−317.79	*LRBA*, *MAB21L2*[b], ***FBXW7***[b], ***SFRP2***[b], ***ANXA10***[b], ***LRAT***, ***PDGFC***[b], *PPID*[b], ***CPE***[b]
Chr4: 171 433 995–173 576 567	4q33	Deletion	7	−37.39	-
Chr4: 175 432 179–176 373 292	4q34.1	Deletion	8	−17.93	*HPGD*[b]
Chr4: 176 393 349–181 031 240	4q34.2	Deletion	7	−93.92	***VEGFC***[b]
Chr4: 181 036 139–183 912 221	4q34.3	Deletion	8	−81.56	-
Chr4: 183 917 599–185 620 740	4q35.1	Deletion	7	−28.21	***IRF2***[b], ***DCTD***[b], ***ING2***[b]
Chr4: 188 066 517–189 969 188	4q35.2	Deletion	8	−31.94	*ZFP42*
Chr5: 31 602 359–32 770 165	5p13.1	Gain	7	17.03	*NPR3*, ***PDZD2***
Chr5: 40 710 736–40 901 620	5p13.1	Gain	8	1.55	***PTGER4***[b], ***PRKAA1***[b]
Chr5: 42 953 413–43 243 995	5p12	Gain	7	3.87	—
Chr5: 43 501 986–43 901 209	5p12	Gain	7	2.68	—
Chr5: 58 935 951–59 201 448	5q12.1	Deletion	6	−6.29	—
Chr5: 141 299 822–141 552 149	5q31.3	Gain	7	2.14	*RNF14*[b]
Chr6: 1 310 265–1 608 630	6p25.3	Gain	7	7.89	—
Chr6: 2 713 210–3 252 280	6p25.2	Gain	8	9.74	***RIPK1***[b], ***NQO2***[b], *SERPINB1*, *SERPINB6*[b], *SERPINB9*
Chr6: 3 391 923–3 411 021	6p25.2	Gain	8	0.98	—
Chr6: 3 663 243–3 673 023	6p25.2	Gain	8	0.90	—
Chr6: 4 026 261–4 342 058	6p25.2	Gain	7	5.45	—
Chr6: 4 840 846–5 220 281	6p25.1	Gain	7	7.36	—
Chr6: 5 493 557–5 966 877	6p25.1	Gain	7	7.00	—
Chr6: 6 476 360–6 654 484	6p25.1	Gain	7	4.22	—
Chr6: 7 647 658–7 683 958	6p24.3	Gain	8	1.18	***BMP6***[b]
Chr6: 10 495 977–10 532 295	6p24.3	Gain	7	0.96	***TFAP2A***[b]
Chr6: 13 593 062–13 758 950	6p23	Gain	7	2.08	—
Chr6: 13 857 081–14 361 384	6p23	Gain	7	6.40	—
Chr6: 15 143 396–15 884 421	6p23	Gain	7	8.47	***CD83***[b]
Chr6: 16 085 720–16 448 755	6p22.3	Gain	7	4.71	—
Chr6: 17 210 737–18 442 309	6p22.3	Gain	7	12.44	*NRN1*, ***DEK***, *BPHL*, ***RIPK1***[b]
Chr7: 158 430 322–158 640 662	7q36.3	Gain	7	3.79	—
Chr10: 71 077 186–71 425 507	10q13.2	Deletion	6	−10.26	—
Chr10: 83 990 316–84 564 355	10q23.1	Deletion	6	−9.53	—
Chr10: 89 574 656–89 676 489	10q23.2	Deletion	7	−1.89	***PTEN***[b]
Chr10: 97 903 413–100 323 090	10q23.33	Deletion	6	−21.38	*BLNK*, *DNTT*, ***FRAT1***[b], ***LOXL4***[b], ***PGAM1***[b], ***SFRP5***[b]
Chr11: 32 104 370–34 401 072	11p13	Gain	6	17.60	***LMO2***, *HIPK3*, ***WT1***[b]
Chr11: 43 653 493–45 414 563	11p11.2	Gain	6	19.81	***CD82***[b], ***EXT2***[b], *ALKBH3*

[a]Number of cases with chromosomal abnormalities identified exclusively in liver metastases and not in their paired primary tumors.
[b]Genes that have been described to be altered/involved in colorectal cancer, and genes that have been commonly associated with the metastatic process are shown in bold italics.

$r^2 = 0.79$; $P < 0.001$), 13q (80 *vs* 70%; $r^2 = 0.78$; $P < 0.001$) and 20q (80 *vs* 75%; $r^2 = 0.80$; $P < 0.001$), as well as deletions of chromosomes 8p (65 *vs* 70%; $r^2 = 0.81$; $P < 0.001$), 17p (75 *vs* 90%; $r^2 = 0.64$; $p = 0.02$) and 18q (75 *vs* 95%; $r^2 = 0.63$; $p = 0.03$) were detected at similar frequencies with both methods.

Microsatellite Status

All primary tumors examined ($n = 20$) showed a normal expression of the MLH1, MSH2 and MSH6 mismatch repair proteins in the nucleus and adjacent non-neoplastic tissue elements.

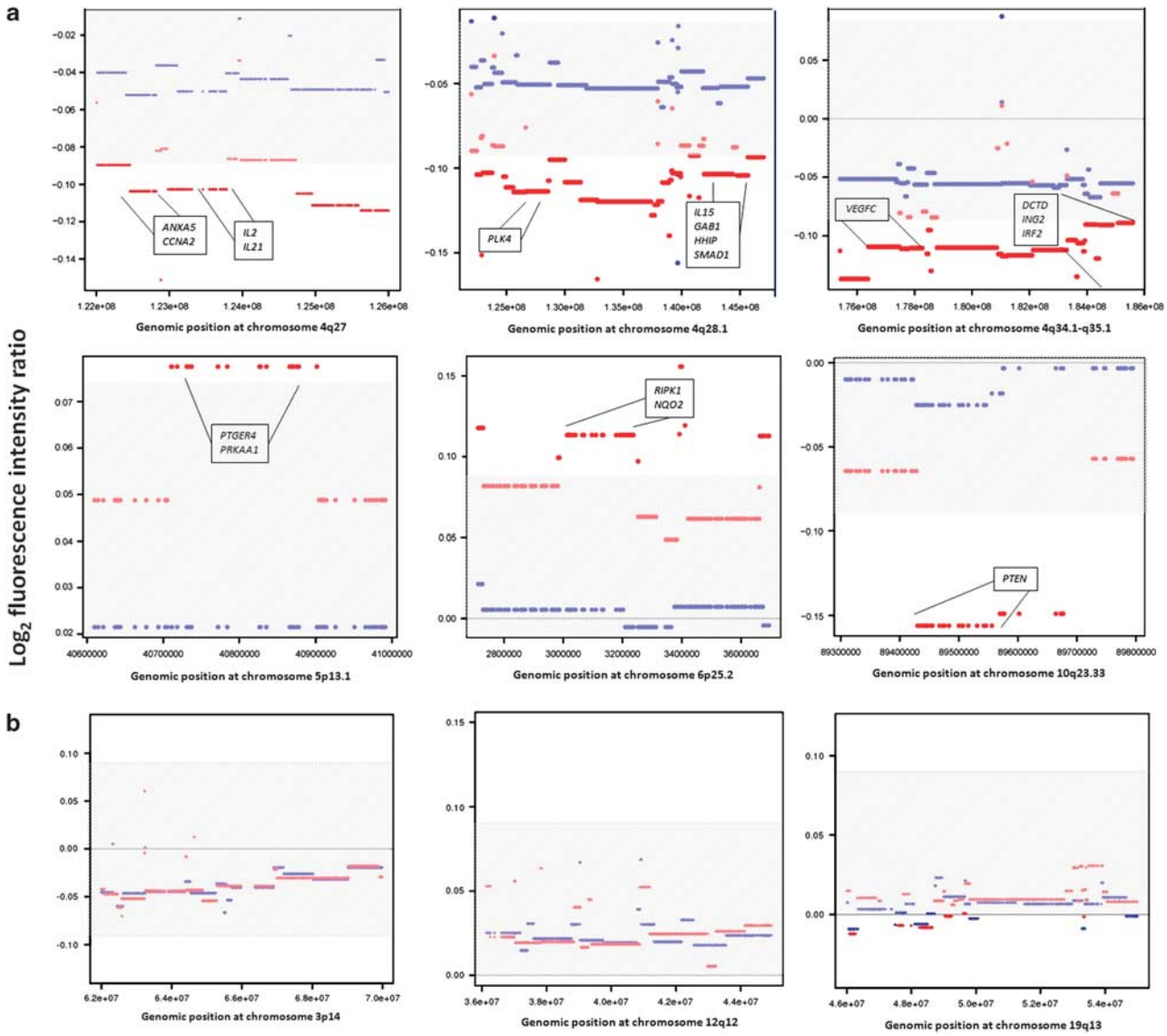


Figure 2 Metastatic colorectal cancer genome for the 20 colorectal carcinoma patients genotyped on the Affymetrix 500 K SNP array platform; copy number changes detected in liver metastases ($n = 8/20$ cases; red color) *versus* their paired primary tumors (blue color) for the 4q27, 4q28.1, 4q34.1, 5p13.1, 6p25.2 and 10q23.33 chromosomal regions (**a**). Genes contained in the newly altered chromosomal regions are listed in italics capital letters. Log_2 ratios >0.09 and <-0.09 (shown as colored background) were used as cut-off thresholds to define the presence of increased and decreased copy number values, respectively. All copy number changes detected between liver metastases *versus* their paired primary tumors showed statistically significant differences ($P<0.001$). As an example, the 3p14, 12q12 and 19q13 chromosomal regions, which did not show any differences between paired primary and metastatic lesions, are also shown (**b**).

Discussion

This study focused on the genetic characterization of liver metastases that occur in the context of primary colorectal carcinoma. To the best of our knowledge, this is the first study that compares the genetic abnormalities found in liver metastases *versus* paired primary colorectal tumors, in which high-resolution 500 K SNP arrays have been systematically used. Overall, primary tumors and their paired metastases from individual patients frequently revealed many common chromosomal changes at both sites; these findings support the existence of a close genetic relationship between primary colorectal tumors and their paired liver metastases, as previously suggested.[17] Genetic changes observed in common in both groups of samples included gains of chromosomes 7, 8q, 13q and 20 and losses of the 1p, 4, 8p, 17p, 18 and 22q chromosomes with normal expression of the MLH1, MSH2 and MSH6 mismatch repair proteins. In line with other studies, all our metastatic tumors showed a higher frequency of these chromosomal abnormalities than primary colorectal tumors,[7,22] and some of these abnormalities, together with deletions on chromosome 15q, have been associated with disease progression.[23] Previous studies in which the genetic abnormalities of colorectal carcinoma have been

investigated by conventional cytogenetics,[24] FISH,[17] CGH,[25] array CGH[26] and low-resolution 50 K SNP arrays[23] have also found that most of these genetic abnormalities are recurrently identified in primary tumors from metastatic colorectal carcinoma. On the basis of the high frequency of these chromosomal abnormalities in both primary and metastatic samples, it could be hypothesized that they reflect a metastatic genetic profile of colorectal carcinoma that could be of great clinical utility for the identification of colorectal carcinoma patients at higher risk of developing liver metastases, already at diagnosis.

Interestingly, high-level genetic amplification was found at specific regions of chromosomes 7, 8, 13 and 20; overall, 43 genes commonly involved/altered in colorectal cancer and/or associated with the metastatic process are coded in these regions. Of note, 17 of these 43 genes have been associated with progression of hepatocellular carcinomas.[27–29] Altogether, these findings could suggest that these genetic abnormalities that are acquired by metastatic colorectal carcinoma cells in the liver could be associated with homing and/or adaptation to the liver microenvironment. Among these genes, overexpression of *TWIST1* has been demonstrated to induce angiogenesis; at the same time, it has been associated with both the development of metastasis in hepatocellular carcinomas[30] and an unfavorable outcome in colorectal carcinoma patients;[31] in turn, increased expression of *IRS2*—commonly found in human hepatocellular carcinoma specimens and hepatoma cell lines—[32] has been associated with colon tumorigenesis, in which it contributes to tumor progression and an aggressive clinical behavior.[33] In line with this hypothesis, preliminary studies on genomic differences detected in primary colorectal carcinomas *versus* paired brain metastases have described a genetic profile consisting of gains of 8q, 12p, 12q, 20p, and loss of 5q in brain metastasis,[34] which is clearly different from that observed in our liver metastases. The different genetic signature associated with liver *versus* brain metastases could mirror the unique adaptation process of metastatic tumor cells to each specific microenvironment. Despite these findings, those three chromosomal regions, which showed the highest levels of amplification, were 13q31.3, 13q34 and 20q13.33, where four known cancer genes (*MIRHG1, COL4A1, COL4A2* and *CDH4*) are coded. To the best of our knowledge, no specific association between amplification of these genes and colorectal carcinoma has been reported so far; conversely, deregulation of these genes has been associated with neuroblastoma,[35] esophageal squamous cell carcinoma[36] and glioblastoma multiforme.[37]

In the present study, we also show the existence of recurrent genetic changes between paired primary and metastatic colorectal tumor cells. Such changes mainly consist of (1) an increased frequency of genetic lesions of chromosomes that have been associated with metastatic colorectal carcinoma (1p, 7p, 8q, 13q, 17p, 18q, 20q) and, more interestingly, (2) acquisition of new chromosomal abnormalities (eg, losses of chromosomes 4 and 10q and gains of chromosomes 5p and 6p). Interestingly, the former abnormalities involved chromosomal regions that encode for up to 11 genes, which have been previously found to be involved in the metastatic process of colorectal carcinoma. As an example, the *ANGPT2* gene (localized in chromosome 8 at 8p23.1) is known to be involved in angiogenic processes and has been previously associated with an invasive/malignant potential;[38] in turn, the *E2F1* gene (20q11.21) has been shown to have a crucial role in the control of cell cycle through down-regulation of tumor suppressor proteins.[39] Similarly, the *ID1* and *BCL2L1* genes (both coded in the same chromosomal region at 20q11.21) are also known to have a role in cell growth, senescence and differentiation, and the carcinogenesis of human colorectal carcinoma,[40] whereas overexpression of the Id-1 protein has been associated with tumor progression in colorectal carcinoma.[41] In turn, Paredes *et al*[11] have recently described that losses of chromosome 17p in metastatic colorectal cancer samples cover larger regions than in primary tumors, suggesting that additional unknown suppressor genes could be present at 17p, in the newly deleted sequences. In line with these findings, we have recurrently identified loss of the *MAP2K4* gene at 17p12 in the great majority of the metastatic samples analyzed. The *MAP2K4* gene is a member of the stress-activated protein kinase signaling cascade involved in the regulation of multiple cellular processes, which among other associations, has been recently suggested to have a functional role as a metastasis-suppressor gene in several malignant tumors, for example, human prostatic cancer,[42] ovarian cancer,[43] as well as breast and pancreatic tumors.[44] Similarly, a potential role for 18q LOH in the development of colorectal cancer with associated liver metastases has been suggested,[10] as well as its potential independent prognostic value,[45] which may depend on the microsatellite instability status.[46] In this regard, chromosomal instability has been associated in colorectal cancer with a worse prognosis, and different groups of tumors have been defined on the basis of the chromosomal instability status.[47] Herein, we identified loss of the 18q22–q23 chromosomal region in the great majority (95%) of the metastatic samples analyzed; interestingly, no clear association could be found between 18q LOH and the microsatellite instability status, because of normal expression of the *MLH1, MSH2* and *MSH6* mismatch repair genes, and potentially, also the relatively limited number of cases studied. Similarly, the sample size and the presence of multiple structural and/or numerical chromosome changes in all liver metastases analyzed precludes the study of chromosomal instability subtypes.

Many patients also showed acquisition of new genetic aberrations, which were not detected in their paired primary tumors. These included gains of chromosomes 2, 5p, 6p, 7q and 11p, and losses of chromosomes 4 and 10q. These results suggest that these chromosomal regions may also have a relevant role in the metastatic process as supported by the fact that some of them—for example, del(4p15.33), del(4q22.3), del(4q27), del(4q28.1), del(4q31), del(4q35.1) and del(10q23)—are known to contain multiple tumor suppressor genes (eg, *PLK4* at 4q28.1, *SFRP2* at 4q31.3, *IRF2* at 4q35.1 and *PTEN* at 10q23.2)[48–51] and genes that are involved in the metastatic process.[52–54] In line with these findings, previous studies in which primary colorectal carcinomas were compared with liver metastases also reported a greater frequency of chromosome 4 losses in late *versus* early stages of the disease.[9,55] However, due to the limited sensitivity of the SNP array technique for the detection of small clones that could already be present in primary tumors, further studies in which such abnormalities are investigated at the single-cell level are required to confirm our findings.

In summary, here we show the existence of relevant genetic differences between paired primary and metastatic colorectal tumors, which mainly consist of (1) an increased frequency of genetic lesions of chromosomes that have been associated with metastatic colorectal cancer (1p, 7p, 8q, 13q, 17p, 18q, 20q) and, more interestingly, (2) acquisition of new chromosomal abnormalities (eg, losses of chromosomes 4 and 10q and gains of chromosomes 5p and 6p). These genetic changes acquired by metastatic tumors may be associated with either the metastatic process and/or adaption of metastatic cells to the liver microenvironment. Further studies in larger series of patients, in which cases with non-resecable liver metastasis are also analyzed, are necessary to dissect the specific role of each of the altered genes and chromosomal regions in the metastatic spread of colorectal tumors. Additional gene expression profile studies are required to validate the proteins associated with copy number alterations in the metastasis *versus* the primary tumor.

Acknowledgements

This work has been partially supported by grants from the Consejería de Sanidad, Junta de Castilla y León, Valladolid, Spain (SAN191/SA09/06, SAN673/SA39/08 and SAN/103/2011), Fundación Memoria de Don Samuel Solórzano Barruso, Salamanca, Spain, Caja de Burgos (Obra Social), Burgos, Spain, Grupo Excelencia de Castilla y León (GR37) and the RTICC from the Instituto de Salud Carlos III (ISCIII), Ministerio de Sanidad y Consumo, Madrid, Spain (RD06/0020/0035-FEDER). JM Sayagués and M González are supported by grants (CP05/00321 and FI08/00721, respectively) from the ISCIII, Ministerio de Ciencia e Innovación, Madrid, Spain.

Disclosure/conflict of interest

The authors declare no conflict of interest.

References

1 Tsai HL, Lu CY, Hsieh JS, *et al.* The prognostic significance of total lymph node harvest in patients with T2-4N0M0 colorectal cancer. J Gastrointest Surg 2007;11:660–665.
2 Macartney-Coxson DP, Hood KA, Shi HJ, *et al.* Metastatic susceptibility locus, an 8p hot-spot for tumour progression disrupted in colorectal liver metastases: 13 candidate genes examined at the DNA, mRNA and protein level. BMC Cancer 2008;8:187.
3 Sugai T, Habano W, Nakamura S, *et al.* Allelic losses of 17p, 5q, and 18q loci in diploid and aneuploid populations of multiploid colorectal carcinomas. Hum Pathol 2000;31:925–930.
4 Zeitoun G, Mourra N, Blanche-Koch H, *et al.* Genomic profile of colon cancer metastases. Anticancer Res 2008;28:3609–3612.
5 Diep CB, Parada LA, Teixeira MR, *et al.* Genetic profiling of colorectal cancer liver metastases by combined comparative genomic hybridization and G-banding analysis. Genes Chromosomes Cancer 2003; 36:189–197.
6 Diep CB, Kleivi K, Ribeiro FR, *et al.* The order of genetic events associated with colorectal cancer progression inferred from meta-analysis of copy number changes. Genes Chromosomes Cancer 2006;45:31–41.
7 Al Mulla F, Keith WN, Pickford IR, *et al.* Comparative genomic hybridization analysis of primary colorectal carcinomas and their synchronous metastases. Genes Chromosomes Cancer 1999;24:306–314.
8 Hu XT, Chen W, Wang D, *et al.* The proteasome subunit PSMA7 located on the 20q13 amplicon is overexpressed and associated with liver metastasis in colorectal cancer. Oncol Rep 2008;19:441–446.
9 Korn WM, Yasutake T, Kuo WL, *et al.* Chromosome arm 20q gains and other genomic alterations in colorectal cancer metastatic to liver, as analyzed by comparative genomic hybridization and fluorescence *in situ* hybridization. Genes Chromosomes Cancer 1999;25:82–90.
10 Tanaka T, Watanabe T, Kazama Y, *et al.* Chromosome 18q deletion and Smad4 protein inactivation correlate with liver metastasis: A study matched for T- and N-classification. Br J Cancer 2006;95:1562–1567.
11 Paredes-Zaglul A, Kang JJ, Essig YP, *et al.* Analysis of colorectal cancer by comparative genomic hybridization: evidence for induction of the metastatic phenotype by loss of tumor suppressor genes. Clin Cancer Res 1998;4:879–886.
12 Blaker H, Graf M, Rieker RJ, *et al.* Comparison of losses of heterozygosity and replication errors in primary colorectal carcinomas and corresponding liver metastases. J Pathol 1999;188:258–262.
13 Walker BA, Morgan GJ. Use of single nucleotide polymorphism-based mapping arrays to detect copy

number changes and loss of heterozygosity in multiple myeloma. Clin Lymphoma Myeloma 2006;7:186–191.
14 Camps J, Grade M, Nguyen QT, *et al.* Chromosomal breakpoints in primary colon cancer cluster at sites of structural variants in the genome. Cancer Res 2008;68:1284–1295.
15 World Health Organization. WHO International Histological Classification of Tumors, Vol 1–25. Geneva, 1967–1981 2nd edn. Springer-Verlag: Berlin, 1988–1992.
16 Greene FL. Current TNM staging of colorectal cancer. Lancet Oncol 2007;8:572–573.
17 Sayagues JM, Abad MM, Barquero H, *et al.* Intratumoral cytogenetic heterogeneity of sporadic colorectal carcinomas suggests several pathways to liver metastasis. J Pathol 2010;221:308–319.
18 Bengtsson H, Irizarry R, Carvalho B, *et al.* Estimation and assessment of raw copy numbers at the single locus level. Bioinformatics 2008;24:759–767.
19 Venkatraman ES, Olshen AB. A faster circular binary segmentation algorithm for the analysis of array CGH data. Bioinformatics 2007;23:657–663.
20 Benjamini Y, Hochberg Y. On the adaptive control of the False Discovery Rate in multiple testing with independent statistics. JEBS 2000;25:60–83.
21 Sayagues JM, Fontanillo C, Abad MM, *et al.* Mapping of genetic abnormalities of primary tumours from metastatic CRC by high-resolution SNP arrays. PLoS ONE 2010;29:e13752.
22 Diep CB, Teixeira MR, Thorstensen L, *et al.* Genome characteristics of primary carcinomas, local recurrences, carcinomatoses, and liver metastases from colorectal cancer patients. Mol Cancer 2004;3:6.
23 Sheffer M, Bacolod MD, Zuk O, *et al.* Association of survival and disease progression with chromosomal instability: a genomic exploration of colorectal cancer. Proc Natl Acad Sci USA 2009;106:7131–7136.
24 Rigola MA, Casadevall C, Bernues M, *et al.* Analysis of kidney tumors by comparative genomic hybridization and conventional cytogenetics. Cancer Genet Cytogenet 2002;137:49–53.
25 De Angelis PM, Clausen OP, Schjolberg A, *et al.* Chromosomal gains and losses in primary colorectal carcinomas detected by CGH and their associations with tumour DNA ploidy, genotypes and phenotypes. Br J Cancer 1999;80:526–535.
26 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 2007;85:293–304.
27 Kanai M, Hamada J, Takada M, *et al.* Aberrant expressions of HOX genes in colorectal and hepatocellular carcinomas. Oncol Rep 2010;23:843–851.
28 Ma S, Guan XY, Lee TK, *et al.* Clinicopathological significance of missing in metastasis B expression in hepatocellular carcinoma. Hum Pathol 2007;38:1201–1206.
29 Fong CW, Chua MS, McKie AB, *et al.* Sprouty 2, an inhibitor of mitogen-activated protein kinase signaling, is down-regulated in hepatocellular carcinoma. Cancer Res 2006;66:2048–2058.
30 Niu RF, Zhang L, Xi GM, *et al.* Up-regulation of Twist induces angiogenesis and correlates with metastasis in hepatocellular carcinoma. J Exp Clin Cancer Res 2007;26:385–394.
31 Okada T, Suehiro Y, Ueno K, *et al.* TWIST1 hypermethylation is observed frequently in colorectal tumors and its overexpression is associated with unfavorable outcomes in patients with colorectal cancer. Genes Chromosomes Cancer 2010;49:452–462.
32 Boissan M, Beurel E, Wendum D, *et al.* Overexpression of insulin receptor substrate-2 in human and murine hepatocellular carcinoma. Am J Pathol 2005;167:869–877.
33 Slattery ML, Samowitz W, Curtin K, *et al.* Associations among IRS1, IRS2, IGF1, and IGFBP3 genetic polymorphisms and colorectal cancer. Cancer Epidemiol Biomarkers Prev 2004;13:1206–1214.
34 Gutenberg A, Gerdes JS, Jung K, *et al.* High chromosomal instability in brain metastases of colorectal carcinoma. Cancer Genet Cytogenet 2010;198:47–51.
35 Wei JS, Johansson P, Chen QR, *et al.* microRNA profiling identifies cancer-specific and prognostic signatures in pediatric malignancies. Clin Cancer Res 2009;15:5560–5568.
36 Chattopadhyay I, Singh A, Phukan R, *et al.* Genome-wide analysis of chromosomal alterations in patients with esophageal squamous cell carcinoma exposed to tobacco and betel quid from high-risk area in India. Mutat Res 2010;696:130–138.
37 Ruano Y, Mollejo M, Ribalta T, *et al.* Identification of novel candidate target genes in amplicons of Glioblastoma multiforme tumors detected by expression and CGH microarray profiling. Mol Cancer 2006;5:39.
38 Ochiumi T, Tanaka S, Oka S, *et al.* Clinical significance of angiopoietin-2 expression at the deepest invasive tumor site of advanced colorectal carcinoma. Int J Oncol 2004;24:539–547.
39 Iwamoto M, Banerjee D, Menon LG, *et al.* Overexpression of E2F-1 in lung and liver metastases of human colon cancer is associated with gene amplification. Cancer Biol Ther 2004;3:395–399.
40 Zhang YL, Pang LQ, Wu Y, *et al.* Significance of Bcl-xL in human colon carcinoma. World J Gastroenterol 2008;14:3069–3073.
41 Zhao ZR, Zhang ZY, Zhang H, *et al.* Overexpression of Id-1 protein is a marker in colorectal cancer progression. Oncol Rep 2008;19:419–424.
42 Kim HL, Vander Griend DJ, Yang X, *et al.* Mitogen-activated protein kinase kinase 4 metastasis suppressor gene expression is inversely related to histological pattern in advancing human prostatic cancers. Cancer Res 2001;61:2833–2837.
43 Yamada SD, Hickson JA, Hrobowski Y, *et al.* Mitogen-activated protein kinase kinase 4 (MKK4) acts as a metastasis suppressor gene in human ovarian carcinoma. Cancer Res 2002;62:6717–6723.
44 Wang L, Pan Y, Dai JL. Evidence of MKK4 prooncogenic activity in breast and pancreatic tumors. Oncogene 2004;23:5978–5985.
45 Watanabe T, Wu TT, Catalano PJ, *et al.* Molecular predictors of survival after adjuvant chemotherapy for colon cancer. N Engl J Med 2001;344:1196–1206.
46 Ogino S, Nosho K, Irahara N, *et al.* Prognostic significance and molecular associations of 18q loss of heterozygosity: a cohort study of microsatellite stable colorectal cancers. J Clin Oncol 2009;27:4591–4598.
47 Walter A, Houlston R, Tomlinson I. Association between chromosomal instability and prognosis in colorectal cancer: a meta-analysis. Gut 2008;57:941–950.

48 Baselga J. The EGFR as a target for anticancer therapy–focus on cetuximab. Eur J Cancer 2001;4: S16–S22.
49 Ko MA, Rosario CO, Hudson JW, *et al.* Plk4 haploinsufficiency causes mitotic infidelity and carcinogenesis. Nat Genet 2005;37:883–888.
50 Wang Y, Liu DP, Chen PP, *et al.* Involvement of IFN regulatory factor (IRF)-1 and IRF-2 in the formation and progression of human esophageal cancers. Cancer Res 2007;67:2535–2543.
51 Veeck J, Noetzel E, Bektas N, *et al.* Promoter hypermethylation of the SFRP2 gene is a high-frequent alteration and tumor-specific epigenetic marker in human breast cancer. Mol Cancer 2008;7:83.
52 Karoui M, Tresallet C, Julie C, *et al.* Loss of heterozygosity on 10q and mutational status of PTEN and BMPR1A in colorectal primary tumours and metastases. Br J Cancer 2004;90:1230–1234.
53 Berghella AM, Contasta I, Pellegrini P, *et al.* Peripheral blood immunological parameters for use as markers of pre-invasive to invasive colorectal cancer. Cancer Biother Radiopharm 2002;17:43–50.
54 Bessard A, Sole V, Bouchaud G, *et al.* High antitumor activity of RLI, an interleukin-15 (IL-15)-IL-15 receptor alpha fusion protein, in metastatic melanoma and colorectal cancer. Mol Cancer Ther 2009;8:2736–2745.
55 Ried T, Just KE, Holtgreve-Grez H, *et al.* Comparative genomic hybridization of formalin-fixed, paraffin-embedded breast tumors reveals different patterns of chromosomal gains and losses in fibroadenomas and diploid and aneuploid carcinomas. Cancer Res 1995; 55:5415–5423.

Supplementary Information accompanies the paper on Modern Pathology website (http://www.nature.com/modpathol)

OPEN ACCESS Freely available online

PLOS ONE

Prognostic Impact of del(17p) and del(22q) as Assessed by Interphase FISH in Sporadic Colorectal Carcinomas

María González-González[1], Luís Muñoz-Bellvis[2], Carlos Mackintosh[3], Celia Fontanillo[4], M. Laura Gutiérrez[1], M. Mar Abad[5], Oscar Bengoechea[5], Cristina Teodosio[1], Emilio Fonseca[6], Manuel Fuentes[1], Javier De Las Rivas[4], Alberto Orfao[1*9], José María Sayagués[1*9]

1 Servicio General de Citometría, Departamento de Medicina and Centro de Investigación del Cáncer (IBMCC-CSIC/USAL), Hospital Universitario de Salamanca-IBSAL, Universidad de Salamanca, Salamanca, Spain, **2** Unidad de Cirugía Hepatobiliopancreática, Departamento de Cirugía, Hospital Universitario de Salamanca-IBSAL, Salamanca, Spain, **3** Centro de Investigación del Cáncer (IBMCC-CSIC/USAL), Universidad de Salamanca, Salamanca, Spain, **4** Grupo de Investigación en Bioinformática y Genómica Funcional, Centro de Investigación del Cáncer (IBMCC-CSIC/USAL), Universidad de Salamanca, Salamanca, Spain, **5** Departamento de Patología, Hospital Universitario de Salamanca-IIBSAL, Salamanca, Spain, **6** Departamento de Oncología Médica, Hospital Universitario de Salamanca-IBSAL, Salamanca, Spain

Abstract

Background: Most sporadic colorectal cancer (sCRC) deaths are caused by metastatic dissemination of the primary tumor. New advances in genetic profiling of sCRC suggest that the primary tumor may contain a cell population with metastatic potential. Here we compare the cytogenetic profile of primary tumors from liver metastatic versus non-metastatic sCRC.

Methodology/Principal Findings: We prospectively analyzed the frequency of numerical/structural abnormalities of chromosomes 1, 7, 8, 13, 14, 17, 18, 20, and 22 by iFISH in 58 sCRC patients: thirty-one non-metastatic (54%) *vs.* 27 metastatic (46%) disease. From a total of 18 probes, significant differences emerged only for the 17p11.2 and 22q11.2 chromosomal regions. Patients with liver metastatic sCRC showed an increased frequency of del(17p11.2) (10% *vs.* 67%;$p<.001$) and del(22q11.2) (0% *vs.* 22%;$p=.02$) versusnon-metastatic cases. Multivariate analysis of prognostic factors for overall survival (OS) showed that the only clinical and cytogenetic parameters that had an independent adverse impact on patient outcome were the presence of del(17p) with a 17p11.2 breakpoint and del(22q11.2). Based on these two cytogenetic variables, patients were classified into three groups: low- (no adverse features), intermediate- (one adverse feature) and high-risk (two adverse features)- with significantly different OS rates at 5-years ($p<.001$): 92%, 53% and 0%, respectively.

Conclusions/Significance: Our results unravel the potential implication of del(17p11.2) in sCRC patients with liver metastasis as this cytogenetic alteration appears to be intrinsically related to an increased metastatic potential and a poor outcome, providing additional prognostic information to that associated with other cytogenetic alterations such as del(22q11.2). Additional prospective studies in larger series of patients would be required to confirm the clinical utility of the new prognostic markers identified.

Citation: González-González M, Muñoz-Bellvis L, Mackintosh C, Fontanillo C, Gutiérrez ML, et al. (2012) Prognostic Impact of del(17p) and del(22q) as Assessed by Interphase FISH in Sporadic Colorectal Carcinomas. PLoS ONE 7(8): e42683. doi:10.1371/journal.pone.0042683

Editor: Hassan Ashktorab, Howard University, United States of America

Received December 22, 2011; **Accepted** July 11, 2012; **Published** August 17, 2012

Funding: This work has been partially supported by grants from the Consejería de Sanidad, Junta de Castilla y León, Valladolid, Spain (SAN673/SA39/08 and SAN/103/2011), Fundación Memoria de Don Samuel Solórzano Barruso, Salamanca, Spain, Caja de Burgos (Obra Social), Burgos, Spain, Grupo Excelencia de Castilla y León (GR37) and the Red Temática de Investigación Cooperativa en Cáncer (RTICC) from the Instituto de Salud Carlos III (ISCIII), Ministerio de Sanidad y Consumo, Madrid, Spain (PI12/02053-FIS and RD06/0020/0035-FEDER). JM Sayagués and M González are supported by grants (CP05/00321 and FI08/00721, respectively) from the ISCIII, Ministerio de Ciencia e Innovación, Madrid, Spain.

Competing Interests: The authors have the following competing interests: The study was partly funded by Caja de Burgos - Obra Social. This does not alter the authors' adherence to all the PLoS ONE policies on sharing data and materials, as detailed online in the guide for authors.

* E-mail: orfao@usal.es (AO); ppmari@usal.es (JMS)

9 These authors contributed equally to this work.

Introduction

Metastatic dissemination of the primary tumor is the major cause of death of sporadic colorectal cancer (sCRC) patients [1]. Metastasis is a complex multi-step process which is driven by sequential accumulation of multiple genetic and molecular alterations and/or epigenetic changes involving one or multiple tumor cell clones. In recent years, data accumulated about the intratumoral pathways of clonal evolution of sCRC associated with chromosomal alterations/instability, indicates that liver metastatic lesions may derive from descendants of a tumor cell clone which is already present in the primary tumor [2]. Advances in genetic profiling of cancer also suggest that the metastatic potential of human tumors is encoded in the bulk of a primary tumor, as metastatic tumors systematically contain those genetic abnormalities observed in the primary tumor sample from the same subject. However, the precise molecular changes associated with the development of sCRC with liver metastasis still remain to be identified [2]. Multiple recurrent chromosomal abnormalities that are found in primary tumours have been associated with

metastatic CRC, including gains of chromosomes 8q, 13q and 20q and losses of the 1p, 8p, 17p, 18q and 22q chromosomal regions [3–5].

In a recent study, we described a detailed map of the genetic abnormalities of primary tumors from sCRC patients with liver metastasis by high-resolution SNP arrays. In this study, we reported the existence of a highly prevalent breakpoint region in the great majority of primary sCRC patients who had synchronous liver metastasis. Such breakpoint region is located in the centromeric region of chromosome 17p, between the genome coordinates 20,156,497 bp and 22,975,771 bp [6]. This breakpoint region has been previously associated with i) a homogeneous genetic profile consisting of a higher frequency of abnormalities of chromosomes 1p, 7, 8, 13q, 17p, 18q, 20q and 22q and ii) an adverse clinical outcome [7]. However, delineation of the minimal common breakpoint region at chromosome 17p11.2 and its potential prognostic value in sCRC tumors, remain to be fully defined.

In the present study we investigated the prognostic value of structural/numerical abnormalities of the most frequently altered chromosomes in liver metastatic colorectal carcinomas from 58 sCRC patients (27 liver metastatic *vs.* 31 non-metastatic tumors) with a long median follow-up, as detected by interphase fluorescence *in situ* hybridization (iFISH). Overall, our results show that the occurrence of del(17p) involving the 17p11.2 breakpoint region is an independent prognostic factor for overall survival, as confirmed in a larger series of 119 patients from the GEO public database. However, we have demonstrated that the combined assessment of del(22q11) and del(17p11.2) increased the predictive value for a liver metastatic tumor.

Materials and Methods

Patients and samples

In the present study, we prospectively analyzed surgical specimens from 58 patients diagnosed with a sCRC between 1999 and 2010 (38 males and 20 females; median age of 69 years, ranging from 38 to 83 years) after informed consent was given by each subject. All patients underwent surgical resection of primary tumor tissues at the Department of Surgery of the University Hospital of Salamanca (Salamanca, Spain) and they were diagnosed and classified according to the WHO criteria [8] prior to any treatment was given. Fourteen primary tumors were localized in the rectum and the other 44 were localized either in the right (caecum, ascending or trasverse) or the left (descending and sigmoid) colon, with an overall mean size of 5.3±2 cm. According to tumor grade, 39 cases were classified as well-differentiated tumors, 15 as moderately- and four as poorly-differentiated carcinomas. In all cases, histopathological grade was confirmed in a second independent evaluation by an experienced pathologist. Median follow-up at the moment of closing this study was of 96 months (range: 12–124 months). The study was approved by the local ethics committee of the University Hospital of Salamanca (Salamanca, Spain) and informed consent was given by each individual, prior to entering the study.

From the 58 cases analyzed, 27 (47%) tumors had liver metastases (group 1; median follow-up of 37 months; pT3–4 pN1–2 M1) identified either at time of colorectal surgery (n = 16) or during the first year after initial diagnosis (n = 11); they all underwent complete surgical resection of both their primary and metastatic CRC. The other 31 (53%) patients corresponded to non-metastatic sCRC selected on the basis of a long follow-up in the absence of liver metastasis (median follow-up of 99 months; pT2–4 pN0 M0) to ensure their non-metastatic nature (group 2).

After histopathological diagnosis was established, part of the primary tumor was used to prepare single-cell suspensions. Once prepared, single cell suspensions were resuspended in methanol/acetic (3/1; vol/vol) and stored at −20°C for further iFISH analyses, as described elsewhere [2]. The remaining tissue was either fixed in formalin and embedded in paraffin, or frozen in liquid nitrogen and stored at room temperature (RT) and at −80°C, respectively. Each individual tissue sample was also evaluated after haematoxylin-eosin staining, to confirm the presence of tumor cells and to evaluate their quantity.

Interphase fluorescence *in situ* hybridization (iFISH) studies

Mixed single-cell suspensions from different samples obtained from each tumor were used for iFISH studies, after fixation in 3/1 methanol/acetic (vol/vol). A set of 18 different probes (Vysis Inc, Downers Grove, IL) specific for those chromosomes and chromosomal regions most frequently gained/amplified and deleted colorectal carcinomas with liver metastases [6], were systematically used in double and triple staining with the Spectrum Orange (SO), Spectrum Green (SG) and Spectrum Aqua (SA) fluorochromes: for chromosome 1, the LSI p58 (1p36) (SO)/TelVysion 1p (SG)/LSI 1q25 (SA) Multi-color probe was employed; for chromosome 7, the LSI D7S486 (7q31) (SO)/CEP 7 (SG) Dual Color probe was used; for chromosome 8, the LSI LPL (8p22) (SO)/CEP 8 (SA)/MYC (8q24) (SG) Multi-color probe was employed; for chromosome 13, the LSI RB1 13q14 (SO)/LSI 13q34 LAMP1 (SG) was used; for chromosome 14, the LSI IGH (14q32.33) Dual Color, Break Apart probe was selected; for chromosome 17, the LSI TP53 (17p13) (SO)/CEP 17 (SA) probe combination was employed; for chromosome 18, the LSI BCL2 (18q21) (SO)/CEP 18 (SA) probe combination was used; for chromosome 20, the LSI ZNF217 (20q13.2) (SO)/CEP 20 (SG) probes were employed, and; for chromosome 22, the LSI BCR (22q11.2) probe was used. We have previously found in primary tumors [6] and their paired liver metastases [9] a high prevalence of gains of chromosomes 7, 8q, 11q, 13q, 20q and X together with losses of the 1p, 8p, 17p and 18q chromosomal regions; in this series of cases, the breakpoints found at the centromeric region of chromosome 17p were variable and were mapped between the genomic coordinates 20,156,497 bp and 22,975,771 bp by SNP's arrays. Herein, we investigated the presence of breakpoints at chromosome 17p11.2 using iFISH probes specifically designed and manufactured for this purpose (Kreatech Diagnostics, Amsterdam, The Netherlands), as schematically described in Figure 1.

The specific methods and procedures used for the iFISH studies have been previously described in detail [2] and for the investigation of the relationships existing between those genes coded at the 17p11.2, 17p13.1 and 22q11.2 chromosomal regions and other genes, the Ingenuity Pathway Analysis software (Ingenuit System®,www.ingenuity.com) was used.

External validation of the prognostic impact of del(17p) and del(22q)

External validation of the prognostic impact of del(17p) and del(22q) was performed in a previously reported group of sCRC patients from which aCGH files (MHP Human 1 Mb) and clinical data were publicly available at the GEO database (accession number GSE12520; genomic markers that predict survivorship in colorectal cancer) [10]. From all cases available in the dataset, we selected those studied with the MHP Human 1 Mb CGH array platform for a total of 109 cases: 81 sCRC from Edinburgh

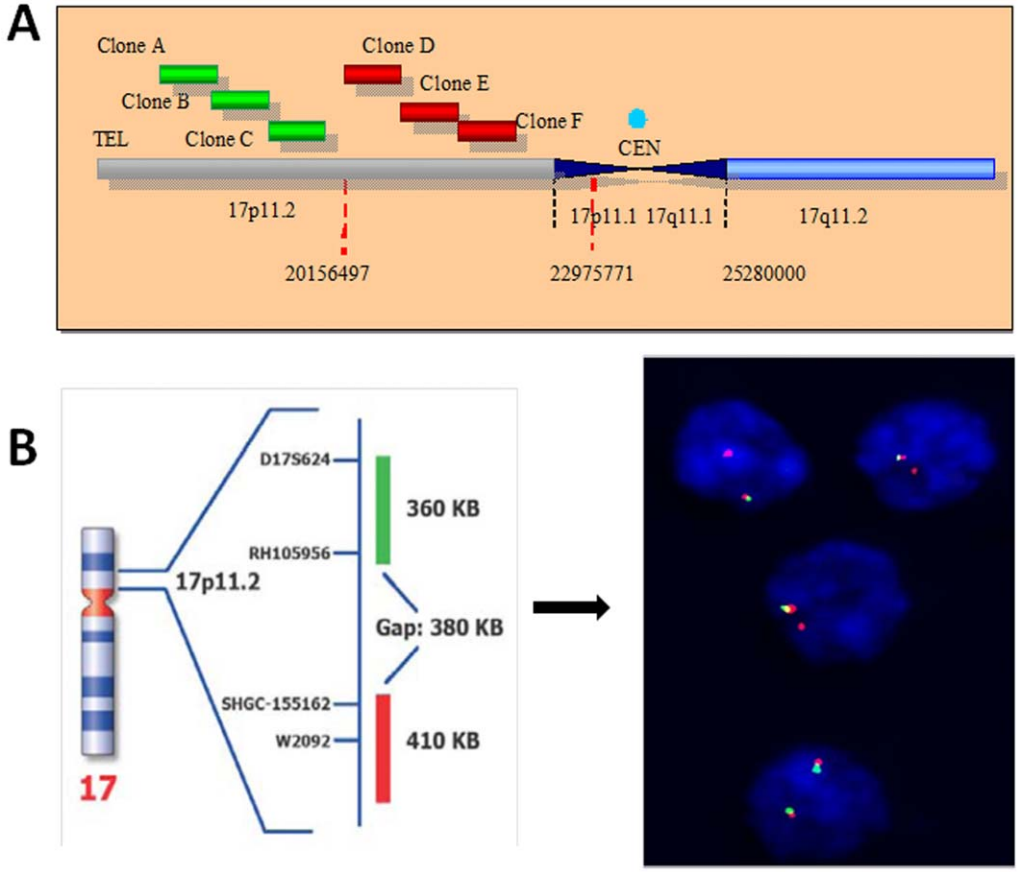


Figure 1. Schematic representation of the chromosome 17p11.2 dual color Break Apart probe combination designed and used for iFISH analysis of this chromosomal region in sCRC. Panel A describes the probe design for which three different clones (A, B and C) directly-labelled with PlatinumBright495 (green signal) and that hybridize to the telomeric part of the 20,156,497 bp region were combined with another three clones (clones D, E and F) directly labelled with PlatinumBright550 and that correspond to sequences harboured centromerically to 20,156,497 bp (red signal), and were produced. The 17p11.2 Break Apart DNA Probe finally consisted of a dual-color assay to detect breakpoints at 17p11.2 using the combination of these 6 fluorescently labelled clones. A positive breakpoint at chromosome 17p11.2 was defined when one or two red/green or yellow fusion signals split into two separate red and green signals. Only red and green signals which were more than one signal diameter apart from each other were counted as reflecting a chromosome break, since based on the probe design a gap of 380 KB exists between the two sets of probes corresponding to the green and the red signals, respectively; two fusion signals identify the two normal chromosomes 17 as illustrated for the lower nuclei shown in panel B. Loss of a green signal in the presence of a single red signal and a fusion signal was interpreted as associated with del(17p) with a 17p11.2 breakpoint (e.g; three upper nuclei in panel B).
doi:10.1371/journal.pone.0042683.g001

(Scotland, UK) and 38 from Hong Kong. Gpr files were pre-processed and normalized as described elsewhere [11]. Patients included in this external validation group were classified according to the Duke's staging system as follows: stage A, 7.5% (n = 8), B, 44.9% (n = 48), C, 39.2% (n = 42) and stage D (metastatic), 8.5% (n = 9). Median of follow up of these patients was 67 months, with a median overall survival of 28.7months (range: 0.3–147.2 months).

Statistical methods

For all continuous variables, mean values and their standard deviation (SD) and range were calculated using the SPSS software package (SPSS 15.0 Inc, Chicago, IL USA); for dichotomic variables, frequencies were reported. In order to evaluate the statistical significance of differences observed between groups, the Student's T and the Mann-Whitney U tests were used for continuous variables, depending on whether they displayed or not a normal distribution, respectively. For qualitative variables, the X^2 test was applied (cross-tab; SPSS). Overall survival (OS) curves were plotted according to the method of Kaplan and Meier, and the log-rank test (one-sided) was used to establish the statistical significance of the differences observed between survival curves (survival; SPSS). Multivariate analysis of prognostic factors for OS was performed using the Cox stepwise regression (forward selection) model (regression, SPSS). For multivariate analysis only those variables showing a significant association with OS in the univariate analysis were included. Statistical significance was considered to be present once P values (or, where appropriate, Pearson-corrected P values) were $<.05$.

Results

Clinical and biological characteristics of liver metastatic versus non-metastatic sporadic colorectal carcinoma (sCRC)

Overall, sCRC cases with liver metastases showed a higher frequency of lymph node metastases ($p \leq .001$) and abnormally increased CEA serum levels ($p \leq .001$) than non-metastatic patients (Table 1). From the prognostic point of view, sCRC with liver metastases also showed a higher frequency of deaths in association

with a significantly shortened patient overall survival (median of 25 months *vs.* not reached, respectively; p≤.001). By contrast, no significant differences were found between liver metastatic *vs.* non-metastatic CRC cases, regarding patient age, gender, tumor localization, histological grade and size, and alkaline phosphatase serum levels (Table 1).

Chromosomal alterations in metastatic *vs* non-metastatic sCRC

For most chromosomes analysed, sCRC with liver metastases showed similar cytogenetic profiles to those of non-metastatic tumors; this included similar (p>.05) frequencies of del(1p) (48% *vs.* 42%), polysomy of chromosome 7 (59% *vs.* 45%), del(8p) associated to gains of 8q (44% *vs.* 26%), polysomy of chromosome 13 (74% *vs.* 58%), del(18q) (52% *vs.* 32%) and gain of chromosome 20q (63% *vs.* 39%) (Table 2). The only statistically significant differences found between liver metastatic and non-metastatic sCRC were those involving chromosomes 17p (p<.001) and 22q (p = .02): all cases showing del(22q) corresponded to liver metastatic tumors (0% *vs.* 22%); del(17p13) was found in 74% of liver metastatic *vs.* 19% of non-metastatic cases; del(17p13) with a breakpoint at 17p11.2 was almost exclusively detected among sCRC with liver metastases (67% *vs.* 10%, p<.001) (Table 2), and; all except one case with del(22q) (n = 5) also demonstrated del(17p11.2) while16 cases which had del(17p11.2) did not carry del(22q). The remaining 36 tumors carried none of the two chromosomal alterations. Interestingly, whenever these two chromosomal alterations were detected, either individually or in combination, they were present in all tumor cells, suggesting they had been acquired in the ancestral tumor cell clone.

Overall, a total of 36 genes are coded at the 17p11.2, 17p13.1 and 22q11.2 chromosomal regions (Table 3); 11 out of these 36 genes (31%) have been found to be involved in cancer. The network of functional interactions among these genes and other related downstream genes implicated in cancer is depicted in Figure 2. As shown in it, such cancer-associated genes deleted in sCRC cases with del(17p11.2) and del(22q11.2) directly related to several well-established biomarkers of sCRC such as the *EGFR*, *BCL2*, *BAX* and *TP53* genes [12–15].

Impact of chromosomal alterations and other disease features of liver metastatic *vs.* non-metastatic sCRC on patient overall survival

Regarding prognosis, the presence of both del(17p13) (p = .04) - including del(17p11.2) (p<.001)- and del(22q11) (p<.001) were associated with a significantly inferior outcome. Other disease features that showed an adverse impact on patient OS were: increased (>7.5 ng/ml) CEA serum levels (p<.001), male gender (p = .04), lymph node involvement (p<.001) and, metastatic liver disease (p<.001) (Figure 3).

Multivariate analysis of the prognostic factors for OS showed that the most informative combination of independent variables to predict an adverse outcome was the presence of del(17p11.2) (p = .04) and del(22q11.2) (p = .002). Based on these two cytoge-

Table 1. Clinical and biological characteristics of liver metastatic (n = 27) versus non-metastatic (n = 31) sporadic colorectal carcinoma (sCRC) patients.

	Liver metastatic sCRC (n = 27)	Non-metastatic sCRC (n = 31)	*p*-value	Total cases (n = 58)
Age (years)*	73 (48–80)	72 (38–83)	NS	72 (38–83)
Gender				
F	11 (41%)	9 (29%)	NS	20 (34%)
M	16 (59%)	22 (71%)		38 (66%)
Tumor Localization				
Rectum	5 (19%)	11 (36%)		16 (28%)
Left colon	13 (48%)	15 (48%)	*NS*	28 (52%)
Right colon	9 (33%)	5 (16%)		14 (20%)
Histological grade				
Well-differentiated	16 (59%)	23 (74%)		39 (67%)
Moderate-differentiated	8 (30%)	7 (22%)	NS	15 (26%)
Poorly-differentiated	3 (11%)	1(4%)		4 (7%)
Histopathology				
pN0	7 (26%)	31 (100%)		38 (66%)
pN1	12 (44%)	0 (0%)	*p≤0.001*	12(21%)
pN2	8 (30%)	0 (0%)		8 (13%)
Tumor Size (cm)#	5 (2.5–9)	5 (2.5–14)	NS	5 (2.5–14)
Serum ALP (mg/dl)	94 (1–330)	108 (55–495)	NS	101 (1–495)
Serum CEA (ng/ml)	45.4 (0.8–4598)	3.2 (0.6–84)	*p≤0.001*	7.2 (0.6–4598)
Deaths	20 (74%)	3 (10%)	*p≤0.001*	23 (40%)
Median OS (months)*	25	Not Reached	*p≤0.001*	Not Reached

*Results expressed as median (range) or
#as number of cases (percentage); NS: statistically not significant (p>.05); F: female; M: male; ALP: alkaline phosphatase; CEA: Carcinoembryonic antigen; OS: overall survival.
doi:10.1371/journal.pone.0042683.t001

Table 2. Chromosomal alterations of primary tumors from liver metastatic (n = 27) versus non-metastatic sCRC patients (n = 31).

	Liver metastatic tumors (n = 27)	Non-metastatic tumors (n = 31)	*p*-value	Total cases (n = 58)
Chromosome 1				
Normal	7 (26%)	14 (45%)		21 (36%)
del(1p)	13 (48%)	13 (42%)	NS	26 (45%)
Polysomy	7 (26%)	4 (13%)		11 (19%)
Chromosome 7				
Normal	5 (19%)	14 (45%)		19 (33%)
del(7q)	5 (19%)	1 (3%)	NS	6 (10%)
q+	1 (3%)	2 (7%)		3 (5%)
Polysomy	16 (59%)	14 (45%)		30 (52%)
Chromosome 8				
Normal	3 (11%)	7 (23%)		10 (17%)
del(8p)	5 (19%)	4 (13%)		9 (15%)
q+	2 (7%)	3 (9%)	NS	5 (9%)
Del(8p)/8q+	12 (44%)	8 (26%)		20 (35%)
Polysomy	5 (19%)	9 (29%)		14 (24%)
Chromosome 13				
Normal	7 (26%)	13 (42%)	NS	20 (35%)
Polysomy	20 (74%)	18 (58%)		38 (65%)
Chromosome 14				
Normal	15 (55%)	19 (61%)		34(59%)
del(14q)	4 (15%)	1 (3%)	NS	5 (9%)
Polysomy	8 (30%)	11 (36%)		19 (32%)
Chromosome 17				
Normal	5 (19%)	20 (65%)		25 (43%)
del(17p)	20 (70%)	6 (19%)	$p<.001$	26 (45%)
Polysomy	2 (7%)	5 (16%)		7 (12%)
Del(17p11.2)	18 (67%)	3 (10%)	$p<.001$	21 (36%)
Chromosmose 18				
Normal	13 (48%)	17 (55%)		30 (52%)
del(18q)	14 (52%)	10 (32%)	NS	29 (50%)
Polysomy	0 (0%)	4 (13%)		4 (7%)
Chromosome 20				
Normal	5 (19%)	12 (39%)		17 (27%)
20q+	17 (63%)	12 (39%)	NS	29 (50%)
Polysomy	5 (19%)	7 (22%)		12 (21%)
Chromosome 22				
Normal	15 (56%)	23 (74%)		38 (66%)
del(22q)	6 (22%)	0 (0%)	$p = .02$	6 (10%)
Polysomy	6 (22%)	8 (26%)		14 (24%)

Results expressed as number of cases and percentage of cases between brackets; NS: statistically not significant ($p>.05$).
doi:10.1371/journal.pone.0042683.t002

netic variables, a scoring system was built to stratify patients into a low- (no adverse features: score 0; n = 24), intermediate- (one adverse feature: score 1; n = 28) and high-risk (two adverse features: score 2; n = 5) groups with significantly different ($p<.001$) OS rates at 5-years: 92%, 53% and 0%, respectively (Figure 3).

Validation of the clinical impact of del(17p11.2) and del(22q) in an independent series of patients

In order to confirm the prognostic impact of the two chromosomal abnormalities described above, we investigated their prognostic impact in an independent series of colorectal cancer patients from the public GEO database (n = 119). Noteworthy, also in this new series, patients whose tumors harboured pericentromeric breakpoints at 17p in the 17p11.2 chromosomal

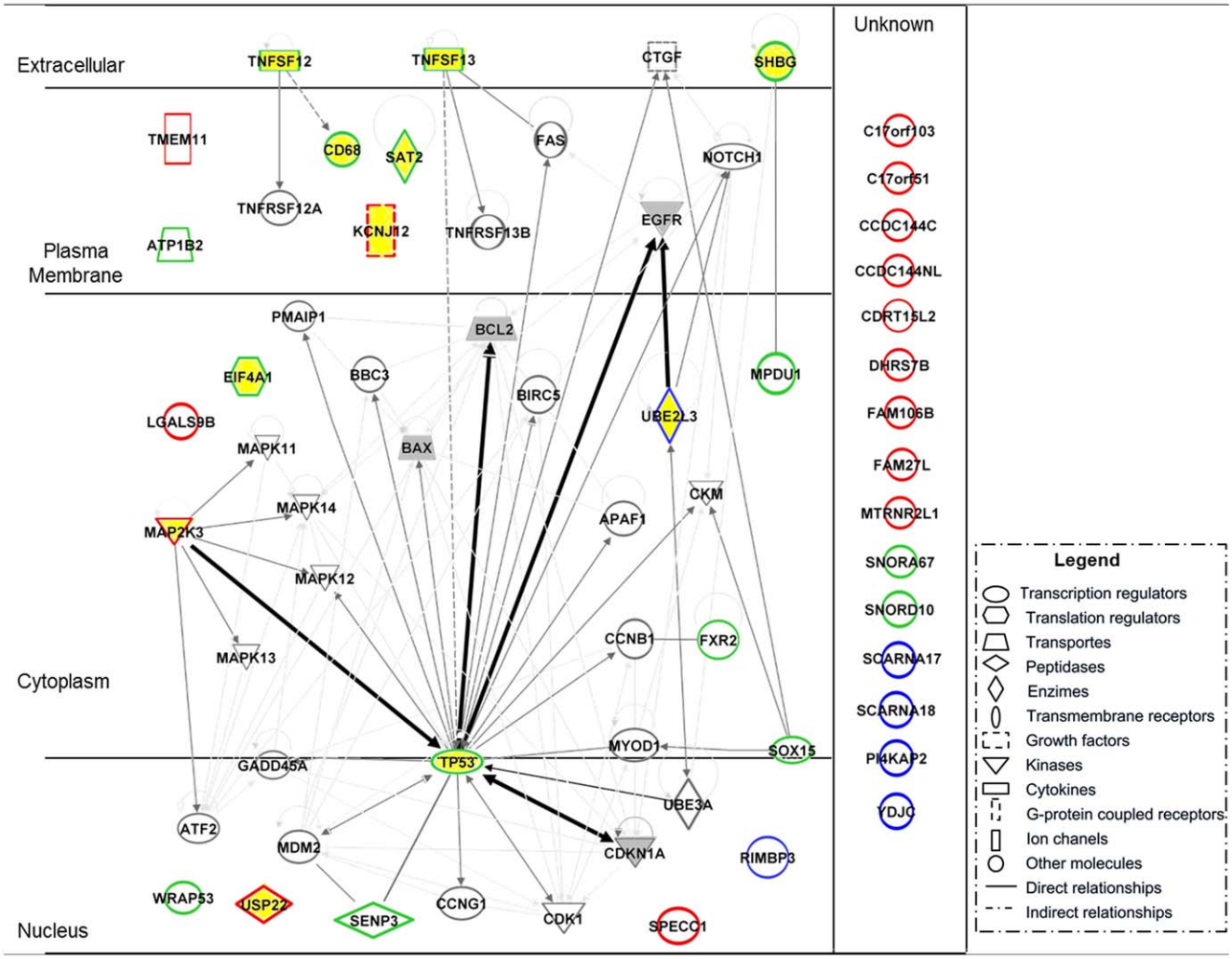


Figure 2. Schematic representation of the network of interactions observed between genes encoded at the 17p11.2 (genes highlighted in red), 17p13.1 (genes highlighted in green) and 22q11.2 (genes highlighted in blue) chromosomal regions, and molecules downstream molecules regulated by these genes which have been associated with cancer or cancer related signalling pathways. Genes highlighted in yellow are encoded at the three chromosomal regions referred above and they have been previously associated with cancer; genes highlighted in grey are considered as biomarkers for sCRC.
doi:10.1371/journal.pone.0042683.g002

region (from 15 to 25 megabases from p-ter) were found to have an inferior clinical outcome than those harbouring del(17p13) alone ($p = .02$ and $p = .04$, respectively). The prognostic impact of del(17p11.2) was even stronger ($p = .01$) when all other tumors which showed pericentromeric deletions, including those with breakpoints in the q-arm close to the centromere (from 15 to 27.5 Mb from p-ter), were considered (Figure S1).

These results support the observations of our dataset and confirm the prognostic impact of del(17p11.2). However, the prognostic impact of del(22q) could not be confirmed ($p>.05$) in this new independent sCRC series of patients.

Discussion

sCRC patients who do not show or develop distant metastasis are often cured by surgical resection of the primary tumor with optional administration of adjuvant therapy. However, when metastasis to the liver and other organs occur, the chances of cure are dramatically reduced. Despite the fact that the understanding of the genetic mechanisms underlying the early stages of both familial [16] and sporadic CRC has significantly advanced in recent years [17], the genetic mechanisms responsible for progression of sCRC to a metastatic phenotype still remain poorly understood. In this study, we investigated the pattern of numerical chromosomal alterations of primary tumors from metastatic sCRC that exhibited synchronous liver metastases versus non-metastatic sCRC. In order to avoid false-negative non-metastatic cases, in this later group only sCRC with a relative long follow up (median follow-up of 99 months) were selected for the non-metastatic tumor group. Similarly, only liver metastatic cases who had undergone complete resection of both their primary and metastatic tumor, were included in the metastatic patient group.

iFISH probes targeting those chromosomal regions more frequently altered in sCRC [6] were specifically applied to the cytogenetic characterization of both patient groups and a new probe for the definition of del(17p) associated with breakpoints at chromosome 17p11.2, was also systematically used. In line with previous observations which show that liver metastatic and non-metastatic sCRC share multiple chromosomal alterations (e.g.

Table 3. List of genes encoded at chromosomal regions identified as being deleted by iFISH probes directed against the 17p11.2 (20156497 bp to 22975771 bp), 17p13.1 (7449445 bp to 7594642 bp) and 22q11.2 (21852397 bp to 21984023 bp) chromosomal regions: gene name, cell localization and function.

Coded name	Gene	Cellular localization	Function
17p11.2			
C17orf103	Chromosome 17 open reading frame 103	Unknown	Unknown
C17orf51	Chromosome 17 open reading frame 51	Unknown	Unknown
CCDC144C	Coiled-coil domain containing 144C	Unknown	Unknown
CCDC144NL	Coiled-coil domain containing 144 family, N-terminal like	Unknown	Unknown
CDRT15L	CMT1A duplicated region transcript 15-like 2	Unknown	Unknown
DHRS7B	Dehydrogenase/reductase (SDR family) member 7B	Unknown	Metabolism
FAM106B	Family with sequence similarity 106, member B	Unknown	Unknown
FAM27L	Family with sequence similarity 27-like	Unknown	Unknown
KCNJ12	Potassium inwardly-rectifying channel, subfamily J, member 12	Membrane	Transport
LGALS9B	Lectin, galactoside-binding, soluble, 9B	Cytoplasm	Cell-cell adhesion
MAP2K3	Mitogen-activated protein kinase kinase 3	Cytoplasm	Cell death
MTRNR2L1	MT-RNR2-like 1	Unknown	Unknown
SPECC1	Sperm antigen with calponin homology and coiled-coil domains 1	Nucleus	Unknown
TMEM11	Transmembrane protein 11	Membrane	Transport
USP22	Ubiquitin specific peptidase 22	Nucleus	Cell cycle
17p13.1			
ATP1B2	ATPase, Na+/K+ transporting, beta 2 polypeptide	Membrane	Metabolism
CD68	CD68 molecule	Membrane	Metabolism
EIF4A1	eukaryotic translation initiation factor 4A1	Cytoplasm	Metabolism
FXR2	fragile X mental retardation, autosomal homolog 2	Cytoplasm	Metabolism
MPDU1	mannose-P-dolichol utilization defect 1	Cytoplasm	Metabolism
SAT2	spermidine/spermine N1-acetyltransferase family member 2	Membrane	Metabolism
SENP3	SUMO1/sentrin/SMT3 specific peptidase 3	Nucleus	Metabolism
SHBG	sex hormone-binding globulin	Extracellular	Cell death
SNORA67	small nucleolar RNA, H/ACA box 67	Unknown	Unknown
SNORD10	small nucleolar RNA, C/D box 10	Unknown	Unknown
SOX15	SRY (sex determining region Y)-box 15	Nucleus	Cell differentiation
TNFSF12	tumor necrosis factor (ligand) superfamily, member 12	Extracellular	Cell death
TNFSF13	tumor necrosis factor (ligand) superfamily, member 13	Extracellular	Cell death
TP53	tumor protein p53	Nucleus	Apoptosis
WRAP53	WD repeat containing, antisense to TP53	Nucleus	Telomerase activity
22q11.2			
PI4KAP2	Phosphatidylinositol 4-kinase, catalytic, alpha pseudogene 2	Unknown	Metabolism
RIMBP3	RIMS binding protein 3	Nucleus	Unknown
SCARNA17	Small Cajal body-specific RNA 17	Unknown	Unknown
SCARNA18	Small Cajal body-specific RNA 18	Unknown	Unknown
UBE2L3	Ubiquitin-conjugating enzyme E2L 3	Cytoplasm	Metabolism
YDJC	YdjC homolog (bacterial)	Unknown	Metabolism

Genes which have been associated with cancer are shown in bold.
doi:10.1371/journal.pone.0042683.t003

gains of chromosomes 7, 8q, 13q and 20q and losses of the 1p, 8p, 14q, 17p, 18q and 22q chromosomes) [7,18–20], here we also found a similar distribution between liver metastatic and non-metastatic tumors for most chromosomal alterations identified. In contrast, del(22q) and del(17p) (particularly when associated with breakpoints at chromosome 17p11.2), were significantly more prevalent or even restricted, to liver metastatic tumors. These later findings support a potential role for both del(17p11.2) and del(22q) in the metastatic process of sCRC to the liver.

Previous reports based on cytogenetic analyses of metastatic disease from colorectal tumors indicated that chromosome 17p is frequently lost in sCRC [21,22]. In line with other studies and using similar methodological approaches, our results showed the presence of del(17p13) in almost half of the sCRC cases studied

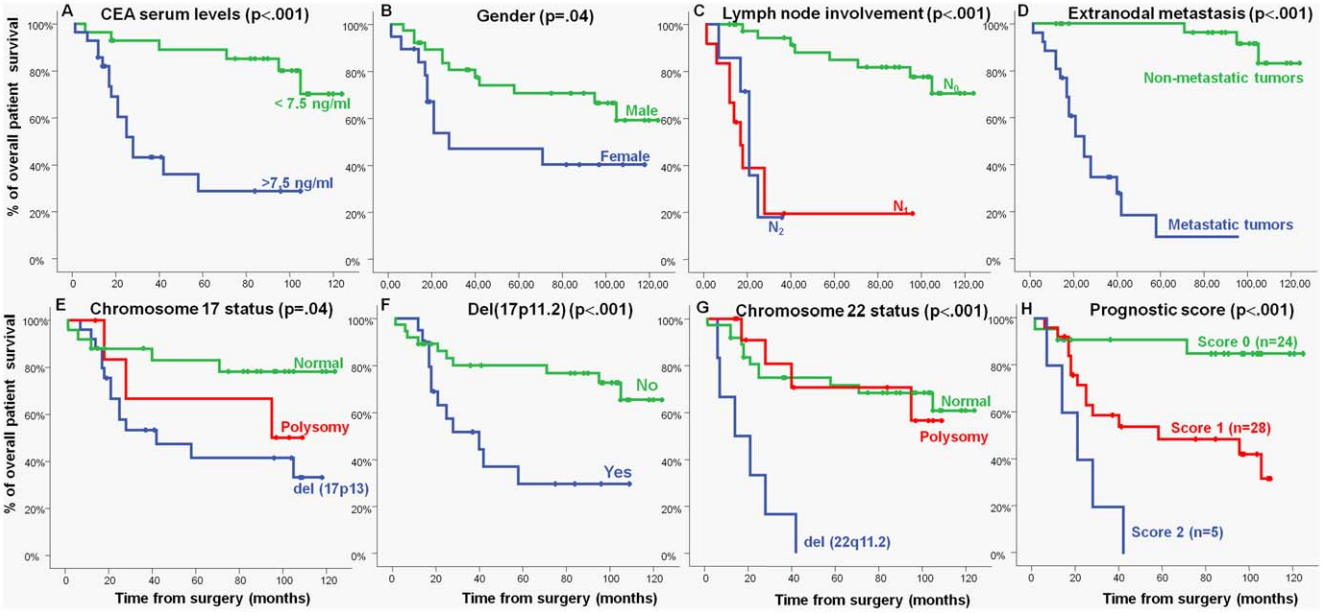


Figure 3. Clinical, biological and genetic characteristics of sCRC patients which showed a significant impact on overall survival in the univariate analysis: (A) carcinoembryonic antigen (CEA), (B) gender, (C) lymph node involvement, (D) occurrence of distant metastasis, (E) chromosome 17 status, (F) del(17p11.2) (G) chromosome 22 status, and (H) prognostic score established on the basis of the two most informative independent prognostic factors -del(17p11.2) and chromosome 22 status; p<.0001-.
doi:10.1371/journal.pone.0042683.g003

[23,24]; The frequency of del(17p13) was also significantly higher in liver metastatic than non-metastatic cases, as has been suggested by other groups [25–27]. It was noted that among cases with del(17p13), occurrence of a breakpoint at chromosome 17p11.2 was mostly restricted to metastatic sCRC. Coinciding with these observations, several authors have previously found that losses of chromosome 17p in metastatic CRC samples cover larger regions than in primary tumors, suggesting that unknown suppressor genes, other than the *TP53* gene, could be involved in the newly deleted 17p sequences [28]. If this is confirmed, then these differences could explain why cases with del(17p) in the absence of *TP53* mutations, also occur in advanced sCRC. Moreover, it provides evidence for the potential existence of new additional tumor suppressor genes (and potentially also oncogenes) coded in the centromeric portion of chromosome 17p, proximal to *TP53*. In this regard, it should be noted that several cancer associated genes (e.g.: *KCNJ12*, *MAP2K3*, and *USP22*) are coded in this chromosomal region, the first gene systematically deleted at this breakpoint region being a gene of unknown function (*FAM27L*). Interestingly, genetic polymorphisms involving this chromosomal region including the FAM27L gene, have been recently associated with an increased risk for chronic myeloid leukemia [29]. Further studies, in which mutations of this gene and deletions at chromosome 17p11.2 are searched for, may indicate their potential role in sCRC liver metastasis. Among other genes the *MAP2K3* gene is also coded in chromosome 17 region found to be commonly deleted in metastatic sCRC. *MAP2K3* is a strong promoter of tumor invasion, progression and short survival in several human cancers [30] and previous studies have shown that decreased expression of *MAP2K3* is associated with human breast infiltrating ductal carcinomas [31]; similarly, non-synonymous coding SNPs downregulating *KCNJ12* expression have been related with rhabdomyosarcomas [32], supporting a potential role for both genes in liver metastatic sCRC. However, in this chromosomal region, also some oncogenes are coded such as the *USP22* gene. Recent studies have shown that aberrant expression of USP22 is associated with liver metastasis and poor prognosis [33], due to the fact that this gene positively regulates cell cycle via both the BMI-1-mediated INK4a/ARF pathway and the Atk signaling pathway [34]. However, the activation and oncogenic role of USP22 in the progression of sCRC is potentially linked to genes encoded in other chromosomal regions such as the *BMI-1* (10p13), *CMYC* (8q24) and *CCND2* (12p13) genes [35].

In addition to del(17p11.2), in this study we also found an association between losses of chromosome 22q and disease outcome, in line with previous observations [26,36,37]. Previous studies based on CGH analysis [38] have shown an association between del(22q) and liver metastasis among sCRC patients; similarly, Yana *et al* [39] showed that del(22q) correlates with the Duke's stage of the disease. Iino *et al* [26] have suggested that LOH at chromosomes 17p, 18q, and 22q, is associated with an increased metastatic potential of sCRC. In the latter study, LOH at chromosome 17p was also significantly associated with vascular invasion, whereas 18q and 22q LOH correlated more with lymphatic dissemination of the disease; importantly, only LOH of chromosome 22q showed a significant association with the presence of lymph node metastasis. Thus, it could be hypothesized that in sCRC, these three chromosomal losses may be specifically associated with the metastatic process. If this holds true, screening for genetic abnormalities of primary sCRC tumors could be useful for predicting the metastatic potential which exists at the time of diagnosis [40]. It should be emphasized that analysis of del(17p11.2) in paired primary tumors and liver metastases from sCRC patients showed either presence or absence of these chromosomal changes in both (paired) tumor samples in all but two cases; in these later two cases, del(17p11.2) was only detected by SNP-arrays in the liver metastatic tumor.

Multivariate analysis of prognostic factors for OS, showed the independent prognostic value of the two chromosomal abnormalities, del(17p) with a breakpoint at 17p11.2 and del(22q);

consequently, coexistence of both chromosomal alterations was associated with a significantly reduced OS *vs.* cases which showed neither of these alterations (OS at 5 years of 0% versus 93%, respectively). Despite the fact that an association has been reported between different chromosomal abnormalities and the prognosis of sCRC [18], to the best of our knowledge this is the first report in which the independent prognostic value of del(17p) with a breakpoint at 17p11.2 and of del(22q) is described. Preliminary results using genome-wide array analyses have shown an association between specific genetic alterations present in primary sCRC tumors and patient survival [10,18,22]. Poulogiannis *et al* (using a DNA microarray platform covering the entire genome at an average of 1 Mb of resolution) identified DNA copy number losses at 18q12.2 to be an independent prognostic marker [10]. In the current study, we have re-analyzed this dataset and confirmed the prognostic value of del(17p) including that of del(17p) with a breakpoint at 17p11.2; in contrast, the clinical impact of del(22q) could not be validated in this series. Although the precise clinical value of del(22q) should be investigated further, validation of our data concerning the prognostic impact of the 17p11.2 breakpoint in an independent dataset (in spite of the substantial differences in the technologies applied in both studies) strengthens the evidence for the clinical relevance of chromosome 17p deletions encompassing genomic regions beyond the *TP53* locus, and points to the potential role of other candidate genes coded at chromosome 17p centromericly to *TP53*. As discussed above, such genes include the *MAP2K3*, *KCNJ12* and *USP22* genes [30–35]. Interestingly, when we searched for direct interactions among the deleted genes and other cancer-associated genes, 30 genes deleted in cases with del(17p), and another 6 genes deleted in cases with del(22q), emerged as directly related to signaling pathways involved in cell growth and proliferation (e.g., *EGFR* and *CDK1A*) as well as in cell death (e.g., *BAX* and *BCL2*). These findings suggest a potential role for the combined deletion of these genes in conferring poor-prognosis to sCRC with coexisting del(17p) and del(22q), possibly due to increased cell proliferation and survival and diminished DNA repair.

In summary, in the present study we show that the presence of del(17p) with a breakpoint at 17p11.2 is an independent adverse prognostic factor for OS of sCRC. When combined with del(22q11.2) it allowed the identification of three groups of sCRC patients with significantly different outcome, which could be predicted at diagnosis. Further prospective studies are required in larger series of sCRC patients to confirm the prognostic value of the combined assessment of del(17p) and del(22q) in primary tumor samples at diagnosis and the precise role of the deleted genes.

Supporting Information

Figure S1 Validation of the impact of chromosome 17 status on overall survival in an independent series of sCRC patients from the GEO database (n = 109): panel A, del(17p13); panels B and C, del(17p) harbouring pericentromeric breakpoints at chromosome 17p and del(17p) harboring a pericentromeric breakpoint at both chromosomes 17p and 17q, respectively.
(TIF)

Author Contributions

Conceived and designed the experiments: AO JMS. Performed the experiments: MGG LMB MLG MMA OB EF. Analyzed the data: MGG CM CF JR MF JMS CT. Contributed reagents/materials/analysis tools: MGG CM CF JR MF JMS. Wrote the paper: MGG LMB CM CF MLG MMA OB EF MF JR AO JMS.

References

1. Sartore-Bianchi A, Moroni M, Veronese S, Carnaghi C, Bajetta E, et al. (2007) Epidermal growth factor receptor gene copy number and clinical outcome of metastatic colorectal cancer treated with panitumumab. J Clin Oncol 25: 3238–3245.
2. Sayagues JM, Abad Mdel M, Melchor HB, Gutierrez ML, Gonzalez-Gonzalez M, et al. (2010) Intratumoural cytogenetic heterogeneity of sporadic colorectal carcinomas suggests several pathways to liver metastasis. J Pathol 221: 308–319.
3. Arnold CN, Goel A, Blum HE, Boland CR (2005) Molecular pathogenesis of colorectal cancer: implications for molecular diagnosis. Cancer 104: 2035–2047.
4. Ashktorab H, Schaffer AA, Daremipouran M, Smoot DT, Lee E, et al. (2010) Distinct genetic alterations in colorectal cancer. PLoS One 5: e8879.
5. Hu XT, Chen W, Wang D, Shi QL, Zhang FB, et al. (2008) The proteasome subunit PSMA7 located on the 20q13 amplicon is overexpressed and associated with liver metastasis in colorectal cancer. Oncol Rep 19: 441–446.
6. Sayagues JM, Fontanillo C, Abad Mdel M, Gonzalez-Gonzalez M, Sarasquete ME, et al. (2010) Mapping of genetic abnormalities of primary tumours from metastatic CRC by high-resolution SNP arrays. PLoS One 5: e13752.
7. Camps J, Grade M, Nguyen QT, Hormann P, Becker S, et al. (2008) Chromosomal breakpoints in primary colon cancer cluster at sites of structural variants in the genome. Cancer Res 68: 1284–1295.
8. World Health Organization. WHO (1981) International Histological Classification of Tumors V-G, 1967–1981; 2nd edn, Berlin: Springer-Verlag, 1988–1992.
9. Munoz-Bellvis L, Fontanillo C, Gonzalez-Gonzalez M, Garcia E, Iglesias M, et al. (2012) Unique genetic profile of sporadic colorectal cancer liver metastasis versus primary tumors as defined by high-density single-nucleotide polymorphism arrays. Mod Pathol.
10. Poulogiannis G, Ichimura K, Hamoudi RA, Luo F, Leung SY, et al. (2010) Prognostic relevance of DNA copy number changes in colorectal cancer. J Pathol 220: 338–347.
11. Mackintosh C, Ordonez JL, Garcia-Dominguez DJ, Sevillano V, Llombart-Bosch A, et al. (2012) 1q gain and CDT2 overexpression underlie an aggressive and highly proliferative form of Ewing sarcoma. Oncogene 31: 1287–1298.
12. Re M, Magliulo G, Tarchini P, Mallardi V, Rubini C, et al. (2011) p53 and BCL-2 over-expression inversely correlates with histological differentiation in occupational ethmoidal intestinal-type sinonasal adenocarcinoma. Int J Immunopathol Pharmacol 24: 603–609.
13. Asghar U, Hawkes E, Cunningham D. (2010) Predictive and prognostic biomarkers for targeted therapy in metastatic colorectal cancer. Clin Colorectal Cancer 9: 274–281.
14. Yashiro M, Hirakawa K, Boland CR (2010) Mutations in TGFbeta-RII and BAX mediate tumor progression in the later stages of colorectal cancer with microsatellite instability. BMC Cancer 10: 303.
15. Seicean R, Crisan D, Boers JE, Mocan T, Seicean A, et al. (2011) The prognostic role of apoptosis mediators in rectal adenocarcinoma. Hepatogastroenterology 58: 1490–1494.
16. Syngal S, Fox EA, Li C, Dovidio M, Eng C, et al. (1999) Interpretation of genetic test results for hereditary nonpolyposis colorectal cancer: implications for clinical predisposition testing. JAMA 282: 247–253.
17. Maltzman T, Knoll K, Martinez ME, Byers T, Stevens BR, et al. (2001) Ki-ras proto-oncogene mutations in sporadic colorectal adenomas: relationship to histologic and clinical characteristics. Gastroenterology 121: 302–309.
18. Sheffer M, Bacolod MD, Zuk O, Giardina SF, Pincas H, et al. (2009) Association of survival and disease progression with chromosomal instability: a genomic exploration of colorectal cancer. Proc Natl Acad Sci U S A 106: 7131–7136.
19. Derks S, Postma C, Carvalho B, van den Bosch SM, Moerkerk PT, et al. (2008) Integrated analysis of chromosomal, microsatellite and epigenetic instability in colorectal cancer identifies specific associations between promoter methylation of pivotal tumour suppressor and DNA repair genes and specific chromosomal alterations. Carcinogenesis 29: 434–439.
20. Popat S, Zhao D, Chen Z, Pan H, Shao Y, et al. (2007) Relationship between chromosome 18q status and colorectal cancer prognosis: a prospective, blinded analysis of 280 patients. Anticancer Res 27: 627–633.
21. Khine K, Smith DR, Goh HS (1994) High frequency of allelic deletion on chromosome 17p in advanced colorectal cancer. Cancer 73: 28–35.
22. Diep CB, Thorstensen L, Meling GI, Skovlund E, Rognum TO, et al. (2003) Genetic tumor markers with prognostic impact in Dukes' stages B and C colorectal cancer patients. J Clin Oncol 21: 820–829.
23. Garcia J, Duran A, Tabernero MD, Garcia Plaza A, Flores Corral T, et al. (2003) Numerical abnormalities of chromosomes 17 and 18 in sporadic colorectal cancer: Incidence and correlation with clinical and biological findings and the prognosis of the disease. Cytometry B Clin Cytom 51: 14–20.

24. Risio M, Casorzo L, Chiecchio L, De Rosa G, Rossini FP (2003) Deletions of 17p are associated with transition from early to advanced colorectal cancer. Cancer Genet Cytogenet 147: 44–49.
25. Losi L, Luppi G, Benhattar J (2004) Assessment of K-ras, Smad4 and p53 gene alterations in colorectal metastases and their role in the metastatic process. Oncol Rep 12: 1221–1225.
26. Iino H, Fukayama M, Maeda Y, Koike M, Mori T, et al. (1994) Molecular genetics for clinical management of colorectal carcinoma. 17p, 18q, and 22q loss of heterozygosity and decreased DCC expression are correlated with the metastatic potential. Cancer 73: 1324–1331.
27. Chang SC, Lin JK, Lin TC, Liang WY (2005) Genetic alteration of p53, but not overexpression of intratumoral p53 protein, or serum p53 antibody is a prognostic factor in sporadic colorectal adenocarcinoma. Int J Oncol 26: 65–75.
28. Paredes-Zaglul A, Kang JJ, Essig YP, Mao W, Irby R, et al. (1998) Analysis of colorectal cancer by comparative genomic hybridization: evidence for induction of the metastatic phenotype by loss of tumor suppressor genes. Clin Cancer Res 4: 879–886.
29. Kim DH, Lee ST, Won HH, Kim S, Kim MJ, et al. (2011) A genome-wide association study identifies novel loci associated with susceptibility to chronic myeloid leukemia. Blood 117: 6906–6911.
30. Gurtner A, Starace G, Norelli G, Piaggio G, Sacchi A, et al. (2010) Mutant p53-induced up-regulation of mitogen-activated protein kinase kinase 3 contributes to gain of function. J Biol Chem 285: 14160–14169.
31. Jia M, Souchelnytskyi N, Hellman U, O'Hare M, Jat PS, et al. (2010) Proteome profiling of immortalization-to-senescence transition of human breast epithelial cells identified MAP2K3 as a senescence-promoting protein which is downregulated in human breast cancer. Proteomics Clin Appl 4: 816–828.
32. Sher RB, Cox GA, Mills KD, Sundberg JP (2011) Rhabdomyosarcomas in aging A/J mice. PLoS One 6: e23498.
33. Liu YL, Yang YM, Xu H, Dong XS (2011) Aberrant expression of USP22 is associated with liver metastasis and poor prognosis of colorectal cancer. J Surg Oncol 103: 283–289.
34. Liu YL, Jiang SX, Yang YM, Xu H, Liu JL, et al. (2012) USP22 acts as an oncogene by the activation of BMI-1-mediated INK4a/ARF pathway and Akt pathway. Cell Biochem Biophys 62: 229–235.
35. Liu YL, Yang YM, Xu H, Dong XS (2010) Increased expression of ubiquitin-specific protease 22 can promote cancer progression and predict therapy failure in human colorectal cancer. J Gastroenterol Hepatol 25: 1800–1805.
36. Castells A, Ino Y, Louis DN, Ramesh V, Gusella JF, et al. (1999) Mapping of a target region of allelic loss to a 0.5-cM interval on chromosome 22q13 in human colorectal cancer. Gastroenterology 117: 831–837.
37. Castells A, Gusella JF, Ramesh V, Rustgi AK (2000) A region of deletion on chromosome 22q13 is common to human breast and colorectal cancers. Cancer Res 60: 2836–2839.
38. Al-Mulla F, Keith WN, Pickford IR, Going JJ, Birnie GD (1999) Comparative genomic hybridization analysis of primary colorectal carcinomas and their synchronous metastases. Genes Chromosomes Cancer 24: 306–314.
39. Yana I, Kurahashi H, Nakamori S, Kameyama M, Nakamura T, et al. (1995) Frequent loss of heterozygosity at telomeric loci on 22q in sporadic colorectal cancers. Int J Cancer 60: 174–177.
40. Ghadimi BM, Grade M, Monkemeyer C, Kulle B, Gaedcke J, et al. (2006) Distinct chromosomal profiles in metastasizing and non-metastasizing colorectal carcinomas. Cell Oncol 28: 273–281.

Genomics 97 (2011) 86–93

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Segmentation of genomic and transcriptomic microarrays data reveals major correlation between DNA copy number aberrations and gene–loci expression

M. Ortiz-Estevez [a], J. De Las Rivas [b], C. Fontanillo [b], A. Rubio [a,*]

[a] Department of Electronics and Communication, CEIT and TECNUN (Univeristy of Navarra), Paseo Manuel Lardizabal 15, 20009, San Sebastian, Spain
[b] Centro de Investigacion del Cancer (CiC-IBMCC), CSIC and USAL, E37007 Salamanca, Spain

ARTICLE INFO

Article history:
Received 4 June 2010
Accepted 22 October 2010
Available online 29 October 2010

Keywords:
Gene expression
Copy number
Microarrays
Segmentation

ABSTRACT

DNA copy number aberrations (CNAs) are genetic alterations common in cancer cells. Their transcriptional consequences are still poorly understood. Based on the fact that DNA copy number (CN) is highly correlated with the genomic position, we have applied a segmentation algorithm to gene expression (GE) to explore its relation with CN. We have found a strong correlation between segmented CN (sCN) and segmented GE (sGE), corroborating that CNAs have clear effects on genome-wide expression. We have found out that most of the recurrent regions of sGE are common to those obtained from sCN analysis. Results for two cancer datasets confirm the known targets of aberrations and provide new candidates to study. The suggested methodology allows to find recurrent aberrations specific to sGE, revealing loci where the expression of the genes is independent from their CNs. R code and additional files are available as supplementary material.

1. Introduction

The presence of genomic aberrations in tumoral cells is a well-known fact. In recent years, several studies have shown that the alteration of DNA copy number (CN) can be related to similar modifications in the expression levels of some specific genes [1–3]. These changes can be amplifications (gains) or deletions (losses) of a region of a chromosome, or even a whole chromosome and they are commonly called DNA copy number aberrations (CNAs). These abnormalities are assumed to affect gene expression (GE) and ultimately some of them may coadjuvate to the development of a particular cancer. However, the relationship between CN and GE is complex and not well understood: there are genes whose expression is not apparently affected by their CNs and genes which show their expression strongly correlated with them. One reason for this unclear relationship is that CN is only one of the several factors that can affect the regulation of GE and gene function in complex metazoans.

Recently, new studies focused on the relationship between CNAs and GE have performed joint analysis based on different strategies. Some of them calculated the correlation between CN and GE gene by gene across samples [2,4,5]. These correlations are not particularly large, although they are significant for many genes. Others, like Tsafrir et al. [6], obtained a correlation along the genome using filtered CN and filtered GE. Witten et al. use a sparsified version of the canonical correlation between CN and GE [7]. Moreover, Jarvinen et al. [8] and Cifola et al. [9] based their experiments on differential expression calculated between groups defined by genomic alterations. Finally, other studies are based on the hypothesis that some genes are grouped in the genome by their functions and, because of this, they consider that CNAs affect groups of cofunctional genes [10,11].

Here, we hypothesize that there should be a common behavior of the genes under the influence of CNAs. From this viewpoint we look for a global consistent relationship between CN and GE. Knowing that CNAs are highly correlated with the position on the genome, we propose that the global GE modifications produced by CNAs should be also correlated with the genomic position.

In order to study and evaluate the relationship between CN and GE, we have used a global approach that does not focus only in gene by gene relationships but that considers the complete genome and treats the loci with a coherent common approach, both when measuring CN and transcriptomic activity. To test the validity of our hypothesis, we have analyzed two different set of samples that have matched CN and GE. One of them is a study of glioblastoma multiforme (GBM) [2] and the other is a study of acute lymphoblastic leukemia (ALL) [3].

Since the CN values of two adjacent positions in the genome are (unless there is a CNA) identical, segmentation of raw CN improves the estimation of real CN. The segmentation algorithm (in the case of CN) identifies contiguous subsets of SNPs in the genome that have the same CN value. It provides “sharp” edges between regions instead of smooth transitions as standard filters do. This characteristic is important when working with CN data because when a region is lost (or amplified) the change between the two sides of the *break point* is not smooth. Therefore, the segmentation methods applied to raw CN data (sCN) give better results than linear or median filters [12] and they are customarily applied. We applied a similar segmentation approach (in the two datasets) to both analysis (GE and CN). Applying

* Corresponding author.
E-mail address: arubio@ceit.es (A. Rubio).

0888-7543/$ – see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2010.10.008

a segmentation algorithm to GE removes (to a certain extent) the possible effects of the regulation of GE that are not related with the genomic position. We found that segmented values of GE (sGE) are strongly correlated with sCN.

Finding recurrent aberrations in segments of the genome has been an active field of research on the last years [13]. Copy number data can be used to find chromosomal regions in which aberrations (deletions or amplifications) frequently appear. It is essential to find the recurrent changes in samples to get a common CNA signature of a given disease. In cancer, full agreement within tumor samples is difficult to find due to tumor heterogeneity. However, if there is a consistent region of aberration which happens more often than expected by chance, it could reveal one of the causes of the detected cancer. These chromosomal regions might include genes which change their expression because of these CNAs and they would be also found as recurrent using sGE data. The recurrent aberrated regions in sCN and sGE in this work have been independently calculated and the results show that most of them appear in the same cytobands in sGE and sCN.

Contrary to this general trend, in some cases we found genome regions or loci where CNAs do not correlate with GE changes, for example where the sGE is significantly altered but the sCN is neutral. These regions can be affected by another level of regulation, e.g., epigenetic methylation/demethylation, or occur in zones with recurrent copy neutral loss of heterozygosity (LOH) [14,15]. However, in our datasets, we did not find a conclusive evidence of this effect.

The calculation of sGE using expression microarray data can be, on its own, an interesting approach to putatively pinpoint loci in the genome affected by different positional factors, such as alteration in the number of copies, epigenetic modifications, LOH or other events linked to the position in the genome.

2. Materials and methods

The analysis of the arrays can be divided into two different parallel processes for CN and GE. Fig. 1 outlines the steps followed in this research.

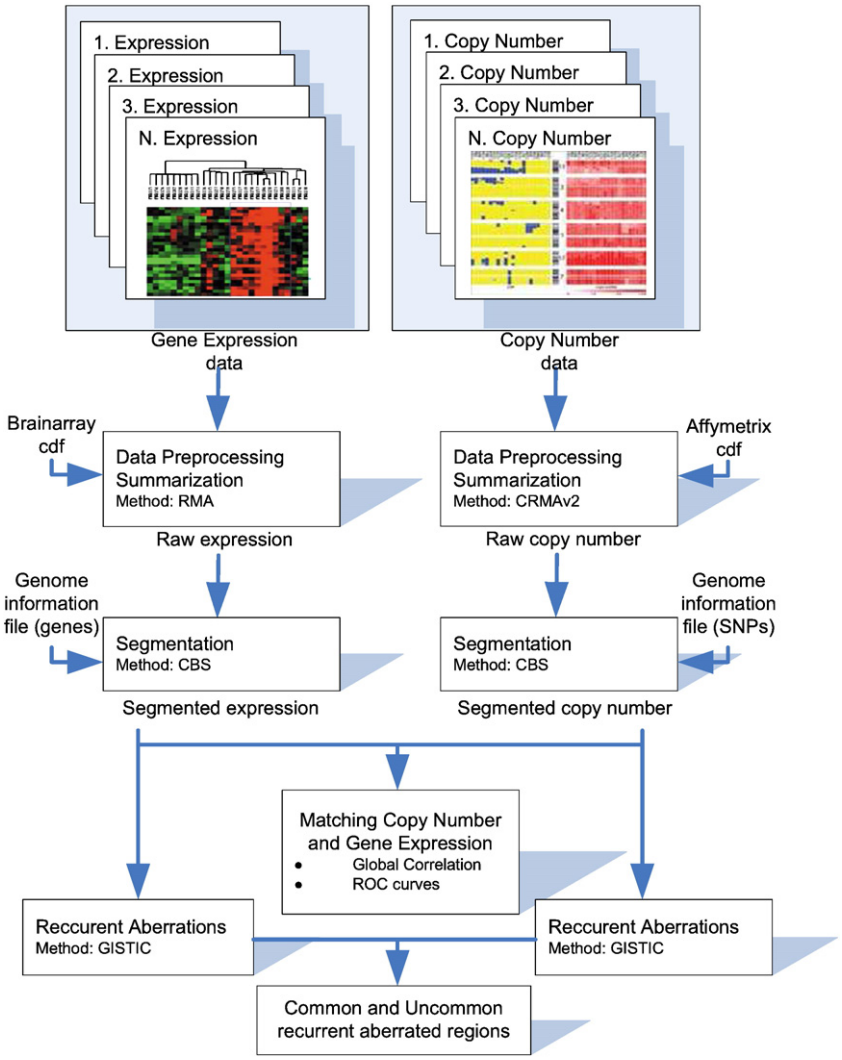


Fig. 1. Analysis work flow for CN and GE. The gene expression arrays are processed with RMA [16] and the SNPs copy number arrays are analyzed using CRMAv2 [17]. The summarization of the probes is done using a specific cdf which has the information on how to group the probes (by genes in the case of expression data or by SNPs for copy number data). Once the data have been processed, a segmentation algorithm is applied dividing the genome in regions of consecutive elements (SNPs or genes) with similar values and assigning a single value, $\log_2$ ratio of CN for SNPs and $\log_2$ ratio of expression for genes. SNP CN and gene expression data can be matched using their locations in the genome as explained in the main text. Once both identifiers are matched, we have computed the correlation (that was strongly significant) and generated the ROC curves that show the similarities between regions over/under expressed and regions with gain/loss of copy numbers. Finally, GISTIC, an algorithm to detect recurrent aberrated regions is applied to the segmented data (sCN and sGE) in order to find altered loci in a significant group of samples and the results are compared. Pipelines for CN and GE data are completely independent.

2.1. Data

Two different studies have been used to validate our method. Both include measurements of genome-wide copy number and expression for each sample. The first dataset is a study of brain tumors carried out by Kotliarov et al. [2]. We used a subgroup of 64 cases (listed in the supplementary material) from the whole dataset related with glioblastoma multiforme (GBM). The second dataset consists of 28 cases of pediatric acute lymphoblastic leukemia (ALL) from the GSE10792 [3]. These leukemia samples are listed in the supplementary material.

2.2. Material and data preprocessing

Gene expression data have been analyzed using RMA [16] over the HGU133 plus2 array. The *chip definition file* (cdf) to perform the analysis was downloaded from version 10 of Brainarray [18], which corresponds to version 46 of Ensembl genes and genomes.

Affymetrix Human Mapping 50K SNP array has been used for the analysis of the CN data. The cdf file needed in this case is the *Affymetrix* GeneChip Mapping 50K Xba240_SNP array cdf. CRMAv2 [17] has been applied to the CN signals in order to obtain the raw CN.

The analysis of both types of data have been performed under R [19] using the aroma.affymetrix package [20].

2.3. Segmentation

A segmentation algorithm divides a set of ordered data into regions of adjacent elements which have similar values. Each region is assigned a single value which represents all the data that belong to it. Segmentation methods are a family of algorithms that were initially applied to image analysis and, more recently, to genomic data.

There are several algorithms to segment genomic data such as circular binary segmentation (CBS) [21,22], CGHseq [23], GLAD [24] or HAAR [25], among others.

CBS has been chosen in this experiment because different independent comparisons [12,26] have proved that it is an accurate method. It is also widely used and implemented both in Matlab and R.

Before the segmentation algorithm is applied, the raw data are normalized dividing by the median of the samples from each element (SNP or gene) and computing its $\log_2$,

$$\Delta\, CN_{i,j} = \log_2\left(\mathrm{CN}_{i,j} / \mathrm{median}\left(CN_{i,1:n}\right)\right) \tag{1}$$

$$\Delta\, GE_{k,j} = \log_2\left(GE_{k,j} / \mathrm{median}\left(GE_{k,1:n}\right)\right), \tag{2}$$

where i and k represent the elements (SNPs or genes, respectively), j is the sample and n is the number of samples analyzed in the experiment.

The input of the segmentation methods are the raw data (i.e., GE or CN values previously calculated) and a list with the name, chromosome and position of each of the probesets of the array.

CBS proceeds as follows. It considers each of the chromosomes as a "ring": both extremes of the chromosomes are assumed to be connected. Each ring is split in two parts and the copy numbers of each of these parts are compared using a t-test

$$Z_{ij} = \frac{\left(S_j - S_i\right)/(j-i) - \left(S_n - S_j + S_i\right)/(n-j+i)}{\sqrt{1/(j-i) + 1/(n-j-i)}} \tag{3}$$

for each pair of positions i, j. S_n is the sum of the raw copy number data from the 1st SNP to the nth SNP. The method is based on the statistic $Z_C = \max_{1<i<j<n}|Z_{ij}|$. If the Z_C is above a threshold, established using a bootstrap method, then a new segment is found. The same algorithm is applied recursively to each of the found segments.

Using bootstrap to select the threshold is time-consuming. The authors of CBS derived an estimation of the threshold one order of magnitude faster [22]. Recently, they have developed an even faster version.

The output of the segmentation method is sGE and sCN for gene expression and copy numbers. respectively. sGE and sCN are matrices with constant values along the positions in the genome for points (genes or SNPs) within the same segment.

As shown in Eq. (2), each gene is previously normalized by the median of its expression across the samples. Therefore, sGE provides regions of the genome that show their expression upregulated or downregulated if compared with the median across the samples. sCN has been normalized in a similar way.

Segmentation of GE shows a specific problem that does not occur with CN data: a gene is itself a "segment" of the genome, not a single point, as with a SNP. RMA (like other summarization algorithms) provides an estimation of the concentration of the whole gene and as indicated before, a segmentation method needs a file with the position of the elements. Then, when dealing with GE data, a single position point has to be assigned to each gene. We decided to use the middle point of the genes as their representative positions (this middle point is calculated for each gene as: $\mathrm{gene}^{\mathrm{pos}}_{\mathrm{middle}} = (\mathrm{gene}^{\mathrm{pos}}_{\mathrm{end}} + \mathrm{gene}^{\mathrm{pos}}_{\mathrm{start}})/2$).

Most of the segmentation methods are able to adapt their accuracy according to the noise level of the data: if the data is clean, narrower segments can be detected. If the data is noisier, only broad segments are statistically significant. Since GE is affected by other factors different from CN, GE data is noisier than CN data and the segmentation algorithm is expected to provide broader segments.

We used CBS algorithm with the default parameters for GE segmentation, adapting the input data to the gene signals as indicated above. The genome information of the genes was generated using ENSMART [27]. The file and the code to generate it are included as supplementary material.

2.4. Recurrent aberrations in chromosomal segments

There are different algorithms to find recurrent segments with aberrations such as STAC [28], SIRAC [29] or GISTIC [30], among others. For a review of different methods, the reader can consult [31]. All of them, based on different statistical techniques, look for regions with aberrations that occur in a significant number of samples.

We have selected GISTIC [30] for our study. GISTIC is a freely available method that distinguishes random background from true aberrations. It takes into account the values of CN and lets the user set the different parameters to find deletions/amplifications and the p-value to determine if an aberration is recurrent.

GISTIC, after a careful analysis of the samples (to exclude duplicates or noisy samples), accepts as input sCN values. In its first stage, a statistic for each SNP is computed as follows:

$$G_i^{\mathrm{amp}} = \frac{1}{n}\sum c_{ij} I\left(c_{ij} > \theta^{\mathrm{amp}}\right), \tag{4}$$

where c_{ij} is the $\log_2$ ratio of the CN, I is an indicator function that equals 1 if its argument is true and 0 otherwise, and θ^{amp} is a threshold to consider that a locus has an amplification. This statistic takes into account both the strength of the amplification as its frequency. Using a semi-exact approximation, to avoid a computer intensive bootstrap, GISTIC identifies the regions of the genome were G_i^{amp} is statistically significant, with FDR correction for multiple hypothesis. Once the recurrent regions are selected, GISTIC identifies the peak or peaks (if the region shows a multimodal distribution). Genes that are located in these peaks are suggested by GISTIC to be the targets of the recurrent amplifications. The same algorithm is used to identify the deletions.

GISTIC was developed to be applied to sCN, but in our experiments, we also applied it to sGE to find recurrent over-expressed or under-expressed regions in the genome. The thresholds established for this experiment are the default values in GISTIC. As a result, GISTIC returns a list of the recurrent regions with their statistical significances.

2.5. Matching DNA copy number and gene expression data

The process to match both identifiers (genes and SNPs) is not straightforward because there are genes that have several SNPs in the array and there are genes with no SNPs. Moreover, there are SNPs in the intergenic regions that do not match any gene coding region.

Our first attempt was a direct assignation using *Ensembl* database [32]. Only the genes that had SNPs assigned in the database were matched and if there was a CNA within a gene, the median of the CN values for all the SNPs belonging to that gene were calculated. With this assignation only 50% of the genes had at least one corresponding SNP in the SNP array. This loss of data drove us to use a different approach.

Considering the sCN, the whole genome is divided into regions depending on their sCN values. Then, using the position of a gene, their corresponding sCN value can be assigned. So, even if a gene has no SNPs, it always lies in a region that has a predicted sCN (by the segmentation method). Fig. 2 illustrates this point. Gene A is located in a segment of the genome with several SNPs mapped to it and a single CN estimation (signal about −0.5). This is the CN value assigned to gene A. Gene B has no SNPs located close to it, but the zone of the genome where it is located has an assigned CN value. Finally, gene C belongs to two different segments. In this last case, we have considered the CN of a single point for each gene located at the center point of its genomic position.

3. Results

As indicated in Section 2, we have used two different datasets of cancer samples that have matched CN and GE data from genome wide microarrays. One dataset is from a GBM study and the other one from an ALL study. Since both studies have matched samples, we have compared sGE with sCN with two different validation methods. Firstly, we have generated a sCN matrix with the dimensions genes by samples in order to check if the CNAs affect the sGE. We have done this using ROC curves. And the second method uses GISTIC to look for recurrent aberrated regions in each of the data-types and check the similarities and differences within the results.

Figs. 3 and 4 show the results from both datasets. Both figures show the recurrent amplified regions (red) and the recurrent over-expressed zones (pink). In GBM (Fig. 3), there are regions such as chromosomes 10, 19, 20 and 22 which are amplified and over-expressed and all of them have been previously published in different articles [33,35]. The recurrent deleted and under-expressed regions from GBM are shown in Fig. 3 in dark and light green (negative part of the plot). For example, arm 4q and regions 6q25, 11p15, 14q and 19p13 have been described as LOH regions because of the deletions [14,15]. All of the LOH regions seem to affect the GE of genes lying in those zones of the genome. On the other hand, there are some recurrent CN aberrated regions that do not seem to affect sGE. This can be due to the fact that they are so small that they are missed by the segmentation method, or because these regions are regulated by other factors that minimize the effect of CN.

From the ALL data we can see in Fig. 4 that chromosomes 4, 6, 14, 17, 18, and 21 are amplified and also found as recurrently over-expressed. Small regions of arm p of chromosomes 7, 16 and 19 are also selected using both types of data. Chromosome 1 seems to have some genes over-expressed independently from sCN. The negative part of the figure shows common aberrations in chromosome arms 7p, 9p, 20q and the whole chromosome 21. These regions are both recurrently deleted and under-expressed. There are also some recurrent CN aberrated regions that do not seem to affect sGE.

3.1. Similarities between sCN and sGE

After performing the matching between sCN and sGE, both types of data have the same dimensions (genes by samples). The Pearson correlation coefficient between them in GBM was 0.60, with a strong statistical significance ($p<2.210^{-16}$). Correlation for the ALL dataset was weaker (0.19) but still strongly significant ($p<2.210^{-16}$).

Figs. 5 and 6 show the ROC curves for three different tests. These tests check how CNAs affect sGE based on ROC curves. Firstly, we generated a ROC curve to test if the CN amplifications affect sGE. In order to do this, we considered that there was an amplification if the measured CN was larger than 2.5. This threshold (for all the samples) gave a set of loci that showed amplifications. After that, different thresholds for sGE data were set. Then, for a particular threshold there were true positives (TP, i.e., loci with amplifications and also over-expressed), false positives (FP, i.e., loci over-expressed and not amplified) and, with equivalent definitions, true negatives (TN) and false negatives (FN). Having these values, the true positive rate (TPR) and the false positive rate (FPR) were calculated as: TPR=TP/(TP+FN) and FPR=FP/(FP+TN). Second, the same reasoning was applied to

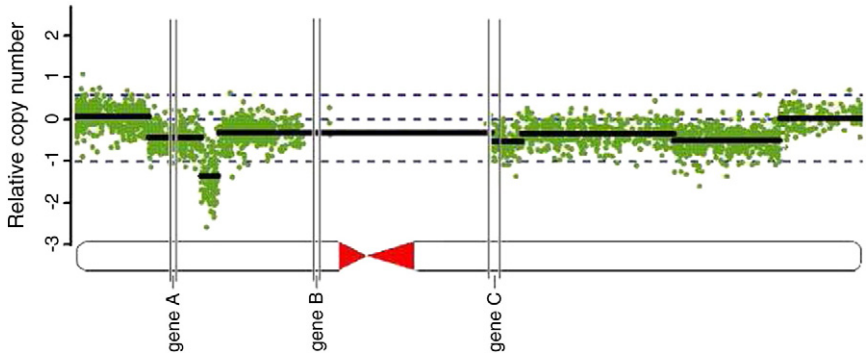


Fig. 2. Mapping between genes and sCN regions. This step is performed in order to assign a CN value to each gene and test how both (CN and GE) matrices behave. On the *Y* axis the $\log_2$ ratio of the CN are represented and on the *X* axis the genomic position. The green dots are the raw copy number values, the black horizontal lines are the segmented copy numbers and the dotted black lines are the "expected" values of a gain to three copies ($\log_2(3/2) = 0.58$), a normal region ($\log_2(2/2) = 0$) and a deletion to one copy ($\log_2(1/2) = -1$). However, these values are not obtained when using real CN data, this can be due to saturation of the probes, normalization methods or contamination with normal tissue. Once the CNs have been segmented, the genome is divided into regions of SNPs with an assigned number of copies. After that, the assignation of a CN value to each gene is performed based on its genomic position. The procedure for different cases is indicated in the figure: gene A includes several SNPs which belong to a region of the segmentation; gene B has no SNPs but its CN can be estimated using the segmented data; finally, gene C displays a possible problem because there are two regions with different CN values within it. In this experiment, if there is a CNAs within a gene we have simply assigned the CN that corresponds to the center point of the gene.

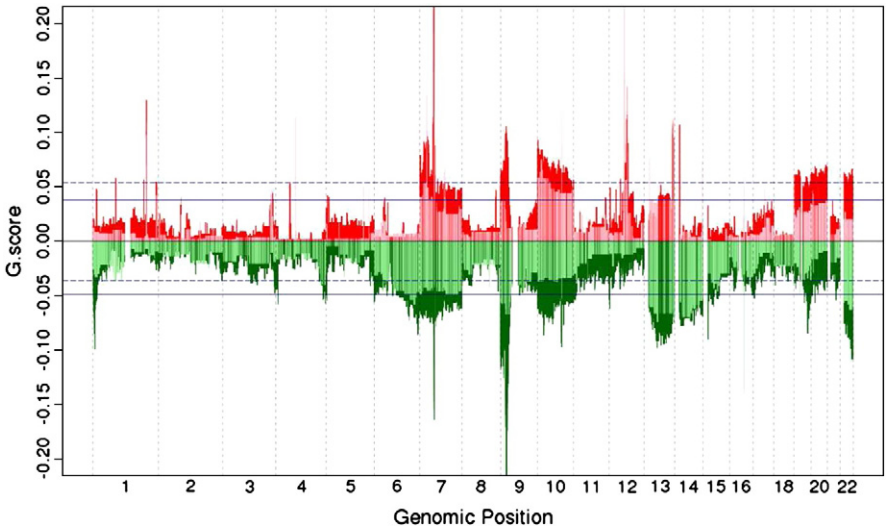


Fig. 3. Output of the GISTIC package for the GBM dataset using sCN and sGE. One of the outputs of GISTIC is a file where the recurrent regions it founds are assigned a *g*-score (value calculated by GISTIC related to the *q*-value). In this figure the positive part of the *Y* axis represents the *g*-scores given to the amplified/over-expressed regions and the negative part represents the (−)*g*-scores given to the deleted/under-expressed regions. We have done this change of sign in order to show both graphs in the same figure. On the other hand, the *X* axis represents the genomic position grouped by chromosomes. There also are four horizontal lines that show the thresholds to consider a *g*-score significant (FDR equals to 25%). The two straight lines highlight the threshold corresponding to sCN, and the dotted lines highlight the ones corresponding to sGE. Depending on the aberration under study (amplifications/over-expressions, deletions/under-expressions) they are respectively plotted in the positive and negative part of the figure. Only the autosomes are shown here. The positive part of Fig. 3 shows the values obtained using GISTIC with both types of data. It shows the recurrent amplified regions (red) and the recurrent over-expressed zones (pink). Regions such as chromosomes 10, 19, 20 and 22 are amplified and over-expressed and all of them have been previously published in different articles [33,34]. The recurrent deleted and under-expressed regions are shown in Fig. 3 in dark and light green (negative part of the plot). For example, arm 4q and regions 6q25, 11p15, 14q and 19p13 have been described as a LOH regions [14,15] and all of them seem to affect the GE of genes lying in those zones of the genome. On the other hand, there are some recurrent CN aberrated regions that do not seem to affect sGE, this can be due to the fact that they are missed by the segmentation method, because of the large level of noise, or because these regions are regulated by other factors that minimize the effect of CN. The *Y* axis were clipped to 0.2 and −0.2 to ease the comparison with the other dataset. The (maximum, minimum) values of the *g*-scores for sCN and sGE are (0.880, −0.391) and (0.426, −0.140) respectively. The names of chromosomes 17, 19 and 21 are omitted owing to lack of space.

check deletions (considering a CN deletion when CN value was equal or lower than 1.5). Finally, a third test to check strong amplifications (four or more copies) is also performed. These figures also include a ROC curve obtained using raw GE (instead of sGE) to test gains.

Fig. 5 shows for all the curves that use sGE, a high TPR (around 0.70) compared to a much lower (less than 0.05) FPR. These curves show a very steep slope for low values of FPR in contrast to the ROC curve obtained using raw GE. They can be interpreted as follows: there are very few loci that show changes in sGE that do not directly correspond to CNAs, i.e., most of the changes in sGE occur due to a change in the CN (although there are some exceptions). However, there are some CNAs that do not show the corresponding alteration in

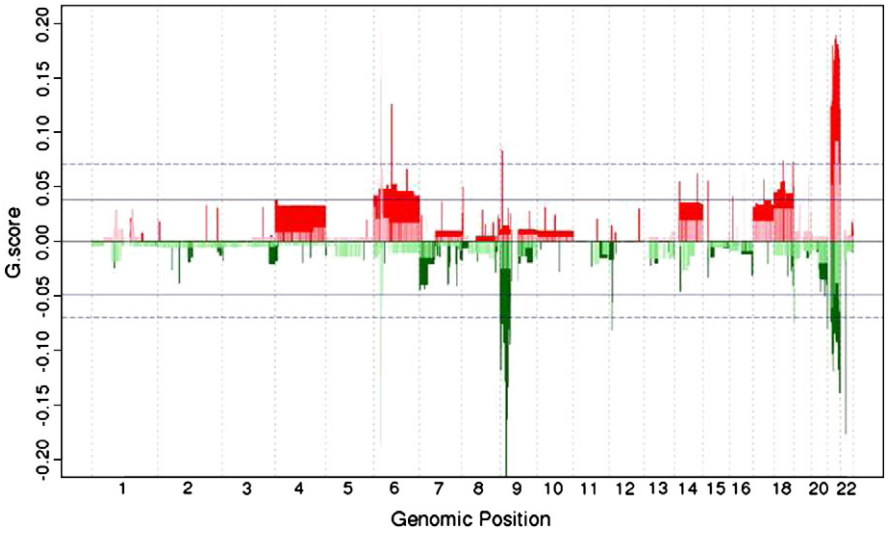


Fig. 4. GISTIC applied to the ALL data. In this figure the positive part of the *Y* axis represents the *g*-scores given to the amplified/over-expressed regions and the negative part represents the (−)*g*-scores given to the deleted/under-expressed regions. Depending on the aberration under study (amplifications/over-expressions, deletions/under-expressions) they are respectively plotted in the positive and negative part of the figure. As in Fig. 3, the recurrent amplified regions are shown in red and the recurrent over-expressed zones in pink, while the recurrent deleted and under-expressed regions are shown in dark and light green (negative part of the plot). In this figure we see recurrent amplifcations/over-expressions in both sCN and sGE. Chromosomes 4, 6, 14, 17, 18 and 21 are amplified and also found as recurrently over-expressed. Small regions of arm p of chromosomes 7, 16 and 19 are also selected using both types of data. Chromosome 1 seems to have some genes over-expressed independent from sCN. The negative part of the figure shows common aberrations in chromosome 7 and 9 arm p, chromosome 20 arm q and the whole chromosome 21. These regions seem to be both deleted and under-expressed. The *Y* axis were clipped to 0.2 and −0.2. The (maximum, minimum) values of the *g*-scores for sCN and sGE are (0.189, −0.370) and (0.200, −0.188) respectively. Chromosome names 17, 19 and 21 are omitted owing to lack of space.

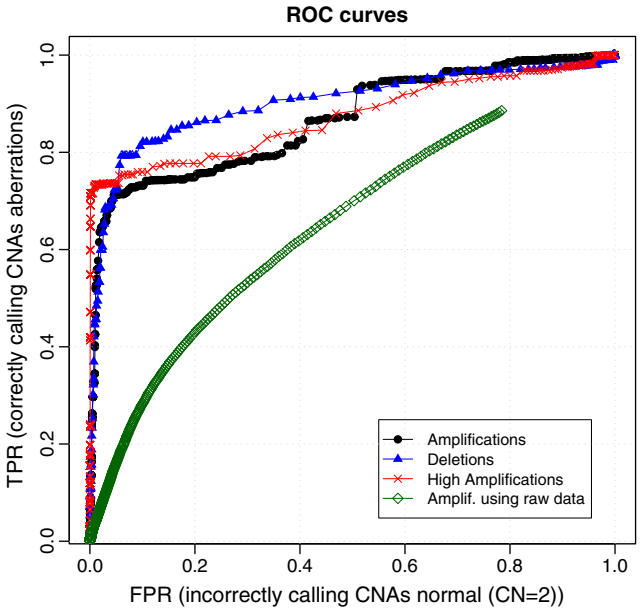


Fig. 5. Using the GBM dataset, we have calculated the ROC curves of the CNAs (amplifications (copy number equal or higher than 2.5), deletions (copy number equal or lower than 1.5) and high amplifications (four or more copies)) that also appear in sGE, and also the same analysis using raw GE. The AUC (area under curve) using raw GE is much smaller than using sGE. Amplifications and high amplifications strongly affect sGE.

sGE. This fact moves the curve downwards for larger FPRs, i.e., the curve approaches 1.0 only for large FPRs. The origin of this fact is two-fold. On one hand, the segmentation algorithm over GE cannot discover narrow segments because of the inherent variability of GE owing to the different regulators. And, on the other hand, not all the CNAs affect GE, and some genes (that do show aberrations in their CN) are regulated by other factors that minimize the effect of CN.

In Fig. 6 the ROC curves obtained from the ALL dataset are shown. As happens in GBM, here the ROC curves demonstrate that the

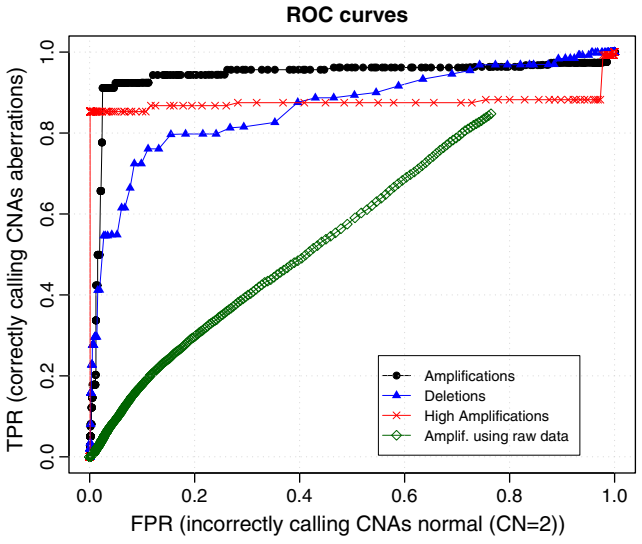


Fig. 6. As done before with GBM, this figure presents the ROC curves of the predictions of CNAs (amplifications (copy number equal or higher than 2.5), deletions (copy number equal or lower than 1.5) and high amplifications (four or more copies)) using segmented gene expression data, and the predictions of amplifications using raw expression data of pediatric acute lymphoblastic leukemia (ALL). TPR (correctly calling CNAs aberrations) is around.9 for a FPR (incorrectly calling CNAs normal (CN = 2)) equal to 0.02 for amplifications and high amplifications (black dots and red crosses). When using raw expression data (green squares) the line we get is almost diagonal.

aberrations (deletions, amplifications and high amplifications) are reflected in segmented GE more than in raw GE. TPR (correctly calling CNAs aberrations) is around 0.9 for a FPR (incorrectly calling CNAs normal (CN = 2)) equal to 0.02 for amplifications and high amplifications (black dots and red crosses). This means that 90% of all the "real CN amplifications" are found using segmented gene expression, or that almost all the regions that are amplified are also over-expressed. In the case of deletions (blue triangles), the slope is a bit lower and we get a TPR of almost 0.8 and a FPR of 0.09. When using raw expression data (green squares) ROC curve is only slightly above the diagonal, i.e., results are only slightly better than those expected by chance.

3.2. Recurrent aberrated regions in sCN and sGE data

Usually researchers are interested in the location of recurrent CNAs, i.e., amplifications or deletions, because they can be the drivers of pathology. In our case, we have used GISTIC to provide the locations of the genome especially enriched in over/under expressions (using sGE data) and also in amplifications/deletions (using sCN data).

Figs. 3 and 4 show the *g*-scores provided by GISTIC for amplified/over-expressed loci (red lines, positive part of the plot) and deleted/under-expressed (green lines, negative part of the plot). It can be seen that the significant recurrent regions (if sCN and sGE are compared) are very similar to each other, i.e., most of the recurrent aberrations affect sGE and most of the recurrent over/under-expressed loci of the genome are related with recurrent aberrations. The list provided by GISTIC with the most significant aberrated loci and a table showing the statistical ratios at different thresholds are given as supplementary material.

3.2.1. Common recurrent altered regions to sCN and sGE

In this section, we illustrate that the regions selected by GISTIC as recurrently over/under-expressed using sGE are consistent with the results from sCN, and also with the information already published about GBM. GBM is the first cancer sequenced by The Cancer Genome Atlas (TCGA [35]) and we have compared our findings with the results published by this consortium.

In the case of common regions to both analysis, it is likely that the reason for the changes in GE are the CNAs, i.e., regions with a deletion tend to be under-expressed and regions with a gain tend to be over-expressed.

Fig. 3 shows that most of the regions selected as recurrent using sCN also appear with sGE data. We independently study both the deletions/under-expressions and the gains/over-expressions.

3.2.1.1. Recurrent deleted and under-expressed regions. In this section we focus on the recurrent deleted and under-expressed regions shown in Figs. 3 and 4 in dark and light green (negative part of the plot).

3.2.1.1.1. GBM. Arm 4q and regions 6q25, 11p15, 14q and 19p13 have been described as deleted regions [14,15] and all of them seem to affect the GE of genes lying in those zones of the genome. The chromosome arm 4q has also been studied as a loss region [36,37] which can be the reason for the change in the GE data.

McLendon et al. [35] describe region 9p21.3 as the most recurrently deleted in GBM, and we show that it is also recurrent in segmented gene expression data as an under-expressed zone. PTEN (10q23.31) is known to be mutated in GBM and to have a homozygous deletion in a high percentage of samples [30,35]. Arm 15q has been reported to be deleted in GBM samples in a study performed by Vranova et al. [38]. Region 13q14 [35] and 22q13 [39] have been reported to be related with the progression of the GBM. All these regions appear to be recurrently under-expressed and deleted in our study.

3.2.1.1.2. ALL. In this dataset, the analysis of the ALL samples based on recurrent deleted/under-expressed regions (Fig. 4) shows that

chromosome 9 has a significant loss of the p arm which is also seen in sGE data. In this way, there is a clear agreement between our analyses and the results published by Bungaro et al. [3]. However, in our dataset we detect higher frequencies of the 9p deletion than Bungaro et al.

3.2.1.2. Recurrent amplified and over-expressed regions. The positive part of Figs. 3 and 4 show the recurrent amplified and over-expressed regions obtained using GISTIC with both datasets respectively. It shows the recurrent amplified regions (red) and the recurrent over-expressed zones (pink).

3.2.1.2.1. GBM. Regions such as chromosomes 10, 19, 20 and 22 are amplified and over-expressed and all of them have been previously published in different articles [34].

Gene EGFR, located at 7p11.2 has been previously found to have been activated in glioblastomas [30,35] and here it appears as recurrent in both sCN and sGE data. A narrow region in 1q32 has been reported as an amplification related with the progression of the gliomas [40]. Weber et al. [33] also associated region 5q34 with the proliferative activity of malignant glioma cell lines.

CDK4 on 12q14 is frequently amplified in GBMs [41]. Knobbe et al. [42] and Van et al. [43] reported region 13q34 as amplified and over-expressed. Finally, Korshunov et al. [44] described cytoband 14q32 as a frequently amplified region in GBMs.

3.2.1.2.2. ALL. Fig. 4 shows that chromosome 21 suffer one of the clearest alteration with a significant gain that can be also seen in sGE data too. Again, there is an agreement between our analyses and the results published by Bungaro et al. [3]. The "chr 21 amplification" affects 64.2% of the cases. Moreover, the method here proposed also allows to detect significant alterations in other chromosomes that were not indicated by Bungaro et al. and that also occur in both data types (sCN and sGE), as the gain of chromosomes 6, 17 and 18 (see Fig. 4) which have already been published [45].

3.2.2. Noncommon recurrent altered regions

In addition to the over/under-expressed regions caused by CNAs, there are zones in sGE which appear to be correlated with the position in the genome but not with the observed CNAs. These regions can appear due to other causes different from sCN (as methylation, LOH with neutral copy number or clusters of genes regulated for the same factor). In our case, a LOH analysis (for loci with neutral CN) was executed and no significant results were found.

3.2.2.1. Recurrent over/under-expressed regions

3.2.2.1.1. GBM. The closest gene to region 15q21.1 is THBS1 which appears to be under-expressed. It has been identified as a methylated tumor suppressor in different cancers [46]. This fact could be the reason why this region appears more clearly when studying sGE than with sCN. There also are a group of loci where different gene families are located that seem to have their expression altered, as MT1 gene family (16q12.2), IRX (5p15.33), NEF (8p21.2) and CXC (4q13.3). These results can be due to a common regulation factor that impacts the whole gene family and can be found using sGE.

3.2.2.1.2. ALL. In the ALL data there are also a number of genomic regions which seem to have the genes within them be affected by a factor related with the position in the genome. These regions are mostly in chromosomes 1, 5, 13 and 19, and most of them have already been studied [47,48].

4. Discussion

In this study, we propose to segment GE data derived from genome-wide expression microarrays. Segmentation of gene expression tends to reduce the effects not related with the position in the genome: if the regulators are not related with the genomic position, upregulation and downregulation of close genes along the genome will tend to cancel out each other. Therefore, sGE is an indirect measurement of the effect on GE of the regulators related with the genomic position. In Section 3, it is shown that one of the effects related with the genomic position and in fact the most important one is CNA.

None of the datasets (both GBM and ALL) includes reference samples. The normalization of the CN and GE estimates has been done using the median of all the samples. This method is valid if most of the samples behave normally. However, in the case of very frequent aberrations, the value of this reference can be biased towards the direction of the alteration. This is the reason why, for example, chromosome 7 in the GBM dataset appears to be both recurrently deleted and amplified (when it is known to be amplified in GBM). If the studies include reference samples it is advisable to use them both for CN and GE normalization, i.e., in Eqs. (1) and (2), the median that appears in the denominator must be performed over the reference samples instead of all the samples.

sGE and sCN data have a close relationship as shown by the global correlation between sGE and sCN which is strongly significant. In addition, the ROC curves of the CNAs and sGE show that, depending on the threshold, it is possible to get specificities and sensitivities over 75% (Figs. 5 and 6). Figs. 3 and 4 also show that most of the recurrent aberrated regions commonly occur in both types of data.

GE, as expected, has a very strong variation across the genome since many factors that affect GE are not related with the genomic position. This additional variability is reflected in the probabilities of recurrent aberrations. As can be seen in Fig. 3, *g*-scores for sCN are larger (more significant) than for segmented expression data. Even though, the predicted recurrent regions are very similar and the overall probabilities provided by sGE are significant. We have also found recurrent "under-expressed segments" not correlated with "CN deletions." These discrepancies can be attributed to local modifications of the genome, for example, a local methylation of the genome as has been published by Stransky et al. [49] located in one of these zones where CNA and GE do not correlate. Other loci not regulated by CN are gene families affected by a common regulation factor.

In summary, at least for an exploratory analysis, sGE provides initial regions to search for possible target genes whose CNAs affect GE. In addition, the combination of sGE and sCN also provides loci uncorrelated GE/CN that can be related to other regulatory events.

Supplementary materials related to this article can be found online at doi:10.1016/j.ygeno.2010.10.008.

Acknowledgments

The authors would like to thank Jose Angel Martinez-Climent for his support and advice in the interpretation of the results.

References

[1] J.R. Pollack, T. Sørlie, C.M. Perou, C.A. Rees, S.S. Jeffrey, P.E. Lonning, R. Tibshirani, D. Botstein, A.-L. Børresen-Dale, P.O. Brown, Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors, Proc. Natl Acad. Sci. USA 99 (2002) 12963–12968.
[2] Y. Kotliarov, M.E. Steed, N. Christopher, J. Walling, Q. Su, A. Center, J. Heiss, M. Rosenblum, T. Mikkelsen, J.C. Zenklusen, H.A. Fine, High-resolution global genomic survey of 178 gliomas reveals novel regions of copy number alteration and allelic imbalances, Cancer Res. 66 (2006) 9428–9436.
[3] S. Bungaro, M. Dell'Orto, A. Zangrando, D. Basso, T. Gorletta, L. Lo Nigro, A. Leszl, B. Young, G. Basso, S. Bicciato, A. Biondi, Integration of genomic and gene expression data of childhood ALL without known aberrations identifies subgroups with specific genetic hallmarks, Genes Chromosom. Cancer (2009).
[4] E. Hyman, P. Kauraniemi, S. Hautaniemi, M. Wolf, S. Mousses, E. Rozenblum, M. Ringnér, G. Sauter, O. Monni, A. Elkahloun, O.-P. Kallioniemi, A. Kallioniemi, Impact of DNA amplification on gene expression patterns in breast cancer, Cancer Res. 62 (2002) 6240–6245.
[5] Y. Kotliarov, S. Kotliarova, N. Charong, A. Li, J. Walling, E. Aquilanti, S. Ahn, M. Steed, Q. Su, A. Center, J. Zenklusen, H. Fine, Correlation analysis between single-nucleotide polymorphism and expression arrays in gliomas identifies potentially relevant target genes, Cancer Res. (2009).

[6] D. Tsafrir, M. Bacolod, Z. Selvanayagam, I. Tsafrir, J. Shia, Z. Zeng, H. Liu, C. Krier, R.F. Stengel, F. Barany, W.L. Gerald, P.B. Paty, E. Domany, D.A. Notterman, Relationship of gene expression and chromosomal abnormalities in colorectal cancer, Cancer Res. 66 (2006) 2129–2137.
[7] D. Witten, R. Tibshirani, T. Hastie, A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis, Biostatistics 10 (2009) 515.
[8] A.-K. Jarvinen, R. Autio, S. Haapa-Paananen, M. Wolf, M. Saarela, R. Grenman, I. Leivo, O. Kallioniemi, A.A. Makitie, O. Monni, Identification of target genes in laryngeal squamous cell carcinoma by high-resolution copy number and gene expression microarray analyses, Oncogene 25 (2006) 6997–7008.
[9] I. Cifola, R. Spinelli, L. Beltrame, C. Peano, E. Fasoli, S. Ferrero, S. Bosari, S. Signorini, F. Rocco, R. Perego, V. Proserpio, F. Raimondo, P. Mocarelli, C. Battaglia, Genome-wide screening of copy number alterations and LOH events in renal cell carcinomas and integration with gene expression profile, Mol. Cancer 7 (2008) 6.
[10] F. Turkheimer, F. Roncaroli, B. Hennuy, C. Herens, M. Nguyen, D. Martin, A. Evrard, V. Bours, J. Boniver, M. Deprez, Chromosomal patterns of gene expression from microarray data: methodology, validation and clinical relevance in gliomas, BMC Bioinform. (2006).
[11] G. Hu, R. Chong, Q. Yang, Y. Wei, M. Blanco, F. Li, M. Reiss, J. Au, B. Haffty, Y. Kang, MTDH activation by 8q22 genomic gain promotes chrmoresistance and metastasis of poor-prognosis breast cancer, Cancer Cell (2009).
[12] W. Lai, M. Johnson, R. Kucherlapati, P. Park, Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data, Bioinformatics 21 (2005) 3763–3770.
[13] D. Pinkel, D. Albertson, Array comparative genomic hybridization and its applications in cancer, Nat. Genet. 37 (Suppl) (2005) S11–S17.
[14] K. Wong, Y. Tsang, Y. Chang, J. Su, A. Di Francesco, D. Meco, R. Riccardi, L. Perlaky, R. Dauser, A. Adesina, M. Bhattacharjee, M. Chintagumpala, Genome-wide allelic imbalance analysis of pediatric gliomas by single nucleotide polymorphic allele array, Cancer Res. (2006).
[15] J. Boulay, U. Stiefel, E. Taylor, B. Dolder, A. Merlo, F. Hirth, Loss of heterozygosity of TRIM3 in malignant gliomas, BMC Cancer 9 (2009) 71.
[16] R. Irizarry, B. Hobbs, F. Collin, Y. Beazer-Barclay, K. Antonellis, U. Scherf, T. Speed, Exploration, normalization, and summaries of high density oligonucleotide array probe level data, Biostatistics 4 (2003).
[17] H. Bengtsson, P. Wirapati, T. Speed, H. Bengtsson, P. Wirapati, T. Speed, A single-array preprocessing method for estimating full-resolution raw copy numbers from all Affymetrix genotyping arrays including GenomeWideSNP 5 & 6, Bioinformatics (2009).
[18] Brainarray, http://brainarray.mbni.med.umich.edu, 2002.
[19] R Development Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2008. ISBN 3-900051-07-0.
[20] H. Bengtsson, K. Simpson, J. Bullard, K. Hansen, aroma.affymetrix: A generic framework in R for analyzing small to very large Affymetrix data sets in bounded memory, Tech Report 745, Department of Statistics, University of California, Berkeley, 2008.
[21] A. Olshen, E. Venkatraman, R. Lucito, M. Wigler, Circular binary segmentation for the analysis of array-based DNA copy number data, Biostatistics 5 (2004) 557–572.
[22] E. Venkatraman, A. Olshen, A faster circular binary segmentation algorithm for the analysis of array CGH data, Bioinformatics 23 (2007) 657–663.
[23] F. Picard, S. Robin, M. Lavielle, C. Vaisse, J. Daudin, A statistical approach for array CGH data analysis, BMC Bioinform. 6 (2005) 27.
[24] P. Hupe, N. Stransky, J. Thiery, F. Radvanyi, E. Barillot, Analysis of array CGH data: from signal ratio to gain and loss of DNA regions, Bioinformatics 20 (2004) 3413–3422.
[25] E. Ben-Yaacov, Y. Eldar, A fast and flexible method for the segmentation of aCGH data, Bioinformatics 24 (2008) i139–i145.
[26] H. Willenbrock, J. Fridlyand, A comparison study: applying segmentation to array CGH data for downstream analyses, Bioinformatics 21 (2005) 4084–4091.
[27] A. Kasprzyk, D. Keefe, D. Smedley, D. London, W. Spooner, C. Melsopp, M. Hammond, P. Rocca-Serra, T. Cox, E. Birney, EnsMart: A generic system for fast and flexible access to biological data, Genome Res. (2004).
[28] S. Diskin, T. Eck, J. Greshock, Y. Mosse, T. Naylor, C.J. Stoeckert Jr., B. Weber, J. Maris, G. Grant, STAC: A method for testing the significance of DNA copy number aberrations across multiple array-CGH experiments, Genome Res. (2006).
[29] C. Lai, H. Horlings, M. van de Vijver, E. van Beers, P. Nederlof, L. Wessels, M. Reinders, SIRAC: Supervised Identification of Regions of Aberration in aCGH datasets, BMC Bioinform. 8 (2007).
[30] R. Beroukhim, G. Getz, L. Nghiemphu, J. Barretina, T. Hsueh, D. Linhart, I. Vivanco, J. Lee, J. Huang, S. Alexander, J. Du, Assesing the significance of chromosomal aberrations in cancer: Methodology and application to glioma, Proc. Natl Acad. Sci. USA (2007).
[31] O. Rueda, R. Diaz-Uriarte, Finding recurrent copy number alteration regions: a review of methods, Curr. Bioinf. 5 (2010) 1–17.
[32] Ensembl, http://www.ensembl.org/index.html, 2000.
[33] R. Weber, M. Sabel, J. Reifenberger, C. Sommer, J. Oberstrass, G. Reifenberger, M. Kiessling, T. Cremer, Characterization of genomic alterations associated with glioma progression by comparative genomic hybridization, Oncogene 13 (1996) 983–994.
[34] E. Burton, K. Lamborn, B. Feuerstein, M. Prados, J. Scott, P. Forsyth, S. Passe, R. Jenkins, K. Aldape, Genetic aberrations defined by comparative genomic hybridization distinguish long-term from typical survivors of glioblastoma, Cancer Res. (2002).
[35] R. McLendon, A. Friedman, D. Bigner, et al., Comprehensive genomic characterization defines human glioblastoma genes and core pathways, Nature 455 (2008) 1061–1068.
[36] Y. Li, C. Tzeng, J. Song, F. Tsia, L. Hsieh, S. Liao, C. Tsai, E. Van Meir, C. Hao, C. Lin, Genomic alterations in human malignant glioma cells associate with the cell resistance to the combination treatment with tumor necrosis factor-related apoptosis-inducing ligand and chemotherapy, Clin. Cancer Res. 12 (2006) 2716–2729.
[37] A. Idbaih, R. Carvalho Silva, E. Crinière, Y. Marie, C. Carpentier, B. Boisselier, S. Taillibert, A. Rousseau, K. Mokhtari, F. Ducray, J. Thillet, Genomic changes in progression of low-grade gliomas, J. Neurooncol. (2008).
[38] V. Vranova, E. Necesalova, P. Kuglak, P. Cejpek, M. Pesakova, E. Budanska, J. Relichova, R. Veselska, Screening of genomic imbalances in gliobastoma multiforme using high-resolution comparative genomic hybridization, Oncol. Rep. (2007).
[39] M. Nakamura, E. Ishida, K. Shimada, M. Kishi, H. Nakase, T. Sakaki, N. Konishi, Frequent LOH on 22q12.3 and TIMP-3 inactivation occur in the progression to secondary glioblastomas, Lab. Investig. 85 (2005) 165–175.
[40] G. Roversi, R. Pfundt, R. Moroni, I. Magnani, S. van Reijmersdal, B. Pollo, H. Straatman, L. Larizza, E. Schoenmakers, Identification of novel genomic markers related to progression to glioblastoma through genomic profiling of 25 primary glioma cell lines, Oncogene (2006).
[41] C. Knobbe, A. Trampe-Kieslich, G. Reifenberger, Genetic alteration and expression of the phosphoinositol-3-kinase/Akt pathway genes PIK3CA and PIKE in human glioblastomas, Neuropathol. Appl. Neurobiol. 31 (2005) 486–490.
[42] G.R. Christiane, B. Knobbe, Genetic alterations and aberrant expression of genes related to the phosphatidyl-lnositol-3'-kinase/protein kinase B (Akt) signal transduction pathway in glioblastomas, Number Brain Pathol. (2003).
[43] J. van den Boom, M. Wolter, R. Kuick, D. Misek, A. Youkilis, D. Wechsler, C. Sommer, G. Reifenberger, S. Hanash, Characterization of gene expression profiles associated with glioma progression using oligonucleotide-based microarray analysis and real-time reverse transcription-polymerase chain reaction, Am. J. Pathol. (2003).
[44] A. Korshunov, R. Sycheva, A. Golanov, Genetically distinct and clinically relevant subtypes of glioblastoma defined by array-based comparative genomic hybridization (array-CGH), Acta Neuropathol. 111 (2006) 465–474.
[45] N. Kawamata, S. Ogawa, M. Zimmermann, M. Kato, M. Sanada, K. Hemminki, G. Yamatomo, Y. Nannya, R. Koehler, T. Flohr, et al., Molecular allelokaryotyping of pediatric acute lymphoblastic leukemias by high-resolution single nucleotide polymorphism oligonucleotide genomic microarray, Blood 111 (2008) 776.
[46] W. Park, J. Park, R. Oh, N. Yoo, S. Lee, M. Shin, H. Lee, S. Han, S. Yoon, S. Kim, C. Choi, P. Kim, A distinct tumor suppressor gene locus on chromosome 15q21.1 in sporadic form of colorectal cancer, Cancer Res. (2000).
[47] J. Davidsson, A. Andersson, K. Paulsson, M. Heidenblad, M. Isaksson, A. Borg, J. Heldrup, M. Behrendtz, I. Panagopoulos, T. Fioretos, et al., Tiling resolution array comparative genomic hybridization, expression and methylation analyses of dup (1q) in Burkitt lymphomas and pediatric high hyperdiploid acute lymphoblastic leukemias reveal clustered near-centromeric breakpoints and overexpression of genes in 1q22-32.3, Hum. Mol. Genet. 16 (2007) 2215.
[48] H. Cave, S. Suciu, C. Preudhomme, B. Poppe, A. Robert, A. Uyttebroeck, M. Malet, P. Boutard, Y. Benoit, L. Mauvieux, et al., Clinical significance of HOX11L2 expression linked to t (5; 14)(q35; q32), of HOX11 expression, and of SIL-TAL fusion in childhood T-cell malignancies: results of EORTC studies 58881 and 58951, Blood 103 (2004) 442.
[49] N. Stransky, C. Vallot, F. Reyal, I. Bernard-Pierrot, S. de Medina, R. Segraves, Y. de Rycke, P. Elvin, A. Cassidy, C. Spraggon, A. Graham, J. Southgate, B. Asselain, Y. Allory, C.C. Abbou, D.G. Albertson, J.-P. Thiery, D.K. Chopin, D. Pinkel, F. Radvanyi, Regional copy number–independent deregulation of transcription in cancer, Nat. Genet. 38 (2006) 1386–1396.

Fontanillo et al. BMC Genomics 2012, 13(Suppl 5):S5
http://www.biomedcentral.com/1471-2164/13/S5/S5

BMC Genomics

RESEARCH Open Access

Combined analysis of genome-wide expression and copy number profiles to identify key altered genomic regions in cancer

Celia Fontanillo[1], Sara Aibar[1], Jose Manuel Sanchez-Santos[2], Javier De Las Rivas[1*]

From X-meeting 2011 - International Conference on the Brazilian Association for Bioinformatics and Computational Biology
Florianópolis, Brazil. 12-15 October 2011

Abstract

Background: Analysis of DNA copy number alterations and gene expression changes in human samples have been used to find potential target genes in complex diseases. Recent studies have combined these two types of data using different strategies, but focusing on finding gene-based relationships. However, it has been proposed that these data can be used to identify key genomic regions, which may enclose causal genes under the assumption that disease-associated gene expression changes are caused by genomic alterations.

Results: Following this proposal, we undertake a new integrative analysis of genome-wide expression and copy number datasets. The analysis is based on the combined location of both types of signals along the genome. Our approach takes into account the genomic location in the copy number (CN) analysis and also in the gene expression (GE) analysis. To achieve this we apply a segmentation algorithm to both types of data using paired samples. Then, we perform a correlation analysis and a frequency analysis of the gene loci in the segmented CN regions and the segmented GE regions; selecting in both cases the statistically significant loci. In this way, we find CN alterations that show strong correspondence with GE changes. We applied our method to a human dataset of 64 Glioblastoma Multiforme samples finding key loci and hotspots that correspond to major alterations previously described for this type of tumors.

Conclusions: Identification of key altered genomic loci constitutes a first step to find the genes that drive the alteration in a malignant state. These driver genes can be found in regions that show high correlation in copy number alterations and expression changes.

Background

Acquisition of somatic genetic alterations plays an important role in the development of cancer. Several systematic efforts have addressed the study of genetic alterations to characterize human cancers [1,2], including: copy-number alterations (CNAs), translocations, insertions or single-nucleotide polymorphisms (SNPs). Most of these approaches are focused on finding frequent alterations, which occur in a high number of cases. According to the *selective pressure* theory, a genomic alteration that confers an advantage to a malignant state is likely to be found in more tumors than expected by chance [3]. However, most methods that look for recurrent aberrations using copy number information find many regions, containing many genes [4,5]. Therefore, to identify recurrently altered genomic regions -biologically relevant- it is necessary to integrate gene and genome information, as proposed by Akavia *et al.* [3]. Several reports have recently shown that integrative strategies can be very useful to identify driver genes, considering the hypothesis that disease-associated gene expression changes are frequently induced by genomic alterations [3,6-10].

* Correspondence: jrivas@usal.es
[1]Cancer Research Center (CIC-IBMCC), Consejo Superior de Investigaciones Científicas (CSIC), Campus Miguel de Unamuno, Salamanca, Spain
Full list of author information is available at the end of the article

Most of these reports are focused on finding gene-based relationships.

Built on these hypotheses -that relate transcriptomic and genomic alterations-, we propose a new integrative method based on the location of both types of signals along the genome. Our method takes into account the genomic loci, both in the copy number (CN) analysis and also in the gene expression (GE) analysis, and applies the segmentation step proposed by Ortiz-Estevez *et al*. [11]. These authors designed a method for robust comparison between CN and GE using paired samples. Such approach is based on a search for correlation between segmented CN regions and segmented GE regions to find the most significant simultaneous alterations. We follow this approach introducing two new steps to asses the matching between CN and GE loci: (i) first, a signal correlation analysis; (ii) second, an alteration frequency analysis. Using these analyses we propose a set of significantly altered genomic regions in the studied pathological state. In order to show the performance and demonstrate the value of our method, we use a dataset of 64 Glioblastoma Multiforme (GBM) samples with paired measurements of GE and CN (taken from [7,8]).

Results and discussion

The method is designed for combined analysis of datasets from two types of genome-wide arrays: DNA genomic microarrays and RNA expression microarrays. These arrays provide copy number and expression quantitative data, respectively. The analysis places both types of signals along the genome, taking into account the gene loci for the CN data and the GE data. The rationale of the method is to search for copy number alterations with a major influence in the expression levels of the genes encoded. As a distinctive element from other integrative approaches we do not consider only SNPs or genes individually. We take into account the gene loci following the strategy described in [11], that is based on the application of the same smoothing and segmentation algorithm to CN and GE in order to establish comparable regions. Once we get the smoothed segments, we perform two independent analyses for each gene loci: a signal correlation analysis and an alteration frequency analysis. (The workflow described in Materials and Methods, presented in last figure, illustrates the procedure of the method including these two independent analyses).

Analysis of correlation between gene expression and copy number levels

The method matches the CN and GE segmented signals within each chromosomal region -i.e., the log2 ratio signals of the corresponding segments- and selects the gene loci that show a significant correlation. These loci can be considered candidate hotspots. In Figure 1 we present the results of this analysis done for the GBM dataset, marking in purple the number of gene loci with *Pearson* Correlation Coefficient $r \geq 0.60$ (that corresponds to a Bonferroni-adjusted p-value < 0.005). Such cutoff ($r \geq 0.60$) includes around 55 % of the human gene loci, providing a good coverage with a significant p-value. Setting more stringent cutoffs reduces the coverage too much: $r \geq 0.70$ includes only ~26 % of the gene loci; $r \geq 0.80$ includes only ~6 %.

The number of probes in the SNP arrays -used to calculate the segmented signals for CN- is large and uniform along the genome. However, in the expression arrays some genomic regions do not have enough allocated gene loci and the number of probes is sparse. This fact is a problem when a GE segment includes outliers (i.e. gene locus which have expression levels very different from the mean of their neighbours). To solve this problem, we look for statistically significant outliers within the GE segments -which were at least in 1/3 of the samples- and we recalculate the signal correlation between their unsegmented GE and the corresponding CN segments. In this way, we find a new set of gene loci with correlation $r \geq 0.60$, which is added to the initial set of candidate hotspots identified. This step of the procedure is important to recover some gene loci with quite significant correlation (e.g. EGFR or SEC61G), which were missed in the first step due to the described problem.

Analysis of frequencies for the categorical states Up-Gain and Down-Loss

The method also proposes to find the genomic regions that present a significant GE and CN alteration in the same direction. To assess this, we included a second selective step based on stratification of the segmented data. The genomic regions are stratified in several categories: up-regulation (**U**), down-regulation (**D**) or no-change (**N**) for expression; and gain (**G**), loss (**L**) or no-change (**N**) for copy number. This approach allows a discretization of the genomic regions into 9 different categories as shown in Figure 2 (inserted table): **U-G**, **N-G**, **D-G**, **U-N**, **N-N**, **D-N**, **U-L**, **N-L**, **D-L**. Figure 2 also presents the empirical cumulative distributions for these 9 categories of the GBM samples per gene loci, counting the frequency of samples for all the gene loci in each category. As expected, the distributions show that the "no change" (i.e. N-N, neutral-neutral) is the most frequent state. The analysis of distributions also finds some regions that show a clear correspondence between GE and CN alterations: i.e. the scenario where GE up-regulation is observed co-located with a CN gain (**U-G** category) and the scenario where GE down-regulation is co-located with a CN loss (**D-L** category). Our interest focuses on these regions, since they are the ones altered in the same way in both types of data. The analysis of the empirical frequency distributions for

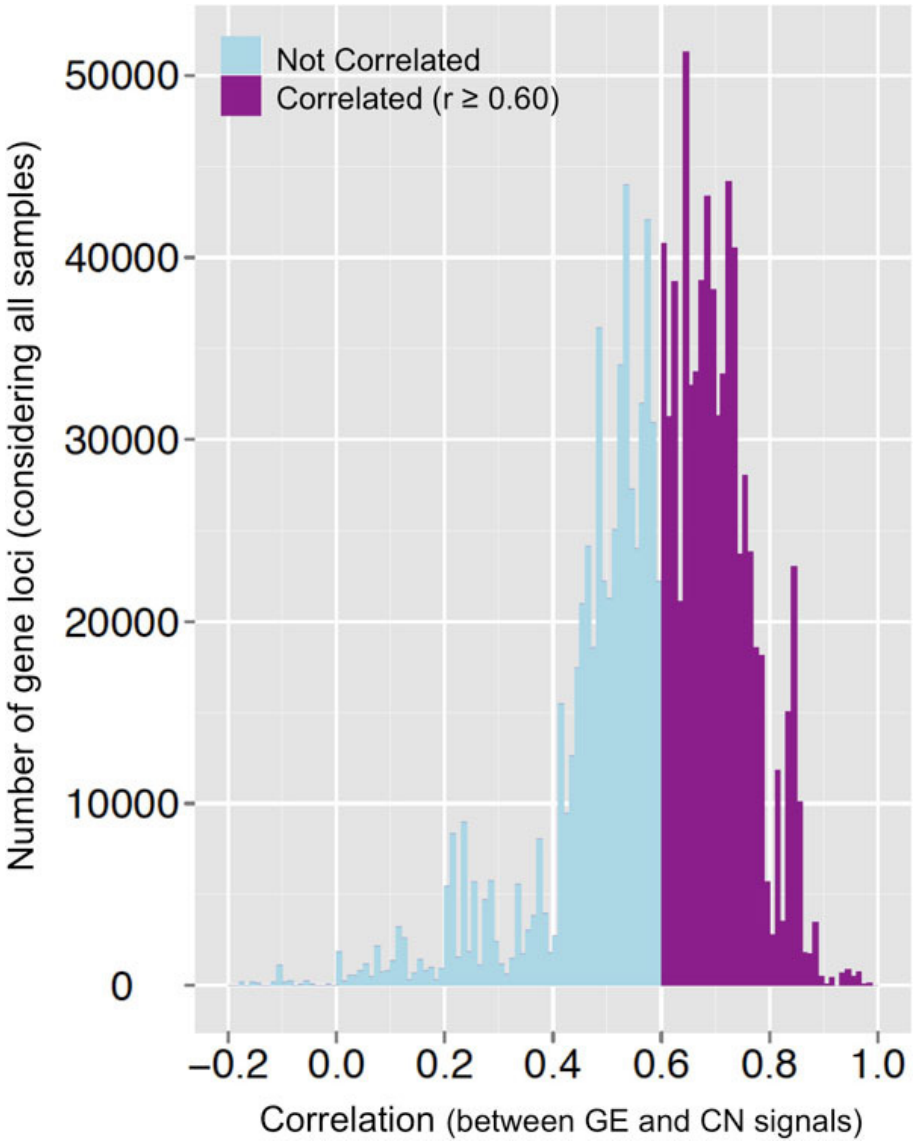


Figure 1 Density distribution of the correlation coefficients between GE and CN for the GBM dataset. Purple represents the number of gene loci that present significant correlation (r ≥ 0.60) between gene expression and copy number signals, counted considering all the samples. Blue are the rest, not considered significant.

the **U-G** and **D-L** categories allows identifying the frequency cutoffs that correspond to the 10% upper quantiles. These cutoffs were: 13 samples for **U-G** and 11 samples for **D-L** (out of 64 in GBM dataset). The set up of these thresholds identifies those genomic regions that are the most frequently assigned to such altered categories (**U-G** and **D-L**) in the studied dataset.

Genome-wide identification of hotspots: candidate key genomic regions

Our method identifies candidate key regions that show high correlation between CN and GE and that are frequently altered in the same direction, in both types of signals. The overlapping between the regions with the most significant correlation and the ones with the highest frequencies of simultaneous alteration (CN and GE) along the genome, will constitute hotspots where putative driver genes are likely to be encoded.

Figure 3 presents the combined view of GE and CN alterations on the complete genome obtained for the GBM dataset. The graph shows the alteration frequency, either in CN or in GE independently, along all the genome (22 human chromosomes). The dark colors correspond to GE up-regulated regions (red) or down-regulated regions (green), and the light colors -placed on top- correspond to CN gains (pale red) and losses (pale green). These results

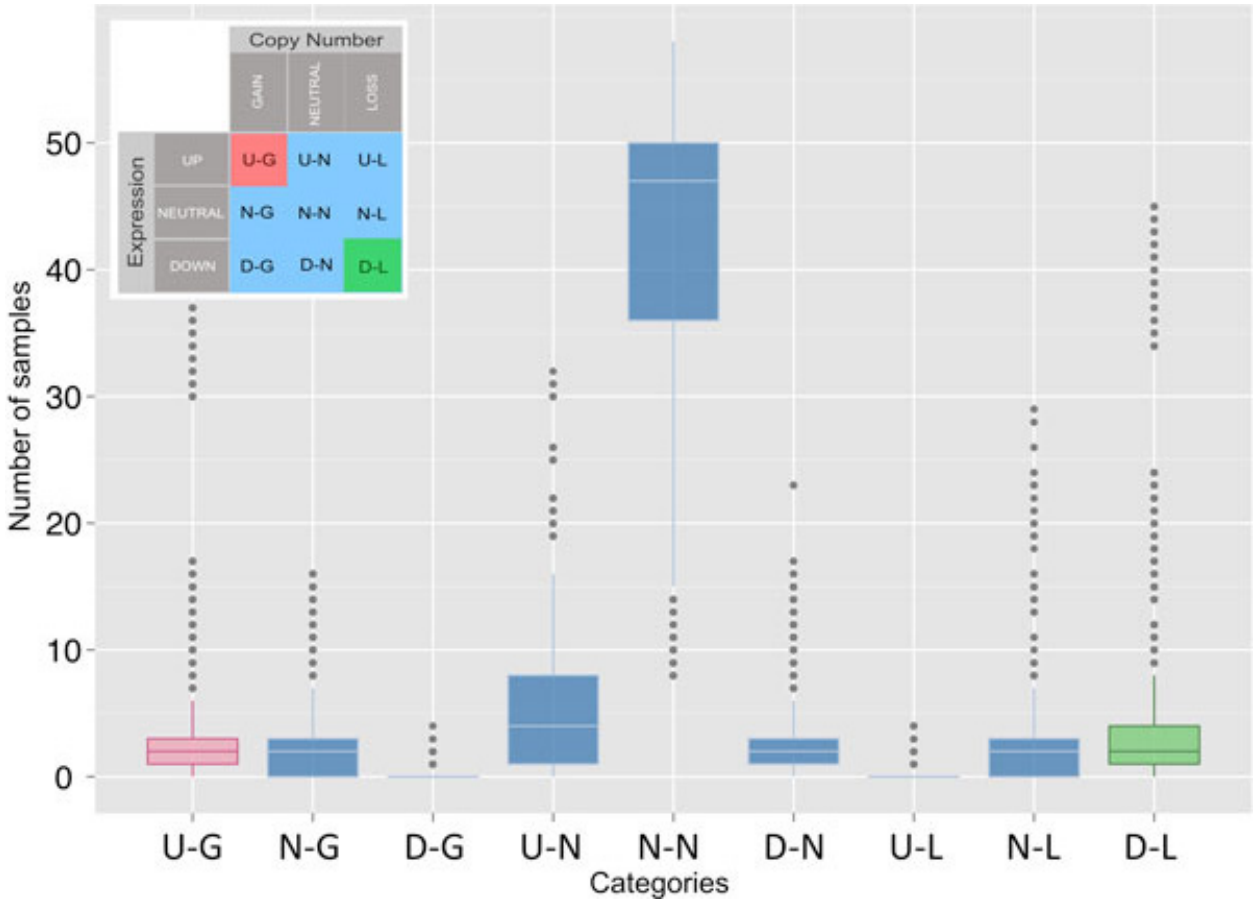


Figure 2 Observed frequency distributions of the 9 categorical states for the GBM dataset: U-G, N-G, D-G, U-N, N-N, D-N, U-L, N-L, D-L. Boxplots corresponding to the distributions of number of samples assigned to each category for all the gene loci. Insert: Contingency table showing the 9 possible categories for the gene expression and copy number. The total number of GBM samples analysed was 64.

show that the method finds the alterations previously described for CN in GBM cancer [12,13]. In fact, the most frequent alterations in glioblastoma are the gain of chromosome 7 and the loss of chromosome 10. Our analysis finds such alterations in CN, and also finds their correlation with GE up-regulation for chromosome 7 and with GE down-regulation for chromosome 10. Figure 4 presents a detailed view of the alterations that occur in chromosome 7. It includes a profile of the regions with significant correlation (purple dots along the chromosome) and a profile of the frequency of **U-G** regions (pale red). They cover nearly the complete chromosome. A figure with the representation for all the 22 human chromosomes for the GBM samples is included as Additional File 3.

Key genomic regions found for the 64 paired GBM cancer samples

As shown in Figure 3, the method presented in this work allows the identification of relevant altered genomic regions suffering significant changes in most of the GBM samples. The results also show that many of the detected CN alterations and GE changes overlap along the genome. These regions can be proposed as relevant "hotspots". In Table 1 and Table 2 we present a detailed description of the common genomic regions found in GBM; indicating the correlation and frequency of the **U-G** regions (Table 1, which includes 19 regions), and the **D-L** regions (Table 2, which includes 24 regions). The tables include the correlation between GE and CN for each region (as average correlation for all the gene loci); and the percentage of samples -frequency- in each region, counting only the samples where simultaneous GE and CN alterations occur: either up gene expression and gain in copy number (**U-G**) or down gene expression and loss in copy number (**D-L**). The regions detected are in the chromosomes that suffer the most significant changes in GBM samples: **U-G**, chr 7 and chr 20; **D-L** chr 10, chr 13, chr 14 and chr 22. The tables also include the genes enclosed in these regions. The most remarkable changes correspond to a

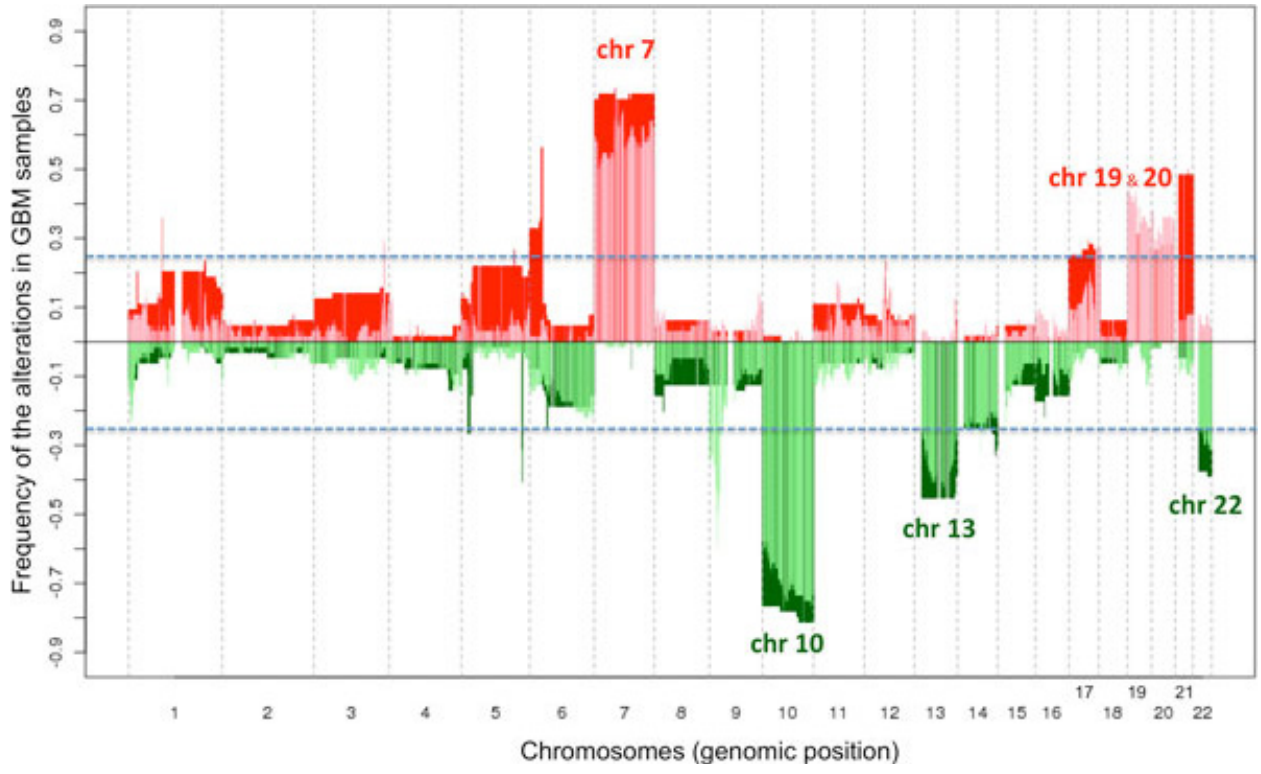


Figure 3 Combined view of GE and CN alterations obtained for the GBM dataset. The regions for all the human chromosomes (chr) that are altered either in CN or in GE are presented along the whole genome keeping the proportional size of the chromosomes. The graph shows the frequency of such alterations in the GBM samples. The colors correspond to GE up-regulated regions (in red) or down-regulated regions (in green), and -plotted on top- the CN gains (in pale red) or the CN losses (in pale green). Blue lines mark the regions that change in more than 25% of the GBM patients. The chromosomes with most significant changes (that present large regions included in the categories **U-G** or **D-L**) are labeled: **U-G** chr 7, chr 19, chr 20; **D-L** chr 10, chr 13, chr 22.

large part of chr 7 (**U-G**) and to a large part of chr 10 (**D-L**). Two important genes are precisely located in these chromosomes: EGFR (in chr 7) usually up-regulated and PTEN (in chr 10) usually down-regulated [12,13]. PTEN is not found in our analysis, but it has been reported an absence of PTEN alterations in more than half of *de novo* glioblastomas and more than 90 % of glioblastomas developed from a pre-existing lower grade gliomas [14], which has been linked to the presence of additional tumor suppressor genes on chr 10, such as LGI1 [15] and MXI1 [16]. We found these two genes in regions 8 and 10 of the **D-L** list (Table 2), and we observed a very variable profile of PTEN in the GBM samples. These facts may indicate that PTEN is not the best genomic marker for this altered

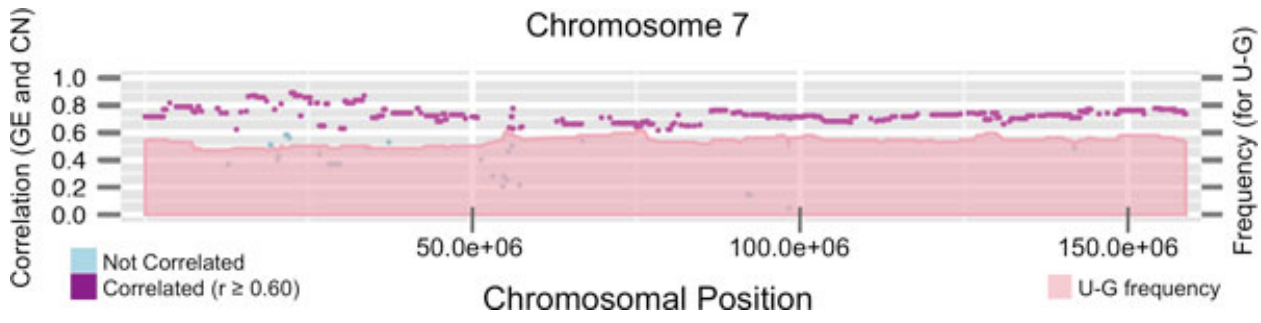


Figure 4 Detailed view of chromosome 7 showing the CN and GE correlation and the U-G category frequency for the GBM dataset. The genomic regions for chromosome 7 are represented in X-axis. Blue and purple dots show the correlation coefficients between CN and GE for each gene loci (purple when r ≥ 0.60). Pink profile represents the frequency values for the Up-Gain category **(U-G)**.

Table 1 Significant U-G regions with the associated genes.

regions	chr	cytobands	start	end	Correlation (average r coefficient)	U-G Frequency (average %)	Number of genes	gene symbols
1	**7**	p22.3,,p22.2,p22.1,...	13912	12407180	**0.75**	**53.13**	**97**	PDGFA,PRKAR1B,HEATR2,...
2	**7**	p21.2,p21.1	13980952	18581782	**0.83**	**48.34**	**16**	ETV1,DGKB,TMEM195,...
3	**7**	p21.1	19741898	19786077	**0.77**	**48.44**	**2**	TWISTNB,TMEM196
4	**7**	p21.1	20735744	20825207	**0.81**	**50.00**	**2**	ABCB5,SP8
5	**7**	p15.3,p15.2	22277336	26372745	**0.85**	**50.00**	**28**	RAPGEF5,IL6,TOMM7,...
6	**7**	p15.2	26682190	27829944	**0.68**	**50.00**	**18**	SKAP2,HOXA1,HOXA2,...
7	**7**	p14.3	29901392	33407268	**0.78**	**50.00**	**29**	WIPF3,SCRN1,FKBP14,...
8	**7**	p14.3,p14.2	34692740	36658380	**0.72**	**48.44**	**13**	NPSR1,DPY19L1,TBX20,...
9	**7**	p14.1,p13,p12.3,...	37856706	50759460	**0.72**	**49.60**	**83**	GPR141,TXNDC3,SFRP4,...
10	**7**	p11.2	54819940	54826939	**0.77 ***	**59.38**	**1**	SEC61G
11	**7**	p11.2	55086725	55275031	**0.77 ***	**60.94**	**1**	EGFR
12	**7**	p11.2	55572215	56043680	**0.70**	**59.06**	**5**	VOPP1,SEPT14,ZNF713,...
13	**7**	p11.2	56125502	56171766	**0.70**	**57.81**	**4**	CCT6A,SUMF2,PHKG1,...
14	**7**	p11.2,p11.1,q11.21	57269897	66582330	**0.66**	**56.56**	**25**	ERV3,VKORC1L1,GUSB,...
15	**7**	q11.22,q11.23,q21.11,...	69660980	91851882	**0.68**	**57.74**	**131**	AUTS2,WBSCR17,CALN1,...
16	**7**	q21.2,q21.3,	92738082	97975672	**0.72**	**56.51**	**36**	SAMD9,SAMD9L,HEPACAM2,...
17	**7**	q22.1,q22.3,q31.1,...	98456252	141707080	**0.71**	**55.94**	**343**	TMEM130,TRRAP,SMURF1,...
18	**7**	q34,q35,q36.1,...	141954920	158879258	**0.75**	**56.26**	**154**	TRBV12-2,TRBC1,PRSS1,...
19	**20**	p13,p12.3,p12.2,...	72762	62897316	**0.83**	**25.62**	**570**	DEFB125,DEFB126,DEFB12,...

Table with the significant Up-regulated and Gained (U-G) regions, indicating: chromosomal bands covered by each region; percentage of samples (%) that are in the U-G category in each region (calculated as average frequency for all the gene loci in the region); correlation between GE and CN for each region (calculated as average correlation of the gene loci in the region); number of genes located in each region. Total number of GBM samples analysed: N = 64. Marked with * the correlations calculated between unsegmented GE and segmented CN. Due to size limitations the table only includes a maximum of 3 cytobands or 3 genes. Complete information corresponding to this U-G regions is included as supplementary material: *Additional-file*-1.

Table 2 Significant D-L regions with the associated genes.

regions	chr	cytobands	start	end	Correlation (average r coefficient)	U-G Frequency (average %)	Number of genes	gene symbols
1	**10**	p14	9365826	11653762	**0.62**	**57.19**	**5**	CUGBP2,C10orf31,USP6NL,...
2	**10**	p13,p12.33,p12.31,...	16746068	24410224	**0.64**	**59.98**	**39**	RSU1,CUBN,TRDMT1,...
3	**10**	p12.1,p11.23,p11.22,...	25189544	47921720	**0.63**	**60.50**	**103**	PRTFDC1,ENKUR,THNSL1,...
4	**10**	p11.2	38238795	38265453	**0.60 ***	**60.94**	**1**	ZNF25
5	**10**	q11.22,,q11.23	49203737	53404614	**0.68**	**62.91**	**42**	FAM25C,BMS1P7,PTPN20C,...
6	**10**	q21.1,q21.2,q21.3,...	59989486	82351987	**0.65**	**64.07**	**152**	IPMK,CISD1,UBE2D1,...
7	**10**	q23.1,	84190870	87742781	**0.66**	**65.63**	**9**	NRG3,GHITM,PCDH21,...
8	**10**	q23.31,q23.32,q23.33	91498034	97024055	**0.62**	**67.19**	**39**	KIF20B,HTR7,RPP30,...
9	**10**	q24.1	97391080	97763109	**0.62**	**66.41**	**6**	ALDH18A1,TCTN3,ENTPD1,...
10	**10**	q24.1,q24.2,q24.31,...	98081203	134474152	**0.67**	**68.70**	**253**	DNTT,OPALIN,TLL2,...
11	**13**	q12.13,q12.2	27693163	28017254	**0.60**	**28.13**	**5**	USP12,RPL21,RASL11A,...
12	**13**	q12.2,,q12.3	28367434	30381664	**0.61**	**30.29**	**13**	GSX1,PDX1,ATP5EP2,...
13	**13**	q13.3,q14.11,q14.12,...	35881808	61059301	**0.73**	**35.48**	**124**	NBEA,MAB21L1,DCLK1,...
14	**13**	q21.32,q21.33	67340716	70478658	**0.62**	**35.94**	**2**	PCDH9,KLHL1
15	**13**	q22.2,q22.3,q31.1	76451567	80912598	**0.63**	**37.40**	**15**	KCTD12,IRG1,CLN5,...
16	**13**	q31.3,	92785210	94467454	**0.62**	**34.38**	**2**	GPC5,GPC6
17	**13**	q32.1,q32.2,q32.3,...	95812885	106130800	**0.64**	**33.35**	**38**	ABCC4,CLDN10,DZIP1,...
18	**13**	q33.3	108170700	108931486	**0.95**	**27.73**	**4**	FAM155A,LIG4,ABHD13,...
19	**13**	q34	110422550	111549152	**0.89**	**23.44**	**9**	IRS2,COL4A1,COL4A2,...
20	**14**	q11.2,q12,q13.1,...	19686156	70583360	**0.75**	**18.44**	**389**	TTC5,CCNB1IP1,PARP2,...
21	**14**	q24.2	71478110	72101080	**0.74**	**17.19**	**2**	PCNX,SIPA1L1
22	**14**	q24.2,,q24.3	73223406	76274521	**0.69**	**17.19**	**52**	DPF3,DCAF4,ZFYVE1,...
23	**14**	q24.3,,q31.1	77825772	80673424	**0.75**	**17.19**	**14**	TMED8,AHSA1,ISM2,...
24	**22**	q11.1,q11.21,q11.22,...	16157622	51224902	**0.73**	**24.93**	**490**	POTEH,CESK1,XKR3,...

Table with the significant Down-regulated and Lost (D-L) regions, indicating: chromosomal bands covered by each region; percentage of samples (%) that are in the D-L category in each region (calculated as average frequency for all the gene loci in the region); correlation between GE and CN for each region (calculated as average correlation of the gene loci in the region); number of genes located in each region. Total number of GBM samples analysed: N = 64. Marked with * : correlations calculated between unsegmented GE and segmented CN. Due to size limitations the table only includes a maximum of 3 cytobands or 3 genes. Complete information corresponding to this D-L regions is included as supplementary material: Additional-file-2.

region. By contrast, we found RB1 tumor suppressor in region 13 of the **D-L** list; and this gene -included in chr 13- is a clear candidate to drive the alteration of tumor cells. With respect to EGFR, it has the highest **U-G** frequency observed (60.9%, Table 1) and therefore the method reveals this gene locus as the most common GE up-regulated and CN gained in the GBM samples. The alteration of EGFR can be associated with other genes that regulate its function, also found by the method. This is the case of VOPP1 and RAB11FIP2. VOPP1 is also known as ECOP (EGFR-coamplified and overexpressed protein) or GASP (Glioblastoma-amplified secreted protein), and is found in region 12 of the **U-G** list (Table 1). RAB11FIP2 is a suppressor of the endocytic internalization of EGFR and it is found in region 10 of the **D-L** list (Table 2) [17]. The presence of these genes in the hotspots found for GBM supports the value of the method described. There are many other interesting genes in the identified altered genomic regions, that can be useful for further investigations on the disease studied.

Complete information corresponding to the genes found in the significant **U-G** regions and **D-L** regions is included respectively as supplementary material in **Additional-file**-1 (for the data corresponding to Table 1) and **Additional-file**-2 (for the data corresponding to Table 2).

Conclusions

The combined analysis of CN and GE data obtained using DNA genome and RNA expression microarrays for paired samples is a very powerful approach to uncover key altered regions in a biological state studied. We present a robust method to find genomic regions that show simultaneous significant changes in both CN and GE. Our calculations applied to a cancer dataset find expected known genomic alterations and many others identified as key altered genomic regions. This approach is also proposed as an adequate strategy to identify driver or causal genes under the hypothesis that disease-associated gene expression changes are frequently induced by genomic alterations.

Materials and methods

Data

In this study we use a dataset of 64 human samples from Glioblastoma Multiforme (GBM) [7] that includes for each sample: *Affymetrix* DNA microarrays applied to detect of genome-wide CN changes and *Affymetrix* RNA expression microarrays applied to detect of GE changes. We used the same subgroup of samples that was previously analysed in Ortiz-Estevez *et al.* [11].

GE and CN normalization and signals calculation

GE data were processed using RMA algorithm [18] applied to the human gene expression microarrays: *Affymetrix* HGU133 plus 2.0 (using the same strategy followed in [19,20]). CRMAv2 algorithm [21] was applied to normalize the raw data and obtain the signals from the *Affymetrix* Human Mapping 500K SNP arrays. The processed signals were divided by the median of the normal samples for each element (SNP or gene) and then the log2 was computed. These log2 ratio signals were smoothed and segmented using Circular Binary Segmentation (CBS) algorithm [22] with the default parameters implemented in the DNAcopy R package.

Correlation between GE and CN

Pearson Correlation Coefficients (r) of the segmented GE and CN data were calculated taking the values of the segmented copy number and gene expression at the central point of the genomic position for each gene. P-values for the correlation coefficient of every gene loci were computed and adjusted by Bonferroni method. The established threshold for the selection of significantly correlated gene loci was correlation coefficient $r \geq 0.60$, which corresponds to adjusted p-value < 0.005. When using the gene loci GE unsegmented signal, the same correlation threshold and p-value cutoff were applied.

Frequency of U-G and D-L alterations

The thresholds that define DNA copy number gains and losses and up and down gene regulation were established applying k-Means algorithm, fixing three clusters (k = 3) on the segmented data, and done independently for the CN data and for the GE data. The CN data values were classified into gained (**G**), lost (**L**) or no-change (**N**) and the GE values were classified as up-regulated (**U**), down-regulated (**D**) or no-change (**N**). The thresholds found by k-Means for CN in the GBM dataset were > 0.19 (of the log2 ratio signals) for gain and < -0.15 for loss. The thresholds found for GE in the GBM were > 0.10 (of the log2 ratio signals) for up-regulation and < -0.12 for down-regulation. A contingency table with the 9 possible categorical states for the two types of data was built for every gene locus. A cutoff threshold was set up for the frequency of up-regulated and gained (**U-G**) and for the down-regulated and lost (**D-L**) categories, based on the empirical cumulative distributions of the categories. Taking into account the gene loci, the significant altered regions were defined as the ones that had a frequency ≥ than the upper 10% quantile of the distribution of **U-G** or the distribution of **D-L**.

General workflow for identification of key regions in the genome

Following the steps described above, we present a general workflow (Figure 5) that illustrates the strategy to achieve a combined paired analysis of datasets from genome-wide microarrays, both for GE and CN.

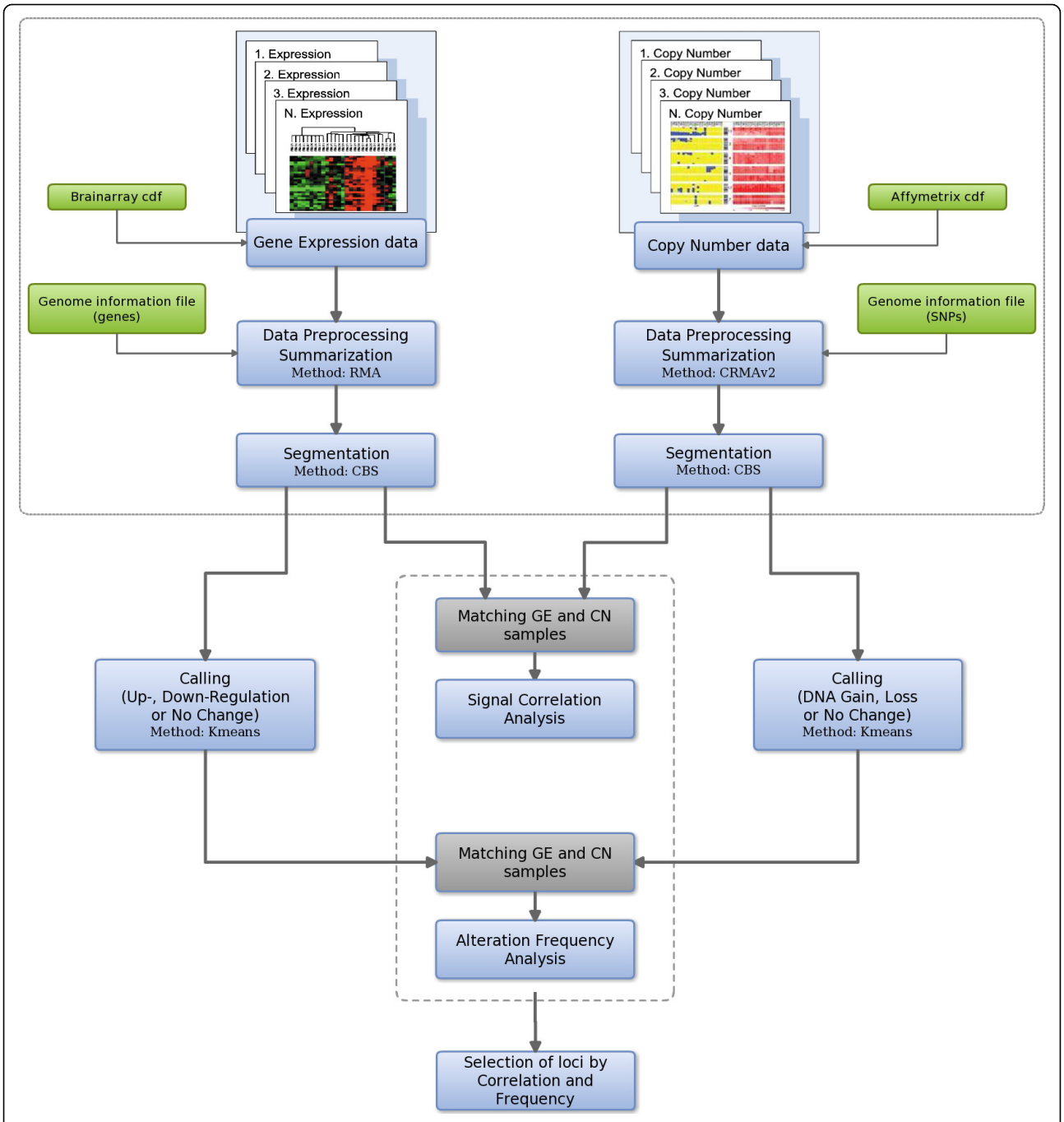


Figure 5 Workflow of the method for the analysis of gene expression and copy number changes. The figure illustrates the strategy to achieve a combined paired analysis of data sets from genome-wide microarrays both for GE and CN. The method searches for high correlation between segmented chromosomal regions that present a significant CN alteration (gain or loss) and segmented regions with significant GE change (up or down). The frequency of alteration is also analysed. The selected loci (hotspots) are regions chosen as significantly correlated and frequently altered in both GE and CN.

The workflow includes the different steps, the applied methods and the progression of the analysis. The strategy designed searches for high correlation between chromosomal regions that present a significant CN alteration (as gain or loss) and regions with significant GE change (as up or down). In this way, it determines which CN alterations have a strong influence on GE patterns. Key regions, i.e. hotspots in the genome, are defined as those regions simultaneously chosen as significantly correlated and frequently altered in both GE and CN.

Additional material

Additional file 1: Spreadsheet with the complete data corresponding to Table 1.

Additional file 2: Spreadsheet with the complete data corresponding to Table 2.

Additional file 3: Detailed view of all the 22 chromosomes showing the CN and GE correlation and the U-G or D-L categories frequency for the GBM dataset. The genomic regions are represented in X-axis. Blue and purple dots show the *correlation* coefficients between CN and GE for each gene loci (purple when r ≥ 0.60). Pink and green profiles represent the *frequency* values for the Up-Gain (U-G) category or the Down-Loss (D-L) category respectively.

Acknowledgements
This work has been supported by funds provided by the Local Government Junta de Castilla y León (JCyL, ref. project: CSI07A09), by the Spanish Government (ISCiii, ref. project PS09/00843) and by the European Commission (Research Grant ref. FP7-HEALTH-2007-223411). SA thanks the JCyL and the European Social Fund (ESF-EU) for a research grant. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.
This article has been published as part of *BMC Genomics* Volume 13 Supplement 5, 2012: Proceedings of the International Conference of the Brazilian Association for Bioinformatics and Computational Biology (X-meeting 2011). The full contents of the supplement are available online at http://www.biomedcentral.com/bmcgenomics/supplements/13/S5.

Author details
[1]Cancer Research Center (CIC-IBMCC), Consejo Superior de Investigaciones Científicas (CSIC), Campus Miguel de Unamuno, Salamanca, Spain.
[2]Department of Statistics, University of Salamanca (USAL), Salamanca, Spain.

Authors' contributions
CF carried out most of the analyses, developed the proposed method and drafted the manuscript. SA helped in the computational analyses and in the presentation of the results. JMSS participated in the design of the study and in the statistical methods applied. JDLR conceived of the study, participated in its design and coordination and wrote the main manuscript. All authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Published: 19 October 2012

References

1. Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, Rahman N, Stratton MR: **A census of human cancer genes.** *Nature Rev Cancer* 2004, **4**:177-83.
2. Stratton MR, Campbell PJ, Futreal PA: **The cancer genome.** *Nature* 2009, **458**:719-24.
3. Akavia UD, Litvin O, Kim J, Sanchez-Garcia F, Kotliar D, Causton HC, Pochanard P, Mozes E, Garraway L a, Pe'er D: **An integrated approach to uncover drivers of cancer.** *Cell* 2010, **143**:1005-17.
4. Beroukhim R, Getz G, Nghiemphu L, Barretina J, Hsueh T, Linhart D, Vivanco I, Lee JC, Huang JH, Alexander S, Du J, Kau T, Thomas RK, Shah K, Soto H, Perner S, Prensner J, Debiasi RM, Demichelis F, Hatton C, Rubin MA, Garraway LA, Nelson SF, Liau L, Mischel PS, Cloughesy TF, Meyerson M, Golub TA, Lander ES, Mellinghoff IK, Sellers WR: **Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma.** *Proc Nat Acad Sci USA* 2007, **104**:20007-12.
5. Beroukhim R, Mermel CH, Porter D, Wei G, Raychaudhuri S, Donovan J, Barretina J, Boehm JS, Dobson J, Urashima M, Mc Henry KT, Pinchback RM, Ligon AH, Cho Y-J, Haery L, Greulich H, Reich M, Winckler W, Lawrence MS, Weir BA, Tanaka KE, Chiang DY, Bass AJ, Loo A, Hoffman C, Prensner J, Liefeld T, Gao Q, Yecies D, Signoretti S, Maher E, Kaye FJ, Sasaki H, Tepper JE, Fletcher JA, Tabernero J, Baselga J, Tsao M-S, Demichelis F, Rubin MA, Janne PA, Daly MJ, Nucera C, Levine RL, Ebert BL, Gabriel S, Rustgi AK, Antonescu CR, Ladanyi M, Letai A, Garraway LA, Loda M, Beer DG, True LD, Okamoto A, Pomeroy SL, Singer S, Golub TR, Lander ES, Getz G, Sellers WR, Meyerson M: **The landscape of somatic copy-number alteration across human cancers.** *Nature* 2010, **463**:899-905.
6. Pollack JR, Sørlie T, Perou CM, Rees CA, Jeffrey SS, Lonning PE, Tibshirani R, Botstein D, Børresen-Dale A-L, Brown PO: **Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors.** *Proc Nat Acad Sci USA* 2002, **99**:12963-8.
7. Kotliarov Y, Steed ME, Christopher N, Walling J, Su Q, Center A, Heiss J, Rosenblum M, Mikkelsen T, Zenklusen JC, Fine HA: **High-resolution global genomic survey of 178 gliomas reveals novel regions of copy number alteration and allelic imbalances.** *Cancer Res* 2006, **66**:9428-36.
8. Kotliarov Y, Kotliarova S, Charong N, Li A, Walling J, Aquilanti E, Ahn S, Steed ME, Su Q, Center A, Zenklusen JC, Fine H a: **Correlation analysis between single-nucleotide polymorphism and expression arrays in gliomas identifies potentially relevant target genes.** *Cancer Res* 2009, **69**:1596-603.
9. Turner N, Lambros MB, Horlings HM, Pearson A, Sharpe R, Natrajan R, Geyer FC, van Kouwenhove M, Kreike B, Mackay A, Ashworth A, van de Vijver MJ, Reis-Filho JS: **Integrative molecular profiling of triple negative breast cancers identifies amplicon drivers and potential therapeutic targets.** *Oncogene* 2010, **29**:2013-23.
10. Kim Y-A, Wuchty S, Przytycka TM: **Identifying causal genes and dysregulated pathways in complex diseases.** *PLoS Computational Biology* 2011, **7**:e1001095.
11. Ortiz-Estevez M, De Las Rivas J, Fontanillo C, Rubio A: **Segmentation of genomic and transcriptomic microarrays data reveals major correlation between DNA copy number aberrations and gene-loci expression.** *Genomics* 2011, **97**:86-93.
12. De Tayrac M, Etcheverry A, Aubry M, Saïkali S, Hamlat A, Quillien V, Le Treut A, Galibert MD, Mosser J: **Integrative genome-wide analysis reveals a robust genomic glioblastoma signature associated with copy number driving changes in gene expression.** *Genes Chromosomes Cancer* 2009, **48**:55-68.
13. Ruano Y, Mollejo M, Ribalta T, Fiaño C, Camacho FI, Gómez E, de Lope AR, Hernández-Moneo JL, Martínez P, Meléndez B: **Identification of novel candidate target genes in amplicons of glioblastoma multiforme tumors detected by expression and CGH microarray profiling.** *Molecular Cancer* 2006, **5**:39.
14. Reifenberger G, Collins VP: **Pathology and genetics of astrocytic gliomas.** *J Mol Med* 2004, **82**:656-670.
15. Chernova OB, Somerville RP, Cowell JK: **A novel gene, LGI1, from 10q24 is rearranged and downregulated in malignant brain tumors.** *Oncogene* 1998, **17**:2873-2881.
16. Wechsler DS, Shelly CA, Petroff CA, Dang CV: **MXI1, a putative tumor suppressor gene, suppresses growth of human glioblastoma cells.** *Cancer Res* 1997, **57**:4905-4912.
17. Cullis DN, Philip B, Baleja JD, Feig LA: **Rab11-FIP2, an adaptor protein connecting cellular components involved in internalization and recycling of epidermal growth factor receptors.** *J Biol Chem* 2002, **277**:49158-49166.
18. Irizarry R a, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP: **Exploration, normalization, and summaries of high density oligonucleotide array probe level data.** *Biostatistics* 2003, **4**:249-64.
19. Vicent S, Luis-Ravelo D, Antón I, García-Tuñón I, Borrás-Cuesta F, Dotor J, De Las Rivas J, Lecanda F: **A novel lung cancer signature mediates metastatic bone colonization by a dual mechanism.** *Cancer Res* 2008, **68**:2275-85.
20. Hernández JA, Rodríguez AE, González M, Benito R, Fontanillo C, Sandoval V, Romero M, Martín-Núñez G, de Coca AG, Fisac R, Galende J, Recio I, Ortuño F, García JL, De Las Rivas J, Gutiérrez NC, San Miguel JF, Hernández JM: **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.
21. Bengtsson H, Wirapati P, Speed TP: **A single-array preprocessing method for estimating full-resolution raw copy numbers from all Affymetrix genotyping arrays including GenomeWideSNP 5 & 6.** *Bioinformatics* 2009, **25**:2149-56.

22. Venkatraman ES, Olshen AB: **A faster circular binary segmentation algorithm for the analysis of array CGH data.** *Bioinformatics* 2007, **23**:657-63.

doi:10.1186/1471-2164-13-S5-S5
Cite this article as: Fontanillo *et al.*: Combined analysis of genome-wide expression and copy number profiles to identify key altered genomic regions in cancer. *BMC Genomics* 2012 **13**(Suppl 5):S5.

OPEN ACCESS Freely available online

PLoS one

Functional Analysis beyond Enrichment: Non-Redundant Reciprocal Linkage of Genes and Biological Terms

Celia Fontanillo[1☯], Ruben Nogales-Cadenas[2☯], Alberto Pascual-Montano[2], Javier De Las Rivas[1]*

1 Cancer Research Center (CiC-IBMCC, CSIC/USAL), Campus Miguel de Unamuno, Salamanca, Spain, **2** National Center of Biotechnology (CNB, CSIC), Campus de Cantoblanco UAM, Madrid, Spain

Abstract

Functional analysis of large sets of genes and proteins is becoming more and more necessary with the increase of experimental biomolecular data at *omic*-scale. Enrichment analysis is by far the most popular available methodology to derive functional implications of sets of cooperating genes. The problem with these techniques relies in the redundancy of resulting information, that in most cases generate lots of trivial results with high risk to mask the reality of key biological events. We present and describe a computational method, called ***GeneTerm Linker***, that filters and links enriched output data identifying sets of associated genes and terms, producing metagroups of coherent biological significance. The method uses fuzzy reciprocal linkage between genes and terms to unravel their functional convergence and associations. The algorithm is tested with a small set of well known interacting proteins from yeast and with a large collection of reference sets from three heterogeneous resources: multiprotein complexes (CORUM), cellular pathways (SGD) and human diseases (OMIM). Statistical *Precision*, *Recall* and balanced *F-score* are calculated showing robust results, even when different levels of random noise are included in the test sets. Although we could not find an equivalent method, we present a comparative analysis with a widely used method that combines enrichment and functional annotation clustering. A web application to use the method here proposed is provided at http://gtlinker.cnb.csic.es.

Citation: Fontanillo C, Nogales-Cadenas R, Pascual-Montano A, De Las Rivas J (2011) Functional Analysis beyond Enrichment: Non-Redundant Reciprocal Linkage of Genes and Biological Terms. PLoS ONE 6(9): e24289. doi:10.1371/journal.pone.0024289

Editor: Debashish Bhattacharya, Rutgers University, United States of America

Received May 18, 2011; **Accepted** August 3, 2011; **Published** September 16, 2011

Funding: Dr. De Las Rivas receives financial support provided by EU FP7-HEALTH-2007-B (project 223411), by Spanish Ministry of Science and Innovation MICINN-ISCiii (projects PI061153 and PS09/00843), and by the Regional Government, Junta de Castilla y Leon JCyL (project CSI07A09). Dr. Pascual-Montano receives financial support provided by MICINN grant BIO2010-17527. Dr. Nogales-Cadenas thanks the Juan de la Cierva Program (MICINN-JDC 2010) and Dr. Fontanillo thanks the CSIC JAE-PREDOC Program. 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

☯ These authors contributed equally to this work.

Introduction

Genome- and proteome-wide analyses performed using high-throughput techniques are providing many collections of genes and proteins that are associated to studies performed over specific sets of samples in definite biological contexts. One of the major challenges of current computational biology is to provide robust automatic methods for a meaningful functional annotation of the long lists of genes or proteins derived from such high-throughput studies. Functional *enrichment analysis* (EA) is at present the most popular available methodology to derive functional implications of sets of cooperating genes. It uses statistical testing to find significant annotations in groups of genes. A recent review of enrichment tools categorizes them in three major classes: *singular* (SEA), *modular* (MEA) and *gene-set* (GSEA) [1]. Modular analysis (MEA) can be considered a second generation of functional enrichment since it uses concurrent gene annotation improving coverage [2,3,4]. Gene set enrichment analysis (GSEA) has become a popular tool to extract biological insight from complete ranked gene lists without the need of pre-selecting top genes [5].

Functional enrichment analysis, however, does not address several key problems associated to the biological annotations: ***(i)*** *Redundancy* of the biological terms, that are repeated in many different annotation resources (e.g. *cell cycle* GO:0007049, *cell cycle* KEGG hsa04110, etc) or that are segregated in very similar terms with the same biological meaning (e.g. GO:0007049 *cell cycle* and GO:0022402 *cell cycle process*). ***(ii)*** *Bias* in the annotation space due to highly frequent use of certain "promiscuous" terms that are unspecific (e.g. GO:0050789 *regulation of biological process* includes more than 44% of all human genes annotated to GO-BP). ***(iii)*** *Inadequate functional annotation* of many genes that are well-known (e.g. NRAS human gene product P01111 is not annotated to GO:0043410 *positive regulation of MAPKKK cascade*, but the role of this gene in the MAPK signaling is well-known, since it is paralogous to gene HRAS, which has a central role in such pathway).

To overcome these limitations and challenges we have developed a new computational method that finds significant and coherent metagroups of genes and terms, performing several steps to eliminate redundant and non-informative data. The method takes the output of an enrichment analysis and produces a simple result that includes genes and co-annotations associated in metagroups. These metagroups are ranked by analysis of their significance and coherence, as a way to find the most relevant functions present in the query gene list. The algorithm is tested with a small set of well known interacting proteins and with a large reference set of data from three heterogeneous resources: mammalian multiprotein complexes (CORUM), yeast cellular

pathways (SGD) and human diseases (OMIM). Statistical *Precision*, *Recall* and balanced *F-score* are calculated for each test, and we observe robust results even introducing different percentages of randomly selected genes in the queries. The computational method can be applied to the output result of any enrichment analysis. We provide a web application to use the method (http://gtlinker.cnb.csic.es) that only needs as input a gene list, because in a first step it runs an enrichment analysis tool [3] implemented within the same workflow.

Results

Analysis of the distributions of terms/genes in different Annotation Spaces

Functional annotation and enrichment analysis relies on the use of biological databases that include groups of genes associated to specific biological functions, such as: metabolic and signaling pathways, cellular processes and apparatus, organisms, etc. Some of the biological databases most used in functional profiling are: GO (repository of gene and gene product ontological attributes across species) [6], KEGG (atlas of biological pathways) [7], UniProt (catalog of structural and functional information on proteins) [8]. In these databases the functions are annotated with specific terms that define and describe the biological roles and actions. They usually apply controlled vocabularies, i. e. structured collections of terms with numerical IDs. As it happens in language evolution, the use of the terms can modulate their meaning, because when some expressions become too trendy, fashionable or promiscuous they can lose significance. In addition, most of these vocabularies are defined to be organism-independent and therefore in some cases they encode global definitions that are not useful to explain very specific biological processes.

We have analyzed and compared the frequency distributions of the biological terms in two worldwide used databases (GO and KEGG). This analysis counts the number of genes assigned to each term and reveals that the distributions are quite uneven, existing a large proportion of terms that include very small number of genes and a considerable amount of outliers assigned to many genes. In fact, for the case of GO-BP (Biological Process), GO-MF (Molecular Function) and GO-CC (Cellular Component) more than 50% of the terms have less than four genes assigned in human (see **Figure 1A**, boxplots of the distributions of GO and KEGG terms assigned to human genes). The distribution is more homogeneous for the case of KEGG terms, which shows a Gaussian-like curve (**Figure 1C and 1D**). The black vertical lines in these plots indicate the percentage of genes per term with respect to the total number of human genes (i.e. 29095 genes using ENSEMBL v57, March 2010). The results show that the most used GO-BP term is assigned to 6.43% human genes (1872 genes assigned to *signal transduction*, GO:0007165). **Figure 1B** presents for each GO category (BP, MF, CC) the three terms most frequently annotated to human genes. Such terms (e.g. term *protein binding*) are outliers in the distributions (**Figure 1A**) and therefore they can be considered terms with low-information-content, too generic to provide clear and meaningful functional annotation on their own.

Identification of over-represented terms to improve functional annotation

The analysis of the distribution of terms indicates that there are some biological annotations that are over-represented, mainly in GO. Such over-representation can be quantified by the deviation from the average number of assignments (red and green vertical lines in **Figure 1C and 1D**). Based on such average ($\bar{X}$) and on the standard deviation (σ_x) of the distributions of terms in each annotation space for each organism, we set up a *Z-score* threshold to identify the outlier terms that had a number of genes assigned (Ng) deviated from average: $Ng > (\bar{X} + n\sigma_x)$. The deviation factor n was set up at 4 for human. This threshold allows identification of the biological terms that are "generic" and "promiscuous", and – on their own– they can be considered not very informative. These generic terms affect a significant proportion of genes. In the case of human, generic GO-BPs include 10,038 genes (34.5% of the total), generic GO-MFs include 12,991 genes (44.6% of the total) and generic GO-CCs include 15,179 genes (52.2% of the total). In the case of KEGG only 2 terms were considered nonspecific and they only affect to 700 genes. All the generic terms were tagged in order to further use them only in the case that they appear in co-occurrence with other terms.

Definition of *GeneTerm-sets* as a type of *Frequent Itemsets*

Most of the enrichment analyses are based in searching for *frequent patterns* of association between biomolecular elements (e.g. genes, proteins) and the corresponding annotations or descriptions found in biological databases. In the data-mining field those patterns are called *frequent itemsets* [9]. A formal mathematical definition of *frequent itemsets* can be as follows: given a set of *items* $I = \{i_1, i_2 \ldots i_n\}$ and a database of *transactions* $T = \{t_1, t_2 \ldots t_m\}$ where each transaction is a subset of I, $F \subseteq I$ is a *frequent itemset* if it is included in a number of transactions greater than a specified threshold, ε. That number of transactions is called the *support* of the itemset.

Translating these concepts to the biological context of enrichment analyses, the *items* will be the "terms" (i.e. the biological annotations) from the different databases, and the *transactions* will be the "genes" (i.e. the biological entities). In this way, it is possible to generalize the *frequent patterns* obtained by any enrichment analysis as a list of annotations related with a subset of genes, both associated by the score or ***p-value*** of the enrichment that measures the strength of the relationship. Formally, such combination of terms/genes/***p-value*** is a *frequent itemset* derived from a functional annotation procedure, and we define such as *GeneTerm-set* element: $E_i = \langle G_i, A_i, p_i \rangle$. Where E_i is the *ith* element in the results, A_i is a set $\{a_1, a_2, \ldots a_n\}$ of biological annotations or terms, G_i is a set of genes $\{g_1, g_2, \ldots g_m\}$ and p_i the ***p-value***. In terms of enrichment A_i is a set of annotations over-represented in a list of genes and G_i is the subset of genes that support that over-representation with a ***p-value*** of p_i. When using singular (SEA) or concurrent modular (MEA) enrichment analyses, the difference in the data structure of the result consists only in the number of elements in A_i, that is 1 in the first case and ≥ 1 in the latest. Most of the enrichment tools provide large lists of these *GeneTerm-set* elements derived from the analysis on different annotation spaces. Such multiple lists are many times very redundant, provided as independent or non-related and including many generic terms. This hampers the extraction of meaningful biological insights because the interpretation of such redundant and complex data sets is quite difficult, time-consuming and daunting, many times dependent on the expertise and the area of interest of the biologists that analyze the lists.

Method: non-redundant reciprocal linkage of *GeneTerm-sets* to go beyond Enrichment

We have developed a computational method to find metagroups of genes and annotations composed by linked *GeneTerm-sets*, eliminating redundant and non-informative elements. The method, called ***GeneTerm Linker*** has 2 major goals: ***(i)*** to provide a robust automatic way to analyse the large

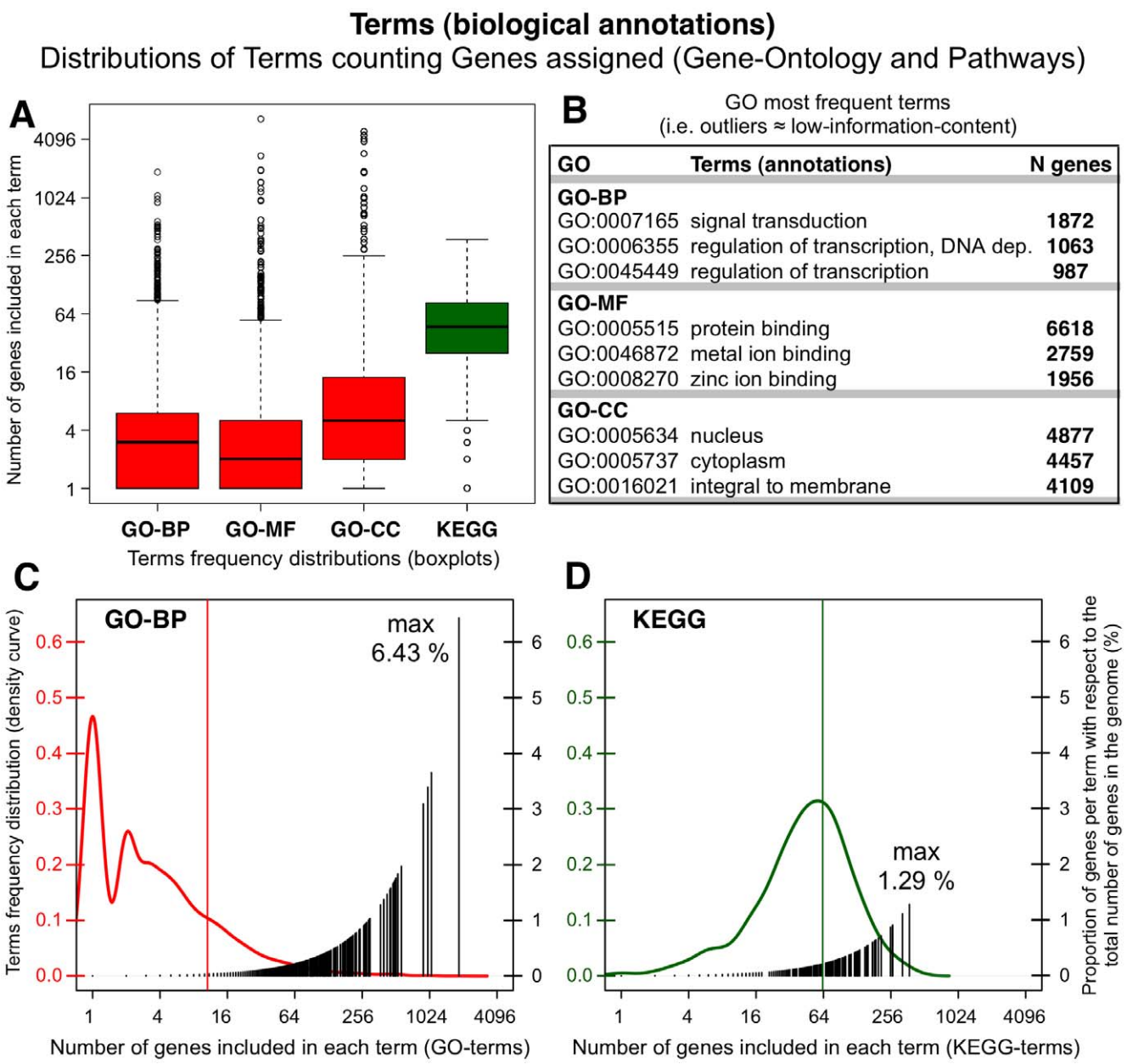


Figure 1. Distributions of biological terms in GO and KEGG databases. Distributions of biological terms from GO and KEGG databases counting the number of genes assigned to each term. The data correspond to human genes. *(A)* Boxplots of the distributions for GO categories (BP, MF, CC) and for KEGG. *(B)* Most frequent GO terms. *(C)* Left: density distribution of GO-BP -marking the average with a red line-; right: proportion of genes per term with respect to the total number of genes (%). *(D)* Same as C for KEGG.
doi:10.1371/journal.pone.0024289.g001

collections of *GeneTerm-sets* produced by enrichment methods; ***(ii)*** to produce significant and coherent metagroups of genes associated to concurrent terms and annotations that describe the specific biological functions of the metagroup. In the following paragraphs we describe the four major procedure steps that the method includes:

Step 1

Filtering *GeneTerm-sets* that only include over-represented terms. As we showed above, those terms whose frequency of appearance in databases is strongly greater than average can provide obvious and non-interesting results, while masking significant functional patterns present in the query genes. Such over-represented terms are considered outliers. Once the outliers are found in each biological annotation category for each organism, the first step of the method consists in removing the *GeneTerm-set* elements that only correlate groups of genes with over-represented terms. If one element in the enrichment result includes outliers in its set of annotations but also contains other terms, the element is not discarded because the generic terms are related with other specific annotations. In this way, given an element $\boldsymbol{E_i}$ from the enrichment result, the whole element will be set aside only if its set of annotations $\boldsymbol{A_i}$ is composed by outliers. This first step of the method significantly reduces the number of elements in the list of results, removing useless information.

Step 2

Retrieve metagroups using reciprocal linkage between *GeneTerm-sets*. The second step of the algorithm creates metagroups of elements that are related by sharing common genes or by sharing common terms. The method is reciprocal because it considers both the genes and the terms included in each *GeneTerm-set*. First, to find the linkage between genes it uses a similarity coefficient that provides a preliminary grouping of *GeneTerm-sets*. Second, to find the linkage between terms it uses a greedy algorithm that explores the annotations to merge the common ones.

Gupta *et al.* showed that the use of the *Jaccard Similarity coefficient* to measure the distance between the transactions that support frequent patterns get better results than the distance between the items, demonstrating its fitness to catch the interactions between those sets in the data and its robustness regardless the size of the data [10]. This is an approach that does not take into account the strength of the relationships between transactions and items, i.e. between genes and terms in our case. Considering these ideas, our method finds the linkage between *GeneTerm-set* elements by creating for each $\boldsymbol{E_i}$ a vector $\boldsymbol{v_i}$ which contains the occurrence of each gene with respect to the whole gene list of the input (in binary numbers 1/0) and incorporates as an additional component the ***p-value*** of each element $\boldsymbol{E_i}$ weighted by factor $\boldsymbol{M}$ (the total number of genes in the list). This additional parameter represents the strength of the relationship within each *GeneTerm-set*. The pair-wise distances between all vectors $\boldsymbol{v_i}$ are calculated using *Cosine Similarity*, a generalization of the *Jaccard Similarity coefficient* for non-binary attributes. Once the similarity is calculated, the distances are analyzed using *Ward's hierarchical clustering* in order to find the linkage between *GeneTerm-sets* (i.e. the clusters formed by the elements). This linkage is considered fuzzy because each gene or combination of genes can be included in several *GeneTerm-sets*. A heuristic threshold consisting of a cutoff set up at a given depth of the cluster tree is used to define the preliminary metagroups. By default the threshold is set up at 20% of the tree depth, but if it is not enough to define metagroups, the algorithm increases the cutoff in 10% steps till at least one metagroup is found. In this way, we identify coherent modules of information based on common genes.

After this process, the algorithm proceeds performing a greedy recursive exploration of terms within the preliminary clusters (pre-metagroups) to merge the ones that share the same terms. At the end of this second step the method provides metagroups where the convergence of genes and terms is maximized. A formal mathematical description of the process is included in the Materials and Methods.

Step 3

Remove redundancy within the selected metagroups. Once the metagroups are created, it is possible to compact and reduce their size by removing the redundant elements included inside each metagroup.

Toivonen *et al.* proposed the concept of *cover* of a set of association rules (a special case of *frequent itemsets*) as the minimal subset that contains all the relationships present in an original set [11]. To avoid losing any item, we extend the concept of *cover* of a collection of *itemsets* (i.e., in our case, a metagroup of *GeneTerm-sets*) with the requirement of *completeness* of the data. In this way, in our algorithm we redefine and apply the concept of *complete cover*. The mathematical description to calculate this parameter is presented in Materials and Methods.

To assess the *complete cover* we do not contemplate only the terms included in the metagroups, but also the genes that support them. Each metagroup is described by the total set of terms and the total set of genes included in their elements. So, to find redundant elements inside a metagroup the method searches for the ones with all its genes and terms included in another elements of the same metagroup. In this search the *GeneTerm-sets* are always ordered by increasing ***p-values*** to eliminate consistently the less significant sets. Following this approach, redundant *GeneTerm-sets* present in the enrichment outputs are found and removed.

Step 4

Calculate significance and coherence of the meta-groups. After the final metagroups have been generated and the redundant *GeneTerm-sets* removed, a series of parameters are calculated to evaluate their significance and coherence. Our assumption is that a functional coherent metagroup should be compact and well separated from other, therefore such coherence tries to measure both the intra-groups compactness and the inter-groups distance.

In order to evaluate the statistical significance a *Hypergeometric test* is performed with all the genes and terms assigned to each metagroup [2,12]. The resultant ***p-values*** are adjusted for multiple tests using the FDR method [13].

In order to assess the compactness (maximum distance in between data points of clusters) and proximity (minimum distance between clusters) the main parameter calculated is the *Silhouette Width*, which ranges from 1 to −1 and measures both the compactness and proximity of multiple groups [14]. The method also calculates the *Diameter*, that is the maximum *Cosine distance* within the *GeneTerm-sets* of each metagroup and ranges from 0 to 1; and the *Similarity Coefficient*, which is [1 – average *Cosine distance*] within the *GeneTerm-sets* of each metagroup and also ranges from 0 to 1. All these distance and similarity calculations are done based on the genes present in the metagroups.

Testing the method with a set of yeast nuclear proteins

We investigate the ability of ***GeneTerm Linker*** method to find metagroups of functionally related genes using as test set of 59 nuclear proteins from yeast (**Figure 2A**) that have been characterized by protein interaction methods and form five well-defined protein complexes [15]. This set had been previously used in the evaluation of a method to find densely connected regions in protein interaction networks [15] and it includes a collection of well-annotated proteins with strong functional links.

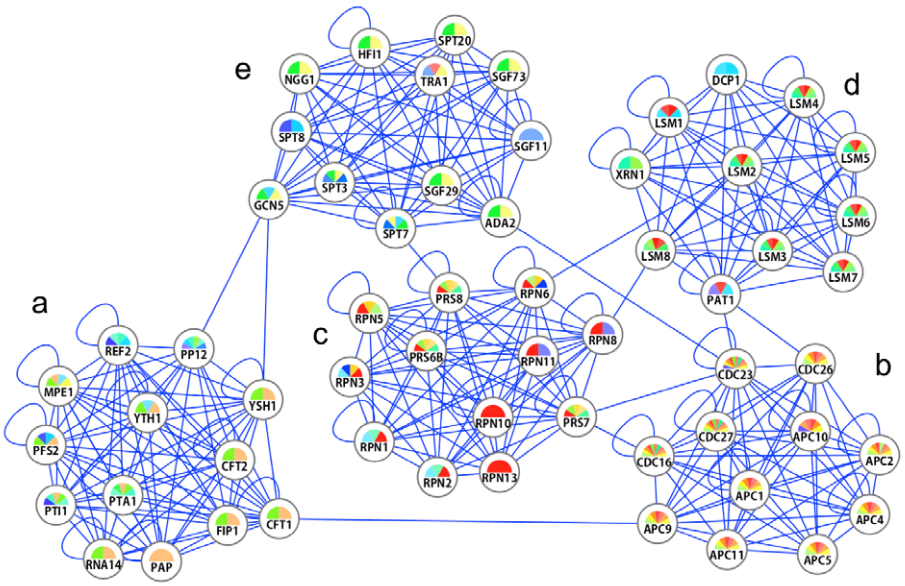
A network of experimentally proven interactions between these proteins was build, using APID and APID2NET [16,17], showing that they form 5 distinct clusters (**Figure 2B**). These clusters constitute a good set for use as a benchmark.

The analysis of the set of yeast proteins is shown in **Figure 2C**. The output of the algorithm shows that five compact metagroups are found, all having a *Silhouette Width*>0.5, that is a good indication of the internal tightness of each metagroup and its external separation from the other metagroups [14]. Moreover, the *Hypergeometric test* also indicates that the metagroups are significant. The size of the 5 metagroups found was: [1] 13 genes and 9 *GeneTerm-sets*; [2] 11 genes and 4 *GeneTerm-sets*; [3] 14 genes and 9 *GeneTerm-sets*; [4] 14 genes and 13 *GeneTerm-sets*; [5] 14 genes and 14 *GeneTerm-sets*. The terms corresponding to each metagroup are presented in **Figure 2C** (co-annotations column), showing the main functions and biological roles found associated to each metagroup (a complete version of this table is included in **Table S1**). Some concurrent terms are synonymous, like in the 3rd metagroup "proteasome complex" (GO:0000502) and "proteasome" (KEGG:03050); but other terms are complementary, like in the 4th metagroup "U4/U6 tri-snRNP complex" (GO:0046540)

A

Protein Complexes (yeast)	Number of proteins
a. mRNA cleavage and polyadenylation specificity factor complex CFT1,CFT2,FIP1,GLC7,MPE1,PAP1,PFS2,PTA1,PTI1,REF2,RNA14,YSH1,YTH1	13
b. anaphase-promoting complex APC1,APC2,APC4,APC5,APC9,APC11,CDC16,CDC23,CDC26,CDC27,DOC1	11
c. proteasome, 19/22S regulator complex RPN1,RPN2,RPN3,RPN5,RPN6,RPN8,RPN10,RPN11,RPN13,RPT1,RPT3,RPT6	12
d. U6 snRNP complex DCP1,KEM1,LSM1,LSM2,LSM3,LSM4,LSM5,LSM6,LSM7,LSM8,PAT1	11
e. SAGA complex ADA2,GCN5,HFI1,NGG1,SGF11,SGF29,SGF73,SPT3,SPT7,SPT8,SPT20,TRA1	12

B



C

Gene Metagroups found (include related groups)	GENES found	GENES in Ref.	p-value	Co-Annotations
Metagroup 1	***Silhouette* Width: 0.529**			
Genes:GLC7,REF2,YTH1,FIP1,PAP1,PFS2,CFT1,RNA14,PTI1, PTA1,MPE1,CFT2,YSH1	**13(59)**	**15(7103)**	**2.26E-26**	GO:0005847:mRNA cleavage & polyadenylation specificity factor complex, GO:0006378:mRNA polyadenylation ...
GLC7,REF2,YTH1,FIP1,PAP1,PFS2,CFT1,RNA14,PTI1,PTA1,MPE1,CFT2,	13(59)	15(7103)	1.03E-24	
GLC7,YTH1,FIP1,PAP1,PFS2,CFT1,RNA14,PTI1,PTA1,MPE1,CFT2,YSH1	12(59)	12(7103)	1.12E-24	
GLC7,YTH1,FIP1,PFS2,CFT1,RNA14,PTI1,PTA1,MPE1,CFT2,YSH1	11(59)	11(7103)	1.11E-22	
...				
Metagroup 2	***Silhouette* Width: 0.823**			
Genes:CDC23,APC5,CDC16,APC2,APC1,DOC1,APC9,APC11, APC4,CDC27,CDC26,GLC7,TRA1	**11(59)**	**11(7103)**	**4.85E-24**	04111: Cell cycle - yeast, GO:0019941: modification-dependent protein catabolic process, GO:0004842: ubiquitin-protein ligase ...
CDC23,APC5,CDC16,APC2,APC1,DOC1,APC9,APC11,APC4,CDC27,CDC	11(59)	11(7103)	1.11E-22	
CDC23,APC5,CDC16,APC2,APC1,APC9,APC11,APC4,CDC27,CDC26	10(59)	10(7103)	1.20E-20	
CDC23,APC5,CDC16,APC2,APC1,DOC1,APC9,APC4,CDC27,CDC26	10(59)	10(7103)	1.20E-20	
...				
Metagroup 3	***Silhouette* Width: 0.609**			
Genes:RPN13,RPN8,RPN1,RPT1,RPN3,RPN10,RPN11,RPN2, RPN5,RPT3,RPT6,RPN6,APC2,DOC1	**14(59)**	**90(7103)**	**8.11E-15**	GO:0006511: ubiquitin-dependent protein catabolic process, GO:0000502: proteasome complex, 03050:Proteasome ...
RPN13,RPN8,RPN1,RPT1,RPN3,RPN10,RPN11,RPN2,RPN5,RPT3,RPT6,I	12(59)	34(7103)	9.79E-17	
RPN13,RPN8,RPN1,RPT1,RPN3,RPN10,RPN11,RPN2,RPN5,RPT6,RPN6	11(59)	30(7103)	1.15E-15	
RPN13,RPN8,RPT1,RPN3,RPN11,RPN5,RPT3,RPN6	8(59)	13(7103)	6.07E-14	
...				
Metagroup 4	***Silhouette* Width: 0.750**			
Genes:LSM5,DCP1,PAP1,LSM3,LSM8,LSM6,LSM1,LSM4,LSM7, LSM2,PTA1,KEM1,PAT1,YSH1	**14(59)**	**93(7103)**	**1.31E-14**	03018: RNA degradation, GO:0030529: ribonucleoprotein complex, GO: 0046540:U4/U6 x U5 tri-snRNP complex ...
LSM5,DCP1,PAP1,LSM3,LSM8,LSM6,LSM1,LSM4,LSM7,LSM2	10(59)	12(7103)	6.27E-19	
LSM5,PAP1,LSM3,LSM8,LSM6,LSM1,LSM4,LSM7,LSM2	9(59)	10(7103)	9.04E-18	
LSM5,LSM3,LSM8,LSM6,LSM4,PTA1,LSM7,LSM2	8(59)	8(7103)	9.04E-17	
...				
Metagroup 5	***Silhouette* Width: 0.570**			
Genes:NGG1,HFI1,TRA1,SPT20,SGF29,SPT7,SGF73,SPT8, SGF11,GCN5,SPT3,ADA2,RPN6,CFT2	**14(59)**	**142(7103)**	**5.26E-12**	GO:0000124: SAGA complex, GO:0016568: chromatin modification, GO:0046695:SLIK (SAGA-like) complex ...
NGG1,HFI1,TRA1,SPT20,SGF29,SPT7,SGF73,SPT8,SGF11,GCN5,SPT3	12(59)	22(7103)	3.04E-19	
NGG1,HFI1,TRA1,SPT20,SPT7,SGF73,SPT8,SGF11,GCN5,SPT3,ADA2	11(59)	19(7103)	3.41E-18	
NGG1,HFI1,TRA1,SPT20,SGF29,SPT7,SGF73,SPT8,GCN5,SPT3	10(59)	16(7103)	4.36E-17	
...				

Figure 2. Analyses of a highly connected set of yeast proteins with *GeneTerm Linker*. Analyses of a set of 59 yeast proteins using the algorithm proposed. *(A)* Lists of the proteins that form 5 known protein complexes. *(B)* Protein interaction network form by such 59 yeast proteins. Each node is a protein and the color scheme corresponds to GO-BP and InterPro terms marked using APID2NET [17]. *(C)* Output of the analysis of the 59 genes with the algorithm proposed (full table in **Table S1**).
doi:10.1371/journal.pone.0024289.g002

and "Like-Sm ribonucleoprotein (LSM) domain" (IPR001163). The overall result shows that the method finds the 5 complexes expected, including in each one all its proteins. In the case of metagroups 3, 4 and 5 some extra proteins are included: APC2 and DOC1 in the 3rd metagroup; PAP1, PTA1 and YSH1 in the 4th metagroup; and RPN6 and CFT2 in the 5th metagroup.

Comparison of the method with another functional annotation approach

To perform a comparative analysis with other methods, we carried out a systematic identification of the gene pairs that compose the test set of five yeast complexes, described above, and all the gene pairs found by the functional association method. In this way, we count all possible gene pairs and all true positive (TP) gene pairs found in the reference complexes, and we can calculate the *Accuracy* (i.e. *Rand statistic*) and the *Jaccard coefficient* defined as:

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)};$$

$$Jaccard\ Coefficient = \frac{(TP)}{(TP+FP+FN)}$$

These parameters measure the relationship between pairs of points using the co-occurrence matrices for the expected partition and the partition generated by a given method [18]. The statistical evaluation was done (see **Table 1**) for the results obtained with our method and for the results obtained with a widely used *Functional Annotation Clustering* (FAC) method developed by DAVID Bioinformatics Resources [4]. This is the only method that we found in the literature that has a similar goal of finding functional modules (that include genes and terms) and use data derived from enrichment analysis.

The results indicate that ***GeneTerm Linker*** method is quite accurate to find the biological complexes present in the test set of 59 yeast nuclear proteins (*Accuracy* = 0.95). Such *Accuracy* drops when using the agglomeration algorithm FAC [4], which by default finds many more groups or modules of genes and terms (15 functional modules). Tuning the parameters of FAC algorithm to find just the 5 expected metagroups the *Accuracy* still does not reach 90% (0.88).

The *Jaccard coefficient* measures the proportion of gene pairs that belong to the same metagroup in both the expected and the computed partition, relative to all pairs that belong to the same metagroup in at least one of the two partitions. This *coefficient* for the case studied was 0.769 using our method and 0.562 using FAC method.

Testing the method with reference sets from three heterogeneous resources: Complexes, Pathways and Diseases

To achieve a more comprehensive evaluation of the method, we did a series of trials with reference sets of gene metagroups defined in three broad biomolecular resources: ***(1)*** sets composed of multiprotein complexes identified in mamals (from CORUM) [19], ***(2)*** sets composed by groups of genes involved in yeast pathways (from SGD) [20], ***(3)*** sets of groups of genes involved in human diseases (from OMIM) [21]. We select from each database ten of sets with at least 8 genes/proteins each (**Figure 3**).

Using this collection of reference gene sets we run the method once for each set, to investigate how many of the reference genes are included in the first, most significant, metagroup found. We performed the analyses using not just each reference metagroup alone, but also mixing it with randomly selected genes to introduce two levels of noise in the set: 20% and 60% (i.e. in order to acquire 20% noise, if the reference group had 10 genes then 2 genes were randomly selected from the whole gene list of such resource and included with the 10 true genes).

The results using ***GeneTerm Linker*** over the whole collection of reference gene sets is shown in **Figure 3**, which presents in each row the most significant metagroup found and its overlap with the corresponding reference gene set used as query. For example, in the case of the first group (1c): the *C complex spliceosome* is composed

Table 1. Comparison of methods: *GeneTerm Linker* and *Functional Annotation Clustering*.

	GeneTerm Linker	DAVID *FAC* (used by default)	DAVID *FAC* (tuned to find 5 groups)
Total groups reference	5	5	5
Total groups found	5	15	5
All possible gene pairs	1711	1711	1711
TP	320	320	254
FN	82	1179	132
FP	0	0	66
TN	1309	212	1259
Jaccard Coefficient	**0.769**	**0.213**	**0.562**
Accuracy	**0.952**	**0.311**	**0.884**

Comparative results for the set of 59 yeast proteins: *Accuracy* and *Jaccard Coefficient* obtained using the present method and using Functional Annotation Clustering (FAC) method with its parameters by default or tuned to find 5 groups.
doi:10.1371/journal.pone.0024289.t001

	COMPLEXES (from CORUM db, human)	GENES in Ref.	GENES Tested	GENES Found	Common GENES	Precision (%)	Recall (%)	F-score (%)	adjusted p.value	TERMS Found	TERMS Found (only first shown)
1c	C complex spliceosome	80	96	68	68	100.00	85.00	91.89	5.25E-138	6	GO:0005681:spliceosomal …
2c	Mediator (transcriptional coactivator) complex	32	39	28	28	100.00	87.50	93.33	3.96E-064	10	GO:0016592:mediator complex …
3c	Proteasome (20S/26S)	22	27	22	22	100.00	100.00	100.00	5.34E-063	12	03050:Proteasome …
4c	RNA polymerase II (RNAPII)	26	32	24	24	100.00	92.31	96.00	1.21E-059	12	GO:0006350:transcription …
5c	F1F0-ATP synthase (EC 3.6.3.14), mitochondrial	16	20	14	14	100.00	87.50	93.33	2.63E-045	12	GO:0005753:mitochondrial ATPase …
6c	DAB complex, transcription preinitiation complex	16	20	16	16	100.00	100.00	100.00	3.20E-042	6	03022:Basal transcription factors …
7c	Exosome	11	14	11	11	100.00	100.00	100.00	2.47E-040	4	GO:0006364:rRNA processing …
8c	eIF3 complex, eukaryotic initiation of translation factor-3	13	16	11	11	100.00	84.62	91.67	8.98E-038	4	GO:0005852:eukar. translation initiation factor …
9c	Nup 107-160 nuclear pore subcomplex	9	11	9	9	100.00	100.00	100.00	1.34E-033	6	GO:0005635:nuclear envelope …
10c	CENP-A NAC-CAD kinetochore complex	13	16	13	13	100.00	100.00	100.00	2.32E-028	2	GO:0005694:chromosome …
	with 20% noise (% of random genes included)		+20% noise		average values=	100.00	93.69	96.62			
	…										
	DISEASES (from OMIM db, human)	**GENES in Ref.**	**GENES Tested**	**GENES Found**	**Common GENES**	**Precision (%)**	**Recall (%)**	**F-score (%)**	**adjusted p.value**	**TERMS Found**	**TERMS Found (only first shown)**
1d	Retinitis pigmentosa	51	62	39	38	97.44	74.51	84.44	1.15E-066	1	GO:0007601:visual perception (BP)
2d	Deafness (autosomal dominant and recesive)	84	101	31	31	100.00	36.90	53.91	4.11E-053	6	GO:0007605:sensory perception of sound …
3d	Cardiomyopathy (dilated, familial and hypertrophic)	44	53	19	19	100.00	43.18	60.32	1.48E-031	5	GO:0008307:structural constituent of muscle …
4d	Epidermolysis bullosa (dystrophica, simplex, junctional)	11	14	11	11	100.00	100.00	100.00	2.77E-024	2	GO:0031581:hemidesmosome assembly …
5d	Congenital disorder of glycosylation (type I and II)	23	28	11	11	100.00	47.83	64.71	1.18E-023	1	00510:N-Glycan biosynthesis
6d	Muscular dystrophy (congenital, limb-girdle, rigid spine)	25	30	11	11	100.00	44.00	61.11	4.45E-019	4	GO:0007517:muscle organ development …
7d	Glycogen storage disease	19	23	9	9	100.00	47.37	64.29	2.53E-018	3	00500:Starch and sucrose metabolism …
8d	Leigh syndrome	8	10	7	7	100.00	87.50	93.33	1.16E-017	10	GO:0006120:mitochondrial electron transport …
9d	Acute Leukemia (lymphoblastic ALLs & myeloid AMLs)	37	45	9	9	100.00	24.32	39.13	7.71E-016	2	05221:Acute myeloid leukemia …
10d	Diabetes mellitus (type 1 or 2, gestational, neonatal)	13	16	5	5	100.00	38.46	55.56	1.01E-010	2	GO:0005975:carbohydrate metabolic process …
	with 20% noise (% of random genes included)		+20% noise		average values=	99.74	54.41	67.68			
	…										
	PATHWAYS (from SGD db, yeast)	**GENES in Ref.**	**GENES Tested**	**GENES Found**	**Common GENES**	**Precision (%)**	**Recall (%)**	**F-score (%)**	**adjusted p.value**	**TERMS Found**	**TERMS Found (only first shown)**
1p	gluconeogenesis	22	27	22	22	100.00	100.00	100.00	7.05E-047	7	GO:0006094:gluconeogenesis …
2p	TCA cycle, aerobic respiration	22	27	23	22	95.65	100.00	97.78	2.36E-037	10	00020:Citrate cycle (TCA cycle) …
3p	sphingolipid metabolism	19	23	16	16	100.00	84.21	91.43	7.64E-033	4	00600:Sphingolipid metabolism …
4p	de novo biosynthesis of purine nucleotides	35	42	21	21	100.00	60.00	75.00	2.02E-030	4	00230:Purine metabolism …
5p	lipid-linked oligosaccharide biosynthesis	12	15	12	12	100.00	100.00	100.00	1.38E-029	5	GO:0005783:endoplasmic reticulum …
6p	ergosterol biosynthesis	11	14	11	11	100.00	100.00	100.00	1.02E-028	7	GO:0006696:ergosterol biosynthetic process …
7p	superpathway of glucose fermentation	14	17	13	13	100.00	92.86	96.30	9.18E-027	5	00010:Glycolysis / Gluconeogenesis …
8p	fatty acid biosynthesis, initial steps	17	21	13	12	92.31	70.59	80.00	9.93E-027	6	GO:0006631:fatty acid metabolic process …
9p	inositol phosphate biosynthesis	19	23	11	11	100.00	57.89	73.33	1.02E-025	3	00562:Inositol phosphate metabolism …
10p	folate biosynthesis	18	22	10	9	90.00	50.00	64.29	6.98E-019	6	GO:0006730:one-carbon metabolic process …
	with 20% noise (% of random genes included)		+20% noise		average values=	97.80	81.56	87.81			
	…										

Figure 3. Analysis of gene sets from 3 biomolecular resources: CORUM, OMIM, SGD. Results of the analysis of thirty gene sets derived from three biomolecular resources: mammalian multiprotein complexes (CORUM), human diseases (OMIM) and yeast cellular pathways (SGD). Each row corresponds to an independent gene set and it includes the result of the functional analysis showing the first metagroup obtained running the method. Each analysis is evaluated with respect to the reference gene sets calculating the *Precision*, *Recall* and *F-score* (in %). The analyses are done introducing 20% random noise; meaning the proportion of random-selected genes added to each query gene set. The number of terms found is indicated in each row. Not all the terms are described due to space restrictions (last column). A complete table, including also the results at 60% random noise and all the information about the specific genes and terms found in each metagroup, is provided as **Table S2**.
doi:10.1371/journal.pone.0024289.g003

of 80 genes, 96 genes are tested (introducing 20% extra randomly selected genes) and the method finds 68 genes, all included in the reference set and functionally linked to 6 terms with a significance of 5.25 e^{-138} (adjusted ***p-value***). Following the same steps, we calculate the results for each one of the thirty reference gene sets. As indicated above these reference sets were taken from three heterogeneous biological sources: complexes (c), diseases (d) and pathways (p). A complete table, including all the results about the specific genes and terms found in each metagroup, is provided as **Table S2**.

Calculating the *Precision*, *Recall* and *F-score* of the method

Since the correct answer is known for each metagroup of the reference gene sets, we can calculate the error rates and estimate the *Precision* and *Recall* of our method. In an information retrieval scenario, *Precision* is defined as the number of relevant document-items retrieved by a search divided by the total number of document-items retrieved by that search, and *Recall* is defined as the number of relevant document-items retrieved by a search divided by the total number of existing relevant document-items (which should have been retrieved). The document-items in our context are the genes. The balanced *F-score* is a measure that combines *Precision* and *Recall* evenly weighted, being the harmonic mean of both. In statistical terminology these parameters –related to type I and type II errors– are defined as:

$$Precision = \frac{(TP)}{(TP+FP)};\ Recall = \frac{(TP)}{(TP+FN)};$$

$$F-score = 2\frac{(Precision \cdot Recall)}{(Precision + Recall)}$$

The *Precision* is a measure of exactness and fidelity, whereas the *Recall* is a measure of completeness. The results (**Figure 3**) reveal that the new functional analysis method proposed is quite precise, because it shows an average *Precision* of 100%, 99.7% and 97.8% in the identification of gene metagroups from protein complexes, diseases and pathways, respectively. Such *Precision* was obtained using a noise level of 20%. This also indicates that it is a rather robust method which allows perturbation in gene lists without losing the major functional signal included in a given metagroup.

The *Recall* –also with 20% noise– was 93.6% and 81.5% for the gene sets obtained for multiprotein complexes and pathways, respectively; and 54.4% for gene sets assigned to protein diseases. This is an interesting observation because it seems that the decrease of the *Recall* follows the same tendency expected if we were considering the strength of "functional units". It is easy to understand that the average cohesion and tightness of the genes associated in multiprotein complexes (i.e. in "molecular machines") should be higher that the cohesion of the genes associated within a pathway, and much stronger that the cohesion of the

genes associated to a disease. In fact, many times there is not a clear functional reason about why a human gene is associated to a given disease [21]. The association is most times heuristic, observational, phenomenological, and not really linked to a known biomolecular cause. This reasoning also provides support to the method, since it shows its power to unravel different types of functional associations, and to disclose cases where the "functional units" holding the linkage between genes are not so well defined.

Finally, it seems that the size of the query groups does not affect the error rates of the method, because sets from 8 to 84 genes were assayed and the values of *Precision* and *Recall* were not dependent on the size. The only need is that each metagroup has to include a minimal number of genes to retrieve enough annotations and terms that allow functional associations. We observed that bellow seven genes it was quite difficult to achieve the linkage between genes and terms, although we do not consider it a critical constraint for high-throughput analysis.

Discussion

Inferring functional linkage between genes and biological terms

Some eloquent studies have asserted that *functional annotation* has become a bottleneck in biomedical science in the current era of high-throughput sequence and structure determination [22,23]. Many genes and gene products are normally annotated by homology, assigning known functions to similar sequences. This procedure can be a potential error-prone which propagates and can contaminate most of the biomolecular databases [23]. The lack of specific knowledge about the biological function of many genes added to a recurrent annotation by simple homology and the frequent use of some terms that become "fashionable" or "promiscuous" under the influence of certain biomedical areas (e.g. cancer) can be a pitfall for many functional enrichment approaches.

Using several information theory principles, we propose a new method for biological functional analysis called ***GeneTerm Linker***, developed with a clear aim of avoiding redundancy and reducing complexity in computational functional annotation, also aiming to combine multiple annotation resources. In **Figure 4** we present a scheme that illustrates the rational followed by ***GeneTerm Linker***. The power of the method is given by the fact that it combines all sources of annotations and biological information regardless of their internal structure in order to provide a single result, in this way it brings together all annotation spaces where a gene list had been interrogated. Lots of efforts have been devoted to use gene ontology (GO) as a main functional annotation space and to find functional similarity metrics in GO using its hierarchical structure and the relationship between its terms. While this is a valid approach, its application cannot be exported to other resources of non-hierarchical but very relevant biological information. As shown in **Figure 4**, our method is able to locate in the same frame terms from GO and from other annotation spaces (KEGG, InterPro, etc) providing metagroups of genes and terms linked with significance scores.

A secondary contribution of our study is to present a comparative analysis of different annotation resources. **Figure 1** reflects that KEGG annotations are more stable and contain less outliers than GO. This is caused by the existence of a thorough curation in KEGG and the fact that GO is, by definition, an ontology resource based on a controlled vocabulary, that many times has to take general broad terms applicable to genes present in very different organisms. We showed that the lack of specificity and the overuse of certain popular terms (e.g. *signal transduction* or *regulation of transcription*, **Figure 1B**) produce a strong influence on the power of the annotation resources and on the quality of their specific application to large query gene lists. Functional characterization of large gene lists, derived from genome-wide experiments, aims ideally to provide a set of annotated groups of genes that should be smaller than the number of genes in the query list [24]. However, currently most researchers in the field realize that is quite difficult to obtain a single and meaningful result using the functional enrichment tools available. The method here proposed (**Figure 4**) solves this problem providing a unique result where the related genes and terms are fuzzy enclosed in metagroups which are evaluated by enrichment, functional coherence and similarity.

In conclusion, after search and comparison with other methods, we can say that the innovation and genuine value of the algorithm presented is to provide a single coherent solution to the problem of functional annotation of lists of genes or proteins. To achieve this, it address the problem of using multiple non-orthogonal and non-homogeneous biological annotation spaces, going beyond enrichment analysis (EA) approaches that provide many lists of genes and annotations usually not integrated, redundant or with low information content. Knowing the use and value of these enrichment approaches, a clear practical problem remains for many biologists that try computer-driven exploration of their candidate gene lists. We expect that the method here presented, ***GeneTerm Linker***, will help to alleviate such difficulties offering a step forward to many gene-based biomedical and biomolecular studies.

Materials and Methods

Reference sets to test the method

A reference set of 59 nuclear proteins from yeast (*Saccharomyces cerevisiae*) that form five well-defined protein complexes [15] was selected as first test set and used in the comparative analysis versus the FAC method [4]. The method was also tested using 30 reference sets of gene metagroups from three biomolecular resources: ***(1)*** CORUM, comprehensive resource of mammalian multiprotein complexes [19]; ***(2)*** SGD, yeast resource that includes a collection of groups of genes involved in cellular pathways [20]; ***(3)*** OMIM, resource that includes groups of genes involved in human diseases [21]. We downloaded these 3 resources and searched for groups composed of at least 8 genes/proteins assigned to specific biological entities within each database, i.e.: assigned to specific multiprotein complexes (c), diseases (d) or pathways (p). Then, we select from each database 10 groups and consider them as reference metagroups in order to test how our method was able to find such groups. The groups are numbered 1c-10c, 1d-10d and 1p-10p. The names of the 10 groups selected from each database are included in **Figure 3** and all the details about the genes included in each reference metagroup are provided in **Table S2**.

Formal definition of *GeneTerm-sets*

The input to the algorithm are elements defined as *GeneTerm-sets* that correspond to combinations of genes/terms/***p-value*** (considered *frequent itemset*) derived from functional annotation enrichment:

$$E_i = \langle G_i, A_i, p_i \rangle$$

$\boldsymbol{E_i}$ *ith* element; $\boldsymbol{G_i}\{g_1, g_2 \cdots g_m\}$ set of genes; $\boldsymbol{A_i}\{a_1, a_2 \cdots a_n\}$ set of terms; $\boldsymbol{p_i}$ ***p-value***

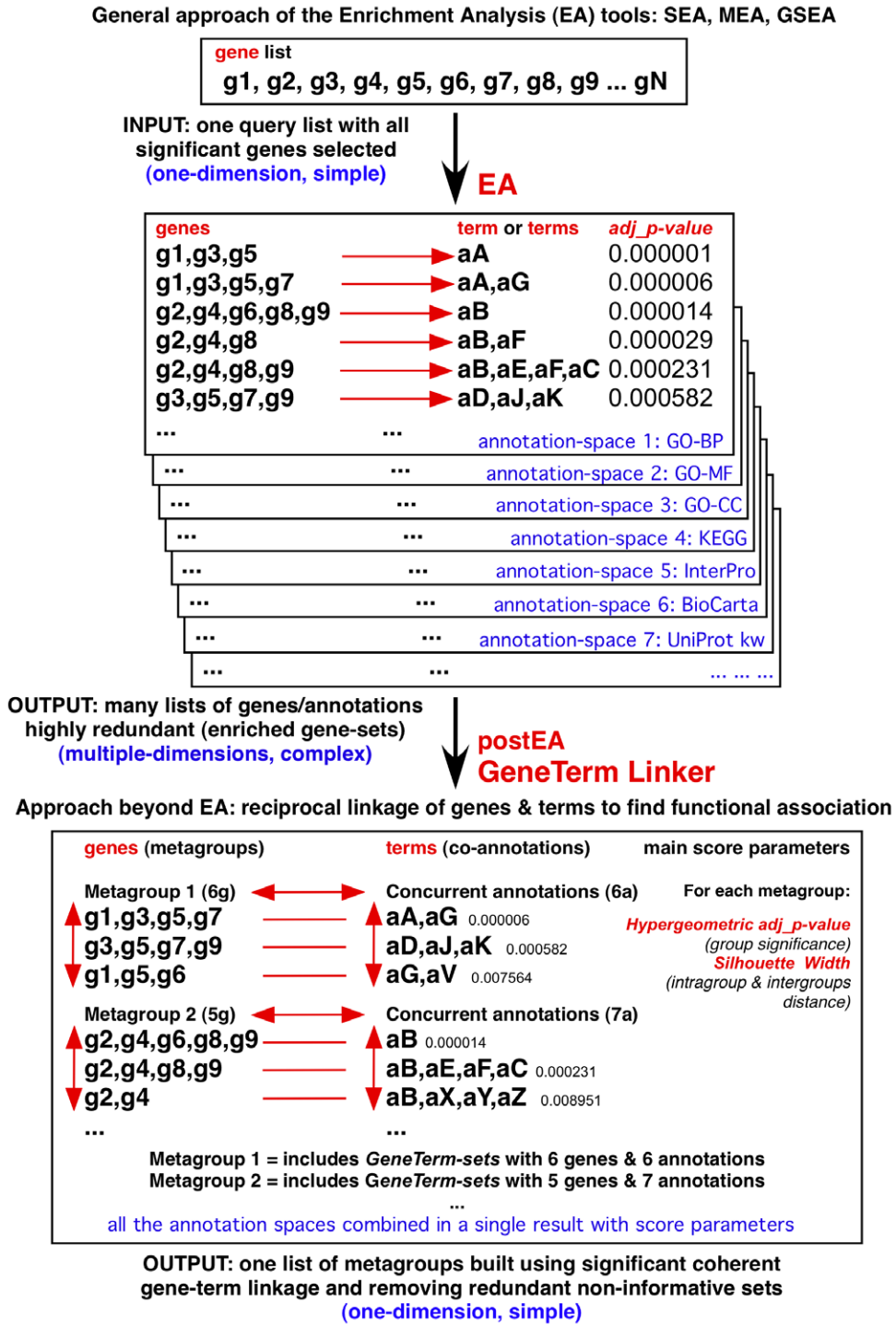


Figure 4. Scheme of the rational followed by *GeneTerm Linker* method. Scheme that illustrates the rational followed by the ***GeneTerm Linker*** method proposed. The method provides a single result combining all annotation spaces where a gene list has been interrogated. The method uses filters for promiscuous and redundant terms/annotations as it is described in the step 1 and 3 of the algorithm.
doi:10.1371/journal.pone.0024289.g004

Mathematical description of the calculation of distances

For each element $\boldsymbol{E_i}$ a vector $\boldsymbol{v_i}$ contains the occurrence of each gene with respect to the whole input gene list and the ***p-value*** of each element $\boldsymbol{E_i}$ weighted by factor $\boldsymbol{M}$ = total number of genes in the list:

$$\mathbf{v}_i = (\delta(g_1,G_i),\delta(g_2,G_i),\ldots,\delta(g_M,G_i),Mp_i)$$

$$\delta(g_k,G_i) = \begin{cases} 1 & g_k \in G_i \\ 0 & g_k \notin G \end{cases}$$

The pair-wise distances between all vectors $\boldsymbol{v_i}$ are calculated using the *Cosine Similarity* that is derived from the *Jaccard Similarity coefficient*:

$$D(E_i,E_j)=1-\cos(\mathbf{v_i},\mathbf{v_j})=1-\frac{\mathbf{v_i}\cdot\mathbf{v_j}}{\|\mathbf{v_i}\|\ \|\mathbf{v_j}\|}$$

Mathematical description of *complete cover* and application to redundancy removal

Each resulting metagroup is formed by a selected collection of *GeneTerm-sets* that keep maximum similarity. The redundancy within the preliminary metagroups is eliminated calculating the *complete cover* of each metagroup (to guarantee the completeness of the data) and then removing the *GeneTerm-sets* that do not include any new gene or any new term. Formally:

given a metagroup $\Gamma=\{E_1,E_2\ldots E_N\}$

and a subset $\Delta\subseteq\Gamma$, Δ is a cover of Γ if

$$\Delta \textit{ is cover of } \Gamma \Leftrightarrow \left(\bigcup_{E_k\in\Delta}\gamma(E_k)=\bigcup_{E_k\in\Gamma}\gamma(E_k)\right)\wedge \left(\bigcup_{E_k\in\Delta}\alpha(E_k)=\bigcup_{E_k\in\Gamma}\alpha(E_k)\right)$$

$$\gamma(E_i)=G_i$$

$$\alpha(E_i)=A_i$$

Supporting Information

Table S1 Complete functional analysis of 59 yeast proteins using *GeneTerm Linker* method. Data file (.xls) containing the complete results provided by ***GeneTerm Linker*** corresponding to the functional analysis of the 59 nuclear yeast proteins (which has been partially presented in **Figure 2C**). The file has two spreadsheets: *(A)* includes a complete view of the same table as **Figure 2C**; *(B)* includes the complete output results provided by ***GeneTerm Linker*** algorithm, showing the five metagroups found with all *GeneTerm-sets* assigned to each metagroup.
(XLS)

Table S2 Complete functional analysis of 30 gene sets from 3 resources (CORUM, OMIM and SGD) using *GeneTerm Linker* method. Data file (.xls) containing the complete results provided by ***GeneTerm Linker*** corresponding to the analysis of 30 gene sets derived from 3 biomolecular resources: CORUM, OMIM and SGD (which has been partially presented in **Figure 3**). Each row corresponds to the functional analysis of one gene set and shows only the *first metagroup* found by the method. All genes and terms found in the *first metagroups* of each gene set are included, together with the statistical parameters (*Precision*, *Recall* and *F-score* in %) and the adjusted ***p-value*** corresponding to such metagroups. Each analysis is done twice for each gene set, introducing 20% or 60% random-selected genes.
(XLS)

Author Contributions

Conceived and designed the experiments: JR CF. Performed the experiments: CF RN. Analyzed the data: CF RN AP JR. Contributed reagents/materials/analysis tools: CF RN AP JR. Wrote the paper: JR. Developed the programming code: CF RN. Implemented the algorithm steps: CF RN. Developed the mathematical formulation: RN AP JR. Developed the web-site to use the method: RN JR. Critical contribution and manuscript correction: CF RN AP.

References

1. Huang DW, Sherman BT, Lempicki RA (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res 37: 1–13.
2. 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.
3. Nogales-Cadenas R, Carmona-Saez P, Vazquez M, Vicente C, Yang X, et al. (2009) GeneCodis: interpreting gene lists through enrichment analysis and integration of diverse biological information. Nucleic Acids Res 37: W317–322.
4. Huang DW, Sherman BT, Lempicki RA (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4: 44–57.
5. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, et al. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 102: 15545–15550.
6. Gene Ontology Consortium (2010) The Gene Ontology in 2010: extensions and refinements. Nucleic Acids Res 38: D331–335.
7. Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M (2010) KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Res 38: D355–360.
8. Apweiler R, Martin MJ, O'Donovan C, Magrane M, Alam-Faruque Y, et al. (2010) The Universal Protein Resource (UniProt) in 2010. Nucleic Acids Res 38: D142–148.
9. Alves R, Rodriguez-Baena DS, Aguilar-Ruiz JS (2010) Gene association analysis: a survey of frequent pattern mining from gene expression data. Brief Bioinform 11: 210–224.
10. Gupta T, Seifoddini H (1990) Production data based similarity coefficient for machine-component grouping decisions in the design of a cellular manufacturing system. International Journal of Production Research 28: 1247–1269.
11. Toivonen H, Klemettinen M, Ronkainen P, Hatonen K, Mannila H (1995) Pruning and grouping discovered association rules. In: MLnet Workshop on Statistics, Machine Learning, and Discovery in Databases, Crete, Greece. pp 47–52.
12. Draghici S, Khatri P, Martins RP, Ostermeier GC, Krawetz SA (2003) Global functional profiling of gene expression. Genomics 81: 98–104.
13. Benjamini Y, Hochberg Y (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J Roy Stat Soc (Ser B) 57: 289–300.
14. Rousseeuw PJ (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics 20: 53–65.
15. Bader GD, Hogue CW (2003) An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinformatics 4: 2.
16. Prieto C, De Las Rivas J (2006) APID: Agile Protein Interaction DataAnalyzer. Nucleic Acids Res 34: W298–302.
17. Hernandez-Toro J, Prieto C, De Las Rivas J (2007) APID2NET: unified interactome graphic analyzer. Bioinformatics 23: 2495–2497.
18. Dalton L, Ballarin V, Brun M (2009) Clustering algorithms: on learning, validation, performance, and applications to genomics. Curr Genomics 10: 430–445.
19. Ruepp A, Waegele B, Lechner M, Brauner B, Dunger-Kaltenbach I, et al. (2010) CORUM: the comprehensive resource of mammalian protein complexes - 2009. Nucleic Acids Res 38: D497–501.
20. Engel SR, Balakrishnan R, Binkley G, Christie KR, Costanzo MC, et al. (2010) Saccharomyces Genome Database provides mutant phenotype data. Nucleic Acids Res 38: D433–436.
21. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res 33: D514–517.
22. Medrano-Soto A, Pal D, Eisenberg D (2008) Inferring molecular function: contributions from functional linkages. Trends Genet 24: 587–590.
23. Llewellyn R, Eisenberg DS (2008) Annotating proteins with generalized functional linkages. Proc Natl Acad Sci U S A 105: 17700–17705.
24. Merico D, Isserlin R, Stueker O, Emili A, Bader GD (2010) Enrichment map: a network-based method for gene-set enrichment visualization and interpretation. PLoS One 5(11): e13984.