Prólogo

A mediados del año 1998 decidí que quería realizar los estudios de tercer ciclo que conllevaban a la obtención del grado de Doctor. Obté a los cursos de Doctorado mediante mi licenciatura en Biología y mediante la especialidad de Biomedicina que encaminé hacia la Genética Molecular. Pero, como hacia algún tiempo que "sospechaba"que la informática y las matemáticas propiciarían un cambio en nuestra concepción de la Genética, decidí realizar la Licenciatura en Ciencias y Técnicas Estadísticas.

Además de los conocimiento teóricos adquiridos, en los ocho años que llevo realizando la tesis doctoral, he estado trabajando en varios centros ya sea como profesora, investigadora o, incluso, como técnico en estadística. Conociendo así muchos modos de pensar y de hacer.

Y, a veces, lo que parece una ventaja puede acabar siendo un gran inconveniente. He revisado unas diez tesis doctorales para ayudarme a escribir ésta. He revisado tesis en Genética, en Estadística y en Informática. Dándome cuenta que no existe un patrón común ente ellas. Así que el resultado de esta tesis doctoral es un extraño cruce entre todas las áreas en las que he trabjado de un modo u otro. Resultando demasiado técnica con un lenguaje demasiado sobrio para una tesis en Genética y, en contrapartida, es una tesis demasiado aplicada para considerarse una tesis en Estadística. La estructura resultante parte de las tesis en Genética pero tomando algunas características de las tesis en Estadística por lo que los modelos y algoritmos desarrollados se encaben dentro del apartado de Resultados.

Por todo ello doy las gracias al tribunal por el esfuerzo añadido que supone leer y valorar esta tesis doctoral.

Agradecimientos

En primer lugar, quiero agradecer esta tesis doctoral a todas las personas con las que he trabajado durante este periodo de tiempo y que, sin lugar a dudas, han contribuido positivamente en la realización de la misma. Durante los ocho años que ha durado mi periodo pre-doctoral he conocido y trabajado con mucha gente por lo que es imposible encaberos a todos en esta lista. Así que pido disculpas a todos los que finalmente no aparezcais.

En segundo lugar quiero agradecer a todos los profesores y amigos que, sin tener nada que ver con este manuscrito, encaminaron mis primeros pasos. Especialmente a Arturo, Bruno, Núria, Marta, Magüí y "mis primas"; Glòria G, Elena y Glòria J por la paciencia y por estar a mi lado, de un modo u otro, todo este tiempo. Y, también a los amigos que ya no están; a Dani por contagiarme sus ganas de vivir y sus inquietudes que, espero, estén conmigo siempre y también a Isabel que aunque no tuve la oportunidad de conocerla bien siempre admiré su coraje.

Agradecerte, José Luis, los maravillosos seis años que he pasado a tu lado y que ya nunca volverán. Por hacerme sentir "normal"después de tanto tiempo viviendo al otro lado. Ojalá la vida, mi vida y las circumstancias hubieran sido distintas. No me cabe la menor duda que hubiéramos podido ser felices pero los caminos a veces, sin quererlo, simplemente se tuercen. Busco algo que no existe y que nunca encontraré; no quiero arrastrarte conmigo mar adentro, tú que puedes busca la manera de ser feliz y yo buscaré que hacer con el resto de una vida que nunca imaginé. No sé si todo tiene fin, así que por si acaso es sólo un *Hasta luego*.

A las chicas del despacho del prefabricado; Cristina, Montse, Pilar, Mar, Esther, Samantha. A Meritxell Girvent, Mireia Valero y Jordi Pérez por el soporte moral.

Quiero agradecer al Departamento de Estadística de la Universidad de Barcelona la oportunidad que me ofrecieron; especialmente al Dr. Jordi Ocaña y al Dr. Carles M. Cuadras. A los compañeros Aurea Grané, Francesc Oliva y Sergi Vives que me enseñaron a trabajar y a preparar las clases. A Regina por que sin ti no sé que hubiera hecho, por todos los maravillos ratos que perdimos (y que perdemos) y a José Antonio (JAS) por los ánimos recibidos en todos los momentos bajos.

A mi guapísima e inteligentísima compañera de doctorado; "Por que Raquel compartir la vida contigo ha sido fantástico, me has hecho crecer como persona (que es lo más importante) y como profesional. Que el tiempo que compartimos ahora me parece corto comparado con todo el tiempo que te he estado echando de menos. Así que ¡¡acaba pronto en Columbia!!"

A mi compañera de fatigas estadísticas y laborales, Georgia.

Al Dr.Roderic Guigó y a su grupo por su colaboración y por toda la ayuda recibida en

mis primeros años de doctorado. Especialmente a Genís Parra, Enrique Blanco, Robert Castelo y Pep Abril. Y, a Enrique Blanco por la amistad recibida y por dirigir mis primeros pasos con Perl, awk y el resto de programillas (aún necesitaré tu ayuda un poquito más ;-))).

Al Dr. Ralph Herwig y a su grupo por la maravillosa estancia en Berlin, por hacerme sentir como en casa. Especialmente a Axel, Reha, Wasco, Mario y Anita.

Al al Dr. Arcadi Navarro y a la Dra. Elodie Gazave por la motivación crítica recibida y por las aportaciones intelectuales.

Al Dr. Xavier Estivill y a su grupo, especialmente al Dr, Lluís Armengol y Manel. Quiero agradecerle a Lluís Armengol su buen hacer como investigador y como persona que sin duda me ha motivado a trabajar más y mejor. A Manel quiero agradecerle que sea una persona de buen trato, su disponibilidad y la ayuda recibida.

A la genete del SERC y del UNIC por acogerme tan bien en mi nueva etapa profesional. Especialmente al Dr. Carlos A. González y a la Dra. Núria Sala. A mis compañeras de despacho; Noemie y Gienesa por el apoyo moral recibido en esta última etapa de mi tesis que ha sido, realmente, muy dura para mí. Y, también, a Joan Valls, Jordi ,Laura P, Laura E y Ramon agradecerles su apoyo, su comprensión y su ayuda. A Paula quiero agradecerle la última conversación en autobús sin la que esta tesis doctoral se hubiera demorado.

Bien, ahora quiero pedir una disculpa y, a la vez, agradecer toda la ayuda recibida emocional y profesionalmente a la gente del grupo del Dr. Lauro Sumoy y del Dr. Luis A. Pérez-Jurado; a Juanjo por introducirme a la bioinformática de las matrices de expresión y CGH, a Eva por encontrar en ella una maravillosa profesional con la que he podido aprender de un modo fácil y sencillo, por su carácter, su manera de ser y por las vidas paralelas. Sin duda Eva ha sido una de las mejores cosas que me he llevado de mi estancia en el PRBB y, espero, que sólo sea el inicio de un largo viaje. Y al resto del equipo; Anna, Fran y Susana.

Del grupo de Luis quiero agradecer y remarcar la amistad recibida de Helena y de Olaya. Por que sin dudarlo ni un segundo son las personas con más chispa que he conocido. Quiero agradecerles que disfruten tanto de simplemente estar investigando, que les guste tanto transmitir el conocimiento adquirido y por poseer las herramientas necesarias para ello.

A Ivon que es, sin dudarlo ni un segundo, una investigadora excepcional de quien auguro una más que brillante carrera científica; "Qué te he de dir? Ha estat molt profitós sentir-te a la vora i només lamento tot el que m'ha quedat per apendre de tu.".

A Anna A por haberme enseñado tanto de Williams y de duplicones y de todo lo demás. Y,.. diría muchas cosas pero como estoy convencida que este es sólo un inicio no un final voy a mandarte un abrazo y dejaré que el futuro me brinde la oportunidad de

agradecerte la amistad recibida.

A Miguel del Campo y a Blanca Gener por acogerme tan bien en su despachito. Por todas las charlas que hemos compartido que me han permitido reencontrarme conmigo misma.

A la gente del labo; a Raquel, Clara, Benja, Jaume, Jesus, Anabel, Verena,...agradecerles los cafés y las tertulias,...

Por último agradercerles, de todo corazón, a mis tres directores de tesis la oportunidad, el apoyo y el aprendizaje recibido. Cuando llegué al PRBB, sinceramente, creía que ya estaba más que preparada para trabajar en cualquier aspecto de estadística aplicada a la Genética. Hoy sé que nunca estaré del todo preparada pero que tengo el resto de la vida para conseguirlo y ello, lo sé, sin duda, gracias a vosotros. A Sergi (director no oficial) quiero agradercerle el esfuerzo que ha supuesto repasar todos los diseños y formulaciones realizadas. Hay gente que ha nacido para cambiar la inercia de las cosas, para plantear y replantear a los demás las situaciones injustas que tan a menudo ocurren en el mundo de la ciencia. Sergi es una de estas personas y quiero agradecerle su disconformismo y sus críticas racionales que sin duda van a llevarnos a todos hacia un mundo un poco mejor. A Lauro por darme la oportunidad de realizar un poquito de laboratorio y por la transmisión de todo el conocimiento adquirido.

Luis, como ves, te tengo en un párrafo a parte. Podría escribir hojas y hojas intentanto explicarte, explicarme y explicar a los demás todo lo que me has aportado pero es totalmente imposible. Tenía tantas ganas de terminar la tesis,... y, ahora, quisiera parar el tiempo, volver atrás y volver a conocerte. No he conocido nunca a nadie como tú, ni lo conoceré. No sé que será de mí, no sé si conseguiré nunca consolidarme como investigadora (lo intentaré con todas mis fuerzas, espero que algún día te sientas orgulloso de haberme dirigido la tesis) pero pase lo que pase será, sin duda, gracias a ti. Has cambiado el rumbo de mi vida, has hecho aparecer cosas en mí que murieron hace mucho mucho tiempo. Gracias por ser una persona tan genial, por todos tus conocimientos, por el esfuerzo que has hecho con todos nosotros,... en fin *Un beso* y hasta la vista!

Finalmente, lamento haber sido tan egoísta y, en vez de escribir sobre vuestro trabajo y sobre lo maravillosos que sois, haberme centrado tanto en lo mucho que me habéis aportado emocional y profesionalmente.

Va por todos vosostros, por hacerme creer que el futuro puede ser mucho mejor.

Índice

1. Introducción

1.1. Preámbulo

El genoma y su complejidad

Los grandes adelantos tecnológicos surgidos en las dos últimas décadas han dado lugar a un espectacular avance en el conocimiento biológico y médico, facilitando la comprensión de la arquitectura genómica de los organismos que ha permitido identificar las causas de algunas enfermedades humanas de origen genético así como identificar nuevos mecanismos evolutivos.

De este modo, hoy en día sabemos que la complejidad de un organismo no es proporcional a la cantidad de material genético que posee y, esto, es debido, en gran parte, a que el genoma de los organismos contiene una alta proporción de repeticiones sin potencial codificante [1]. La abundacia relativa de estas repeticiones (ver Tabla 1) da lugar a grandes diferencias de tamaño entre los genomas de organismos eucariotas con una complejidad funcional similar (por ejemplo *Drosophila Melanogaster* y *Podisma Pedestris* poseen una compleijidad equivalente pero el genoma de *Podisma Pedestris* de 18.000 Mb es 100 veces mayor que el de *Drosophila Melanogaster* de 180 Mb). Así el porcentage de ADN codificante presente en un genoma varía entre especies, llegando a representar una pequeña parte en el caso de los mamíferos donde menos de un 5 % de su genoma es codificante. El hecho de que la mayor parte del genoma humano esté compuesto por repeticiones las convierte en una importante fuente de variabilidad y de plasticidad genómica. El conjunto y disposición de repeticiones varía entre individuos de la misma especie y por ello se han utilizado en estudios de asociación (segregación de un/os polimorfismo/s entre casos y controles no relacionados) y en estudios de ligamiento (cosegragación de la enfermedad con polimorfismo/s entre sujetos relacionados). Además, éstas pueden afectar a la transcripción y traducción de genes mediante inserción produciendo enfermedades de tipo genético, susceptibilidad variable a enfermedades complejas o cambios evolutivos.

Las duplicaciones segmentarias y su implicación en evolución

La reciente secuenciación de genomas de varios organismos dio a conocer otro tipo de repeticiones en bajo número de copias llamadas LCRs (*Low-Copy Repeats*) o DS (*Duplicaciones Segmentarias*).

Las DSs se definen como repeticiones en bajo número de copias que tienen una longitud mínima entre 1-5 kb (varía según el autor) y una longitud máxima de 500 kb con una identidad nucleotídica >90 % [2, 3, 4]. Su elevada identidad nucleotídica indica que se formaron en la historia reciente de cada una de las especies que las contienen. Éstas pueden contener secuencias repetitivas, genes, estructuras intrón-exón reconocibles, pseudogenes y otros elementos funcionales, además de ADN no codificante. Hasta el momento se han identificado DSs en todas las especies eucariotas secuenciadas, incluyendo levaduras [5, 6], plantas [7], peces [8], aves [9] y mamíferos [10, 11, 12]. Así, el proyecto de secuenciación del

Tabla 1: **Elementos repetitivos no codificantes en el genoma humano**

Elementos repetitivos clásicos	Tipo			Longitud	N [a](%) Total estimado	Distribución
Bloques de repeticiones en tándem	ADN satélite			$10^2 - 10^6$ pb	? (3 %)	centrómeros y telómeros
	Minisatélites			6-100 pb en bloques de 20 a 50 repeticiones	$1,5 * 10^5$ (3 %)	subtelómeros
	Microsatélites			1-6 bp y longitud total < 200 pb	10^6 (3 %)	regiones distales del cromosoma
Repeticiones dispersas	No-LTR	LINE	L1	6-8 kb	$5,5 * 10^5$ (15 %)	Regiones ricas en AT
			L2	6-8 kb	$3 * 10^5$ (6 %)	Uniformemente distribuidos
		SINE	Alu	100-300 bp	$1,3 * 10^6$ (11 %)	Regiones ricas en GC
			MIR	300-400 bp	$3 * 10^5$ (2-3 %)	Regiones ricas en GC
	LTR			3-11 kb	$4,5 * 10^5$ (8 %)	Relativo a la superfamilia considerada

[a]: N ;Número de copias.
(%); Porcentaje del genoma ocupado por este tipo de repeticiones
según Feuk et al 2006 [13], Medstrand et al 2003 [14]
En esta tabla no se ha tenido en cuenta el ADN repetitivo que codifica para ARNr

genoma humano dio a conocer que aproximadamente el 5-6 % del genoma humano está formado por DSs o duplicones, que surgieron en los últimos 35 millones de años de evolución.

Las DSs se clasifican, según la posición de las copias respecto a la original, en intercromosómicas si se encuentran en cromosomas distintos y en intracromosómicas si se hallan dentro del mismo cromosoma, siendo este último grupo el más abundante. Además, su distribución a lo largo del genoma humano no es homogénea, los cromosomas 7, 9, 15, 16, 17, 19, 22 e Y están significativamente enriquecidos en duplicaciones intra e intercromosómicas mientras que los cromosomas 2, 3, 4, 5, 8, 14 y 20 tienen un contenido reducido [2]. Y, a su vez, las regiones pericentroméricas y subteloméricas están sensiblemente enriquecidas [4].

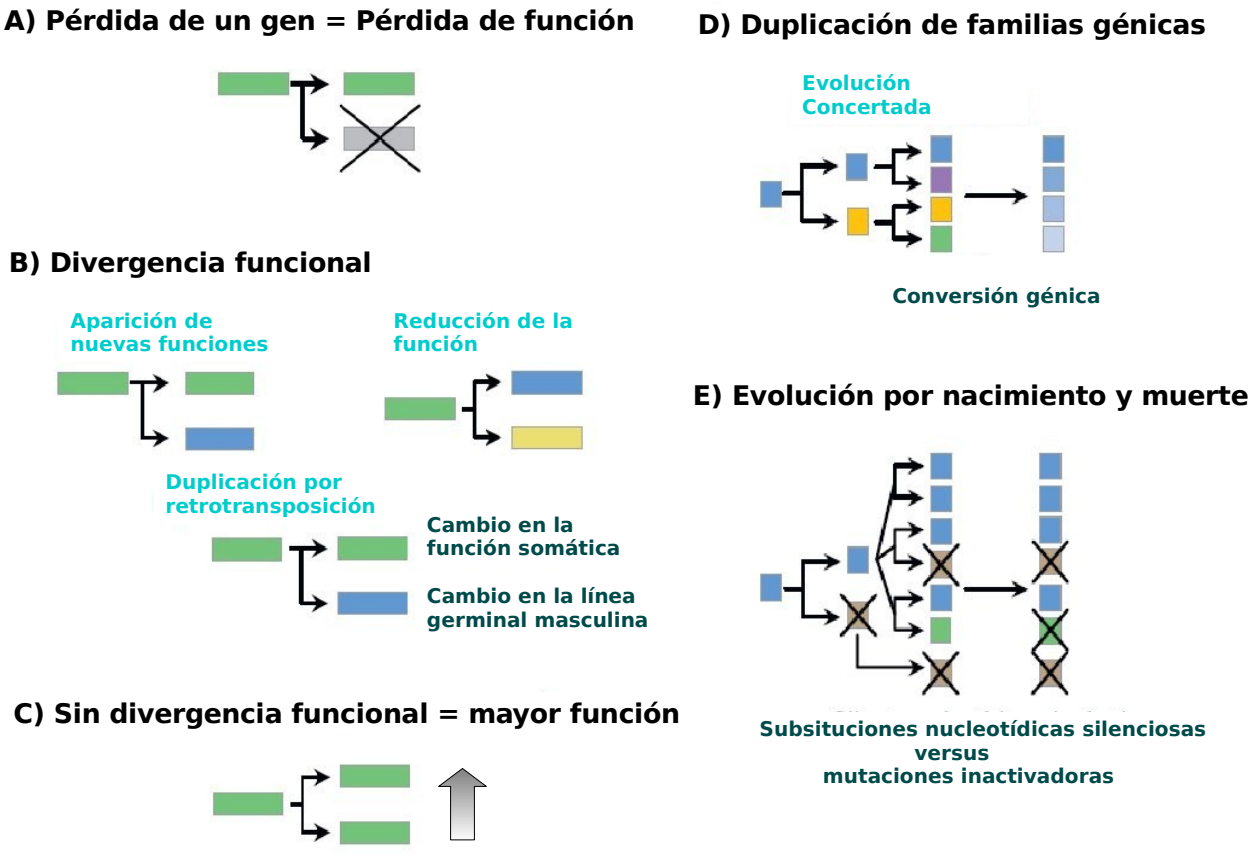

Figura 1: **Mecanismos evolutivos que pueden estar mediados por DS**. (A-C) Duplicación de genes individuales que puede producir A) no funcionalidad de la copia, B) neofuncionalización; pocas veces la copia y el ancestro evolucionan y divergen en función o bien C) la copia y el gen complementan su función, subfuncionalización. Tanto la neofuncionalización como subfuncionalización pueden producirse por retrotransposición. (D-E) Duplicación de familias génicas. D) Después de varias duplicaciones la conversión génica homogeniza las secuencias de los miembros de esta familia. E) Ejemplifica un proceso de equilibrio entre el gen y las copias mediante mutaciones. Figura adaptada de Conrad et al 2007 [15]

Los mecanismos que las generan son aún desconocidos aunque se han propuesto dos

modelos; uno mediado por repeticiones tipo *Alu* que podría explicar, al menos, la aparición de aproximadamente el 12 % de todos los duplicones humanos [16] y otro basado en la presencia de *lugares frágiles* caracterizados por la poca estabilidad de la hélice del ADN [17]. NAHR actúa como un tercer mecanismo generador de DSs, una vez éstas se han formado.

Se cree que la aparición de nuevas DSs tiene un rol importante en la especiación, ya que la duplicación génica facilita la generación de nuevas funciones [18] (ver Figura 1). La comparación de secuencias entre distintas especies ha permitido constatar que este fenómeno ha dado lugar a las α y β globinas y al clúster de genes HOX presentes en todos los vertebrados.

Las DSs y su implicación en enfermedades humanas

Más allá de intentar conocer los mecanismos evolutivos que han dado lugar a la especie humana la existencia de estos bloques comunes inter e intracromosómicos predispone al genoma humano a reordenamientos cromosómicos que pueden ser recurrentes mediados por recombinación homóloga no alélica (NAHR) debido al alto grado de homologia existentes entre ellos [19, 20] (ver Figura 2).

En estos reordenamientos cromosómicos mediados por NAHR puede haber pequeñas pérdidas (deleciones) y/o ganancias (amplificaciones) de material genético y ello es la causa de más de 25 enfermedades de origen genético con una incidencia global de 1 cada 1.000 nacidos vivos. Entre ellas se encuentran el síndrome de Williams-Beuren (WBS), el síndrome de Angelman/Prader-Willi (AS/PWS) y el síndrome de Smith-Magenis (SMS), entre otras (ver sección 1.3.1). Esta lista se incrementa constantemente ya que la aparición de nuevas técnicas de estudio permite detectar nuevos síndromes.

Reordenamientos patogénicos versus variantes polimórficas

Un reordenamiento cromosómico es el resultado de una rotura cromosómica y su posterior unión anómala. El resultado de un reordenamiento cromosómico puede ser una alteración cromosómica equilibrada, cuando el contenido total de material genético se conserva, o desequilibrada, si se gana o se pierde material.

Entre este tipo de alteraciones se encuentran las *duplicaciones* (ganancia de material genético que puede ser debido a la presencia de un cromosoma extra, a la presencia de un cromosoma marcador o la presencia de una región duplicada en el mismo cromosoma o en otro), *deleciones* (pérdida de material genético en una región o cromosoma), *inversiones* (una región del cromosoma se presenta en sentido inverso al habitual), *translocaciones* (intercambio de material genético entre dos cromosomas que puede ser balanceado o no), *isocromosomas* (cuando un cromosoma presenta dos brazos idénticos), *cromosomas en anillo* (producidos por la unión de los extremos cuando se produce una rotura a ambos lados del centrómero) y *cromosomas marcadores* (presencia de un cromosoma extra que se corresponde con una región o parte de otro/s cromosoma/s de origen desconocido) o

cromosoma derivativo (si se conoce el origen).

Se denomina reordenamiento citogenético cuando es visible en cariotipo rutinario al microscopio (> 5 o 10 Mb) y críptico cuando es necesario aplicar técnicas más sensibles para su visualización, normalmente por debajo de 5 Mb.

Los reordenamientos equlibrados no implican ganancias/pérdidas de ADN por ello no son detectables mediante la mayoría de técnicas basadas en cuantificación de la fluorescencia. El efecto y número de estos reordenamientos, fundamentalmente inversiones, sobre el fenotipo se considera subestimado debido a las dificultades existentes en su detección [13]. La importancia de la detección de las mismas reside en el hecho de que algunas inversiones se han identificado como factor de riesgo o causa de enfermedades. Así se ha demostrado que, por ejemplo, en los padres de pacientes afectados por el síndrome de Williams-Beuren o de Angelman entre otros, poseen una prevalencia mayor de inversiones en la región reordenada [21, 22] e, incluso, algunas inversiones han sido identificadas como causa directa en enfermedades (y por ello este tipo de inversiones no se consideran polimórficas) entre las que se encuentran una inversión recurrente de 400 kb de longitud en el factor VIII que se halla en el 40 % de los pacientes afectados de Hemofilia A e inversiones del gen IDS relacionado con el síndrome de Hunter [23]. Además se ha demostrado que una inversión de 900 kb en el cromosoma 17q21.31 se halla bajo selección positiva en europeos. Dicha inversión predispone a la deleción de la región [24]. Todos estos ejemplos muestran la importancia de conocer el número y disposición de las mismas en la población general.

Las técnicas de cuantificación de ADN a gran escala desarrolladas recientemente han dado a conocer la existencia de otro tipo de polimorfismos. Así en individuos fenotípicamente normales, además de inversiones, cambios de un solo nucleótido (SNPs) y variabilidad en secuencias repetitivas se hallan también numerosas deleciones y duplicaciones submicroscópicas y reordenamientos complejos llamados conjuntamente variaciones en el número de copias (CNVs)[25, 26, 27, 28].

En subsiguientes estudios miles de CNVs han sido descritas englobando más del 20 % (**??** base de datos tacg) del genoma humano, pudiendo dar lugar a una divergencia en longitud entre dos genomas de dos individuos sanos no relacionados de 9 Mb [13, 29, 30]. Recientemente se ha descubierto que algunas de estas nuevas variantes polimórficas confieren susceptibilidad en caracteres multifactoriales además de contribuir a la variabilidad fenotípica interindividual. Por ejemplo deleciones del gen UGT2B17, que está vinculado al metabolismo de la testosterona, predisponen al cáncer de próstata encontrándose diferencias interindividuales y entre poblaciones [31, 32, 33, 34], un número bajo de copias de FCGR3 predispone a glomerulonefritis [35], el incremento en GSK3β predispone a padecer trastorno bipolar [36] y una duplicación de 0,5 Mb en el cromosoma 21 que contiene el gen APP se ha asociado a Alzheimer [37]. De este modo, se cree que, las CNVs conjuntamente con los SNPs son las principales fuentes de variabilidad genética interindividual responsables de una divergencia de aproximadamente un 1 % y un 0,1 % respectivamente [38] entre dos genomas escogidos al azar. Así, CNVs y SNPs se combinan con factores

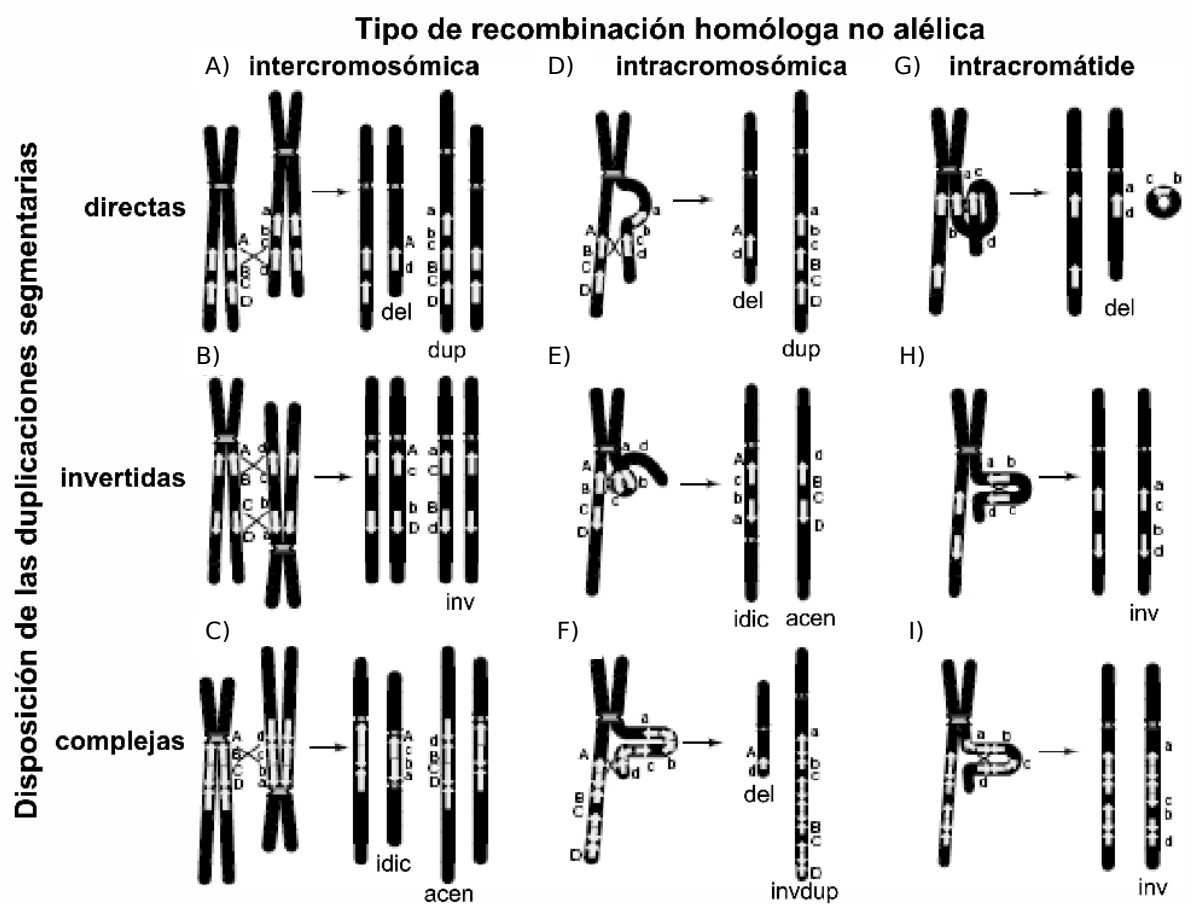

Figura 2: **Mecanismos moleculares de producción de algunos reordenamientos cromosómicos estructurales, mediados por recombinación homóloga no alélica** Los distintos tipos de recombinación (intercromosómica, intracromosómica o intracromátide) dependen de la disposición de las secuencias facilitadoras o DSs, generándose deleciones (A, D, F, G), duplicaciones (A, D, F), inversiones (B, F, H, I), cromosomas acéntricos o dicéntricos (C, E), o reordenamientos complejos (F, I). Figura adaptada de Stankiewicz y Lupski 2002 [20]

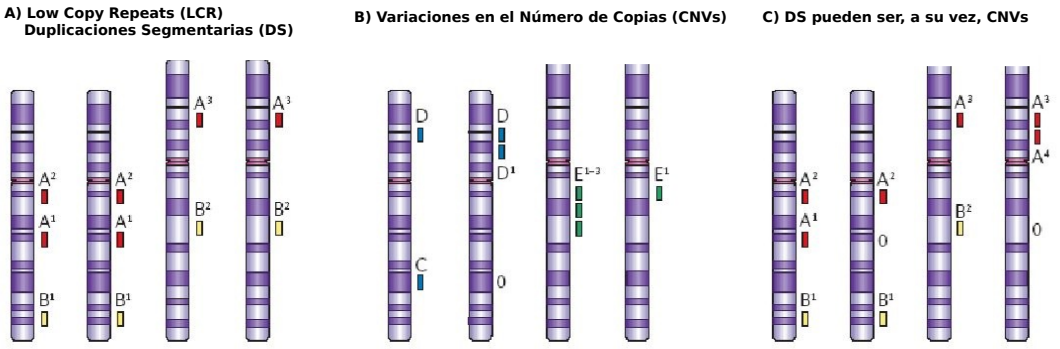

Figura 3: **Relación entre DS y CNV**. En A) están representadas las DSs. Estas DSs se presentan en los dos juegos de cromosomas tanto las que son de tipo intracromosómico como intercromosómico. En B) se representan las CNVs que representan, al contrario que las DSs, un número variable de copias. En C) se representa un caso más complejo en la que se representan segmentos que puede estar presentes en DSs y CNVs a la vez. Figura adaptada de Feuk et al 2006 [13]

medioambientales para dar lugar al fenotipo del individuo.

Variantes de secuencia

Los SNPs, trás su descubrimiento en los años 1960-1980 (en los años '60 se describieron, por primera vez, cambios polimórficos en las secuencias proteícas que, posteriormente en los años '80, se demostró que eran causadas por SNPs), han sido ampliamente estudiados demostrándose su importante papel en la variabilidad genética y fenotípica interindividual además de hallarse asociaciones a enfermedades y a su susceptibilidad. En la actualidad son muy utilizados en la realización de estudios de asociación, en el estudio de caracteres complejos, en análisis de ligamiento... Pero el conocimiento de la localización y contenido de las DSs y CNVs ha puesto de relieve que una parte importante de estos SNPs son, en realidad, cambios de uno o pocos nucleótidos que se dan entre secuencias parálogas contenidas en estas repeticiones llamadas también PSVs (*Paralogue Sequence Variant*). Se calcula que alrededor de un 20 % de todos los SNPs depositados en la base de datos del NCBI (*National Center for Biotechnology Information*) son, en realidad, PSVs [38]. Otros estudios [39] señalan que únicamente un 20 % de todos los SNPs situados sobre duplicones son PSVs y que alrededor de un 30 % son MSVs (*Multiple Sequence Variant*). El término MSV se emplea cuando cada una de las copias parálogas contiene a su vez SNPs o bien cuando pueden coexistir SNPs con deleciones o inserciones de un solo nucleótido, ver Figura 4 [39].

El rol de las PSV/MSVs en enfermedades de origen genético así como su contribución a la variabilidad genómica interindividual es todavía desconocido. También se desconoce el número exacto y disposición de estas variantes en el genoma humano y por ello es de gran importancia desarrollar métodos bioinformáticos y experimentales que permitan su detección y análisis.

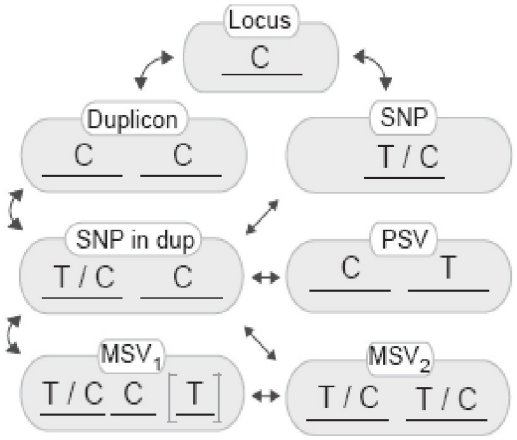

Figura 4: **Relación entre SNP, PSV y MSV.** Figura adaptada de Fredman et al 2004 [39]

Las variantes estructurales y su repercusión en la expresión génica

Más allá de conocer los cambios genómicos que se producen es necesario establecer una relación entre éstos y la expresión génica. En este sentido, en un primer estudio realizado con ARN procedente de líneas celulares limfablastoides de 270 individuos no relacionados (panel HAPMAP [29, 30]) realizado por Stranger et al 2007 [40] concluye que el 18 % de los cambios de expresión detectados se corresponden con CNVs.

Aún conociendo el importante papel de las DSs y CNVs sobre la variabilidad interindividual las tecnologías existentes hoy en día que permiten su cuantificación están en fase de desarrollo.

1.2. Métodos de análisis genómico

1.2.1. Matrices basadas en Hibridación Genómica Comparada (aCGH)

Una técnica desarrollada recientemente llamada aCGH [41, 42, 43] permite identificar regiones genómicas con CNVs mediante un sólo experimento al analizar a la vez miles de loci representados por sondas en una matriz. Estas sondas pueden ser, entre otras, BACs (*Bacterial Artificial Chromosome*, también reciben el nombre de clones) u oligonucleótidos.

La técnica aCGH consiste en comparar, mediante hibridación competitiva, el ADN genómico procedente de una muestra problema contra un ADN de referencia que suele ser un *pool* de controles del mismo género. Cada una de las muestras se marca con un fluorocromo distinto y son hibridadas simultáneamente en un portaobjetos de vidrio que contiene las sondas correspondientes (ver Figura 5). El número de copias relativo entre ambas muestras se determina mediante la valoración de la capacidad de cada muestra a unirse a la misma sonda impresa en el portaobjetos, es decir el cociente o ratio de fluorescencia entre los dos ADNs es indicativo del número de copias relativo. Existen, al menos, dos pasos críticos en la fabricación y análisis de una matriz aCGH:

- La fase de impresión que es el procedimiento mediante el cual se disponen e inmovilizan las sondas en los portaobjetos de vidrio.
- La fase de hibridación que es el procedimiento mediante el cual los ADNs de la muestra problema y la muestra de referencia son marcados con un flurocromo distinto y depositados sobre el portaobjetos. Éstos se incuban durante un tiempo suficiente en condiciones de temperatura y de concentración de sales para permitir la unión específica de cada molécula marcada con su sonda complementaria.

El proceso de hibridación es el paso más importante en la detección de CNVs. Si la hibridación no ha funcionado habrá un gran número de falsos positivos y falsos negativos dando lugar a la no validación por técnicas alternativas suponiendo un gran coste económico y de tiempo o bien a la no detección. Por ello, es necesario desarrollar métodos que permitan conocer la calidad del proceso y la fiabilidad de los datos obtenidos.

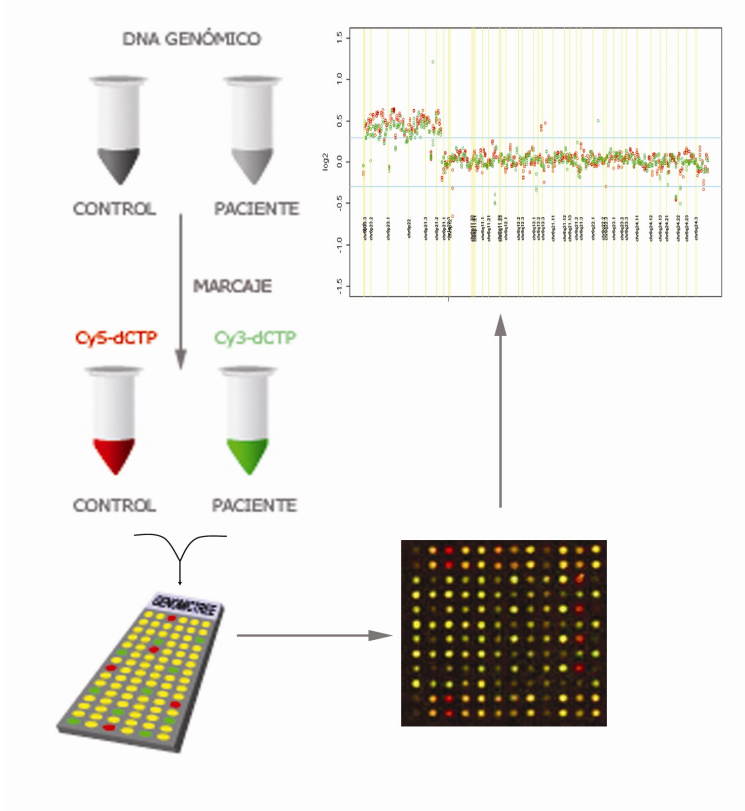

Figura 5: **Proceso de marcaje e hibridación** de las muestras test o paciente y referencia o control en una matriz aCGH hasta obtener una imagen que representa el número relativo de copias.

Esta técnica aún está sujeta a un alto grado de variación que responde a la complejidad de las sondas (en cuanto a secuencia y comportamiento de hibridación en condiciones uniformes que deben ser subóptimas), por lo tanto es difícil determinar cuan inequívocos son los resultados obtenidos y por eso, en muchos casos, es conveniente la validación mediante otras técnicas como FISH (*Fluorescent In Situ Hybridation* mucho menos sensible), MLPA (*Multiple ligation-dependent probe amplification*)o qPCR (*quantitative Polymerase Chain Reaction*). Datos previos propios y de otros grupos indican que el principal problema es la existencia de un gran número de falsos positivos y la posibilidad de que exista también un gran número de falsos negativos. Realizar réplicas técnicas y biológicas reduciría el número de falsos positivos y falsos negativos sin embargo no siempre es posible debido al tiempo necesario para realizar los experimentos (según Drazinic et al 2005 [44], un técnico puede realizar entre 6-12 hibridaciones cada tres días) y debido al coste económico de cada uno de los experimentos (más de 120 € por experimento) sin tener en cuenta el coste de las herramientas informáticas necesarias para el análisis.

Una vez realizado el proceso de hibridación, existen ciertos parámetros que permiten conocer la calidad de los datos obtenidos como, por ejemplo, la representación gráfica MA (donde se representa el ratio de fluorescencia obtenido versus las intensidades medias). Aún cuando pueda asegurarse que todo ha funcionado correctamente, el análisis de los datos es bastante complejo. Las variables de estudio, las cuantificaciones de intensidad de hibridación que se suponen proporcionales a la cantidad de ADN para cada secuencia en las muestras analizadas, son proporcionadas por un programa informático a partir de la

lectura de las imágenes (ver Figura 6).

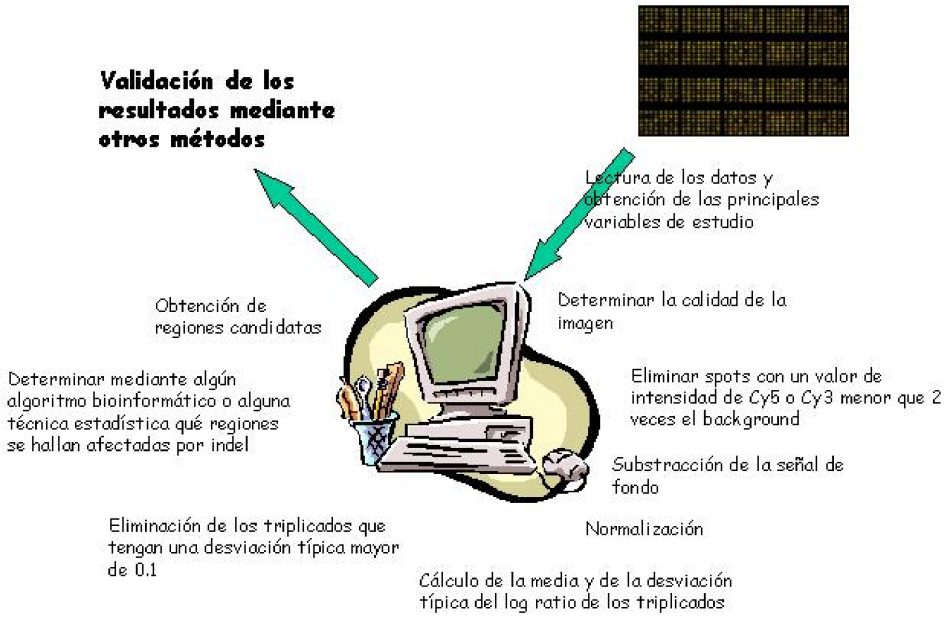

Figura 6: Fases de análisis de las imágenes producidas tras la hibridación en una matriz aCGH. Las sondas (BACs u oligonucleótidos) se hallan inmovilizadas en el portaobjeto de vidrio. Cada una de las regiones del portaobjetos dónde se ha impreso una sonda recibe el nombre de *spot*.

El análisis de los datos incluye el preprocesado de los datos crudos (normalización) y la aplicación de métodos bioinformáticos y/o estadísticos que permiten predecir los cambios de dosis.

El proceso de normalización consiste en ajustar el ratio observado en función de la media geométrica de las intensidades de los dos canales asumiendo que el ratio debe ser igual a 1 para la mayoría de las sondas [45, 46]. El ajuste sobre las intensidades permite minimizar el efecto diferencial debido a la eficiencia de marcaje y de fluorescencia. Este proceso permite la eliminación de artefactos producidos durante la fabricación de la matriz aunque estudios recientes han demostrado que, en el caso de aCGH, no es totalmente efectivo. Muchos autores realizan la substracción del ruido de fondo (BG o *background*) de la señal de los *spots* en esta fase. La substracción del ruido de fondo consiste en medir la cantidad de señal no específica emitida por el portaobjetos alrededor del *spot*. Se considera que el *spot* posee el mismo grado de señal inespecífica y, por lo tanto, se le resta a la señal del *spot*, aunque algunos autores se han manifestado en contra de este procedimiento [47, 48]

El método más sencillo y más utilizado en la literatura para detectar CNVs consiste en establecer unos puntos de corte aplicados sobre el ratio comunes a todas las hibridaciones. Se considera que existe una alteración cuando el valor absoluto de la variable respuesta está por encima del punto de corte elegido.

Sin embargo, actualmente, para determinar con mayor rigor regiones con ganancia o pérdida de material genómico a partir de datos obtenidos por aCGH se están desarrollando numerosos métodos basados en computación intensiva o en técnicas estadísticas conocidas. Como, por ejemplo, los basados en la detección de cambios de tendencia [49, 50, 51, 52, 53], Cadenas de Markov ocultas (HMM) [54, 55], el algoritmo de Smith-Waterman [56], métodos bayesianos [57],... aunque ninguno de ellos parece ser totalmente efectivo. Los mejores resultados han sido obtenidos por los métodos de segmentación [58, 59] y, especialmente, por el método de segmentación binaria circular presentado por Olshen et al 2004 [52]. Este método es muy efectivo cuando se buscan lesiones relativamente grandes sin información previa sobre las sondas.

Aún así, los resultados que se obtienen de estas matrices aCGH son de gran relevancia en investigación clínica [60], [61] con aplicaciones asistenciales casi inmediatas, y por ello cada vez más investigadores trabajan con esta técnica. En la Figura 7 se muestra el incremento del número de publicaciones donde se utilizó esta técnica (datos obtenidos de la revisión bibliográfica realizada en la sección 1.6).

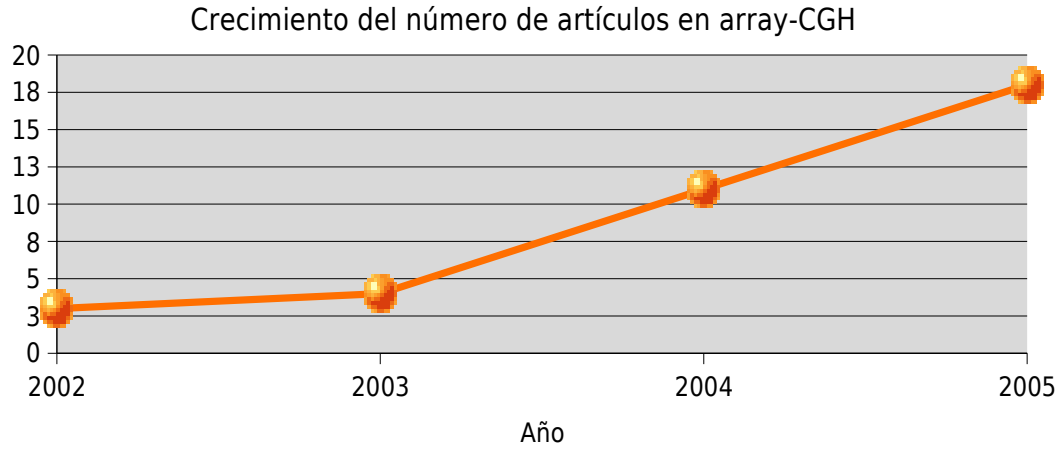

Figura 7: **Crecimiento del número de artículos por año** aparecidos en pubmed entrre los años 2002-2005 sobre aCGH

La estanderización de los métodos de análisis de aCGH permitiría su aplicación de manera homogénea entre diferentes laboratorios así como la utilización compartida de los datos generados para estudios a mayor escala y meta-análisis.

1.2.2. Matrices aCGH basadas en Cromosomas Bacterianos Artificiales (BAC aCGH)

Las sondas impresas en este tipo matriz son insertos de ADN humano que pueden proceder de Cromosomas Bacterianos Artificiales o BACs y tienen una longitud media de

aproximadamente 150 kb. Existen librerías disponibles publicamente de estos BACs como son la librería de CHORI (Children's Hospital Oakland Research Instiute) y la librería del *Sanger Institute*.

Cada uno de los insertos se corresponden con una región del genoma humano concreta. Para realizar matrices de fabricación propia el experimentador realiza un diseño específico de las sondas según sus propósitos y las imprimé una o varias veces sobre el portaobjetos.

1.2.3. Matrices aCGH basadas en oligonucleótidos (oligo aCGH)

Alternativamente las sondas pueden ser oligonucleótidos que tienen una longitud entre 21 y 60-mers. Las matrices oligo aCGH pueden ser diseñadas por el experimentador o bien se usan las sintetizadas in situ por casas comerciales que realizan el diseño de las sondas basándose en el conocimiento existente sobre la composición del genoma humano (i.e. mayor densidad en regiones con alto contenido génico y menor densidad en regiones con polimorfismos en el número de copias) y optimizadas para minimizar posibles problemas durante la hibridación derivados de las características de la secuencia. Del mismo modo, las sondas pueden estar impresas una o varias veces sobre el portaobjetos.

Además para el estudio de SNPs se han desarrollado plataformas específicas tanto en Affymetrix [62] como en Illumina [63, 64] que permiten detectar, a la vez, CNVs.

1.2.4. Otros métodos de cuantificación

Como alternativa y/o para la validación de los datos obtenidos por aCGH, existen otras técnicas para detección de CNVs a menor escala. No existe ninguna técnica alternativa que sea un buen *gold standard*, es decir que confirme con un 100 % de fiabilidad los resultados obtenidos por aCGH. Cada una de las técnicas que se describen a continuación tienen un grado de resolución asociado que, en algunos casos, es bastante distinto de la técnica aCGH y, para MLPA o PCR cuantitativa, se desconoce el porcentaje de falsos positivos y negativos asociados.

Southern blot

Es una de las técnicas más utilizada en los laboratorios de diagnóstico genético. Esta técnica consiste en fragmentar el ADN mediante enzimas de restricción. La existencia de un reordenamiento o aneuploidía puede producir la aparición o desaparición de una diana de restricción provocando la aparición de nuevas bandas. La introducción del gel de electroforesis en campo pulsado (PFGE) incrementó el rango de medidas permitiendo la detección de amplificaciones y deleciones. Además, esta técnica permite la semicuantificación mediante la cuantificación de fluorescencia obtenida hibridando una sonda marcada en relación a la misma cuantificación en locus y muestra normales. Las mayores limitaciones de esta técnica es la baja resolución, baja cobertura y dependencia de la existencia de dianas para los enzimas de restricción [65].

FISH

La hibridación in situ con fluorescencia [66, 67] es un método habitual en citogenética para estudiar aneuploidías completas o regionales. Este método se utiliza para detectar deleciones y amplificaciones que van desde centenares de kb hasta pocas Mb. Para incrementar la resolución del método se pueden emplear sondas específicas de la región de interés marcadas sobre núcleos en interfase. La principal limitación de este método son las derivadas del tamaño de las sondas y la imposibilidad de escanear todo el genoma.

CGH Convencional

La hibridación genómica comparada (CGH) se utiliza para detectar aneuploidías o aneusomías a lo largo de todo el genoma. Esta técnica consiste en comparar un individuo con cariotipo conocido normal contra un individuo de desconocido cariotipo o de conocido cariotipo anómalo que requiere de caracterización. Las muestras de ADN de estos dos individuos son marcados con dos fluorocromos distintos y aplicados sobre una preparación de cromosomas en metafase de un individuo con conocido cariotipo normal. La diferencia de intensidades entre los ADNs marcados es indicativo del número de copias presente en un segmento del genoma [68, 69]. CGH tiene una resolución de 5-10 Mb.

PCR cuantitativa

La técnica PCR (*Polimerase Chain Reaction*) se basa en la amplificación de un fragmento específico de ADN mediante un par de oligonucleótidos cebadores complementarios al segmento de interés. El enzima ADN polimerasa es el encargado de realizar la amplificación debido a su capacidad de fabricar una cadena de ADN complementaria a otra ya existente.

El término PCR cuantativa se refiere a la posibilidad de detectar a tiempo real la amplificación del ADN de interés en la fase exponencial de amplificación dado a que la cantidad de producto detectado es directamente proporcional al número de moléculas iniciales en la muestra. Para la cuantificación directa es necesario introducir un patrón de concentraciones conocidas de la secuencia analizada mientras que para la cuantificación relativa es necesario introducir un gen control de número de copias invariante.

El número de moléculas distintas que pueden ser interrogadas en una misma reacción es limitado por el número de flurocromos disponibles y la capacidad de detección del instrumento. La limitación principal de esta técnica es la necesidad de optimizarla para cada juego de sondas [65].

MLPA

Multiplex ligation-dependent probe amplification (MLPA) fue descrito por primera vez por Schouten et al. 2002 [70]. El método fue diseñado para detectar aneuploidías en un amplio espectro de condiciones mediante la cuantificación relativa de más de 45 sondas de

ADN en una sola reacción. La técnica MLPA consiste en diseñar parejas de oligonucleótidos de distinto tamaño que hibridan en unos puntos específicos adyacentes a la secuencia diana. Una vez han hibridado los dos oligonucleótidos se unen mediante una ligasa y es amplificada mediante una PCR con un cebador marcado. Todas las sondas tienen la mismas secuencias finales, ello permite la amplificación simultánea. De cada sonda se obtiene un producto resultante de la amplificación de un único tamaño entre 130-480 pb. Este tamaño diferencial permite su posterior separación e identificación. La cuantificación relativa de cada producto de PCR es proporcional al número de copias de cada sonda. Los resultados finales son cuantificaciones relativas obtenidas respecto a controles normales. Se ha detectado que la técnica puede fallar cuando existen mutaciones o polimorfismos cercanos al sitio de ligación.

Segregación de microsatélites

Los microsatélites o repeticiones cortas en tándem (STR) son *loci* de ADN polimórfico que contienen secuencias de nucleótidos repetitivas. Las repeticiones suelen estar formadas por 2-7 nucleótidos de longitud. En cada *locus* puede haber un número distinto de repeticiones compuestas por unidades de repetición de la misma longitud. El análisis de microsatélites se lleva a cabo mediante amplificación por PCR utilizando cebadores marcados fluorescentemente. Los productos derivados de la PCR son analizados por electroforesis que realiza una separación por tamaño y, debido a que el número de repeticiones en cada *locus* puede ser distinto, pueden observarse alelos de distintas longitudes. En los estudios de tríos (pacientes y padres) es posible determinar la existencia de una pérdida o ganancia de material genético mediante la detección de una segregación anómala de los alelos materno y paterno [65]. Esta técnica también llevarse a cabo con cualquier otro poliformismo.

1.3. Enfermedades causadas por reordenamientos genómicos

Los reordenamientos genómicos puede causar enfermedades debidos a la pérdida o duplicación de genes sensibles a dosis y/o su región promotora y/o debido a un cambio de la posición de referencia entre gen y su región reguladora (se denominan cambios posicionales). Los reordenamientos genómicos se consideran recurrentes si se ven afectadas las mismas regiones en distintos pacientes y son mediadas por el mismo mecanismo mutacional en regiones genómicas susceptibles. Los reordenamientos no recurrentes son lesiones esporádicas raramente explicables por factores genómicos de susceptibilidad. Actualmente se considera que la reparación por unión de extremos no homólogos contribuiría a producir parte de las alteraciones no recurrentes.

1.3.1. Trastornos genómicos recurrentes

Son un grupo de enfermedades que ocurren en regiones del genoma especialmente inestables y ricas en DSs. Consisten mayoritariamente en deleciones y duplicaciones que afectan, comúnmente, a intervalos genómicos concretos con tamaño idéntico entre pacientes y que ocurren con una frecuencia relativamente elevada. Los trastornos genómicos

recurrentes afectan, globalmente, a aproximadamente 1 de cada 1.000 recién nacidos. En algunos casos, se han detectado polimorfismos estructurales en la región, como inversiones paracéntricas, que son más prevalentes entre progenitores de afectados. Hasta la actualidad, estas inversiones polimórficas se han descrito como alelos de susceptibilidad en progenitores de pacientes con síndrome de Angelman, síndrome de Williams y deleción 17q21 [22, 23, 71].

Las regiones subteloméricas son ricas en DSs (estas regiones comprenden el 0,1 % del genoma y sin embargo contienen el 40 % de todas las DSs) y como consecuencia sufren trastornos genómicos recurrentes con mayor frecuencia. Algunos de ellos se han asociado con retraso mental idiopático [72].

Los reordenamientos causantes de trastornos genómicos recurrentes están mediados por NAHR entre secuencias homólogas (ver Figura 1). Las DSs predisponen a un mal alineamiento cromosómico durante la división celular y causan reordenamientos inter (translocaciones) o intracromosómicos (deleciones, duplicaciones, inversiones, isocromosomas) por recombinación desigual. En la Tabla 2 se detallan los reordenamientos recurrentes más frecuentes.

Aunque las regiones alteradas incluyen decenas de genes, en la mayoría de estos síndromes de microdeleción o microduplicación uno o pocos genes son los responsables de la enfermedad mientras que el resto de genes afectados modifican el fenotipo final. Esta hipótesis ha podido ser contrastada en los síndromes mediados por NAHR más frecuentes como son el Síndrome de Angelman / Prader-Willi, el síndrome de Smith-Magenis, el síndrome de DiGeorge y el síndrome de Williams-Beuren. En la Figura 8 se muestran pacientes afectados por estos síndromes.

El **síndrome de Angelman** (AS, OMIM 105830) y el **síndrome de Prader-Willi** (PWS, OMIM 176270) están causados, principalmente (70 % de los casos), por una deleción en la banda cromosómica 15q11.2-q13 con un tamaño de 4,5 Mb. El origen parental del cromosoma afectado determina el síndrome fenotípico del paciente que está asociado a la impronta genómica [73, 74]. La prevalencia de estos síndromes se sitúa alrededor de 1 cada 10.000 nacidos vivos. El gen UBE3A es el responsable del fenotipo en AS mientras que en PWS hay varios genes. AS se caracteriza por la presencia de retraso mental y motor, ataxia, hipotonia, epilepsia y problemas severos de comunicación oral. Los rasgos faciales más característicos de estos pacientes es la presencia de una mandíbula alargada y una expresión con la boca abierta que permite ver la lengua (ver Figura 8) mientras que los rasgos más característicos en PWS son una disminución de la actividad fetal, obesidad, hipotonia muscular, retraso mental, baja talla, hipogonadismo y manos y pies pequeñas (ver Figura 8).

El **síndrome de Smith-Magenis** (SMS, OMIM 182290) está causado, en la mayor parte de los casos, por una deleción de 3,7 Mb situada en la banda cromosómica 17p11.2. SMS afecta a 1 de cada 25.000 nacidos vivos. El gen responsable del fenotipo es RAI1[75]. Este síndrome se caracteriza por la presencia de problemas en el lenguaje y en

el comportamiento así como hiperactividad y retraso mental moderado. Algunos pacientes presentan rasgos craneofaciales y braquidactilia.

El **síndrome de DiGeorge** (DiG, OMIM 188400) es el síndrome de microdeleción más frecuente ya que afecta a 1 de cada 4.000 nacidos vivos. En la mayoría de los casos está recombinación produce una deleción de 3 Mb en la banda 22q11.2. El gen responsable de la mayor parte de los fenotipos es TBX1 [76]. Los rasgos fenotípicos más característicos comprenden hipocalcemia, hipoplasia paratiroidea, defectos cardíacos y migración de la cresta neuronal.

Figura 8: **Pacientes afectados por síndromes mediados por NAHR**. Arriba y de izquierda a derecha se muestran pacientes afectados por Síndrome de Angelman, Síndrome de Prader-Willi y Síndrome de Smith-Magenis. Abajo y de izquierda a derecha se muestran pacientes afectados por Síndrome de DiGeorge y Síndrome de Williams-Beuren.

El **síndrome de Williams-Beurem** (WBS, OMIM 194050) está causado por una deleción en la banda cromosómica 7q11.23 de 1,55 Mb de longitud. WBS es un trastorno del desarrollo que tiene una prevalencia estimada de 1 en cada 7.500 nacidos vivos. Su fenotipo se caracteriza por rasgos faciales específicos, retraso mental de moderado a leve y un déficit cognitivo que englobla problemas en algunas áreas como psicomotricidad e integración visual y una relativa preservación de otras como el lenguaje y la musicalidad. Poseen una personalidad amigable, ocasionalmente presentan hipercalcemia en la infancia y vasculopatía con estenosis supravalvular aórtica [77].

A pesar de haber una clara correlación entre diagnóstico clínico y la presencia de la deleción en hemicigosis en 7q11.23, se desconoce la contribución de cada gen afectado (entre 26 y 28 genes aproximadamente) al fenotipo final del paciente. Actualmente, las

líneas de investigación abiertas en el estudio de WBS tienen el objetivo de conocer la contribución exacta de cada gen a la patogénesis y fisiopatología para identificar métodos terapeúticos.

Tabla 2: Enfermedades recurrentes causadas por NAHR entre DSs

Síndrome	OMIM	Tipo de reordenamiento cromosómico	Locus	Tamaño (Mb)	Ref
Monosomía 1p36	607872	Deleción	1p36	10,5	[79] [80]
del(1)(q21)		Deleción	1q21	3	[81]
Nefronoptisis familiar juvenil	256100	Deleción	2q13	0,29	[82]
del(3q29)	609425	Deleción	3q29	1,5	[83]
Sotos	117550	Deleción	5q35	2,2	[84]
Williams-Beuren	194050	Deleción	7q11.23	1,55	[77]
dup(7)(q11.23;q11.23)	609757	Duplicación	7q11.23	1,55	[78]
del(8p23)	600576	Deleción	8p23.1	3	[85]
Prader-Willi	176270	Deleción	15q11-q13	4	[74]
Angelman	105830	Deleción	15q11-q13	4	[73]
dup(15)(q11;q13)/Autismo	608636	Duplicación	15q11-q13	4	[86, 87]
del(15)(q13.3)		Deleción	15q13.3	1,5	[88]
del(15)(q24)		Deleción	15q24	3,7	[89]
Smith-Magenis	182290	Deleción	17p11.2	5	[75]
dup(17)(p11.2;p11.2)	610883	Duplicación	17p11.2	5	[90]
Neurofibromatosis tipo I	601097	Deleción	17p11.2	1,5	[91]
Charcot-Marie-Tooth Tipo 1A	118220	Duplicación	17p11.2	1,5	[92]
Neuropatía hereditaria con parálisis por presión	162500	Deleción	17p12	1,5	[93]
del(17)(q21)	610443	Deleción	17q21.31	< 1	[71]
dup(22)(q11)	608363	Duplicación	22q11 3	3	[94]
DiGeorge/Velocardiofacial	188400	Deleción	22q11.2	3	[76]
Ojos de gato	115470	Cromosoma marcador supernumerario	22q11.2	3	[95]
Deficiencia en esteroide sulfatasa	308100	Deleción	Xp22.32	0,4	[96]

1.3.2. Enfermedades de etiología desconocida candidatas a estar mediadas por reordenamientos crípticos

El avance tecnológico que ha acontecido en los últimos años permite detectar con mayor grado de resolución microdeleciones y microduplicaciones. Ello hace que enfermedades de etiología desconocida con cariotipo normal como retraso mental, autismo u otros síndromes polimalformativos entre los que se encuentra el síndrome de Kabuki sean susceptibles de ser estudiadas mediante estas nuevas técnicas. Los primeros estudios realizados han puesto de manifiesto la existencia de más cuadros causados por deleciones recurrentes en regiones genómicas inestables ricas en DSs. Aunque este tipo de técnicas permiten detectar, también, reordenamientos no recurrentes involucrados en cambios de dosis de ADN.

Retraso Mental idiopático

El retraso mental es un trastorno común que padece entre 1-3 % de la población general [97] y que afecta a hombres y a mujeres en una proporción desigual, especialmente en los grados de moderado y severo (IQ < 50), siendo los hombres los más afectados. Ello puede ser indicativo de que muchos de los genes que pueden estar asociados a retraso mental se hallan en el cromosoma X.

Cuando existe únicamente un trastorno cognitivo como manifestación de la enfermedad se denomina retraso mental idiopático. Si existen otras manifestaciones clínicas se denomina retraso mental sindrómico. La etiología del retraso mental es todavía desconocida aunque se sabe que muchos casos son debidos a alteraciones de tipo genético mientras que otros son debidos a factores ambientales.

Estudios recientes han demostrado que reordenamientos teloméricos son la causa de entre un 5-7 % de todos los casos de retraso mental [98]. Los trastornos genómicos recurrentes que se asocian con retraso mental sindrómico ya se han descrito en el apartado anterior.

Autismo

El autismo o el trastorno del espectro autista (ASD, OMIM 209850) es una enfermedad común con una prevalencia, que varía según el método utilitzado para calcularla, entre 3,3 y 16 casos cada 10.000 nacidos vivos en la *Unión Europea*. El autismo es un trastorno neurológico con una heredabilidad superior al 90 % [99] pero que comprende diversas etiologías. Por ello, durante años se le ha considerado como un conjunto de enfermedades distintas con un fenotipo común caracterizado por comportamiento social atípico, problemas en el habla y en la comunicación no verbal, patrones inusuales de intereses muy restringidos y conductas repetitivas. Recientes avances tecnológicos en el área de la anatomía y función cerebral han demostrado la presencia de un patrón común en los trastornos del espectro autista que se basa en una desconexión parcial de ciertas áreas del cerebro con el lóbulo frontal que se produce durante desarrollo [100].

La existencia de una base genética en los trastornos del espectro autista viene, también, reforzada por una mayor prevelancia de casos de ASD en enfermedades neurológicas de etiología conocida como son el síndrome de X frágil, síndrome de Rett, síndrome de Smith-Opitz-Lemi y síndrome de Down.

Se considera que deben haber al menos 15 genes involudrados en ASD lo que conferiría el alto grado de heterogeneidad observado [101].

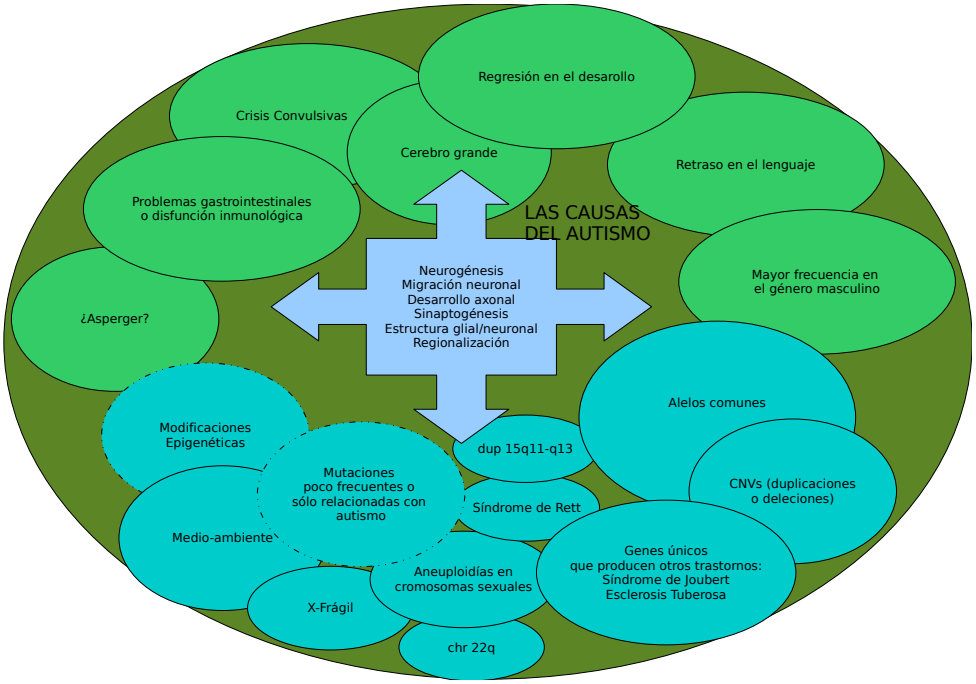

Figura 9: **Fenotipo asociado con ASD y sus posibles causas**. En esta Figura adapatada de [100] se representan en verde los rasgos fenotípicos más característicos en ASD (el tamaño del globo indica su frecuencia) y en azul las posibles causas de éstos rasgos.

Se han descrito CNVs de novo en distintas regiones del genoma en pacientes con ASD y, globalmente, se estima que entre el 10 %-15 % de los casos ASD pueden ser debidos a reordenamientos cromosómicos, alguno de ellos recurrentes com la dup15q.11-q13 materna o la deleción 16p11.2 descrita recientemente ([102]).

Síndrome de Kabuki

El síndrome de Kabuki (KS, OMIM 147920) es un trastorno congénito establecido de manera independiente por Niikawa [103] y Kuroki [104] con una incidencia estimada de 1 cada 32.000 nacidos vivos. KS se caracteriza por rasgos faciales específicos que incluyen fisura palpebral mongoloide, cejas arqueadas muy finas, columela corta con la punta de la nariz aplanada y orejas prominentes. Los afectados por KS también presentan retraso mental entre moderado y leve, retraso en el crecimiento post-natal con una baja estatura al final del crecimiento, alteraciones esqueléticas y patrones dermatológicos inusuales con clinodactilia en el quinto dedo.

La mayoría de casos descritos son esporádicos y en la misma proporción entre los dos géneros. Aunque la causa de KS es desconocida, se han descrito numerosas alteraciones citogenéticas. Matsumoto and Niikawa [105] especulaban con la existencia de una posible microdeleción o microduplicación que involucrara diversos genes contiguos como causa del desorden, dado que los pacientes con KS muestran un amplio espectro de alteraciones clínicas multisistémicas y que la mayoría de los casos son esporádicos.

Milunski et al 2003 [106] realizaron hibridación genómica comparada (CGH) en portaobjetos cromosómicos y encontraron una duplicación de 3,5 Mb en 8p23.1-p22 en 6

pacientes no relacionados sugiriendo que esta alteración podría ser la causa del desorden pero está alteración no ha sido encontrada en otros pacientes indicando que este reordenamiento no es común a todos los KS [107].

1.4. CNVs y susceptibilidad a enfermedades

Aunque la mayor parte de las CNVs y variantes estructurales presentes en algunas regiones genómicas no tienen una consecuencia fenotípica evidente, aquellas que tengan como consecuencia cambios de dosis génica pueden causar enfermedades genéticas o susceptibilidad a enfermedades complejas ya sean solas o en combinación con otras o con factores medioambientales. En general, existe una correlación entre la dosis génica y la expresión relativa de los genes [38], ver Tabla 3.

Las variantes estructurales pueden afectar directamente la expresión génica (ver Figura 10):

1. **Modificando directamente la dosis génica**: Los cambios de dosis génica pueden afectar directamente a la expresión. Y aunque estos genes no estén directamente relacionados con una enfermedad determinada pueden afectar o alterar la susceptibilidad o riesgo a padecer una enfermedad de caracter multifactorial. En este sentido se han realizado, desde el descubrimiento de las CNVs, numerosos estudios dirigidos con genes candidatos obteniendo un gran éxito. En la Tabla 3 se muestran algunos casos donde el riesgo a padecer una enfermedad se ve alterado mediante cambios de dosis producidos por CNVs. En esta tabla se resumen los estudios publicados a fecha de diciembre del 2007.

2. **Mediante un efecto posicional**: Más allá de los efectos sobre la expresión génica causados por un cambio de dosis deben ser tenidos en cuenta aquellos mecanismos en que la expresión de un gen puede cambiar debido a un cambio posicional respecto a sus estructuras reguladoras que pueden estar hasta a 1 Mb de distancia [13]. Un ejemplo contrastado es una traslocación que causa una disrupción del gen HDAC9 en la banda 7p21.1 y su recíproco punto de rotura en el cromosoma 1, aproximadamente a unas 500 kb de distancia del gen TGFB2. Los pacientes que tienen está traslocación padecen la anomalía de Peter que produce un defecto en la cámara anterior del ojo. Un ratón *knockout* (ratón transgénico obtenido mediante la eliminación de uno o pocos genes) para el gen TGFB2 posee la misma alteración por lo que la causa de la patología está asociada al gen TGFB2 más que a la disrupción del gen HDAC9 [108].

Tabla 3: Alteraciones del riesgo a padecer enfermedades multifactoriales mediadas por CNVs

Gen	Locus	Tamaño (kb)	Fenotipo	Tipo	Efecto	Mecanismo	Copias	Ref
RHD	1p36.11	60	Sensibilidad al RH	?	Dosis	Deleción	0-2	[109]
FCGR3	1q23.3	> 5	Glomeronefritis, lupus sistémico erimatoso	Riesgo	Dosis	Deleción	0-14	[35]
UGT2B17	4q13	150	Cáncer de próstata	Riesgo	Dosis	Deleción	0-2	[31, 32, 33, 34]
GSK3B	3q13.3	?	Desorden bipolar	Riesgo	Posicional	Duplicación	1-3	[36]
SNCA	4p15	>1,600	Parkinson	Riesgo	Dosis	Duplicación	2-3	[110]
C4A/ C4B	6p21.1/ 6p22.3	∼100	Lupus Sistémico erimatoso	Riesgo	Posicional	Deleción	0-5/ 0-4	[111]
LPA	6q25.3	5,5	Alteraciones coronarias	Riesgo	Dosis	Deleción	2-38	[112]
MYB	6	500	Leucemia limfoblastoide aguda	Riesgo	Dosis	Duplicación	2-3	[113]
PRSS1	7	605	Pancreatitis heriditaria	Riesgo	Dosis	Duplicación	2-3	[114]
MTUS1 (exón 4)	8p21.3	∼1,2	Cáncer de mama heriditario	Protector	Posicional	Deleción	0-2	[115]
DEFB4	8p23.1	20	Enfermedad de Crohn	Riesgo	Dosis	Deleción	2-10	[116]
DEFA1/ DEFA3	8p23.1	240	Susceptibilidad a infecciones	Riesgo	Dosis	Deleción	2-12	[117]
CCL3L1/ CCL4L1	17q12	?	Infección HIV	Riesgo	Dosis	Deleción	0-14	[87]
CCL3L1	17q12	?	Artritis reumatoide, Diabetes tipo I	Riesgo	Dosis	Duplicación	0-14	[119]
CYP2A6	19q13.2	?	Alteración del metabolismo de nicotina	Riesgo	Dosis	Duplicación	1-5	[120]
APP	21	500	Alzheimer	Riesgo	Dosis	Duplicación	2-3	[37]
OPN1LW/ OPN1MW	Xq28	∼14	Ceguera al rojo/azul	Riesgo	Dosis	Deleción	0-4/ 0-7	[121]
CFHR1/ CFHR3	1q23	85	Degeneración macular relacionada con la edad	Protector	Dosis	Deleción	0-2	[122]

?; El tamaño de la CNV encontrada en estos casos es deconocida, el estudio cubre sólo los genes candidatos.

Las modificaciones fenotípicas derivadas de estos efectos pueden clasificarse en:

1. **Variación a la penetrancia de un rasgo**: Existen enfermedades con un grado variable de penetrancia (entendiendo como penetrancia la fracción de los individuos con un genotipo determinado que presentan el mismo fenotipo) o expresividad (entendiendo como expresividad el grado de afectación presentado por los pacientes con el mismo genotipo). Estos cambios en la penetrancia y/o expresividad han sido observados en distintas enfermedades como consecuencia de CNVs. Entre ellas se

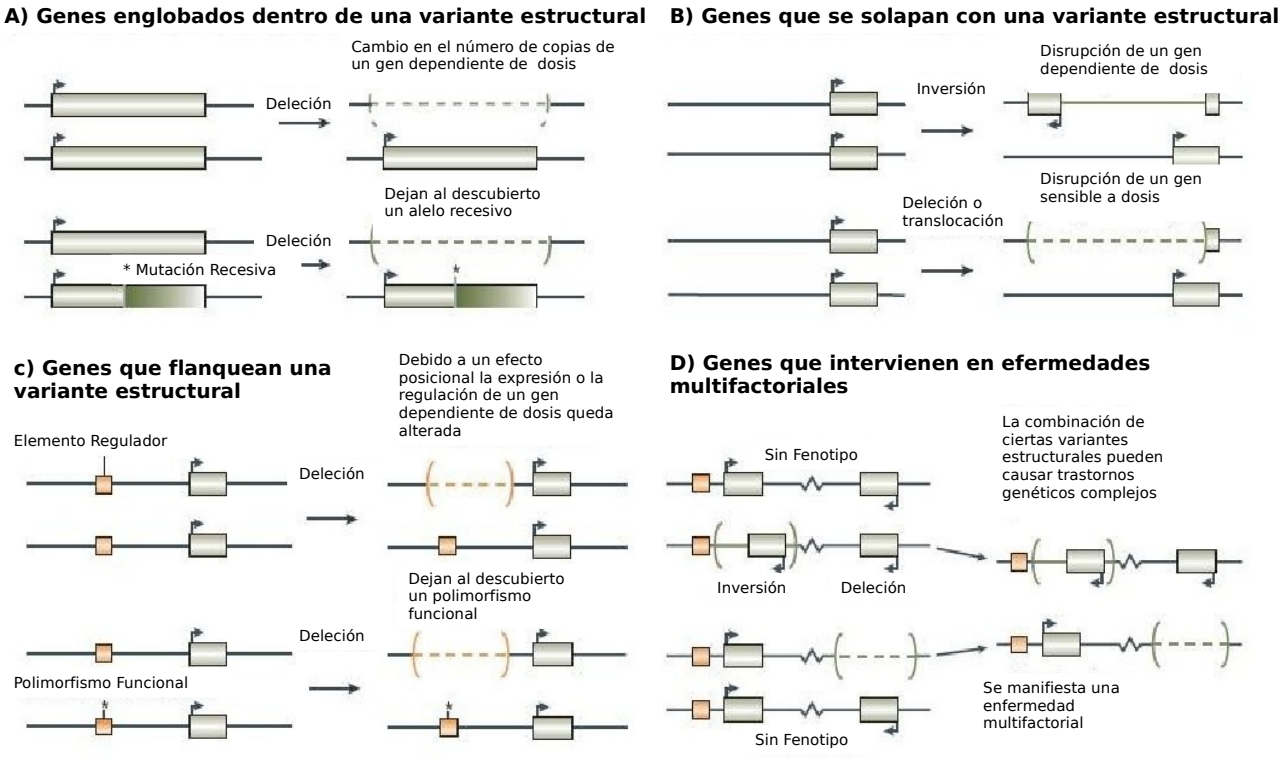

Figura 10: **Relación entre CNVs, fenotipo y enfermedad** A) cambios de dosis géncia por deleción pueden causar enfermedad directamente o bien por que la deleción deja al descubierto una enfermedad de tipo recesivo. B)Inversiones pueden truncar un gen directamente. Deleciones y translocaciones también pueden producir pérdidas de parte de un gen. C) Las variantes estructurales o CNVs pueden afectar a la expresión de genes distales. Este fenómeno se conoce como efecto posicional. Algunas veces estas variantes estructurales pueden producir la sobre expresión de algunos genes, D) CNVs y variantes estructurales pueden producir efectos protectores o de riesgo en enfermedades multifactoriales. Figura adaptada de Feuk et al 2006 [13]

encuentran el síndrome de DiGeorge y su recíproca duplicación que producen cambios en la expresividad. Otro ejemplo es el caso de una variante aminoacídica en el factor del complemento H y en la membrana del cofactor (CFH) que predispone a degeneración macular relacionada con la edad. El riesgo ocasionado por un cambio aminoacídico es modificado por una variación en el número de copias de la región que contiene gen CFH y que incluye también los genes CFHR1 y CFHR3. De hecho Hughes *et al* [122] demostraron que deleciones en los genes CFHR1 y CFHR3 protege contra la degeneración macular.

2. **Modificación de los rasgos fenotípicos en síndromes causados por aneuploidías**: Los genes y los elementos directamente relacionados con varios de los fenotipos observados en trisomías comunes son bastante desconocidos. Se han descrito CNVs que comprenden 3,5 Mb del cromosoma 21, 10,3 Mb del cromosoma 13 y 6,5 Mb del cromosoma 18 en individuos sanos sin causar los fenotipos asociados a las trisomías de estos cromosomas. Estos cambios de dosis génica podrían encontrarse en los pacientes con estas trisomías y podrían ser los responsables de la gran diversidad de fenotipos observados en este tipo de pacientes [38].

3. **Reversión de los efectos causados por mutaciones**: CNVs que contienen genes asociados a enfermedades mediadas por mutaciones puntuales pueden revertir estos efectos debido a la presencia de copias sanas en CNVs [38].

Otros roles de las CNVs

La presencia de CNVs pueden producir susceptibilidad a nuevas mutaciones en el otro alelo y la heterocigosiad para CNVs predispone al mal alineamiento cromosómico y a nuevos reordenamientos [123].

Categorías específicas de genes parecen estar sobrerepresentados en CNVs entre las que se incluyen aquellos genes que interaccionan directamente con el medioambiente como por ejemplo los receptores olfatorios y los genes responsables de desencadenar una respuesta a estímulos externos (i.e. GSTT1 y GSTM1, el citocromo P450 y el componente del complemento C4). Mientras que otras son detectadas con una frecuencia inferior al 1 % de la población general.

Por todo ello, conocer la relación entre los genes incluidas en éstas y las enfermedades permitiría definir nuevas vías fisiopatológicas. Además, el efecto real de las CNVs sobre las diversas enfermedades multifactoriales conocidas depende del conocimiento exacto del número de copias relativo presente en una muestra. Y, en general, es difícil asociar un cambio de dosis en el ADN con la expresión génica de un gen concreto debido, en muchos casos, a la falta del tejido o del estado de desarrollo apropiado.

1.5. PSVs y su relación con enfermedades

Las regiones con DSs son puntos calientes para NAHR causante tanto de reordenamientos cromosómicos (intercambio recíproco) como de fenómenos de conversión génica.

Se considera que en el genoma humano existen entre 25.000-35.000 genes funcionales (secuencia de ADN que tiene la capacidad de transcribirse) de los cuales un 5 % se sitúan sobre DSs. Las copias adicionales de algunos genes en DSs reciben el nombre de pseudogenes cuando no son funcionales. Durante el mecanismo de conversión génica existe una transferencia unidireccional de material génico entre la secuencia donadora (que puede ser el gen o el pseudogen) hacia su homólogo llamado aceptor. Si durante esta recombinación un segmento con modificaciones no sinónimas presentes en el pseudogen acaba en el gen este puede volverse afuncional provocando una enfermedad o viceversa, es decir, que un pseudogen puede volverse funcional siendo la causa de la enfermedad.

Debido a que no existe modificación en el número de copias, los fenómenos de conversión génica no son detectables por los sistemas experimentales que se utilizan para CNVs basados en aCGH y se desconoce cual es su ocurrencia real a nivel genómico.

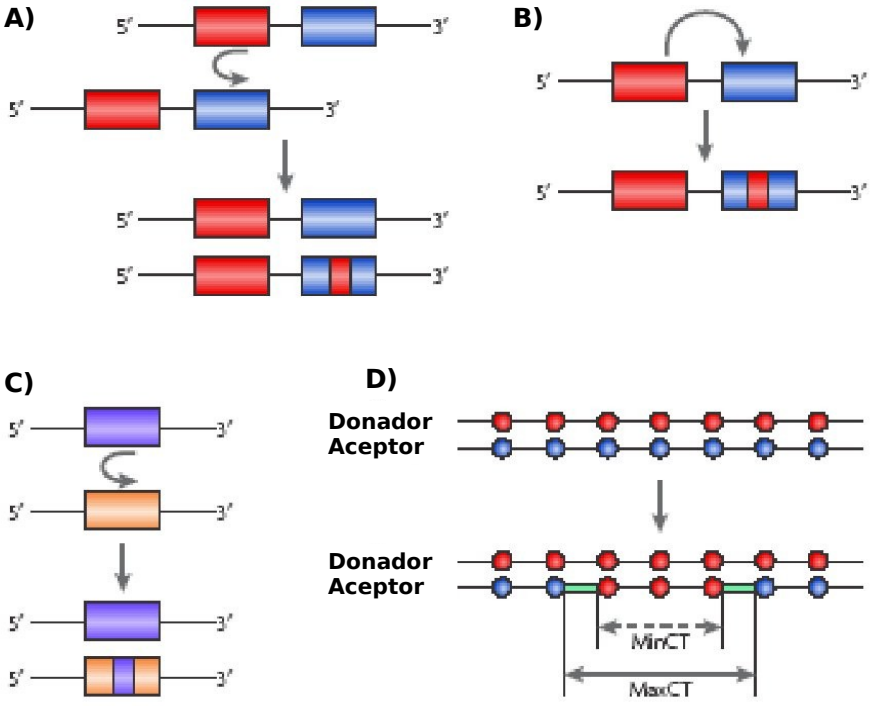

Figura 11: **Conversión génica** A) Conversión génica no-alélica en trans muestra como secuencias parálogas que se hallan en distintas cromátidas o cromosomas pueden recombinarse entre sí (interloci); B) conversión génica no-alélica en cis entre secuencias parálogas que se hallan en la misma cromátida; C) conversión génica interalélica entre cromosomas homólogos; D) la región potencialmente susceptible a recombinación génica está marcada en verde y minCT y maxCT muestran el rango mínimo y máximo respectivamente. Figura adaptada de Chen et al 2007 [124]

Cuando la transferencia por conversión ocurre de pseudogen no funcional a gen, puede aparecer una mutación patogénica por conversión responsable de la enfermedad. Si la

conversión es en dirección contraria (de gen a pseudogen) puede haber una ganancia de función.

Las mutaciones más comunes asociadas a enfermedades son:

- Cambios no sinónimos de una o de pocas pb.
- Desaparición (o aparición) de una zona de *splicing*.
- Inserciones/deleciones que modifican la pauta de lectura.
- Aparición/desaparición de secuencias silenciadoras de *splicing*.

En la Tabla 4 se resumen varias enfermedades causadas por fenómenos de conversión génica.

Tabla 4: Enfermedades causadas por conversión génica

Síndrome	Gen Donador	Gen Aceptor	Locus	Ref
Gaucher	GBAP	GBA	1q21	[125, 126, 127, 128]
Hemólisis urémica atípica	CFHR1	CFH	1q32	[129]
Atrofia muscular espinal	SMN2	SMN1	5q13.2	[130]
Hiperplasia adrenal congénita	CYP21A1P	CYP21A2	6p21.3	[131]
Shwachaman Bodian Diamond	SBDSP	SBDS	7q11.22	[132]
Granulomatoso crónico	NCF1B o NCF1C	NCF1	7q11.23	[133]
Pancreatitis crónica	PRSS2	PRSS1	7q35	[134]
Deficiencia en corticosterona metiloxidasa II	CYP11B1	CYP11B2	8q21-q22	[135]
Microcitosis moderada	HBB	HBD	11p15.5	[136]
Persistencia de hemoglobina fetal heriditaria	HBG2	HBG1	11p15.5	[137]
Defectos en el tubo neuronal	FOLR1P	FOLR1	11q13.3-q14.1	[138]
Policistocis autosómica dominante	?	PKD1	16p13.3	[139, 140]
Baja estatura	GH2	GH1	17q22-q24	[141]
Cataratas autosómica dominante	CRYBP1	CRYBB2	22q11.2-q12.1	[142, 143]
von Willwbrand	VWFP	VWF	22q11.22-11.23	[144]
Agamaglobulinemia crónica granulomatosa	IGLL3	IGLL1	22q11.23	[145, 146]
Ceguera al azul	OPNIMW	OPN1LW	Xq28	[147]

?; desconocido

1.6. Estado al inicio del proyecto. Revisión bibliográfica en PUBMED de estudios previos

Hace años que se desarrollan técnicas que permiten analizar con fiabilidad pérdidas y/o ganancias de material genético. Pero la mayoría de estas técnicas se han especializado en detectar grandes cambios. Durante el año 2005 se demostró que la técnica aCGH permitía detectar microganancias y micropérdidas de material genético ([148]) relacionadas con enfermedades o bien polimorfismos poblacionales, pero con elevado número de falsos positivos y, seguramente, falsos negativos.

El rápido incremento de la utilización de esta técnica para relacionar anomalías cromosómicas crípticas con enfermedades previamente diagnosticadas hizo conveniente revisar la bibliografía existente (ver gráfico 1) al objeto de:

- Conocer cuales son las condiciones experimentales más utilizadas en la realización de un aCGH.
- Conocer cuales son las variables, estadísticos y/o técnicas estadísticas empleadas en el análisis. de datos después del filtrado y normalización de los datos crudos obtenidos.
- Conocer los valores aproximados de las variables y/o estadísticos descriptivos .
- Averiguar si es posible comparar los resultados obtenidos por distintos grupos.

Se incluyeron aquellos artículos que estudiaban muestras humanas no relacionados con cáncer donde se utilizaba la técnica aCGH para el diagnóstico de enfermedades conocidas o para describir las causas genéticas de enfermedades de origen desconocido o para hallar polimorfismos poblacionales de pérdidas y/o ganancias de material genético.
La revisión se realizó con la finalidad de conocer si nuestros propios resultados, que fueron obtenidos de un estudio piloto, eran comparables con los que se habían publicado.

Se revisaron los 15 artículos disponibles en la literatura sobre la temática de interés de entre todos los identificados por la búsqueda.

Condiciones de impresión

Los portaobjetos de tipo Ultragaps de Corning (N=7 estudios) o Sigma (N=3) fueron los más utilizados.
Las soluciones de impresión más utilizadas fueron SSC (N=10) y DMSO (N=3) en distintas concentraciones (en dos casos al 50 % y en un caso al 80 %).

El resto de experimentadores enviaron sus muestras a casas comerciales donde se realizaron los experimentos.

Las concentraciones de ADN durante la fase de impresión variaba entre 100 ng/μl de ADN y 2.000 ng/μl de ADN. Según Schoumans *et al* [36], cuando se trabaja con concentraciones superiores a 800 ng/μl de ADN impreso ello produce saturación e implica

obtener falsos negativos y artefactos en la imagen.

No se ha hallado una relación clara entre tipo de porta, solución de impresión y concentración de ADN.

Condiciones de la hibridación

El marcaje de las muestras de realizó en todos los casos con Cy5-Cy3 y se utilizó Cot-1 para bloquear secuencias altamente repetitivas como Alu, LINES, SINES,...

En diez de los estudios se realizaron réplicas del experimento, y en cinco casos, las réplicas consistieron en marcar los ADNs de modo inverso al primer experimento, es decir, giraron los flurocromos en el subsiguiente experimento.

En cinco casos se hibridó la muestra problema contra un *pool* de referencia, en siete casos se hibridó contra una única muestra y en tres casos se desconoce. En cinco casos la referencia fue del género contrario, en cuatro casos se hibridó contra el mismo género y en seis casos se desconoce.

Descripción de las matrices aCGH utilizadas

Las matrices aCGH utilizadas dependen de los objetivos del estudio. La mayor parte de ellas se han fabricado de manera casera.

Descripción de las muestras hibridadas

En 14 artículos las muestras hibridadas fueron pacientes con distintas patologías germinales y en un caso se hibridaron muestras de abortos espontáneos.

Preprocesado de los datos y análisis

Las imágenes fueron obtenidas por diferentes programas comerciales; *GenePix* en ocho casos, *SPOT* en cinco, *GeneSensor Reading* en un caso y en un caso no aparecía la información.

En ocho estudios se realizó un normalización distinta a la que realizan por defecto los programas de lectura de imágenes.

En seis casos se realizó substracción del ruido de fondo. Las señales obtenidas por los *spots* se consideran válidas si existe una buena circularidad (un caso) y reducida variabilidad entre réplicas de la misma sonda presentes en el portaobjetos (cuatro casos).

La variable respuesta fue, en nueve casos, el logaritmo del ratio entre las intensidades de los dos canales en base 2 y, en el resto, el ratio entre las intensidades de los dos canales.

Tabla 5: **Lista de artículos revisados**

AUTOR	AÑO	REVISTA	TITULO	N BACs	Chr	Min	N	Ref
Rauen KA	2002	Am J Med Genet	Additional patient with del(12)(q21.2q22): further evidence for a candidate region for cardio facio cutaneuos syndrome	2464	Todos	0,8Mb	2	[101]
Veltman JA	2002	Am J Hum Genet	High-throughput analysis of subtelomeric chromosome rearrangements by use of array comparative genomic hybridization	80	Todos	?	20	[60]
Erickson RP	2003	Am J Med Genet A	Does chromosome 22 have anything to do with sex determination: further studies on a 46XX,22q11.2 male	329	22	?	3	[100]
Veltman JA	2003	Am J Hum Genet	Definition of a critical region on chromosome 18 for congenital aural atresia by array CGH	114	18	<4Mb	20	[149]
Goto T	2004	Congenit Anom (Kyoto)	Large fontanelles are a shared feature of haploinsufficiency of RUNX2 and its co-activador CBFB	2500	Todos	1,2Mb	1	[33]
Schaeffer AJ	2004	Am J Hum Genet	Comparative genomic hybridization-array analysis enhances the detection of aneuploidies and submicroscopic imbalace in spontaneuos miscarriages	287	Todos	?	41	[38]
Shaw CJ	2004	J Med Genet	Comparative genomic hybridisation using a proximal 17p BAC/PAC array rearrangements responsible for four genomic disorders	?	Específicos	0,8Mb	25	[55]
Shaw-Smith C	2004	J Med Genet	Microarray based comparative genomic hybridisation (array-CGH) detects patients with learning disability	?	Todos	?	50	[39]
Bejjani BA	2005	Am J Med Genet A	Use of targeted array-based CGH for the clinical diagnosis of chromosomal imbalance:is less more?	589	Todos	< 0,8Mb	86	[31]
Schoumans J	2005	Eur J Hum Genet	Genome-wide screening using array-CGH does not reveal microdeletion / microduplication inn children with kabuki syndrome	2600	Todos	?	10	[36]
Knijnenburg J	2005	Am J Med Genet A	Insights from genomic microarrays into structural chromosome rearrangements	3500	Todos	6,3Mb	4	[34]
Locke DP	2004	J Med Genet	BAC microarray analysis of 15q11-q13 rearrangements and the impact of segmental duplications	18	15q11-q13	?	19	[40]
Prescott K	2005	Hum Genet	A novel 5q11.2 deletion detected by microarray comparative genomic hybridisation in a child referred as a case of suspected 22q11 deletion syndrome	?	Todos/22	5Mb	6	[35]
Ren H	2005	Hum Mutat	BAC-based PCR fragment microarray: high-resolution detection of chromosomal deletion and duplication breakpoints	177	Específicos	¡0,1Mb	2	[18]
Zhang X	2005	Am J Hum Genet	High-resolution mapping of genotype-phenotype relationships in cri du chat syndrome using array comparative hybridization	1750-2000	Todos	1,2Mb	94	[150]

?; dato no proporcionado

Los valores esperados para el logaritmo en base dos del ratio de las intensidades son: 0 cuando no hay lesión, -1 para deleciones en heterocigosis y 0,5 para ganancias en heterocigosis. En los artículos que utilizaron el logaritmo en base dos emplearon puntos de corte para la deteción de deleciones en heterocigosis entre -0,3 y -0,6, para ganancias en heterocigosis entre 0,3 y 0,6. En los artículos que se utilizó el ratio se consideró deleción en heterocigosis entre 0,6 y 0,8 y amplificación en heterocigosis entre 1,28 y 1,4.

Las desviaciones típicas obtenidas por cada portaobjetos oscilaban entre 0,1 y 0,2.

Sólo en tres artículos utilizan algún método estadístico o bioinformático para detectar regiones con ganancias y/o pérdidas.

En 13 artículos realizan una posterior validación y en dos casos se desconoce.

Pocos artículos dan datos de sensibilidad y especificidad en su estudio pero se cita que en regiones donde las aneusomías parciales constan de varias Mb el aCGH tiene una sensibilidad superior al 90 %.

Algunos de los artículos han estudiado casos de mosaicismos con aCGH. Los resultados revelan que valores intermedios de logaritmo en base dos pueden estar asociados a mosaicismo y puede detectarse si la anomalía se presenta en almenos un 50 % de las células [149, 151].

Los BACs que contienen DSs presentan una gran variabilidad y la interpretación de estos datos es difícil [151].

2. Objetivos

1. Conocer cuales son las fuentes de variación más importantes en aCGH así como sus posibles causas con la finalidad de desarrollar métodos que permitan evaluar la calidad de los datos obtenidos por aCGH antes de su validación mediante otras técnicas.

2. Revisar y desarrollar métodos estadísticos y/o bioinformáticos que permitan determinar con un alto grado de fiabilidad la localización y tamaño de las CNVs detectables por aCGH.

3. Correlación de los datos obtenidos por aCGH (alteración en el número de copias de ADN) y aExpr (niveles de expresión génica) para la identificación de vías de regulación afectadas.

4. Análisis bioinformático de PSVs localizadas en DS del genoma humano para detectar variantes funcionales y diseñar herramientas para su análisis a gran escala.

3. Material y Métodos

3.1. Muestras

Las muestras de pacientes y controles fueron obtenidas de distintos proyectos de investigación mediante consentimiento informado y evaluado por el Comité de Ética en Investigación Clínica (CEIC).

Se utilizaron muestras de ADN procedente de sangre periférica de; (i) 20 controles sanos (10 de género masculino y 10 de género femenino), (ii) de 98 pacientes con cariotipo normal que presentaban retraso mental y fenotipos dismórficos y (iii) de un paciente de género masculino diagnosticado de síndrome velocardiofacial, muestra 00-18.

Se utilizaron muestras de ADN y de ARN procedentes de líneas celulares limfoblastoides inmortalizadas mediante Epstein-Barr. Estas muestras procedían de; (i) cuatro pacientes diagnosticados de WBS de género masculino (grupo sw compuesto por los individuos s3, sw5, sw263 y sw266) y (ii) dos individuos con una deleción menor (grupo nw compuesto por los individuos nw10 de género masculino y nw35 de género femenino) que presentaban un fenotipo parcial.

Se utilizaron distintos *pools* de referencia según el estudio realizado. En cada estudio se disponía de un *pool* de género masculino y de un *pool* de género femenino. Así, se utilizó ADN de cinco líneas celulares limfoblastoides de género masculino y de cinco líneas celulares limfoblastoides de género femenino, ambas inmortalizadas mediante Epstein-Barr, como *pools* de referencia en el estudio piloto y en el estudio de impresión de matrices aCGH. Mientras que para el resto de experimentos con matrices aCGH se utilizó el ADN de 100 individuos sanos de raza caucásica no relacionados entre sí (50 de género masculino y 50 de género femenino). Finalmente el ARN procedente cinco líneas celulares limfoblastoides de género masculino y de cinco líneas celulares limfoblastoides de género femenino, ambas inmortalizadas mediante Epstein-Barr, se utilizaron como *pool* de referencia en los experimentos de expresión.

3.2. Datos públicos

Se analizaron los datos de 47 meduloblastomas hibridados contra un *pool* de controles de género opuesto recogidos de la base de datos GEO [87] (GSE2139) [152]. A cada muestra se le asignó un género a partir de los ratios obtenidos por los cromosomas sexuales y se eliminaron cuatro muestras que no presentaban los ratios esperados para el género masculino ni para el género femenino.

3.3. Extracción de ADN

El ADN genómico de los controles y de los pacientes fue extraído de sangre periférica utilizando el kit *Puregene DNA Purification System* (Gentra Systems, Minneapolis,

MN). La concentracion y el grado de pureza de las mismas fue determinado mediante el espectrofotómetro *NanoDrop ®ND-1000*.

3.4. Extracción de ARN

El ARN de las líneas celulares empleadas fue extraído siguiendo el protocolo de purificación con *RNeasy (QIAGEN)*. Las muestras fueron cuantificadas mediante espectrofotometría utilizando *NanoDrop ®ND-1000* y su integridad fue monotorizada mediante nanoelectroforesis capilar (*lab-on-achip, bioanalyzer, Agilent*).

3.5. Plataformas basadas en matrices

3.5.1. Diseño de las plataformas de BACs aCGH

En este trabajo se desarrollaron tres plataformas basadas en matrices aCGH fabricadas íntegramente en el *Laboratorio de Microarrays del CRG*.

Matriz aCGH 0,2K

Consta de 228 fragmentos de ADN clonados en BACs que cubren regiones flanqueadas por DSs y regiones subteloméricas. Ello conlleva a que la distribución de estos BACs a lo largo del genoma no sea homogénea. Estos insertos provienen de la librería humana 32K de BACs de *CHORI (Children's Hospital Oakland Research Institute)*. La localización de estos BACs es mayoritariamente autosómica con 210 representantes versus los 18 BACs localizados en cromosomas sexuales (15 de ellos en el cromosoma X y 3 en el cromosoma Y). Los BACs están impresos por cuadriplicados en cada portaobjetos.

Matriz aCGH pruebas de impresión

Consta de 14 BACs del cromosoma 5 (BACs localizados en cromosomas autosómicos) y 8 BACs del cromosoma X (BACs localizados en cromosomas sexuales). Los BACs están impresos por duplicados en cada portaobjetos.

Matriz aCGH 5,2K

Consta de 5.222 fragmentos de ADN clonados en BACs que se corresponden, preferentemente, con regiones cromosómicas flanqueadas por DSs en tándem y que contienen, al menos, un gen en la región de copia única o bien son fragmentos con la posibilidad de conversión entre gen y un pseudogen no funcional. En cualquier caso la distancia entre BACs es inferior a 5 Mb (por debajo del límite de resolución citogenético).

Las regiones candidatas han sido seleccionadas de una genoteca de BACs con cobertura casi completa del genoma (32.000 clones, genoteca 32K, CHORI). Se seleccionaron aproximadamente 2.500 clones para replicar y conservar. También se dispone de aproximadamente 2.500 clones del *Sanger Institute* (validados por FISH y espaciados cada Mb del genoma) y aproximadamente de 300 clones que cubren las regiones subteloméricas de

todos los cromosomas. En total se dispone de un chip de 5.222 clones cuya distribución y principales características se han resumido en la Tabla 6. En la medida de lo posible, se han utilizado BACs cuya localización ha sido comprobada por hibridación *in situ* mediante fluorescencia (FISH) o por secuenciación verificada, al menos, en sus extremos. La distribución de los clones en el portaobjetos es aleatoria y están impresos por triplicado.

Tabla 6: Resumen de las características y diseño de la matriz aCGH 5,2K

Chr	**Long (bp)**	**pb Cubiertas (%)**		**nBAC**	**nBAC en DS(%)**	
chr1	245.522.847	54.056.316	(22 %)	412	51	(12 %)
chr2	243.018.229	55.732.536	(23 %)	375	47	(13 %)
chr3	199.505.740	51.088.223	(26 %)	343	24	(7 %)
chr4	191.411.218	46.198.484	(24 %)	314	29	(9 %)
chr5	180.857.866	42.658.510	(24 %)	331	30	(9 %)
chr6	170.975.699	32.311.128	(19 %)	260	23	(9 %)
chr7	158.628.139	46.286.542	(29 %)	385	86	(22 %)
chr8	146.274.826	36.143.542	(25 %)	249	25	(10 %)
chr9	138.429.268	27.337.252	(20 %)	04	29	(14 %)
chr10	135.413.628	27.952.355	(21 %)	200	47	(24 %)
chr11	134.452.384	35.327.610	(26 %)	255	29	(11 %)
chr12	132.449.811	29.705.595	(22 %)	209	22	(11 %)
chr13	114.142.980	24.634.364	(22 %)	181	16	(9 %)
chr14	106.368.585	20.428.351	(19 %)	134	11	(8 %)
chr15	100.338.915	23.855.563	(24 %)	160	35	(22 %)
chr16	88.827.254	21.594.020	(24 %)	154	29	(19 %)
chr17	78.774.742	21.186.398	(27 %)	165	27	(16 %)
chr18	76.117.153	17.520.838	(23 %)	123	8	(7 %)
chr19	63.811.651	17.150.897	(27 %)	132	35	(27 %)
chr20	62.435.964	11.260.292	(18 %)	105	12	(11 %)
chr21	46.944.323	8.552.737	(18 %)	70	7	(10 %)
chr22	49.554.710	10.430.325	(21 %)	98	29	(30 %)
chrX	154.824.264	38.045.303	(25 %)	307	75	(24 %)
chrY	57.701.691	6.225.107	(11 %)	48	35	(73 %)
Total	3.076.781.887	705.682.288	(23 %)	5.214	761	(15 %)

3.5.2. Fabricación y condiciones de hibridación de las matrices aCGH con BACs

Matriz aCGH de 0,2K

La amplificación de los BACs se llevó a cabo mediante DOP-PCR tal y como se describe en Fiegler et al. 2003 [109]. Las amplificaciones se purificaron utilizando placas de purificación MontageHTS (Millipore) siguiendo las instrucciones del fabricante. El ADN se cuantificó con Picogreen y se secó por centrifugación desecante al vacío *SpeedVAC (Savant)*. Los BACs fueron resuspendidos en 50 % DMSO y desnaturalizado 2 min a 95°C a una concentración final de 400 ng/μl y fueron impresos por cuadriplicado utilizando un *spotter* de Affymetrix GMS417.

El proceso de hibridación se realizó marcando 400 ng de ADN de la muestra test con dCTP-Cy5 (Cy5) y 400 ng de la muestra de referencia con dCTP-Cy3 (Cy3) (en la hibridación directa; HD. El marcaje se realizó con el *kit* de marcaje *Bioprime (Invitrogen)* con las modificaciones descritas en Snijders et al 2001 [153]. Después de limpiar la reacción (Invitrogen) se dejó precipitar con 100μg de Cot-1 durante una hora con NaAc 3M, pH 5,2 (1/10 partes del volumen obtenido en la reacción) y 500 μl etanol 100 % frío a -20°C se centrifugó 30 min a 4°C, se eliminó el sobrenadante y el *pellet*. Finalmente el *pellet* fue resuspendido con 36 μl DIGEasy Hyb, 4,5 μl de ADN de esperma de salmón (20 mg/ml) y 4,5 μl de tRNA de levadura (10 mg/ml). Se realiza una desnaturalización durante 10 min a 85°C y después se deja durante 1h a 45°C para bloquear secuencias repetitivas. La hibridación se realizó en cámaras de Corning durante 40h. Una vez terminada, se realizan dos lavados con 4x 0,1xSSC; uno con 0,1 % SDS y el segundo sin y, finalmente, se realiza un lavado rápido en agua y se seca mediante centrifugación (5 min a 1.500 rpm).

Los portaobjetos fueron escaneados con el escáner de Agilent G2565BA a 10μm de resolución con 100 % de láser y fotomultiplicador.

Matriz pruebas de impresión

El proceso de amplifiación e impresión de los BACs se realizó tal y como se ha descrito en el apartado anterior pero, en este caso, la impresión se realizó con el robot *spotter* VersArrayPro de Bio-Rad.

La impresión de las matrices se llevó a cabo utilizando diferentes soluciones de resuspensión (50 % DMSO, 150mM Phosphate, 3xSSC, PRONTO-Epoxy, PRONTO-Amino y Schott-Epoxy) a distintas concentraciones de ADN impreso (100 ng/μl, 150 ng/μl, 200 ng/μl y 400 ng/μl).

La hibridación sobre se llevó a cabo siguiendo el protocolo de Wang NJ *et al.* 2004 [154] que utilizó sulfato de dextrano en la solución de hibridación. Los portaobjetos fueron analizados con el escáner de Agilent G2565BA y las imágenes fueron cuantificadas con el *software GenePix Pro 6.0* utilizando la opción de *irregular spot finding features* con el sistema de marcado por defecto. Esta opción ajusta la forma del *spot* cuando se realiza la lectura de su señal y la lectura del ruido de fondo que existe alrededor de él. En todos los casos se realizaron hibridaciones directas, HD, (marcaje de la muestra con Cy5 y del *pool* de referencia con Cy3) y hibridaciones reversas, HR, o *dye-swap* (itercambiando los fluorocromos).

Matriz aCGH 5,2K

El proceso de amplifiación e impresión de los BACs se realizó tal y como se ha descrito en el apartado anterior.

La impresión de las matrices se llevó a cabo utilizando 50 % DMSO como solución de resuspensión a una concentración de ADN impreso de 400 ng/μl.

La hibridación sobre los portaobjetos Ultragaps de Corning se llevó a cabo siguiendo el protocolo de Wang NJ *et al.* 2004 [154] que utilizó sulfato de dextrano en la solución de hibridación. Los portaobjetos fueron analizados con el escáner de Agilent G2565BA y las imágenes fueron cuantificadas con el *software GenePix Pro 6.0* utilizando la opción de *irregular spot finding features* con el sistema de marcado por defecto. Esta opción ajusta la forma del *spot* cuando se realiza la lectura de su señal y la lectura del ruido de fondo que existe alrededor de él. En todos los casos se realizaron hibridaciones directas, HD, (marcaje de la muestra con Cy5 y del *pool* de referencia con Cy3) y hibridaciones reversas, HR, o *dye-swap* (itercambiando los fluorocromos).

3.5.3. Protocolo de hibridación de oligo aCGH

Se utilizaron dos matrices aCGH de oligonucleótidos comerciales (oligo aCGH) de 60-mer: Agilent G4410B y Agilent G4411B. En ambos casos se partió de 1.000 ng de ADN por muestra y se siguió las instrucciones de la casa comercial versión 4 con *Bioprime arrayCGH Labelling Kit* (Ref. 18095-011, Invitrogen), después de los lavados, los portaobjetos fueron analizados con el escáner *Agilent Microarray 2565BA* y las imágenes fueron cuantificadas con el *software Genepix Pro 6.0* utilizando *irregular spot finding features* con el sistema de marcado por defecto para las matrices G4410B. Mientras que para las matrices G4411B, la cuantificación de las imágenes se realizó con el *software Feature Extraction* de Agilent con las opciones estándar recomendadas por la casa comercial. Las hibridaciones HD se realizaron marcando la muestra con Cy5 y del *pool* de referencia con Cy3 y en algunos casos se realizaron hibridaciones HR.

3.5.4. Protocolo de hibridación de matrices de expresión (aExpr)

Se ha utilizado una plataforma basada en matrices de oligonucleótidos de 44K de cobertura (ExprAg44K). Antes de la realización de la hibridación, se analizó la calidad de las muestras y el *pool* de referencia mediante electroforesis capilar de Agilent (*Bioanalyzer*) y únicamente se hibridaron aquellas muestras con un RIN (*RNA Integrity Number*) superior a 8. Posteriormente se marcaron 500 ng de ARN total de la muestra problema y del ARN de referencia utilizando el *Low Input linear RNA Amplification Kit* siguiendo las instrucciones de la casa comercial.

Las hibridaciones se realizaron siguiendo el protocolo v2.2, dejando los portaobjetos 18 h a 60 °C a 4 rpm y realizando los siguientes lavados en agitación:

- 6xSSPE (Invitrogen) + 0,1 % N-LauroylSarcosine (Sigma) durante un minuto.
- 0,2xSSPE + 0,01 % N-LauroylSarcosine durante un minuto.
- Actonitrilo durante un minuto.
- *Stabilization and Drying Solution*, 30 segundos.

Los portaobjetos fueron analizados con el escáner de Agilent G2565BA y las imágenes fueron cuantificadas con el *software Genepix Pro 6.0* utilizando la opción de *irregular spot*

finding features con el sistema de marcado por defecto. En todos los casos se realizaron hibridaciones HD y HR. Las HD se realizaron marcando la muestra test con Cy5 y del *pool* de referencia con Cy3 y las HR se realizaron itercambiando los fluorocromos.

3.6. Métodos de normalización y preprocesado de los datos

Los métodos de normalización variaron según el estudio que se llevó a cabo. Así, en el estudio piloto y en el estudio de impresión de matrices aCGH se aplicó el método de normalización *loess* con una amplitud de ventana de 0,3 y con substracción del ruido de fondo. Y, además, se eliminaron aquellos *spots* con una variabilidad superior a 0,1 entre réplicas de un mismo portaobjetos. En el estudio de las fuentes de variación se aplicaron cuatro métodos distintos de normalización sobre los mismos datos con y sin susbtracción del ruido de fondo con una amplitud de ventana de 0,3; *loess, print-tip loess, loess loc y loess loc sacle*. En el resto de secciones se aplicó *print-tip loess* sin substracción del ruido de fondo para BAC aCGH. Mientras que en el caso de matrices oligo aCGH y aExpr se aplicó el método de normalización *loess* sin substracción del ruido de fondo.

En todos los casos se excluyeron aquellos *spots* con una media de la señal inferior a dos veces el ruido de fondo (FG/BG) en ambos canales.

3.7. Validación de CNVs

La validación de CNVs se llevó a cabo cruzando los resultados de las posiciones de las sondas de las matrices aCGH de BACs y oligo aCGH. Las coordenadas de las posiciones de los BACs y sondas proceden del hg18.

El estudio de las regiones con CNVs se realizó a partir de los *browsers:*

http://projects.tcag.ca/humandup

http://davinci.crg.es/cgi-gbrowse/hg18.

3.8. Fuentes de variación en aCGH y sus causas

3.8.1. Fuentes de variación asociadas al proceso de fabricación e hibridación

En las Figuras 12 y 13 se esquematizan los distintos diseños experimentales realizados en el estudio de las fuentes de variación en aCGH.

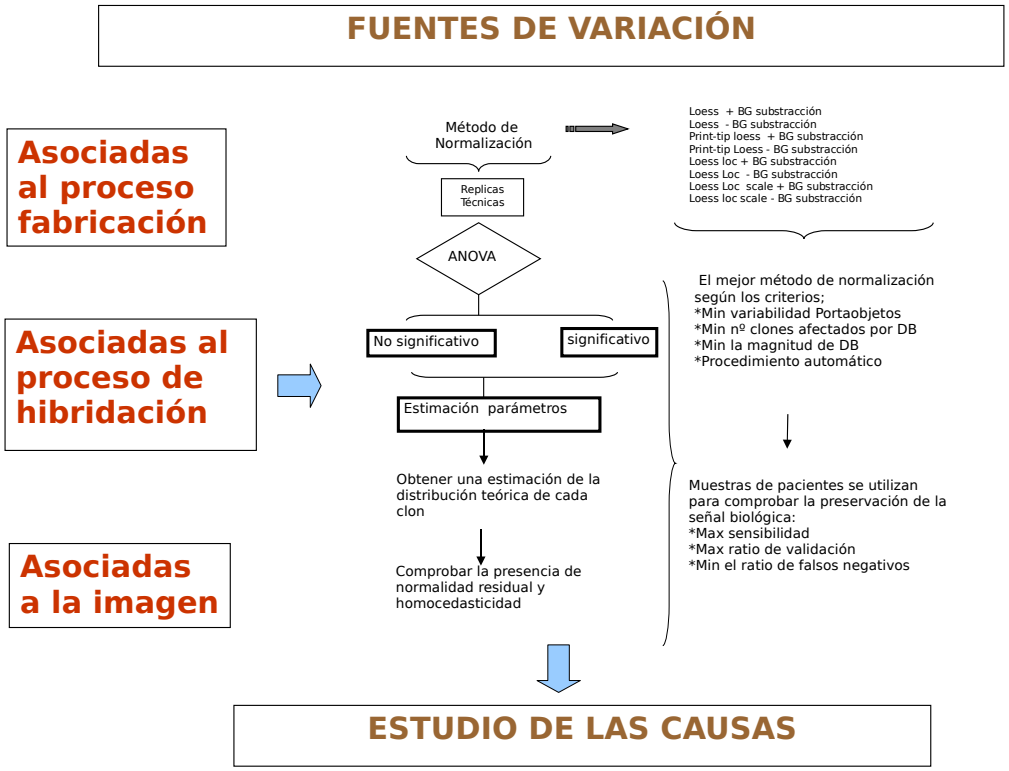

Figura 12: **Fuentes de variación**. En este esquema se representa el proceso utilizado en el estudio de las matrices aCGH y como se relacionan entre sí. El objetivo es la detección de CNVs. Para ello se eligieron las condiciones óptimas en el proceso impresión (fabricación) y se estudiaron las fuentes de variación asociadas al proceso de hibridación con la intención de obtener un estándard de calidad.

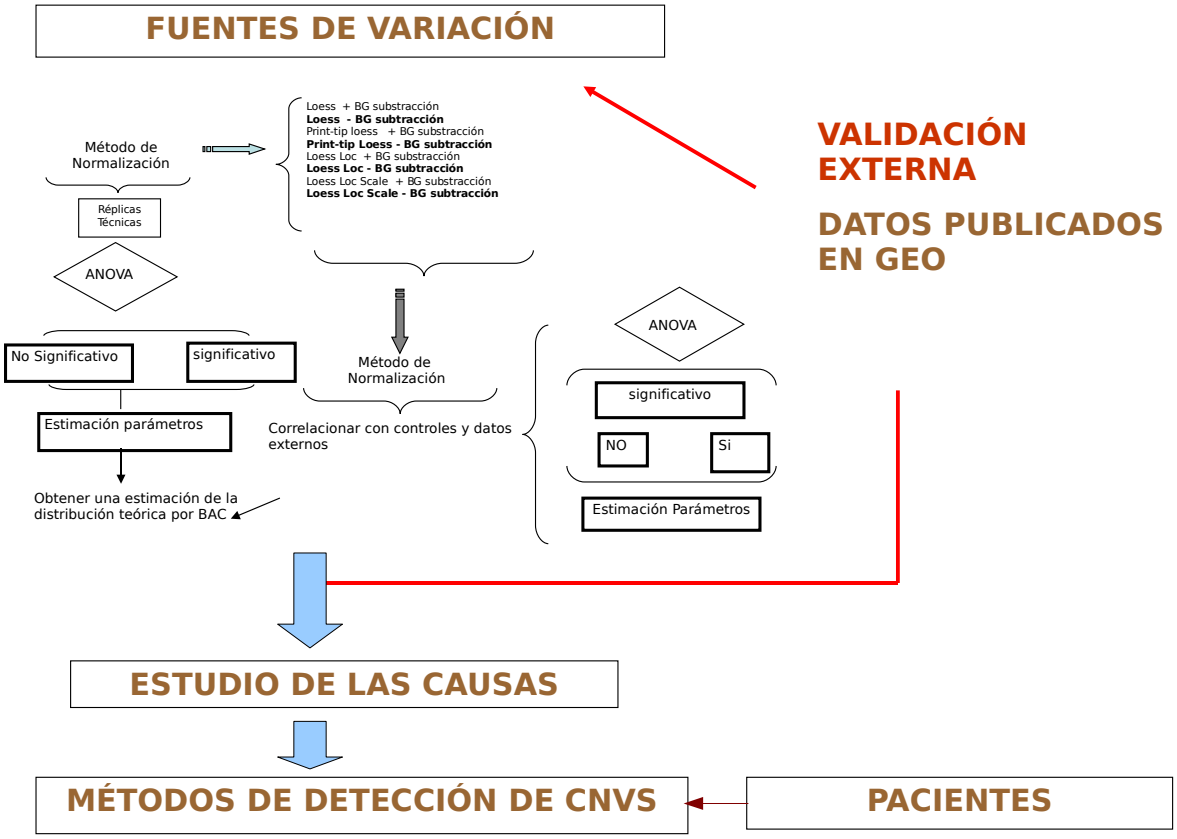

Figura 13: **Detección de errores sistemáticos**. Se realizó otro modelo ANOVA sobre dos sets de datos distintos con la intención de validar los resultados obtenidos en el estudio sobre las fuentes de variación asociadas al proceso de hibridación. Esta información se utilizó para desarrollar métodos más eficientes en la detección de CNVs

Condiciones experimentales óptimas para la impresión de matrices aCGH

Mediante un modelo ANOVA se evaluaron las siguientes condiciones;

1. Concentración de ADN impresión (Conc); 100 ng/μl, 150 ng/μl, 200 ng/μl y 400 ng/μl
2. Solución de resuspensión utilizada en la impresión (Sol): 50 %DMSO, 3xSSC, 150mM Phosphate, PRONTO-Amino, PRONTO-Epoxy y Schott-Epoxy
3. Intercambio de fluorocromos (DB); HD y HR
4. BACs analizados (BAC); 14 BACs tomados al azar en cromosomas autosómicos y 8 BACs tomados al azar en cromosomas sexuales.
5. Portaobjetos sobre el que se ha realizado la hibridación (P); en este estudio se considera como bloque aleatorio anidado a DB. Se dispone de 12 portaobjetos para HD y 12 portaobjetos para HR.
6. Tipo de portaobjetos que podía ser Ultragaps o Codelink. Se utilizaron 12 portaobjetos (6 HD y 6 HR) para cada tipo.

En cada portaobjetos había dos réplicas de cada BAC.

Detección de fuentes de variación asociadas a la fiabilidad de la medida

Para este estudio se diseñó y formuló un modelo ANOVA jerárquico. Los efectos significativos fueron estimados mediante un modelo lineal jerárquico de efectos mixtos (ver anexo 9.3) utilizando el algoritmo descrito por Shafer 2007 [155].

Se estudiaron cuatro posibles fuentes de variación: (i) el sesgo producido por el intercambio de fluorocromos (*Dye-Bias*, DB), (ii) el efecto técnico (evaluado mediante dos técnicos que realizaron todos los experimentos simultámeamente y de manera indepen-diente), (iii) el efecto día (bajo este efecto se quería evaluar las pequeñas diferencias que pueden darse en un laboratorio en distintos momentos, los experimentos se realizaron en dos días diferentes) y (iv) el portaobjetos (P) sobre el cual se realiza el experimento y que es resultado de la fabricación propia de la matriz. El efecto P está anidado al efecto DB, técnico y día. Se realizaron cuatro réplicas (equivalente a cuatro portaobjetos) en la mismas condiciones de DB, técnico y día, realizando un total de 32 hibridaciones de una misma muestra control de género femenino (experimento 32x).

Además, este estudio permitió valorar el efecto de distintos métodos de normalización sobre la variabilidad de cada portaobjetos y sobre las fuentes de variación estudiadas. El efecto de la normalización sobre la variabilidad asociada a cada portaobjetos se valoró mediante un análisis descriptivo utlizando la desviación estándar y el MAD (*Median Absolute Deviance*) mientras que el efecto de la normalización sobre las fuentes de variación se valoró mediante la capacidad de cada método de minimizarlas.

3.8.2. Estudio del efecto DB y sus causas

El efecto DB como error sistemático. Detección del efecto DB en otros experimentos

Los datos de 19 controles sanos (experimento 19c) y de 43 muestras (experimento 47m) obtenidas de GEO (GSE2139) fueron analizadas mediante modelos ANOVA donde los factores considerados fueron DB, muestra y género. Posteriormente se aplicó un modelo lineal jerárquico mixto con la misma estructura para estimar los valores de los efectos [155] significativos.

El estudio se repitió con distintos métodos de normalización. Y se realizó un análisis descriptivo de la variabilidad de cada portaobjetos según el método de normalización empleado. Los estadísticos descriptivos utilizados fueron la desviación típica y MAD. Además se estudió el efecto de la normalización sobre el efecto DB.

El efecto DB asociado a la estructura primaria del ADN

En este apartado se utilizó el método de normalización *loess loc scale* para clasificar

los BACs en tres grupos:

- Clones con un gran efecto positivo de DB (afinidad al fluorocromo Cy3).
- Clones con un gran efecto negativo de DB (afinidad al fluorocromo Cy5).
- Clones sin efecto DB.

A partir de la secuencia génica de cada clon se calculó; el número de islas CpG y el número de genes así como sus respectivas longitudes, la longitud de CNVs conocidas incluídas en él y la longitud de las DSs contenidas. También se obtuvo información sobre variables de tipo cualitativo como; la presencia o ausencia de DSs, de CNVs, de genes y la distribución cromosómica de los clones.

Se estudió la posible asociación entre tipo de clon y las características estudiadas de la secuencia de estos clones. Para ello se aplicó el test no paramétrico U de Mann-Whitney para las variables de tipo cuantitativo que permite comparar dos grupos independientes y el test Kruskall-Wallis para las variables de tipo cuantitativo que permite comparar más de dos grupos independientes. Se aplicó test de Fisher para analizar todas las variables cualitativas excepto para la distribución cromosómica. En este caso se aplicó una variante del test chi-cuadrado donde el pvalor se obtiene mediante remuestreo (tamaño de remuestra de 10.000) solventando, así, los problemas derivados del pequeño tamaño muestral.

El contenido de CNV fue obtendido de la base de datos TCAG (última actualizacion de Octubre del 2006, accesible en:

http://projects.tcag.ca/variation/download.html

Las islas CpG fueron identificadas utilizando el programa Newcpgreport de EMBOSS

http://bioweb.pasteur.fr/docs/EMBOSS/newcpgreport.html

Los genes y las duplicaciones segmentarias fueron definidas a partir de su anotación en NCBI Build 36

http://genome.ucsc.edu.

Las longitudes de las distintas variables fueron calculadas como la suma del total de pares de bases (pb) incluidas en cada BAC.

El estado de purificación del ADN y el efecto DB

Se realizó una segunda extracción de ADN de la muestra control utilizada en el experimento 32x. El protocolo de extracción de ADN de GENTRA fue modificado en dos pasos limitantes; (i) en el tratamiento con RNAsa y (ii) en la cantidad utilizada de solución de precipitación de proteínas tal y como se especifica en la Tabla 7. En la Tabla 7 se puede

observar, mediante los ratios 260/230 (contaminación proteíca) y 280/230 (contaminación salina), los distintos grados de purificación del ADN conseguidos.

Finalmente se optó por hibridar las muestras de ADN procedentes de las condiciones 1, 3, 9 y 12 en la matriz aCGH 5,2K.

El efecto del grado de purificación se midió en relación a la capacidad de detectar las CNVs conocidas en esta muestra (sensibilidad) y en relación al número de datos no concordantes entre las hibridaciones HD y HR.

3.8.3. Fuentes de variación asociadas a la imagen

Se utilizó el *software* de libre distribución *RealSpot* [156] para evaluar las imágenes de los *spots*.

La visualización de la imagen de los *spots* permite dividir las formas de estos en, al menos, cinco categorías: normal (Wt), aureola (A), donut (D) y artefacto (F) y no presente (Ms), ver anexo 9.4 para más detalles

Clasificación de *spots* a partir de las imágenes obtenidas por GenePix

Se utilizaron seis hibridaciones (tres HD y tres HR) realizadas en seis portaobjetos distintos con la matriz aCGH 5,2K de la muestra 00-18. Se eligieron al azar 600 *spots* (100 por cada portaobjetos) y fueron valorados por la misma persona en dos rondas distintas. El orden de cada portaobjetos en las rondas fue elegido aleatorio.

La clasificación de las formas de los *spots* mediante las variables obtenidas por GenePix se realizó con un análisis discriminante.

Fiabilidad entre observadores en la evaluación de la imagen

Se utilizaron los datos procedentes de una hibridación en un portaobjetos del experimento 32x. Cuatro observadores independienetes valoraron las mismas imágenes de los *spots* en cinco rondas de distintos tamaños muestrales distribuidos al azar. Las rondas estaban compuestas por los siguientes tamaños muestrales; ronda1 160 *spots*, ronda2 80 *spots*, ronda3 200 *spots*, ronda4 40 *spots* y ronda5 120 *spots*. El intervalo temporal entre rondas fue de 24h como mínimo. En cada una de las rondas había 20 *spots* en común con otra ronda. Estos *spots* se utilizaron para obtener una medida de la concordancia intra-individual que se calculó con el índice Kappa, este índice también se utilizó para evaluar la concordancia entre evaluadores.

Tabla 7: **Modificación de la calidad del ADN** mediante modificaciones en el tratamiento con RNAsa y en las cantidades de la solución de precipitación de proteínas utilizadas.

Cond	V[a] Sangre	RBC[b]	*Cell lysis*	RNAsa	Prot Prec[c]	Isopropanol	Etanol 70 %	DNA hyd[d]	Vf[e]	ng[f]	Conc DNA[g]	260/280	260/230
1	3 ml	9μl	3μl	15μl	1μl	3μl	3μl	250μl	250ml	65000	426,81	1,86	1,88
2	0,9ml	2,7μl	900μl	4,5μl	300μl	900μl	900μl	100μl	100ml	40800	329,14	1,84	1,59
3	0,9ml	2,7μl	900μl	NO	300μl	900μl	900μl	100μl	100ml	40400	373,31	1,84	1,7
4	0,9ml	2,7μl	900μl	4,5μl	200μl	900μl	900μl	100μl	100ml	28800	389,49	1,7	0,91
5	0,9ml	2,7μl	900μl	NO	200μl	900μl	900μl	250μl	250ml	85750	61,41	1,83	0,7
6	0,9ml	2,7μl	900μl	4,5μl	400μl	900μl	900ml	250μl	250ml	14250	30,30	1,5	0,3
7	0,9ml	2,7μl	900μl	NO	400μl	900ml	900μl	100μl	100ml	2800	20,72	1,72	0,17
4	0,9ml	2,7μl	900μl	4,5μl	200μl	900μl	900μl	100μl	100ml	41500	328,34	1,7	0,94
5	0,9ml	2,7μl	900μl	NO	200μl	900μl	900μl	100μl	100ml	35000	336,53	1,8	1,27
6	0,9ml	2,7μl	900μl	4,5μl	400μl	900μl	900μl	100μl	100ml	32800	346,75	1,86	1,66
7	0,9ml	2,7μl	900μl	NO	400μl	900μl	900μl	100μl	100ml		6,95	1,44	0,11
8	0,9ml	2,7μl	900μl	4,5μl	150μl	900μl	900μl	100μl	100ml	27700	322,60	1,75	1,07
9	0,9ml	2,7μl	900μl	NO	150μl	900μl	900μl	100μl	100ml	43500	245,70	1,68	0,81
10	0,9ml	2,7μl	900μl	4,5μl	100μl	900μl	900μl	100μl	100ml	34900	311,85	1,78	1,17
11	0,9ml	2,7μl	900μl	NO	100μl	900μl	900μl	100μl	100ml	89400	323,41	1,84	1,55
12	900μl	2,7μl	900μl	4,5ml	500μl	900μl	900μl	100μl	100ml	36600	325,17	1,45	0,47
13	900μl	2,7μl	900μl	NO	500μl	900μl	900μl	100μl	100ml	12500	73,29	1,39	0,27
14	3ml	9μl	3μl	15μl	1ml	3ml	3ml	250μl	250ml	263750	304,51	1,86	1,8

[a]: (V) Volumen
[b]: (RBC)
[c]: (Prot Prec) Precipitación proteíca
[d]: (DNA hyd)
[e]: (Vf) Volumen final
[f]: (ng) ng totales
[g]: (Conc ADN) Concentración ADN

Datos atípicos y las formas de los *spots*

Se utilizaron los datos procedentes de cinco hibridaciones (dos HD y tres HR) del experimento 32x. Cuatro observadores independientes valoraron las imágenes en cinco rondas con un intervalo temporal entre rondas de 24h como mínimo. Los mismos *spots* (N=99) fueron valorados en cada ronda y en cada portaobjetos. Cada uno de estos *spots* están asociados a datos atípicos en, al menos, una hibridación.

Los valores atípicos se identificaron mediante los residuos estudentizados (con valor absoluto > 3) obtenidos de los modelos lineales mixtos para la muestra 32x de BACs que tenían un pvalor para la normalidad ajustando por Bonferroni inferior a 0,05.

Se aplicó el test chi-cuadrado para determinar si existía asociación entre la forma del *spot* y concordancia entre *spots* con ser o no valor atípico. El tipo de forma y ser o no valor atípico se comparó con las variables obtenidas por GenePix mediante el test no paramétrico U de Mann-Whitney con la intención de obtener una regla de clasificación de *spots* automática. Las variables significativas se agruparon entre sí en la construcción de una puntuación que reflejara la calidad. Posteriormente se realizó una validación cruzada (*cross-validation*) aplicando la puntuación de calidad sobre otro set de datos atípicos de otros dos portaobjetos del experimento 32x escogidos al azar (una HD y una HR).

3.9. Métodos de detección de CNVs

Se han aplicado diversos métodos de detección de CNVs sobre las matrices aCGH con BACs; (i) estanderización de los datos mediante la substracción de la mediana y dividiendo por el el rango intercuartílico (IQR) divido por 1,349, (ii) puntos de corte basados en la media de las réplicas y en la variabilidad de las réplicas (PCmed), (iii) puntos de corte basados en el valor mínimo (duplicaciones) o máximo (deleciones) de un conjunto de sondas replicadas (PCmin), (iv) aplicación de un Intervalo de Confianza (IC) sobre las sondas para detectar datos atípicos (v) aplicación del método *Circular Binary Segmentation* (CBS) [52] y (vi) la aplicación de los métodos PCmed, PCmin y CBS después de substraer los valores esperados de los datos caracterizados mediante las estimaciones realizadas por el modelo lineal mixto sobre el set de datos 19c o bien (para BACs con datos perdidos) mediante las estimaciones realizadas en este mismo set de datos mediante remuestreo (métodos combinados).

Sobre las matrices oligo aCGH se aplicaron los siguientes métodos; (i) aplicación de un método basado en puntos de corte considerando CNVs aquellas regiones con 3 sondas sobre 3 consecutivas con valores de M en valor absoluto mayores a 0,2 situadas en el mismo brazo y cromosoma. O bien regiones con 3 sobre 5 sondas consecutivas o sus múltiplos que cumplan este requisito y (ii) aplicación del método CBS con los siguientes parámetros; realización de suavizado, sd=3, alpha=0,01.

Además para la detección y caracterización de CNVs se aplicó un método gráfico llamado *Cromoplot* que consiste en representar los valores de M ordenados mediante su

posición cromosómica a lo largo de un cromosoma.

Los distintos métodos utilizados en la detección de CNVs se compararon entre ellos mediante gráficos; (i) en datos reales se representó la sensibilidad en el eje de ordenadas (entendiendo como sensibilidad la capacidad de detectar 8 BACs conocidos como CNVs en la muestra control del experimento 32x) y la eficiencia en el eje de abcisas (entendiendo como eficiencia el número de sondas conocidas reales sobre el total detectado) y (ii) los datos obtenidos mediante simulación se representaron mediante curbas ROC (*Receive operator Curbe*) [157], en ellas se representa la sensibilidad en el eje de ordenadas (número de sondas positivas sobre el total de sondas positivas) y 1-especificidad en el eje de ordenadas (entendiento como especificidad el número de sondas detectadas negativas sobre el total negativo).

Estos métodos se aplicaron sobre las hibridaciones HD y HR. Se denominó global a la mera agrupación de los resultados obtenidos por ambas hibridaciones.

3.10. Análisis transcriptómico en individuos con aneusomías

Para detectar pérdidas y ganancias en los pacientes con WBS se aplicó el método combinado con PCmin situando el punto de corte en $|0, 2|$.

En los casos en los que se han aplicado ANOVAs el gen se ha tratado como bloque o bien se ha realizado un ANOVA por cada gen. Y, en los casos en los que se ha aplicado el test t-student y SAM [158], los valores obtenidos por las distintas sondas se han colapsado mediante la media aritmética en genes.

Para detectar sobre o infraexpresión en las regiones WBS, *upstream* (UPS), *downstream* (DWS) y otras regiones control (C1, C2 y EXT) se aplicaron modelos ANOVA.

Para detectar sobre o infraexpresión global se aplicó; (i) dos veces (una vez para HD y otra para HR) el test t-student sobre el grupo sw y sobre el grupo sw + nw considerando significativos aquellos genes con pvalores menores de 0,01 (ii) dos veces (una vez para HD y otra para HR) SAM sobre el grupo sw y sobre el grupo sw + nw encontrando significativos aquellos genes con un FDR 5 % y (iii) ANOVA para detectar diferencias entre grupos aplicando la corrección de Benjamini-Hochberg (BH) [159] para multitest (iv) el método clúster K-means sobre los genes en los que se hallaron diferencias significativas entre grupos con ANOVA. Se emplearon *heatmaps* para representar los genes diferencialmente expresados entre nw y sw.

Una vez identificados los genes estadísticamente interesantes se utilizó; (i) la base de datos REACTOME [160] para identificar las vías significativamente dereguladas según el test hipergeométrico implementado en dicha base de datos (ii) Biomart [161] para identificar la función de estos genes cuando no resultaron significativos en REACTOME.

3.11. Localización de variantes parálogas de secuencia (PSV)

Este estudio se realizó a fecha de Enero del 2008. El material utilizado ha sido las secuencias anotadas de los transcritos génicos antes de realizar el *splicing* y las secuencias anotadas de los genes procesados (región codificante). Las secuencias se han obtenido de la base de datos Biomart [161] (htpp://www.biomart.org) creada conjuntamente por Cold Spring Harbor y EBI [162] a partir del built 36 o hg18 que toma como referencia la base de datos de genes de la especie *Homo sapiens* de SANGER. La anotación completa del genoma utilizada para realizar la búsqueda de secuencias con elevada similitud a los genes obtenidos procede de genome.uscs.edu y se utilizó la misma versión del genoma. La anotación de las DS proviene de la base de datos TACG (http://projects.tcag.ca) y de NCBI Build 36 (http://genome.ucsc.edu). Los programas utilizados se describen a continuación:

- BLAT instalado localmente [163].
- El lenguaje awk que permite aplicar filtros sobre los resultados obtenidos en BLAT y R [164].
- El algoritmo ha sido escrito íntegramente en *Perl* y para ello se han utilizado los modulos de *Bioperl*; *PrimarySeq*, *SimpleAlign*, *AlignIO* y *StandAloneBlast*.

3.12. Consideraciones generales sobre el análisis estadístico

En el análisis estadístico se ha utilizado el *software* de libre distribución R [164].

La normalidad de los residuos de todos los modelos ANOVA fue testada mediante el test de Shapiro-Wilks y la homocedasticidad se validó aplicando el test de Bartlett y/o de manera gráfica mediante boxplots. En las pruebas post-hoc se aplicó el test de Tukey-Schefee.

Las representaciones espaciales se realizaron con la librería GeoR [165]

4. Resultados

4.1. Estudio piloto

Se estudiaron 98 individuos (64 de género masculino y 34 de género femenino) con cariotipo normal que presentaban retraso mental y fenotipos dismórficos. Siete de estos individuos eran muestras de pacientes con patología y etiología conocida (tres con AS, dos con PWS, uno con WBS y un individuo con SMS). Por cada muestra se realizó una hibridación HD.

En una primera aproximación, la detección de regiones alteradas (con cambios en la dosis génica) se llevó a cabo mediante la aplicación de un umbral fijo para todas las hibridaciones sobre la variable respuesta (M). Considerando la presencia de una alteración si $|M| > 0{,}3$. Así se seleccionaron varias regiones candidatas por individuo que se intentaron validar mediante otras técnicas obteniendo un porcentaje de validación de un $\sim 30\,\%$.

Este porcentaje de validación es relativamente bajo y, con la finalidad de hallar un método alternativo, se estudió el comportamiento de M así como las variables asociadas a calidad obtenidas por el programa GenePix que realiza la lectura de las imágenes.

En la Figura 14 se representan los valores de M por BAC. Estos *boxplots* muestran como los valores de M, para algunos BACs, no están centrados en el cero (ver BAC 87) y también se observa diferencias en la dispersión entre BACs (ver BAC 111).

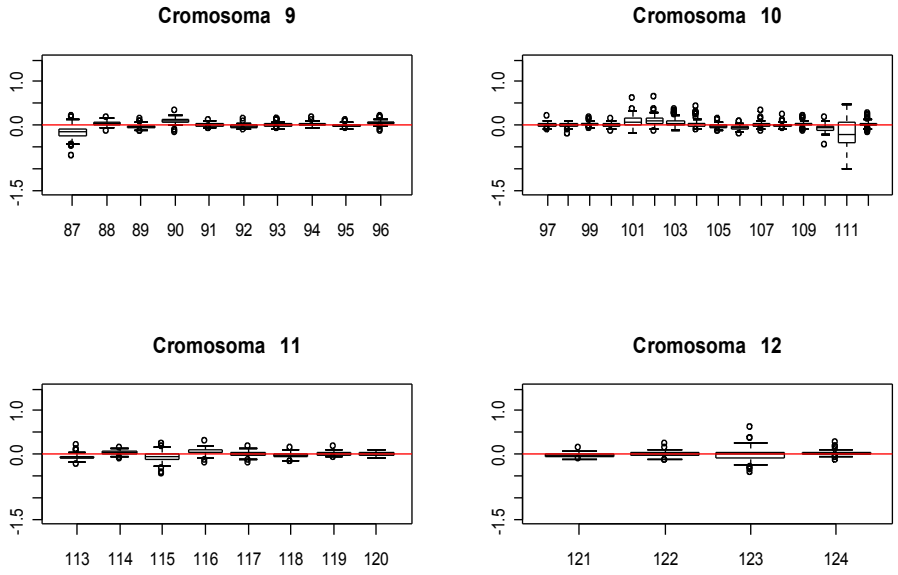

Figura 14: **Cada boxplot representa un BAC**, en el eje de coordenadas aparece el número de orden del BAC sobre los 228 utilizados. En el eje de ordenadas se muestra los valores de log-ratio que ha obtenido cada BAC en los 98 experimentos realizados.

A partir del estudio de las distintas variables obtenidas por el programa GenePix se dedujo la presencia de una distribución espacial diferencial de los fluorocromos en los portaobjetos (ver Figura 15).

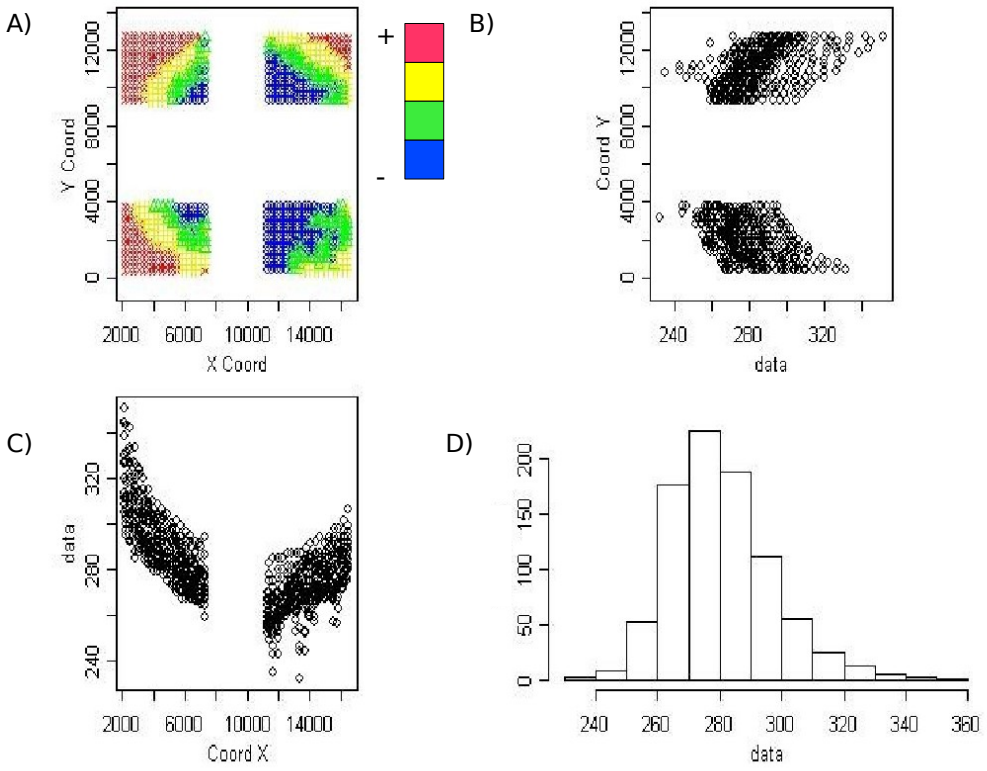

Figura 15: **Distribución de la mediana de la intensidad ruido de fondo o background local** para el fluorocromo Cy5. A) Se representa la distribución de la mediana del ruido de fondo asociado al fluorocromo Cy5 a lo largo del portaobjetos y en ella se observan cambios en la intensidad. Estos cambios reflejan zonas periféricas con mayor intensidad y una parte central con una intensidad mucho menor. B) Se observa la distribución de la mediana del ruido de fondo asociado al fluorocromo Cy5 a lo largo del eje vertical del portaobjetos. En este gráfico se observa como el valor de la mediana del fluorocromo Cy5 varía de forma no aleatoria a lo largo del eje vertical. C) Se observa la distribución de la mediana del ruido de fondo asociado al fluorocromo Cy5 a lo largo del eje horizontal del portaobjetos dónde también se observan cambios no aleatorios en función de la posición. D) se presenta un histograma de los datos.

Bajo la hipótesis de una afinidad diferencial a los fluorocromos según la posición de los BACs en el portaobjetos, se desarrolló un nuevo método para la detección de regiones alteradas. Este método consistía en estanderizar los valores de la variable respuesta por cada clon tal y como se muestra en la ecuación 1. Debido a que se esperaba encontrar en en algunos BACs alteraciones que podían estar relacionadas con la enfermedad a estudiar se utilizó la mediana e IQR dividido entre 1,349 en lugar de la media y la desviación estándar respectivamente.

$$\begin{array}{c} y_{ij} = \mu_i + e_j \quad \imath = 1, ..., 210 \quad j = 1, ..., 98 \\ E(e_i) = 0 \quad Var(e) = \sigma_i^2 \end{array} \tag{1}$$

Dónde y_{ij} es la media de los triplicados del logaritmo en base 2 del ratio del canal Cy5 versus el canal Cy3 normalizado (M). i los BACs analizados y j las muestras estudiadas. En el caso de los 18 BACs situados en cromosomas sexuales la estanderización se

realizó únicamente sobre los individuos del mismo género

Finalmente se representaron los valores de M y los resultantes de la estanderización en *cromoplots* (ver Figura 16). Se realizó un *cromoplot* por cada hibridación o muestra y ello permitió observar distintos grados de variabilidad de la respuesta M que conllevó a considerar también la variabilidad del perfil en la detección de una alteración. Así, para que un clon fuera considerado alterado debía obtener un valor absoluto estanderizado mayor de 2,5 y, además, estar fuera del límite descrito por la variabilidad de cada perfil (variabilidad dependiente de la hibridación). Este límite se estableció en 3 veces el IQR dividido entre 1,349.

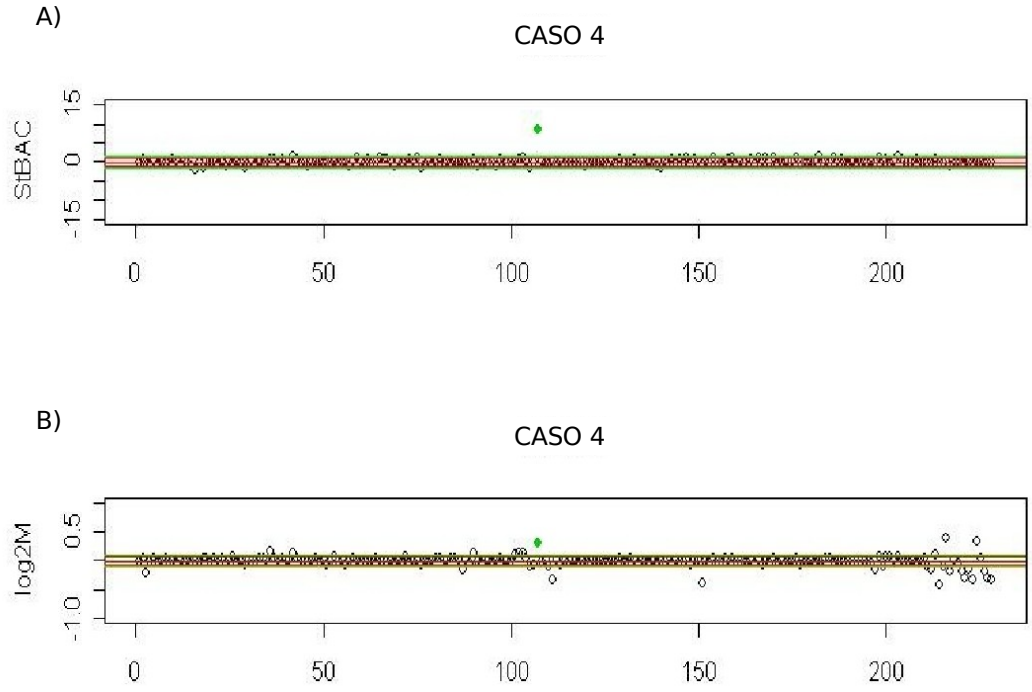

Figura 16: **Perfiles antes y después de estanderizar** A) Perfil estanderizado y B) Perfil de una hibridación antes de estanderizar. En los gráficos se representan los valores del log2 ratio de los 228 BACs ordenados según su posición en el genoma. Las línea roja representa el valor cero indicativo de la ausencia de alteración mientras que las líneas verdes representan 2,5 veces el IQR divido entre 1,349. El punto verde se corresponde con una alteración cromosómica real validada por FISH.

El método de detección basado en estanderizar permitió validar más del 70 % de los BACs considerados alterados incrementando notablemente el ratio de validación ($\sim 30\,\%$) respecto a aplicar un umbral fijo sobre el valor M.

La sensibilidad del método se midió a partir de las siete muestras de pacientes que padecen síndromes de microdeleciones conocidos. Cinco de estos pacientes presentan una deleción en heterocigosis en la banda cromosómica 15q11-q13 (tres pacientes con AS y dos pacientes con PWS). En los cinco casos la técnica fue lo suficientemente sensible para detectar la deleción y los puntos de rotura asociados. La matriz tenía tres clones

que cubrían 4,5 Mb de la región 15q11 (RP11-289D12, RP11-98D02 y RP11-10K20). Se detectó correctamente los puntos de rotura de cada una de ellas; tres muestras tenían los puntos de rotura situados en el clon 1 y clon 2 y las otras dos muestras en el clon 2 y el clon 3. El paciente con WBS presenta una deleción de la región 7q11.2 que estaba cubierta por un clon; RP11-41F22. Esta deleción en heterocigosis también fue detectada. En el caso del paciente con SMS sólo uno (RP11-189D22) de los dos clones (RP11-189D22, RP11-312O19) que cubrían la región de 2 Mb situada en la banda 17p11 fue detectado. La técnica ha permitido detectar deleciones con un tamaño entre 1,5 Mb y 4,0 Mb con una sensibilidad del 100 % en la detección de regiones alteradas y en 6 de los 7 (86 %) casos se situaron correctamente los puntos de rotura.

En el caso de los pacientes con retraso mental y caracteres dismórficos de etiología desconocida y cariotipo normal (91 casos) se confirmaron aquellos reordenamientos no detectados como CNVs en las bases de datos públicas a fecha de Julio del 2005 (ver Tabla 8). En la Tabla 8 se ha actualizado esta información hasta Marzo del 2008.

Tabla 8: Alteraciones cromosómicas detectadas y validadas

Caso	Síndrome asociado	N BAC [a]	Tipo de reordenamiento detectado	Locus	Tamaño (Mb)	CNVs? [b]	Técnica empleada en la validación
1	MR	1	Deleción	18q23	0,5	Sí	FISH
2	Silver-Rusell	4	Duplicación	7q36	10	No	Microsatélites
2	Silver-Rusell	2	Deleción	10q	8	Sí	Microsatélites
3	Silver-Rusell	1	Duplicación	7p12.2	1-3	No	Microsatélites
4	Autismo	1	Duplicación	15q11.2	4	No	FISH
5	Autismo	1	Deleción	15q11	0,4	Sí	FISH
6	Autismo	1	Duplicación	10q11	11	Sí	FISH

[a]: Número de BACs detectados con la plataforma desarrollada para esta sección.

[b]: Indica la presencia de CNVs detectadas en la misma región a fecha de Marzo 2008.

4.2. Fuentes de variación asociadas al proceso de fabricación e hibridación de matrices aCGH

4.2.1. Condiciones experimentales óptimas para el proceso de impresión

Este apartado se divide en dos partes; (i) formulación y desarrollo del modelo matemático utilizado y (ii) los resultados obtenidos a partir de la aplicación de dicho modelo.

Formulación y desarrollo del modelo ANOVA

En la ecuación 7 se presenta el modelo ANOVA resuelto en esta sección.

$$y_{biklpn} = BAC_b + DB_i + Conc_k + Sol_l + P(DB)_p + BAC:DB_{bi} + DB:Conc_{ik} + +DB:Sol_{il} + Conc:Sol_{kl} BAC:Conc_{bk} + +BAC:Solbl + BAC:DB:Conc_{bik} + BAC:DB:Sol_{bil} + BAC:Conc:Sol_{bkl} + BAC:DB:Conc:Sol_{bikl} + e_{biklpr}$$

$BAC_j \sim N(0, \sigma_{BAC})$	$H_0 : \sigma_{BAC} = 0$	$\forall b = 1, B$
$\sum_{i=1}^{I} DB_i = 0$	$H_0 : DB_i = 0$	$\forall i = 1, I$
$\sum_{k=1}^{K} Conc_k = 0$	$H_0 : Conc_k = 0$	$\forall k = 1, K$
$\sum_{l=1}^{L} Sol_l = 0$	$H_0 : Sol_l = 0$	$\forall l = 1, L$
$P(DB)_i \sim N(0, \sigma_P)$	$H_0 : \sigma_P = 0$	$\forall p = 1, P$
$BAC:DB_{bi} \sim N(0, \sigma_{BAC:DB})$	$H_0 : \sigma_{BAC:DB}$	$\forall b = 1, B; i = 1, I$
$\sum_{i}^{I}\sum_{k}^{K} DB:Conc_{ik}$	$H_0 : DB:Conc_{ik} = 0$	$\forall i = 1, I; k = 1, K$
$\sum_{i=1}^{I}\sum_{l=1}^{L} DB:Sol_{il}$	$H_0 : DB:Sol_{il} = 0$	$\forall i = 1, I; l = 1, L$
$\sum_{k=1}^{K}\sum_{l=1}^{L} Conc:Sol_{kl}$	$H_0 : Conc:Sol_{kl} = 0$	$\forall k = 1, K; l = 1, L$
$BAC:Conc_{bk} \sim N(0, \sigma_{BAC:Conc}$	$H_0 : \sigma_{BAC:Conc}$	$\forall b = 1, B; k = 1, K$
$BAC:Sol_{bl} \sim N(0, \sigma_{BAC:Sol})$	$H_0 : \sigma_{BAC:Sol}$	$\forall b = 1, B; l = 1, L$
$\sum_{i=1}^{I}\sum_{k=1}^{K}\sum_{l=1}^{L} DB:Conc:Sol_{ikl}$	$H_0 : DB:Conc:Sol_{ikl} = 0$	$\forall i = 1, I; k = 1, K; l = 1, L$
$BAC:DB:Conc_{bil} \sim N(0, \sigma_{BAC:DB:Conc})$	$H_0 : \sigma_{BAC:DB:Conc}$	$b = 1, B; i = 1, I; k = 1, K$
$BAC:DB:Sol_{bil} \sim N(0, \sigma_{BAC:DB:Sol})$	$H_0 : \sigma_{BAC:DB:Sol}$	$b = 1, B; i = 1, I; l = 1, L$
$BAC:Conc:Sol_{bkl} \sim N(0, \sigma_{BAC:Conc:Sol})$	$H_0 : \sigma_{BAC:Conc:Sol}$	$b = 1, B; k = 1, K, l = 1, L$
$BAC:DB:Conc:Sol_{bikl} \sim N(0, \sigma_{BAC:DB:Conc:Sol})$	$H_0 : \sigma_{BAC:DB:Conc:Sol}$	$b = 1, B; i = 1, I; k = 1, K; l = 1, L$
$e_{biklpr} \sim N(0, \sigma)$	$H_0 : \sigma = 0$	$\forall r = 1, R$

(2)

El modelo ANOVA formulado se resolvió según se presenta en la Tabla ANOVA 9.

Tabla 9: **Tabla ANOVA:**Descomposición de la varianza para condiciones óptimas

Factor	gdl [a]	SS [b]	MS [c]	F [d]
BAC	$(B-1)$	SS_{BAC}	MS_{BAC}	$MS_{BAC}/**$
DB	$(I-1)$	SS_{DB}	MS_{DB}	$MS_{DB}/MS_{P(DB)}$
Conc	$(K-1)$	SS_{Conc}	MS_{Conc}	$MS_{Conc}/MS_{BAC:Conc}$
Sol	$(L-1)$	SS_{Sol}	MS_{Sol}	$MS_{Sol}/MS_{BAC:Sol}$
P(DB)	$(P-1)d$	$SS_{P(DB)}$	$MS_{P(DB)}$	$MS_{P(DB)}/MS_E$
BAC:DB	$(B-1)(I-1)$	$SS_{BAC:DB}$	$MS_{BAC:DB}$	$MS_{BAC:DB}/MS_E$
DB:Conc	$(I-1)(K-1)$	$SS_{DB:Conc}$	$MS_{DB:Conc}$	$MS_{DB:Conc}/MS_{BAC:DB:Conc}$
DB:Sol	$(I-1)(L-1)$	$SS_{DB:Sol}$	$MS_{DB:Sol}$	$MS_{DB:Sol}/MS_{BAC:DB:Sol}$
Conc:Sol	$(K-1)(L-1)$	$SS_{Conc:Sol}$	$MS_{Conc:Sol}$	$MS_{Conc:Sol}/MS_{BAC:Conc:Sol}$
BAC:Sol	$(B-1)(L-1)$	$SS_{BAC:Sol}$	$MS_{BAC:Sol}$	$MS_{BAC:Sol}/MS_E$
BAC:Conc	$(B-1)(K-1)$	$SS_{BAC:Conc}$	$MS_{BAC:Conc}$	$MS_{BAC:Conc}/MS_E$
DB:Conc:Sol	$(I-1)(K-1)(L-1)$	$SS_{DB:Conc:Sol}$	$MS_{DB:Conc:Sol}$	$MS_{DB:Conc:Sol}/MS_{BAC:DB:Conc:Sol}$
BAC:DB:Conc	$(B-1)(I-1)(K-1)$	$SS_{BAC:DB:Conc}$	$MS_{BAC:DB:Conc}$	$MS_{BAC:DB:Conc}/MS_E$
BAC:DB:Sol	$(B-1)(I-1)(L-1)$	$SS_{BAC:DB:Sol}$	$MS_{BAC:DB:Sol}$	$MS_{BAC:DB:Sol}/MS_E$
BAC:Conc:Sol	$(B-1)(K-1)(L-1)$	$SS_{BAC:Conc:Sol}$	$MS_{BAC:Conc:Sol}$	$MS_{BAC:Conc:Sol}/MS_E$
BAC:DB:Conc:Sol	$(B-1)(I-1)(K-1)(L-1)$	$SS_{BAC:DB:Conc:Sol}$	$MS_{BAC:DB:Conc:Sol}$	$MS_{BAC:DB:Conc:Sol}/MS_E$
Residuo	$(nobs-1)-\sum gdl$	SS_E	MS_E	
Total	$(nobs-1)$			

[a]: (gdl) grados de libertad
[b]: (SS) Suma de Cuadrados
[c]: (MS) Cuadrados Medios
[d]: (F) Estadístico F de Fisher
$**$: Cuasi F-Ratios

Aplicación del Modelo ANOVA y resultados derivados

En este estudio, las muestras hibridadas fueron el *pool* de referencia femenino (en HD con Cy5) contra el pool de referencia masculino (en HD con Cy3).

Con la intención de evitar heterocedasticidad se analizaron los portaobjetos Ultragaps y Codelink en modelos ANOVA distintos. A su vez, se estudiaron los BACs situados en cromosomas autosómicos y los BACs situados en cromosomas sexuales por separado. Por ello se realizaron cuatro modelos ANOVA formulados a partir del mismo modelo (ver ecuación 7) y resueltos según Tabla ANOVA 9.

En los portaobjetos tipo Codelink hay una importante perdida de *spots* cuando se utilizan las soluciones 50 %DMSO, PRONTO-Amino y PRONTO-Epoxy como puede observarse en la Figura 17. Ello sugiere que no deberían emplearse estas soluciones sobre este tipo de portaobjetos. Así, antes de proceder al análisis de los datos, se eliminó, en los modelos Codelink, las soluciones 50 %DMSO, PRONTO-Amino y PRONTO-Epoxy.

En los modelos Ultragap no se eliminó ninguna solución pero la pérdida del bloque F (ver Figura 18) obligó a realizar dos modelos distintos; (i) en un modelo se consideraron todas las soluciones pero no se consideró el nivel 400ng/μl de concentración de ADN impreso mientras que (ii) en el segundo modelo se consideraron todas las concentraciones de ADN impreso pero no se consideró la solución PRONTO-Epoxy.

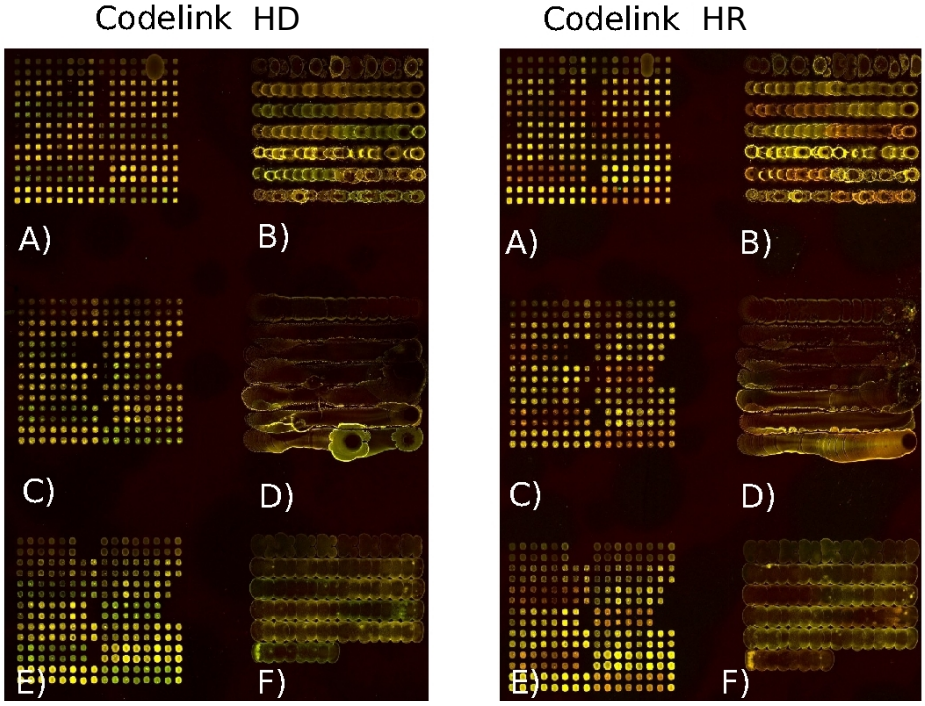

Figura 17: **Imágenes de los portaobjetos utilizados para Codelink**: A la derecha la imagen de la hibridación directa (HD) y a la izquierda la hibridación reversa (HR). Las soluciones utilizadas fueron A) Schott-Epoxy B) 50 %DMSO C)3xSSC D) PRONTO-Amino E) 150mM Phosphate F) PRONTO-Epoxy

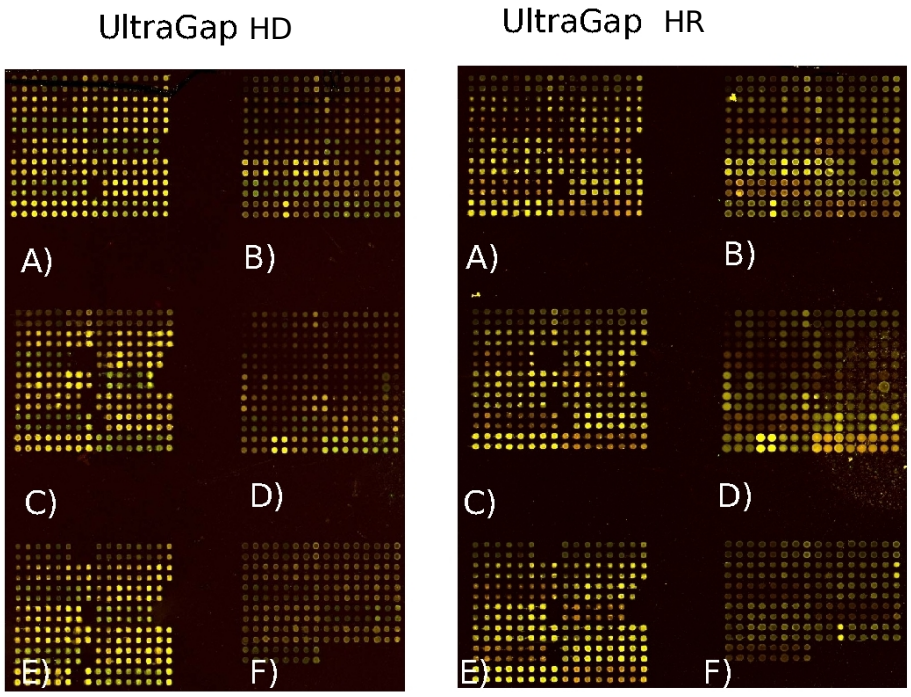

Figura 18: **Imágenes de los portaobjetos utilizados para Ultragaps**: A la derecha la imagen de la hibridación directa (HD) y a la izquierda la hibridación reversa (HR). Las soluciones utilizadas fueron A) schott epoxy B) 50 %DMSO C)3xSSC D) PRONTO-Amino E) 150mM Phosphate F) PRONTO-Epoxy

En la Tabla 10 se presentan los resultados obtenidos para Ultragaps en cromosomas autosómicos sin la solución de impresión PRONTO-Epoxy. Según se observa en la Tabla 10 hay diferencias significativas entre la concentración de ADN empleada y en la solución de resuspensión. Para establecer cuales eran los niveles que se asemejaban más a las condiciones óptimas se han realizado pruebas post-hoc utilizando el criterio de Tukey-Scheffé. Así, se observa que las principales diferencias en los niveles de concentración de ADN impreso se hallan entre los niveles bajos (100-150ng/μl, no hay diferencias significativas entre ellos) versus los niveles altos (200-400ng/μl, no hay diferencias significativas entre ellos).En la Figura 19 A) puede apreciarse que cuando se incrementa la concentración de ADN impreso más cercana es la variable respuesta a los valores esperados (cero en este caso). Las principales diferencias entre las soluciones se hallan en PRONTO-Amino versus 50 %DMSO y 150 mM Phosphate siendo PRONTO-Amino la menos apropiada tal y como puede observarse en la Figura 19 A).

Tabla 10: **Resultados obtenidos con Ultragap de Corning para BACs en cromosomas autosómicos**. Sin la solución PRONTO-Epoxy.

Factor	gdl[a]	SS[b]	MS[c]	F[d]	pvalor
DB	1	0,0007	0,001	0,111	0,750
BAC	13	0,131	0,010	3,036	0,000
Conc	3	0,050	0,017	3,971	0,015
Sol	4	0,105	0,026	5,184	0,001
P(DB)	10	0,038	0,004	1,267	0,243
BAC:DB	13	2,086	0,160	48,404	<0,001
DB:Conc	3	0,458	0,153	14,371	<0,001
DB:Sol	4	0,072	0,018	2,754	0,038
Conc:Sol	12	0,095	0,008	1,365	0,189
BAC:Conc	39	0,163	0,004	1,257	0,131
BAC:Sol	52	0,264	0,005	1,532	0,008
DB:Conc:Sol	12	0,209	0,017	2,323	0,009
BAC:DB:Conc	39	0,414	0,011	3,204	<0,001
BAC:DB:Sol	52	0,339	0,007	1,965	<0,001
BAC:Conc:Sol	144	0,836	0,006	1,751	<0,001
BAC:DB:Conc:Sol	138	1,033	0,007	2,257	<0,001
Residuo	5614	18,610	0,003		

[a]: (gdl) grados de libertad
[b]: (SS) Suma de Cuadrados
[c]: (MS) Cuadrados Medios
[d]: (F) Estadístico F de Fisher

En un subsiguiente modelo, realizado en Ultragaps sobre BACs situados en cromosomas autosómicos, se consideraron todas las soluciones sin la concentración de 400ng/μl. Los resultados fueron similares a los que se obtuvieron mediante el modelo anterior y se observó un comportamiento parecido a 3xSSC por parte de la solución PRONTO-Epoxy.

En la Tabla 11 se presentan los resultados para el portaobjetos Ultragaps considerando BACs situados en cromosomas sexuales sin la concentración de 400ng/μl. Según el criterio de Tukey-Scheffeé (aplicado únicamente para aquellos factores que resultaron significativos según se muestra en la Tabla ANOVA 11) la solución menos apropiada es PRONTO-Epoxy (ver Figura 19 B). Y ello es debido a que toma valores alejados de los esperados (el valor esperado es en este caso -1) mientras que las más apropiadas son 50 %DMSO, 150mM Phosphate y 3xSSC, siendo 50 %DMSO la más indicada con diferencias significativas respecto al resto.

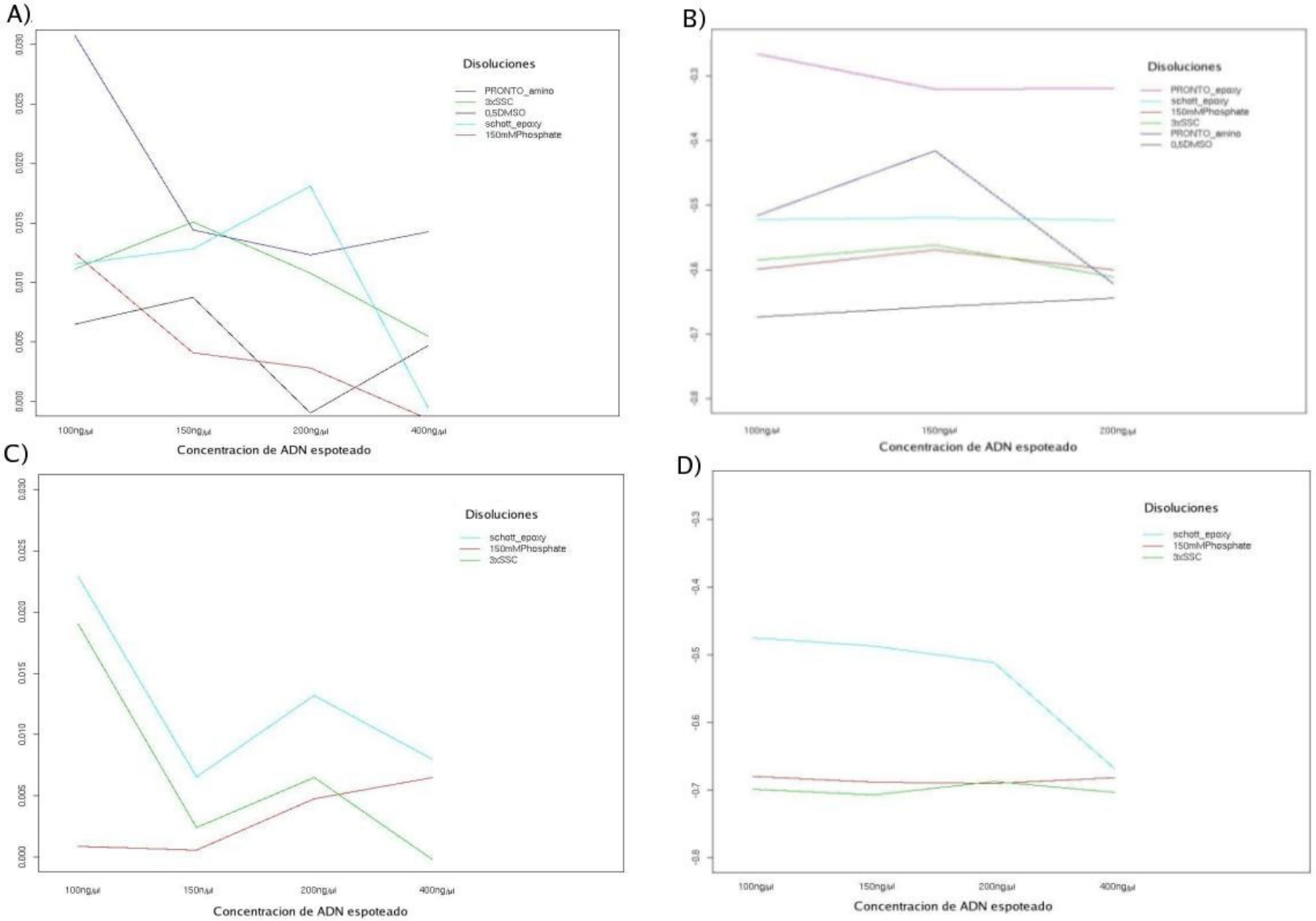

Figura 19: **Gráficos de interacción** para Ultragaps (arriba) y Codelink (abajo): A) Solución versus ADN impreso en Ultragaps y cromosomas autosómicos B) Solución versus ADN impreso en Ultragaps y cromosomas sexuales C) Solución versus ADN impreso en Codelink y cromosomas autosómicos D) Solución versus ADN impresos en Codelink y cromosomas sexuales

Tabla 11: **Resultados obtenidos con Ultragap de Corning para BACs en cromosomas sexuales.** Sin la concentración de 400ng/μl

Factor	gdl[a]	SS[b]	MS[c]	F[d]	pvalor
DB	1	1,652	1,652	2,597	0,158
BAC	7	11,288	1,613	141,009	<0,001
Conc	2	0,664	0,332	22,739	<0,001
Sol	5	31,658	6,331	43,809	<0,001
P(DB)	10	3,816	0,382	34,727	<0,001
BAC:DB	7	0,219	0,031	2,735	0,008
DB:Conc	2	0,215	0,108	3,590	0,055
DB:Sol	5	3,597	0,719	21,452	<0,001
BAC:Conc	14	0,205	0,015	1,278	0,213
BAC:Sol	35	5,058	0,144	12,638	<0,001
Conc:Sol	10	2,077	0,208	1,175	0,325
DB:Conc:Sol	10	0,320	0,032	1,194	0,320
BAC:DB:Conc	14	0,420	0,030	2,622	<0,001
BAC:DB:Sol	34	1,140	0,034	2,933	<0,001
BAC:Conc:Sol	61	10,782	0,177	15,456	<0,001
BAC:DB:Conc:Sol	47	1,268	0,027	2,341	<0,001
Residuo	2387	27,298	0,011		

[a]: (gdl) grados de libertad
[b]: (SS) Suma de Cuadrados
[c]: (MS) Cuadrados Medios
[d]: (F) Estadístico F de Fisher

En los modelos realizados para el portaobjetos Codelink se emplearon sólo tres soluciones (3xSCC, 150mM Phophate y Schott Eppoxy) debido a la péridida de *spots* Figura 17.

Según se muestra en la Tabla 12, en portaobjetos Codelink sobre BACs situados en cromosomas autosómicos, no se observaron diferencias significativas entre las soluciones empleadas pero sí se detectaron diferencias entre las concentraciones de 100ng/μl y 400ng/μl de ADN impreso. Según se observa en la Figura 19 C) la concentración de 400ng/μl de ADN impreso toma valores más cercanos a cero.

En el modelo ANOVA realizado en portaobjetos Codelink sobre BACs situados en cromosomas sexuales (ver Tabla 13) se observaron diferencias significativas entre las soluciones empleadas siendo Schott Epoxy la que obtuvo peores resultados. En cuanto a la concentración de ADN impreso se obtuvieron los mismos resultados que en el modelo realizado sobre cromosomas autosómicos.

Se consideró 50 %DMSO la solución con mejores resultados para Ultragaps mientras que para Codelink se consideró 3xSSC como la solución con mejores resultados. Como concentración óptima de ADN impreso se consideró 400ng/μl en ambos casos. Se evaluó el efecto de estas dos combinaciones sobre la variabilidad asociada a los portaobjetos y sobre el DB en dos modelos independientes. El modelo aplicado es equivalente al formulado en la ecuación 7 sin considerar los efectos Sol y Conc. Los portaobjetos Ultragaps de Corning con 50 %DMSO presentaban un menor efecto DB (efecto no significativo) y una menor

variabilidad que los portaobjetos de Codelink con 3xSSC.

Tabla 12: **Resultados obtenidos con Codelink para BACs en cromosomas autosómicos.**

Factor	gdl[a]	SS[b]	MS[c]	F[d]	pvlor
DB	1	0,017	0,017	2,110	0,197
BAC	13	0,123	0,009	4,163	<0,001
Conc	3	0,068	0,023	3,748	0,019
Sol	2	0,064	0,032	3,185	0,058
P(DB)	10	0,048	0,005	3,500	0,005
BAC:DB	13	1,814	0,140	61,574	<0,001
DB:Conc	3	0,231	0,077	8,967	<0,001
DB:Sol	2	0,032	0,016	2,960	0,069
BAC:Conc	39	0,236	0,006	2,667	<0,001
BAC:Sol	26	0,261	0,010	4,442	<0,001
Conc:Sol	6	0,065	0,011	0,252	0,957
DB:Conc:Sol	6	0,541	0,090	33,327	< 0,001
BAC:DB:Conc	39	0,334	0,009	3,783	<0,001
BAC:DB:Sol	26	0,140	0,005	2,383	<0,001
BAC:Conc:Sol	75	3,241	0,043	19,070	<0,001
BAC:DB:Conc:Sol	73	0,197	0,003	1,194	0,127
Residuo	3545	8,034	0,002		

[a]: (gdl) grados de libertad
[b]: (SS) Suma de Cuadrados
[c]: (MS) Cuadrados Medios
[d]: (F) Estadístico F de Fisher

Tabla 13: **Resultados obtenidos con Codelink para BACs en cromosomas sexuales.**

Factor	gdl[a]	SS[b]	MS[c]	F[d]	pavlor
P(DB)	10	3,809	0,381	38,090	<0,001
DB	1	10,174	10,174	16,020	0,007
BAC	7	19,056	2,722	286,606	<0,001
Conc	3	2,191	0,731	14,571	<0,001
Sol	2	10,930	5,465	26,057	<0,001
BAC:DB:	7	0,690	0,099	10,372	<0,001
BAC:Conc	20	1,003	0,050	5,279	<0,001
BAC:Sol	14	2,936	0,210	22,081	<0,001
DB:Conc	3	0,291	0,097	7,815	0,001
DB:Sol	2	0,106	0,053	2,699	0,102
Conc:Sol	6	2,857	0,476	11,794	<0,001
DB:Conc:Sol	6	0,536	0,089	7,997	<0,001
BAC:DB:Conc	20	0,248	0,012	1,308	0,162
BAC:DB:Sol	14	0,276	0,020	2,077	0,011
BAC:Conc:Sol	35	1,413	0,040	4,251	<0,001
BAC:DB:Conc:Sol	35	0,391	0,011	1,177	0,221
Residuo	1848	17,553	0,010		

[a]: (gdl) grados de libertad
[b]: (SS) Suma de Cuadrados
[c]: (MS) Cuadrados Medios
[d]: (F) Estadístico F de Fisher

4.2.2. Detección de fuentes de variación asociadas a la fiabilidad de la medida

Este apartado se divide en dos partes;(i) formulación y desarrollo del modelo matemático utilizado y (ii) los resultados obtenidos de la aplicación de dicho modelo.

Formulación y desarrollo del modelo ANOVA

En la ecuación 11 se presenta el modelo ANOVA resuelto en esta sección.

$$y_{bijksr} = \mu_b + DB_{bi} + Tecnico_{bt} + Dia_{ba} + P\left(DB, Tecnico, Dia\right)_{bp(ita)} + e_{bitapr}$$

$$\begin{array}{lll}
\sum_{i=1}^{I} DB_{bi} = 0 & H_0 : DB_{bi} = 0 & \forall i = 1, I \\
Tecnico_{tb} \sim N(0, \sigma_T) & H_0 : \sigma_T = 0 & \forall t = 1, T \\
Dia_{ab} \sim N(0, \sigma_D) & H_0 : \sigma_D = 0 & \forall a = 1, A \\
P_{p(ita)b} \sim N(0, \sigma_P) & H_0 : \sigma_P = 0 & \forall p = 1, P \\
e_{itaprb} \sim N(0, \sigma) & H_0 : \sigma = 0 & \forall r = 1, R
\end{array}$$

El modelo se resolvió según la Tabla ANOVA 14.

Tabla 14: **Tabla ANOVA:** Descomposición de la varianza para la fiabilidad de la medida

Factor	gdl [a]	SS [b]	MS [c]	F [d]
DB	$(I-1)$	SS_{DB}	MS_{DB}	MS_{DB}/MS_P
Tecnico	$(T-1)$	SS_T	MS_T	MS_T/MS_P
Dia	$(A-1)$	SS_D	MS_D	MS_D/MS_P
P	$I*T*A*(P-1)$	SS_P	MS_P	MS_P/MS_R
Residuo	$[I*T*A*P*R-1]-$ $[(I-1)+(T-1)+(A-$ $1)+(I*T*A*(P-1))]$	SS_R	MS_R	
Total	$I*T*A*P*R-1$			

[a]:(gdl) grados de libertad
[b]:(SS) Suma de Cuadrados
[c]: (MS) Cuadrados Medios
[d]: (F) Estadístico F de Fisher

Las esperanzas de los cuadrados medios (ver ecuación 11), las sumas de cuadrados (ver ecuación 9), los cuadrados medios (ver ecuación 10), grados de libertad y F-ratios fueron deducidas aplicando el algoritmo de Benett-Franklin sobre el modelo completo (ver anexo).

Aplicación del modelo ANOVA y resultados derivados

En este estudio se hibridaron 32 réplicas técnicas (en 32 portaobjetos) de la misma muestra en distintas condiciones según el modelo descrito en la ecuación 11 (experimento 32x).

En primer lugar se realizó un análisis descriptivo de cada portaobjetos sobre los valores de M según el tipo de normalización. En la Tabla 15 se han representado el promedio, el máximo y el mínimo de la desviación típica y MAD obtendidos de los 32 portaobjetos utilizados. En esta tabla se puede observar que el método de normalización *loess* obtiene valores más altos de dispersión que el resto y que, en general, la no substracción del ruido de fondo minimiza la variabilidad asociada a cada portaobjetos.

Tabla 15: **Variabilidad asociada a cada portaobjetos;** para el experimento 32x según el tipo de normalización utilizado.

Estadístico descriptivo		Loess		Print-tip loess		Loess loc		Loess loc scale	
Substracción BG [a]		SI	NO	SI	NO	SI	NO	SI	NO
Desviación estándar (sd)	media	0,094	0,085	0,077	0,077	0,092	0,076	0,084	0,075
	min	0,077	0,069	0,063	0,063	0,075	0,064	0,070	0,063
	max	0,121	0,1116	0,108	0,108	0,127	0,104	0,1233	0,108
Median absolute deviance (mad)	media	0,059	0,057	0,048	0,048	0,060	0,045	0,047	0,045
	min	0,047	0,044	0,039	0,039	0,047	0,036	0,038	0,037
	max	0,084	0,082	0,072	0,071	0,085	0,067	0,069	0,067

[a]: Ruido de fondo o background

Los resultados obtenidos del modelo ANOVA indican que los factores más influyentes en la fiabilidad de la medida son el portaobjetos y el efecto DB siendo prácticamente despreciables las variabilidades aportadas por el efecto *día y técnico* (ver Tabla 16) ya que el número de veces que éstos resultaron significativos es parecido al error tipo I esperado (por ejemplo un error tipo I de 0,05, equivalente al utilizado, indica que aproximadamente un 5 % de los tests aplicados pueden ser significativos por azar). Si se analiza únicamente qué factores y en qué proporción resultaron significativos (Tabla 16) no se observan grandes diferencias entre los métodos de normalización propuestos.

Tabla 16: **Número de tests ANOVA significativos** según las fuente de variabilidad testadas.

	Loess		Print-tip loess		Loess loc		Loess loc scale	
Substracción BG [a]	SI	NO	SI	NO	SI	NO	SI	NO
P	4436 (89 %)	4471 (90 %)	4250 (85 %)	4231 (85 %)	4187 (84 %)	4214 (85 %)	4104 (82 %)	4151 (83 %)
DB	3602 (72 %)	3577 (72 %)	3667 (74 %)	3660 (73 %)	3604 (72 %)	3602 (72 %)	3634 (73 %)	3641 (73 %)
Día	629 (13 %)	676 (14 %)	856 (17 %)	862 (17 %)	956 (19 %)	920 (18 %)	993 (20 %)	967 (19 %)
Técnico	83 (2 %)	109 (2 %)	166 (3 %)	169 (3 %)	165 (3 %)	189 (4 %)	162 (3 %)	170 (3 %)
Total	4983	4983	4983	4983	4983	4983	4983	4983

[a]: Substracción del ruido de fondo o *background*

En cambio, sí que se observan diferencias sobre la estimación de la magnitud de los efectos que resultaron significativos entre los métodos de normalización. En la Figura 20

se observa que el método de normalización *loess* obtiene valores absolutos de DB mayores que el resto de métodos de normalización y que este efecto tiende a ser menor cuando no se realiza substracción del ruido de fondo.

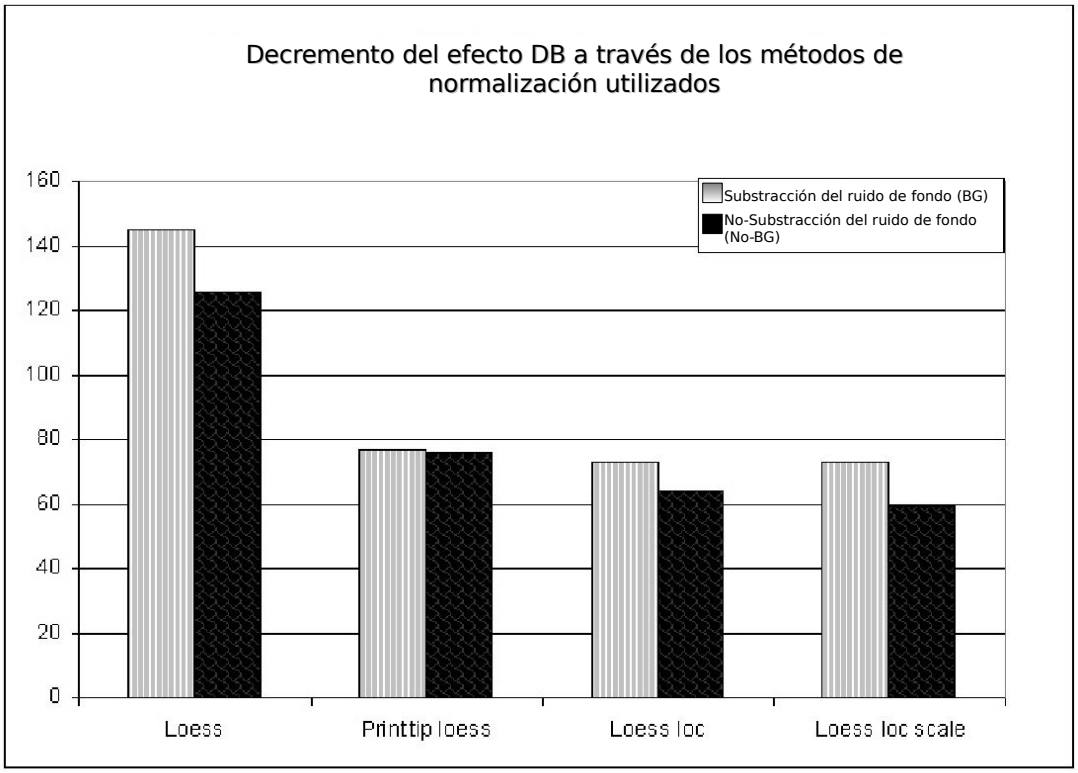

Figura 20: **Decremento del DB a través de los procesos de normalización**. En el eje de ordenadas se representa el número de BACs con valores de $|DB|$ >0,2. En el eje de abcisas se representan los distintos tipos de normalización.

Se utilizó la librería *GeoR* [**?**] para representar la magnitud del DB obtenida por cada método de normalización en las coordenadas reales de un portaobjeto, ver Figura 21 y 22.

En la Figura 21 puede observarse la presencia de una zona interior con valores altos de DB (una magnitud positiva indica tendencia a captar el fluorocromo Cy3) y la presencia de una región periférica con valores más bajos (valores negativos indican una tendencia a captar el fluorocromo Cy5). Del mismo modo que en la Figura 20, puede apreciarse como los métodos *print-tip loess, loess loc y loess loc scale* corrigen mejor este efecto que *loess* siendo *loess loc scale* el más eficaz.

En este estudio, la realización de la substracción incrementa la variabilidad (Tabla 15) y la magnitud del efecto DB (Figura 20). Ello pone de manifiesto que no debe existir una correlación entre éste y la señal del *spot*. En la Figura 23 se ha representado la mediana de la intensidad de la señal del ruido de fondo y la señal del *spot* para cada canal. En ellas puede observarse como el ruido de fondo toma un marcado efecto espacial que no se observa en la señal del *spot* (FG o *Foreground*) indicando que no hay correlación entre ambas variables. Si se comparan la Figura 21 y 23 puede verse que efecto espacial del

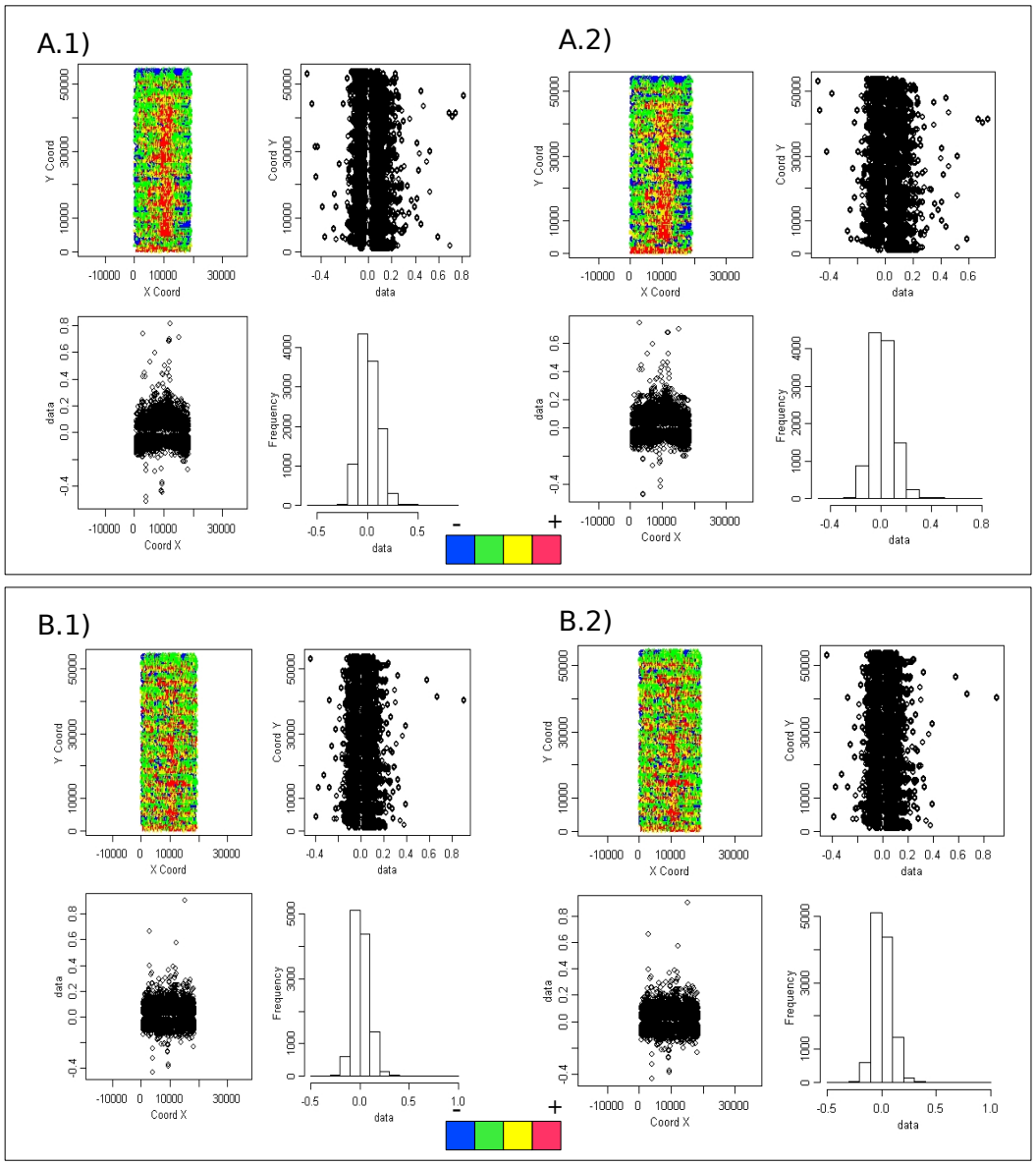

Figura 21: **Distribución de la magnitud del DB en el experimento 32x**; A.1) para el procedimiento *loess* con substracción del ruido de fondo y A.2) sin substracción del ruido de fondo. B.1) el procedimiento *print-tip loess* con substracción del ruido de fondo B.2) y sin substracción del ruido de fondo.

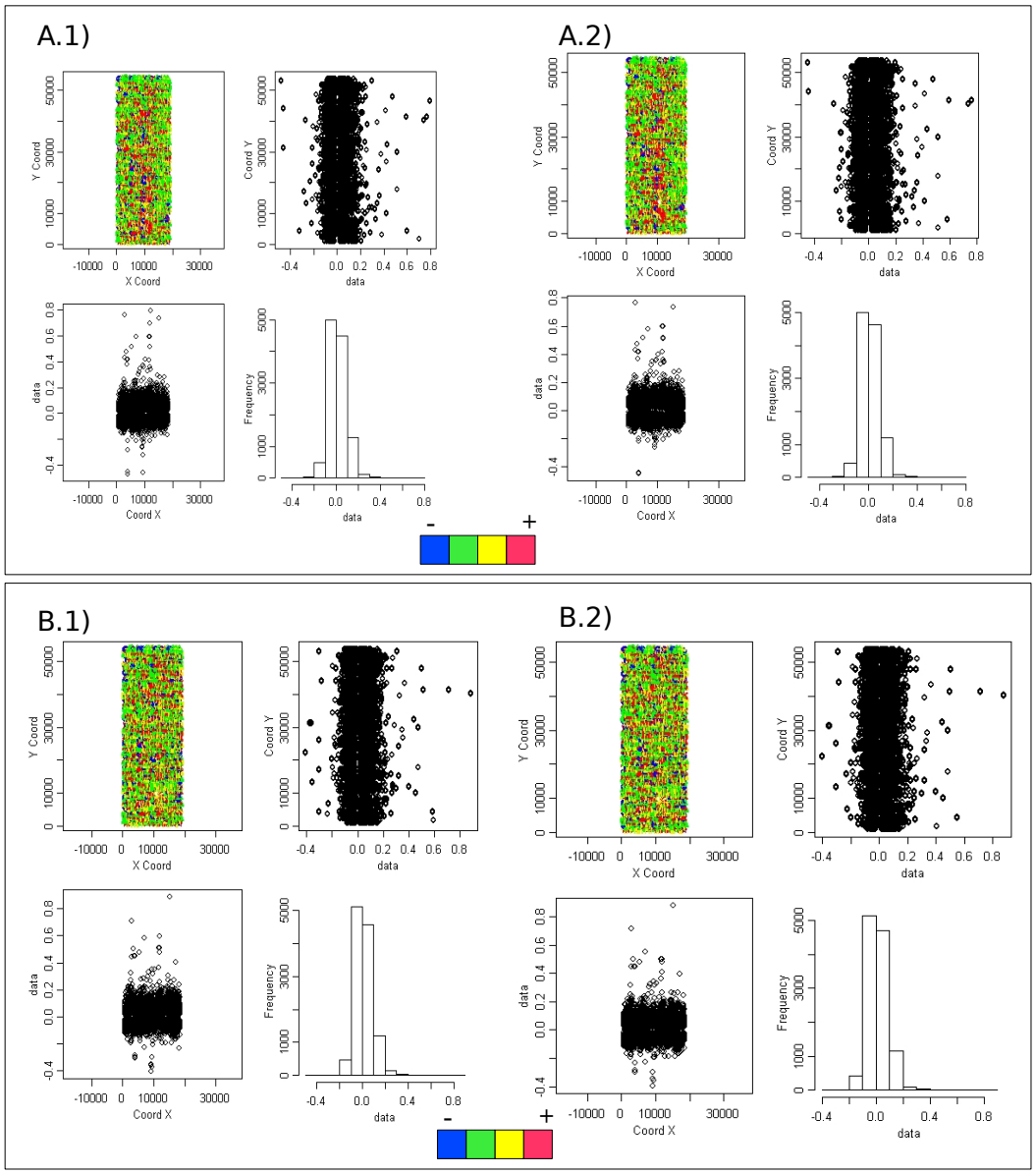

Figura 22: **Distribución de la magnitud del DB en el experimento 32x**; A.1) el procedimiento *loess loc* con substracción del ruido de fondo y A.2) sin substracción del ruido de fondo. B.1) el procdimiento *loess loc scale* con substracción del ruido de fondo y B.2) sin substracción del ruido de fondo.

ruido de fondo tiene una estructura espacial parecida al efecto DB pero la substracción en vez de minimizar el efecto DB lo incrementa.

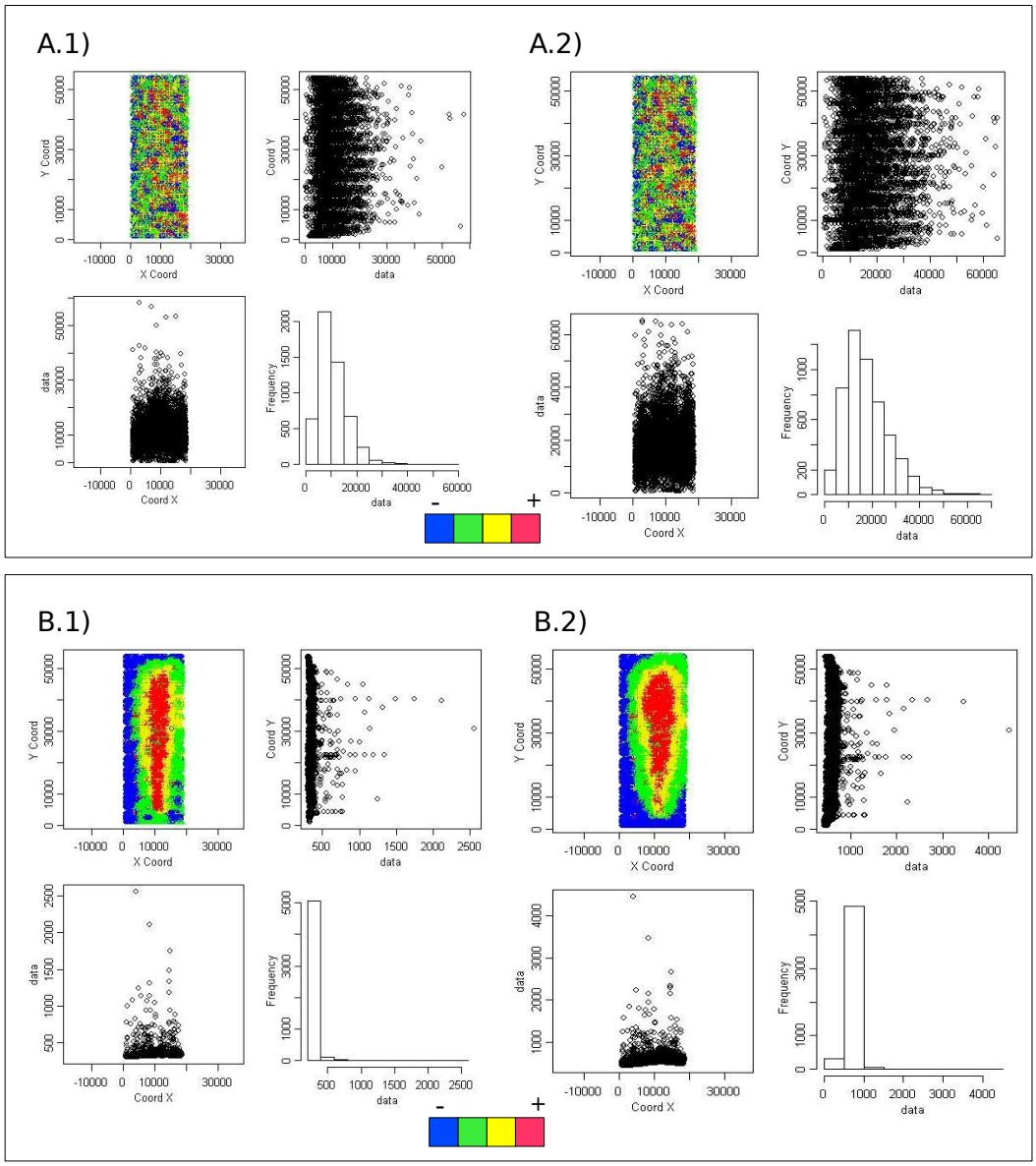

Figura 23: **Distribución de la mediana FG o señal del *spot* y del ruido de fondo en el experimento 32x**. A.1) FG para el canal Cy5, A.2) FG para el canal Cy3, B.1) BG para el canal Cy5 y B.2) BG para el canal Cy3.

4.3. Estudio del efecto DB y sus causas

4.3.1. El efecto DB como error sistemático. Detección del efecto DB en otros experimentos

Este apartado se divide en dos partes;(i) formulación y desarrollo del modelo matemático y (ii) los resultados obtenidos de aplicar dicho modelo.

Formulación y desarrollo del modelo ANOVA

$$
\begin{aligned}
& y_{bimur} = \mu_b + DB_{bi} + Muestra_{bm} + Genero_{bu} + e_{bimur} \\
& \sum_{i=1}^{I} DB_{bi} = 0 \qquad H_0 : DB_{bi} = 0 \qquad \forall i = 1, I \\
& \sum_{m=1}^{M} Muestra_{bm} = 0 \quad H_0 : Muestra_{bm} = 0 \quad \forall m = 1, M \\
& \sum_{u=1}^{U} Genero_{bu} = 0 \qquad H_0 : Genero_{bu} = 0 \qquad \forall u = 1, U \\
& e_{bimur} \sim N(0, \sigma) \qquad\qquad\qquad\qquad \forall r = 1, R
\end{aligned}
\tag{3}
$$

El modelo se resolvió según se indica en la Tabla ANOVA 17.

Tabla 17: **Tabla ANOVA:** Descomposición de la varianza para el estudio del error sistemático

Factor	gdl[a]	SS[b]	MS[c]	F[d]
DB	$(I-1)$	SS_{DB}	MS_{DB}	MS_{DB}/MS_R
Muestra	$(M-1)-(U-1)$	$SS_{Muestra}$	$MS_{Muestra}$	$MS_{Muestra}/MS_R$
Genero	$(U-1)$	SS_{Genero}	MS_{Genero}	MS_{Genero}/MS_R
Residuo	$[I*M*U*R-1]-[(I-1)+(M-1)-(U-1)+(U-1)]$	SS_R	MS_R	
Total	$I*M*U*R-1$			

[a]: (gdl) grados de libertad [c]: (MS) Cuadrados Medios
[b]: (SS) Suma de Cuadrados [d]: (F) Estadístico F de Fisher

Las estimaciones de los cuadrados medios (ver ecuación 13) así como las sumas de cuadrados (ver ecuación 14) se representan a continuación.

Aplicación del Modelo ANOVA y resultados derivados

El modelo formulado en la ecuación 11 se aplicó sobre 19 muestras de controles sanos (10 de género masculino y 9 de género femenino). Cada muestra se hibridó dos veces (HD y HR) por lo que se realizaron un total de 38 hibridaciones en 38 portaobjetos distintos (experimento 19c).

Se evaluó la variabilidad asociada a cada portaobjetos según el tipo de normalización y se obtuvieron resultados equivalentes a los presentados en el apartado anterior (ver Tabla 18).

Tabla 18: **Variabilidad asociada a cada portaobjetos;** para el experimento 19c (38 hibridaciones).

Estadístico descriptivo		Loess	Print-tip loess	Loess loc	Loess loc scale
Desviación estándar (sd)	media	0,068	0,063	0,064	0,064
	min	0,057	0,052	0,053	0,052
	max	0,087	0,079	0,080	0,081
Median absolute deviance (mad)	media	0,049	0,044	0,043	0,043
	min	0,043	0,038	0,037	0,037
	max	0,058	0,052	0,051	0,051

El efecto DB resultó ser la principal fuente de variabilidad. Se compararon las estimaciones obtenidas para el efecto DB en este experimento con las estimaciones obtenidas en el experimento 32x con la finalidad de evaluar si el efecto observado puede considerarse sistemático. Los principales resultados se resumieron en la Tabla 19.

Tabla 19: **Resumen de los principales resultados del modelo para el experimento 19c** y su correlación con el experimento 32x.

Proceso de normalización	N	Grupo Control 19c			Resultados en común 32x 19c						
		N Sign [a]	abs(DB) >0,1	abs(DB) >0,2	N Sign [a]	Spear-man	R^2	PO[b]	Pend[c]	abs(DB) >0,1	abs(DB) >0,2
Loess	4940	4262	900	119	3443	0,79	0,64	0,004	0,73	481	59
Print-tip	4940	4279	730	81	3549	0,85	0,70	0,003	0,80	420	32
Loess loc	4940	4324	750	89	3477	0,85	0,72	0,003	0,87	388	29
Loess loc scale	4940	4334	728	88	3512	0,85	0,71	0,003	0,86	363	32

[a]:Número de test ANOVA con el efecto DB significativo

[b]:Punto de ordenadas

[c]:Pendiente

Existe una buena correlación entre ambas estimaciones ($R^2 > 0{,}70$) indicando que el efecto observado puede ser sistemático. Además, la estimación realizada por regresión lineal es similar a la recta teórica esperada en presencia de error sistemático; con punto de ordenadas en el cero y pendiente en el uno (ver Tabla 19).

Para determinar si DB es una fuente de variabilidad importante en otros laboratorios se buscaron datos comparables a los obtenidos en nuestro laboratorio en la base de datos pública GEO. Se encontraron tres sets de datos en los que se había utilizado matrices aCGH basadas en BACs y en los cuales se hubiese realizado hibridaciones de tipo HD y HR. Pero solamente en dos de estos sets de datos se pudieron obtener los datos crudos en un formato similar al nuestro para ser analizado. Estos datos fueron preprocesados del mismo modo que los datos del experimento 19c y se les aplicó el modelo descrito en la ecuación 11. En la Figura 24 se representa las variabilidades residuales obtenidas por estos modelos. En el set de datos GEO2 se aprecia un residuo muy grande que hace que estos datos no puedan ser comparados con los datos obtenidos en el CRG. Este residuo puede ser debido a la presencia de otras fuentes de variabilidad no controladas en este modelo como pudiera ser una baja calidad del ADN obtenido (i.e posible contaminación proteíca).

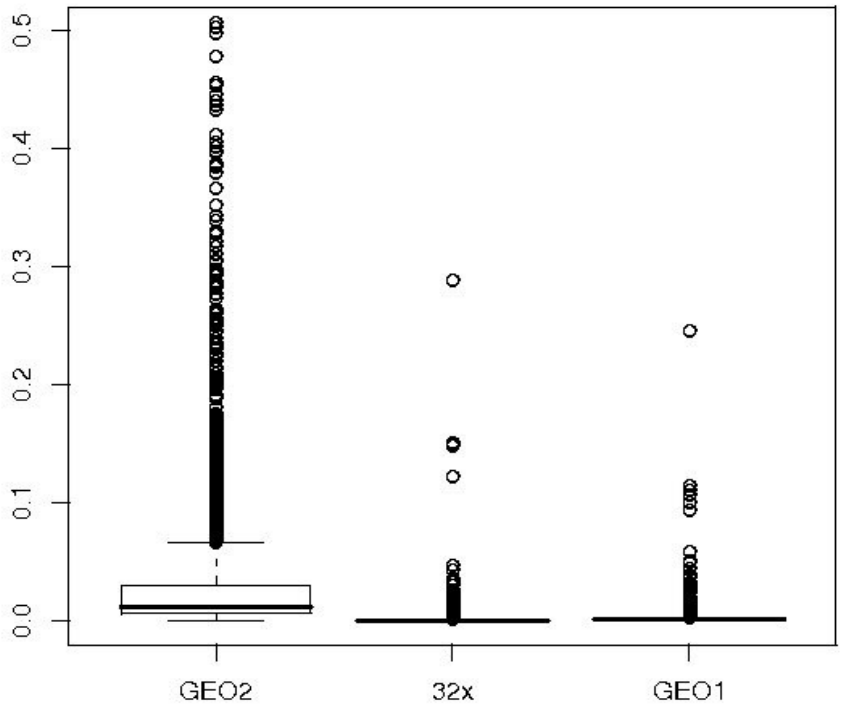

Figura 24: **Variabilidad residual de los experimentos**

Se estudió únicamente el set de datos GEO1 al cual nos referiremos como 47m. Se analizaron 43 muestras (con 86 portaobjetos asociados) de este experimento. De igual modo, se evaluó el efecto de los distintos tipos de normalización sobre la variabilidad del portaobjetos (ver Tabla 20) obteniendo resultados similares a los anteriores. En estos datos puede apreciarse mayores variabilidades asociadas a los portaobjetos que en el experimento 32x y que en el experimento 19c.

Tabla 20: **Variabilidad asociada a cada portaobjetos;** para el experimento 47m según el tipo de normalización utilizado.

Estadístico descriptivo		Loess	Print-tip loess	Loess loc	Loess loc scale
Desviación estándar (sd)	media	0,211	0,205	0,206	0,208
	min	0,148	0,140	0,145	0,145
	max	0,322	0,316	0,318	0,322
Median absolute deviance (mad)	media	0,108	0,095	0,095	0,094
	min	0,046	0,040	0,040	0,041
	max	0,239	0,222	0,226	0,222

Finalmente, se compararon las estimaciones obtenidas para el efecto DB en el experimento 47m con las estimaciones obtendidas en el experimento 32x y, aunque el efecto DB resultó igualmente significativo en los modelos ANOVA, las estimaciones de la magnitud del efecto DB fueron bastante menores que las obtenidas en el experimento 32x y 19c. En el experimento 47m la magnitud del efecto DB tomó valores entre -0,2 y 0,15 mientras que en los experimentos 32x y 19c la magnitud del efecto DB tomó valores entre -0,4 y 0,8. No se halló una clara correlación entre las magnitudes del DB entre el experimento 47m y 32x pero sí se aprecia una tendencia similar cuando este se categoriza (ver Figura 25, el efecto DB puede categorizarse en tres niveles; DB 0 (afinidad al fluorocromo Cy3), DB $>$ 0 (afinidad al fluorcromo Cy5) y sin efecto DB.

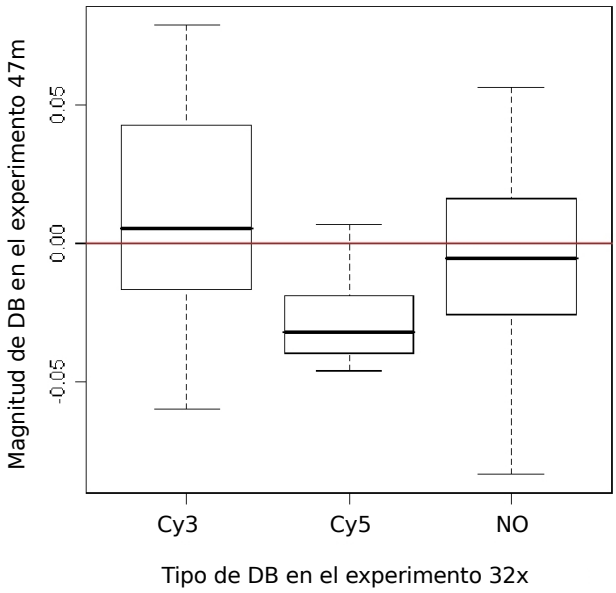

Figura 25: **Correlación del efecto DB entre el experimento 47m y 32x**

Se observó una estructura espacial similar a las obtenidas por los experimentos 32x y 19c de la magnitud DB sobre el portaobjetos cuando se aplicaba el método de normalización *loess*. Del mismo modo, la aplicación del método *print-tip loess* minimiza dicho efecto. Ello puede apreciarse en la Figura 26

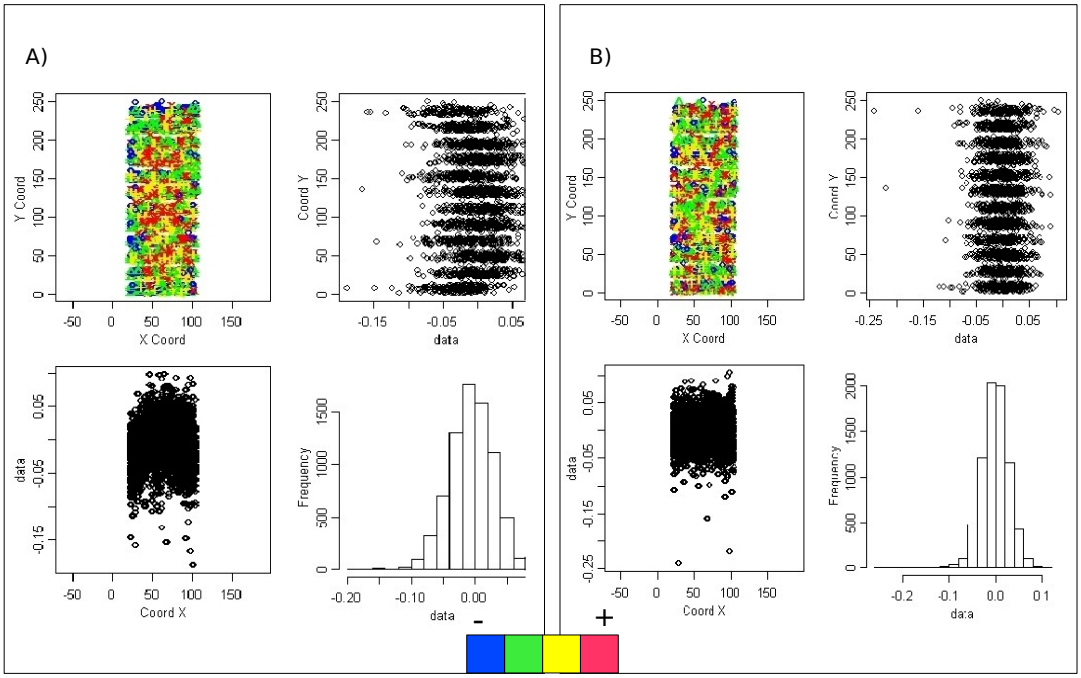

Figura 26: **Distribución de la magnitud del DB en el experimento 47m**. A) método de normalización *loess* sin substracción del ruido de fondo B) *print-tip loess* sin substracción del ruido de fondo.

La Figura 27 se muestra una estructura espacial similar a la observada anteriormente para la señal del *spot* y el ruido de fondo y no hay correlación entre estas dos medidas. En esta figura, además, puede observarse que las intensidades obtenidas por el ruido de fondo y por la señal del *spot* son menores que las obtenidas en el experimento 32x (ver Figura 23). La relación entre estas intensidades entre el set 32x y 47m es similar a la relación obtenida para la magnitud del DB; el ruido de fondo tiene una mediana alrededor de 500 en el experimento 32x mientras que en el experimento 47m tiene una mediana alrededor de 300 y la señal del *spot* tiene una mediana alrededor de 2000 en el experimento 32x mientras que en el experimento 47m está alrededor de 1000.

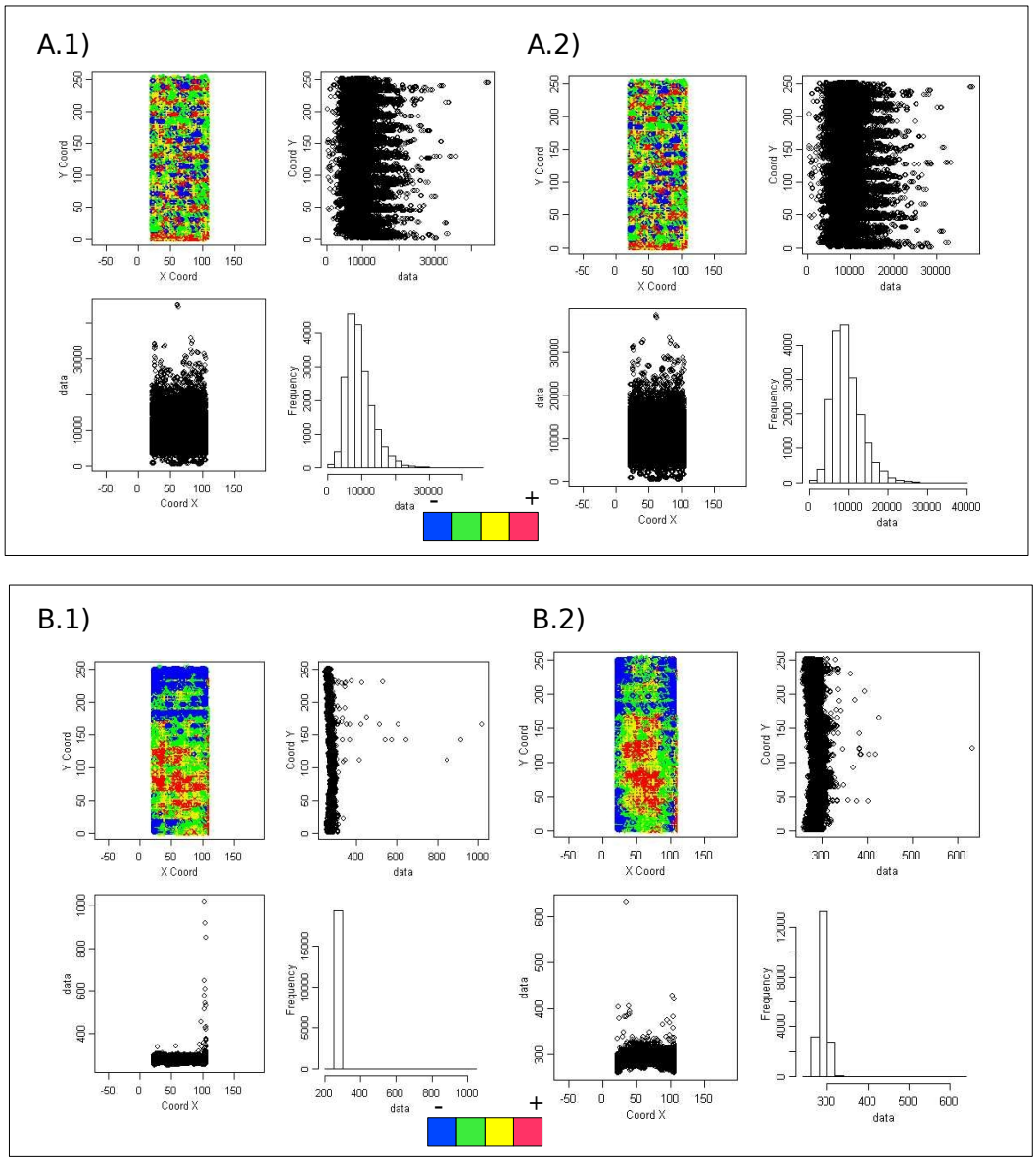

Figura 27: **Distribución espacial de la mediana de la intensidad de los fluorocromos en el experimento 47m.** A.1) FG para el canal Cy5, A.2) FG para el canal Cy3, B.1) BG para el canal Cy5 y B.2) BG para el canal Cy3.

4.3.2. El efecto DB asociado a la secuencia del ADN

Se estudió si la tendencia de algunos clones a captar en mayor medida un fluorocromo respecto a otro estaba asociado a la secuencia de estos. Para determinar esta asociación se tomaron las estimaciones del efecto DB obtenidas mediante el método de normalización *loess loc scale* con la finalidad de minimizar el efecto espacial asociado. Se tomaron los clones que obtuvieron valores absolutos mayores de 0,1 tanto en el experimento 32x como en el experimento 19c. Este método permitió identificar 174 clones con una clara tendencia a captar el fluorocromo Cy3 (una cantidad tres veces mayor a lo esperable por azar), 22 clones con una clara tendencia a captar el fluorocromo Cy5 y 215 clones sin efecto DB (considerando como clon libre de efecto DB si no resultó significativo este efecto ni en el experimento 32x ni en el experimento 19c). Los clones que presentaban tendencia hacia el Cy5 no fueron estudiados en este primer análisis debido a su reducido tamaño muestral.

Los principales resultados se presentan en las Tablas 21, 22 y 23. Los clones analizados presentaban una longitud aproximada de 150 kb sin diferencias significativas entre grupos (ver Tabla 21).

Tabla 21: **Resumen de los resultados obtenidos para la prueba U-Mann Whitney** en relación a características de la secuencia de ADN y el DB.

		EFECTO DB			pvalor adj[a]	
		No (N=215)	Cy3 (N=174)	Total (N =389)	No	Si
BAC L [b]	media ± sd mediana ± mad	149003 ± 45532 161972 ± 29817	151689 ± 52374 153496 ± 30880	150204 ± 48665 160335 ± 30747	0,500	0,500
DS L [b]	media ± sd mediana ± mad	27346 ± 59472 0 ± 0	48384 ± 82413 0 ± 0	36757 ± 71336 0 ± 0	0,003	0,006
CpG island N[c]	media ± sd mediana ± mad	3,74 ± 4,10 2 ± 1,48	8,80 ± 9,10 5 ± 5,93	5,86 ± 7,11 3 ± 2,96	<0,001	<0,001
CpG Island L[b]	media ± sd mediana ± mad	1336 ± 1099 714 ± 1010	3415 ± 3801 2122 ± 2691	2265 ± 2999 1064 ± 1503	<0,001	<0,001
CNV L [b]	media ± sd mediana ± mad	19012 ± 45277 0 ± 0	33443 ± 60972 0 ± 0	25467 ± 53290 0 ± 0	0,0134	0,022
Gen N [c]	media ± sd mediana ± mad	1,61 ± 2,41 1 ± 1,48	3,9 ± 4,3 2 ± 2,97	2,64 ± 3,59 1 ± 1,48	<0,001	<0,001
Gen L [c]	media ± sd mediana ± mad	143060 ± 285849 29456 ± 43671	120555 ± 167843 68306 ± 101270	132994 ± 240318 50245 ± 74493	0,164	0,197

[a]: Indica si el pvalor está ajustado por multi test
[b]: Longitud en pb
[c]: Número

Se hallaron diferencias significativas, entre los clones con tendencia a captar Cy3 y clones sin DB, en el número de islas CpG que contenían y en su longitud (ver Tabla 21). También se hallaron diferencias significativas en el número de genes pero, en cambio, no se hallaron diferencias significativas en el número de pares de bases (pb) de cada clon contenido en genes (longitud génica). Además, se encontró asociación entre número de pb

implicados en DSs y entre el número de pb implicados en CNVs descritas.

También se hallaron diferencias significativas en el porcentage de G+C entre ambos grupos. Se encontró un porcentage más alto de G+C (47 %) en el grupo con tendencia al Cy3 mientras que en el grupo sin DB presentaban un porcentage sensiblemente menor (40 %). Como cabía esperar existe una correlación mayor del 70 % entre el contenido de G+C, el número de islas CpG y el número de genes.

Las mismas asociaciones se pueden apreciar en el análisis cualitativo (Tabla 22); la tendencia al Cy3 de algunos clones está asociado a la presencia de DSs y a una mayor densidad de genes. No hay una distribución homogénea en los cromosomas entre los clones con Cy3 y sin DB indicando que hay regiones del genoma más propensas a presentar este efecto (i.e. existe un claro enriquicimiento de clones con tendencia al Cy3 en el cromosoma 19). El enriquicimiento de G+C que existe en este cromosoma explicaría esta asociación. Y, aunque la longitud del clon implicada en CNVs obtuvo un pvalor significativo, no se ha observado la misma tendencia en el análisis cualitativo si bien este resultado puede estar enmascarado por la presencia de CNVs detectadas con tecnologías distintas por lo que sólo sería comparable el tamaño de estas ya que CNVs de mayor tamaño indicarían la utilización de la misma técnica.

Se estudiaron otro tipo de repeticiones presentes en los BACs (ver Tabla 23) y se hallaron asociaciones significativas con los elementos *Alu* y L1. Cuyas asociaciones pueden ser explicadas por la correlación entre éstas y el contenido en G+C e islas CpG.

Análogamente se estudiaron 2.372 clones procedentes del experimento 47m que eran comunes a la matriz aCGH 5,2K. Estos clones fueron clasificados en tendencia al Cy3 (N=607), tendencia al Cy5 (N=999) y sin tendencia (N=676). Se obtuvieron diferencias significativas entre estos y su contenido en islas CpG (Kruskal-Wallis test $< 0{,}0001$). Se realizaron comparaciones múltiples utilizando el test U de Mann-Whitney corregido por BH y se hallaron diferencias significativas entre los tres niveles. Clones con tendencia en Cy5 presentaban un menor contenido de islas CpG ($3{,}6 \pm 4{,}21$), clones sin DB presentaban valores intermedios ($4{,}08 \pm 4{,}66$) y clones con tendencia al Cy3 presentaban una mayor cantidad de islas CpG ($5{,}04 \pm 5{,}74$).

Tabla 22: **Resumen de los resultados obtenidos para la prueba Chi-cuadrado o test exacto de Fisher** en relación a características de la secuencia de ADN y el DB.

		EFECTO DB			pvalor adj[a]	
		NO (N=215) N (%)	Cy3 (N=174) N (%)	Total (N =389) N (%)	Si	No
DS	No	178 (82,8 %)	128 (73,6 %)	306 (78,7 %)	0,034	0,035
	Si	37 (17,2 %)	46 (26,4 %)	83 (21,3 %)		
	Total	215 (100,0 %)	174 (100,0 %)	389 (100,0 %)		
CNVs[b]	No	55 (25,6 %)	62 (35,6 %)	117 (30,1 %)	0,035	0,035
	Si	160 (74,4 %)	112 (64,4 %)	272 (69,9 %)		
	Total	215 (100,0 %)	174 (100,0 %)	389 (100,0 %)		
GENES	No	91 (42,3 %)	55 (31,6 %)	146 (34,3 %)	0,035	0,035
	Si	124 (57,7 %)	119 (68,4 %)	243 (65.7 %)		
	Total	215 (100,0 %)	174 (100,0 %)	389 (100,0 %)		
Chr	1	16 (7,4 %)	11 (6,3 %)	27 (6,9 %)	<0,001	<0,001
	2	19 (8,8 %)	4 (2,3 %)	23 (5,9 %)		
	3	9 (4,2 %)	5 (2,9 %)	14 (3,6 %)		
	4	15 (7,0 %)	4 (2,3 %)	19 (4,9 %)		
	5	17 (7,9 %)	6 (3,4 %)	23 (5,9 %)		
	6	7 (3,3 %)	11 (6,3 %)	18 (4,6 %)		
	7	23 (10,7 %)	10 (5,7 %)	33 (8,5 %)		
	8	13 (6,0 %)	8 (4,6 %)	21 (5,4 %)		
	9	2 (0,9 %)	3 (1,7 %)	5 (1,3 %)		
	10	15 (7,0 %)	3 (1,7 %)	18 (4,6 %)		
	11	10 (4,7 %)	15 (8,6 %)	25 (6,4 %)		
	12	8 (3,7 %)	7 (4,0 %)	15 (3,9 %)		
	13	4 (1,9 %)	5 (2,9 %)	9 (2,3 %)		
	14	3 (1,4 %)	4 (2,3 %)	7 (1,8 %)		
	15	6 (2,8 %)	4 (2,3 %)	10 (2,6 %)		
	16	5 (2,3 %)	8 (4,6 %)	13 (3,3 %)		
	17	4 (1,9 %)	6 (3,4 %)	10 (2,6 %)		
	18	4 (1,9 %)	1 (0,6 %)	5 (1,3 %)		
	19	4 (1,9 %)	25 (14,4 %)	29 (7,5 %)		
	20	4 (1,9 %)	10 (5,7 %)	14 (3,6 %)		
	21	3 (1,4 %)	1 (0,6 %)	4 (1,0 %)		
	22	5 (2,3 %)	5 (2,9 %)	10 (2,6 %)		
	23	14 (6,5 %)	16 (9,2 %)	30 (7,7 %)		
	24	5 (2,3 %)	2 (1,1 %)	7 (1,8 %)		
	Total	215 (100,0 %)	174 (100,0 %)	389 (100,0 %)		

[a]: Indica si el pvalor está ajustado por multi test

[b]: Presencia de CNVs en la base de datos tcag a fecha de Setiembre 2007

Tabla 23: **Resumen de los resultados obtenidos para la prueba U-Mann Whitney** en relación al ADN repetitivo y el DB.

		Efecto DB			pvalor adj[a]	
		No (N=215)	Cy3 (N=174)	Total (N =389)	Si	No
SINE N[b]	media ± sd	91,8 ± 65,9	127,0 ± 92,6	105,9 ± 79,7	$< 0,001$	$< 0,001$
	mediana ± mad	74,0 ± 43,0	107,0 ± 81,6	85,0 ± 59,3		
SINE L[c]	media ± sd	20434 ± 15905	28546 ± 22127	23672 ± 19059	$< 0,001$	$< 0,001$
	mediana ± mad	15431 ± 9799	22601 ± 17373	18551 ± 13124		
Alu N[b]	media ± sd	61,3 ± 61,2	91,9 ± 84,3	73,7 ± 72,7	$< 0,001$	$< 0,001$
	mediana ± mad	38,0 ± 28,2	64,0 ± 56,3	45,0 ± 35,6		
Alu L[c]	media ± sd	16119 ± 15564	23693 ± 21378	19187 ± 18438	$< 0,001$	$< 0,001$
	mediana ± mad	10502 ± 7715	16406 ± 14353	12003 ± 9677		
MIR N[b]	media ± sd	30,5 ± 20,1	35,2 ± 28,7	32,1 ± 24,5	0,679	0,788
	mediana ± mad	28,0 ± 16,3	27,0 ± 22,2	27,0 ± 19,3		
MIR L[c]	media ± sd	4315 ± 2882	4852 ± 4082	4486 ± 3488	0,996	0,996
	mediana ± mad	3861 ± 2321	3589 ± 3214	3686 ± 2753		
LINE N[b]	media ± sd	73,7 ± 28,1	69,1 ± 32,0	72,1 ± 29,8	0,041	0,071
	mediana ± mad	75,5 ± 24,5	67,0 ± 28,1	72,0 ± 28,2		
LINE L[c]	media ± sd	32722 ± 16988	24806 ± 17764	29708 ± 17892	$< 0,001$	$< 0,001$
	mediana ± mad	29534 ± 16175	19827 ± 14541	26505 ± 16921		
L1 N[b]	media ± sd	48,1 ± 21,5	41,7 ± 23,9	45,9 ± 23,1	0,002	0,004
	mediana ± mad	46 ± 21,5	38,0 ± 23,7	45 ± 22,2		
L1 L[c]	media ± sd	26775 ± 16161	18944 ± 17274	23819 ± 17321	$< 0,001$	$< 0,001$
	mediana ± mad	14130 ± 13173	22580 ± 14147	20106 ± 15243		
L2 N[b]	media ± sd	21,8 ± 12,2	24,4 ± 17,2	22,7 ± 14,6	0,485	0,617
	mediana ± mad	21,0 ± 11,9	22 ± 16,3	21,0 ± 13,3		
L2 L[c]	media ± sd	5160 ± 3141	5340 ± 3899	5219 ± 3461	0,871	0,938
	mediana ± mad	4824 ± 2873	4775 ± 3629	4775 ± 3202		
LTR N[b]	media ± sd	34,0 ± 18,4	33,6 ± 25,3	34,0 ± 21,4	0,149	0,208
	mediana ± mad	32 ± 19,3	31,0 ± 23,7	32,0 ± 20,8		
LTR L[c]	media ± sd	12846 ± 7831	12727 ± 10916	13034 ± 9345	0,147	0,208
	mediana ± mad	11467 ± 7710	11080 ± 8986	11594 ± 8540		

[a]:Indica si el pvalor está corregido por multi test

[b]:Número

[c]:Longitud

4.3.3. El estado de purificación del ADN y el efecto DB

El análisis de datos procedentes de otros laboratorios puso de relieve la presencia de otras fuentes de variación que añaden variabilidad al residuo y, en consecuencia, lo incrementan (ver Figura 24). Una de las principales diferencias, entre las muestras hibridadas en el CRG y las muestras disponibles en la base de datos, era la procedencia del ADN y, por lo tanto, el protocolo de purificación empleado. Este hecho pudo haber dado lugar a ADNs con distinto grado de purificación siendo ésta una variable no controlada en los modelos realizados que podría incrementar los residuos. Dada la imposibilidad de obtener datos sobre el estado de purificación del ADN de las muestras procedentes de otros laboratorios, se estudiaron cuatro réplicas del experimento 32x obtenidas de una segunda extracción de sangre alterando distintos pasos del protocolo de purificación de ADN que dió lugar a distintas combinaciones de pureza.

Los resultados obtenidos con este experimento indican (ver Tabla 24) que hay una correlación entre la calidad de la muestra y la sensibilidad del método llegando incluso a la no detección en casos con gran contaminación proteíca. Además, tal y como cabía esperar, el grado de purificación óptimo estuvo situado alrededor de 1,8 en el ratio 260/230. La contaminación con ARN (condición 1) y la contaminación proteíca (condición 3) desbalancean el número de falsos positivos hacía el flurocromo Cy3.

Tabla 24: **Variación en la sensibilidad y número de falsos positivos** en la hibridación directa según la calidad del ADN.

Condición	Estado del ADN		Sensibilidad		N Falsos Positivos	
	260/280	260/230	Completo (N=4)	Parcial (N=4)	N DEL	N AMPL
1	1,86	1,87	50 %	25 %	13	1
3	1,80	1,74	100 %	50 %	9	7
9	1,60	0,89	75 %	50 %	35	6
12	1,33	0,42	0 %	0 %	90	7

4.4. Fuentes de variación asociadas a la imagen

Se realizaron tres estudios independientes con la finalidad de valorar la calidad de las imágenes obtenidas por cada hibridación.

4.4.1. Clasificación de *spots* a partir de las imágenes obtenidas por GenePix

En el primer estudio se valoró la capacidad de etiquetar correctamente los mismos *spots* por la misma persona en rondas diferentes habiendo dejado un espacio temporal igual o superior a cuatro días entre cada una ellas. La correlación entre rondas obtenida en los seis portaobjetos analizados se presentan en la Tabla 25.

Tabla 25: **Se muestra el grado de concordancia** entre las asignaciones realizadas por el mismo obervador en dos rondas distintas en porcentaje.

GPR valorado	Primer Orden de Evaluación	Segundo Orden de Evaluación	Correctamente Clasificados
MG 13302100 pM Cy5 00-18 Cy3 DS.gpr	1 (9/11/2005)	5 (21/11/2005)	75 %
MG 13309485 00-18 Cy5 pF Cy3.gpr	2 (14/11/2005)	2 (18/11/2005)	78 %
MG 13309747 pool100 Cy5 00-18 Cy3 DS.gpr	3 (14/11/2005)	3 (18/11/2005)	67 %
MG 13311280 00-18 Cy5 pM Cy3.gpr	4 (14/11/2005)	4 (21/11/2005)	72 %
MG 13311630 00-18 Cy5 pool Cy3.gpr	5 (14/11/2005)	6 (21/11/2005)	62 %
MG 13311739 pF Cy5 00 18 Cy3 DS.gpr	6 (14/11/2005)	1 (18/11/2005)	73 %

El estadístico Kappa global fue de 0,59 indicando la presencia de una correlación significativa entre las dos rondas pero moderada.

Para estudiar la relación entre calidad y forma de los *spots* se eliminaron los *spots* poco frecuentes (insuficientes datos para ser evaluados); F (substracción de 3 datos) y Ms (substracción de 17 datos). Finalmente sólo se estudiaron aquellos *spots* que fueron clasificados en la misma categoría las dos veces (N=407).

Se aplicó un test chi-square para conocer si el porcentaje de tipos de *spots* es homogéneo entre los portaobjetos estudiados (cada una de las réplicas técnicas). El pvalor obtenido fue de 0,003, indicando que existen diferencias en el porcentaje de los tipos de *spots* hallados en cada portaobjetos.

Posteriormente se realizó un análisis discriminante entre los grupos de *spots* identificados y las variables obtenidas por el programa GenePix. Se obtuvieron dos funciones discriminantes sobre los grupos estudiados (D, A y W). Las variables significativas que permiten clasificar mayor número de *spots* correctamente fueron F532 Median - B532, F635 Median - B635, F532 CV y F635 CV. Los valores de Lambda de Wilks para las variables F532 Median - B532 y F635 Median - B635 fueron de 0,84 mientras que para F532 CV y F635 CV fueron de 0,61 y 0,60 respectivamente. El porcentaje de casos correc-

tamente clasificados fue de 72,5 %.

Estas funciones discriminantes pueden ser interpretadas fácilmente ya que en el eje de abcisas (Función 1) los grupos de *spots* se separan por su coeficiente de variación. Los valores bajos en los coeficientes de variación se corresponden con la variedad W y los valores altos con D y A. En la función discriminante del eje de ordenadas (Función 2) los grupos se separan a través de la intensidad de cada canal menos la intensidad del background o ruido de fondo. Este segundo eje permite separar las variedades A y D. Así parece que la variedad A presenta, en general, intensidades más bajas que la D (ver biplot en la Figura 28).

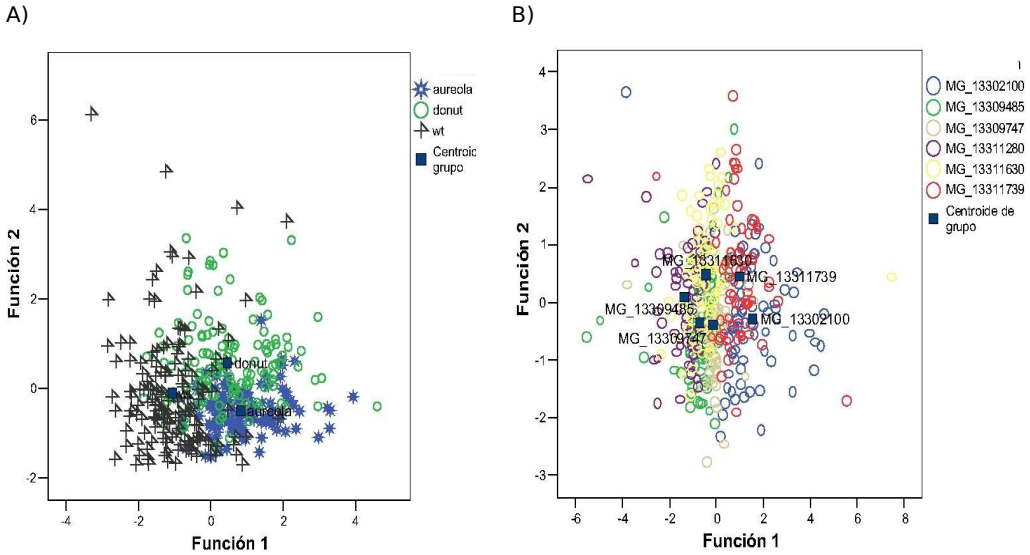

Figura 28: **Biplots para *spots* a la izquierda y para portaobjetos a la izquierda**

Se realizó un análisis discriminante con la intención de conocer la capacidad de clasificación de cada uno de estos portaobjetos por las variables obtenidas en el anterior análisis discriminante. Se obtuvieron cuatro funciones discriminantes (una por cada variable a estudiar) sobre los seis portaobjetos. Las variables F635 Median - B635 y F532 Median - B532 no fueron significativas y, por lo tanto, obtuvieron valores de Lambda de Wilks próximos a 1 (0,98 y 0,97 respectivamente). Las variables F532 CV y F635 CV resultaron significativas con valores de Lambda de Wilks de 0,87. Estas funciones clasifican correctamente el 55 % de los casos mientras que por azar se esperaba un 33 %. En la Figura 28 B) puede observarse el *biplot* de los portaobjetos respecto las dos primeras funciones discriminantes.

Finalmente, se realizó un ANOVA de efectos fijos dónde M era la respuesta y los factores analizados fueron el portaobjetos y la forma asignada. No se hallaron diferencias significativas entre formas ni en la interacción forma-portaobjetos pero sí que se hallaron diferencias significativas entre portaobjetos.

4.4.2. Fiabilidad entre observadores en la evaluación de la imagen

Después de la aplicación del protocolo de análisis de imágenes, se hallaron correlaciones intraindividuales mayores que en el apartado anterior (en promedio Kappa=82 %) y, como cabía esperar, más altas que las interindividuales (en promedio, Kappa=60 %).

Se examinaron las variables dispensadas por el programa Genepix para conocer si existía alguna asociación significativa con el tipo de forma. Este estudio se llevó a cabo con 144 BACs que tenían formas equivalentes en sus replicados (40 A y 104 D) y que fueron clasificados por los cuatro observadores dentro de la misma categoría. Las variables que mostraron asociación significativa aplicando el test U-Mann-Whitney fueron: *F635 Median, F532 Median, F635 Sd, F532 Sd, circularidad, F635 Median -B635, F532 Median -B532, F635 total intensidad, F532 total intensidad, SNR635 y SNR532.* Así las formas donut (D) se asocian con intensidades mayores que en aureolas (A), mayor dispersión y menor circularidad.

Estas mismas asociaciones fueron obervadas cuando se tomaron los BACs que habían sido clasificados dentro de la misma categoría por al menos tres observadores.

Se analizó el tamaño de la ronda con la capacidad de clasificar correctamente los *spots* sin obtener una asociación estadísticamente significativa.

4.4.3. Datos atípicos y las formas de los *spots*

Del experimento 32x se obtuvieron un conjunto de *spots* que podían considerarse datos atípicos o anómalos. Sobre estos, se eligieron 99 *spots* que se correspondían con datos atípicos en al menos un portaobjetos de los cinco elegidos para realizar el análisis (N=495).

396 *spots* (158 A, 168 D, 39 F, 7 Ms y 24 Wt) fueron clasificados en la misma categoría por al menos tres de los cuatro evaluadores. Se halló una asociación significativa entre las formas y el grupo (dato atípico versus no dato atípico) indicando una presencia más alta de F en el grupo de datos atípicos. También se valoró la concordancia entre las réplicas versus el grupo obteniendo, igualmente, una asociación significativa con un enriquicimiento de datos no concordantes en la categoría de datos atípicos. Pero no se halló una correlación clara con el resto de formas, de hecho en 66 casos de los 99 los BACs presentaban las mismas formas tanto cuando se habían clasificado como dato atípico como cuando no era así.

Se evaluó la relación entre los datos atípicos y las variables proporcionadas por el programa GenePix. Se encontraron las siguientes asociaciones positivas aplicando el test U-Mann-Whitney; *F635Median, F635Median-B635, F635CV, F532CV, B635CV, B532CV, circularidad, SNR.* A partir de estos resultados se calculó la mediana para cada grupo y se creó un índice de calidad basado en el siguiente criterio:

- Si F635 está por encima de 16.070 o F635Mean -B635 está por encima de 15.600 se puntuará con 1
- Si F635CV está por encima de 46 o B635CV por encima de 26 se puntuará con 1

- Si F532CV está por encima de 46 o B532CV por encima de 26 se puntuará con 1
- Si Circularidad está por debajo de 90 se puntuará con 1
- Si SNR635 o SNR532 están por debajo de 80 se puntuará con 1

La suma de estos puntuaciones permite clasificar los *spots* en 6 categorías de calidad dónde es 0 buena calidad y 5 mala calidad, y por ello, puede considerarse una regla a utilizar en la validación de los *spots*.

En el gráfico boxplot A) de la Figura 29 se observa una clara diferenciación entre los datos atípicos y datos no atípicos según la puntuación recibida. En el gráfico *boxplot* B) se oberva un incremento de la variabilidad en M cuando la calidad del *spot* baja indicando que una baja calidad del *spot* predispone a la obtención de datos con valores anómalos.

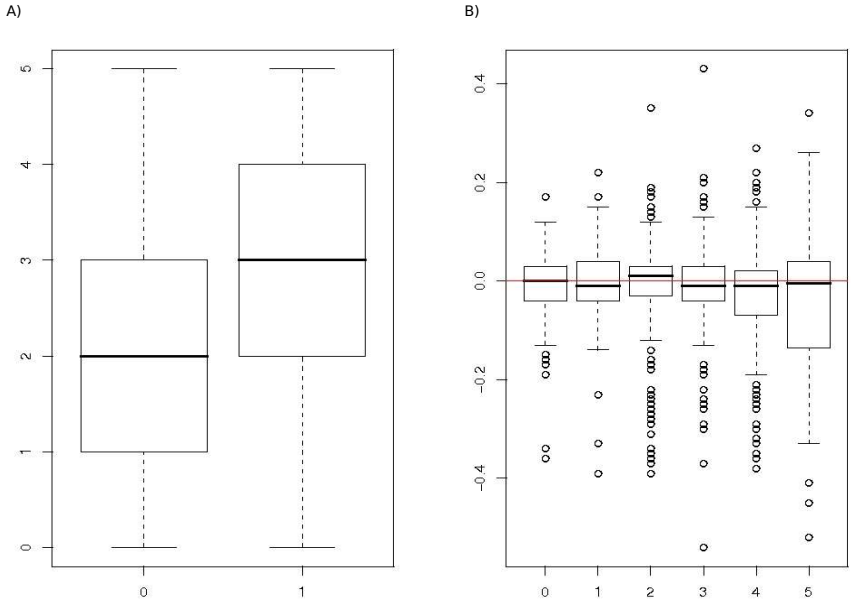

Figura 29: **Efecto de la calidad en la presencia de datos atípicos y en el valor de la respuesta** A) Clasificación de los datos atípicos en abcisas (0=No, 1=Sí) según el score de calidad ordenadas, B) En el eje de ordenadas se representa M y en el eje de abcisas se representa la calidad de 0, alta, a 5, baja,.

En la Figura 30 se muestra la cross-validación realizada mediante un set de datos con valores atípicos independiente (es decir, no utilizado en la creación de la puntuación). En este gráfico se observa, de igual modo, una mayor tendencia a tomar valores atípicos (fuera del rango de normalidad representado por las líneas verdes) cuando la puntuación de calidad posee un valor de 3 o mayor.

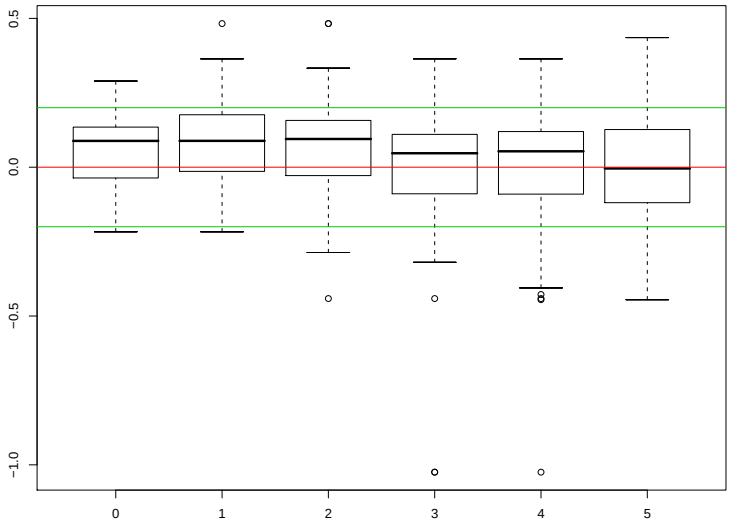

Figura 30: ***Cross-validación*** **mediante un grupo de valores atípicos situados en otros portaobjetos**

4.5. Métodos para la detección de CNVs

En este apartado se ha desarrollado dos métodos para la detección de variaciones en el número de copias. Uno de ellos está basado en IC que incorporan la información obtenida de los modelos ANOVA. El segundo método, es un método combinado que utiliza la información obtenida de los ANOVAs con los métodos basados en puntos de corte y CBS.

4.5.1. Desarrollo de un método basado en intervalos de confianza (IC)

En este apartado se ha realizado; (i) la estimación de los parámetros de centralización y de dispersión para los ICs y (ii) la comparación de este método con otros descritos en la literatura.

Caracterización de la medida de centralización

Según los resultado previos obtenidos en esta tesis doctoral, se ha considerado cuatro distribuciones relacionadas con cada medida de centralización resultantes de la combinación del tipo de hibridación (HD y HR) y género (femenino y masculino).

Se han considerado dos estrategias para la obtención de estas estimaciones; (i) mediante estimación máximo verosímil a partir de los modelos ANOVA para los experimentos 32x y 19c según está implementado en la librería *lmm* y (ii) mediante remuestreo.

La estimación máximo-verosímil utilizada para el experimento 32x considera el efecto DB como fijo y el resto de efectos como aleatorios, mientras que la estimación realizada en este apartado para el experimento 19c considera DB y Género como fijos pero el efecto muestra es considerado como aleatorio.

El algoritmo de remuestreo se ha realizado sobre los experimentos 32x, 19c y 6p:

1. Por cada BAC
2. Generar un número B de remuestras
3. Tomar al azar G muestras con reemplazamiento HD y G muestras con reemplazamiento HR
4. Tomar al azar r valores con reemplazamiento de cada portaobjetos
5. Generar 1 media
6. Generar G medias
7. Obtener B medias

Para el experimento 32x, se han generado B=100 remuestras tomando G=16. Para el experimento 19c, se han generado B=100 remuestras tomando G=10 para género masculino y G=9 para género femenino. Para el experimento 6p se han generado B=100

remuestras tomando G=3.

Las estimaciones obtenidas mediante el método máximo-verosímil y las obtenidas mediante remuestreo se comparan en la Tabla 26 utilizando la correlación de Pearson. En esta tabla se puede apreciar la presencia del efecto DB ya que se observa una pérdida de correlación entre HD y HR llegando a producir, incluso, una asociación negativa. También se aprecia una mayor correlación entre los valores del mismo género.

Tabla 26: **Correlaciones entre las estimaciones** obtenidas para las distintas distribuciones y con distintos métodos

Test	32x HD	32x HR	19c pF HD	19c pF HR	19c pM HD	19c pM HR
lmm HD (32x)	1	0,130	0,829	0,322	0,370	−0,533
lmm HR (32x)	0,130	1	−0,069	0,782	−0,599	0,493
REMUESTREO HD (32x)	0,996	0,135	0,830	0,079	0,374	−0,541
REMUESTREO HR (32x)	0,254	0,993	0,012	0,751	−0,557	0,426
REMUESTREO pF HD (19c)	0,868	0,054	0,980	0,033	0,518	−0,642
REMUESTREO pF HR (19c)	0,098	0,795	−0,092	0,989	−0,760	0,668
REMUESTREO pM HD (19c)	0,370	−0,641	0,604	−0,763	0,984	−0,763
REMUESTREO pM HR (19c)	−0,498	0,519	−0,687	0,990	−0,777	0,990
lmm pF HD (19c)	0,829	-0,069	1	−0,086	0,628	−0,713
lmm pF HR (19c)	0,090	0,782	−0,086	1	−0,760	0,676
lmm pM HD (19c)	0,370	−0,599	0,628	−0,713	1	−0,778
lmm pM HR (19c)	−0,533	0,493	−0,713	0,676	−0,778	1
REMUESTREO HD (6p)	0,297	−0,411	0,411	−0,568	0,681	−0,615
REMUESTREO HR (6p)	−0,252	0,456	−0,384	0,317	−0,440	0,452

En general, los resultados muestran una buena correlación entre las estimaciones máximo-verosímiles y las estimaciones realizadas mediante remuestro. Se optó por considerar las estimaciones máximo-verosímiles para la creación del IC excepto en aquellos casos en los que la estimación máximo-verosímil no fue posible (debido a la presencia de datos pérdidos) en los que se utilizó la estimación obtenida por remuestreo.

Caracterización de la medida de dispersión

Mediante estimaciones máximo-verosimiles de los modelos 32x y 19c se ha obtenido una estimación por cada BAC de su variabilidad que está asocida a un error aleatorio no sistemático y al factor portaobjetos. La aportación de la muestra dentro de la variabilidad de cada BAC no se considera ya que en ello influye los polimorfismos de copias presentes.

No se obtuvo una buena correlación para la variabilidad de cada clon considerando las estimaciones producidas por los modelos 32x y 19c. En general, las variabilidades asociadas al experimento 32x fueron mayores que en las halladas para el experimento 19c.

Además se demostró que existe una relación entre la variabilidad del clon y la presencia del efecto DB así como con la presencia de CNVs (ver Figura 31 y Figura 32).

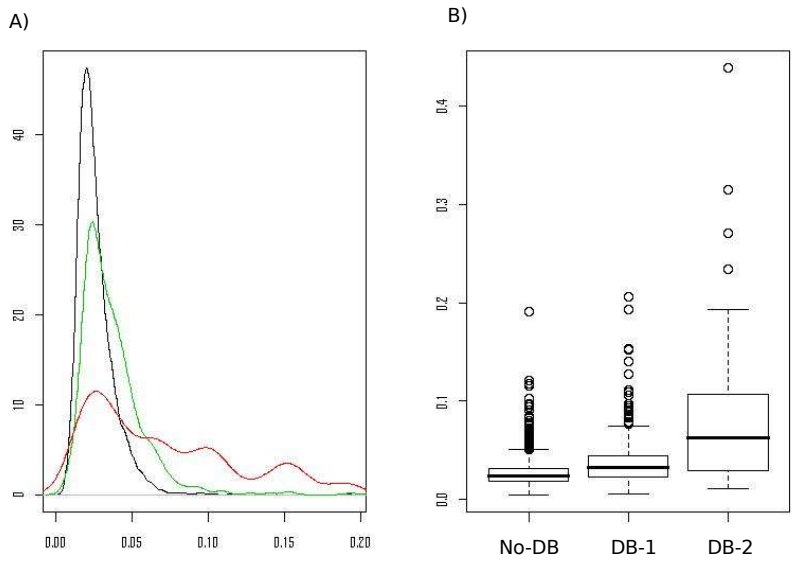

Figura 31: **El efecto DB y la variabilidad**; A) muestra un grafico de densidad dónde la línea negra representa la distribución de la desviación típica en BACs sin efecto DB, en verde BACs con efecto moderado de DB (valor absoluto mayor de 0,1) y en rojo BACs con gran efecto DB (valor absoluto mayor de 0,2). En el gráfico B) puede apreciarse la misma escala. Datos obtenidos del experimento 32x.

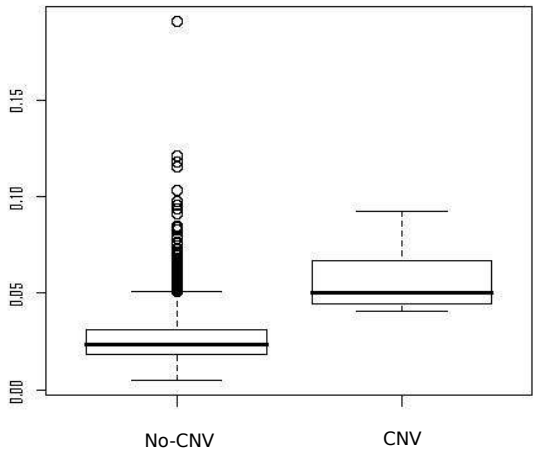

Figura 32: **La presencia de CNVs incrementa la variabilidad del BAC**.Datos obtenidos del experimento 32x.

Se consideró un único valor de dispersión común a todos los BACs. Se tomó como la medida más fiable la mediana de la distribución de la desviación típica asociada a cada clon en el experimento 32x dividida por la raíz cuadrada del número de réplicas existente en cada portaobjetos (0,027).

Caracterización del IC

El IC teórico se representa en la ecuación 4.

$$\mu_{b,g,d} \pm *Z_{\alpha/2} * \sigma$$

$$\begin{aligned} &\forall b = 1, ..., 5{,}222 \\ &\forall g = \{pF, pM\} \\ &\forall d = \{HD, HR\} \end{aligned} \tag{4}$$

Dónde pF representa género femenio y pM género masculino.

Se obtiene una medida de centralización, μ_{bgd}, por cada BAC, género y tipo de hibridación. μ_{bgd} fue estimado ($\bar{log}_{b,g,d}$) mediante el modelo 19c o mediante el método de remuestreo en presencia de datos perdidos.

$Z_{\alpha/2}$ puede representar los percentiles de una distribución normal o bien unos puntos de corte.

No se encontraron evidencias de que existiera una variabilidad asociada cada BAC, género o tipo de hibridación por lo que se consideró un único valor de σ representado por $sd/\sqrt{n}$.

Se considera duplicación si $y_{.,b,g,d}$ (promedio de las réplicas de un portaobjetos) es mayor que $\bar{log}_{b,g,d} + Z_{\alpha/2} * sd/\sqrt{n}$

Se considera deleción si $y_{.,b,g,d}$ (promedio de las réplicas de un portaobjetos) es menor que $\bar{log}_{b,g,d} - Z_{\alpha/2} * sd/\sqrt{n}$

Comparación de métodos en la detección de CNVs

En esta sección se comparan los siguientes métodos de detección de CNVs:

1. Método 1 basado en puntos de corte; PCmed. Se considera alteración si la media de las réplicas presentes en el mismo portaobjetos superan un cierto umbral y si la variabilidad entre estas es inferior a otro umbral. Los puntos de corte o umbrales para las medias en valor absoluto fueron; 0,2; 0,25 y 0,3 y para la desviación típica fueron de 0,1; 0,05 y 0,15.

2. Método 2 basado en puntos de corte; PCmin. Se considera alteración si el valor mínimo de una de las réplica supera; 0,2; 0,25 y 0,3 o bien si el máximo no supera -0,2, -0,25 o -0,3.

3. CBS con un punto de corte de en valor absoluto fue de 0,2 con una probabilidad para cada segmento de 0,01 y 0,001, a 3, 5 y 7 desviaciones típicas con y sin suavizado.

4. IC a 4, 5, 6, 7 y 8 desviaciones típicas.

En la Figura 33 se presentan los resultados obtenidos por cada método en cuanto a su capacidad de detectar 8 BACs alterados en el experimento 32x (eje de ordenadas) respecto a la eficiencia. El método CBS sin suavizar es el que obtiene peores resultados mientras que el método basado en IC se muestra más adecuado ya que obtiene valores de sensibilidad muy constantes en función de la eficiencia en HD. El método PCmed basado en puntos de corte es más sensible que el método PCmin aunque éste último toma valores de eficiencia mayores.

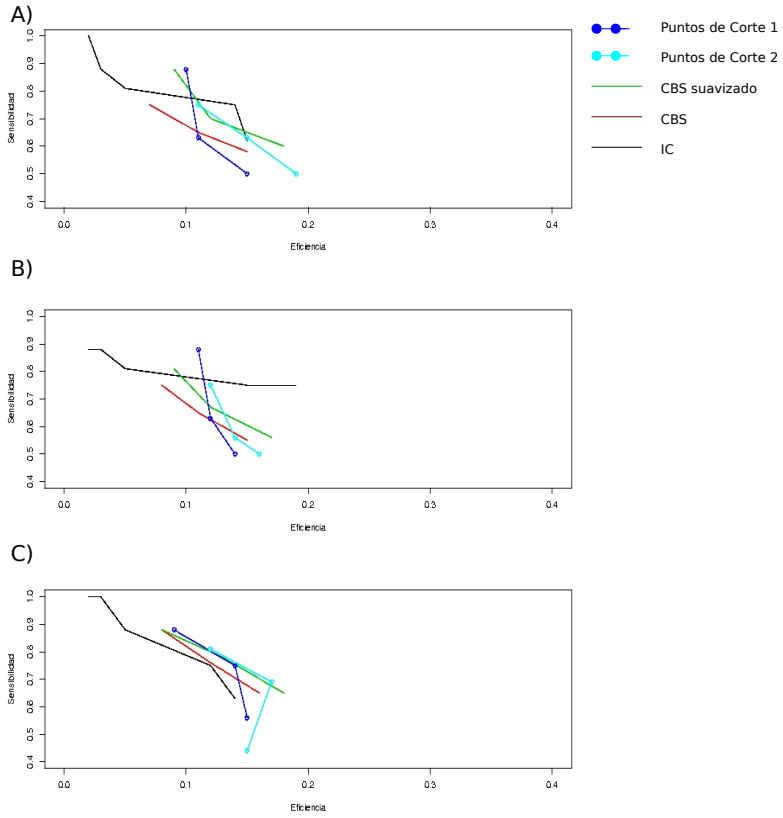

Figura 33: **Comparación de los métodos propuestos**; A)Arriba los resultados para todas las hibridaciones, B) Resultados para HD, C) Resultados para HR.

4.5.2. Desarrollo de métodos combinados

En esta sección se aplican los métodos basados en puntos de corte (PCmed y PCmin) y CBS una vez se han corregido los datos restando a cada BAC el valor estimado según hibridación y género.

Los resultados muestran un incremento de la eficiencia sin que se produzca un decremento de la sensibilidad. Según se puede observar en la Figura 34 el método CBS suavizado y sin suavizar presenta resultados muy similares. También se hallan valores muy similares entre los dos métodos basados en puntos de corte. Cabe destacar que, dados valores similares de sensibilidad, la hibridación HD es más eficiente que HR.

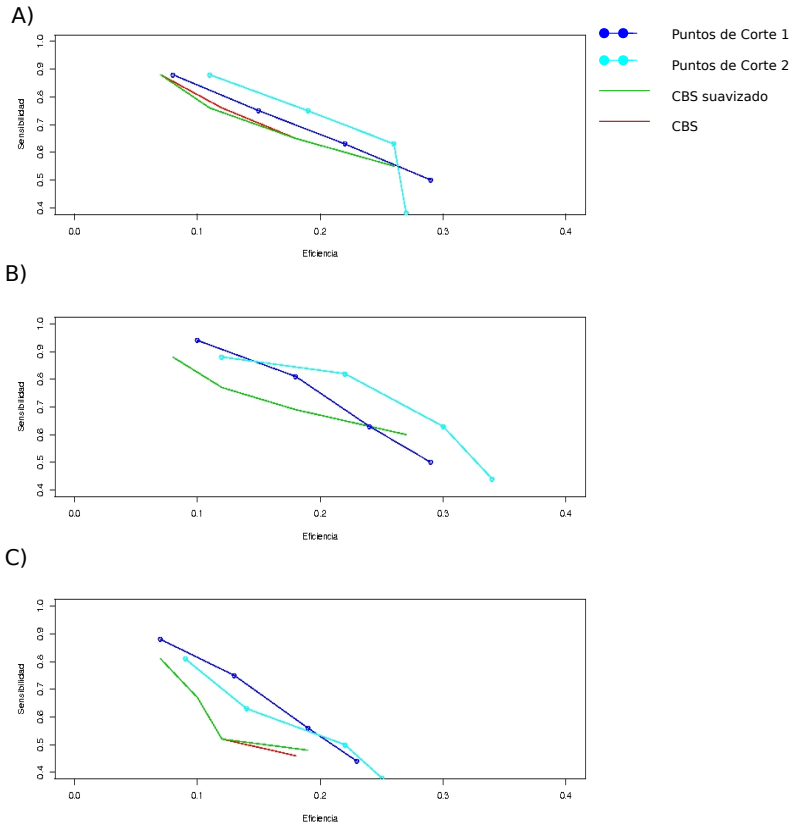

Figura 34: **Comparación de los métodos combinados**; A)Arriba los resultados para todas las hibridaciones, B) Resultados para HD, C) Resultados para HR.

4.5.3. Rendimiento de los métodos propuestos mediante simulación

Este apartado se divide en dos partes; (i) desarrollo de un algoritmo de simulación y (ii) los principales resultados obtenidos.

Desarrollo de un algoritmo de simulación

Teniendo en cuenta todo lo expuesto en apartados anteriores se ha desarrollado el siguiente algoritmo:

1. Se elige el género de la muestra
2. Por cada muestra se genera HD y HR (que tendrán las mismas CNVs)
3. Se obtiene la desviación típica de cada hibridación mediante una distribución gamma de parámetros; forma de 0,027 y ratio de 1.
4. Se genera un valor aleatorio uniforme entre 0 y 1
5. Si el clon anterior no está alterado y el valor es menor de 0,01 o bien si el clon anterior está alterado y el valor generado es menor de 0,2 se genera otro valor aleatorio entre 0 y 1:
   - si el valor es menor de 0,5 se genera una deleción mediante un valor Uniforme dentro del intervalo -1 y -0,2

- si el valor es mayor de 0,5 se genera una ampliación mediante un valor Uniforme dentro del intervalo 0,2 y 0,6

En otro caso se genera un cero.

6. Se genera n réplicas de cada clon mediante una distribución normal de media la estimación según género y tipo de hibridación más el valor generado en el paso anterior y con desviación típica la generada en el paso 3.

Mediante este algoritmo se han generado 20 muestras de género masculino y 20 muestras de género femenino. Se han generado un promedio de 28 deleciones y 28 amplificaciones por cada muestra.

Principales resultados

En la Figura 35 se compara el perfil de una hibridación real (del experimento 32x) con una hibridación simulada mediante el algoritmo descrito. En ella se puede observar como el nivel de ruido por perfil son similares entre los datos reales y los datos simulados.

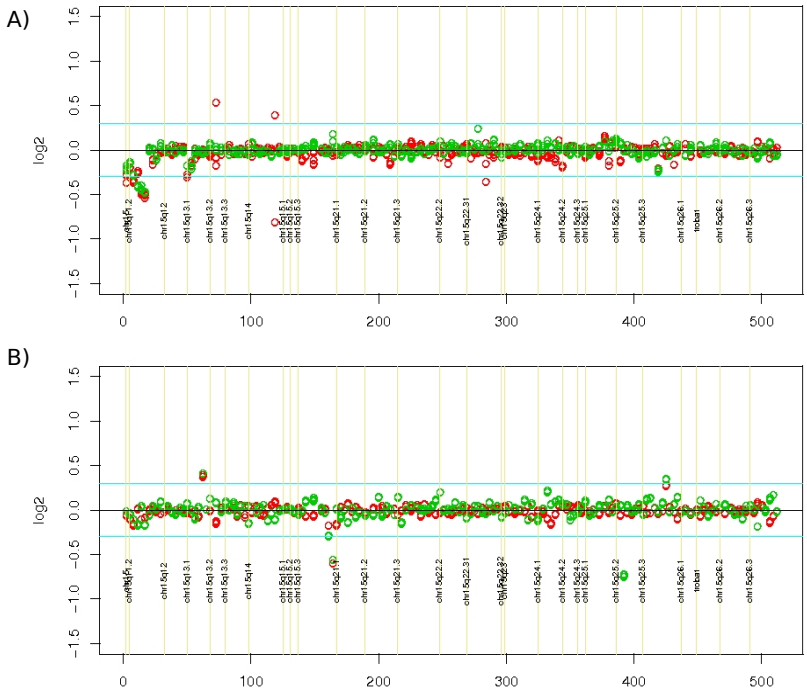

Figura 35: **Perfil real versus perfil creado mediante simulación**; A) Perfil real con una deleción en la región p-terminal del cromosoma 15 B) Perfil creado mediante simulación

Los perfiles simulados HD tienen una correlación superior 70 % con el perfil generador mientras que las hibridaciones HR tienen una correlación cercana al 60 %. Dichas correlaciones se observan tanto para el perfil de género femenino como para el perfil de género masculino. La variabilidad mediana de cada perfil es de 0,089 y oscila entre 0,079 y 0,194. Estas variabilidades parecen superiores a las encontradas en los perfiles reales pero el número de CNVs por muestra simulada es mucho mayor que las halladas en las

muestras reales referidas en este trabajo. La variabilidad por perfil sin estas CNVs se sitúa en mediana alrededor de 0,07 coincidiendo con los resultados reales y la visualización de los perfiles reales y simulados muestran variabilidades cercanas, ver Figura 35.

En la Figura 36 se compara la capacidad de detección de CNVs de los métodos PCmed, PCmin y CBS con y sin suavizado en muestras simuladas mediante curbas ROC. Los resultados obtenidos son parecidos a los obtenidos con datos reales; los mejores resultados se consiguen con PCmed y PCmin. El método CBS con suavizado resulta más eficiente que el método CBS sin suavizado. Del mismo modo se observa que la hibridación HD es más sensible y eficiente que la hibridación HR.

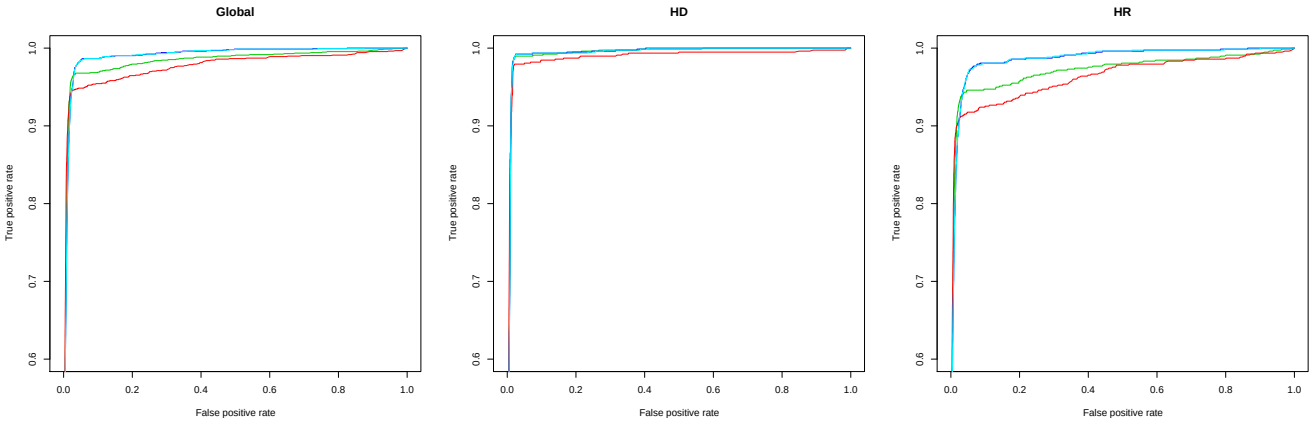

Figura 36: **Curbas ROC de métodos propuestos en la detección de CNVs**; A) Resultados globales, B) Resultados para HD, C) Resultados para HR. En azul marino se representa el método PCmed, en azul claro PCmin, en verde CBS suavizado y en rojo CBS no suavizado.

4.6. Concordancia entre plataformas en la detección de CNVs

La muestra utilizada en el experimento 32x se hibridó cuatro veces en Agilent 44K (2 en HD y 2 en HR) y se realizó otra hibridación con Agilent 244K. Ello permitió evaluar la conocordancia entre plataformas en la detección de CNVs.

En la Tabla 27 se resume las alteraciones validadas entre plataformas presentes en la muestra control. Este estudio permitió detectar una nueva CNV presente en el cromosoma X no descrita previamente (ver Figura 37). La información aportada por Agilent indica que al menos más del 50 % del BAC está alterado. En este BAC están contenidas cuatro DSs. La duplicación DC0353 guarda un 91 % de homología con DC0354. La duplicación DC0355 guarda un 91 % de homología con DC0356. Se tratan, pues, de DSs intracromosómicas. No existen otras DSs en esta región.

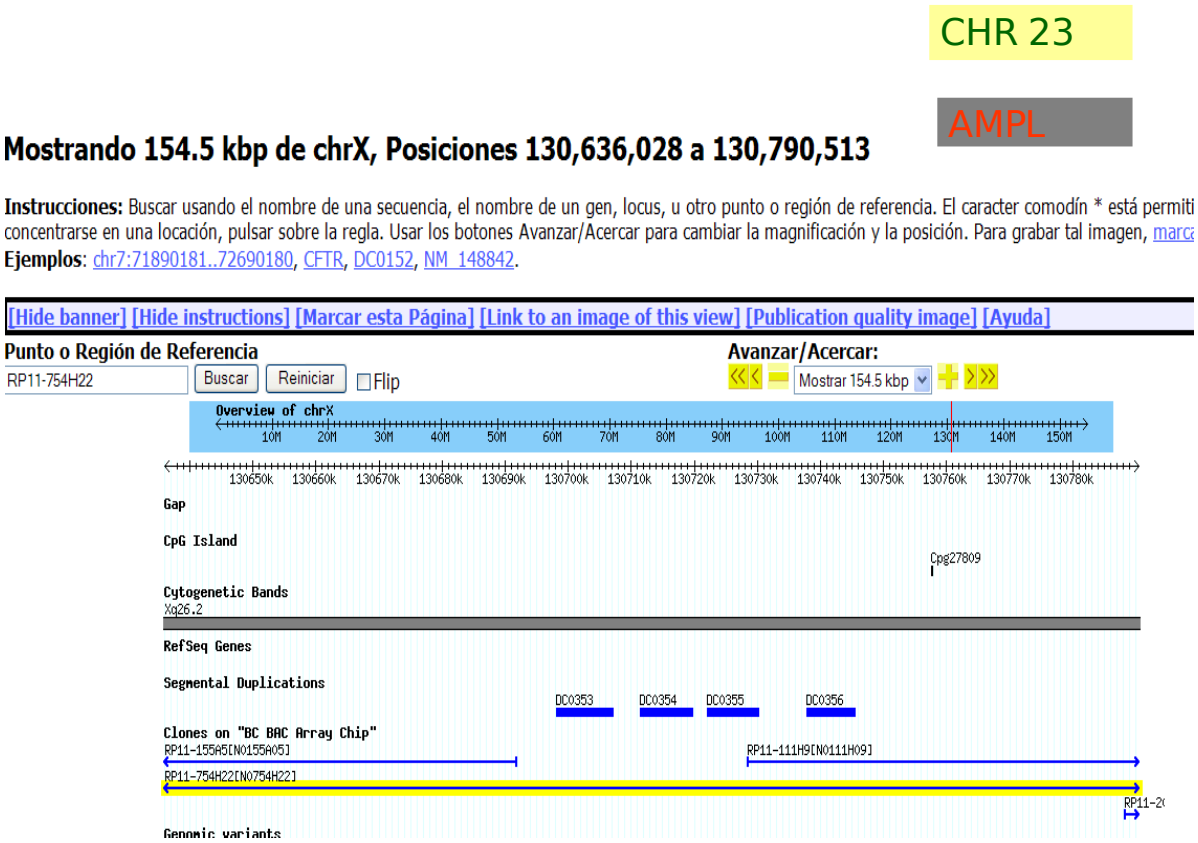

Figura 37: **Detalle de la región del cromosoma X** dónde se ha hallado una nueva CNV.

Con las plataformas Agilent sólo dos regiones más, con una longitud mayor a 100 kb, puedieron ser validadas entre sí. Para tres BACs; RP11-365H22, RP1-215P15 y RP11-410M8 no había sondas de Agilent que cubrieran dichas regiones y la región que cubría el BAC RP11-342G18 sólo había dos sondas en Agilent 244K con resultado negativo. Tres BACs más (RP11-694E12, RP11-483E23, RP11-550A14) estaban situados en regiones con DSs con pocas sondas de Agilent que cubrieran la región (menos de 5) y concentradas en una parte concreta del BAC que no permitió descartar la existencia de una CNV. Otros

Tabla 27: **Resumen de las CNVs halladas en un muestra control** Se presentan los datos hallados mediante tres plataformas distintas; BACs, Agilent 44K y Agilent 244K.

BAC	Chr	Inicio	Fin	Tipo	Longitud (pb)	CNV descrita	BACs aCGH			Agilent 44K		Agilent 244K	
							media	%	Min % Alt	media	N	media	N
RP11-449I24	14	19289488	19485285	DEL	195.798	SI	-0,46	100 %	100 %			-0,4	26
RP11-32B5	15	19154639	19296842	DEL	142.204	SI	-0,30	81 %	100 %			-0,44	11
RP11-2F9	15	19791326	19970462	DEL	179.137	SI	-0,33	49 %	82 %			-0,62	16
RP11-603B24	15	19970467	20097557	DEL	127.090	SI	-0,45	94 %	86 %			-0,7	8
RP11-264M14	16	32209877	32374541	DEL	164.665	SI	-0,25	25 %	65 %			-0,23	10
RP11-530B22	16	68631384	68800116	DEL	168.733	SI	-0,23	52 %	53 %	-0,19	5	-0,25	12
RP11-402H1	23	102974612	103207477	AMPL	232.866	SI	0,22	23 %	58 %	0,45	4	0,3	26
RP11-754H22	23	130636028	130790513	AMPL	154.486	NO	0,20	13 %	57 %			0,3	8
–	4	68723666	69806372	AMPL	1.082.706	SI	–	–	–	0,19	5	0,33	11
–	19	10779961	10901508	DEL	121.548	SI	–	–	–	-0,28	5	-0,27	7

seis BACs no se validaron aún con suficientes sondas Agilent cubriendo la región.

En Agilent 44k se detectaron 116 regiones candidatas a tener una posible CNV (ver la Figura 38). De estas, sólo 22 de ellas se detectaron en más de una réplica (20 %) y 8 de estas regiones (36 %) pertenecían a regiones con DB acusado. Las regiones que pudieron ser validadas entre plataformas Agilent poseían una longitud equivalente a la que puede detectada por un BAC.

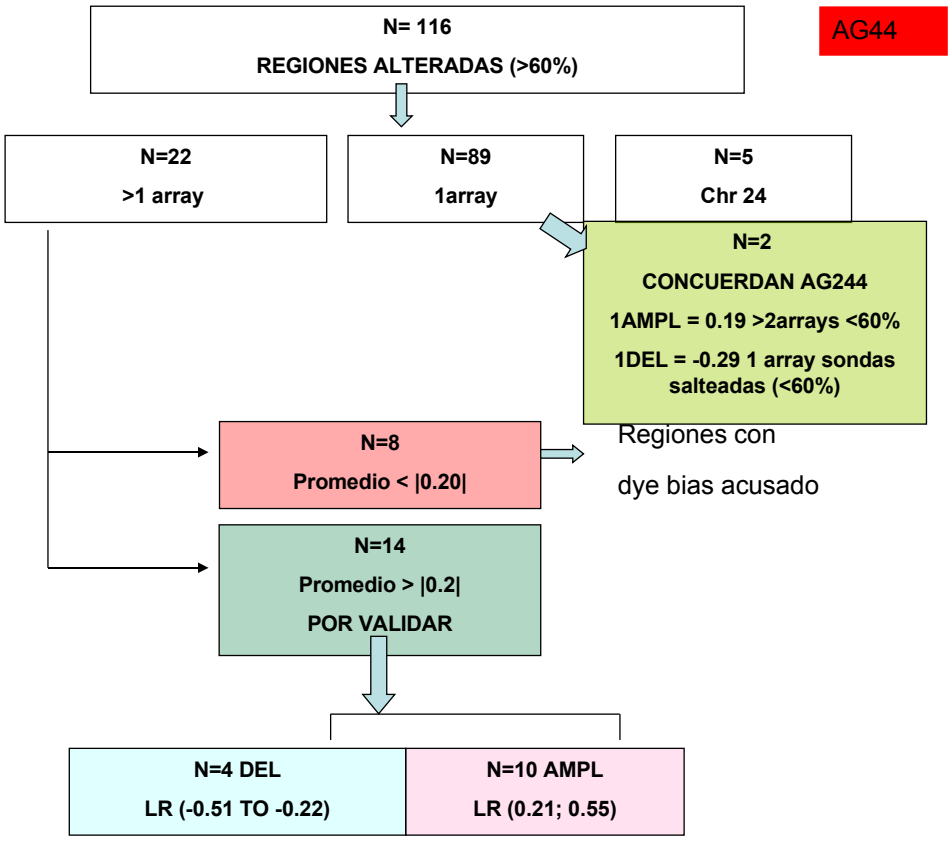

Figura 38: **Muestra las regiones detectadas en Agilent 44K** como posibles regiones (N) a contener CNV.

4.7. Estudio de la expresión génica en aneuploidías

Se ha utilizado el síndrome de Williams-Beuren (WBS) como modelo para establecer una correlación entre los datos obtenidos mediante matrices aCGH con la expresión génica global obtenida a partir de aExpr. Este estudio, además, permite definir las posibles vías alteradas en estos pacientes.

Se ha analizado muestras de cinco pacientes de género masculino y una paciente de género femenino; cuatro varones con fenotipo clásico WBS (sw3, sw5, sw263 y sw266) y un paciente de género masculino (nw10) y una paciente de género femenino (nw35) con fenotipo parcial que incluye un coeficiente intelectual (CI) dentro de la normalidad y ausencia de problemas de integración visual espacial. Los seis pacientes presentaban deleciones heterocigotas en 7q11.23 caracterizadas previamente (ver Figura 39).

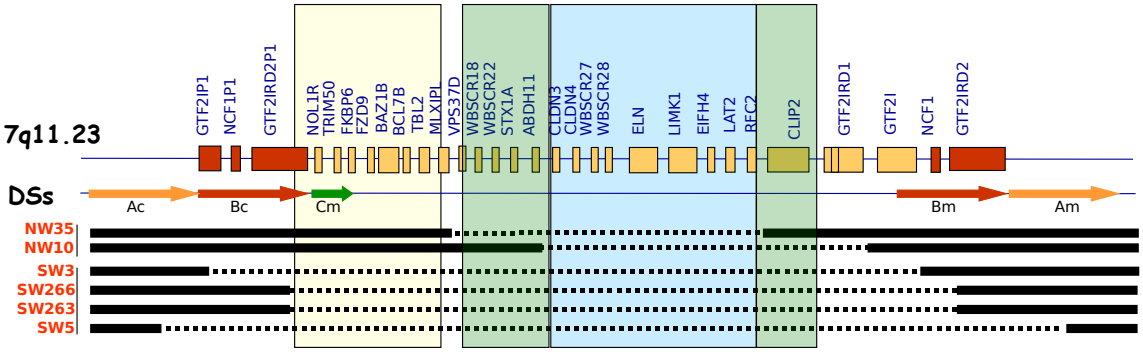

Figura 39: **Se describe el tamaño de las deleciones presentes en los pacientes estudiados**. La línea discontínua simboliza la presencia de deleción. El recuadro amarillo señala una región dónde los individuos sw presentan deleción pero no el grupo nw. El recuadro verde representa una región dónde sólo un individuo nw (nw10) no presenta la deleción. La recuadro azul representa una región delecionada común a las seis muestras bajo estudio.

4.7.1. Detección de ganancias y pérdidas en WBS

Con la finalidad de detectar las alteraciones en el número de copias presentes en las líneas celulares procedentes de los cinco pacientes varones, se han realizado dos hibridaciones por cada muestra (HD y HR) en matrices aCGH. Se ha empleado dos métodos para la detección de alteraciones; (i) se ha empleado el método combinado PCmin y (ii) se ha aplicado un modelo ANOVA que ha permitido obtener una estimación del efecto grupo entre los pacientes analizados y un set de 10 controles sanos de género masculino.

En la Figura 40 se puede observar que el método combinado PCmin ha permitido detectar correctamente la posición de la deleción en 7q11.23 en cada una de las muestras estudiadas. Así mismo, la aplicación de la técnica aCGH sobre las líneas celulares procedentes de los cinco pacientes de género masculino ha permitido detectar otras variaciones que pueden ser debidas a CNVs presentes en los individuos bajo estudio o bien pueden ser debidas a reordenamientos producidos en la línea celular.

BAC	Chr	Posición	Banda	nw10	sw3	sw5	sw263	sw266	Nalt	ANOVA	CNVs	ExprAg44	DS
RP11-15J7	2	88979593	chr2p11.2		↓	↓	↓	↓	4	-0,78	Sí	Sí	Sí
RP11-294I20	2	88997519	chr2p11.2			↓	↓	↓	3	-0,27	Sí	Sí	Sí
RP11-316G9	2	89665752	chr2p11.2				↓	↓	2	-0,07	Sí	Sí	Sí
RP11-101D2	7	72259198	chr7q11.23			↓		↓	2	-0,22	No	Sí	Sí
RP11-329B5	7	73174290	chr7q11.23	↓	↓	↓	↓	↓	5	-0,26	No	Sí	Sí
RP11-196F10	7	73317260	chr7q11.23	↓	↓	↓	↓	↓	5	-0,47	No	Sí	Sí
RP11-180C6	7	73441012	chr7q11.23		↓	↓	↓	↓	4	-0,31	No	Sí	Sí
RP4-771P4	7	73521013	chr7q11.23			↓			1	-0,18	No	Sí	Sí
RP11-694L21	8	86724325	chr8q21.2		↓	↓	↓		3	-0,28	Sí	No	Sí
RP11-48I22	8	86757502	chr8q21.2		↓	↓			2	-0,25	Sí	No	Sí
RP11-279F15	13	56562261	chr13q21.1	↑	↓	↑	↓		4	0,01	Sí	No	No
RP11-118E23	15	18473375	chr15q11.2		↓	↓		↓	3	-0,32	Sí	Sí	Sí
RP11-717D19	15	18633824	chr15q11.2			↓		↓	2	-0,22	Sí	Sí	Sí
RP11-32B5	15	19154639	chr15q11.2			↓		↓	2	-0,26	Sí	Sí	Sí
RP11-2F9	15	19791326	chr15q11.2			↓		↓	2	-0,22	Sí	Sí	Sí
RP11-603B24	15	19970467	chr15q11.2			↓		↓	2	-0,28	Sí	Sí	Sí
RP11-264M14	16	33282423	chr16p11.2	↓			↓		2	-0,19	Sí	Sí	Sí
RP11-545B4	16	33501545	chr16p11.2				↓		1	-0,1	Sí	Sí	Sí
RP11-488I20	16	34359841	chr16p11.2			↓			1	-0,18	Sí	Sí	Sí
RP11-274A17	16	34728914	chr16p11.1				↓		1	-0,09	Sí	Sí	Sí
CTC-251H24	19	18202507	chr19p13.11		↑	↑			2	0,58	No	Sí	No
CTC-260F20	19	19418918	chr19p13.11		↑	↑			2	0,28	No	Sí	No
RP11-50L23	22	21395304	chr22q11.22		↓		↓	↓	3	-0,37	Sí	Sí	Sí
CTD-2149G5	23	152843981	chrXq28	↑	↓				2	0,19	Sí	Sí	Sí
RP11-333O6	23	152880550	chrXq28	↑	↓				2	0,19	Sí	Sí	Sí
RP11-330B2	23	152880550	chrXq28	↑	↓				2	0,17	Sí	Sí	Sí
CTD-2238E23	23	152939205	chrXq28	↑					1	0,12	Sí	Sí	Sí
Total Alt				15	42	55	16	21					

Figura 40: **Regiones recurrentes con pérdidas y/o ganancias**.BAC (Nombre de la sonda utilizada para la detección), Chr (Cromosoma), Posición (Posición inicial), Banda (Banda cromosómica), nw10, sw3, sw5, sw263, sw266 (pacientes estudiados), Nalt (Número de veces que se ha detectado una alteración), ANOVA (Estimación del efecto grupo), CNVs (presencia de CNVs en la región, Marzo 2008), ExprAg44 (Presencia de sondas en la matriz de expresión de la región), DS (Presencia de DS flanqueando la región).

Se ha hallado una correlación entre la pérdida de material genético en la región 7q11.23 y la expresión de esta región (ver Figura 41). Sin embargo no se ha hallado una clara relación entre el resto de regiones con pérdidas y/o ganancias de material genético y la expresión. La expresión no se ha podido medir en las regiones 8q21.2 y 13q21.1 debido a la falta de sondas en la matriz aExpr. En la región 22q11.22 sólo había dos sondas por lo que los resultados obtenidos no son concluyentes y, en el resto de regiones o no se han detectado cambios en la expresión (19p13.11 y 16p11.2) o bien se han detectado cambios en la expresión entre individuos (2p11.2, 15q11.2, 22q.11.22 y Xq28) pero no presentan un perfil concordante con los datos aCGH. Es decir, no es posible asociar una pérdida o ganancia de material genético con una expresión diferencial respecto los individuos sin cambios.

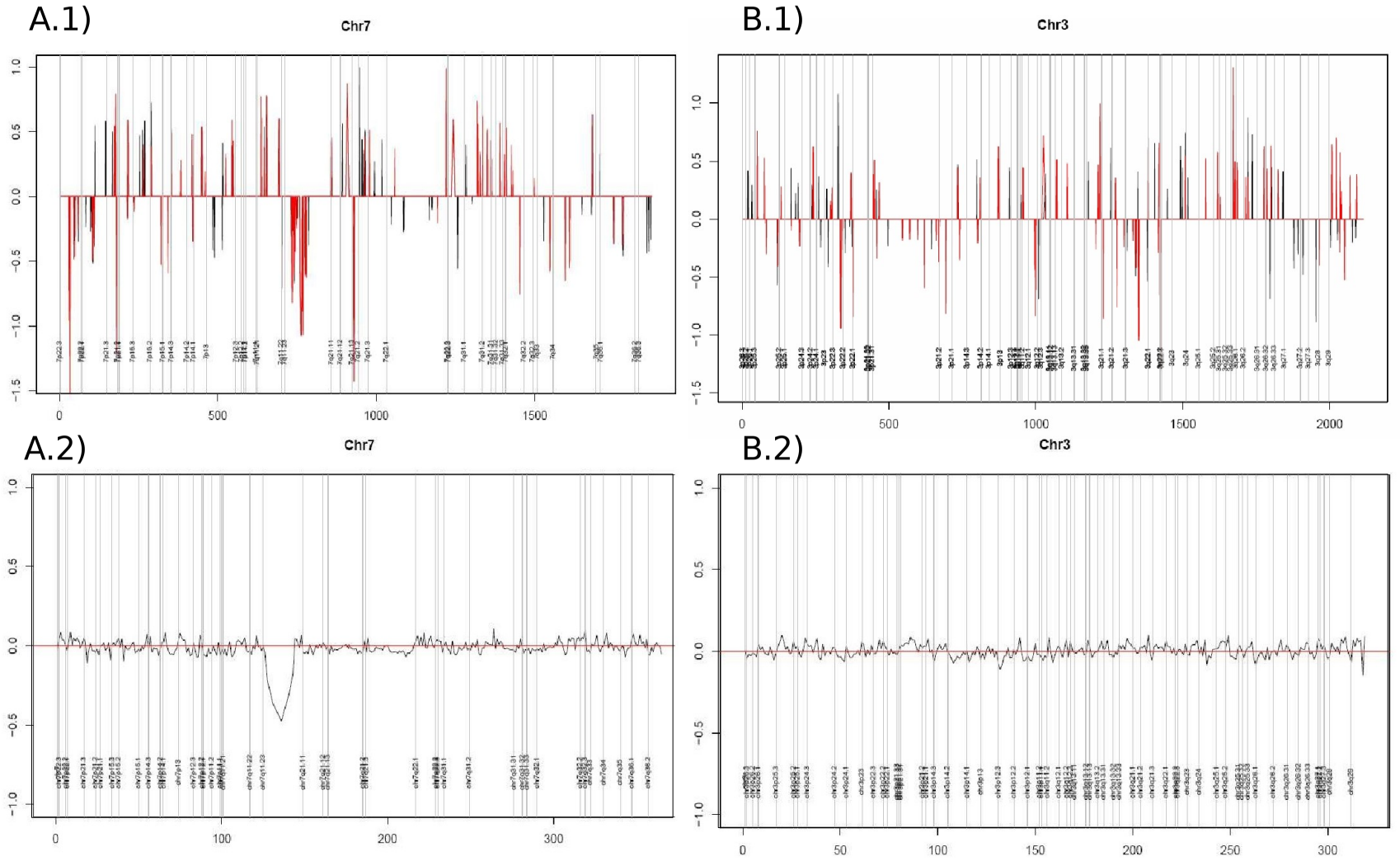

Figura 41: **Cambios en la expresión génica producidos por una aneuploidía parcial**; En la Figura 41 se representan dos cromosomas, a la izquierda se observa el cromosoma 7 afectado por una deleción en la banda 7q11.23 y, a la derecha se observa el cromosoma 3 no afectado por ninguna ganancia o pérdida. En esta Figura se puede observar el tamaño de la deleción caracterizado mediante aCGH (parte inferior) y en la parte superior la media de la expresión de las sondas significativamente diferentes de 0 mediante el test t-Student. A.1) se puede apreciar la presencia de una región con mayor densidad de sondas significativas en el cromosoma 7 correspondiente con B.1) la pérdida de material génico de la banda 7q11.23. Este efecto no es observable en el cromosma 3 (B.1 y B.2).

4.7.2. Estudio de la expresión génica en la región WBS y en las regiones flanqueantes

Estudio comparativo de la expresión génica entre regiones

Este apartado se divide en dos partes; (i) formulación y desarrollo del modelo ANOVA y (ii) resultados obtenidos de la aplicación del modelo.

Formulación y desarrollo del modelo ANOVA

$$y_{giemr} = \mu + Gen_g + DB_i + Muestra_m + Region_e + DB:Region_{ie} + \\ + DB:Muestra_{im} + Region:Muestra_{em} + DB:Region:Muestra_{iem} + e_{giemr}$$

$$\begin{array}{lll}
\sum_{g=1}^{G} Gen_g = 0 & H_0 : Gen_g = 0 & \forall g = 1, G \\
\sum_{i=1}^{I} DB_i = 0 & H_0 : DB_i = 0 & \forall i = 1, I \\
\sum_{m=1}^{M} Muestra_m = 0 & H_0 : Muestra_j = 0 & \forall m = 1, M \\
\sum_{e=1}^{E} Region_e = 0 & H_0 : Region_k = 0 & \forall e = 1, E \\
\sum_{i=1}^{I}\sum_{e=1}^{E} DB:Region_{ie} = 0 & H_0 : DB:Region_{ie} & \forall i = 1, I; e = 1, E \\
\sum_{i=1}^{I}\sum_{m=1}^{M} DB:Muestra_{im} = 0 & H_0 : DB:Muestra_{im} & \forall i = 1, I; m = 1, M \\
\sum_{e=1}^{E}\sum_{m=1}^{M} Region:Muestra_{me} = 0 & H_0 : Region:Muestra_{me} & \forall e = 1, E; m = 1, M \\
\sum_{i=1}^{I}\sum_{e=1}^{E}\sum_{m=1}^{M} DB:Region:Muestra_{iem} = 0 & H_0 : DB:Region:Muestra_{iem} = 0 & \forall i = 1, I; e = 1, E; m = 1, M
\end{array} \quad (5)$$

Tabla 28: **Tabla ANOVA; expresión en la región WBS y flanqueantes**

Factor	gdl[a]	SS[b]	MS[c]	F[d]
Gen	$(G-1)$	SS_{Gen}	MS_{Gen}	MS_{Gen}/MS_R
Muestra	$(M-1)$	$SS_{Muestra}$	$MS_{Muestra}$	$MS_{Muestra}/MS_R$
Region	$(E-1)$	SS_{Region}	MS_{Region}	MS_{Region}/MS_R
DB:Region	$(G-1)(E-1)$	$SS_{DB:Region}$	$MS_{DB:Region}$	$MS_{DB:Region}/MS_R$
DB:Muestra	$(I-1)(M-1)$	$SS_{DB:Muestra}$	$MS_{DB:Muestra}$	$MS_{DB:Muestra}/MS_R$
Region:Muestra	$(M-1)(E-1)$	$SS_{Region:Muestra}$	$MS_{Region:Muestra}$	$MS_{Region:Muestra}/MS_R$
DB:Region:Muestra	$(I-1)(E-1)(M-1)$	$SS_{DB:Region:Muestra}$	$MS_{DB:Region:Muestra}$	$MS_{DB:Region:Muestra}/MS_R$
Residuo	$nobs - sum$	SS_R	MS_R	
Total	$nobs - 1$			

Dónde gen actúa como un bloque englobando todas las sondas que le representan (de una a cuatro). Todos los factores se consideran fijos excepto el bloque, Gen. La variable respuesta es el logaritmo en base dos del ratio entre la muestra test y el *pool* de referencia.

Aplicación del modelo ANOVA y resultados derivados

Mediante el modelo descrito en la ecuación 5 se ha estudiado la relación entre la expresión génica de 21 genes incluidos en la región de WBS, de las regiones flanqueantes a ambos lados de la deleción hasta un máximo de 2 Mb (UPS representada por 3 genes y DWS representada por 11 genes), de dos regiones de características similares (es decir regiones flanqueadas por DS; región C1 representada por 10 genes y región C2 representada por 10 genes) y un conjunto de genes escogidos al azar (EXT representada por 8 genes). Para este fin se han utilizado únicamente las muestras con fenotipo WBS (N=4)

La resolución del modelo se presenta en la Tabla 29 de dónde se deduce que existen diferencias significativas entre regiones. En la Figura 42 se representan los valores de expresión obtenidos por cada región.

Las pruebas post-hoc de Tukey-Scheffé no detectan diferencias significativas entre las regiones UPS, DWS y EXT. Indicando que no existen pruebas sólidas de una deregulación en la expresión génica para las regiones flanqueantes mayor de la que se pueda dar en otras regiones del genoma.

No se ha hallado un diferencias significativas entre las interacciones DB:Región ni DB:Muestra. En un modelo reducido realizado sin estas interacciones se ha hallado los mismos resultados.

Tabla 29: **Tabla ANOVA:**Resultados para la expresión diferencial entre regiones.

Factor	gdl[a]	SS[b]	MS[c]	F[d]	Pvalor
Región	5	24,38	4,88	86,78	$< 0{,}001$
DB	1	0,59	0,59	10,56	0,001
Muestra	3	0,76	0,25	4,52	0,004
Gen	57	31,31	0,55	9,78	$< 0{,}001$
DB:Muestra	3	0,02	0,01	0,14	0,938
DB:Región	5	2,76	0,55	9,81	$< 0{,}001$
Region:Muestra	15	2,3	0,15	2,73	$< 0{,}001$
DB:Region:Muestra	15	0,2	0,01	0,24	0,999
Residuo	743	41,74	0,06		

[a]: (gdl) grados de libertad
[b]: (SS) Suma de Cuadrados
[c]: (MS) Cuadrados Medios
[d]: (F) Estadístico F de Fisher

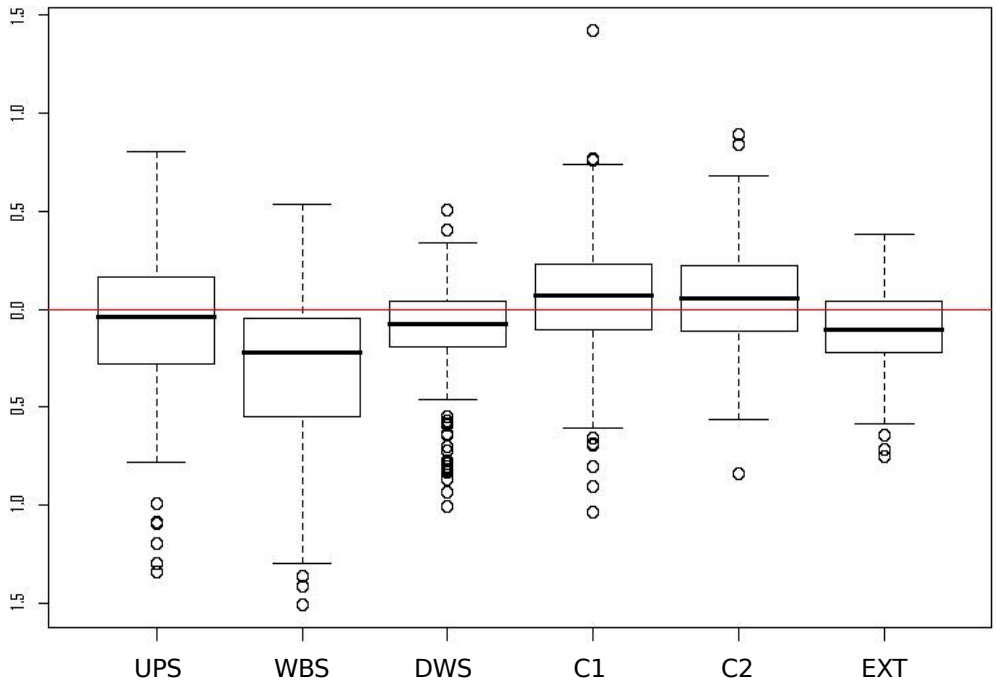

Figura 42: **Muestra las diferencias existentes en la expresión entre regiones**

Estudio de la expresión génica en la región WBS y flanqueantes

Este apartado se divide en dos partes; (i) formulación y desarrollo del modelo ANOVA y (ii) resultados obtenidos de la aplicación del modelo.

Formulación del Modelo ANOVA

Nótese que en el modelo de la ecuación 6 se han eliminado las interacciones DB:Grupo y DB:Muestra, al resultar ésta última no significativa en el modelo anterior realizado sobre los mismos datos.

$$y_{gimur} = \mu + Gen_g + DB_i + Muestra_m + Grupo_u + Gen:Grupo_{gu} + \\ + Gen:Muestra_{gm} + e_{gimur}$$

$$\begin{array}{lll} \sum_{g}^{G} Gen_g = 0 & H_0 : Gen_g = 0 & \forall g = 1...G \\ \sum_{i}^{I} DB_i = 0 & H_0 : DB_i = 0 & \forall i = 1, I \\ \sum_{m}^{M} Muestra_m = 0 & H_0 : Muestra_j = 0 & \forall m = 1, M \\ \sum_{u}^{U} Grupo_u = 0 & H_0 : Grupo_u = 0 & \forall u = 1, U \\ \sum_{g}^{G}\sum_{m}^{M} Gen:Muestra_{gm} = 0 & H_0 : Gen:Grupo_{gm} = 0 & g = 1, G; m = 1, M \\ \sum_{g}^{G}\sum_{u}^{U} Gen:Grupo_{gu} = 0 & H_0 : Gen:Grupo_{gu} = 0 & g = 1, G; u = 1, U \\ e_{gimur} \sim N(0, \sigma) & & \end{array} \quad (6)$$

Tabla ANOVA de resolución del modelo

Tabla 30: **Tabla ANOVA: Relación entre sw vs nw y regiones**:

Factor	gdl	SS	MS	F
Gen	$(G-1)$	SS_{Gen}	MS_{Gen}	MS_{Gen}/MS_R
Muestra	$(M-1)-(U-1)$	$SS_{Muestra}$	$MS_{Muestra}$	$MS_{Muestra}/MS_R$
DB	$(I-1)$	SS_{DB}	MS_{DB}	MS_{DB}/MS_R
Grupo	$(U-1)$	SS_{Grupo}	MS_{Grupo}	MS_{Grupo}/MS_R
Gen:Muestra	$(G-1)[(M-1)-(U-1)]$	$SS_{Gen:Muestra}$	$MS_{Gen:Muestra}$	$MS_{Gen:Muestra}/MS_R$
Gen:Grupo	$(G-1)(U-1)$	$SS_{Gen:Grupo}$	$MS_{Gen:Grupo}$	$MS_{Gen:Grupo}/MS_R$
Residuo	nobs-sum	SS_R	MS_R	
Total	nobs-sum	SS_R	MS_R	

Aplicación del Modelo ANOVA y resultados derivados

Se realizó un modelo separado para cada región (UPS, DWS y WBS) con el objetivo de conocer el comportamiento de las muestras y de los grupos en cada una de ellas. Con esta finalidad se utilizaron seis muestras que se clasificaron en dos grupos; con fenotipo WBS (sw, N=4) y con fenotipo WBS parcial (nw, N=2). El modelo empleado se representa en la ecuación 6.

En la región UPS (ver Tabla 31) se hallan diferencias significativas entre genes. Estas diferencias son debidas al gen AUTS2 que aparece deregulado en todas las muestras excepto en nw35. AUTS2 está sobrexpresado en la muestra sw266 e infraexpresado en el resto.

Tabla 31: **Tabla ANOVA para la región UPS**

Factor	gdl [a]	SS [b]	MS [c]	F [d]	Pvalor
Gen	2	2,941	1,471	33,878	$< 0,001$
Grupo	1	0,009	0,009	0,199	0,658
Muestra	4	0,961	0,240	5,535	0,001
DB	1	1,128	1,128	25,980	$< 0,001$
Gen:Muestra	8	4,513	0,564	12,995	$< 0,001$
Gen:Grupo	2	0,507	0,254	5,844	0,006
Residuo	41	1,780	0,043		

[a]: (gdl) grados de libertad
[b]: (SS) Suma de Cuadrados
[c]: (MS) Cuadrados Medios
[d]: (F) Estadístico F de Fisher

Aplicando la ecuación 6 para la región DWS se aprecian diferencias significativas entre genes y muestras sin que exista un efecto grupo (ver Tabla 32). El comportamiento deregulado es más acusado en las muestras nw10 y sw5. En esta región se han detectado dos genes deregulados; (i) TMEM120A que está infraexpresado en todas las muestras excepto en nw10 (ii) HSP1B está infraexpresado en todas las muestras.

Tabla 32: **Tabla ANOVA para la región DWS**

Factor	gdl[1]	SS[2]	MS[3]	F[4]	Pvalor
Gen	10	3,950	0,395	13,770	$< 0,001$
Grupo	1	0,087	0,087	3,026	0,083
Muestra	4	0,397	0,099	3,461	0,009
DB	1	0,057	0,057	2,000	0,159
Gen:Grupo	10	0,474	0,047	1,654	0,095
Gen:Muestra	40	3,958	0,099	3,501	$< 0,001$
Residuo	173	4,9618	0,0287		

[a]: grados de libertad
[b]: Suma de Cuadrados
[c]: Cuadrados Medios
[d]: Estadístico F de Fisher

En la región WBS se observan diferencias significativas entre muestras en el sentido esperado, es decir una infraexpresión mayor se corresponde con deleciones más grandes

de la región. Aunque la muestra sw266, que presenta una deleción menor que sw3, tiene niveles de infraexpresión mayor de lo que le correspondería.

Tabla 33: **Tabla ANOVA para la región WBS**

Factor	gdl[a]	SS[b]	MS[c]	F[d]	Pvalor
Gen	20	27,583	1,379	18,925	$< 0,001$
Grupo	1	1,048	1,048	14,382	$< 0,001$
Muestra	4	1,688	0,422	5,790	$< 0,001$
DB	1	0,015	0,015	0,209	0,648
Gen:Grupo	20	5,018	0,251	3,443	$< 0,001$
Gen:Muestra	80	5,855	0,073	1,004	0,475
Residuo	353	25,723	0,073		

[a]: grados de libertad
[b]: Suma de Cuadrados
[c]: Cuadrados Medios
[d]: Estadístico F de Fisher

En el gráfico de interacción grupo y gen que se muestra en la Figura 43 puede apreciarse dos regiones con expresión diferencial entre sw y nw (región amarilla y región verde).

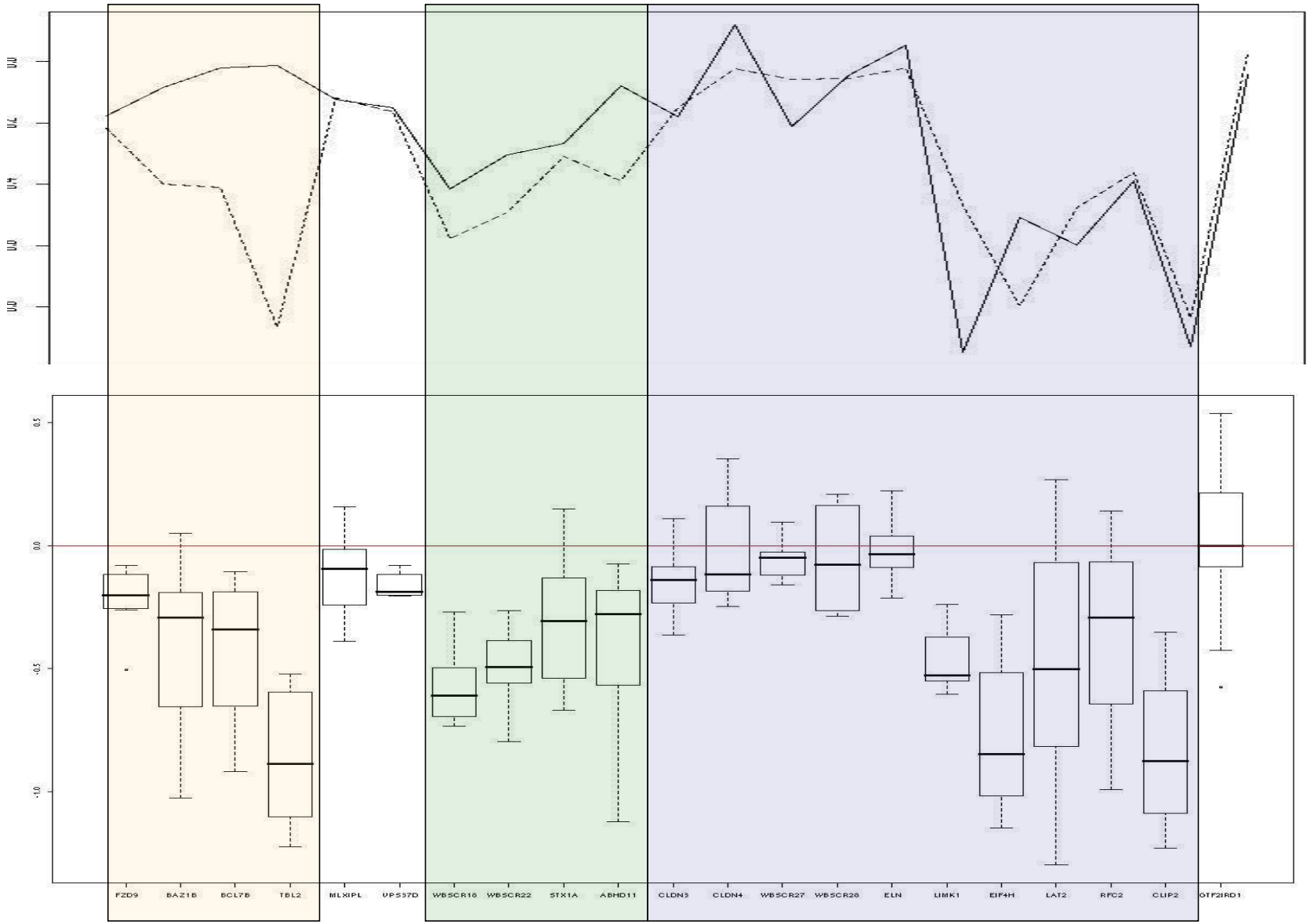

Figura 43: **Muestra la interacción grupo (sw versus nw) y gen**. Arriba se muestra en línea contínua el perfil para nw y en línea discontinúa el perfil para sw. Abajo se muestra la expresión para cada gen sólo para el grupo sw. En amarillo se representa una primera región con diferencias significativas entre gen y grupo, en este bloque las muestras nw no presentan deleción mientras que las muestras sw sí. La región verde muestra una interacción gen y grupo con deleción común para todas las muestras sw y deleción en una muestra nw. En la región azul todas las muestras sw y nw presentan una deleción.

4.7.3. Análisis transcriptómico global

En la Figura 46 se muestra el procedimiento utilizado en el análisis de la expresión génica así como los principales resultados obtenidos.

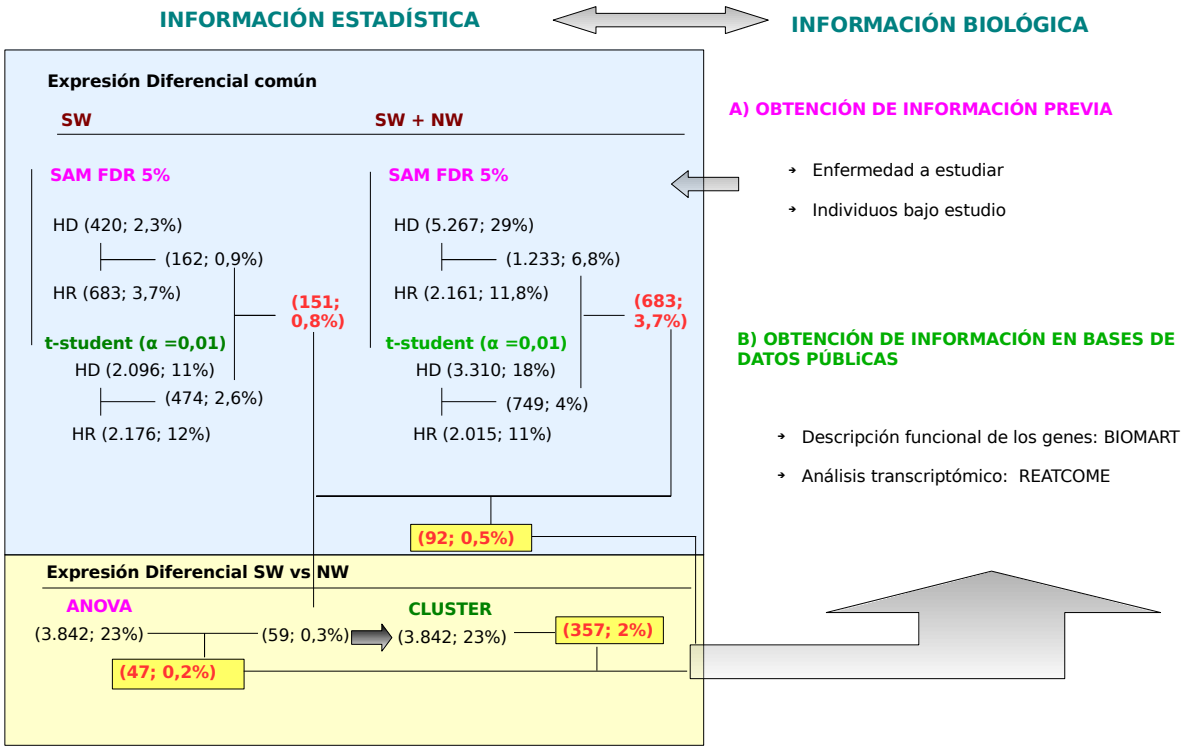

Figura 44: **Expresión géncia global**. Se muestran los principales resultados obtenidos de los distintos tests aplicados sobre 18.228 genes presentes en la matriz de expresión utilizada.Para conocer que vías se hallan alteradas en todos los individuos estudiados se ha aplicado el test SAM con un FDR del 5 % y el test t-student tomando como significativos aquellos genes que obtuvieran un pvalor menor de 0,01. Mientras que para conocer que vías se hallan diferencialmente expresadas entre estos dos grupos (sw versus nw) se aplicó un test ANOVA para cada gen (g) según ecuación

La proporción de genes sobreexpresados e infraexpresados según los distintos tests aplicados se detallan en las Figuras 46 y 45. El 63 % de los genes diferencialmente expresados en el grupo sw + nw (N=92) están sobre-expresados (ver Figura 45).

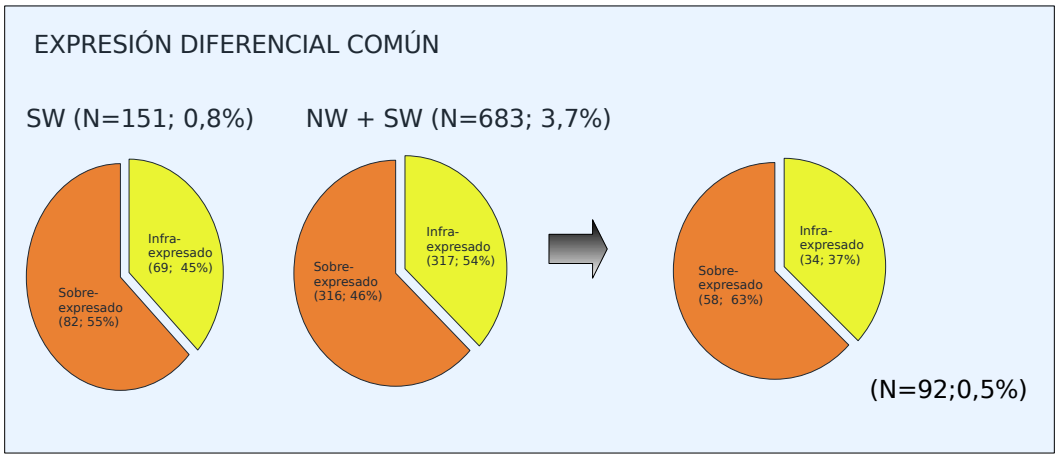

Figura 45: **Expresión diferencial común (sw +nw)**. Se muestran los principales resultados obtenidos de los distintos tests aplicados sobre 18.228 genes en los grupos sw y sw + nw.

En cambio, cuando se compara el grupo sw con nw un 64 % de los genes en sw están infra-expresados versus el grupo nw.

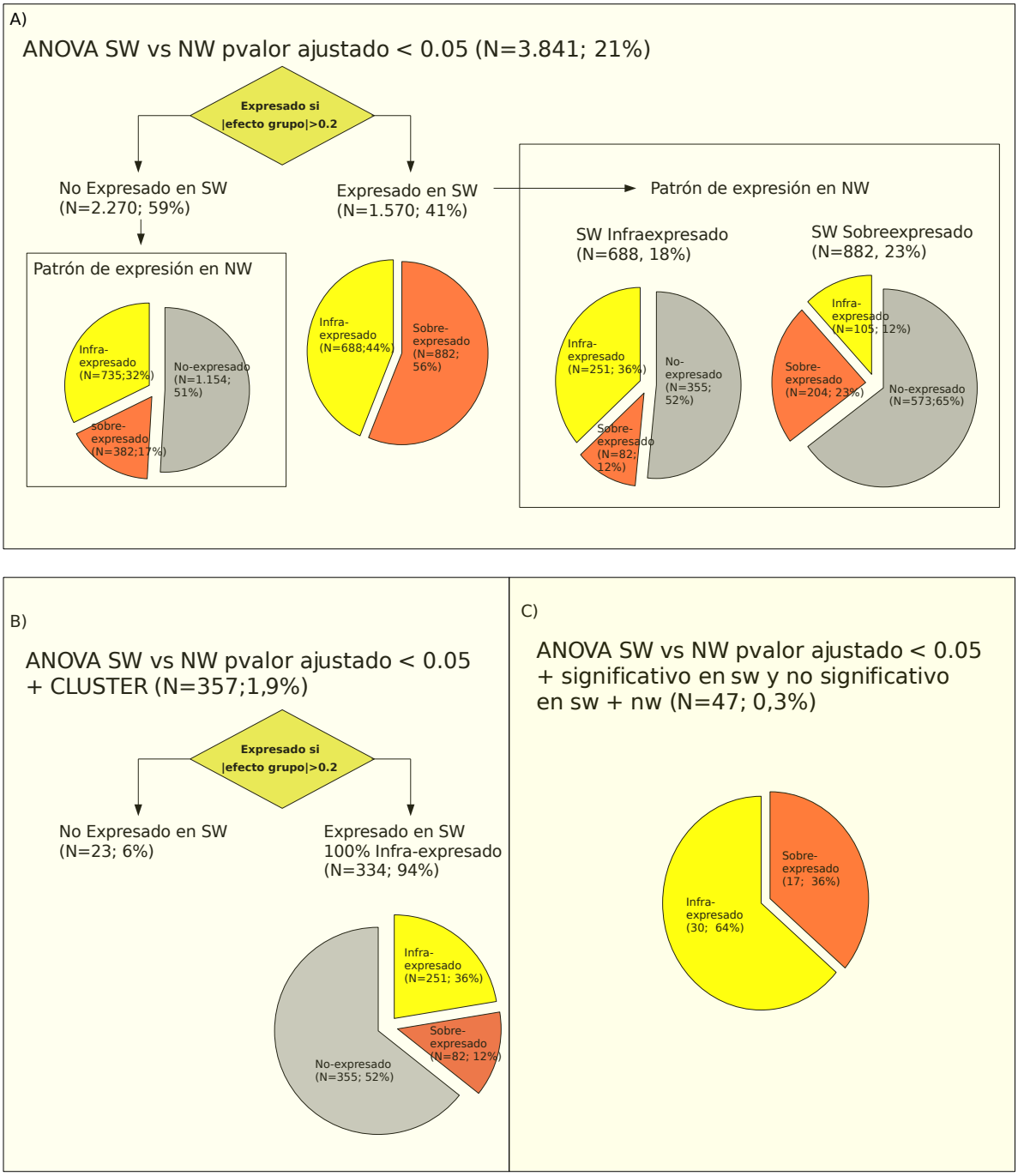

Figura 46: **Expresión diferencial grupo sw versus nw**. Se ha aplicado un punto de corte ($> |0,2|$) sobre la estimación del efecto grupo en los genes diferencialmente expresados (N=3.841). Considerando, así, que los genes diferencialmente expresados son aquellos genes significativos para ANOVA y con una magnitud entre grupos superior al umbral. Está restricción elimina 1.154 (30 %) de los genes que resultaron significativos según ANOVA

.

Finalmente, los sets de N=92, N=47 y N=357 se enviaron a la base de datos REACTOME y mediante la herramienta *Skypainter* se evaluó las vías más representadas en ellos. Las principales vías detectadas se muestran en la Tabla 34. Las vías más representadas

en el set de datos N= 92 (sw + nw) fueron la glicolisis y la gluconeogénesis así como el metabolismo de las lipoproteínas. Las vías más representadas en los sets de datos N=47 y N=357 (sw versus nw) fueron el ciclo de Cori, el transporte celular, el sistema inmune innato y el metabolismo de nucleótidos.

Tabla 34: **REACTOME: Principales vías metabólicas afectadas**

Datos Evaluados	Sign	Vía metabólica	Genes Candidatos	Procesos biológicos
A (N=92)	No	Metabolismo de las pequeñas moléculas	GYS1;PKM2	Gluconeogénesis/ Glicolisis
A (N=92)	No	Metabolismo de las lipoproteínas	MCEE;AGPAT3	
B (N=47)	Sí	Metabolismo de las pequeñas moléculas	PFKP;LDHC;PGK1; TPI1;ENO1	Gluconeogénesis / Glicolisis
B (N=47)	Sí	Transporte Celular	KIF14;SNX4;MAP1B; USO1;BCL7B	Formación de microtúbulos/ Contracción celular
C (N=357)	Sí	Metabolismo de las pequeñas moléculas	PGK1;TPI1;PFKP; ENO1;GPI;GAPDH; PGM1;ALDOC;PFKB4	Gluconeogénesis / Glicolisis
C (N=357)	Sí	Sistema inmune innato	CFD;IGHG1;IGKV4-1	Cascada del Complemento
C (N=357)	Sí	Metabolismo de nucleótidos	EIF4H;TK1;NME	
C (N=357)	No	Metabolismo de lipoproteínas	FABP6	
C (N=357)	No	Ciclo NOTCH	JAG1	
C (N=357)	No	Metabolismo de la vitamina K	VKORC1	

A; Datos procedentes de genes diferencialmente expresados en sw, sw+nw
B; Datos procedentes de genes diferencialmente expresados en sw vs nw según ANOVA y sw
C; Datos procedentes de genes diferencialmente expresados en sw vs nw según ANOVA y Cluster

Finalmente, se estudió la posible relación con la enfermedad de los genes hallados en estos sets de datos pero que no formaban parte de vías significativas. Los principales resultados se muestran en la Tabla 35.

Tabla 35: **Descripción de otros genes candidatos**

Datos Evaluados	Gen	Descripción
B (N=47)	PHIP	Receptor de insulina
B (N=47)	PYROXD1	Actividad en la cadena de electrones
B (N=47)	UTS2	Actúa en la transmisión del impulso sináptico y regulador de la presión sanguínea
B (N=47)	PAFAH1B3	Implicado en el desarrollo del sistema nervioso
B (N=47)	WBSCR22	Actividad metil-transferasa
B (N=47)	TAF9	Factor de transcripción
B (N=47)	CCDC881	Factor de transcripción
B (N=47)	SCML1	Factor de transcripción
B (N=47)	GTF2I	Factor de transcripción

B; Datos procedentes de genes diferencialmente expresados en sw vs nw según ANOVA y sw
C; Datos procedentes de genes diferencialmente expresados en sw vs nw según ANOVA y Cluster

En los *heatmaps* de la Figura 47 se representan los clústers de los genes diferencialmente expresados entre los grupos sw y nw. En estos *heatmaps* puede observarse la expresión de cada gen a través de las muestras analizadas. En ellos se han representado el set de datos N=47 y los genes más representativos entre nw y sw en el set=357.

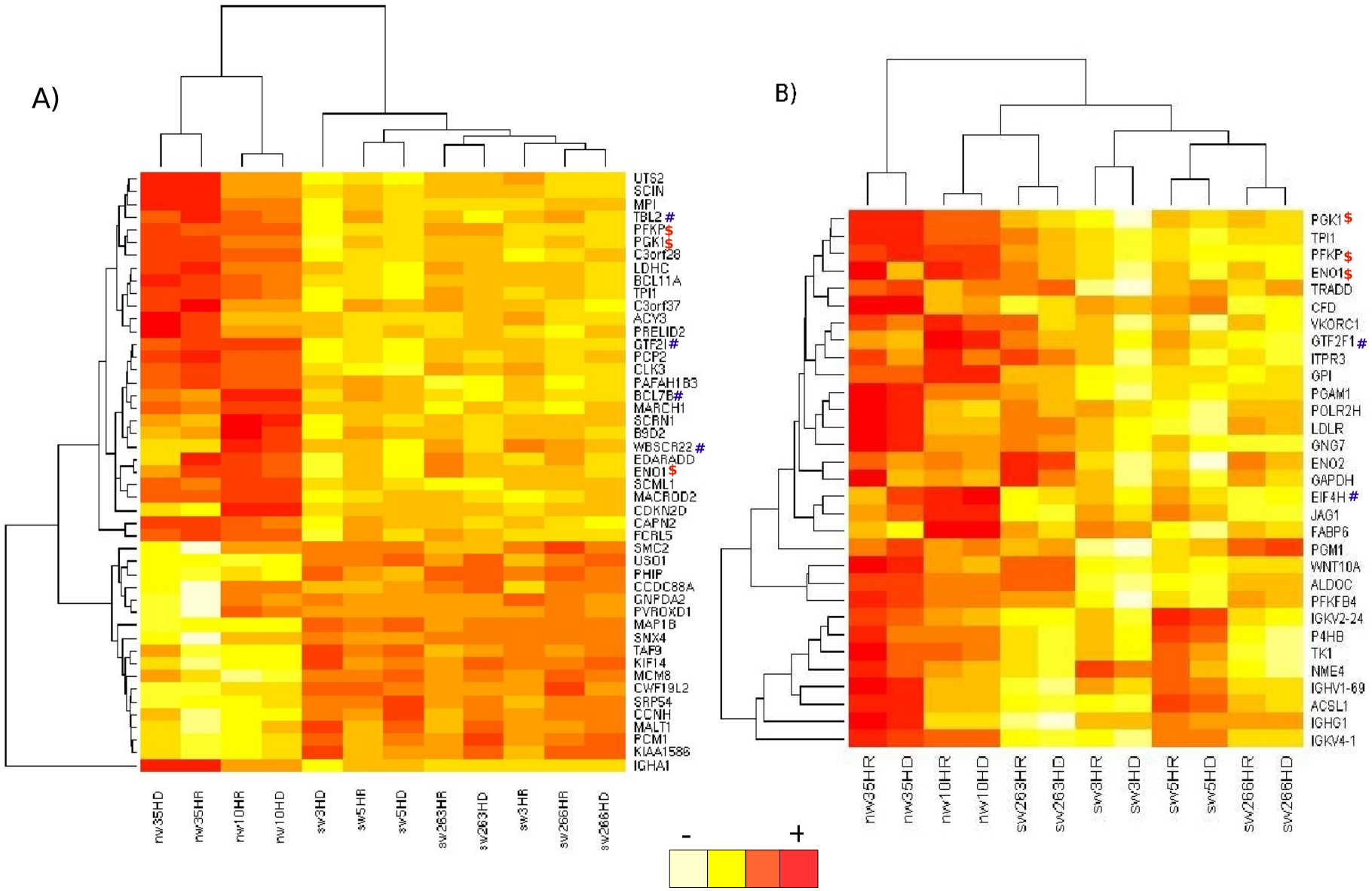

Figura 47: **Genes indentificados con expresión diferencial entre sw y nw** A) Se representan 47 genes significativos para t-test y SAM en sw con diferencias significativas en ANOVA test entre sw y nw. B) Se representan los genes procedentes de vías con sentido biológico indentificados medianta ANOVA y cluster. (♯) Genes de la región WBS,($) Genes en común entre *heatmaps*

4.8. Identificación de PSVs

Para detectar la secuencia y posición de PSVs funcionales a lo largo del genoma humano se han identificado copias génicas;

- con una longitud mayor de 300 pb
- con una identidad copia-gen mayor del 95 %
- que no se correspondan con retrotransposones
- que presentan PSVs funcionales en secuencia codificante o en secuencias de *splicing*.

Algoritmo desarrollado para la detección de PSVs funcionales

El algoritmo desarrollado se basa en tres pasos:

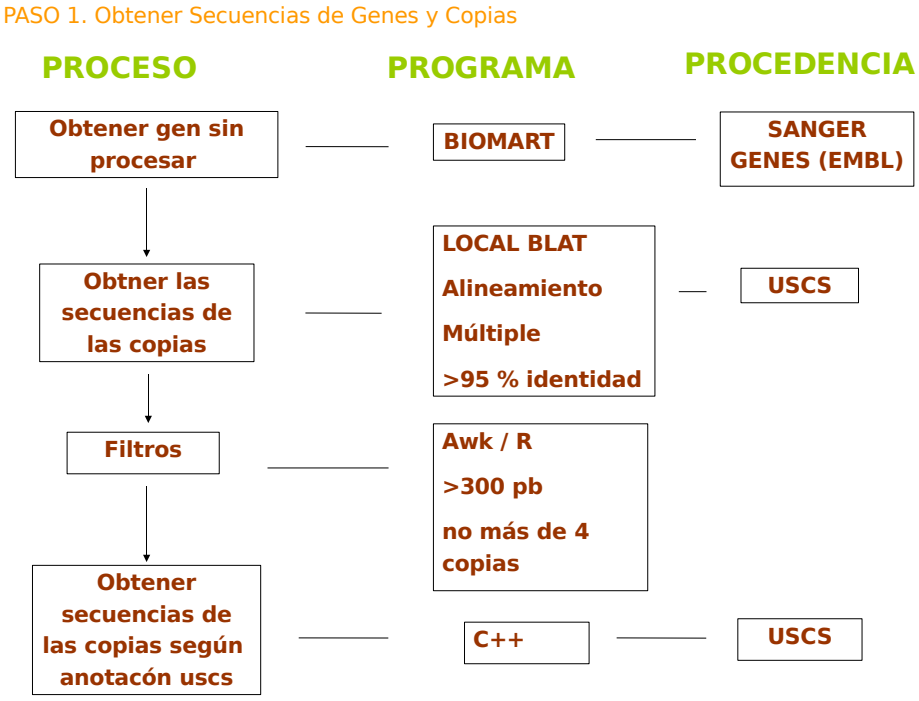

Figura 48: **En este paso se obtienen las secuencias** de los genes no procesados a partir de la base de datos *Biomart* [161]. Se realiza un BLAT [163] de estos genes sobre el genoma y se obtienen las secuencias de las copias que cumplen las condiciones mencionadas.

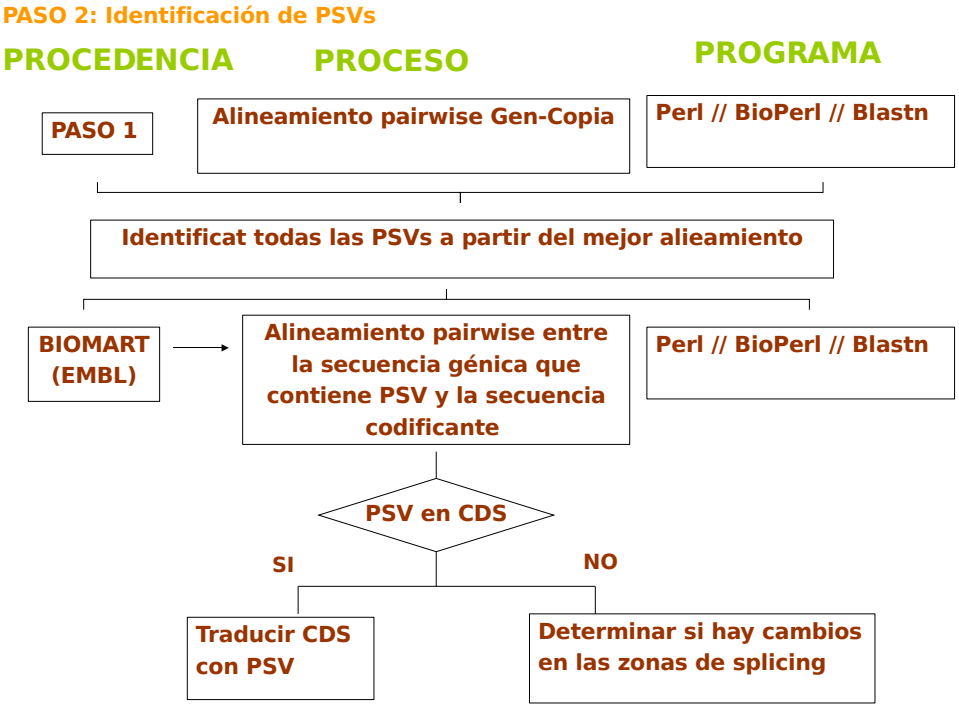

Figura 49: **En este segundo paso se ha realizado un alineamiento *pair-wise* mediante BLAST** entre la secuencia del gen no procesado y cada una de sus copias que elimina los alineamientos por bloque pudiendo evitar, así, en gran parte la retrotransposición. En este paso se obtienenen secuencias de aproximadamente 200pb que contienen los *mis-match* gen-copia en la posición central. Estas secuencias son alineadas mediante un BLAST con la secuencia del gen procesado. Así se puede identificar la posición que ocupa la PSVs en el gen y ello permite conocer si produce un cambio no sinónimo o si produce el silenciamiento de un sitio de *splicing*.

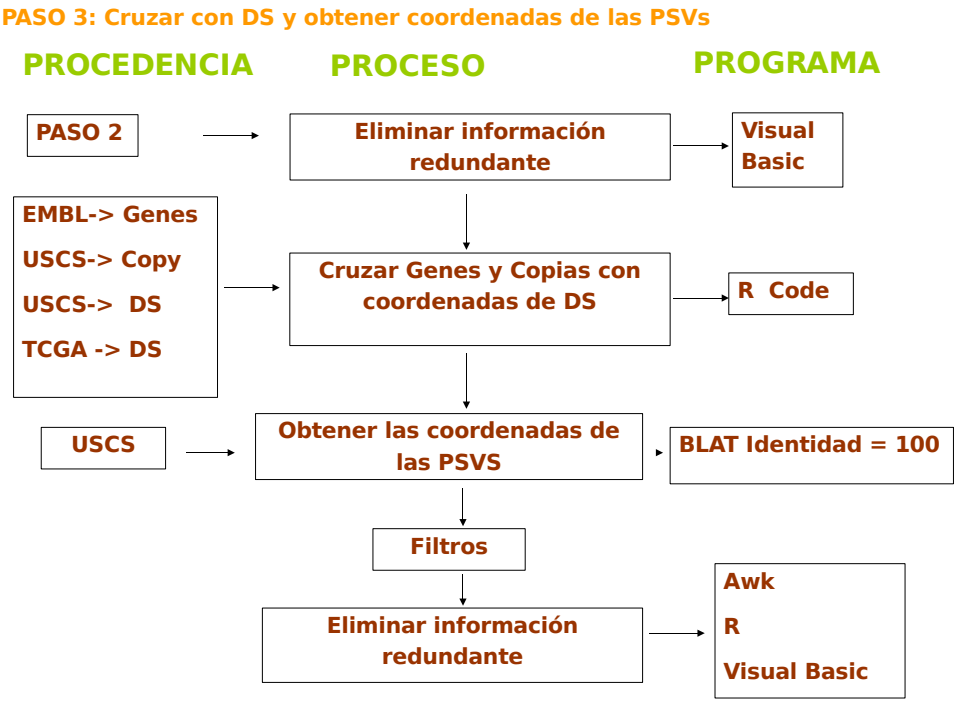

Figura 50: **Se representan los filtros utilizados**. En este paso se han aplicado varios filtros con la intención de obtener información no redundante. Se han eliminado aquellas entradas que procedan de genes no situados en DS.

Principales resultados obtenidos

Este análisis se realizó a fecha de 25 de Enero del 2008. Se testaron 31.079 genes recogidos de la base de datos *Biomart* [161]. Se localizaron 24.489 PSVs funcionales (paso 2) en 1.780 genes. Se detectaron 1.344 entradas únicas que se correspondían con MSVs. De estas, 8 fueron localizadas en los límites intrón-exón (0,6 %).

Una vez filtrados los resultados no presentes en DS o bien redundantes se obtuvieron 13.037 PSVs en 1.115 genes. Los genes que contienen estas PSVs tienen una longitud que varía entre 800 pb y 1,5 Mb con una media alrededor de 30 kb. Las secuencias procesadas tienen una longitud media de 1.000 pb.

En la Tabla 36 se resume el número de copias por gen considerado y el número de PSVs en gen-copias consideradas.

Tabla 36: **Resumen entre el número de copias y número de PSVs por gen y copia**

Copias por Gen	NGenes	NCopias	PSVs por gen	N Genes
1	751	751	[1,5]	978
2	234	468	[6,10]	283
3	83	249	[11,20]	253
4	47	188	[21,]	142
Total	1.115	1.656	Total	1.656

2.034 (16 %)de las 13.037 PSVs están causadas por indel. 11.378 (87 %) PSVs se diferencian por un solo cambio entre gen y copia. En 1.668 (13 %) PSVs la diferencia constaba en más de un cambio. En 47 casos había una base desconocida en el cambio. Las frecuencias de cambio más altas son entre A-G y entre T-C. Se han detectado un total de 57 cambios causantes de la desaparición de un sitio de *splicing*.

5. Discusión

5.1. Estudio piloto

En la literatura, hay autores [152] que realizan réplicas del experimento intercambiando los fluorocromos. Ello es debido a que, por causas aún desconocidas, ciertos clones, en determinadas condiciones, tienden a presentar afinidad hacía un flurocromo determinado traduciéndose en la observación de una ganancia o pérdida falsa. La realización del experimento inverso permite observar si existe realmente una lesión o si el valor observado es debido a una afinidad diferencial. Esta afinidad puede ser debida a determinadas características de los clones que la manifiestan o bien puede ser debida a problemas surgidos durante el proceso de fabricación e hibridación. En este estudio, así como en otros estudios previos [49, 50, 54, 56], se observó la presencia de una distribución espacial de los fluorocromos a lo largo del portaobjetos. Algunos autores han relacionado este efecto espacial con la detección de valores atípicos que, seguramente, está relacionado con el proceso de fabricación y/o hibridación y que puede llegar a producir errores sistemáticos.

En este estudio piloto se realizó un único experimento (HD) y se encontró un alto porcentaje de falsos positivos que se corregía, en gran parte, mediante la estanderización de los valores (indicando la presencia de un error sistemático). La estanderización es viable debido a que los pacientes presentaban un espectro grande de enfermedades y a que no se esperaba alteraciones en todos los casos. Y, aunque la estanderización resultó ser más eficiente en cuanto a sensibilidad y porcentaje de validación que los métodos basados en umbrales o puntos de corte, ésta tiene, también, efectos negativos como la dilución de CNVs frecuentes.

Aún así, este estudio permitió identificar las alteraciones presentes en pacientes con síndromes conocidos y situar correctamente los puntos de rotura de estos. Además se identificó siete regiones candidatas entre los 91 pacientes (7,8 %) estudiados que presentaban retraso mental y/o síndromes dismórficos. El porcentaje de casos identificados es similar a los detectados en estudios previos, por ejemplo en el estudio de 100 casos de retraso mental de Vries et al 2003 [98] se encontró un 10 % de alteraciones submicroscópicas utilizando una matriz aCGH de BACs de 32.447 clones que cubría todo el genoma; Schoumans et al 2005 [36] detectó un 9,8 % de reordenamientos utilizando la matriz comercial de *Spectral genomics* que contiene 2.600 BACs espaciados aproximadamente a una Mb.

5.2. Fuentes de variación en aCGH y sus causas

Con la intención de evitar los efectos adversos que se producen durante la fabricación e hibridación, se optó por realizar un diseño experimental sobre el proceso de impresión de las matrices aCGH. Concretamente, se quería conocer la combinación de; (i) solución de impresión, (ii) concentración de ADN impresa y (iii) tipo de portaobjetos que:

- Maximizara el número de *spots* fiables. Se considera un *spot* fiable cuando; (i) la señal del *spot* respecto el ruido de fondo en, al menos un canal, es mayor de dos veces

y (ii) la variabilidad entre las réplicas del mismo BAC en el mismo portaobjetos no sea mayor de 0,1

- Minimizara la variabilidad del sistema
- Minimizara el efecto DB
- Minimizara la diferencia entre los valores esperados y los promedios obtenidos; para BACs en cromosomas autosómicos la esperanza es 0 mientras que para los BACs en cromosomas sexuales es -1

Las soluciones 50 % DMSO, PRONTO-Amino y PRONTO-Epoxy producen un gran número de *spots* no fiables en los portaobjetos tipo Codelink. Así que se descartaron estas soluciones para este tipo de portaobjetos.

La variabilidad residual de los modelos en los que se ha utilizado como portaobjetos Ultragap es menor que en los modelos con Codelink, además en Ultragaps hay menos fuentes de variación asociadas (menos factores estadísticamente significativos). Por ello se consideró más recomendable utilizar portaobjetos de tipo Ultragaps.

El efecto DB en sí mismo no resultó significativo en los modelos evaluados. Este efecto depende del BAC que se esté considerando (el efecto DB no es significativo pero la interacción sí). Ello implica que hay BACs con más tendencia a presentar problemas de DB. Existe la creencia que este tipo de problemas se corrige mediante el proceso de normalización, así que, este hecho podría estar relacionado con una normalización poco adecuada. Por ello se consideró estudiar el efecto de la normalización sobre el efecto DB en un diseño experimental propuesto para detectar las fuentes de variación más importantes asociadas al proceso de hibridación. De hecho, existe gran controversia entre cuales son los mejores métodos de normalización en matrices aCGH. Los métodos más usados en aCGH son global *loess* y *print-tip loess* [166] con substracción del ruido de fondo que habían sido propuestos, inicialmente, para matrices aExpr. Los métodos de normalización dependen del parámetro amplitud de ventana. Habitualmente se escoge una amplitud de ventana común a los distintos portaobjetos a pesar de no estar recomendado [46] pero distintos estudios muestran que no existen grandes diferencias en el proceso de normalización al usar una u otra amplitud de ventana. En Khojastech *et al* 2005 [166] el intervalo recomendado está entre 0,1 y 0,4 así que se eligió una amplitud de ventana equivalente a 0,3. En este diseño experimental se quiso:

1. Determinar cual es el mejor proceso de normalización, considerando mejor aquel que:
    - Minimice el número de clones afectados por DB
    - Minimice la magnitud del efecto DB
    - Controle el posible efecto espacial del portaobjetos
    - Minimice la variabilidad del portaobjetos
    - Minimice la perdida de información biológica

- Minimice el número de falsos positivos y de falsos negativos
- Permita la máxima automatización del proceso de normalización.

2. Conocer si existe un efecto DB sistemático

3. Conocer las causas del efecto DB si existe.

En anteriores trabajos se ha intentado conocer cual es el mejor tipo de normalización para los datos procedentes de aCGH. Sin embargo estos trabajos toman como único objetivo la minimización de la variabilidad asociada a cada portaobjetos sin tener en cuenta otras fuentes de variabilidad tanto o más importantes [155, 166].Y por ello, el estudio de este tipo de sistemas mediante diseño experimental (como se ha propuesto en este trabajo) permite incrementar la reproducibilidad y la calidad de los datos obtenidos. Así la principal aportación en este trabajo es la de desarrollar una metodología estadística que permite identificar fuentes de variación y, que a la vez, sea interpretable pudiendo discernir apropiadamente la señal y separarla del ruido experimental y/o biológico.

Muchos experimentadores realizan substracción del ruido de fondo sobre el proceso de normalización. Aunque ello es contradictorio con los principales resultados encontrados; Khojasteh *et al* 2005 [166] observó una ligera mejoría, es decir, menor variabilidad y mayor repetibilidad, cuando no se realizaba substracción del ruido de fondo aunque ello no parecía afectar a la detección de CNVs. Los resultados hallados en este trabajo indican que, basándose en la variabilidad de los portaobjetos, los mejores resultados se obtienen sin la substracción del ruido de fondo y, especialmente, utilizando el método *loess loc* y *loess loc scale*. Resultados similares han sido reportados por Fiegler H., *et al* 2006 [167]. Otros efectos no desados asociados a la substracción del ruido de fondo fueron descritos por Qin *et al.* 2004 [168] en matrices aExpr mostrando que la substracción del ruido de fondo incrementa la variabilidad de las sondas y dificulta la detección de genes expresados. Sharpf *et al* 2007 [169] demuestran que la substracción del ruido de fondo sólo es necesaria cuando existe una buena correlación entre éste y la señal del *spot*. En este estudio se ha comprobado la ausencia de correlación entre FG y BG y, por ello, la substracción del ruido de fondo incrementa, incluso, el número de falsos positivos. Por lo tanto, no debería utilizarse. Además, se ha demostrado en este trabajo la presencia de una estructura espacial que incrementa el efecto DB pero que puede corregirse, al menos en parte, mediante normalización; *Print-tip loess* realiza una corrección espacial por bloques que minimiza el efecto DB tanto bajo substracción del ruido de fondo como no, ello le confiere cierta robustez en frente de *loess loc* y *loess loc scale*. Khojasteh *et al.* 2005 [166] recomendaba, en su estudio, la aplicación del método *loess loc*. Este estudio indica que, efectivamente, *loess loc* y *loess loc scale* corrigen mejor los efectos espaciales y reducen el número de falsos positivos pero también puede observarse como los valores esperados en el caso de existencia de una CNV están más alejados de lo esperable reduciendo, así, la sensibilidad. Aunque se considera que pueden ser especialmente adecuados cuando exista un gran efecto espacial que repercuta en un gran número de falsos positivos.

Aún aplicando el método de normalización *loess loc scale* se ha observado estructuras persistentes de DB. Y, además, cuando se compararon los resultados obtenidos entre los

experimentos 19c y 32x este efecto permanecía. Demostrándose, así, que se trata de un error sistemático. Este hecho refuerza la idea de que es necesario conocer mejor el sistema aCGH mediante replicación con la intención de conocer y definir una distribución por cada sonda. Ello permitiría la detección de CNVs basándose en una distribución de referencia por cada sonda ya que, basándonos en la idea de la presencia de un efecto DB, no parece conveniente aplicar a cada sonda el mismo punto de corte. Enfoques parecidos se han propuesto para matrices aExpr [170].

Las mismas estructuras espaciales se han observado en datasets depositados en bases de datos públicas aunque con diferencias significativas en la magnitud del DB. La magnitud del DB parece estar relacionada con la intensidad de la señal, indicando que a intensidades altas se produce un mayor efecto DB. Por contrapartida, las intensidades bajas incrementan, notablemente, la variabilidad asociada a cada portaobjetos descrito previamente por Wu Z *et al* 2004 [171]

Ello indica que aún existiendo los mismos errores sistemáticos en distintos laboratorios el efecto real sobre los datos puede variar. Es también interesante observar el alto grado de variabilidad detectado entre laboratorios indicando que otras fuentes de variabilidad están actuando. En base a estos resultados se puede concluir que o bien existen los mismos errores sistemáticos con distintos efectos sobre la variable respuesta o que distintos errores sistemáticos afectan a distintos laboratorios. Sea como fuere este hecho refuerza la necesidad de cada laboratorio realice sus propios tests de fiabilidad y adecuar los test estadísticos a aplicar. Ello también implica la necesidad de analizar los datasets por separado mediante la aplicación de métodos basados en meta-análisis [172] y no como si se trataran de un solo dataset.

En este estudio también se ha demostrado que un alto contenido en islas CpG y densidad génica está claramente relacionado con el DB así como un alto contenido en *Alu* y repeticiones L1 están también relacionadas con el DB. Aunque su efecto es menos acusado y la asociación observada puede ser debida a la correlación existente entre este tipo de repeticiones y el contenido G+C [14].

Posibles explicaciones para estas asociaciones pueden ser la estructura de la cromatina, la estructura tridimensional de estas secuencias,...y sus propias características físicoquímicas que les puede hacer responder de distinta manera durante el proceso de marcaje e hibridación. Sin embargo, no se descarta que otros factores puedan estar asociados ya que los mismos clones muestran diferente comportamiento bajo condiciones diferentes y diferentes muestras.

Finalmente podemos concluir que, en función de los resultados obtenidos, para optimizar la técnica, deben elegirse sondas que contengan menos de 4 islas CpG (o entre un 37 % y un 42 % de G+C). La relación entre el G+C y su relación con la eficiencia en el marcaje fue descrita por primera vez por Kallioniemi *et al* [68]. En este estudio se ha demostrado que el mismo efecto se halla en las matrices aCGH y que no queda completamente corregido mediante los procedimientos usuales de normalización. Sharp *et*

al 2007 [173] observaron un efecto negativo del contenido en G+C sobre las sondas de oligonucleótidos por ello se cree que este efecto pudiera llegar incluso a afectar a nuevas técnicas de genotipado a gran escala que se basan en sondas que pueden estar sujetas a las mismas propiedades físicas y químicas.

Locke *et al* 2004 [40] describió que las sondas contenidas en DSs incrementan el efecto DB. Los análisis cualitativo y cuantitativo realizados en este estudio corroboran los resultados obtenidos por Locke *et al* 2004 [40]. En este estudio también se ha hallado una fuerte relación entre el DB y el contenido en CNVs. Ello puede sugerir la presencia de una sobrestimación en el número de CNVs que existe hoy en día en las bases de datos.

Otra fuente de variación detectada en este trabajo es la calidad del ADN que influye directamente sobre la calidad de los resultados obtenidos. Dada esta limitación existe una creciente necesidad de conocer cuales son los estándars para la óptima realización de los experimentos como son la contaminación proteíca y el contenido en sales.

Fuentes de variación asociadas a la imagen

La presencia de formas anómalas en los *spots* se ha descrito como una importante fuente de variación en trabajos previos [174], [175], [176], [177]. En este trabajo se ha demostrado que estas formas pueden ser reconocidas por distintos evaluadores y correctamente clasificadas mediante funciones discrimininantes aunque no se encontró ninguna asociación significativa entre estas formas y la presencia de señales anómalas ni sobre los valores de la variable respuesta M. Sin embargo si que se observó una asociación significativa con la presencia de artefactos dentro del *spot* y con la presencia de formas irregulares.

Por último se estudió la posible relación entre las señales anómalas detectadas y las variables obtenidas mediante el programa GenePix. Las variables que resultaron significativas fueron clasificadas en cinco categorías que fueron utilizadas para puntuar los *spots* en función de su calidad. Dicha puntuación puede es muy valiosa para determinar un orden de preferencia para la posterior validación de CNVs detectadas.

5.3. Métodos para la detección de CNVs

En la literatura se utilizan muy a menudo dos métodos; (i) basado en puntos de corte para la media de las réplicas presentes en un portaobjetos pero con la restricción de que la variabilidad entre éstas no supere un cierto umbral y (ii) el método CBS. Sin embargo, durante el transcurso de esta tesis doctoral, se constató que este método basado en puntos de corte fallaba ante la presencia de deleciones en homozigosis; en estos casos un canal desaparece y el ratio se vuelve más ruidoso llegando a tomar altas variabilidades entre réplicas. CBS ha mostrado ser el mejor método para detectar regiones con un mínimo tres sondas alteradas, debido a que el algoritmo se basa en la detección de segmentos. Y, también, en estudios previos se ha demostrado que el método falla cuando se buscan regiones alteradas en mosaico o que envuelvan menos de tres sondas en cuyo caso parece más adecuado utilizar HMM. También se ha demostrado que el método HMM funciona

mejor cuando se dispone información previa sobre los valores de las sondas [59].

En este trabajo se ha aplicado un método basado en IC que se ha construido a partir de la información aportada por los modelos ANOVA. Con el propósito de estimar los valores del IC se consideraron cuatro distribuciones distintas según género de la muestra hibridada y tipo de hibridación (HD y HR). La presencia de distintas distribuciones según género puede ser explicada por la hibridación cruzada en distintos lugares del genoma causadas por las DSs presentes en el cromosoma X e Y. Mientras que las diferencias presentes en el tipo de hibridación son debidas a la afinidad diferencial que presentan algunos BACs.

En los experimentos reales con matrices aCGH o aExpr no es posible conocer la sensibilidad y especificidad asociado a cada método ya que no es posible conocer con absoluta certeza las alteraciones presentes. Irizarry et al 2005 [178] propusieron un método alternativo a las curbas ROC [157] en aExpr. El método propuesto basa en comparar el grado de concordancia entre los genes más deregulados (tomando diferente número de genes cada vez, 10,50,...) respecto una de las hibridaciones que se toma como referencia. En matrices aCGH este método no es aplicable ya que el número de alteraciones presentes es menor (normalmente entre 15-20). Por ello se ha desarrollado un nuevo tipo de gráfico donde se representa en el eje de ordenadas la sensibilidad como el número de alteraciones conocidas detectadas respecto la eficiencia, resumiendo eficiencia como el número de alteraciones conocidas alteradas respecto el total de alteradas detectadas (cuanto mayor número de falsos positivos existan menor será la eficiencia). Se compararon los métodos PCmin, PCmed, CBS suavizado, CBS y IC mediante estos gráficos. CBS suavizado y CBS obtuvieron los peores resultados, y se hallaron resultados similares para PCmin y PCmed. Posteriormente se aplicaron estos métodos después de realizar una substracción de los valores esperados por cada BAC (métodos combinados). Esta aplicación permitió doblar la eficiencia sin modificar la sensibilidad de los métodos. Además, se hallaron mejores resultados para las hibridaciones HD que para las hibridaciones HR en todos los casos (mayor eficiencia para la misma sensibilidad). Estos resultados se corroboraron mediante curbas ROC obtenidas por simulación de datos.

5.4. Concordancia entre plataformas en la detección de CNVs

Los estudios realizados sobre concordancia entre plataformas han intentando discernir entre la calidad de datos aportados por aCGH de BACs y por oligo aCGH. Ylstra B *et al* 2006 [179] compararon estas dos plataformas y los principales resultados obtenidos indican una mayor fiabilidad y reproducibilidad en aCGH de BACs versus oligo aCGH. Resultados similares se han obtenido en este estudio cuando se ha comparado la matriz aCGH 5,2K con la matriz oligo aCGH de Agilent 44K. Más allá de la reproducibilidad y fiabilidad, el tamaño de las CNVs que han podido ser validadas con Agilent son similares a los tamaños de los BACs debido a la necesidad de considerar regiones cubiertas por tres o más sondas por ello la resolución, a la práctica, es equivalente [179]. Además, aún cuando la resolución es máxima, Agilent 244K, hay regiones del genoma que no quedan cubiertas (i.e 9 BACs dónde no se ha podido demostrar en oligo aCGH de Agilent 244K

la presencia o no de CNV debido a la falta de sondas). Los resultados de Redon *et al* [29] 2006 muestran la misma falta de solapamiento entre dos plataformas mayores; una matriz aCGH 32K de BACs 500K de oligonucleótidos.

Incluso algunos de los problemas detectados en la matrix 5,2K aCGH también se han observado en oligo aCGH. Así, también, se ha observado un efecto DB (es decir algunas sondas tomaban valores de duplicación o deleción en función del marcaje realizado). Este efecto podría estar asociado a la presencia de secuencias con alto contenido G+C [173] ya que el mismo efecto y causa se ha detectado previamente en otro tipo de plataformas como Affymetrix, desarrollando métodos de normalización específicos para evitarlos (GC-RMA). Del mismo modo, para evitar estos efectos en aCGH podría ser ventajoso aplicar determinados pesos a las sondas en función del contenido en G+C durante el proceso de normalización.

Las CNVs encontradas mediante este y otros proyectos son depositadas en bases de datos públicas (i.e. TCAG). Las CNVs que en ellas se encuentran varían en tamaño según el tipo de plataforma utilizada. Teniendo en cuenta las fuentes de variación detectadas en esta tesis doctoral como sus causas y teniendo en cuenta el hecho de que muchos investigadores no realizan réplicas de los experimentos, se cree que sería conveniente proceder a una revisión de estas bases de datos.

Aunque ha habido un incremento constante en la aplicación de la técnica aCGH sigue siendo una incóngita el rango de normalidad en las variaciones en el número de copias por cada individuo sano, así como su efecto real sobre la expresión y como se relacionan entre ellas para producir enfermedad. En la muestra control utilizada (procedente del experimento 32x) se han validado 10 regiones distintas del genoma con alteración pero aún podrían haber muchas más si se validaran los resultados más consistentes (6 regiones más mediante BACs, 14 mediante a44 y 15 mediante a244). El hecho de haber encontrado una CNV no descrita previamente con sólo una muestra control indica el poco conocimiento existente sobre CNVs.

5.5. Expresión en aneusomías parciales

Hughes *et al* 2000 [180] relacionó las aneusomías con cambios en la expresión génica. Sin embargo, estudios recientes más completos han demostrado que las ganancias y pérdidas no siempre influyen sobre la expresión de los genes incluidos en ellas y que no todos los genes responden de igual manera [181]. Es más, en la mayoría de síndromes de microdeleción o microduplicación se ha indentificado uno o pocos genes responsables de las características fenotípicas observadas mientras que el resto de genes pueden actuar como modificadores. Y, en otros casos, la relación puede ser indirecta debido a un efecto a larga distancia causada por elementos reguladores en *cis* o efectos en trans de manera secudaria [182].

Expresión en la región WBS

Tal y como se esperaba existen claras diferencias en la expresión de los genes FZD9, BAZ1B, BCL7B y TBL2 en nw vs sw ya que estos genes se hallan incluidos dentro de la región no delecionada en los individuos nw. También se visualizan diferencias entre grupos debidas a la deleción en nw35 de los genes; WBSCR18, WBSCR22, STX1A y ABDH11. No se hallaron diferencias entre grupos sobre la expresión de los genes MLXIPL y VPS37D que, aparentemente, se hallan infraexpresados en todas las muestras. En la región común delecionada se hallan infraexpresados los genes LIMK1, ETF4H, LAT2, RFC2 y CLIP2. CLIP2 no está delecionado en una de las muestras nw pero aún así está afectada su expresión que puede ser debido a la deleción de alguna de las secuencias reguladoras (presencia de un efecto posicional). Inesperadamente no se hallaron infraexpresados los genes CLDN3, CLDN4, WBSCR27, WBSCR28 y ELN presentes en la región común delecionada aunque hay evidencias, en estudios previos, de que existe un efecto en la expresión de, al menos, ELN en WBS [77]. Se conoce que algunos genes no se expresan en líneas celulares limfoblastoides por ello no se puede excluir que su expresión esté realmente afectada.

Expresión en las regiones flanqueantes

Merla *et al* 2006 [183] propusieron que los genes que flanquean la región WBS ven modificada su expresión. Se estudió esta hipótesis a partir de los resultados de aExpr. Se representó el perfil de la expresión génica en el cromosoma 7 en relación a la región delecionada y, en él, se puede observar como existe una clara infraexpresión de los genes contenidos en la región y, aunque, existe deregulación a ambos lados de la región delecionada, no existen envidencias de que sea mayor que en otros cromosomas sin ningún tipo de alteración (como por ejemplo en el cromosoma 3).

Para verificar estos resultados se comparó la región WBS y las regiones flanqueantes a ambos lados (hasta un máximo de 2 Mb) con dos regiones de características similares, es decir flanqueadas mediante DS, y con una región externa tomada al azar. Los resultados encontrados indican que, aún existiendo una tendencia a la baja en las regiones flanqueantes, éste no es significativamente diferente de la región externa tomada al azar como control.

De los tres genes *upstream* analizados, se halló deregulación en el gen AUTS2 aunque los resultados son variables entre muestras con un genotipo similar. En el artículo de Merla *et al* 2006 [183] estudiaron la expresión mediante qPCR de este gen respecto a un grupo de controles encontrando un pvalor próximo a las diferencias significativas (p=0,06). Ello indica que este gen se halla deregulado aunque, seguramente, no asociado al fenotipo observado ya que en algunas muestras se halla infraexpresado mientras que en otras se halla sobreexpresado.

En la región *downstream* tampoco se hallaron diferencias significativas entre grupos ni en la interacción gen:grupo. Sólo dos genes se hallaron claramente deregulados sobre los 11 estudiados; el gen TMEM120A y el gen HSP1 que no fueron analizados en el arículo

de Merla *et al* 2006 [183]. Ello parece indicar que la expresión de los genes flanqueantes, en general, no contribuyen significativamente al fenotipo de los individuos WBS.

En Merla *et al* 2006 [183] tomaron regiones flanqueantes de mayor tamaño; 8,4 Mb *upstream* y 11,9 Mb *downstream*. Por ello se decidió comparar gen a gen los resultados obtenidos entre ambos estudios. La concordancia de resultados entre el estudio presentado en esta tesis doctoral y el estudio de Merla *et al* 2006 [183] es del 91 % (es decir una correspondencia de resultados de 31 genes sobre 34). En 11 de estos 31 casos se hallaron diferencias entre los grupos nw vs sw y en WBS vs controles; ASL, BAZ1B, BCL7B, BTG3, EIFH4, GBAS, GTF2I, LIMK1, TBL2, UFD1L y WBSCR22. En cuatro genes, sin diferencias significativas entre nw vs sw, se halló una infraexpresión o sobreexpresión en nuestro estudio mientras que existían diferencias entre WBS vs controles en el estudio de Merla *et al* 2006 [183]; CLIP2, LAT2, RFC2 y STX1A. Para los 16 genes restantes se obtuvieron resultados negativos en ambos estudios. En Merla *et al* 2006 [183] se hallaron diferencias significativas WBS vs controles en tres genes que no pudieron ser reproducidas en el presente estudio; HIP1, KCTD7 y MDH2.

Detección de vías afectadas

El gen MLXIPL, infraexpresado en todas las muestras estudiadas (sw + nw), se ha vinculado, recientemente, con el metabolimo de los triglicéridos. Se ha demostrado que ciertas variantes del gen MLXIPL incrementan el riesgo a padecer una enfermedad cardíaca **??**. Una de las vías posiblemente alterada, según el set de datos N=92 común a sw y nw, es el metabolimo lípidico pudiendo ser MLXIPL el gen relacionado. Otra vía alterada es el metabolimo glucídico (gluconeogénesis y glicolisis) que podría explicar el elevado riesgo de padecer diabetes de los pacientes con WBS. MLXIPL se ha vinculado anteriormente en pacientes con WBS con el riesgo a padecer diabetes.

El gen BCL7B, situado en la región WBS, está diferencialmente delecionado y expresado en las muestras sw vs nw. Este gen puede ser responsable de las alteraciones significativas en el transporte y contracción celular halladas en REACTOME.

Otros genes detectados, como diferencialmente expresados entre sw vs nw, son UTS2 y PAFAH1B3 que pueden ser buenos candidatos para WBS; (i) UTS2 es un vasoconstrictor que se expresa exclusivamente en cerebro y también se ha relacionado con riesgo coronario así como con diabetes ([184, 185] y (ii) PAFAH1B3 se expresa en cerebro e interviene en el desarrollo del sistema nervioso ([186]). También, de manera no tan evidente, podría estar alterado el ciclo de NOTCH con JAG1 como principal componente diferencialmente expresado en esta vía entre grupos.

5.6. Aplicación de los modelos ANOVA en Genética molecular

Algunas consideraciones sobre el tamaño muestral

Una de las cuestiones más importantes en el estudio de datos procedentes de matrices, ya sea para estudios de calidad o para realizar comparaciones entre grupos (i.e. casos vs controles), es conocer el número de réplicas técnicas y/o biológicas necesarias para obtener un cierto poder estadístico. El número de réplicas necesarias para realizar un buen diseño experimental en matrices aCGH y aExpr se ha discutido en muchos trabajos [187, 188, 189]. La mayoría de estudios apuntan a un mínimo de ocho réplicas, ello garantiza un buen poder estadístico aunque este número es muchas veces difícil de conseguir; (i) incrementa el coste de los experimentos y (ii) no existe suficiente material biológico. En el experimento 32x se han realizado cuatro réplicas técnicas en las mismas condiciones que aseguran 16 réplicas para cada efecto principal consiguiendo, así, un buen tamaño muestral para la identificación de los efectos principales. Cuando se desea comparar grupos, la evidencia indica la necesidad de trabajar con un mínimo de cinco réplicas biológicas por grupo [170]. Aunque en síndromes con una prevalencia baja no es posible conseguir suficientes muestras. Una alternativa es la realización de réplicas técnicas intercambiando los fluorocromos que puede ayudar a reducir el número de falsos positivos tan a menudo relacionados con este tipo de experimentos. En ciertos casos el motivo no es la disponibilidad de la muestra sino no encarecer extremadamante el coste de estos experimentos. Y, a veces, ello conlleva a la realización de *pools* de muestras. Un argumento a favor de este tipo de soluciones es que analizar cinco *pools* de tres muestras es más representativo que la realización de cinco hibridaciones procedentes de cinco individuos pero el problema reside en que la mayoría de los investigadores realizan un sólo *pool* por grupo no siendo este procedimiento correcto desde un punto de vista estadístico.

El diseño experimental

Evitar la confusión por factores externos es primordial debido a los pequeños tamaños muestrales con los que se suele trabajar. Es, evidentemente, conveniente poder realizar todos los experimentos el mismo día por el mismo técnico y por el mismo *batch* de reactivo [170]. Aunque en este estudio se ha demostrado que, al menos para matrices aCGH basadas en BACs, no existe un evidente sesgo acontecido por estos factores ya que se consigue una reproducibilidad muy alta aún cuando las hibridaciones se hayan realizado meses más tarde. Es un buen resultado ya que la realización de experimentos por el mismo técnico, día y *batch* suele ser inasumible a la práctica. Por otro lado el diseño experimental permite obtener estándars de perfiles de hibridación que puedan irse modificando mediante la introducción de nuevos datos (por métodos bayesianos) [170].

Los modelos ANOVA

Los modelos ANOVA son estructuras altamente flexibles, diseñables por cada usuario y en cada condición como aquí se ha expuest. Los modelos ANOVA permiten la incorporación de otras estructuras de error en el análisis de matrices aCGH y aExpr.

Los ANOVAs clásicos incorporan toda la información en un solo modelo. Este tipo de estructura en matrices aCGH y aExpr es totalmente intractable debido al gran número de genes que se analizan simúltaneamente (varios miles). Kerr *et al* 2003 [188] desarrollo diversos modelos basados en la tecnología ANOVA y propuso trabajar con un modelo por cada gen considerando que era equivalente. Este mismo resultado se ha hallado en este trabajo mediante el estudio de los genes de la región WBS en un modelo ANOVA clásico versus un modelo ANOVA por cada gen.

Principales inconvenientes o limitaciones

Las principales limitacionse están relacionadas con el residuo de los modelos. Si este residuo es pequeño, como en el caso de la comparación entre los grupos sw y nw en el ANOVA para aCGH, se tiende a hallar un elevado número de diferencias significativas que rara vez se correlaciona con una diferencia biológica entre grupos (magnitud del efecto grupo) por ello se recomienda estimar la magnitud de la diferencia encontrada (efecto grupo) y basar los resultados en el pvalor y en la magnitud.

5.7. Identificación de PSVs funcionales

Tras el descrubrimiento de las DSs [3] se detectó que gran parte de los SNPs depositados en las bases de datos eran, en realidad, pequeños cambios nucleotídicos entre secuencias parálogas (PSVs) y no entre variantes alélicas del mismo gen (SNPs) [190].

Hay dos mecanismos principales mediante los cuales se duplican regiones codificantes; retrotransposición y NAHR. Pero, la retrotransposición raramente ocasiona copias funcionales de mRNA [191]. En cambio NAHR, proporciona copias intactas. Estas copias permiten la aceptación de un alto nivel de modificación (mediante PSVs y otros mecanismos) pudiendi llegar, incluso, a desarrollar nuevas funciones sin que las funciones básicas se vean comprometidas. Por ello, su estudio, puede ser clave para comprender algunos mecanismos evolutivos o de enfermedad aunque se espera que la mayoría de genes parálogos acumulen mutaciones a un ritmo neutral y que estos acaben degenerados como pseudogenes no procesados.

Las substituciones nucleotídicas en genes (o mutaciones) se ha clasifican dentro de dos categorías; cambios sinónimos si no producen un cambio aminoacídico y cambios no-sinónimos si los producen. Las substituciones sinónimas ocurren con mayor frecuencia (Ks) que las no sinónimas (Ka) de tal manera que el ratio Ka/Ks es una medida de la distancia evolutiva; cuanto más grande sea el ratio mayor distancia evolutiva entre genes o mayor presión evolutiva hay entre ellos.[192]. Estos cambios entre copias parálogas pueden producir conversión génica no alélica [193]. Cuando está conversión se produce un cambio no sinónimo entre copia y gen o viceversa puede producir enfermedad [124].

La identificación del número relativo entre estos pseudogenes o copias versus los genes originales permitiría conocer qué pseudogenes están bajo presión evolutiva así como

realizar estudios de asociación casos-controles para detectar nuevas enfermedades génicas causadas por este proceso. Pero para ello es necesario conocer su posición en el genoma así como los cambios esperados entre ellos. Así, pues, en este trabajo se ha desarrollado un algoritmo que permite detectar la posición y tipo de cambio en PSVs. En esta primera aproximación, el trabajo se ha centrado en detectar PSVs que produzacan cambios no sinónimos o bien que sean silenciadoras de *splicing*. Es conocido que alteraciones en otras regiones altamente conservadas pueden producir enfermedad (i.e. la conversión génica entre SMN1 y SMN2 se produce en un cambio sinónimo pero afecta a una secuencia silenciadora de splicing [130], se ha demostrado que cambios en la región de splicing puede producir enfermedad aún conservando los dinucleótidos GC y AG [194].)

Se estima que el genoma humano no contiene más de 2.000 pseudogenes no procesados [**?**]. En este estudio se han identificado 1.780 genes con PSVs funcionales (entendiendo como funcionales aquellas PSVs que modifican la pauta de lectura, producen un cambio aminoacídico, aparición de un sitio de STOP o silencian un splicing mediante la modificación de los dinucleótidos GC y AG) aunque no todos ellos son pseudogenes y, además, algunas de estas entradas están duplicadas (es decir genes con copias funcionales ambas están registradas en la base de datos *Biomart* [161]; i.e NCF1 y NCF1C).

En un segundo paso se han eliminado aquellas entradas gen-copia no situadas en DSs. Después de aplicar este filtro y eliminar las entradas dobles se han hallado 1.115 genes con PSVs funcionales. Los cambios más frecuentes entre éstas se dan en un solo nucleótido, y, dentro de estos los más frecuentes son los cambios entre purinas y pirimidinas (26 % y 35 % respectivamente). Estudios clásicos sobre conversión génica muestran como se favorece a unos alelos respecto a otros (GC sobre AT) [193]. En nuestro estudio se ha detectado el doble de cambios en la dirección GC que en AT. Ello es concordante con la hipótesis de que el proceso de enriquicimiento en GC está presente en secuencias bajo conversión génica [193]. En este estudio también se han identificado inserciones y/o deleciones de uno o varios pares de bases (el límite se sitúa en 30 pb) que suponen 2.034 (16 % del total de PSVs identificadas).

En esta tesis doctoral se ha descrito un método altamente específico que permite la detección de PSVs funcionales. El método descrito en este apartado ha permitido indentificar PSVs funcionales anteriormente descritas que cumplían los criterios especificados con un 100 % de sensibilidad.

6. Conclusiones

A continuación se resumen las principales aportaciones de esta tesis doctoral.

- Se ha demostrado que los modelos basados en ANOVAs son herramientas muy potentes y versátiles que permiten caracterizar fuentes de variación experimental y detectar CNVs con un alto grado de sensibilidad. Además, las estimaciones de los efectos significativos proporcionan perfiles para cada tipo de hibridación que pueden ser utilizados como medida de calidad o bien conjugarlos con métodos estándar en la detección de CNVs e incrementar así la sensibilidad y la especificidad.

- Se ha realizado un diseño experimental que ha permitido detectar como principales fuentes de variación el efecto *Dye-Bias* DB (observable cuando se giran los flurocromos entre HD y HR), el portaobjetos (sobre el cual se realiza la hibridación) y la calidad del ADN utilizado (una calidad mayor reduce el número de falsos positivos y falsos negativos).

- Se ha demostrado que el efecto DB es sistemático y que está asociado a efectos espaciales presentes en el portaobjetos y a las características propias de las secuencias de estos BACs, especialmente al contenido en G+C.

- No se ha encontrado asociación de las formas aureola y donut con la presencia de datos atípicos. Aunque sí se ha hallado asociación con formas irregulares y artefactos. La detección automática de los datos atípicos mediante el estudio de las variables incluidas en GenePix no es exacta aunque se pueden desarrollar reglas que permiten obtener información sobre la calidad de los datos.

- En el estudio sobre la expresión génica de la región WBS no se ha encontrado evidencia estadística de que las regiones flanqueantes tengan deregulada su expresión. Pero se ha hallado una clara relación entre la deleción de la región y la expresión de los genes contenidos en ella. En este mismo estudio se han detectado algunas vías afectadas que pueden estar relacionadas con el fenotipo de WBS como la diabetes, el riesgo coronario y con problemas en el desarrollo cerebral.

- Se ha desarrollado un algoritmo bioinformático que permite la detección de PSVs funcionales a lo largo del genoma humano. El algoritmo ha permitido detectar PSVs conocidas relacionadas con enfermedad así como otras no-relacionadas con ninguna enfermedad pero que son candidatas a ello.

7. Abreviaturas

A: *Spot* tipo aureola

aCGH: matriz basada en Hibridación Genòmica Comparada (*Comparative Genome Hybridization*)

ADN: Ácido desoxiribonucleico

aExpr: matriz de expresión

aExprAg44: matriz de expresión Agilent 44K para *Homo sapiens*

ANOVA: Analisis de la Varianza

AS: Síndrome de Angelman

ASD: *Autism spectrum disorder*. Transtorno del espectro autista o TEA.

ARN: Ácido ribonucleico

BAC: *Bacterial Artificial Chromosome* o Cromosoma Artificial Bacteriano.

BG: *Background* o ruido de fondo.

BH: Corrección para multi-test Benjamini-Hochberg.

CBS: *Circulary Binary Segmentation* Algoritmo que permite detectar CNDs.

Chr: cromosoma

CHORI: *Children's Hospital Oakland Research Institute*

CI: Coeficiente Intelectual

CND: *Copy number difference* Diferencia en el número de copias. Término usualmente aplicado cuando se tiene desconocimiento sobre su patogenicidad.

CNV: *Copy number variation* o Variación en el número de copias. Término usualmente aplicado a polimorfismos

Conc: Concentración

CRG: Centro de Regulación Genómica

D: *Spot* tipo donut

DB: *Dye Bias*: Es la tendencia a captar un flurocromo sobre otro. Este efecto se observa al realizar dos hibridaciones distintas sobre la misma muestra test obteniendo valores distintos según el flurocromo con que fue marcada.

DiG: Síndrome diGeorge

DS: Duplicación Segmentaria, duplicones o LCR.

DWS: *down stream*

EXT: Región externa

F: *Spot* artefactual

FG: Señal del *spot*

FG/BG ratio: señal versus ruido de fondo (calculado como la media de FG respecto la mediana de BG)

FDR: *False Discovery Rate*

FISH: *Fluorescent in situ hybridization* Hibridación in situ con fluorescencia.

GEO: *Gene Expression Omnibus*

HD: Hibridación directa en los experimentos de dos colores dónde la muestra test se marca con el fluorocromo Cy5 y la muestra de referencia con el fluorocromo Cy3.

HMM: *Hiden Markov Model*. Algoritmo que permite detectar CNDs.

HR: Hibridación reversa en los experimentos de dos colores dónde la muestra test se marca con el fluorocromo Cy3 y la muestra de referencia con el fluorocromo Cy5.

IC: Intervalo de Confianza

IQR: *Inter Quantile Range*, Rango intercuartílco

LCR: *Low Copy Repeat*, Repeticiones en bajo número de copias también llamadas DS o duplicones

LINE:

KS: Síndrome de Kabuki

M: el logaritmo en base dos del ratio entre la muestra test y referencia normalizado.

MAD: *Median Absolute Deviance*

MLPA: *Multiplex Ligation-dependent Probe Amplification*

MSV: *Multi Site Variation*

M; log en base 2 del ratio entre los dos canales normalizado.

NAHR: Recombinación homóloga no alélica

oligo aCGH: matriz aCGH basada en oligonuecleótidos

P: Efecto portaobjetos en los modelos ANOVA.

PCmed: Método basado en puntos de corte para el valor medio de las réplicas de un portaobjetos

PCmin: Método basado en puntos de corte para el valor mínimo de las réplicas de un portaobjetos

PSV: *Paralogue Sequence Variant*

pb, kb, Mb; pares de base, kilobases y Megabases

qPCR: *quantitative Polymerase Chain Reaction*

ROC: *Receive Operator Curbe*

RIN: *RNA Integrity Number*

SD: desviación típica.

SNP: *Single Site Polimorphism*

STR: *Sequence Tandem Repeat*

Sol: Solución de resuspensión

UPS: *Up stream*

WBS: William-Beuren Syndrome

Wt: *spot* normal

8. Bibliografía

Referencias

[1] Hartl DL. Molecular melodies in high and low C. *Nat Rev Genet* 2000;**1**(2):145-149.

[2] Bailey JA, Gu Z, Clark RA, Reinert RV, Samonte S, Shwartz MD *et al.* Recent Segmental Duplications in the Human Genome. *Science* 2002;**297**(5583):1003-1007.

[3] Cheung J, Wilson M, Zhang J, Khaja R, MacDonald JR, Heng HHQ *et al.* Recent segmental and gene duplications in the mouse genome. *Genome Biology* 2003;**4**(8):R47.

[4] Zhang L, Lu HHS, Chung W, Yang J, Li W. Patterns of Segmental Duplication in the Human Genome. *Mol Biol Evol* 2005;**22**(1):135-141.

[5] Wolfe KH, Shields DC. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* 1997; **387**(6634):708-713.

[6] Kellis M, Birren BW, Lander ES. Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. *Nature* 2004; **428**(6983):617-624.

[7] Hughes AL, Friedman R, Ekollu V, Rose JR. Non-random association of transposable elements with duplicated genomic blocks in Arabidopsis thaliana. *Mol Phylogenet Evol* 2003; **29**(3):410-416.

[8] Aparicio S, Chapman J, Stupka E, Putnam N, Chia JM, Dehal P *, et al.*. Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes. *Science* 2002; **297**(5585):1301-1310.

[9] Hillier LW, Miller W, Birney E, Warren W, Hardison RC, Pointing CP *et al.* Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* 2004; **432**(7018):695-716.

[10] Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J*, et al.* Initial sequencing and analysis of the human genome. *Nature* 2001; **409**(6822):860-921.

[11] Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P *et al.*. Initial sequencing and comparative analysis of the mouse genome. *Nature* 2002; **420**(6915):520-562.

[12] Gibbs RA, Weinstock GM, Metzker ML, Muzny DM, Sodergren EJ, Scherer S *et al.* Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* 2004; **428**(6982):493-521.

[13] Feuk L, Carson AR, Scherer SW. Structural variation in the human genome. *Nat Rev Genet* 2006;**7** (2):85-97.

[14] Medstrand P, van de Lagemaat LN, Mager DL. Retroelement distributions in the human genome: variations associated with age and proximity to genes. *Genome Res* 2002; **12**:1483-1495.

[15] Conrad B, Antonarakis SE. Gene duplication: a drive for phenotypic diversity and cause of human disease. *Annu Rev Genomics Hum Genet* 2007;**8**:17-35.

[16] Bailey JA, Eichler EE. Genome-wide detection and analysis of recent segmental duplications within mammalian organisms.*Cold Spring Harb Symp Quant Biol* 2003; **68**:115-124.

[17] Zhou Y, Mishra B. Quantifying the mechanisms for segmental duplications in mammalian genomes by statistical analysis and modeling. *Proc Natl Acad Sci USA* 2005; **102**(11):4051-4056.

[18] Susumu O. Evolution bye Gene Duplication. New York: Springer, 1970.

[19] Lupski JR. Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits. *Trends in Genetics* 1998;**14**(10):417-422.

[20] Stankiewicz P, Lupski JR. Genome architecture, rearrangements and genomic disorders. *Trends in Genetics* 2002;**18**(2):74-82.

[21] Osborne et al. A 1.5 million-base pair inversion polymorphism in families with Williams-Beuren syndrome. *Nat Genet* 2001;**9**:321-325

[22] Gimelli Genomic inversions of human chromosome 15q11-q13 in mothers of Angelman syndrome patients with class II (BP2/3) deletions. *Human Molec Genet* 2003;**12**:849-858.

[23] Lakich D, Kazazian H H Jr, Antonarakis, SE, Gitschier J. Inversions disrupting the factor VIII gene are a common cause of severe haemophilia A. *Nat Genet* 1993;**5**:236-241.

[24] Stefansson H, Helgason A, Thorleifsson G, Steinthorsdottir V, Masson G, Barnard J *et al.* A common inversion under selection in Europeans. *Nature Genet* 2005;**37**:129-137.

[25] Tuzun E, Sharp AJ, Bailey JA, Kaul R, Morrison VA, Pertz LM *et al.* Fine-scale structural variation of the human genome. *Nat Genet* 2005;**37**(7):727-732.

[26] Sharp AJ, Locke DP, McGrath SD, *, et al.* Segmental Duplications and Copy-Number Variation in the Human Genome. *Am J Hum Genet* 2005;**77**(1):78-88.

[27] Sebat J, Lakshmi B, Troge J, Alexander J, Young P, Lundin P *et al.* Large-Scale Copy Number Polymorphism in the Human Genome. *Science* 2004;**305**(5683):525-528.

[28] Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y *et al.* Detection of large-scale variation in the human genome. *Nat Genet* 2004;**36**(9):949-951.

[29] Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD , et al.Global variation in copy number in the human genome. *Nature* 2006; **444**(7118):444-454.

[30] Wong KK, Deleeuw RJ, Dosanjh NS, Kimm LR, Cheng Z, Horsman DE , et al.A comprehensive analysis of common copy-number variations in the human genome. *Am J Hum Genet* 2007; **80**:91-104.

[31] Gallagher CJ, Kadlubar FF, Muscat JE, Ambrosone CB, Lang NP, Lazarus P. The UGT2B17 gene deletion polymorphism and risk of prostate cancer. A case-control study in Caucasians. *Cancer Detect Prev* 2007; **31**(4):310-315.

[32] Park JY, Tanner JP, Sellers TA, Huang Y, Stevens CK, Dossett N, Shankar RA, Zachariah B, Heysek R, Pow-Sang J. Association between polymorphisms in HSD3B1 and UGT2B17 and prostate cancer risk. *Urology* 2007; **70**(2):374-379.

[33] Karypidis AH, Olsson M, Andersson SO, Rane A, Ekstrom L. Deletion polymorphism of the UGT2B17 gene is associated with increased risk for prostate cancer and correlated to gene expression in the prostate. *Pharmacogenomics J* 2007. In press

[34] Park J, Chen L, Ratnashinge L, Sellers TA, Tanner JP, Lee JH *et al.* Deletion polymorphism of UDP-glucuronosyltransferase 2B17 and risk of prostate cancer in African American and Caucasian men. *Cancer Epidemiol Biomarkers Prev* 2006; **15**(8):1473-1478.

[35] Aitman TJ, Dong R, Vyse TJ, Norsworthy PJ, Johnson MD, Smith J *et al.* Copy number polymorphism in Fcgr3 predisposes to glomerulonephritis in rats and humans. *Nature* 2006; **439**(7078):851-855.

[36] Lachman HM, Pedrosa E, Petruolo OA, Cockerham M, Papolos A, Novak T *et al.* Increase in GSK3beta gene copy number variation in bipolar disorder. *Am J Med Genet B Neuropsychiatr Genet* 2007; **144**(3):259-265.

[37] Rovelet-Lecrux A, Hannequin D, Raux G, Le Meur N, Laquerriere A, Vital A *et al.* APP locus duplication causes autosomal dominant early-onset Alzheimer disease with cerebral amyloid angiopathy. *Nat Genet* 2006; **38**(1):24-26.

[38] Beckmann JS, Estivill X, Antonarakis SE. Copy number variants and genetic traits: closer to the resolution of phenotypic to genotypic variability. *Nat Rev Genet* 2007; **8**(8):639-646.

[39] Fredman D, White SJ, Potter S, Eichler EE, Den Dunnen JT, Brookes AJ. Complex SNP-related sequence variation in segmental genome duplications. *Nat Genet* 2004; **36**(8):861-866.

[40] Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, Thorne N, Redon R, Bird CP, de Grassi A, Lee C, Tyler-Smith C, Carter N, Scherer SW, Tavare S, Deloukas P, Hurles ME, Dermitzakis ET. Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* 2007; **315**(5813):848-853.

[41] Solinas-Toldo S, Lampel S, Stilgenbauer S, Nickolenko J, Benner A, Dohner H et al. Matrix-based comparative genomic hybridization: biochips to screen for genomic imbalances. Genes Chromosomes. *Cancer* 1997;**20**(4):399-407.

[42] Pollack JR, Perou CM, Alizadeh AA, Eisen MB, Pergamenschikov A, Williams CF et al. Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nat Genet* 1999; **23**(1):41-46.

[43] Carter NP, Fiegler H, Piper J. Comparative analysis of comparative genomic hybridization microarray technologies: report of a workshop sponsored by the Wellcome Trust. *Cytometry* 2002;**49**(2):43-48.

[44] Drazinic CM, Ercan-Sencicek AG, Gault LM, Hisama FM, Qumsiyeh MB, Nowak NJ et al. Rapid array-based genomic characterization of a subtle structural abnormality: a patient with psychosis and der(18)t(5;18)(p14.1;p11.23). *Am J Med Genet A* 2005; **134**(3):282-289.

[45] Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J et al.: Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res* 2002; **30**:e15.

[46] Smyth GK, Speed T: Normalization of cDNA microarray data. *Methods* 2003; **31**:265-273.

[47] Kerr MK, Churchill GA: Experimental design for gene expression microarrays. *Biostatistics* 2001; **2**: 183-201.

[48] Scharpf RB, Iacobuzio-Donahue CA, Sneddon JB, Parmigiani G: When should one subtract background fluorescence in two color microarrays? *Biostatistics* 2007; **8**(4):695-707.

[49] Hsu L, Self SG, Grove D, Randolph T, Wang K, Delrow JJ et al. Denoising array-based comparative genomic hybridization data using wavelets. *Biostatistics* 2005;**6**(2):211-226.

[50] Jong K, Marchiori E, Meijer G, Vaart AV, Ylstra B. Breakpoint identification and smoothing of array comparative genomic hybridization data. *Bioinformatics* 2004;**20**(18):3636-3637.

[51] Hupe P, Stransky N, Thiery JP, Radvanyi F, Barillot E. Analysis of array CGH data: from signal ratio to gain and loss of DNA regions. *Bioinformatics* 2004;**20**(18):3413-3422.

[52] Olshen AB, Venkatraman ES, Lucito R, Wigler M. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* 2004;**5**(4):557-572.

[53] Picard F, Robin S, Lavielle M, Vaisse C, Daudin JJ. A statistical approach for array CGH data analysis. *BMC Bioinformatics* 2005;**6**(1):27.

[54] Fridlyand J, Snijders AM, Pinkel D, Albertson DG, and Jain A. Hidden Markov models approach to the analysis of array CGH data.*J Multivariate Anal* 2004 **90**; 132-153.

[55] Chen QR, Bilke S, Khan J. High-resolution cDNA microarray-based comparative genomic hybridization analysis in neuroblastoma. *Cancer Lett* 2005; **228**(1-2):71-81.

[56] Price TS, Regan R, Mott R, Hedman A, Honey B, Daniels RJ et al. SW-ARRAY: a dynamic programming solution for the identification of copy-number changes in genomic DNA using array comparative genome hybridization data. *Nucleic Acids Res* 2005; **33**(11):3455-3464.

[57] Daruwala RS, Rudra A, Ostrer H, Lucito R, Wigler M, Mishra B. A versatile statistical analysis algorithm to detect genome copy number variation. *Proc Natl Acad Sci USA* 2004; **101**(46):16292-16297.

[58] Lai WR, Johnson MD, Kucherlapati, Park J. Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. *Bioinformatics* 2005;**21**(19):3763-3770.

[59] Willenbrock H, Fridlyand J. A comparison study: applying segmentation to array CGH data for downstream analyses. *Bioinformatics* 2005; **21**(22):4084-4091.

[60] Veltman JA, Schoenmakers EF, Eussen BH*, et al.* High-throughput analysis of subtelomeric chromosome rearrangements by use of array-based comparative genomic hybridization. *Am J Hum Genet* 2002; **70**(5):1269-1276.

[61] Vissers LE, de Vries BB, Osoegawa K*, et al.* Array-based comparative genomic hybridization for the genomewide detection of submicroscopic chromosomal abnormalities. *Am J Hum Genet* 2003;**73**(6):1261-1270.

[62] Bengtsson H, Irizarry R, Carvalho B, Speed TP. Estimation and assessment of raw copy numbers at the single locus level. *Bioinformatics* 2008; in press.

[63] Fan JB, Gunderson KL, Bibikova M, Yeakley JM, Chen J, Wickham Garcia E *et al.* Illumina universal bead arrays. *Methods Enzymol* 2006;**410**:57-73.

[64] Steemers FJ, Gunderson KL. Whole genome genotyping technologies on the BeadArray platform. *Biotechnol J* 2007;**2**(1):41-49.

[65] Rooms L, Reyniers E, Kooy RF. Subtelomeric rearrangements in the mentally retarded: a comparison of detection methods. *Hum Mutat* 2005;**25**(6):513-524.

[66] Knight SJ, Horsley SW, Regan R, Lawrie NM, Maher EJ, Cardy DLN, Flint J, Kearney L. Development and clinical application of an innovate fluorescence in situ hybridization technique which detects submicroscopic rearrangements involving telomeres. *Eur J Human Genet* 1997; **5**:1-8.

[67] Knight SJ, Regan R, Nicod A*, et al.* Subtle chromosomal rearrangements in children with unexplained mental retardation.*Lancet* 1999; **354**(9191):1676-1681.

[68] Kallionemi A, Kallionemi OP, Sudar D, Rutovitz D, Grey JW, Waldman F, Pinkel D. Comparative Genomic Hybridization for molecular cytogenetic analysis of solid tumors. *Science* 1992; **258**:818-821.

[69] Levy B, Dunn TM, Kaffe S, Kardon N, Hirschorn K. Clinical applications of comparative genomic hybridization. *Genet Med* 1998;**1**:4-12.

[70] Shouten JP, McElgunn CJ, Waaijer R, Zwijnenburg D, Diepvens F, Pals G.2002. Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification. *Nucleic Acids Res* 2002;**30**:e57.

[71] Koolen DA, Vissers L, Pfundt R, de Leeuw N, Knight SJL, Regan R, *et al* A new chromosome 17q21.31 microdeletion syndrome associated with a common inversion polymorphism. *Nature Genet* 2006. **38**: 999-1001

[72] Flint J, Wilkie AO, Buckle VJ*, et al.* The detection of subtelomeric chromosomal rearrangements in idiopathic mental retardation. *Nat Genet* 1995;**9**(2):132-140.

[73] Williams CA, Beaudet AL, Clayton-Smith J, Knoll JH, Kyllerman M, Laan LA, *et al.* Angelman syndrome 2005: updated consensus for diagnostic criteria. *Am J Med Genet A* 2006; **140**(5):413-418.

[74] Wattendorf DJ, Muenke M. Prader-Willi syndrome. *Am Fam Physician* 2005;**72**(5):827-830.

[75] Gropman AL, Elsea S, Duncan WC Jr, Smith AC. New developments in Smith-Magenis syndrome (del 17p11.2). *Curr Opin Neurol* 2007;**20**(2):125-134.

[76] Pavlicek A, House R, Gentles AJ, Jurka J, Morrow BE. Traffic of genetic information between segmental duplications flanking the typical 22q11.2 deletion in velo-cardio-facial syndrome/DiGeorge syndrome. *Genome Res* 2005;**15**(11):1487-1495.

[77] Antonell A, Del Campo M, Flores R, Campuzano V, Perez-Jurado LA. Williams syndrome: its clinical aspects and molecular bases. *Rev Neurol* 2006; **42**(Suppl 1):S69-75.

[78] Somerville MJ, Mervis CB, Young EJ, Seo EJ, del Campo M, *et al.* Severe expressive-language delay related to duplication of the Williams-Beuren locus. *N Engl J Med* 2005;**353**:1694-1701.

[79] Shaffer LG, Lupski JR. Molecular mechanisms for constitutional chromosomal rearrangements in humans. *Annu Rev Genet* 2000; **34**: 297-329, .

[80] Shapira SK, McCaskill C, Northrup H, Spikes AS, Elder F F B, Sutton VR, *et al.* Chromosome 1p36 deletions: the clinical phenotype and molecular characterization of a common newly delineated syndrome. *Am J Hum Genet* 1997. **61**: 642-650.

[81] Christiansen J, Dyck JD, Elyas BG, Lilley M, Bamforth JS, Hicks M , *et al.* Chromosome 1q21.1 contiguous gene deletion is associated with congenital heart disease. *Circ Res.* 2004; **94**(11):1429-35.

[82] Ala-Mello S, Koskimies O, Rapola J, Kääriäinen H. Nephronophthisis in Finland: epidemiology and comparison of genetically classified subgroups. *Eur J Hum Genet* 1999; **7**(2):205-211.

[83] Willatt L, Cox J, Barber J, Cabanas ED, Collins A, Donnai D, *et al.* 3q29 microdeletion syndrome: clinical and molecular characterization of a new syndrome. *Am J Hum Genet* 2005;**77**(1):154-160.

[84] Tatton-Brown K, Douglas J, Coleman K, Baujat G, Chandler K, Clarke A, *et al.* Multiple mechanisms are implicated in the generation of 5q35 microdeletions in Sotos syndrome. *J Med Genet* 2005;**42**(4):307-313.

[85] Pehlivan T, Pober BR, Brueckner M, Garrett S, Slaugh R, Van Rheeden R, Wilson DB, Watson MS, Hing AV. GATA4 haploinsufficiency in patients with interstitial deletion of chromosome region 8p23.1 and congenital heart disease. *Am J Med Genet* 1999;**83**(3):201-206.

[86] Bonati MT, Finelli P, Giardino D, Gottardi G, Roberts W, Larizza L. Trisomy 15q25.2-qter in an autistic child: genotype-phenotype correlations. *Am J Med Genet* 2005; **133A**: 184-188.

[87] Shao Y, Cuccaro ML, Hauser ER, Raiford KL, Menold MM, Wolpert CM,*et al.* Fine mapping of autistic disorder to chromosome 15q11-q13 by use of phenotypic subtypes. *Am J Hum Genet* 2003; **72**: 539-548, .

[88] Sharp AJ, Mefford HC, Li K, Baker C, Skinner C, Stevenson RE, *et al.* A recurrent 15q13.3 microdeletion syndrome associated with mental retardation and seizures. *Nat Genet.* 2008; **40**(3):322-8.

[89] Klopocki E, Graul-Neumann LM, Grieben U, Tönnies H, Ropers HH, Horn D, Mundlos S, Ullmann R. A further case of the recurrent 15q24 microdeletion syndrome, detected by array CGH. *Eur J Pediatr* 2007

[90] Potocki L, Chen KS, Park SS, Osterholm DE, Withers MA, Kimonis V, Summers AM, Meschino WS, Anyane-Yeboa K, Kashork CD *et al.* Molecular mechanism for duplication 17p11.2-the homologous recombination reciprocal of the Smith-Magenis microdeletion. *Nat Genet* 2000;**24**(1):84-87.

[91] Patel K, Frost G, Rossell R, Pizer B, Gee A, Sugimoto T, Phimister E, Kemshead J. Expression of the neural cell adhesion molecule (NCAM) on the haemopoietic cell line Nalm-16. *Leuk Res* 1992; **16**(3):307-315.

[92] Blair IP, Nash J, Gordon MJ, Nicholson GA. Prevalence and origin of de novo duplications in Charcot-Marie-Tooth disease type 1A: first report of a de novo duplication with a maternal origin. *Am J Hum Genet* 1996; **58**(3):472-476.

[93] Sessa A, Ghiggeri GM, Turco AE. Autosomal dominant polycystic kidney disease: clinical and genetic aspects. *J Nephrol* 1997; **10**(6):295-310.

[94] Yobb TM, Somerville MJ, Willatt L, Firth H V, Harrison K, MacKenzie J,*et al.* Microduplication and triplication of 22q11.2: a highly variable syndrome. *Am J Hum Genet* 2005; **76**: 865-876, .

[95] Mears AJ, el-Shanti H, Murray JC, McDermid HE, Patil SR. Minute supernumerary ring chromosome 22 associated with cat eye syndrome: further delineation of the critical region. *Am J Hum Genet* 1995; **57**(3):667-73.

[96] Shapiro LJ, Yen P, Pomerantz D, Martin E, Rolewic L, Mohandas T. Molecular studies of deletions at the human steroid sulfatase locus. *Proc Natl Acad Sci USA* 1989; **86**(21):8477-8481.

[97] Chelly J, Mandel JL. Monogenic causes of X-linked mental retardation. *Nat Rev Genet* 2001;**2**(9):669-680.

[98] De Vries BB, Winter R, Schinzel A*, et al.* Telomeres: a diagnosis at the end of the chromosomes. *J Med Genet* 2003;**40**(6):385-398.

[99] Bailey W, Popovich B, Jones KL. Monozygotic twins discordant for the Russell-Silver syndrome. *Am J Med Genet* 1995; **58**(2):101-105.

[100] Geschwind DH, Levitt P. Autism spectrum disorders: developmental disconnection syndromes. *Curr Opin Neurobiol* 2007; **17**(1):103-111.

[101] Risch N, Spiker D, Lotspeich L, Nouri N, Hinds D, Hallmayer J *et al.* A genomic screen of autism: evidence for a multilocus etiology. *Am J Hum Genet* 1999; **65**(2):493-507.

[102] Barnby G, Abbott A, Sykes N, Morris A, Weeks DE, Mott R, *et al*; International Molecular Genetics Study of Autism Consortium (IMGSAC) : Candidate-gene screening and association analysis at the autism-susceptibility locus on chromosome 16p: evidence of association at GRIN2A and ABAT. *Am J Hum Genet* 2005; **76**: 950-966.

[103] Niikawa N, Matsuura N, Fukushima Y, et al. Kabuki make-up syndrome: a syndrome of mental retardation, unusual facies, large and protruding ears, and postnatal growth deficiency. *J Pediatr* 1981;**99**(4):565-569.

[104] Kuroki Y, Suzuki Y, Chyo H, et al. A new malformation syndrome of long palpebral fissures, large ears, depressed nasal tip, and skeletal anomalies associated with postnatal dwarfism and mental retardation. *J Pediatr* 1981;**99**(4):570-573.

[105] Matsumoto N, Niikawa N. Kabuki make-up syndrome: a review. *Am J Med Genet C Semin Med Genet* 2003;**117**(1):57-65.

[106] Milunsky JM HX. Unmasking Kabuki syndrome: chromosome 8p22-23.1 duplication revealed by comparative genomic hybridization and BAC-FISH. *Clin Genet* 2003; **64**:509-516.

[107] Sanlaville D, Genevieve D, Bernardin C, Amiel J, Baumann C, de Blois MC, *et al.*Failure to detect an 8p22-8p23.1 duplication in patients with Kabuki (Niikawa-Kuroki) syndrome. *Eur J Hum Genet* 2005

[108] David D, Santos IM, Johnson K, Tuddenham EG, McVey JH. Analysis of the consequences of premature termination codons within factor VIII coding sequences. *J Thromb Haemost* 2003; **1**(1):139-146.

[109] Wagner FF, Flegel WA. RHD gene deletion ocurred in the Rhesus box. *Blood* 2000; **95**: 3662-3668.

[110] Singleton AB, Farrer M, Johnson J, Singleton A, Hague S, Kachergus J, *et al.* alpha-Synuclein locus triplication causes Parkinson's disease. *Science* 2003 ;**302**(5646):841

[111] Yang Y, Chung EK, Wu YL, Savelli SL, Nagaraja HN, Zhou B, *et al.* Gene copy-number variation and associated polymorphisms of complement component C4 in human systemic lupus erythematosus (SLE): low copy number is a risk factor for and high copy number is a protective factor against SLE susceptibility in European Americans. *Am J Hum Genet* 2007; **80**(6):1037-1054.

[112] Kraft HG, Lingenhel A, Köchl S, Hoppichler F, Kronenberg F, Abe A, *et al.* Apolipoprotein(a) kringle IV repeat number predicts risk for coronary heart disease. *Arterioscler Thromb Vasc Biol* 1996; **16**(6):713-719

[113] Lahortiga I, De Keersmaecker K, Van Vlierberghe P, Graux C, Cauwelier B, Lambert F, *et al.* Duplication of the MYB oncogene in T cell acute lymphoblastic leukemia. *Nat Genet* 2007; **39**(5):593-595.

[114] Le Maréchal C, Masson E, Chen JM, Morel F, Ruszniewski P, Levy P, Férec C. Hereditary pancreatitis caused by triplication of the trypsinogen locus. *Nat Genet* 2006;**38**(12):1372-1374.

[115] Frank B, Bermejo JL, Hemminki K, Sutter C, Wappenschmidt B, Meindl A, *et al.* Copy number variant in the candidate tumor suppressor gene MTUS1 and familial breast cancer risk. *Carcinogenesis* 2007; **28**(7):1442-1445.

[116] Fellermann K, Stange DE, Schaeffeler E, Schmalzl H, Wehkamp J, Bevins CL, *et al.* A chromosome 8 gene-cluster polymorphism with low human beta-defensin 2 gene copy number predisposes to Crohn disease of the colon. *Am J Hum Genet* 2006; **79**(3):439-448.

[117] Aldred PM, Hollox EJ, Armour JA. Copy number polymorphism and expression level variation of the human alpha-defensin genes DEFA1 and DEFA3. *Hum Mol Genet* 2005; **14**(14):2045-2052.

[118] Gonzalez E, Kulkarni H, Bolivar H, Mangano A, Sanchez R, Catano G *et al.* The influence of CCL3L1 gene-containing segmental duplications on HIV/ AIDS succeptibility. *Science* 2005; **307**:1434-1440.

[119] McKinney C, Merriman ME, Chapman PT, Gow PJ, Harrison AA, Highton J,*et al.* Evidence for an influence of chemokine ligand 3-like 1 (CCL3L1) gene copy number on susceptibility to rheumatoid arthritis. *Ann Rheum Dis* 2008; **67**(3):409-413.

[120] Ingelman-Sundberg M, Sim SC, Gomez A, Rodriguez-Antona C. Influence of cytochrome P450 polymorphisms on drug therapies: pharmacogenetic, pharmacoepigenetic and clinical aspects. *Pharmacol Ther* 2007; **116**(3):496-526.

[121] Deeb SS. Genetics of variation in human color vision and the retinal cone mosaic. *Curr Opin Genet Dev* 2006; **16**(3):301-307.

[122] Hughes AE, et al. A common CFH haplotype with deletion of CFHR1 and CHFR3 is associated to lower risk of age-related macular degeneration. *Nat Genet* 2006. **38**;1173-1177.

[123] Cuscó I, del Campo M, Vilardell M, González E, Gener B, Galán E, Toledo L, Pérez-Jurado LA. Array-CGH in patients with Kabuki-like phenotype: identification of two patients with complex rearrangements including 2q37 deletions and no other recurrent aberration. *BMC Med Genet* 2008; **11**:9-27.

[124] Chen JM, Cooper DN, Chuzhanova N, Férec C, Patrinos GP. Gene conversion: mechanisms, evolution and human disease. *Nat Rev Genet* 2007; **8**(10):762-75.

[125] Hatton CE, Cooper A, Whitehouse C, Wraith JE. Mutation analysis in 46 British and Irish patients with Gaucher's disease. *Arch Dis Child* 1997;**77**:17-22

[126] Latham T, Grabowski GA, Theophilus BD, Smith FI. Complex alleles of the acid β-glucosidase in Gaucher disease. *Am J Hum Genet* 1990;**47**:79-86.

[127] Eyal N, Wilder S, Horowitz M. Prevalent and rare mutations among Gaucher patients. *Gene* 1990;**96**:277-283

[128] Hong CM, Ohashi T, Yu XJ, Weiler S, Barranger JA. Sequence ot two alleles responsible for Gaucher disease. *DNA Cell Biol* 1990;**9**:233-241

[129] Heinen S et al. De novo gene conversion in the RCA gene cluster (1q32) causes mutations in complement factor H associated with atypical hemolyc uremic syndrome. *Hum Mutat* 2006;**27**:292-293.

[130] Bussaglia E, Clermont O, Tizzano E, Lefebvre S, Bürglen L, Cruaud C, Urtizberea JA, Colomer J, Munnich A, Baiget M, et al. A frame-shift deletion in the survival motor neuron gene in Spanish spinal muscular atrophy patients. *Nature Genet* 1995;**11**:335-337

[131] Lee HH, Tsai FJ, Lee YJ, Yang YC. Diversity of the CYP21A2 gene: a 6.2kb TaqI fragment and 3.2 TaqI fragment mistaken as CYP21A1P. *Mol Genet Metab* 2006;**88**:58-65.

[132] Nicolis E, Bonizzato A, Assael BM, Cipolli M. Identification of novel mutations in patients with Schwachman Diamond syndrome. 2005 *Hum Mutat*;**25**:410

[133] Vázquez N, Lehrnbecher T, Chen R, Christensen BL, Gallin JI, Malech H, Holland S, Zhu S, Chanock SJ. Mutational analysis of patients with p47-phox deficient chronic granulomatous disease: the significance of recombination events between the p47-phox gene (NCF1) and its highly homologous pseudogenes. *Exp Hematol* 2001;**29**:234-243

[134] Teich N, et al. Gene conversion between functional trypsinogen genes PRSS1 and PRSS2 associated to chronic pancreatitis in six-year-old girl. *Hum Mutat* 2005; **25**:343-347

[135] Nicod J, Dick B, Frey FJ, Ferari P. Mutation analysis of CYP11B1 and CYP11B2 in patients with increased 18-hydroxycortisol production. *Mol Cell Endocrinol* 2004 ;**214**:167-174

[136] Adams JG, Marrison WT, Steinberg MH. Hemogloblin Parchman: double crossover within a single human gene. *Science* 1982;**218**:291-293

[137] Patrinos GP, Kollia P, Loutradi-Anagnostou A, Loukopoulos D, Papadakis MN. The Cretan type of non-deletional hereditary persistence of fetal hemoglobin —Aγ-158 C ¿T— results from two independent gene conversion events. *Hum Genet* 1998;**102**:629-634

[138] De Marco P, Moroni A, Merello E, de Franchis R, Andreussi L, Finnell RH, Barber RC, Cama A, Capra V. Folate pathway alterations in patients with neural tube defects. *Am J Med Genet* 2000;**95**:216-223

[139] Watnick TJ, Gandolph MA, Weber H, Neumann HP, Germino GG. Gene conversion is a likely cause of mutation in PKD1. 1998 *Hum Mol Genet*;**7**:1239-1243

[140] Inoue S, Inoue K, Utsunomiya M, Nozaki J, Yamada Y, Iwasa T, Mori E, Yoshinaga T, Koizumi A. Mutation analysis in PKD1 of Japanese autosomal dominant polycystic kidney disease patients. 2002. *Hum Mutat*;**19**:622-628

[141] Millar DS, Lewis MD, Horan M, Newsway V, Easter TE, Gregory JW, *et al.* Novel mutations of the growth hormone 1 (GH1) gene disclosed by modulation of the clinical selection criteria for individuals with short stature. *Hum Mutat* 2003;**21**:424-440

[142] Fardella CE, Rodriguez H, Montero J, Zhang G, Vignolo P, Rojas A, Villarroel L, Miller WL. Gene conversion in the CYP11B2 gene encoding P450c11AS is associated with, but not cause, the syndrome of corticosterone methyloxidase II deficiency. *J Clin Endocrinol Metab* 1996;**81**:321-326

[143] Vanita, Sarhadi V, Reis A, Jung M, Singh D, Sperling K, Singh JR, Bürger J. A unique form of autosomal dominant cataract explained bye gene conversion between β-crystallin B2 and its pseudogen. *J Med Genet* 2001;**38**:392-396

[144] Eikenboom JC, Castman G, Vos HL, Bertina RM, Rodeghiero F. Characterization of the genetic defects in recessive type I and type 3 von Willebrand disease patients of italian origin. *Thromb Haemost* 1998;**79**:709-717

[145] Minegishi Y, Lavoie A, Cunningham-Rundles C, Bédard PM, Hébert J, Côté L, *et al.* Mutations in the human λ/14.1 gene results in B cell deficiency and agammaglobulinemia. *J Exp Med* 1998;**187**:71-77

[146] Roesler J, Curnutte JT, Rae J, Barrett D, Patino P, Chanock SJ, Goerlach A. Recombination events between the p47-phox gene and its highly homologous pseudogenes are the main cause of autosomal recessive chronic granulomatous disease. *Blood* 2000;**95**:2150-2156

[147] Reyniers E, Van Thienen MN, Meire F, De Boulle K, Devries K, Kestelijn P, Willems PJ. Gene conversion between red and defective green opsin gene in blue cone monochromacy. 1995. *Genomics*;**29**:323-328

[148] Vermeesch JR, Melotte C, Froyen G, Van Vooren S, Dutta B, Maas N et al. Molecular karyotyping: array CGH quality criteria for constitutional genetic diagnosis. *J Histochem Cytochem* 2005;**53**(3):413-422.

[149] Veltman JA, Jonkers Y, Nuijten I, Janssen I, van d, V, Huys E et al. Definition of a critical region on chromosome 18 for congenital aural atresia by arrayCGH. *Am J Hum Genet* 2003;**72**(6):1578-1584.

[150] Zhang X, Snijders A, Segraves R, Zhang X, Niebuhr A, Albertson D, *et al.* High-resolution mapping of genotype-phenotype relationships in cri du chat syndrome using array comparative genomic hybridization. *Am J Hum Genet* 2005; **76**(2):312-326.

[151] Shaw CJ, Shaw CA, Yu W, Stankiewicz P, White LD, Beaudet AL, Lupski JR. Comparative genomic hybridisation using a proximal 17p BAC/PAC array detects rearrangements responsible for four genomic disorders. *J Med Genet* 2004;**41**(2):113-119.

[152] Mendrzyk F, Radlwimmer B, Joos S, Kokocinski F, Benner A, Stange DE et al.: Genomic and protein expression profiling identifies CDK6 as novel independent prognostic marker in medulloblastoma. *J Clin Oncol* 2005; **23**: 8853-8862.

[153] Snijders AM, Nowak N, Segraves R, Blackwood S, Brown N, Conroy J,*et al.* Assembly of microarrays for genome-wide measurement of DNA copy number. *Nat Genet* 2001; **29**(3):263-4.

[154] Wang NJ, Liu D, Parokonny AS, Schanen NC: High-resolution molecular characterization of 15q11-q13 rearrangements by array comparative genomic hybridization (array CGH) with detection of gene dosage. *Am J Hum Genet* 2004, **75**: 267-281.

[155] Shafer JL. The lmm package. *R package* 2007.

[156] Chen Z, Liu L. RealSpot: software validating results from DNA microarray data analysis with spot images. Physiol *Genomics* 2005; **21**(2):284-91.

[157] Pepe MS. Receiver Operating Characteristic Methodology. *J Am Stat Assoc* 2000;**95**449:308-311

[158] Schwender, H., Krause, A., and Ickstadt, K. Identifying Interesting Genes with siggenes. *RNews* 2006; **6**(5): 45-50.

[159] Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Roy Stat Soc B* 1995; **B57**:289-300.

[160] Joshi-Tope G, Gillespie M, Vastrik I, Deustachio P, Schmidt E, Bono BD *et al.* REACTOME: a Knowledgebase of biological pathways. *Nucl Acids Res* 2005; **33**:428-432.

[161] Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, Brazma A, Huber W. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics* 2005; **21**(16):3439-40.

[162] Birney E, Andrews D, Caccamo M, Chen Y, Clarke L, Coates G, *et al*: Ensembl 2006.*Nucleic Acids Res* 2006; **34**:D556-561.

[163] Kent WJ. BLAT-The BLAST-Like Alignment Tool. *Genome Res* 2002; **12**:656-664.

[164] Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, *et al.* Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biology*, **5**:R80.

[165] Ribeiro PJ, Diggle PJ geoR: A package for geostatistical analysis. *R-NEWS* 2001; **1**(2):1609-3631.

[166] Khojasteh M, Lam WL, Ward RK, Macaulay C: A stepwise framework for the normalization of array CGH data. *BMC Bioinformatics* 2005; **6**:274-288.

[167] Fiegler H, Redon R, Andrews D, Scott C, Andrews R, Carder C *et al.* Accurate and reliable high-throughput detection of copy number variation in the human genome. *Genome Res* 2006; **16**: 1566-1574.

[168] Qin LX, Kerr KF. Empirical evaluation of data transformations and ranking statistics for microarray analysis. *Nucleic Acids Res* 2004; **32**: 5471-5479.

[169] Scharpf RB, Iacobuzio-Donahue CA, Sneddon JB, Parmigiani G: When should one subtract background fluorescence in two color microarrays? *Biostatistics* 2007, in press

[170] Allison DB, Cui X, Page GP, Sabripour M. Microarray data analysis: from disarray to consolidation and consensus. *Nature Rev* 2006; **7**; 55-65.

[171] Wu Z, Irizarry RA, Gentleman R, Martinez-Murillo F, Spencer F. A model-based background adjustment for oligonucleotide Expression Arrays. *J Am Stat Assoc* 2004; **99** (458); 909-917.

[172] Hong F, Breitling R. A comparison of meta-analysis methods for detecting differentially expressed genes in microarray experiments. *Bioinformatics*. 2008; **24**(3):374-82.

[173] Sharp AJ, Itsara A, Cheung Z, Alkan C, Schwartz S, Eichler EE. Optimal design of nucleotide microarrays for mesurement of DNA copy number. *Hum Mol Gen* 2007. In press

[174] Mary-Huard T, Daudin JJ, Robin S, Bitton F, Cabannes E, Hilson P. Spotting effect in microarray experiments. *BMC Bioinformatics*. 2004; **19**(5):63.

[175] Li Q, Fraley C, Bumgarner RE, Yeung KY, Raftery AE. Donuts, scratches and blanks: robust model-based segmentation of microarray images. *Bioinformatics* 2005;**21**(12):2875-82.

[176] Daskalakis A, Cavouras D, Bougioukos P, Kostopoulos S, Glotsos D, Kalatzis I, *et al.* Improving gene quantification by adjustable spot-image restoration. *Bioinformatics* 2007; **23**(17):2265-2272.

[177] Balagurunathan Y, Wang N, Dougherty ER, Nguyen D, Chen Y, Bittner ML, Trent J, Carroll R. Noise factor analysis for cDNA microarrays. *J Biomed Opt* 2004; **9**(4):663-678.

[178] Irizarry RA, Warren D, Spencer F, Kim IF, Biswal S, Frank BC *et al.* Multiple-laboratory comparison of microarray platforms. **Nat Methods** 2005;**2**(5):345-350.

[179] Ylstra B, van d, I, Carvalho B, Brakenhoff RH, Meijer GA: BAC to the future! or oligonucleotides: a perspective for micro array comparative genomic hybridization (array CGH). Nucleic Acids Res 2006; 34: 445-450.

[180] Hughes TR, Roberts CJ, Dai H, Jones AR, Meyer MR, Slade D, Burchard J, Dow S, Ward TR, Kidd MJ, Friend SH, Marton MJ. Widespread aneuploidy revealed by DNA microarray expression profiling. Nat Genet. 2000 Jul;25(3):333-7.

[181] Melendez B, Diaz-Uriarte R, Cuadros M, Martínez-Ramírez A, Fernández-Piqueras J, Dopazo A, *et al.* Gene expression analysis of chromosomal regions with gain or loss of genetic material detected by comparative genomic hybridization. *Gen Chrom Can* 2004; **41**(4):353-365.

[182] Morley M, Molony CM, Weber TM, Devlin JL, Ewens KG, Spielman RS, Cheung VG. Genetic analysis of genome-wide variation in human gene expression. *Nature* 2004; **430**(7001):743-747.

[183] Merla G, Howald C, Henrichsen CN, Lyle R, Wyss C, Zabot MT, Antonarakis SE, Reymond A. Submicroscopic deletion in patients with Williams-Beuren syndrome influences expression levels of the nonhemizygous flanking genes. *Am J Hum Genet* 2006; **79**(2):332-341.

[184] Valdez GR, Inoue K, Koob GF, Rivier J, Vale W, Zorrilla EP. Human urocortin II: mild locomotor suppressive and delayed anxiolytic-like effects of a novel corticotropin-releasing factor related peptide. *Brain Res* 2002;**943**(1):142-150.

[185] Ong KL, Wong LY, Man YB, Leung RY, Song YQ, Lam KS, Cheung BM. Haplotypes in the urotensin II gene and urotensin II receptor gene are associated with insulin resistance and impaired glucose tolerance. *Peptides* 2006; **27**(7):1659-16667.

[186] Nothwang HG, Kim HG, Aoki J, Geisterfer M, Kübart S, Wegner RD, *et al.*Functional hemizygosity of PAFAH1B3 due to a PAFAH1B3-CLK2 fusion gene in a female with mental retardation, ataxia and atrophy of the brain.*Hum Mol Genet* 2001; **10**(8):797-806.

[187] Pan W, Lin J, Le CT. How many replicates of arrays are required to detect gene expression changes in microarray experiments? A mixture model approach. *Genome Biol* 2002; **3**(5):

[188] Kerr MK. Design considerations for efficient and effective microarray studies. *Biometrics* 2003; **59**: 822-828.

[189] Wernisch L. Can replication save noisy microarray data?. *Comp Funct Genom* 2002; **3**:372-374

[190] Estivill X, Cheung J, Pujana MA, Nakabayashi K, Scherer SW, Tsui LC. Chromosomal regions containing high-density and ambigously mapped putative single nucleotide polymorphisms (SNPs) correlate with segmental duplications in the human genome. *Hum Mol Gen* 2002; **11**(17):1987-1995.

[191] Brosius J. The contribution of RNAs and retrotransposition to evolutionary novelties. *Genetica* 2003; **118**

[192] Nekrutenko A, Makova KD, Li WH. The Ka/Ks Ratio Test for Assessing the Protein-Coding Potential of Genomic Regions: An Empirical and Simulation Study. *Genome Res* 2001; **12**: 198-202.

[193] Marais G. Biased gene conversion: implications for genome and sex evolution. *Trends Genet* 2003; **19**(6): 330-338.

[194] Buratti E, Baralle M, Baralle FE. Defective splicing, disease and therapy: searching for master checkpoints in exon definition. *Nucl Ac Res* 2006; **34**(12):3494-3510.

9. Anexos

9.1. Especificaciones de los modelos

Por cada modelo ANOVA se muestra; (i) el modelo aplicado junto con la parametrización y las hipótesis, (ii) Las esperanzas de los cuadrados medios (iii) los cuadrados medios y (iv) las sumas de cuadrados.

9.1.1. Condiciones óptimas para el proceso de impresión

$$y_{biklpn} = BAC_b + DB_i + Conc_k + Sol_l + P(DB)_p + BAC:DB_{bi} + DB:Conc_{ik} + +DB:Sol_{il} + Conc:Sol_{kl} BAC:Conc_{bk} + \\ +BAC:Solbl + BAC:DB:Conc_{bik} + BAC:DB:Sol_{bil} + BAC:Conc:Sol_{bkl} + BAC:DB:Conc:Sol_{bikl} + e_{biklpr}$$

$BAC_j \sim N(0, \sigma_{BAC})$	$H_0 : \sigma_{BAC} = 0$	$\forall b = 1, B$
$\sum_{i=1}^{I} DB_i = 0$	$H_0 : DB_i = 0$	$\forall i = 1, I$
$\sum_{k=1}^{K} Conc_k = 0$	$H_0 : Conc_k = 0$	$\forall k = 1, K$
$\sum_{l=1}^{L} Sol_l = 0$	$H_0 : Sol_l = 0$	$\forall l = 1, L$
$P(DB)_i \sim N(0, \sigma_P)$	$H_0 : \sigma_P = 0$	$\forall p = 1, P$
$BAC:DB_{bi} \sim N(0, \sigma_{BAC:DB})$	$H_0 : \sigma_{BAC:DB}$	$\forall b = 1, B; i = 1, I$
$\sum_{i}^{I}\sum_{k}^{K} DB:Conc_{ik}$	$H_0 : DB:Conc_{ik} = 0$	$\forall i = 1, I; k = 1, K$
$\sum_{i=1}^{I}\sum_{l=1}^{L} DB:Sol_{il}$	$H_0 : DB:Sol_{il} = 0$	$\forall i = 1, I; l = 1, L$
$\sum_{k=1}^{K}\sum_{l=1}^{L} Conc:Sol_{kl}$	$H_0 : Conc:Sol_{kl} = 0$	$\forall k = 1, K; l = 1, L$
$BAC:Conc_{bk} \sim N(0, \sigma_{BAC:Conc}$	$H_0 : \sigma_{BAC:Conc}$	$\forall b = 1, B; k = 1, K$
$BAC:Sol_{bl} \sim N(0, \sigma_{BAC:Sol})$	$H_0 : \sigma_{BAC:Sol}$	$\forall b = 1, B; l = 1, L$
$\sum_{i=1}^{I}\sum_{k=1}^{K}\sum_{l=1}^{L} DB:Conc:Sol_{ikl}$	$H_0 : DB:Conc:Sol_{ikl} = 0$	$\forall i = 1, I; k = 1, K; l = 1, L$
$BAC:DB:Conc_{bil} \sim N(0, \sigma_{BAC:DB:Conc})$	$H_0 : \sigma_{BAC:DB:Conc}$	$b = 1, B; i = 1, I; k = 1, K$
$BAC:DB:Sol_{bil} \sim N(0, \sigma_{BAC:DB:Sol})$	$H_0 : \sigma_{BAC:DB:Sol}$	$b = 1, B; i = 1, I; l = 1, L$
$BAC:Conc:Sol_{bkl} \sim N(0, \sigma_{BAC:Conc:Sol})$	$H_0 : \sigma_{BAC:Conc:Sol}$	$b = 1, B; k = 1, K, l = 1, L$
$BAC:DB:Conc:Sol_{bikl} \sim N(0, \sigma_{BAC:DB:Conc:Sol})$	$H_0 : \sigma_{BAC:DB:Conc:Sol}$	$b = 1, B; i = 1, I; k = 1, K; l = 1, L$
$e_{biklpr} \sim N(0, \sigma)$	$H_0 : \sigma = 0$	$\forall r = 1, R$

(7)

$$
\begin{aligned}
E\left(MS_{DB}\right) &= \sigma^2 + R*\sigma^2_P + K*L*R*\sigma^2_{BAC:DB} + \frac{B*K*L*R*\sum_{i=1}^{I} DB_i^2}{I-1} \\
E\left(MS_{BAC}\right) &= \sigma^2 + I*K*L*R*\sigma^2_{BAC} \\
E\left(MS_{Conc}\right) &= \sigma^2 + I*L*R*\sigma^2_{BAC:Conc} + \frac{B*I*L*R*\sum_{k=1}^{k} Conc_k^2}{K-1} \\
E\left(MS_{Sol}\right) &= \sigma^2 + I*K*R*\sigma^2_{BAC:Sol} + \frac{B*I*K*R*\sum_{l=1}^{L} Sol_l^2}{L-1} \\
E\left(MS_{P(DB)}\right) &= \sigma^2 + R*\sigma^2_P \\
E\left(MS_{BAC:DB}\right) &= \sigma^2 + K*L*R*\sigma^2_{BAC:DB} \\
E\left(MS_{BAC:Conc}\right) &= \sigma^2 + I*L*R*\sigma^2_{BAC:Conc} \\
E\left(MS_{BAC:Sol}\right) &= \sigma^2 + I*K*R*\sigma^2_{BAC:Sol} \\
E\left(MS_{Conc:Sol}\right) &= \sigma^2 + R*\sigma^2_{BAC:DB:Conc:Sol} + I*R*\sigma^2_{BAC:Conc:Sol} + \\
&\quad + \frac{B*K*R*\sum_{i=1}^{I} DB_i^2 \sum_{l=1}^{L} Sol_l^2}{(I-1)(L-1)} \\
E\left(MS_{DB:Conc}\right) &= \sigma^2 + R*\sigma^2_{BAC:DB:Conc:Sol} + L*R*\sigma^2_{BAC:DB:Conc} + \\
&\quad + \frac{B*L*R*\sum_{i=1}^{I} DB_i^2 \sum_{k=1}^{K} Conc_k^2}{(I-1)(K-1)} \\
E\left(MS_{DB:Sol}\right) &= \sigma^2 + +B*K*R*\sigma^2_{DB:Sol} \\
E\left(MS_{BAC:DB:Conc}\right) &= \sigma^2 + L*R*\sigma^2_{BAC:DB:Conc} \\
E\left(MS_{BAC:DB:Sol}\right) &= \sigma^2 + K*R*\sigma^2_{BAC:DB:Sol} \\
E\left(MS_{DB:Conc:Sol}\right) &= \sigma^2 + R*\sigma^2_{BAC:DB:Conc:Sol} + B*R*\sigma^2_{DB:Conc:Sol} \\
E\left(MS_{BAC:Conc:Sol}\right) &= \sigma^2 + I*R*\sigma^2_{BAC:Conc:Sol} \\
E\left(MS_{BAC:DB:Conc:Sol}\right) &= \sigma^2 + R*\sigma^2_{BAC:DB:Conc:Sol} \\
E\left(MS_{R}\right) &= \sigma^2
\end{aligned}
$$

$$
\begin{array}{lcl}
MS_{DB} & = & SS_{DB}/\left(I-1\right) \\
MS_{BAC} & = & SS_{BAC}/(B-1) \\
MS_{Conc} & = & SS_{Conc}/(K-1) \\
MS_{Sol} & = & SS_{Sol}/(L-1) \\
MS_{P(DB)} & = & SS_{P}/(P-1)I \\
MS_{DB:Sol} & = & SS_{DB:Sol}/(I-1)(L-1) \\
MS_{DB:Conc} & = & SS_{DB:Conc}/(I-1)(K-1) \\
MS_{BAC:Sol} & = & SS_{BAC:Sol}/(B-1)(L-1) \\
MS_{BAC:Conc} & = & SS_{BAC:Conc}/(B-1)(K-1) \\
MS_{BAC:DB:Sol} & = & SS_{BAC:DB:Sol}/(B-1)(I-1)(L-1) \\
MS_{BAC:DB:Conc} & = & SS_{BAC:DB:Conc}/(B-1)(I-1)(K-1) \\
MS_{BAC:Conc:Sol} & = & SS_{BAC:Conc:Sol}/(B-1)(K-1)(L-1) \\
MS_{BAC:DB:Conc:Sol} & = & SS_{BAC:DB:Conc:Sol}/(B-1)(I-1)(K-1)(L-1)
\end{array}
\tag{8}
$$

9.1.2. Detección de fuentes de variación asociadas a la fiabilidad de la medida

El modelo reducido

$$y_{bijksr} = \mu_b + DB_{bi} + Tecnico_{bt} + Dia_{ba} + P\left(DB, Tecnico, Dia\right)_{bp(ita)} + e_{bitapr}$$

$$
\begin{array}{lll}
\sum_{i=1}^{I} DB_{bi} = 0 & H_0 : DB_{bi} = 0 & \forall i = 1, I \\
Tecnico_{tb} \sim N(0, \sigma_T) & H_0 : \sigma_T = 0 & \forall t = 1, T \\
Dia_{ab} \sim N(0, \sigma_D) & H_0 : \sigma_D = 0 & \forall a = 1, A \\
P_{p(ita)b} \sim N(0, \sigma_P) & H_0 : \sigma_P = 0 & \forall p = 1, P \\
e_{itaprb} \sim N(0, \sigma) & H_0 : \sigma = 0 & \forall r = 1, R
\end{array}
$$

$$
\begin{array}{lcl}
E\left(MS_{DB}\right) & = & \sigma^2 + R * \sigma_P^2 + \frac{T*A*P*R* \sum_{i=1}^{I} DB_i^2}{I-1} \\
E\left(MS_{T}\right) & = & \sigma^2 + R * \sigma_P^2 + I * A * P * R * \sigma_T^2 \\
E\left(MS_{D}\right) & = & \sigma^2 + R * \sigma_P^2 + I * T * P * R * \sigma_D^2 \\
E\left(MS_{P(DB)}\right) & = & \sigma^2 + R * \sigma_P^2 \\
E\left(MS_{R}\right) & = & \sigma^2
\end{array}
$$

$$
\begin{aligned}
MS_{DB} &= SS_{DB}/(I-1) \\
MS_T &= SS_T/(T-1) \\
MS_D &= SS_D/(A-1) \\
MS_{P(DB)} &= SS_P/I*T*A*(P-1) \\
MS_R &= SS_R/I*T*A*P*(R-1)
\end{aligned} \tag{9}
$$

$$
\begin{aligned}
SS_{total} &= \sum_{i=1}^{I}\sum_{t=1}^{T}\sum_{a=1}^{A}\sum_{p=1}^{P}\sum_{r=1}^{R}(y_{bitapr} - \overline{y_{b.....}})^2 \\
SS_{DB} &= T*A*P*N*\sum_{i=1}^{I}(\overline{y_{bi....}} - \overline{y_{b.....}})^2 \\
SS_T &= I*A*P*N*\sum_{t=1}^{T}(\overline{y_{b.t...}} - \overline{y_{b.....}})^2 \\
SS_D &= I*T*P*N*\sum_{a=1}^{A}(\overline{y_{b..a..}} - \overline{y_{b.....}})^2 \\
SS_P &= R*\left[\sum_{i=1}^{I}\sum_{t=1}^{T}\sum_{a=1}^{A}\sum_{p=1}^{P}(\overline{y_{bitap.}} - \overline{y_{b.....}})^2\right]
\end{aligned} \tag{10}
$$

El modelo completo

$y_{bitapr} = \mu_g + DB_{bi} + Tecnico_{bt} + Dia_{ba} + (DB:Tecnico)_{bit} + (Tecnico:Dia)_{bta} +$
$+(DB:Tecnico:Dia)_{bita} + P\,(DB, Tecnico, Dia)_{bp(ita)} + e_{bpita}$

$$
\begin{aligned}
&\sum_{i=1}^{I} DB_i = 0 && i = 1, I \\
&Tecnico_{bt} \sim N(0, \sigma_T) && t = 1, T \\
&Dia_{ba} \sim N(0, \sigma_D) && k = 1, K \\
&DB:Tecnico_{bit} \sim N(0, \sigma_{DB:T}) \\
&DB:Dia_{bia} \sim N(0, \sigma_{DB:D}) \\
&DB:Tecnico:Dia_{bita} \sim N(0, \sigma_{DB:T:D}) \\
&P_{bp(ita)} \sim N(0, \sigma_P) && p = 1, P \\
&e_{bp} \sim N(0, \sigma) && r = 1, R
\end{aligned}
$$

$$
\begin{aligned}
E\left(MS_{DB}\right) &= \sigma^2 + R*\sigma^2_P + P*R*\sigma^2_{DB:D:T} + T*P*R*\sigma^2_{DB:D} + I*P*R*\sigma^2_{D:T} + \\
&\quad + A*P*R*\sigma^2_{DB:T} + \frac{T*A*P*R* \sum_{i=1}^{I} DB_i{}^2}{I-1} \\
E\left(MS_T\right) &= \sigma^2 + R*\sigma^2_P + I*P*R*\sigma^2_{D:T} + I*A*P*R*\sigma^2_T \\
E\left(MS_D\right) &= \sigma^2 + R*\sigma^2_P + I*P*R*\sigma^2_{D:T} + I*T*P*R*\sigma^2_D \\
E\left(MS_{DB:T}\right) &= \sigma^2 + R*\sigma^2_P + P*R*\sigma^2_{DB:D:T} + A*P*R*\sigma^2_{DB:T} \\
E\left(MS_{DB:D}\right) &= \sigma^2 + R*\sigma^2_P + P*R*\sigma^2_{DB:D:T} + T*P*R*\sigma^2_{DB:D} \\
E\left(MS_{D:T}\right) &= \sigma^2 + R*\sigma^2_P + I*P*R*\sigma^2_{D:T} \\
E\left(MS_{DB:D:T}\right) &= \sigma^2 + R*\sigma^2_P + P*R*\sigma^2_{DB:D:T} \\
E\left(MS_P\right) &= \sigma^2 + R*\sigma^2_P \\
E\left(MS_R\right) &= \sigma^2
\end{aligned}
$$

Tabla 37: **Tabla ANOVA para el modelo completo**. Fuentes de variación

FACTORES	SS [1]	g.d.l [2]	MS [3]	F Exp	F DIST
DB	SS_{DB}	(I-1)	MS_{DB}	MS_{DB} /(**)	F[(I-1), (**)]
Técnico	SS_T	(T-1)	MS_T	MS_T /$MS_{D:T}$	F[(T-1), (T-1)*(A-1))]
Día	SS_D	(A-1)	MS_D	MS_D /$MS_{D:T}$	F[(A-1), (T-1)*(A-1)]
DB:Técnico	$SS_{DB:T}$	(I-1)*(T-1)	$MS_{DB:T}$	$MS_{DB:T}$ /$MS_{DB:D:T}$	F(I-1)*(T-1),(I-1)*(T-1)*(A-1)]
DB:Día	$SS_{DB:D}$	(I-1)*(A-1)	$MS_{DB:D}$	$MS_{DB:D}$ /$MS_{DB:D:T}$	F[(I-1)*(A-1), (I-1)*(T-1)*(A-1)]
Día:Técnico	$SS_{D:T}$	(A-1)*(T-1)	$MS_{D:T}$	$MS_{D:T}$ /MS_P	F[(I-1)*(T-1), I*T*A*(P-1)]
DB:Día:Técnico	$SS_{DB:D:T}$	(I-1)*(T-1)*(A-1)	$MS_{DB:D:T}$	$MS_{DB:D:T}$ /MS_P	F[(I-1)*(T-1)*(A-1),I*T*A*(P-1)]
P(DB, Técnico, Día)	SS_P	I*T*A*(P-1)	MS_P	MS_P /MS_R	F[I*T*A*(P-1), I*T*A*P*(R-1)]
Residuo	SS_R	I*T*A*P*(R-1)	MS_R		
Total	SS_{Tot}	I*T*A*P*R-1			

[1] (SS); Suma de Cuadrados
[2] (gdl); Suma de Cuadrados
[3] (MS); Cuadrados Medios
(**); Cuasi-F ratios calculados a partir de la apropiada combinación lineal de MS y g.d.l

9.1.3. Estudio del efecto DB y sus causas

$$
\begin{aligned}
& y_{bimur} = \mu_b + DB_{bi} + Muestra_{bm} + Genero_{bu} + e_{bimur} \\
& \sum_{i=1}^{I} DB_{bi} = 0 \qquad H_0 : DB_{bi} = 0 \qquad \forall i = 1, I \\
& \sum_{m=1}^{M} Muestra_{bm} = 0 \quad H_0 : Muestra_{bm} = 0 \quad \forall m = 1, M \\
& \sum_{u=1}^{U} Genero_{bu} = 0 \qquad H_0 : Genero_{bu} = 0 \qquad \forall u = 1, U \\
& e_{bimur} \sim N(0, \sigma) \qquad \forall r = 1, R
\end{aligned}
\tag{11}
$$

$$
\begin{aligned}
MS_{DB} &= \sigma^2 + \frac{M*U*R* \sum_{i=1}^{I} DB_i^2}{I-1} \\
MS_{Muestra} &= \sigma^2 + \frac{I*U*R* \sum_{m=1}^{M} Muestra_i^2}{M-1} \\
MS_{Genero} &= \sigma^2 + \frac{I*M*R* \sum_{u=1}^{U} Genero_i^2}{U-1} \\
MS_R &= \sigma^2
\end{aligned}
\tag{12}
$$

$$
\begin{aligned}
&MS_{DB} = SS_{DB}/(I-1) \\
&MS_{Muestra} = (SS_{Muestra} - SS_{Genero}) \,/\, [(M-1)-(U-1)] \\
&MS_{Genero} = SS_{Genero}/(U-1) \\
&MS_R = SS_R/\,[(I*M*U-1)-[(I-1)+(M-1)-(U-1)+(U-1)]]
\end{aligned}
\tag{13}
$$

$$
\begin{aligned}
SS_{total} &= \sum_{i=1}^{I}\sum_{m=1}^{M}\sum_{u=1}^{U}\sum_{r=1}^{R}(y_{bimur} - \overline{y_{b....}})^2 = SS_{DB} + SS_{Muestra} + SS_{Genero} + SS_R \\
SS_{DB} &= M*R*\sum_{i=1}^{I}(\overline{y_{bi...}} - \overline{y_{b....}})^2 \\
SS_{Muestra} &= M*R*\sum_{m=1}^{M}(\overline{y_{b.j..}} - \overline{y_{b....}})^2
\end{aligned}
\tag{14}
$$

La suma de cuadrados se pondera tal y como se describe a continuación (ver ecuación 15) para desigual número de réplicas por bloque.

$$
\begin{aligned}
&si \quad i=1 \quad Ut_1 = I^2/\sum_{u=1}^{U} N_{1u} \\
&si \quad i=2 \quad Ut_2 = I^2/\sum_{u=1}^{U} N_{2u} \\
&y^*_{b....} = \sum_{u=1}^{U} Ut_u * \overline{y_{b..u.}}/\sum_{u=1}^{U} Ut_u \\
&SS_{Genero} = \sum_{u=1}^{U} Ut_u * (\overline{y_{b..u.}} - y^*_{b....})^2
\end{aligned}
\tag{15}
$$

9.2. Cálculo del tamaño muestral en diseño experimental

El cálculo del tamño muestral en micromatrices viene referido a conocer el número de réplicas necesarias para testar una condición de manera independiente del número de genes que sean evaluados.

En general el tamaño muestral necesario depende de:

1. La magnitud mínima de cambio que se desea conocer
2. La probabilidad de detectar el cambio (poder estadístico 1- Error Tipo II)
3. El Error Tipo I

Los diseños experimentales clásicos aplicados a matrices (aExpr y aCGH) cubren dos aspectos:

1. Estimación de las fuentes de variación asociadas a micromatrices.
2. Detectar regiones con diferencial de dosis génica o detectar genes diferencialmente expresados.

Si puede considerarse que el diseño está completamente aleatorizado, el cálculo del número de réplicas necesarias en diseños balanceados se reduce a conocer el número de réplicas necesarias para un factor;

$$\alpha_0 = \frac{E(R_0)}{G_0} \tag{16}$$

Dónde α_0 es el error tipo I para cualquier gen g pertenciente a G_0. $E(R_0)$ es el número de falsos positivos esperados al aplicar el mismo procedimiento a G_0 genes o sondas. G_0 es el número de genes del conjunto de genes G global dónde no se esperan cambios. Este valor es desconocido.

Entonces el número de réplicas necesarias puede escribirse como:

$$n = \left(\frac{z_a + z_b}{|\mu_1|/2 * \sigma} \right)^2 \tag{17}$$

Dónde n es el número de réplicas de cada grupo.μ_1 es la diferencia mínima que se desea detectar entre dos dos niveles del factor. z_a and z_b son los percentiles a $(1 - \alpha/2)$ y b (la potencia deseada) de la distribución normal de media 0 y desviación estándar 1. σ es la variabilidad residual del modelo ANOVA que puede ser estimado a partir del cuadrado medio del error. Habitualmente esta estimación es proporcionada por un estudio piloto anterior al estudio en cuestión. Ello asume que cada uno de los grupos evaluados posee la misma variabilidad.

La bondad de estos cálculos sólo puede garantizarse para la estimación de los efectos principales por ello existen varios trabajos sobre el cálculo del tamaño muestral apropiado para este tipo de análisis. En todos ellos se concluye que el número mínimo de réplicas por combinación debe estar entre 4 y 8.

9.3. Modelos mixtos para datos jerarquizados; la librería *lmm*

La librería *lmm* permite la estimación máximo-verosímil de los parámetros de un modelo lineal con efectos aleatorios jerarquizados sobre efectos fijos y aleatorios. Los modelos que se pueden ajustar mediante esta librería se definen como:

$$Y_{n_i*1} = X_{n_i*p} \,\%* \,\%\beta + Z_{n_i*q} \,\%* \,\%b_{n_i*1} + e_{n_i*1} \forall i = 1, ..., m \tag{18}$$

Dónde cada elemento y representa los valores observados por cada unidad que puede ser un sujeto o un unidad de cluster. En esta tesis doctoral la unidad jerarquizada son los portaobjetos. La matriz X representab los efectos fijos mientras que la matriz Z representa los valores aleatorios. $b_{n_i*1} \sim N(0, psi_i)$ contiene las variabilidades asociadas a los efectos aleatorios. Los errores se distribuyen según $e_{n_i*1} \sim N(0, \sigma^2 * V_{n_i*n_i})$. La matriz $V_{n_i*n_i}$ utilizada en este trabajo es la matriz identidad caracterizada por unos en la diagonal principal y ceros en el resto de elementos.

La estimación de los efectos se ha realizado mediante el método *fastrml.lm* implementado en la librería *lmm* que realiza una estimación maximo-verosímil con restricciones (RML). Este procedimiento tiene un alto grado de convergencia.

Según se ha descrito en esta sección, las matrices empleadas para el experimento 32x se representan a continuación.

$$y_{i,j,k,s,r} = \begin{bmatrix} log_2 x_{1,1,1,1,1} \\ . \\ . \\ . \\ . \\ . \\ . \\ . \\ . \\ log_2 x_{2,2,2,32,3} \end{bmatrix} \; subj = \begin{bmatrix} 1 \\ 2 \\ . \\ . \\ . \\ . \\ . \\ . \\ . \\ 32 \end{bmatrix} \; X = \begin{bmatrix} 1 & 0 \\ . & . \\ . & . \\ . & . \\ 1 & 0 \\ 1 & 1 \\ . & . \\ . & . \\ . & . \\ 1 & 1 \end{bmatrix} \; Z = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \\ 1 & . & . \\ 1 & . & . \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \end{bmatrix} \tag{19}$$

9.4. Protocolo de clasificación de las imágenes

La hoja de recogida de datos consta de 4 columnas;

1. Nombre del BAC
2. Tipo del *spot*
3. Forma del *spot*
4. Concordancia entre réplicas

Por cada BAC se evaluan todas sus réplicas. En el caso de que cada una de las réplicas tenga una forma distinta en la hoja de recogida constará la forma con peor comportamiento según la clasificación siguiente;

- Aspecto esperado (Wt); presenta una coloración totalmente uniforme.
- Aureolas (A); existe un halo de mayor brillantez alrededor del spot pero este es muy homogéneo en su interior.
- Donut (D); presenta una zona central sin color o con un cambio de intensidad muy pronunciado.
- Artefacto (F); presenta pequeñas zonas de alta intensidad en su interior.
- Missing (Ms); el *spot* no está presente

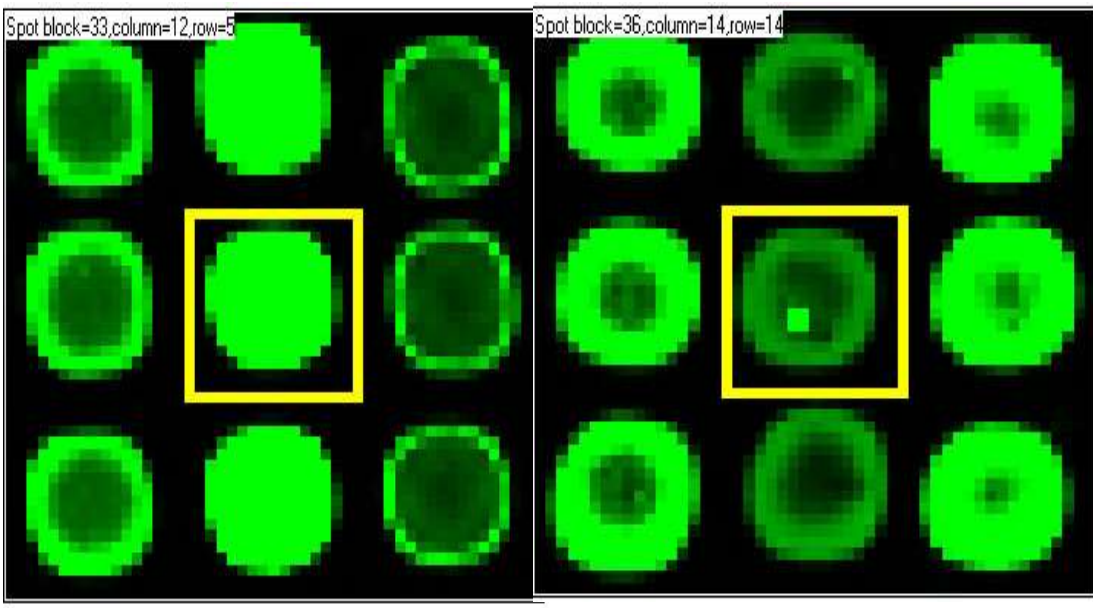

Figura 51: **Tipos de espots**; se muestran las distintas formas de los espots que son evaluadas en este estudio. La primera columna muestra el tipo aureola (A) que se distingue de la clase donut por conservar coloración en su interior. En la segunda columna puede observarse el tipo normal (Wt). En la tercera columna se puede observar el tipo donut (D). En las columnas 4 y 6 se puede observar distintos tipos de aureolas. La columna 5 muestra distintos tipos de donuts y marcado con el recuadro se puede observar el tipo artefacto (F). Todos estos *spots* muestran formas regulares.

Se considera que la calidad de la imagen va en orden decreciente según las etiquetas anteriores.

Estas etiquetas pueden resultar altamente subjetivas, ante la duda se etiquetará con el grado asociado a una peor calidad de la imagen.

La forma del *spot* (0=Regular, 1=Irregular) indica si el *spot* presenta una forma circular (forma regular) o no (forma irregular).

Puesto que puede no existir concordancia entre los tipos y formas según la réplica también se indicará la concordancia entre las réplicas (0=NO, 1=SI).

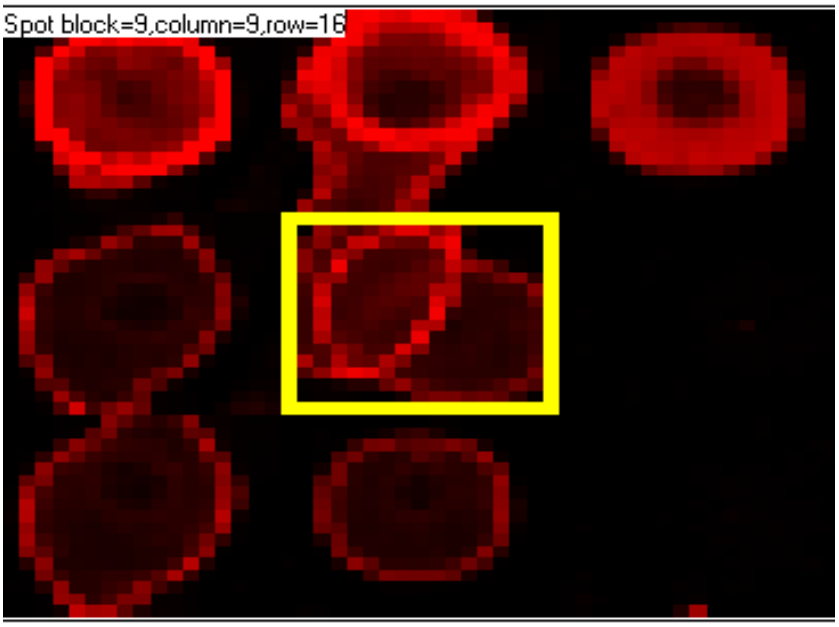

Figura 52: **Formas de los *spots*;** Se muestran un conjunto de formas irregulares que pueden coexistir con distintas formas de los *spots*.

9.5. El método CBS; la librería *DNACopy*

En un primer paso CBS detecta puntos en los que se produce un posible cambio en el número de copias. Estos cambios son utilizados para partir el genoma en segmentos con un número de copias equivalente. La variable utilizada para este procedimiento es el logaritmo en base dos del ratio entre ambos canales normalizado.

El procedimiento de segmentación binaria aplica el test recursivamente hasta que no se detectan más cambios en ninguno de los segmentos obtenidos.

El método CBS considera que cada segmento puede ser cortado por ambos extremos formando un círculo para evitar los problemas que surgían en segmentación binaria como por ejemplo no encontrar un pequeño segmento dentro de un gran segmento con igual número de copias.

El estadístico de test está basado en el ratio de verosimilitud y contrasta la hipótesis de que el arco de i+1 hasta j y su complementario tienen igual media frente la alternativa, que son diferentes.

Siendo $X_1, ..., X_n$ son los log ratios eb base 2 de las intensidades, las cuales están indexadas por sus localizaciones de las n posiciones que están siendo estudiadas. $S_i = X_1 + ... + Xi, \forall 1 \leq i \leq n$ son las sumas parciales. Si los datos están normalmente distribuidos con varianza conocida El estadístico (likelihood ratio) permite testar la hipótesis nula de no cambio versus la alternativa que indica la existencia de un cambio en una localización desconocida i es dada por el $Z_B = max_{1 \leq i < n} |Z_i|$.

$$Z_i = \left\{ \frac{1}{i} + \frac{1}{(n-1)} \right\}^{1/2} \left\{ \frac{S_i}{i} - \frac{(S_n - S_i)}{(n-1)} \right\} \tag{20}$$

La hipótesis nula de no cambio es rechazada si el estadístico excede el α th cuantil de la distribución nula de Z_B Esta distribución puede ser obtenida mediante simulación.

El método CBS propone una modificación. Si se considera que el segmento puede unirse mediante sus extremos formando un círculo. El likelihood test stadistico para testar la hipótesis nula de que el arco formado de i+1 a j y su complementario tienen diferentes medias:

$$Z_{ij} = \left\{ \frac{1}{(j-1)} + \frac{1}{n-j+i} \right\}^{(-1/2)} \left\{ \frac{(S_j - S_i)}{j-1} - \frac{S_n - S_j + S_i}{n-j+i} \right\} \tag{21}$$

Esta modificación permite que el estadístico de test sea $Z_C = max_{1 \leq i < j \leq n} |Z_{ij}|$.

De igual modo se considera que existe un cambio si Z_C supera un cierto umbral. Si X_i se distribuyen normalmente este umbral puede ser obtenido mediante simulación.

Bajo la hipótesis nula de no presencia de cambios X_i están identicamente distribuídas. Así se puede generar una distribución de referencia mediante permutación. Dónde $X_i^*, ..., X_n^*$ es la permutación y $Z_C^* = max|Z_{ij}^*|$ el estadístico resultante. El umbral se define como el percentil 1-α de la distribución de Z_C^*.

El método CBS presenta dos modificaciones adicionales:

1. Suavizado o *smoothig* antes de la segmentación

2. Corregir regiones con tendencias

El método de suavizado permite minimizar el efecto que puedan producir sobre la estimación del número de copias la presenica de datos atípicos procedentes de artefactos de la imagen o por otro tipo de errores técnicos. Este tipo de errores se producen como un sólo punto. La región de suavizado para cada i se define como $i-R, ..., i, i+R$ dónde R es un entero pequeño (de 2 a 5). Si m_i es la mediana de los datos y σ la desviación estándar de todos los datos. Si dado un j que dónde de X_i a X_j se excede de $L * \sigma$ entonces se reemplaza X_i con $m_i + (X_i - X_j) * M * \sigma$.

La segunda modificación que se introduce permite eliminar tendencias locales que pueden presentar los datos por motivos desconocidos que no son indicativos de cambios en el número de copias. Después de la aplicación del método, se calcula la suma de las desviaciones al cuadrado de cada punto del segmento respecto la media del segmento; $SS(C)$. Después se computan todas las sumas de cuadrados ordenadas tales que satisfacen $SS(1), ..., SS(C-1)$. Se identifican como alterados el conjunto de segmentos $(SS(c'))$ que cumplen $c' = min\left\{c : \left[\frac{SS(c)}{SS(C)}\right] < \gamma\right\}$.

Índice de tablas

Índice de figuras