Universidad de Alcalá

D. ADRIANO JIMÉNEZ ESCRIG, DOCTOR EN MEDICINA Y PROFESOR ASOCIADO DEL DEPARTAMENTO DE MEDICINA DE LA UNIVERSIDAD DE ALCALÁ

CERTIFICA:

Que ISABEL GOBERNADO FERRANDO, Licenciada en Medicina y Especialista en Psiquiatría, ha realizado bajo mi dirección el estudio titulado: "SECUENCIACION DE EXOMA COMPLETO EN TRASTORNO BIPOLAR AUTOSOMAL DOMINANTE: AFECTACIÓN DEL GEN PERIOD3-RITMO CIRCADIANO".

Este trabajo reúne los requisitos necesarios para ser defendido como Tesis Doctoral,

Para que así conste, firmo el presente certificado en Madrid a 22 de abril de 2013.

Fdo. Prof. Jiménez-Escrig

UNIVERSIDAD DE ALCALÁ FACULTAD DE MEDICINA HOSPITAL UNIVERSITARIO RAMÓN Y CAJAL

Universidad de Alcalá

Crta. Colmenar, km. 9,100
28034 Madrid
Teléfonos: 91 3368550 – 91 3368
Fax: 91 3368545
Jeronimo.saiz@uahes

PABELLÓN DOCENTE
HOSPITAL RAMÓN Y CAJAL

**Prof. Jerónimo Saiz Ruiz, Catedrático de Psiquiatría del Departamento de Medicina y Especialidades Médicas de la Facultad de Medicina y CC. De la Salud de la Universidad de Alcalá**

**CERTIFICA:** Que **Dña. Isabel Gobernado Ferrando**, ha realizado la Tesis Doctoral titulada:

**"SECUENCIACIÓN DE EXOMA COMPLETO EN TRASTORNO BIPOLAR AUTOSÓMICO DOMINANTE: AFECTACIÓN DEL GEN PERIOD3-RITMO CIRCADIANO"**

bajo mi codirección, considerando que reúne los méritos de calidad, originalidad y rigor científico suficientes para optar al grado de Doctor.

Y para que conste y surta a los efectos oportunos, expido el presente certificado en Madrid a veintiuno de junio de dos mil trece.

Jerónimo Saiz Ruiz
Codirector de la Tesis Doctoral

UNIVERSIDAD DE ALCALÁ
FACULTAD DE MEDICINA
HOSPITAL UNIVERSITARIO
RAMÓN Y CAJAL

Universidad de Alcalá

DEPARTAMENTO DE MEDICINA
Y ESPECIALIDADES MÉDICAS
Facultad de Medicina
Campus Universitario
28805 Alcalá de Henares (Madrid)
Teléfono: 918854536–4790 / Fax: 918856455-4594
e.mail: dpto.especialidades@uah.es

D. Agustín Albillos Martínez, Catedrático de la Universidad de Alcalá y Director en funciones del Departamento de Medicina y Especialidades Médicas

**INFORMA**

Que la Tesis Doctoral titulada *"SECUENCIACIÓN DE EXOMA COMPLETO EN TRASTORNO BIPOLAR AUTOSÓMICO DOMINANTE: AFECTACIÓN DEL GEN PERIOD3-RITMO CIRCADIANO"*, presentada por Dª Isabel Gobernado Ferrando, cumple con todos los requisitos científicos y metodológicos para ser defendida ante un Tribunal.

Alcalá de Henares, 6 de junio de 2013

EL DIRECTOR DEL DEPARTAMENTO


Agustín Albillos Martínez

**UNIVERSIDAD DE ALCALÁ**

**FACULTAD DE MEDICINA**

**PROGRAMA DE DOCTORADO: PSIQUIATRÍA Y PSICOLOGÍA MÉDICA**

**TESIS DOCTORAL**

# SECUENCIACIÓN DE EXOMA COMPLETO EN TRASTORNO BIPOLAR AUTOSÓMICO DOMINANTE: AFECTACIÓN DEL GEN PERIOD3-RITMO CIRCADIANO.

ISABEL GOBERNADO FERRANDO

**DIRECTOR: Prof. D. Adriano Jiménez Escrig.**
**CO-DIRECTOR: Prof. D. Jerónimo Sáiz Ruiz.**

*Con mi más sincero agradecimiento*

*para todos los que han hecho posible este trabajo.*

# ÍNDICE

# ÍNDICE DE GRÁFICOS

# 1. JUSTIFICACIÓN E HIPÓTESIS DE TRABAJO.

En los últimos años estamos asistiendo a un importante crecimiento en el conocimiento del genoma humano y su papel en la génesis y el desarrollo de las enfermedades. Tras el rápido avance inicial, con los descubrimientos de los genes implicados en las enfermedades de herencia mendeliana, los investigadores se encontraron con el escollo de las enfermedades de herencia compleja. Tras distintos intentos de descifrar su base genética, actualmente la investigación en este campo se centra en dos hipótesis de trabajo:

-Hipótesis de enfermedad común-variaciones comunes. Postula que las enfermedades complejas estarían producidas por el efecto acumulativo de múltiples variantes genéticas comunes en la población, cada una de ellas con un efecto pequeño.

-Hipótesis de enfermedad común-múltiples variantes raras. Defiende que estas enfermedades estarían causadas por variaciones genéticas poco frecuentes pero con mucha penetrancia. Estas variaciones podrían ser mutaciones puntuales, variaciones en el número de copia (*copy number variants* o CNVs), o inserciones o delecciones de mayor longitud.

Es probable que ambas hipótesis sean ciertas, contribuyendo las dos en cierta medida a constituir la realidad genética de las enfermedades complejas (figura 1-1).

Para testar la primera hipótesis se desarrollaron los estudios de asociación de genoma completo (*genomewide association studies* o GWAS), que permiten rastrear varios miles o millones de variaciones comunes del genoma por individuo en forma de SNPs (polimorfismos de nucleótido único) con el fin de establecer asociaciones entre estas variantes frecuentes y fenotipos concretos. Para la segunda se utilizan los estudios de familias y ligamiento o, más recientemente, la secuenciación genómica. El problema de los primeros es el escaso número de individuos afectos de los que se dispone habitualmente, lo que hace que adolezcan de escasa potencia. La secuenciación genómica, por el contrario, identifica todas las variaciones presentes en el genoma del individuo, permitiendo relacionarlas con posibles enfermedades sin necesidad de hipótesis previas o muestras amplias. El problema de esta técnica han sido sus limitaciones por coste y duración, lo que había restringido su uso a estudios aislados en los que se secuenciaban áreas concretas y pequeñas del genoma. Estas áreas seleccionadas están habitualmente guiadas por los estudios de ligamiento, lo que introducía factores de confusión y le restaban eficiencia. Esta situación ha cambiado con el desarrollo de las técnicas actuales de secuenciación masiva paralela o "de nueva generación", que permiten la secuenciación del ADN a gran velocidad y con un coste progresivamente menor (Harismendy et al., 2009; Metzker, 2010).

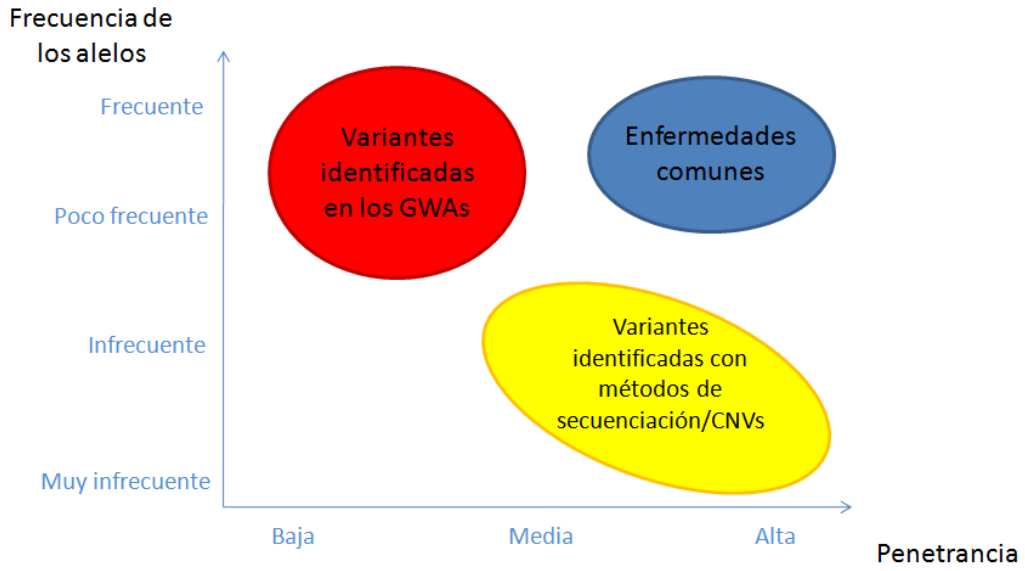



**Figura 1-1.** Modificado de Owen y cols., 2009(Owen, Williams, & O'Donovan, 2009).

Como el coste actual de la secuenciación de un genoma humano completo todavía es elevado, se han planteado estrategias para minimizarlo sin perder eficiencia. Ng y cols. (Ng et al., 2009) publicaron en 2009 el primer estudio en el que secuenciaban exclusivamente el exoma (parte codificante del genoma), unas 30 megabases (el 1% del total del genoma). Ya que la inmensa mayoría de las mutaciones relacionadas con las enfermedades conocidas actualmente se encuentran en esta parte codificante, este acercamiento supone un importante ahorro de tiempo y, principalmente, económico. Presenta además otra ventaja. Hay que tener en cuenta que en la secuenciación del exoma se encuentran unas 15000-25000 variaciones por individuo, dependiendo de la definición de exoma y el origen (los individuos de ascendencia africana tienen típicamente más variaciones que los de ascendencia europea, por ejemplo), mientras que en el genoma completo hablamos de unos 4 millones de variaciones por individuo (Ng et al., 2009). Dado que existe escaso consenso sobre cómo interpretar las variaciones halladas en las secuencias no codificantes, la secuenciación exómica limita mucho las incógnitas a la hora de interpretar los resultados.

De forma simplificada, la secuenciación del exoma completo consiste en identificar, seleccionar y enriquecer las secuencias que corresponden al exoma, desechando el resto del material genético, y después secuenciarlas con las técnicas de segunda generación.

Hay numerosos artículos publicados que han utilizado la secuenciación de exoma completo para encontrar las mutaciones responsables de distintas enfermedades (Ng et al., 2009; Choi et al., 2009; Robinson, 2010; Johnson et al., 2010; Bolze et al., 2010; Wang et al., 2010; Haack et al., 2010; Bonnefond et al., 2010; Musunuru et al., 2010; Johnson, Gibbs, Van, Houlden, & Singleton, 2010; Gilissen et al., 2010; Bilguvar et al., 2010; Rios, Stein, Shendure, Hobbs, & Cohen, 2010; Ng et al., 2010; Lalonde et al., 2010; Wu et al., 2011; Min et al., 2011; Clayton-Smith et al., 2011; Benitez et al., 2011; Vissers et al., 2011; Pierson et al., 2011; Theis et al., 2011; Lam, Guo, Wilson, Kohl, & Wong, 2011; Doi et al., 2011; Ozgul et al., 2011; Weedon et al., 2011; Watkins et al., 2011; Raffan et al., 2011; Wang et al., 2011a; Takata et al., 2011a;

Chen et al., 2011b; Wang et al., 2011b; Jimenez-Escrig, Gobernado, Garcia-Villanueva, & Sanchez-Herranz, 2012; Daoud et al., 2012; Marti-Masso et al., 2012). Ha resultado especialmente útil en el descubrimiento de mutaciones *de novo* y aquellas presentes en familias con escaso número de afectos en las que resultaba inviable la utilización de otro tipo de estudios. Hasta la fecha, en el campo de la psiquiatría se ha aplicado en la búsqueda de mutaciones *de novo* en el Síndrome de Tourette, la esquizofrenia y el autismo (Xu et al., 2011; O'Roak et al., 2011; Sundaram et al., 2011; Sanders et al., 2012), con hallazgos pendientes de replicación.

**Posibles limitaciones.**

Para aplicar la secuenciación de exoma completo en la búsqueda de mutaciones potencialmente patógenas hay que aceptar como ciertas las siguientes asunciones (Ng et al., 2009):

- La mutación buscada se encuentra en áreas codificantes del genoma.

Si estuviera en un área intrónica no sería posible encontrarla, ya que no la estaríamos secuenciando. Sabemos que hay áreas reguladoras, cuyas mutaciones podrían estar relacionadas con enfermedades, que se pueden encontrar a muchas kilobases de distancia del gen, fuera de las áreas exómicas (Birney et al., 2007; Robinson, Krawitz, & Mundlos, 2011b). Es más, hay que tener en cuenta que los *loci* que constituyen el exoma no están del todo bien definidos. Habitualmente se utilizaban para crear las bibliotecas todas las secuencias incluidas en la base de datos CCDS (*Consensus Coding Sequence*). Posteriormente se comprobó que no era completa, y más recientemente se están utilizando otras bases que incluyen más fragmentos, como la RefSeq, la Ensembl, el set GENCODE o los genes de la UCSC browser (Coffey et al., 2011; Kuhn, Haussler, & Kent, 2012; Pruitt, Tatusova, Brown, & Maglott, 2012; Flicek et al., 2012).

- Una mutación es suficiente para causar la enfermedad.

Se asume que la responsable del fenotipo de interés es una mutación única, común a todos los afectos.

- La mutación buscada tiene una alta penetrancia y es poco frecuente.

Esto supone que los portadores de la mutación presentarán el fenotipo de interés en mayor o menor grado, y que la mutación no está presente o lo está en muy baja frecuencia en la población general.

Otras limitaciones están relacionadas con las propias técnicas empleadas. La captura de las áreas de interés, independientemente de la técnica que se aplique, no es perfecta, quedando aproximadamente un 8% de secuencias exómicas sin capturar. Algunas mutaciones, como las variaciones por número de tripletes, no las identifica y con otras, como las variaciones de número de copia, puede presentar problemas (Singleton, 2011; Biesecker, Shianna, & Mullikin, 2011). Por otro lado, la propia secuenciación puede introducir errores de lectura, especialmente en áreas complicadas o de baja cobertura, lo que podría dificultar el hallazgo de resultados.

Finalmente, también se pueden cometer errores debidos al alto número de variantes encontradas, asignando la enfermedad a una variante no causante o filtrando en exceso y obviando la mutación causal.

**Hipótesis de trabajo.**

En el caso del trastorno bipolar (TBP) los estudios epidemiológicos sugieren un claro componente genético. No obstante, hasta el momento actual no se ha conseguido describir ninguna mutación o conjunto de variaciones que se relacionen de forma inequívoca con dicho trastorno. Se han dado múltiples explicaciones para ello: el escaso tamaño y la heterogeneidad de las muestras utilizadas en los estudios, derivada esta última de los rudimentarios métodos diagnósticos actuales, la necesidad de contar con hipótesis etiopatogénicas de las que carecemos, etc.

Basándonos en los hallazgos de los estudios de asociación de genoma completo (GWAS), que encuentran en el TBP una mayor cantidad de áreas de asociación en los exomas (Lehne, Lewis, & Schlitt, 2011; Smith et al., 2011a), dado que parte de los problemas antes mencionados quedan resueltos con la secuenciación de exoma completo, y pese a las limitaciones anteriormente expuestas, planteamos la hipótesis de que esta técnica podría ser útil para ayudar a aclarar la genética que subyace al trastorno bipolar.

Partiendo de la hipótesis de enfermedad común-variantes raras con alta penetrancia, se pretende secuenciar a individuos de una familia con varios miembros afectos de TBP con el objetivo de encontrar la variación genética común origen del trastorno.

# 2. GENÉTICA DEL TRASTORNO BIPOLAR

Descrito por Kraepelin hace más de un siglo, el trastorno bipolar (TBP) es una enfermedad caracterizada por episodios de depresión y de manía, intercalados habitualmente por periodos intermedios de ánimo normal (figura 2-1). Los episodios depresivos se caracterizan por la presencia de ánimo depresivo durante al menos dos semanas, junto con otros síntomas como disminución de intereses o capacidad de disfrutar, sensación de falta de energía o alteraciones del sueño y el apetito. Estos síntomas producen en el paciente malestar o deterioro significativo. La manía se caracteriza por ánimo extremadamente elevado, con sensación de mayor energía, pautas inhabituales de pensamiento y conducta y, en ocasiones, psicosis (American Psychiatric Association, 2000).

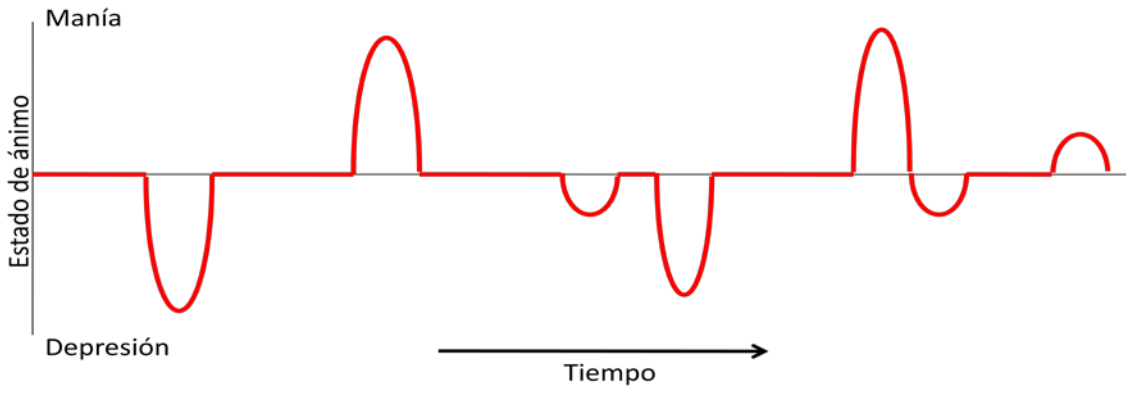



**Figura 2-1**. Ejemplo de la evolución del estado de ánimo en un paciente con TBP.

El trastorno bipolar tiene una prevalencia a lo largo de la vida de 1 por cada 100 individuos (Wyatt & Henter, 1995; Craddock & Jones, 1999; Merikangas & Low, 2004; Merikangas et al., 2007), aunque algunos autores lo cifran en el 4% o hasta el 5% utilizando fenotipos atenuados o el llamado "espectro bipolar" (Akiskal et al., 2000; Kessler et al., 2005). La enfermedad se caracteriza por elevadas tasas de recurrencia, ocasionales síntomas residuales, deterioro cognitivo y empeoramiento de la calidad de vida (Martinowich, Schloesser, & Manji, 2009), y supone un importante coste económico (Kessler et al., 2006).

## A. Evidencia de una base genética. Estudios de familias, gemelos y adopción.

La epidemiología clínica permite valorar cómo se distribuyen las enfermedades en la población. Los estudios de agregación familiar permiten determinar si una patología es más frecuente entre los individuos de una misma familia que entre la población general, suponiendo un punto inicial para determinar si tiene o no una etiología genética. La presencia de agregación familiar no constituye una evidencia definitiva, ya que los individuos de una misma familia pueden estar expuestos a un mismo factor ambiental. Ayudan a clarificar esto último los estudios de adopción y de gemelos. Los estudios de gemelos examinan el grado de concordancia respecto a la aparición de una determinada enfermedad en gemelos monocigóticos (con igual material genético) comparados con los gemelos dicigóticos (diferente material genético, similares factores ambientales) (Fañañas Saura & Sáiz Ruiz, 2000).

Los estudios de familias y gemelos muestran una evidencia clara de un componente genético en el TBP, siendo una de las enfermedades psiquiátricas con mayor heredabilidad y la que comporta mayor riesgo relativo genético para los familiares de primer grado (Shih, Belmonte, & Zandi, 2004; Hales RE, Yudofsky SC, & Gabbard GO, 2008). La concordancia entre gemelos monocigóticos varía según los estudios, situándose entre un 40 y un 70%, y llegando al 93% en algún estudio. Entre

gemelos dicigóticos y familiares de primer grado las cifras también son variables, situándose entre el 5% y el 60% según los estudios. También es mayor el riesgo de otros trastornos psiquiátricos en los familiares de los probandos, principalmente depresión unipolar (Bertelsen, Harvald, & Hauge, 1977; Mendlewicz & Rainer, 1977; Kendler, Pedersen, Neale, & Mathe, 1995; Cardno et al., 1999; Merikangas et al., 2002; Smoller & Finn, 2003; Kieseppa, Partonen, Haukka, Kaprio, & Lonnqvist, 2004).

## B. En busca del gen mutado. Estudios de ligamiento y asociación. Genes candidatos.

Una vez se determina que una enfermedad tiene un origen genético, el siguiente paso es determinar cómo se hereda y cuál es el gen o genes implicados. Para encontrar estos genes causantes, se han llevado a cabo numerosos estudios de ligamiento y asociación.

Los **estudios de ligamiento** se basan en el hecho de que los *loci* genéticos situados a escasa distancia se segregan juntos, es decir, permanecen juntos tras la meiosis. Si un marcador genético cuya ubicación es conocida se hereda junto con una enfermedad determinada, se puede postular que el gen que produce la enfermedad se encuentra en las proximidades de este marcador. Así, los estudios de ligamiento describen zonas del genoma en las que podría existir un gen relacionado con la enfermedad en cuestión. Posteriormente, se analizan estas zonas del genoma en busca de genes candidatos. Resultan muy útiles en la búsqueda de genes causantes de enfermedades monogénicas o de herencia mendeliana, pero dan poca información si la herencia es compleja (Fañañas Saura et al., 2000).

Se han realizado múltiples estudios de ligamiento en el TBP, describiéndose resultados positivos en casi todos los cromosomas (Hayden & Nurnberger, Jr., 2006). Entre los hallazgos más replicados están la traslocación balanceada en DISC1 descrita en 1990 que se relaciona con esquizofrenia y trastorno bipolar (St et al., 1990; Marx,

2007; Chubb, Bradshaw, Soares, Porteous, & Millar, 2008), la delección en el cromosoma 22q11 relacionada con psicosis en el contexto del síndrome velo-cardio-facial (Papolos et al., 1996; Murphy, 2002), o la delección del cromosoma 15 relacionada con sintomatología psiquiátrica y el síndrome de Prader-Willi (Soni et al., 2008).

Los meta-análisis realizados señalan como áreas de ligamiento las áreas cromosómicas 6q21-q25 y 8q24 (McQueen et al., 2005) y 13q y 22q (Badner & Gershon, 2002). Segurado en su meta-análisis (Segurado et al., 2003) no encuentra ligamiento estadísticamente significativo en ningún *loci*, siendo las áreas de mayor asociación 9q22.3-21.1, 10q11.21-22.1, 14q24.1-32.12 y zonas del cromosoma 18.

Los **estudios de asociación** analizan si determinados alelos son más frecuentes en el grupo de afectos que en el de sanos (estudios caso-control). Se utilizan generalmente como marcadores genéticos los SNPs. Estas variaciones pueden ser en sí mismas funcionales y estar relacionadas con la fisiopatología de la enfermedad, pero en la mayoría de los casos son utilizadas sólo para localizar los verdaderos sitios relevantes. Los estudios de asociación primitivos analizaban uno o varios genes candidatos señalados por datos indirectos (evidencia posicional de los estudios de ligamiento o desde hipótesis neurobiológicas). Se describieron múltiples variaciones que podrían estar relacionadas con el TBP. No obstante, los resultados de estos estudios deben de ser tomados con cautela debido a sus múltiples debilidades (Chanock et al., 2007; Zollner & Pritchard, 2007; Sullivan, 2007).

Basándose en los resultados de los estudios descritos previamente se han propuesto decenas de **genes candidatos**, sin que se haya encontrado ningún resultado que haya sido replicado de forma definitiva (Perez de Castro et al., 1995; Jones & Craddock, 2001; Nievergelt et al., 2006; Escamilla & Zavala, 2008; Serretti & Mandelli, 2008; Martinowich et al., 2009; Craddock & Sklar, 2009; Barnett & Smoller, 2009; Le-Niculescu et al., 2009; Etain, Milhiet, Bellivier, & Leboyer, 2011; Nurnberger, Jr., 2012). En negrita se señalan los que actualmente tienen mayor nivel de evidencia:

- Genes relacionados con la transmisión monoaminérgica: DRD1, DRD2, DRD4, DAT1, SLC6A3, HTTLPR, HTR2A, **SLC6A4**, GRIA1, GRIN2B, GRM1, GRM3, GRM4, GRIK1, GRIK4, CHMA7, **COMT**, **TPH-2**, **DAOA (G72/G30)**, **MAOA**, OPRM1, NOS1, SYN3.
- Genes relacionados con los ritmos circadianos: CLOCK, BMAL1, PERIOD3, ARNTL, TIMELESS, RORA, RORB, RXRG.
- Genes relacionados con el neurodesarrollo y neurotropismo: **DISC1**, **BDNF**, NCAM1, PPARD, NRG1, FGF12, PTN, ELAVL2, NAV2, NBP, MIT1l, Olig2, QKI.
- Genes relacionados con la adhesión y transducción: ANK2, APP, CD44, CDH13, CLSTN2, EPHA5, NCAM1, PLXNA2, SYNE1, CELSR1, GNA12, PARD3, IMPA2, ADCY1, PTPRT, PIK3R1, PRKCE, PDE10A, **GRK3** (modula la expresión de receptores acoplados a proteínas G).
- Genes relacionados con el ciclo celular: A2BP1, ATXN1, **GSK3b**, STK24.
- Genes relacionados con canales iónicos: DPP10, KCNAB1, KCNB1, KCND2, KCNK1, **CACNAC1A**, CACNB2, CAMK2A, DCAMKl1/DCLK1, RYR3, SLC8A1, TRPM3.
- Otros: **P2RX7** (codifica para una ATPasa calcio-dependiente), **FKBP5** (modula receptores de corticoides), SOD1 (superóxido-dismutasa 1), HMOX1, Ak3l1 y ACACB (función mitocondrial), CREBBP, FOXP1, NF1B, NR3C1, RXRG, ZHX2.

## C. Enfermedades de herencia compleja. Perspectivas actuales.

El TBP se considera actualmente una enfermedad de herencia compleja. Se denomina herencia compleja a aquella que se escapa del modelo mutación-enfermedad de la herencia mendeliana. Como se describió en el capítulo 1, desde la entrada del siglo XXI el estudio de la genética de estas enfermedades se orienta desde de dos hipótesis distintas de trabajo:

-Enfermedad común-variaciones comunes. Postula que las enfermedades complejas estarían producidas por el efecto acumulativo de múltiples variantes genéticas comunes en la población, cada una de ellas con un efecto pequeño.

-Enfermedad común-múltiples variantes raras. Defiende que estas enfermedades estarían causadas por variaciones genéticas poco frecuentes pero con mucha penetrancia.

Ambas hipótesis podrían ser ciertas, explicando la segunda los casos familiares y algunos esporádicos debidos a mutaciones *de novo*. Qué genes concretos están implicados, cuáles son las influencias ambientales y de qué manera actúan sobre la expresión genética y el desarrollo de la enfermedad es todavía desconocido.

MODELO ENFERMEDAD COMÚN-VARIACIONES COMUNES.

Craddock (Craddock, Khodel, Van, & Reich, 1995) propuso en 1995 su modelo matemático, sugiriendo que la herencia del TBP se explica mejor con un modelo de varios genes de susceptibilidad que interaccionan entre ellos por fenómenos de epistasis, apoyando el modelo enfermedad común-variaciones comunes. Para el estudio de estas variaciones en grandes grupos de población se desarrollaron en 2006 los estudios de asociación genómica amplia (GWAS en inglés).

La técnica de **asociación genómica amplia (*wide genome association o GWAS*)** puede rastrear actualmente unos 2,5 millones de SNPs a lo largo de todo el genoma con microarrays de genotipación (microarrays por sondas que reconocen áreas concretas del ADN), con el fin de establecer posibles sitios de asociación con el fenotipo de estudio. Se supone que si un número lo suficientemente grande de casos se compara con controles adecuados con una densidad suficiente de marcadores genéticos (típicamente SNPs), los alelos que confieren riesgo para la enfermedad deberían poder detectarse como una desviación de su frecuencia frente a la que se observa en los controles. Estos estudios conllevan un error inherente a su altísimo número de hipótesis (cada uno de los SNPs), por lo que exigen para ser significativa una $p<5x10e-8$, basándose en el error estadístico tipo I (Nurnberger, Jr., 2012). Se exige además su replicación en una población independiente.

Los GWAS tienen las ventajas del diseño de los estudios de asociación (que tienen la potencia suficiente para detectar efectos pequeños) y no precisan de unas hipótesis

iniciales de trabajo o de un conocimiento de la fisiopatología de la enfermedad. Su relativa velocidad y coste permite realizar estas búsquedas en muestras muy amplias de individuos. En cualquier caso, no se debe olvidar que la asociación estadística no muestra causalidad. Se necesita posteriormente identificar las mutaciones, caracterizarlas y realizar experimentación biológica y estudios clínicos.

Se han llevado a cabo numerosos GWAS y meta-análisis en pacientes con TBP con hallazgos interesantes (Wellcome Trust Case Control Consortium, 2007; Sklar et al., 2008; Ferreira et al., 2008; Baum et al., 2008a; Baum et al., 2008b; Scott et al., 2009; Hattori et al., 2009; Smith et al., 2009; Oedegaard et al., 2010; Djurovic et al., 2010; Yosifova et al., 2011; Lee et al., 2011; Cichon et al., 2011; Chen et al., 2011a; Smith et al., 2011b; Vassos et al., 2012), aunque la falta de consistencia en la literatura publicada es la norma (Seifuddin et al., 2012). Los genes encontrados con las variantes de riesgo más destacadas se recogen en la tabla 2-1. Los *odds ratio* (OR) determinados para cada alelo de riesgo son del orden de 1.2-1.4, lo que sugiere la existencia de múltiples alelos de bajo riesgo en la génesis de la enfermedad. Es de destacar que ningún gen señalado por estudios previos de ligamiento o genes candidatos aparece en estos GWAS (Barnett et al., 2009).

**Tabla 2-1. Genes candidatos más destacados hallados en los GWAS.**

| Gen candidato | Cr | Descripción | Referencias |
|---|---|---|---|
| **ANK3** | 10 | Gen de la ankirina-G, de la familia de proteínas que une las proteínas de membrana al citoesqueleto. Entres sus funciones: movilidad, proliferación y contacto celular, mantenimiento de los dominios especializados de membrana y regulación de canales de sodio. Su actividad podría tener relación con la modulación de las conexiones límbico-frontales. | (Ferreira et al., 2008; Smith et al., 2009; Schulze et al., 2009; Ripke et al., 2011; Takata et al., 2011b; Linke et al., 2012b) |
| **CACNA1C** | 12 | Gen de la subunidad a-1C de los canales de calcio voltaje-dependientes tipo L. Parece tener relación con la función de los circuitos fronto-subcorticales. | (Ferreira et al., 2008; Bigos et al., 2010; Wang, McIntosh, He, Gelernter, & Blumberg, 2011; Ripke et al., 2011; Jogia et al., 2011; Perrier et al., 2011) |
| **DGKH** | 13 | Gen de la Diacil-glicerol kinasa. Esta molécula participa en el sistema del fosfatidil-inositol, lugar de acción del litio. | (Baum et al., 2008a) |
| **NCAN** | 19 | Neurocan. Proteoglucano implicado en la adhesión y migración celular. Su inhibición en modelos animales origina comportamientos mania-*like*. | (Ripke et al., 2011; Cichon et al., 2011; Miro et al., 2012) |

Cr- Cromosoma.

MODELO ENFERMEDAD COMÚN-MÚLTIPLES VARIANTES RARAS.

Redon, en 2006, describe que en el genoma humano hay alrededor de 1400 polimorfismos o variantes largas, que denomina **variaciones en el número de copia (*Copy Number Variants* o CNVs**), especialmente frecuentes en genes que tienen que ver con la neurofisiología (Redon et al., 2006). Las CNVs son delecciones, inserciones, duplicaciones y otras anomalías citogenéticas demasiado cortas para ser vistas a través de un microscopio pero que pueden incluir desde 1kb hasta cientos o miles de pares de bases. Estas CNVs modifican la expresión genética, alterando los genes o el número de copias de éstos. En 2008 se encuentra que estas CNVs son especialmente frecuentes en patologías que tienen relación con el neurodesarrollo, como la esquizofrenia, el retraso mental o el autismo (Walsh et al., 2008; Girirajan, Campbell, & Eichler, 2011). Pueden ser heredadas o *de novo*, estas últimas de aparición con frecuencia significativa en patologías de inicio en la infancia.

Los resultados en pacientes con TBP son variables, existiendo estudios que encuentran que las CNVs son más frecuentes en el grupo de pacientes que en el de controles, especialmente en aquellos pacientes con edad de aparición más precoz (Zhang et al., 2009; Malhotra et al., 2011; Priebe et al., 2012), y otros estudios que no encuentran diferencias entre ambos grupos (Grozeva et al., 2010; Olsen et al., 2011; Bergen et al., 2012).

Para dar significado biológico a estas CNVs es necesario caracterizarlas y relacionarlas con *loci* y genes específicos. Hasta el momento algunos estudios han implicado al gen GSK3beta (*glycogen synthase kinase 3 beta*) (Lachman et al., 2007) y genes relacionados con la neurotransmisión glutamatérgica (Wilson et al., 2006).

En relación con la disminución de costes asociada a los secuenciadores de nueva generación, se están realizando de forma cada vez más frecuente **estudios de secuenciación a gran escala**, de todo el genoma o sólo del exoma (descrito en el capítulo 1). Esta técnica es especialmente útil en la detección de variantes raras (presentes en menos del 1% de la población), ya sean mutaciones de nucleótido único, CNVs u otros (Alkan et al., 2009). Proyectos internacionales de secuenciación a gran

escala para aportar una base de datos de variantes de referencia, como el Proyecto de los 1000 genomas, están ya avanzados.

Hasta el momento actual no hay ningún estudio publicado que aplique esta técnica en pacientes con TBP.

**D. <u>Limitaciones de los estudios genéticos en el TBP; futuras perspectivas:</u>** (Escamilla et al., 2008; Porteous, 2008).

Pese a los esfuerzos invertidos en descifrar las bases genéticas del TBP, siguen sin encontrarse hallazgos claros, y la mayoría de ellos no son replicados en otros estudios. Se han propuesto distintos motivos para explicar esta "herencia perdida" además de diversas soluciones, que se resumen en la tabla 2-3.

LIMITACIONES RELACIONADAS CON LA SELECCIÓN DE LA MUESTRA.

**Tamaño muestral.**

En la mayoría de los estudios de los que se dispone las muestras son pequeñas, por lo que adolecen de escasa potencia. Además, para detectar efectos pequeños (como las variantes que se buscan con los GWAS) las muestras deben ser aún mayores. Se calcula que para alcanzar significación estadística con *odds ratio* (OR) del orden de 1.2-1.4 se necesitan muestras de entre 10000 y 20000 sujetos (Cichon et al., 2009).

Para solucionar este problema, se han puesto en marcha estudios multicéntricos y se realizan meta-análisis. Una de las iniciativas más importantes actualmente es la *National Institute of Mental Health Bipolar Genetics Initiative*, en la que colaboran diversas Universidades estadounidenses para poner en común sus muestras y resultados. El *Psychiatric GWAS Consortium* es también un ejemplo de colaboración para combinar las muestras de pacientes de diversos grupos investigadores. Igualmente se están creando bases de datos genéticas sin ánimo de lucro a las que pueden tener acceso los investigadores en la materia.

**Falta de homogeneidad.**

Existe una importante dificultad para realizar una buena definición de los casos, resultando en la utilización de muestras muy heterogéneas, lo que dificulta el hallazgo

de significación estadística en estudios genéticos. Este problema se debe principalmente a la clasificación actual de las enfermedades mentales (DSM IV TR o CIE10), que definen los trastornos por criterios puramente clínicos o basándose en la tradición nosológica, consensos de expertos y utilidad clínica, y no en etiología o neurofisiología. Numerosos estudios muestran que entre los familiares de pacientes no sólo es más frecuente el TBP, sino también otros diagnósticos psiquiátricos (como esquizofrenia, trastorno esquizoafectivo o depresión unipolar), mostrando una susceptibilidad genética compartida entre estos trastornos (lo que está también sustentado por estudios genéticos) y demostrando que los conceptos diagnósticos actuales no tienen relación directa con las causas genéticas de la enfermedad (Craddock, O'Donovan, & Owen, 2005). Igualmente, no es posible diferenciar entre trastornos primarios, con más probabilidad relacionados con factores genéticos, o secundarios.

Se han propuesto distintas soluciones:

**Segregación por subtipos** en los análisis. Los pacientes se agrupan en distintos subtipos, entre los que se han propuesto:

-Subtipos de TBP: tipo I, tipo II, trastorno esquizoafectivo, ciclotimia.

-Comorbilidades (Kerner, Lambert, & Muthen, 2011): trastorno por déficit de atención e hiperactividad (TDAH), trastorno de pánico (Barnett et al., 2009), alcoholismo (Lydall et al., 2011).

-Curso de la enfermedad: edad de inicio, polaridad del primer episodio, frecuencia de episodios, presencia de psicosis, suicidio.

-Otras: respuesta a litio (Grof, Duffy, Alda, & Hajek, 2009).

**Utilización de endofenotipos.**

En los trastornos de herencia compleja el fenotipo es resultado de la combinación de varios genes con distintas variables del entorno. En un intento de homogeneizar las muestras utilizadas en los estudios de genética, en los años 80 se definieron los endofenotipos. Un endofenotipo es un fenotipo intermedio e interno que llena el hueco en la cadena causal entre los genes y la enfermedad. Puede ser una característica neurofisiológica, neuropsicológica, cognitiva, neuroanatómica, bioquímica o endocrinológica (Gottesman & Shields, 1973) (figura 2-2).

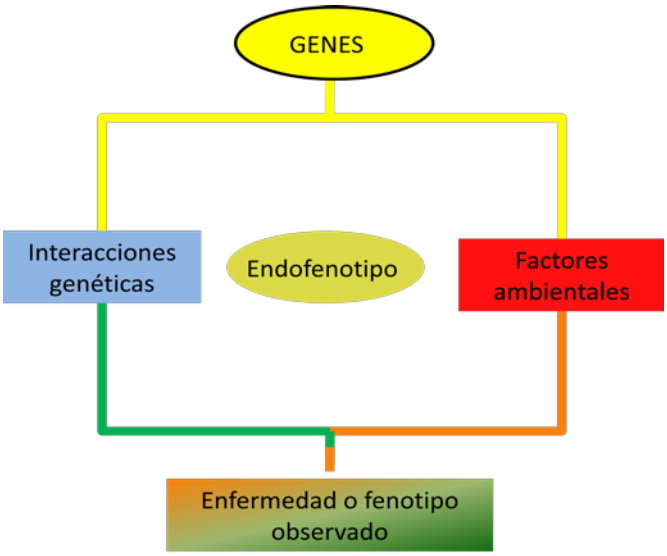



**Figura 2-2.** El endofenotipo representa un fenotipo intermedio entre la expresión de los genes y el resultado final o fenotipo observado.

Los endofenotipos que se utilizan para realizar estudios genéticos deberían cumplir los siguientes criterios (Gershon & Goldin, 1986):

1. Asociarse con la enfermedad.
2. Ser heredables.
3. Ser independientes del curso de la enfermedad.
4. Segregarse junto con la enfermedad en las familias.
5. Encontrarse en los familiares de los afectados en mayor proporción que en la población general.
6. Ser medibles.

En el TBP se han sugerido diversos endofenotipos, recogidos en la tabla 2-2. Actualmente los esfuerzos se están centrando en la búsqueda de endofenotipos a través de estudios que combinan la genética y la neuroimagen.

## LIMITACIONES TÉCNICAS

Las técnicas utilizadas en los estudios genéticos tienen sus propias limitaciones y la descripción en detalle de los problemas de cada una de ellas queda más allá del propósito de esta tesis. Como ejemplos se pueden nombrar la ausencia de protocolos

establecidos y de una metodología homogénea en muchas de ellas o la ausencia de cobertura de algunas áreas del genoma tanto con los chips utilizados en los GWAS como con las técnicas de secuenciación masiva. Las limitaciones de estas últimas técnicas se exponen en los capítulos 1 y 3.

**Tabla 2-2. Endofenotipos propuestos para el estudio del TBP.**

| | | | |
|---|---|---|---|
| Neuroimagen | Estructural | RM (resonancia magnética).<br>Disminución en el volumen de distintas áreas cerebrales: ínsula, áreas del cíngulo, amígdala, áreas frontales.<br>Presencia de hiperintensidades en la sustancia blanca, de significado incierto. | (McDonald et al., 2004; Kempton, Geddes, Ettinger, Williams, & Grasby, 2008; Arnone et al., 2009; Beyer, Young, Kuchibhatla, & Krishnan, 2009; Hallahan et al., 2011; De et al., 2012) |
| | | DTI (*Diffusion Tensor Imaging*).<br>Alteración de las fibras de sustancia blanca en distintas áreas: cíngulo anterior, cuerpo calloso, giro parahipocampal. | (Wang et al., 2009; Cui et al., 2011; Vederine, Wessa, Leboyer, & Houenou, 2011; Chen et al., 2012) |
| | Funcional | Hiperactivación de áreas del sistema límbico e hipoactivación de áreas prefrontales durante el procesamiento emocional. | (Delvecchio, Sugranyes, & Frangou, 2012; Hummer et al., 2012; Townsend et al., 2012; Garrett et al., 2012; Kim et al., 2012; Linke et al., 2012a) |
| Neuropsicología | | Alteración de mecanismos atencionales.<br>Menor rendimiento en tareas de memoria verbal.<br>Lentitud en el procesamiento de la información. | (Seidman et al., 2002; Clark, Sarna, & Goodwin, 2005; Clark, Kempton, Scarna, Grasby, & Goodwin, 2005; Klimes-Dougan, Ronsaville, Wiggs, & Martinez, 2006; Balanza-Martinez et al., 2008; Hill, Harris, Herbener, Pavuluri, & Sweeney, 2008; Kulkarni, Jain, Janardhan Reddy, Kumar, & Kandavel, 2010; Schulze et al., 2011) |
| Otros | | Respuesta a la depleción de triptófano. | (Quintin et al., 2001; Sobczak, Honig, Nicolson, & Riedel, 2002) |
| | | Respuesta a la deprivación de sueño. | (Wehr, Sack, & Rosenthal, 1987) |
| | | Supresión de melatonina por la luz. | (Nurnberger, Jr. et al., 2000) |
| | | Rasgos de personalidad/temperamento. | (Savitz & Ramesar, 2006; Savitz, van der, & Ramesar, 2008; Vazquez et al., 2008) |

En las enfermedades de herencia compleja la interpretación de los resultados no siempre es sencilla, dado que distintos mecanismos genéticos pueden funcionar como factores de confusión y hay que tenerlos en cuenta (Jimenez Escrig, 2007):

- Penetrancia variable.

El grado de penetrancia determina la probabilidad con la que un genotipo determinado se expresa fenotípicamente, es decir, la probabilidad de que un individuo que tenga el alelo que determina una enfermedad concreta la manifieste. La ausencia de penetrancia completa complica la definición de los casos, pudiendo encontrarse los alelos con la mutación de estudio en pacientes sanos.

- Fenómenos de heterogeneidad y pleoitropismo.

Distintos genes pueden producir el mismo fenotipo (heterogeneidad) y la misma mutación genética puede producir distintos fenotipos (pleiotropismo).

- Fenómenos de epistasis.

El efecto de un determinado gen está influenciado por los efectos de otros genes.

- *Imprinting*.

Mecanismo por el que un gen o grupo de genes tienen una expresión diferente si son heredados del padre o de la madre. Se debe, entre otros, a fenómenos de metilación.

- Necesidad de un determinado factor ambiental para el desarrollo de la enfermedad.

La existencia de una mutación no sería suficiente para que la enfermedad se manifestara, teniendo que estar presente además un determinado factor ambiental. Estas interacciones genes-ambiente pueden modificar la trascripción o la expresión genética por mecanismos epigenéticos y, en algunos casos, transmitirse a la descendencia.

Como posibles factores ambientales relacionados con el TBP se han descrito el estrés físico o psicosocial, determinados rasgos de personalidad, como la obsesividad y la introspección, y el consumo de tóxicos (Glassner & Haldipur, 1983; Miklowitz, Goldstein, Nuechterlein, Snyder, & Mintz, 1988; Swendsen, Hammen, Heller, & Gitlin, 1995).

- Efectos epigenéticos.

La diferente metilación y acetilación de las histonas del ADN durante el desarrollo embrionario modifican la expresión genética.

No hay estudios epigenéticos en el TBP, aunque algunos modelos animales de depresión han mostrado evidencia de que los cambios en los patrones de metilación debidos a factores ambientales modifican la expresión genética. En el modelo de rechazo social (*social defeat*), en los animales con ausencia de socialización se produce una disminución de expresión de BDNF (*Brain-Derived Neurotrophic Factor*) y CREB (*cAMP Response Element-Binding*), disminuyendo el crecimiento neuronal, efecto reversible con tratamiento antidepresivo (Tsankova et al., 2006). En el modelo de Champagne, la ausencia de cuidado maternal en ratas aumenta la metilación del promotor del receptor de estrógenos en las crías y, por tanto, disminuye la expresión de ese receptor, asociando, entre otros, una mayor reactividad al estrés. Esa alteración en la metilación se transmite a la descendencia (Champagne & Meaney, 2001; Bredy, Grant, Champagne, & Meaney, 2003; Weaver et al., 2004).

En los estudios genéticos actuales se tienen en cuenta todas estas variables a la hora de analizar los resultados, y se van desarrollando herramientas que facilitan su integración.

En el caso de los fenómenos de epistasis, para facilitar la compresión de los hallazgos de los GWAS se han propuesto el método de puntuación poligénica (*poligenic score method*) (Purcell et al., 2009) y el análisis de rutas (Wang, Li, & Bucan, 2007). En el método de puntuación poligénica se asigna una puntuación a cada alelo de riesgo y posteriormente se suma la puntuación de cada uno de los que un individuo posee para calcular su riesgo de padecer la enfermedad. El análisis de rutas parte de la premisa de que múltiples genes intervienen en la susceptibilidad a padecer una enfermedad. Estos genes pertenecen a vías comunes (rutas metabólicas, de señalización celular…) e interaccionan unos con otros por fenómenos de epistasis. Por tanto, para evaluar el riesgo que aporta cada variación genética en concreto, habría que tener en cuenta el conjunto de la ruta bioquímica. En el TBP se han realizado diversos análisis de ruta, encontrando asociaciones significativas con la regulación de la neurotransmisión dopaminérgica, genes de la vía de las cadherinas y de la vía de señalización Wnt (Torkamani, Topol, & Schork, 2008; Pandey et al., 2012).

En el caso de los estudios de secuenciación a gran escala, se han desarrollado estrategias de filtrado de las múltiples mutaciones encontradas (descritas en el capítulo 4), que tienen en cuenta los factores arriba descritos. Igualmente, los secuenciadores de nueva generación (como el *zeromode-waveguide*, actualmente en desarrollo) son capaces de leer patrones de metilación, lo que permitirá un avance rápido del conocimiento de cómo estos mecanismos influyen la expresión genética.

Para complicar algo más la situación, nuestro conocimiento del genoma es todavía limitado. No todo el genoma es accesible a las técnicas de las que disponemos actualmente, y existen amplias áreas del genoma aún sin explorar. Cada día se describen nuevos elementos reguladores, como el reciente descubrimiento de los micro-RNAs no codificantes (Thum et al., 2008). Proyectos de investigación genética como el ENCODE (*Encyclopedia of DNA Elements*), señalan la existencia de importantes

elementos reguladores de la expresión genética en las áreas no codificantes, aún escasamente estudiadas.

Las nuevas herramientas de estudio del genoma (estudios de secuenciación de genoma completo, GWAS, modelos animales, estudio directo del mRNA en los tejidos, estudios de expresión de proteínas) van permitiéndonos ampliar nuestro conocimiento sobre la compleja regulación de la expresión de los genes, lo que nos permitirá en el futuro ser capaces de utilizar la información genética de forma eficiente.

**Tabla 2-3. Resumen de las limitaciones de los estudios genéticos en el TBP.**

| | Limitaciones | Soluciones |
|---|---|---|
| Relacionadas con la selección de la muestra | Escaso tamaño muestral. | Ampliar tamaños muestrales. Bases de datos comunes. Proyectos multicéntricos; metaanálisis. |
| | Falta de homogeneidad. | Segregación por subtipos.<br>Utilización de endofenotipos. |
| Limitaciones técnicas | Falta de desarrollo de las técnicas y exceso de volumen de datos. | Mejora de los equipos y desarrollo de protocolos comunes. |
| | Complejidad genética de las enfermedades de herencia compleja. | Tenerlo en cuenta en el análisis de los resultados: método de puntuación poligénica; análisis de rutas.<br>Lectura de patrones de metilación. |
| | Limitaciones en el conocimiento del genoma. | Proyecto ENCODE y otros.<br>Nuevas herramientas de estudio del genoma. |

## E. Conclusiones.

Hay que reconocer que el avance en el estudio de las bases genéticas del TBP ha sido hasta ahora bastante decepcionante, principalmente por los escasos resultados, que incluso teniendo en algunos casos significación estadística, carecen por completo de importancia clínica. Sin embargo, el día en el que el psiquiatra sea capaz de hacer diagnósticos genéticos en la clínica diaria está cada vez más cercano.

Basándonos en el conocimiento actual de otras enfermedades complejas como la enfermedad de Alzheimer, lo más probable es que seamos capaces en la próxima década de describir un grupo de mutaciones relacionadas con el TBP (y otras psicosis),

de elevada penetrancia, presentes en grupos familiares y en algunos casos esporádicos (mutaciones *de novo*), junto con otra serie de mutaciones que confieran un riesgo determinado de padecer la enfermedad. La caracterización de esas mutaciones nos permitirá avanzar en el conocimiento de las bases bioquímicas subyacentes a este trastorno, pudiendo desarrollar mejores tratamientos. Esto permitirá ir abandonando progresivamente el modelo diagnóstico actual, basado en síntomas, e ir entendiendo el TBP como un fenotipo complejo, compuesto por múltiples alteraciones conductuales, anímicas y biológicas que son resultado del funcionamiento de complejas redes bioquímicas y neuronales destinadas a capacitar al ser humano para adaptarse al medio.

# 3. SECUENCIACIÓN MASIVA PARALELA: REVISIÓN DE LAS TÉCNICAS ACTUALES.

La genética como ciencia nace en el siglo XIX con las observaciones de Mendel de los patrones de herencia de algunos caracteres morfológicos en las plantas de guisantes. En 1869 descubre en el núcleo de células sanguíneas lo que hoy conocemos como ADN, y en 1953 James F. Watson y Francis Crick postulan el modelo de doble hélice. En el año 1977, en paralelo, y con técnicas distintas, Gilbert y Maxam, y Fred Sanger (Maxam & Gilbert, 1977; Sanger, Nicklen, & Coulson, 1977), llevan a cabo la primera secuenciación de ADN. En abril de 1983, Kary Mullis da a conocer la técnica de reacción en cadena de la polimerasa o PCR, recibiendo por ello el Premio Nobel de Química en el año 1993.

MÉTODOS DE SECUENCIACIÓN DE PRIMERA GENERACIÓN.

Basándose en la técnica descrita por Sanger, los procesos manuales se fueron sustituyendo por métodos automatizados basados en electroforesis capilar o microarrays, más rápidos y con menor coste. En 2003 el proyecto Genoma Humano finaliza la secuenciación completa del genoma humano basándose en estas técnicas, con un coste total de casi tres billones de dólares y una duración de 13 años (Davis, 1990; Collins, Morgan, & Patrinos, 2003; International Human Genome Sequencing Consortium, 2004).

MÉTODOS DE SECUENCIACIÓN DE NUEVA GENERACIÓN.

En 2005 comienzan a desarrollarse técnicas de secuenciación que no están basadas en la técnica de Sanger. A partir de ese momento comienza la carrera de avances para lograr el objetivo de secuenciar un genoma completo por un precio de 1000 dólares. El objetivo es incrementar la longitud de las lecturas y minimizar el tiempo, convirtiendo la secuenciación en una herramienta lo suficientemente económica para ser utilizada en la rutina clínica. Posibles aplicaciones serían programas de *screening* de enfermedades, valoración pronóstica o como guía para tratamientos médicos personalizados.

Existen actualmente en el mercado distintas plataformas comerciales de secuenciación denominadas de "secuenciación masiva paralela", ya que son capaces de secuenciar múltiples cadenas de ADN al mismo tiempo. Las distintas plataformas se basan en técnicas diferentes y utilizan distintos métodos para la preparación de las muestras.

Dado que el objetivo de esta tesis no es repasar con detalle esta tecnología, se describirán exclusivamente las técnicas en las que se basan las actuales plataformas comerciales, nombrando al final aquellas de reciente desarrollo y prometedor futuro o "secuenciación de nueva generación". Igualmente, se expondrán brevemente los distintos métodos de preparación del ADN para ser secuenciado y, finalmente, se repasarán métodos de enriquecimiento y captura.

**A. <u>PREPARACIÓN DE LOS MOLDES DE ADN</u>** (Voelkerding, Dames, & Durtschi, 2009; Metzker, 2010)

Según la técnica de secuenciación que se vaya a utilizar será necesario preparar la muestra de ADN de una manera u otra. Para todas ellas se requiere la ruptura del ADN en fragmentos de distinto tamaño que posteriormente se unen en sus extremos a moléculas llamadas adaptadores. El conjunto de todos los fragmentos de ADN unidos a adaptadores se conoce como biblioteca.

**Moldes con amplificación clonal**

Algunos sistemas de imagen o detección necesitan de la utilización de numerosas copias de cada trozo de ADN para tener suficiente señal (al aumentar la relación señal ruido o *signal to noise ratio*). Para ello se utilizan métodos de amplificación, que consiguen crear millones de copias idénticas a la molécula de ADN original. Existen dos formas de llevar a cabo esta amplificación sin necesidad de utilizar células, lo que evita la pérdida arbitraria de secuencias asociada a los métodos de clonación bacteriana: la reacción en cadena de la polimerasa (PCR) en emulsión (emPCR) y la amplificación en fase sólida.

PCR en emulsión.

Se crea una biblioteca de moldes de ADN mediante la fragmentación aleatoria de la cadena original. Se unen adaptadores a sus extremos y se incluyen en una emulsión oleo-acuosa. Esta emulsión contiene microesferas recubiertas con moléculas que hibridan con los adaptadores y además actúan de cebadores. Tiene además todo lo necesario para llevar a cabo una PCR. Se crean las condiciones para que en cada esfera hibride únicamente un fragmento de ADN. Sobre éste se llevan a cabo múltiples PCR que generan copias clonales de ese fragmento concreto. Finalmente, todas las microesferas se inmovilizan en una superficie para ser secuenciadas (figura 3-1). Este tipo de amplificación es la que utilizan las plataformas Roche/454 Life Sciences o SOLiD.

Amplificación en fase sólida.

Se unen cebadores directos e inversos a un portaobjetos de forma covalente. Los distintos fragmentos de la librería de ADN, con sus extremos unidos a adaptadores, hibridan con los cebadores, iniciándose la reacción de PCR y la formación de nuevas cadenas desde ambos extremos en forma de “puente”. Se repite la operación en el área circundante gracias a los cebadores, hasta conseguir muchas localizaciones físicamente aisladas que contienen múltiples copias idénticas de un fragmento único (figura 3-2). Este método es el utilizado por Solexa (de la que ahora es propietaria la empresa Illumina).

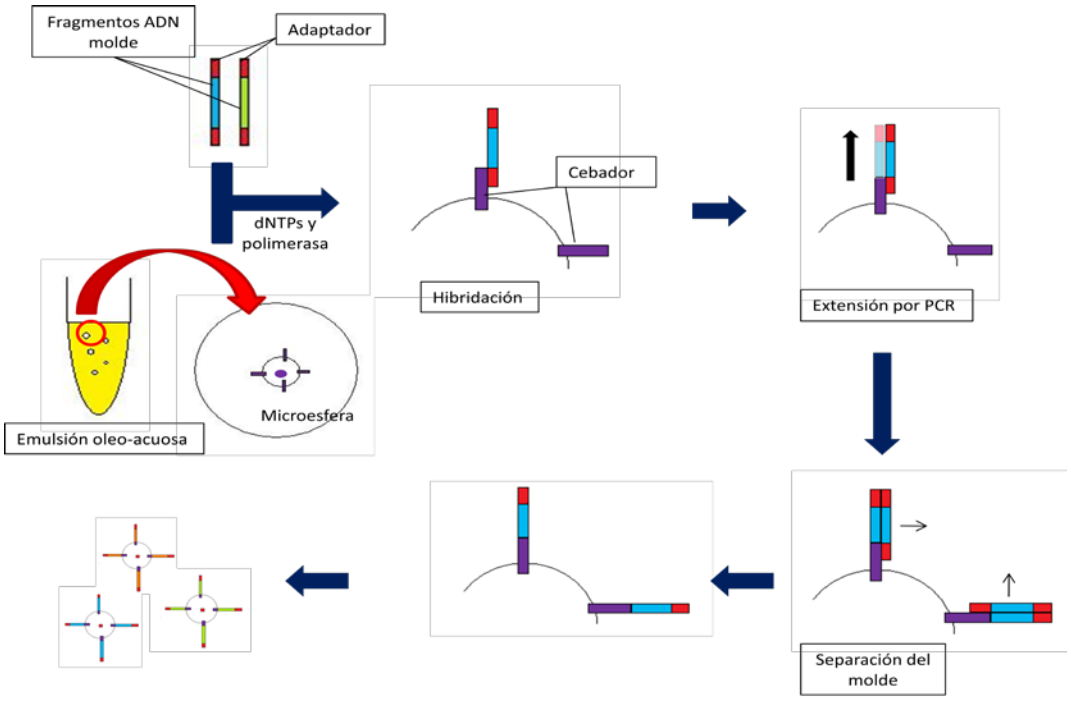



**Figura 3-1.** PCR en emulsión. Modificada de Metzker, 2010 (Metzker, 2010).

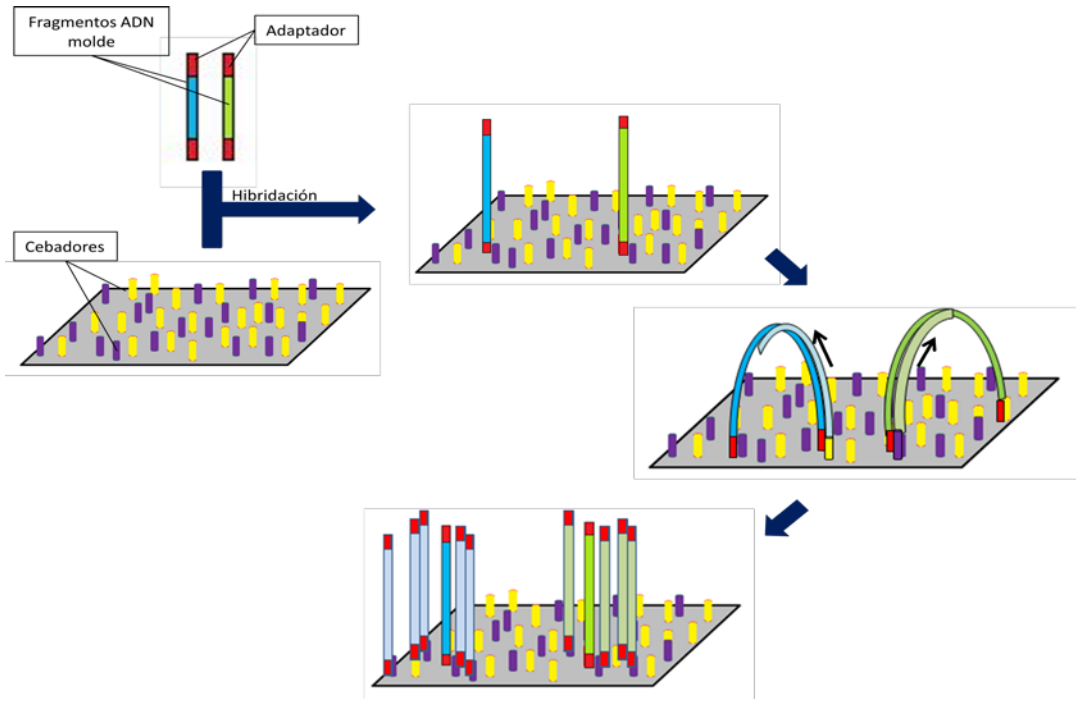



**Figura 3-2.** Amplificación en fase sólida (simplificación). Modificada de Metzker 2010 (Metzker, 2010). Los fragmentos de ADN (azul y verde) se hibridan con los cebadores del portaobjetos (amarillos y morados) gracias a los adaptadores añadidos a sus extremos (rojos). Comienza la PCR formándose las nuevas cadenas (en color más claro), idénticas a la original, "doblándose" las cadenas de ADN en forma de "puente". Cada una de estas cadenas recién formadas sirve como molde para la fabricación de nuevas cadenas, también idénticas a la inicial. Finalmente quedan "agregados" de cadenas idénticas unidas covalentemente con la superficie sólida.

**Moldes de molécula única**

Necesitan menos cantidad de ADN inicial y evitan los problemas asociados a la PCR, como la introducción de mutaciones o los problemas de amplificación de las áreas ricas en AT o GC.

Los fragmentos de ADN se pueden inmovilizar en soportes sólidos mediante tres técnicas:

-Los adaptadores unidos a los extremos de cada fragmento de la biblioteca de ADN hibridan con los cebadores que están unidos de forma covalente a la superficie sólida.

-Los propios fragmentos de ADN se unen a la superficie sólida de forma covalente.

Ambos métodos son los utilizados en la plataforma Helicos BioSciences.

-Se unen al soporte sólido las moléculas de polimerasa, distribuidas en el espacio. A cada una de ellas se le une un fragmento de ADN junto con un cebador.

Este último método es el que utiliza Pacific Biosciences y se describe en patentes de Life/VisiGen.

**B. <u>TÉCNICAS DE SECUENCIACIÓN Y DETECCIÓN</u>** (Shendure, Mitra, Varma, & Church, 2004; Metzker, 2005; ten, Jr. & Grody, 2008; Tucker, Marra, & Friedman, 2009; Voelkerding et al., 2009; Metzker, 2010; Schadt, Turner, & Kasarskis, 2010; Mardis, 2011)

TÉCNICAS POR ADICIÓN DE NUCLEÓTIDO ÚNICO

Se basan en una secuenciación por síntesis (dependiente de ADN polimerasa). Utilizan ciclos de lectura y lavado, y son más rápidas que la técnica de Sanger ya que secuencian un gran número de cadenas de ADN al mismo tiempo.

**Pirosecuenciación:**

El ejemplo más desarrollado es la pirosecuenciación, descrita por primera vez por Hyman en 1988 (Hyman, 1988) y desarrollada por el equipo de Nyrén y Ronaghi, del Instituto de Tecnología de Estocolmo (Ronaghi, Karamohamed, Pettersson, Uhlen,

& Nyren, 1996; Ronaghi, Pettersson, Uhlen, & Nyren, 1998; Ronaghi, Nygren, Lundeberg, & Nyren, 1999).

Esta técnica consiste en la adición secuencial de nucleótidos a una solución que contiene los moldes de ADN y ADN polimerasa. Cuando el dNTP (deoxiribonucleósido trifosfato) añadido corresponde con el complementario de la cadena molde se activa la ADN polimerasa, une el dNTP a la cadena y se libera un pirofosfato por cada uno de los nucleótidos añadidos. Este pirofosfato se convierte en luz visible gracias a sulfurilasas y luciferasas. Esta luz, de intensidad variable según el número de nucleótidos incorporados, es recogida y cuantificada, revelando tras ser analizada la secuencia original de ADN.

La luz emitida en cada ciclo se recoge en una serie de picos llamado pirograma, que corresponde al orden de los nucleótidos complementarios a la cadena molde (figura 3-3). La pausa en la secuenciación tras la incorporación de cada dNTP a la cadena se consigue añadiendo éstos a la solución de forma secuencial, limitando la cantidad disponible para cada reacción.

Esta tecnología es la que incorpora la plataforma Roche/454 Life Sciences. Integra la pirosecuenciación con la plataforma PicoTiterPlate, desarrollada por ellos (Margulies et al., 2005). Esta plataforma lleva a cabo una PCR en emulsión sobre el PicoTiterPlate, una placa grabada con cientos de miles de pocillos de unos 40 micrometros. Las microesferas con las copias de la biblioteca de ADN se introducen de forma individual en los pocillos junto con los enzimas necesarios para llevar a cabo la pirosecuenciación, llevándose a cabo cientos de miles de reacciones al mismo tiempo y recogiéndose la luz emitida por cada una de ellas con una cámara CCD (*charge-coupled device* o “dispositivo de carga acoplada”).

La ventaja de esta técnica es que obtiene longitudes de lectura semejantes a las de la técnica de Sanger. Sin embargo, presenta dos inconvenientes importantes:

1. Se realiza sobre una biblioteca de fragmentos amplificados con PCR. Esto puede introducir errores en la lectura a través de los fenómenos de desfase, lo que aumenta el ruido y limita la longitud de las lecturas.

2. La lectura de áreas de repetición de homopolímeros es poco exacta, dando con frecuencia errores.

En esta plataforma, las inserciones son el error más frecuente, seguido de las delecciones. No se producen sustituciones, por lo que sería la de elección para la validación de una secuencia dada.

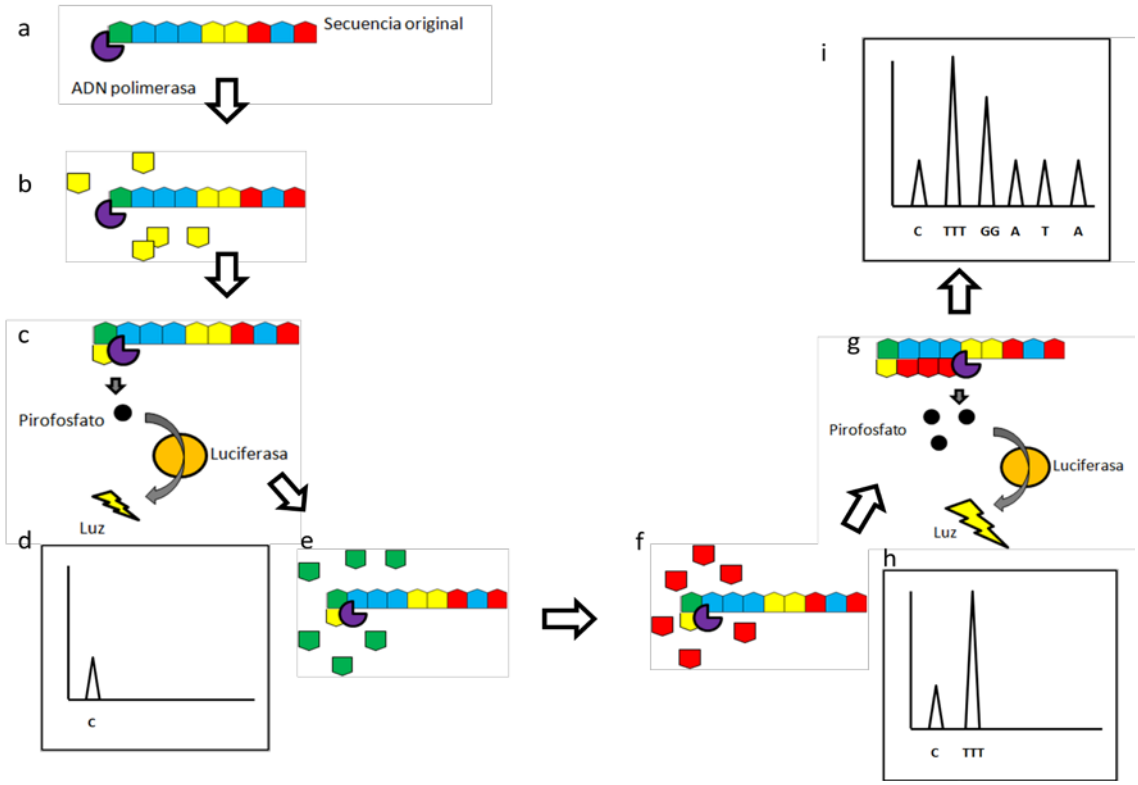



**Figura 3-3.** Pirosecuenciación. a) La polimerasa está unida al fragmento de ADN que se quiere secuenciar y preparada para crear la cadena complementaria. b) Se incorpora a la solución un tipo de dNTP (amarillo). c) Como corresponde con el complementario de la cadena original, la ADN polimerasa lo une y libera un pirofosfato. Éste se trasforma en ATP por la acción de una sulfurilasa (no mostrado), y este ATP en luz gracias a una luciferasa. d) La luz es recogida y trasformada en un gráfico que muestra un pico cuya altura corresponde a la intensidad de la luz emitida. e) Se añade otro tipo de dNTP a la solución (verde). Como no corresponde con el complementario, se lava y no sucede nada. f) Se añade un nuevo tipo de dNTP (rojo), que en este caso sí corresponde con el complementario. g) Se incorporan 3 nucleótidos a la cadena, por lo que la luz emitida tiene el triple de intensidad. h) Esto se refleja en el pirograma con un pico mayor que el anterior. i) El proceso se repite hasta finalizar la lectura.

**Secuenciador semiconductor (Ion Torrent's *sequencer*).**

Su funcionamiento es semejante al de la pirosecuenciación, pero en lugar de transformar el pirofosfato liberado en la incorporación de cada nucleótido a la cadena en luz, mide el cambio de pH producido por el protón liberado.

TÉCNICAS DE TERMINACIÓN REVERSIBLE CÍCLICA (TRC):

Basadas también en la ADN polimerasa y en ciclos de secuenciación. Utilizan terminadores reversibles, que son nucleótidos modificados de forma que se les une una molécula fluorescente que no permite continuar la incorporación del siguiente nucleótido a la cadena hasta que no es retirada.

Los ciclos se componen de: incorporación del nuevo nucleótido protegido con un grupo fluorescente a la cadena en formación, lavado del resto de nucleótidos no unidos, lectura de la señal fluorescente y posterior retirada del grupo protector para permitir la unión del siguiente nucleótido.

Las principales ventajas de estas técnicas es que limitan la posibilidad de lecturas erróneas al eliminar los grupos fluorescentes previos y, además, detectan mejor las secuencias homopolímeras que la pirosecuenciación.

Existen en el mercado dos plataformas de secuenciación basadas en esta técnica:

**-Illumina/Solexa** (Bentley et al., 2008)**:**

Combina la amplificación en fase sólida descrita previamente con una secuenciación de terminación reversible cíclica con nucleótidos modificados marcados con 4 colores. Se incorporan los cuatro nucleótidos (A, C, T y G) marcados cada uno de ellos con un color diferente a la placa a la que están unidos los fragmentos de ADN amplificados. Se produce la unión del nucleótido complementario a cada una de las cadenas en formación. Tras la unión, se retiran los nucleótidos sobrantes y se recoge la imagen mediante láseres. Este proceso se va repitiendo y finalmente las imágenes se analizan para identificar la secuencia completa (figura 3-4).

Se utiliza la amplificación clonal para amplificar la señal y minimizar el ruido. La señal fluorescente observada será el consenso de las emitidas por todos los nucleótidos añadidos a cada grupo de cadenas idénticas de ADN en cada uno de los ciclos.

Los inconvenientes principales de esta técnica son:

1. Precisa de amplificación (en este caso PCR en fase sólida), lo que introduce errores, principalmente infrarrepresentación de regiones homopolímeras.
2. Sufre fenómenos de desfase. En cada ciclo de incorporación se van introduciendo errores, que se van acumulando. Como la señal observada es el consenso de la emitida por las cadenas clonales, va aumentando progresivamente el ruido de la señal fluorescente. Esto limita la longitud de las lecturas.

El error más común son las sustituciones, siendo más frecuentes cuando el nucleótido anterior es una guanina.

**-Helicos BioSciences** (Harris et al., 2008)**:**

Este grupo comercializa el secuenciador HeliScope (Braslavsky, Hebert, Kartalov, & Quake, 2003). Se basa en una secuenciación reversible cíclica utilizando marcadores fluorescentes de un solo color. Escanea miles de millones de moléculas individuales de ADN fijadas a una superficie mientras crecen utilizando un cebador, una polimerasa modificada y análogos de nucleótidos marcados (figura 3-5) llamados *Virtual Terminator nucleotides* (VTn). A diferencia de los análogos de nucleótidos utilizados en la plataforma de Solexa, estos VTn no tienen bloqueado el extremo 3', estrategia más eficiente, ya que requiere únicamente la ruptura de un enlace para continuar con la síntesis de la cadena, precisando los primeros la ruptura de dos enlaces.

Como importante ventaja frente a la plataforma de Illumina/Solexa, no requiere ningún tipo de amplificación de la muestra, lo que anula los problemas asociados a esta técnica. Sin embargo, presenta la desventaja de generar lecturas muy cortas, lo que dificulta el ensamblaje posterior de éstas. Esto limita su uso en para la secuenciación de novo.

Los errores más frecuentes son las delecciones en las secuencias repetitivas.

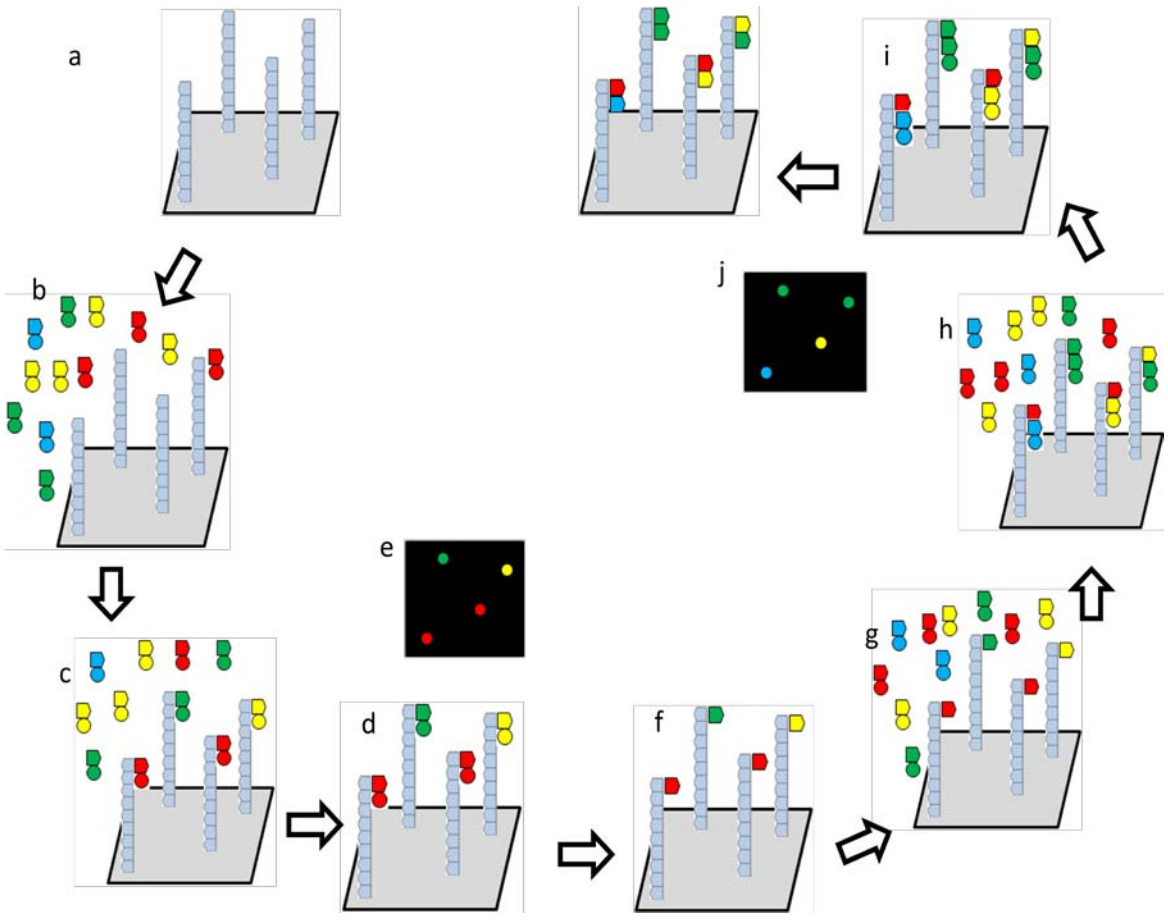



**Figura 3-4.** Secuenciación TRC con cuatro colores. a) La biblioteca de ADN está unida de forma covalente a una superficie sólida. Para simplificar la imagen se muestra una cadena única, aunque en realidad son grupos de cadenas iguales. b) Se introducen en la solución los cuatro nucleótidos modificados (A, G, T y C) cada uno marcado con una molécula fluorescente de un color (círculos rojo, verde, amarillo y azul). c) Los nucleótidos complementarios se unen a las cadenas, uno en cada una, parándose la síntesis, ya que los grupos adheridos, que son los que contienen el fluorescente, bloquean el extremo 3' y no permiten continuar añadiendo nucleótidos a la cadena. d) Los nucleótidos sobrantes son lavados. e) Se recoge la imagen. f) El grupo fluorescente unido al extremo 3' es eliminado, pudiendo reiniciarse la síntesis. g, h ,i) Se vuelven a introducir a la solución nucleótidos marcados, repitiéndose el proceso y j) obteniéndose una imagen por ciclo, hasta que termina la secuenciación.

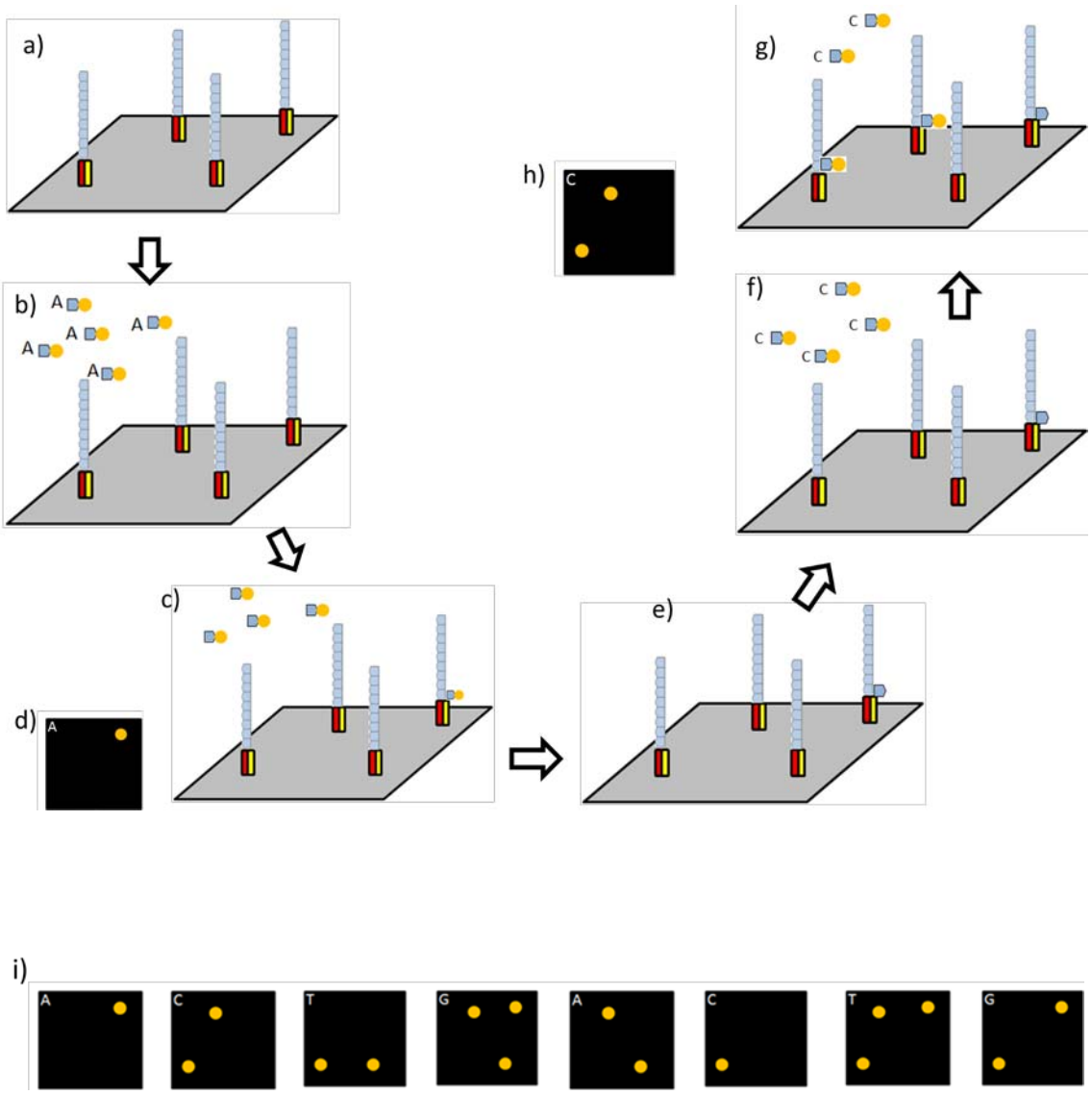



**Figura 3-5.** Secuenciación TRC con un solo color. a) Se parte de una superficie sólida a la que están unidos de forma covalente cebadores (amarillo) sobre los que ha hibridado la biblioteca de ADN gracias a adaptadores (rojo). b) Se añade a la solución un único tipo de nucleótido marcado modificado (por ejemplo, adenina). c) La adenina marcada se une a aquellas cadenas en la que es el nucleótido complementario, y se frena la síntesis. Tras el lavado de los nucleótidos sobrantes, d) se recoge la imagen. e) Se elimina el grupo fluorescente, permitiendo continuar con la síntesis. f) Se añade otro nucleótido (por ejemplo, citosina), que g) se une a los fragmentos de ADN en el que es el complementario. h) Se obtiene una nueva imagen. Este ciclo se repite, añadiendo de forma secuencial los distintos nucleótidos. i) La lectura de la secuencia de imágenes revela la secuencia de ADN, que en este ejemplo sería: fragmento delante izquierda: CTCTG; fragmento delante derecha: TGA; fragmento detrás izquierda: CGAT; fragmento detrás derecha: AGTG.

SECUENCIACIÓN POR LIGACIÓN

Utilizada con éxito para la secuenciación del genoma de E. coli por Shendure y cols. en 2005 (Shendure et al., 2005), este tipo de secuenciación utiliza una ADN ligasa en lugar de una ADN polimerasa para crear la cadena complementaria. Se basa en la utilización de sondas de ADN de 1 o 2 nucleótidos, marcadas con fluorescencia, que hibridan con el fragmento a secuenciar. Una simplificación de este método se muestra en la figura 3-6.

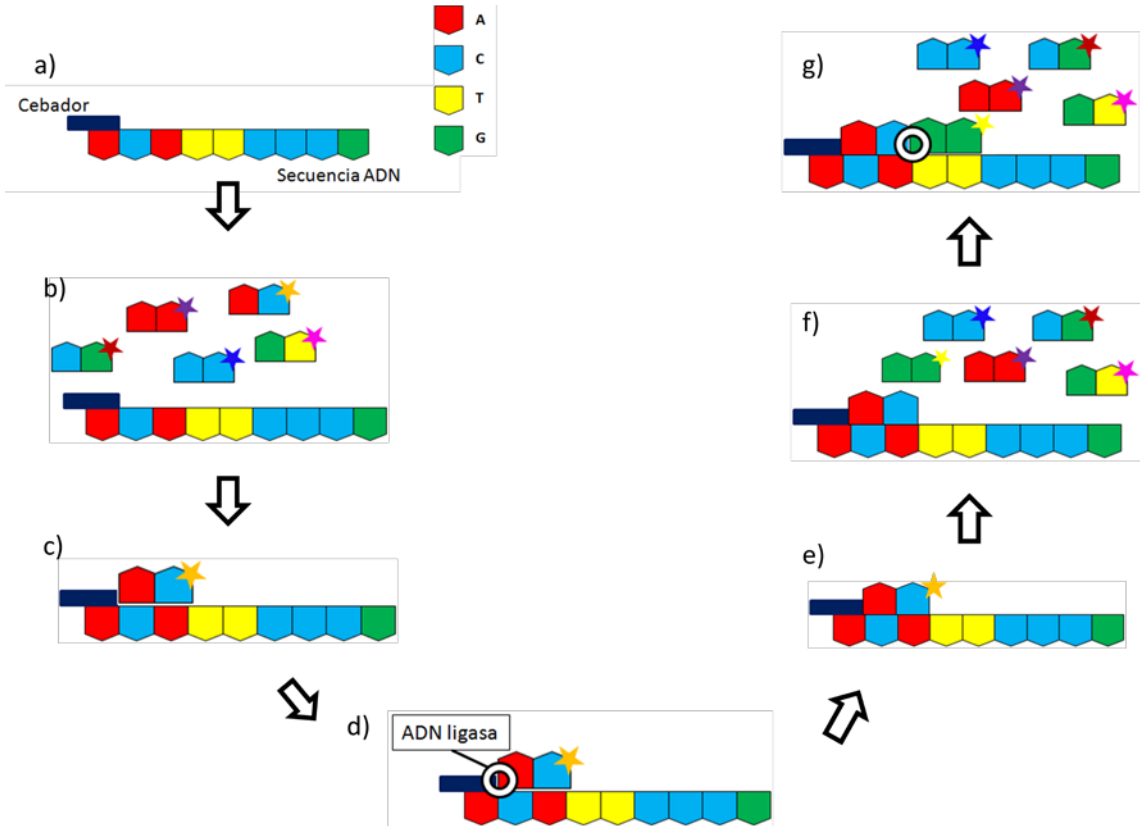



**Figura 3-6**. a) Se une un cebador al fragmento de ADN a secuenciar. b) Se introducen las sondas de 2 nucleótidos, cada una de ellas marcadas con una molécula fluorescente de un color diferente. c) La que corresponde a la secuencia complementaria de la cadena de ADN que sigue al lugar de unión del cebador se sitúa en su lugar y d) es unida al cebador gracias a la ADN ligasa. e) El resto de sondas son lavadas y se recoge la imagen. En este ejemplo fluorescencia naranja, que corresponde a la sonda adenina-citosina. f) Se elimina la molécula fluorescente y se vuelven a introducir sondas marcadas. g) Se repite este ciclo de ligación y lectura de la fluorescencia una y otra vez hasta completar el fragmento a secuenciar.

Esta técnica es la que utiliza la plataforma SOLiD (*Sequencing by Oligonucleotide Ligation and Detection*) de Applied Biosystems (Valouev et al., 2008). La plataforma SOLiD utiliza una PCR en emulsión para amplificar la biblioteca de ADN. Posteriormente hibrida un cebador a la posición n de cada fragmento de ADN a secuenciar. Su peculiaridad es que utiliza sondas de 8 nucleótidos. En estas sondas, los dos primeros nucleótidos son los que sirven para unirse a la cadena de ADN complementaria, los 3 siguientes son nucleótidos degenerados y los 3 últimos son eliminados junto con la molécula fluorescente. El ciclo de unión con ligasa se repite 10 veces. Para el siguiente ciclo se une un nuevo cebador, en esta ocasión en la posición n-1 y se repiten los 10 ciclos (figura 3-7). Estos 10 ciclos de ligación se realizan un total de 5 veces. Esto permite finalmente obtener 2 lecturas de cada nucleótido de la cadena original. Las distintas lecturas se ordenan, permitiendo descifrar la secuencia.

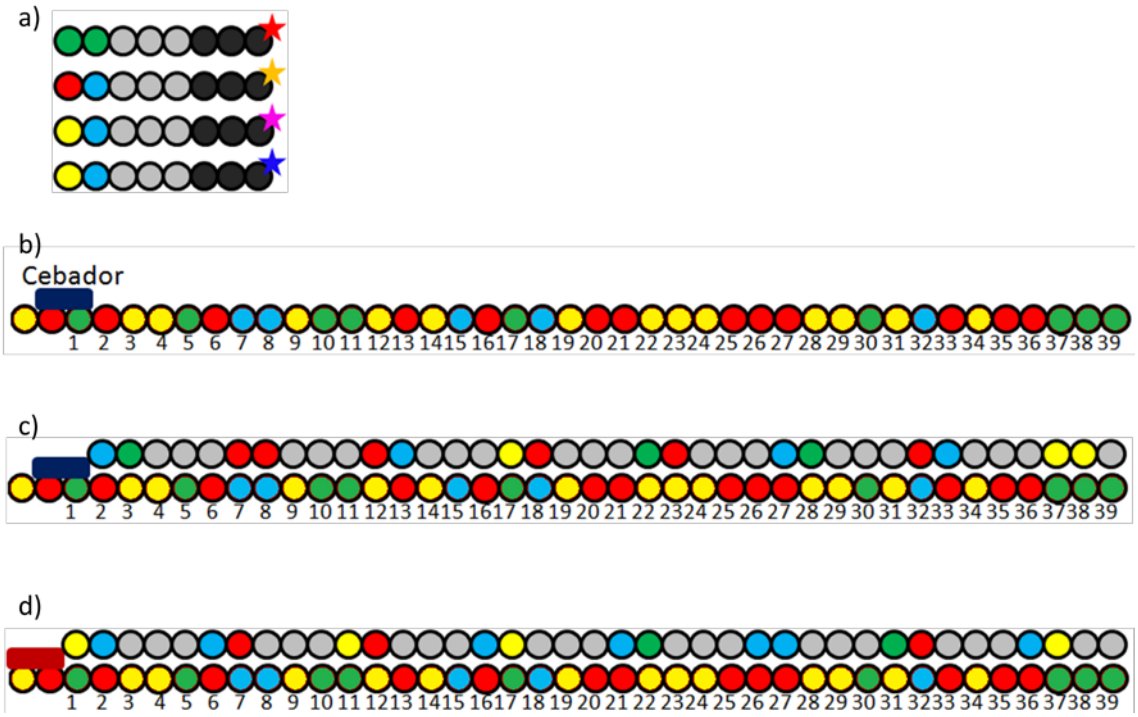



**Figura 3-7**. Plataforma SOLID. a) Sondas de 8 nucleótidos. b) El cebador se une al nucleótido de la posición 1. c) En los 10 primeros ciclos de ligación se obtiene lectura de los nucleótidos en las posiciones 2 y 3, 7 y 8, 12 y 13, 17 y 18, 22 y 23, 27 y 28, 32 y 33, 37 y 38, 42 y 43, 47 y 48. Las dos posiciones de lectura corresponden a los nucleótidos de las sondas; las tres posiciones no leídas corresponden a los nucleótidos degenerados. d) Para el siguiente ciclo se une un nuevo cebador, pero en este caso en la posición (n-1). En los siguientes 10 ciclos se leen los nucleótidos de las posiciones 1 y 2, 6 y 7, 11 y 12, 16 y 17, etc.

Como importante ventaja, la práctica ausencia de sustituciones, ya que lee dos veces cada base. Sus principales desventajas son la mala lectura de regiones ricas en AT y GC.

SECUENCIACIÓN EN TIEMPO REAL

Basadas también en la ADN polimerasa, estas técnicas no frenan la secuenciación tras la incorporación de cada base para llevar a cabo la lectura ni invierten tiempo en ciclos de lavado, por lo que son más rápidas. Tampoco requieren amplificación, lo que evita las desventajas asociadas a este paso y no sufren fenómenos de desfase.

***Zero-mode waveguide.***

Se trata de un dispositivo que permite limitar el campo de observación lo suficiente como para captar exclusivamente la fluorescencia emitida por un nucleótido, el que se está incorporando de forma sucesiva la ADN polimerasa mientras sintetiza la cadena.

Este dispositivo ha sido desarrollado por Pacific Biosciences e incorporado en su plataforma SMRT (*Single Molecule Real Time*). Moléculas individuales de ADN polimerasa son ancladas a la superficie inferior de detectores *zero-mode waveguide* (ZMW), pocillos de unos 30nm de diámetro, y van incorporando nucleótidos fosfato marcados con fluorescente de cuatro colores (uno por cada base) a la cadena a tiempo real. Los pulsos de fluorescencia emitidos con cada incorporación de un nucleótido a la cadena se van leyendo, mientras este proceso tiene lugar en miles de detectores ZMW de forma simultánea.

Tiene una tasa de error de más del 5%, principalmente en forma de inserciones y delecciones, en relación con las limitaciones de la propia técnica (ausencia de lectura por incorporación muy seguida de dos nucleótidos, errores por lecturas de nucleótidos erróneos presentes en el sitio activo de la ADN polimerasa…). Sin embargo, gracias a la repetición de lecturas, la precisión puede aumentar hasta más de un 99.999%. Realiza

lecturas muy largas, de hasta 10.000 pares de bases, lo que facilita su posterior ensamblaje.

Dado que aporta información cinética de la actividad de la polimerasa, podría utilizarse para determinar patrones metilación, lo que abre una posibilidad de mucho interés para la investigación.

**Técnica FRET** (*Fluorescence Resonance Energy Transfer*) de pares sencillos (Braslavsky et al., 2003).

Fenómeno por el que la excitación de un cromóforo puede pasar a otro cercano. Si los cromóforos son fluorescentes, la transferencia de energía produce la aparición de fluorescencia. Esta interacción depende de la distancia entre ambos, lo que permite utilizarla para demostrar interacción de moléculas.

Life/VisiGen ha desarrollado ADN polimerasas con un marcador fluorescente que cuando se encuentra próximo a un nucleótido unido a un cromóforo receptor produce fluorescencia a través de este sistema.

**Tabla 3-1. Comparativa entre distintas plataformas de secuenciación.**

| Plataforma | Técnica | Longitud lecturas | Tasa error | Tpo/run | Ventajas | Inconvenientes |
|---|---|---|---|---|---|---|
| **Roche 454** | Pirosecuenciación<br>PCR emulsión | Largas | 1% | Horas | Gran longitud lecturas.<br>Secuenciación rápida.<br>No produce sustituciones. | Mala cobertura de secuencias homopolímeras.<br>Más cara. |
| **Illumina/Solexa HiSeq v3** | ADN polimerasa<br>Amplificación fase sólida. | Cortas | <0.1% | Días | La más utlizada por su coste-eficacia. | Error más frecuente: sustituciones. |
| **SOLiD 5500** | ADN ligasa<br>PCR emulsión | Cortas | <0.01% | Días | Tasa de errores muy baja. | Lenta. |
| **Pacific Biosciences** | Sintesis<br>No necesita preparación de la muestra | Muy largas | 15% | Minutos | Gran longitud de lecturas. | Frecuentes inserciones y delecciones. |

OTROS

<u>Técnicas de secuenciación basadas en nanoporos</u> (Branton et al., 2008)

Se basan en la identificación de las distintas bases de la cadena de ADN gracias a una señal óptica o por la variación que se produce en una corriente eléctrica al pasar la cadena a través de un nanoporo anclado a una membrana.

Secuenciación con nanoporos por detección eléctrica (figura 3-8).

En desarrollo por Oxford Nanopore. Se ancla una exonucleasa en la superficie externa de un poro de 1,5nm situado en una membrana lipídica, que va rompiendo la cadena nucleótido a nucleótido. Estos nucleótidos pasan a través del nanoporo, obstruyéndolo en distintos grados y obteniéndose fluctuaciones en su conductancia eléctrica, lo que es detectado y analizado.

La tasa de error es de un 40% en una lectura única, llegando hasta 0,1% en 15 lecturas.

Secuenciación con nanoporos por traslocación (figura 3-9).

Una cadena sencilla de ADN pasa a través del poro sin necesidad de ser fragmentada. Las variaciones en la conductancia eléctrica que produce cada base al pasar por el poro permiten determinar la secuencia.

Secuenciación con nanoporos de lectura óptica (figura 3-10).

Se anclan moléculas fluorescentes a las bases y se va leyendo la luz emitida por a medida la cadena de ADN pasa por el poro con una cámara CCD.

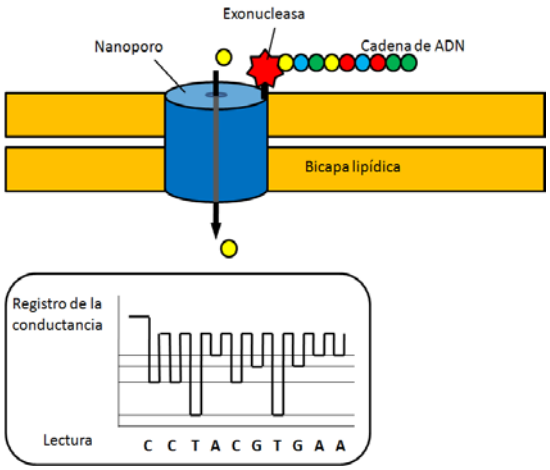



**Figura 3-8.** Secuenciación con nanoporos por detección eléctrica (simplificación).

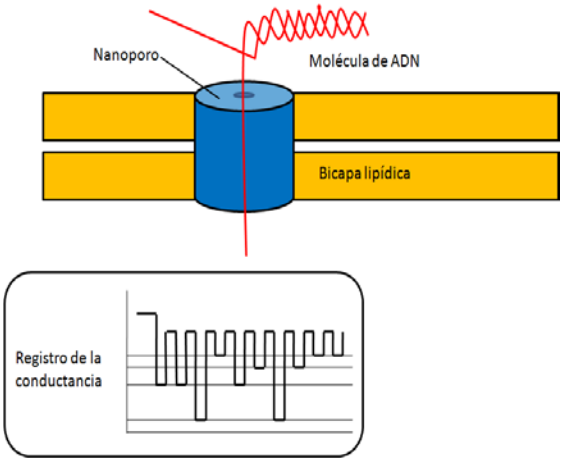



**Figura 3-9.** Secuenciación con nanoporos por traslocación (simplificación).

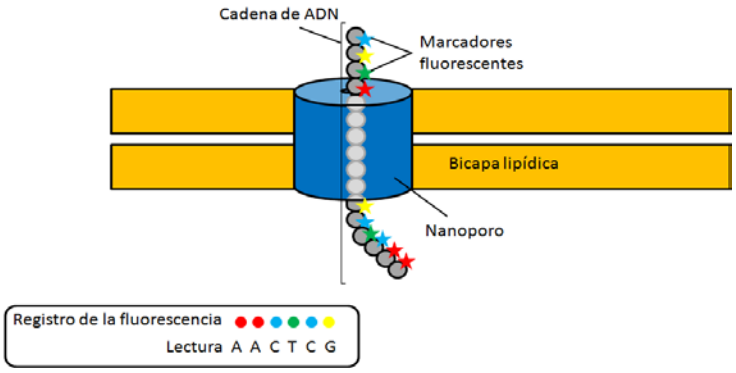



**Figura 3-10.** Secuenciación con nanoporos de lectura óptica (simplificación).

Para ampliar la información consultar la página http://www.nanoporetech.com/.

Observación directa del ADN con técnicas de microscopía

Otra tecnología, encuadrada en los secuenciadores de tercera generación, es la desarrollada por compañías como Halcyon o ZS Genetics, que utiliza la microscopía electrónica y permite leer la secuencia del ADN directamente por métodos ópticos.

Transistor IBM

IBM ha creado un dispositivo formado por una estructura artificial de nanoporos formados por capas alternas de un material dieléctrico y metal. La cadena de ADN pasa a través de estos poros, y las alteraciones en la corriente producidas por el paso de cada una de las bases permite identificarlas y, de esa forma, descifrar la secuencia de la cadena. La velocidad de paso de la cadena de ADN a través de los poros se puede controlar modulando la corriente aplicada a este material.

## C. TÉCNICAS DE CAPTURA O ENRIQUECIMIENTO DE LAS ÁREAS DE INTERÉS

No siempre el objetivo de los estudios es el genoma completo, sino que en ocasiones interesa secuenciar un único gen, un cromosoma o algunos fragmentos concretos del ADN. Para poder seleccionar exclusivamente las áreas de interés antes de la secuenciación con las nuevas plataformas se han desarrollado una serie de técnicas de captura, descritas a continuación:

**Microarrays o hibridación en fase sólida** (Albert et al., 2007; Hodges et al., 2007; Okou et al., 2007; Okou et al., 2009; Summerer, 2009)**.**

Se construyen microarrays con oligonucleótidos unidos covalentemente a una superficie sólida con secuencias complementarias a aquellas que se quieren seleccionar. El ADN se fragmenta y las áreas de interés hibridan en los oligonucleótidos, quedando unidas a la superficie sólida. Las secuencias sobrantes son eliminadas. Si fuera necesario, estas secuencias seleccionadas son amplificadas por PCR (figura 3-11).

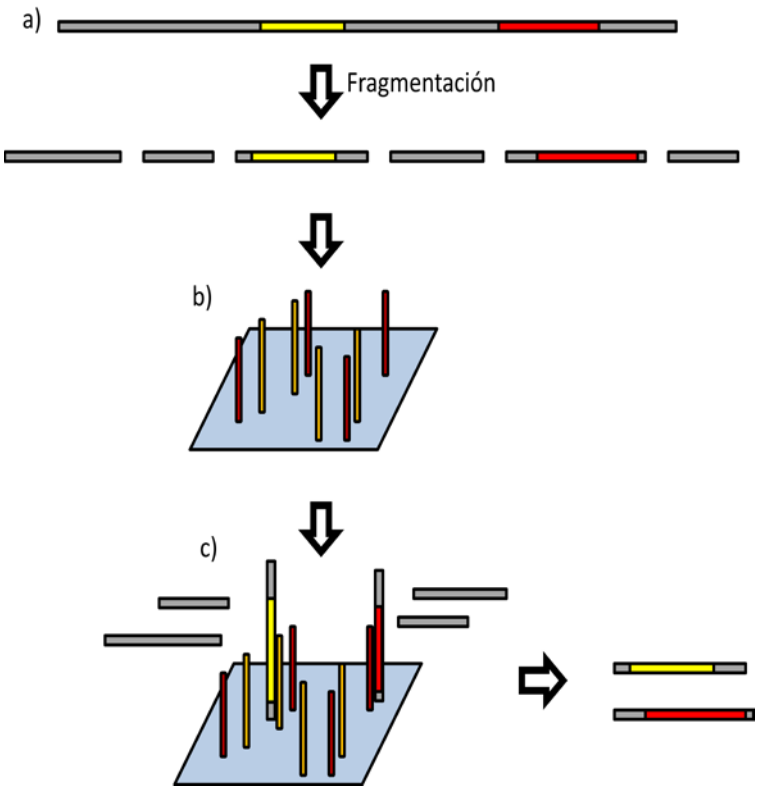



**Figura 3-11.** Hibridación en fase sólida. a) Partimos de una molécula de ADN que contiene las áreas de interés (amarillo y rojo). Esta molécula se fragmenta y
b) se expone a un array, superficie sólida a la que están unidos oligonucleótidos con secuencias complementarias a las de interés. c) Se produce la hibridación de las secuencias seleccionadas, lavándose el resto.

Roche/NimbleGen ha desarrollado y comercializa estos microarrays, entre otros, el desarrollado para la captura del exoma (SeqCap EZ *Exome capture system*). Los estudios realizados han combinado esta técnica con las plataformas de Illumina y Roche/454, obteniéndose buenos resultados.

**Hibridación en solución** (Dahl, Gullberg, Stenberg, Landegren, & Nilsson, 2005; Stenberg, Dahl, Landegren, & Nilsson, 2005; Akhras et al., 2007; Porreca et al., 2007; Summerer, 2009; Turner, Lee, Ng, Nickerson, & Shendure, 2009)**.**

Existen dos formas de hibridación en solución:

-Técnicas basadas en moléculas de ADN circular.

Hay varias versiones o variantes de la técnica de sondas de inversión molecular o "*molecular inversion probe*" (MIP), todas ellas basadas en el mismo principio. Se crean oligonucleótidos de ADN con la característica de que las secuencias de sus extremos son complementarias a las secuencias de los extremos del fragmento de ADN que se desea seleccionar. Estos oligunucleótidos se combinan con el ADN, hibridando sus extremos con las áreas complementarias de la cadena. Se rellena el hueco gracias a la ADN polimerasa, formando un ADN circular, y se elimina todo el ADN sobrante, quedando exclusivamente las secuencias de interés incluidas en los fragmentos de ADN circular (figura 3-12).

Estudios publicados que combinan esta técnica con plataformas de secuenciación de nueva generación describen la captura del 90-98% de las secuencias de interés con una profundidad en el rango de 10X para algo más del 50% de ellas, todavía en inferioridad respecto a la técnica de microarrays (Mamanova et al., 2010). Otro inconveniente es que, dado que la biblioteca generada está formada por los fragmentos de interés completos (en lugar de realizarse una fragmentación aleatoria del ADN), en el caso de fragmentos largos y utilizando plataformas con longitudes de lectura cortas podría haber un déficit de cobertura de las secuencias intermedias. Finalmente, en las lecturas finales tras la secuenciación se incluirán numerosas secuencias repetidas comunes que corresponden a adaptadores u otras secuencias pertenecientes a los oligonucleótidos, lo que limita la calidad de los datos y dificulta su análisis informático.

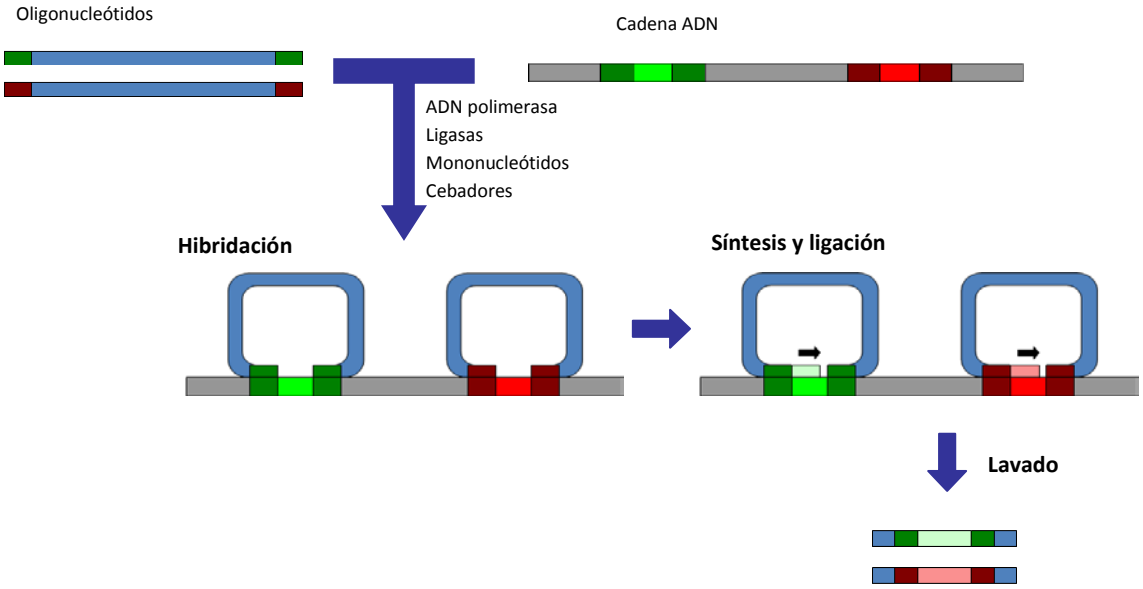



**Figura 3-12.** Sondas de inversión molecular. Se crean los oligonucleótidos de ADN por síntesis en microarray (Agilent), formados por una secuencia común (azul), las secuencias de los extremos, complementarias a las secuencias de los extremos del fragmento de ADN que se desea seleccionar (rojo y verde oscuro), y adaptadores (no se muestran), que permiten la amplificación de estos oligonucleótidos por PCR y luego son eliminados. Estos oligunucleótidos se añaden a una solución en la que está el ADN diana con las áreas que se desean seleccionar (rojo y verde claro), cebadores, ADN polimerasa, mononucleótidos y ligasas. Los oligonucleótidos hibridan con sus cadenas complementarias, los extremos de las áreas de interés. El hueco que queda se rellena con la ADN polimerasa y se forma un ADN circular con la intervención de enzimas ligasas. Posteriormente se elimina todo el ADN lineal que queda en la solución gracias a exonucleasas específicas, quedando exclusivamente el fragmento deseado, que posteriormente se fragmenta para convertirlo en un ADN lineal y poder iniciar los procesos siguientes.

-Sondas de captura de ARN biotinizado (Gnirke et al., 2009).

Se crean oligonucleótidos en un microarray con secuencias complementarias a las áreas de interés. A partir de estas cadenas, por transcripción, se generan ARNs que hibridan con las áreas del ADN que se desean seleccionar (figura 3-13).

Desarrollada por Agilent, actualmente se comercializan diversos kits de captura de exoma completo, entre otros el desarrollado por ellos (SureSelect Human All Exon).

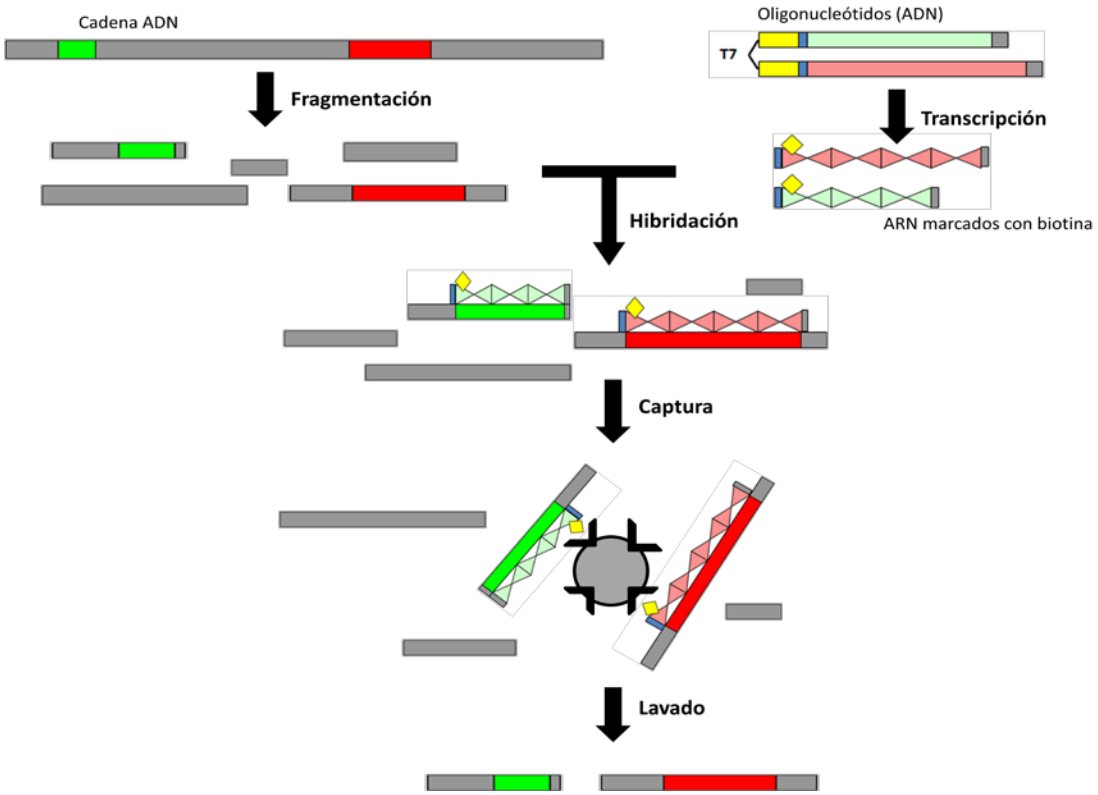



**Figura 3-13.** Sondas de captura de ARN biotinizado. Se crean los oligonucleótidos con secuencias de unas 170 pb complementarias a las áreas de interés (verde y rojo) y cebadores universales, y son ligados a adaptadores con la secuencia promotora de polimerasa T7 (amarillo). Se inicia una transcripción **in** *vitro*, y tomando como molde cada oligonucleótido, se crea un ARN que, gracias a la secuencia T7, incorpora una uridina trifosfato (UTP) marcada en el extremo 5' con una biotina (rombo amarillo). Estos ARN marcados se incorporan en la solución que contiene la biblioteca con el ADN fragmentado, hibridándose con las secuencias de interés. Los híbridos formados ARN-ADN son capturados con unas microesferas magnéticas revestidas de estreptavidina. Posteriormente se digieren los ARN, quedando exclusivamente los fragmentos seleccionados de ADN.

Esta técnica, que combina la flexibilidad y economía de la síntesis de oligonucleótidos en un array con la cinética de la hibridación en solución, se ha utilizado con la plataforma de Illumina, con buenos resultados. Sin embargo, presenta el inconveniente de la longitud de las sondas, habitualmente más largas que la mayoría de los exones (unos 120bp de media). Esto hace que se seleccionen numerosos fragmentos de ADN situados más allá de los extremos de las áreas de interés, que podría resultar, en el caso de plataformas de secuenciación de lecturas cortas, con una menor cobertura de las áreas centrales de estos fragmentos y una sobrerrepresentación de las áreas de los extremos y adyacentes.

Si se comparan la captura en array y en solución, se observa que con la primera se obtienen lecturas de mejor calidad y mejores coberturas, y aunque las técnicas en solución generan librerías con menos lecturas duplicadas, a la hora de la alineación no resultan muy diferentes (Sulonen et al., 2011). Otras comparativas encuentran que ambas técnicas presentan resultados son similares (Mamanova et al., 2010).

**Tecnología de PCR en microgotas** (Tewhey et al., 2009; Metzker, 2010)**.**

Utiliza la plataforma RainStorm, desarrollada por RainDance Technologies, dispositivo que crea microgotas acuosas en una solución de aceite que contienen cebadores directos e inversos. Cada una de las gotas (hasta 4000 por experimento) contiene la pareja de cebadores específica que corresponde a cada región de interés del ADN. Estas gotas se combinan en este mismo dispositivo con otras microgotas que contienen el ADN fragmentado junto con todo lo necesario para llevar a cabo una PCR, en una proporción de 1:1. Ambos tipos de microgotas se combinan gracias a impulsos eléctricos. En estas nuevas gotas creadas se lleva a cabo una PCR, obteniéndose los fragmentos de interés.

Se han llevado a cabo estudios con capturas de un 80% de los fragmentos de interés con coberturas de 25x en más del 90% de ellas. La uniformidad era semejante a las otras técnicas descritas.

# 4. MATERIAL Y MÉTODOS.

## A. MUESTRA

Se utilizó en el estudio la familia mostrada en la figura 4-1. Se seleccionaron los individuos II, III y IV, descartándose para el análisis inicial al individuo I dada la proximidad genética con el individuo III (padre e hijo, compartirían el 50% de las variaciones encontradas). Todos ellos habían sido atendidos en distintos dispositivos de Salud Mental y diagnosticados de TBP tipo I con criterios DSMIV-TR. Se utilizó para la confirmación del diagnóstico y la búsqueda de comorbilidades la entrevista semiestructurada MINI INTERNATIONAL NEUROPSYCHIATRIC INTERVIEW, versión en español 5.0, realizándose posteriormente una entrevista clínica por parte de un psiquiatra adjunto.

Se obtuvo el consentimiento informado por escrito de todos los participantes en el estudio.

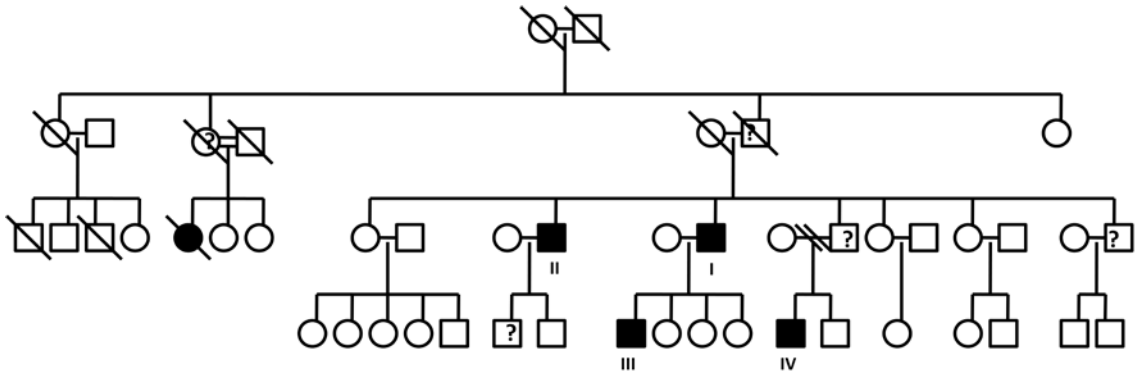



**Figura 4-1. Árbol familiar. ? representa individuos posiblemente afectados.**

**Características clínicas.**

El sujeto número II recibió el diagnóstico de TBP a los 35 años en el contexto de un ingreso hospitalario por un episodio maniaco con síntomas psicóticos. Desde entonces ha tenido otros 3 ingresos, todos ellos por episodios maniacos. En seguimiento en Salud Mental, recibe litio como tratamiento estabilizador.

El sujeto número III debutó a los 27 años, presentando un episodio maniaco con síntomas psicóticos que requirió un ingreso hospitalario de más de dos meses de duración. Asociado al estado de ánimo exaltado presentaba verborrea, pensamiento ideo-fugaz, aumento de actividad y planes e insomnio casi global. Presentaba una ideación delirante de contenido místico-religioso. El año previo había presentado un episodio de menos de una semana de duración de ánimo exaltado, insomnio y sensación de aumento de energía, durante el que había tenido "grandes ideas para inventos", que se autolimitó sin precisar ninguna intervención. En ningún caso hubo consumo de tóxicos.
En tratamiento con ácido valproico desde entonces, continúa en seguimiento en Salud Mental, sin haber requerido nuevos ingresos.

El sujeto número IV presentó a los 22 años un episodio maniaco con síntomas psicóticos que requirió ingreso hospitalario. El ánimo era exaltado-irritable, acompañado de un aumento de actividad y planes, aceleración del pensamiento, e insomnio casi global. Asociaba ideación delirante de contenido mixto, místico-religioso y paranoide. Recibió litio para el control del cuadro. Abandonó el tratamiento unos meses tras el alta hospitalaria. Desde entonces no ha continuado con el seguimiento psiquiátrico ni recibe tratamiento farmacológico. Explorando su evolución, se objetivan claros episodios de sintomatología depresiva e hipomaniaca que en ningún caso han deteriorado el funcionamiento del paciente lo suficiente como para obligarle a consultar de nuevo en Salud Mental ni requerir ingreso psiquiátrico. No hay consumo de tóxicos.

## B. EXTRACCIÓN DEL ADN.

A cada individuo se le extrajeron 12ml de sangre. Las muestras se conservaron en tubos con EDTA (ácido etilendiaminotetraacético) hasta que se procedió a la extracción del ADN. Éste se obtuvo de linfocitos sanguíneos utilizando el QIAamp DNA Blood Maxi Kit (Quiagen, Valencia, California, USA) (http://www.qiagen.com/). Los distintos pasos del proceso se resumen en la figura 4-2 (QIAGEN, 2012).

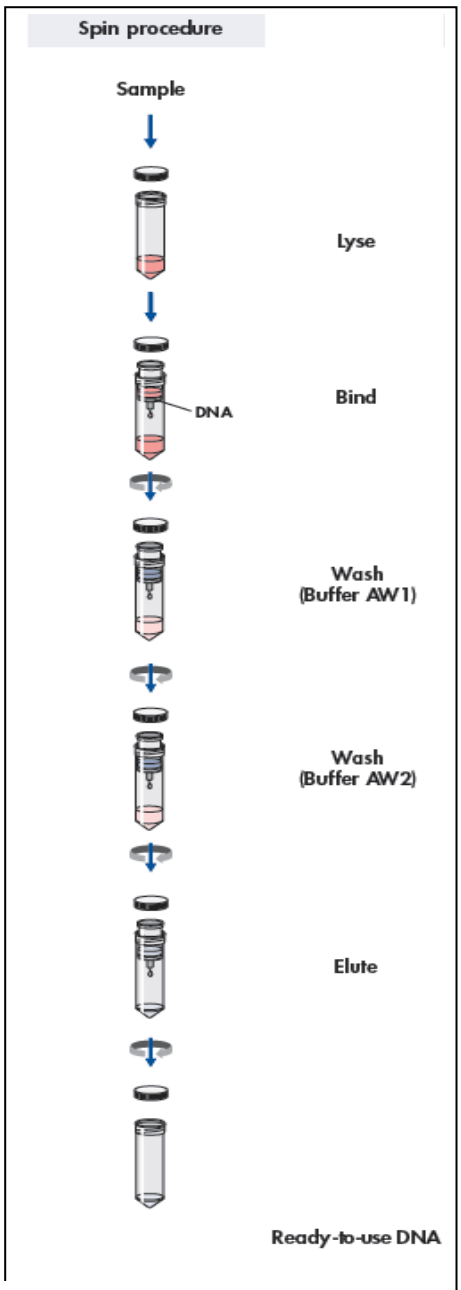



**Figura 4-2.** Purificación de ADN con el QIAamp DNA Blood Maxi kit.

**Lisis.**

Para llevar a cabo la extracción se parte de una muestra de entre 8 y 10ml de sangre total. El primer paso es la lisis de las células sanguíneas para la extracción del material genético. Para ello se incuba la sangre a 70ºC un mínimo de 30' en una solución que rompe las membranas de las células (Buffer AL) junto con una proteasa K para desnaturalizar las proteínas (ver cuadro 4-1). En el paso siguiente se añade etanol a la solución para precipitar el ADN, quedando todo preparado para iniciar la separación del ADN por filtración.

**Filtrado.**

La solución atraviesa una columna que contiene una membrana de sílice, ayudándose de una centrifugación a 3000x. Esta membrana atrapa las moléculas de ADN dejando pasar el resto de sustancias (cromatografía de adsorción). Esto sucede ya que los ácidos nucleicos están cubiertos por una capa de moléculas de agua que los hace solubles en soluciones acuosas. Con la adición de iones caotrópicos (destructores de estructuras

complejas), se destruye esta capa hidratante, creando un entorno hidrofóbico que permite a las moléculas de ADN unirse a la membrana.

Para mejorar la pureza del ADN, la membrana se somete a distintos lavados mediante centrifugación con tampones con etanol.

**Elución.**

Posteriormente, y una vez desechados los productos sobrantes, se separan las moléculas de ADN de la membrana con una centrifugación a 4000x en dos pasos sucesivos, utilizando un tampón ligeramente alcalino (buffer AE) o agua destilada a temperatura ambiente, que permitiría recuperar la capa hidratante de los ácidos nucleicos liberándolos así de la membrana.

De esta forma se obtiene el ADN purificado en fragmentos de hasta 50 kb (la mayoría fragmentos de entre 20 y 30kb) a una concentración de 30-80 µg/ml, obteniéndose finalmente entre 60 y 120 µg de ADN.

**Cuadro 4-1. Relación de productos que incluye el QIAamp DNA Blood Maxi kit.**

| |
|---|
| Proteasa QUIAGEN. Patentada por QUIAGEN. Carece de actividad DNasa o RNasa. |
| Etanol 96-100% |
| Columna de filtrado basada en membrana de sílice. |
| Buffer de lisis (AL) |
| Buffers de lavado AW1 y AW2 |
| Buffer de elución (AE) |

**Control de calidad.**

Antes de enviar la muestra al laboratorio para realizar la secuenciación, se comprueba que su pureza y concentración sean las adecuadas. Para ello se realiza un análisis en un espectrofotómetro de luz UV.

- Concentración. Las distintas bases que conforman el ADN son capaces de absorber la luz UV con una longitud de onda de 260nm. La lectura de la densidad óptica (OD) a 260nm permite calcular la concentración de ácidos nucleicos en la muestra al compararla con valores de referencia.

- Pureza. Se realizan medidas con una longitud de onda de 260 y 280nm. La proporción entre la lectura a 260nm (que refleja la cantidad de ADN) y 280nm (que representa la cantidad de proteínas) proporciona una estimación de la pureza de la muestra. Preparaciones puras tienen un OD260/OD280 de 1.8-2.0. Si hay contaminación los valores obtenidos serán menores.


**Figura 4-3.** Espectrofotómetro.

La compañía Otogenetics Inc., que realiza la secuenciación, realiza también un control de calidad inicial, reflejando la concentración de ADN de cada una de las muestras y su calidad. Para ello realiza, además de una espectrofotometría, una electroforesis, en la que debe aparecer una banda que refleja la existencia en la muestra de ADN de alto peso molecular (figura 4-4).

otogenetics corporation

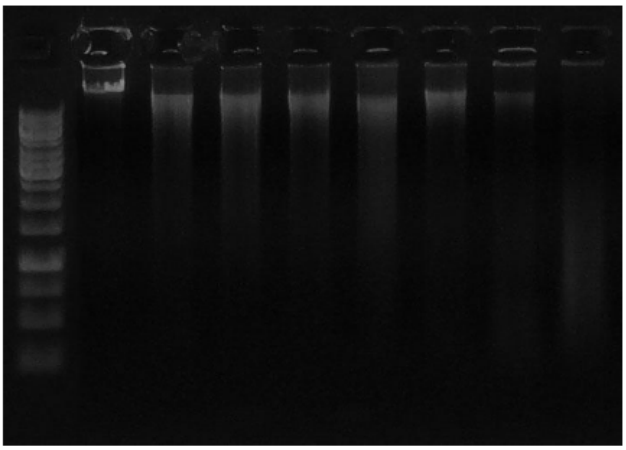
www.otogenetics.com
Toll-free: 1-855-otogene (686-4363)
sales@otogenetics.com

400 Pinnacle Way, Suite#
Norcross, GA 30071
USA

## Sample Initial QC Report

Quote#: 300-919 QC by: SZ&JJ Approved by: NM Date: 12/14/12

| Position In Gel | Storage No. | Sample name | ID on tube | Nanodrop Conc.(ng/ul) | OD 260/280 | Remaining Vol.(ul) | Load Vol.(ul) | QC result |
|---|---|---|---|---|---|---|---|---|
| **Lane1** | Ot6754 | [redacted] | [redacted] | 121.3 | 1.88 | 190 | 1.0 | Pass |
| **Lane2** | Ot6755 | [redacted] | [redacted] | 33.0 | 1.91 | 990 | 3.0 | Pass |
| **Lane3** | Ot6756 | [redacted] | [redacted] | 38.7 | 1.90 | 990 | 3.0 | Pass |
| **Lane4** | Ot6757 | [redacted] | [redacted] | 28.9 | 1.96 | 990 | 3.0 | Pass |
| **Lane5** | Ot6758 | [redacted] | [redacted] | 72.7 | 1.93 | 990 | 1.5 | Pass |
| **Lane6** | Ot6759 | [redacted] | [redacted] | 27.2 | 2.00 | 990 | 3.0 | Pass |
| **Lane7** | Ot6760 | [redacted] | [redacted] | 38.4 | 2.04 | 990 | 2.0 | Pass |
| **Lane8** | Ot6761 | [redacted] | [redacted] | 87.1 | 2.00 | 450 | 1.5 | Marginal* |



Note:* Can not guarantee read depth or library quality.

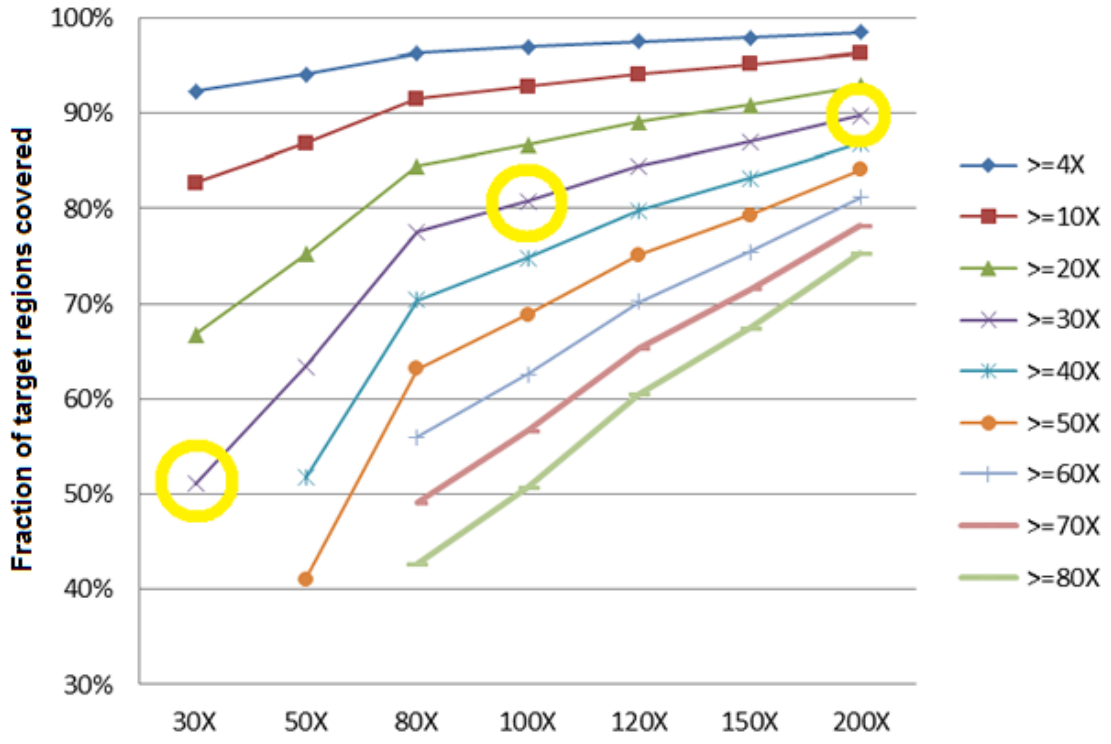
Page **1** of **1**

**Figura 4-4.** Informe de calidad de las muestras. Todas son válidas salvo la situada en la línea 8, en la que no se observa la banda de ADN de alto peso molecular. Se han ocultado aquellos datos que pudieran identificar a los sujetos del estudio.

### C. SECUENCIACIÓN EXÓMICA.

La secuenciación de exoma completo la realizó la compañía Otogenetics Inc., utilizando para la captura y el enriquecimiento del exoma los kits SeqCap EZ Exome de Nimblegen V2.0 y el TruSeq de Illumina. La preparación de la muestra y la secuenciación se llevó a cabo con la plataforma HiSeq2000 de Illumina. Todas estas técnicas se han revisado en el capítulo 3.

Se solicitó a la empresa una cobertura de 50X. Se eligió por su relación calidad-precio. A mayor cobertura el precio es más elevado y, como se puede observar en la gráfica 4-1, con una cobertura media de 50X se obtiene una cobertura mayor o igual a 4X (cobertura mínima aceptable) en el 95% de las secuencias, mientras que en más de un 85% de ellas se obtienen coberturas de 10X (adecuadas), siendo los datos de suficiente calidad para los objetivos que perseguimos en este estudio.




**Gráfica 4-1.** Profundidad de cobertura de las secuencias de interés.

## D. ANÁLISIS DE LOS DATOS.

Para llevar a cabo en análisis de los datos se utilizaron las herramientas informáticas recogidas en el cuadro 4-2. El proceso se resume en la figura 4-5.

**Cuadro 4-2. Herramientas informáticas utilizadas para el análisis de los datos.**

| |
|---|
| **Hardware:** |
| PC de sobremesa Intel Core i7 2600 CPU 3,40 GHz x8. 64 bits. 16 Gigas RAM. |
| **Software:** |
| Sistema operativo: Linux Ubuntu 12 04 LTS. |
| **Programas:** |
| -FastQC v0.10.1 y bedtools. |
| -Picard v1.7. |
| -GATK v1.6. |
| -ANNOVAR v2013Feb21. |
| -IGV v2.2.5. |
| -KGGSeq. |

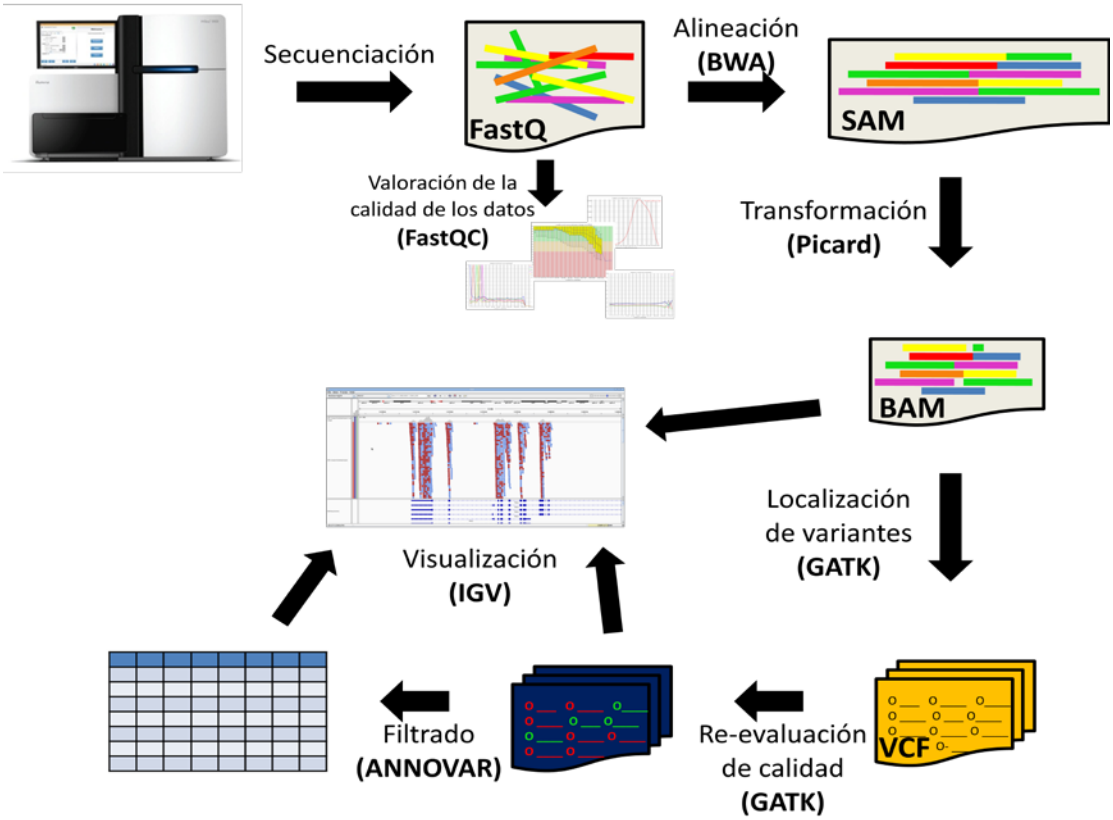



**Figura 4-5.** Simplificación gráfica del proceso de análisis.

**Control de la calidad de la descarga.**

El laboratorio que realiza la secuenciación cuelga los datos en internet a disposición del usuario que, a través de la introducción de una clave, se los descarga en su terminal. Para comprobar que la descarga está completa, utilizamos el comando `md5sum –c md5.txt` en Linux, comprobando que el número que obtenemos y el que consta en el archivo md5.txt de la descarga son el mismo.

**Comprobación de la calidad de los datos.**

El formato en el que se reciben desde el laboratorio las lecturas de la secuenciación se conoce como FastQ (figura 4-6). De cada sujeto se reciben 2 archivos, ya que de cada fragmento de la librería se realizan lecturas desde ambos extremos (*pair ends reads*). En este tipo de archivos, la calidad de las lecturas se expresa en base logarítmica. Así, una calidad Q20 significa una posibilidad de 0.01 de haber leído una base erróneamente. Q30 traduciría una posibilidad de error de 0.001 y así sucesivamente.

```
@EAS100R:136:FC706VJ:2:2104:15343:197393 1:Y:18:ATCAG
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

**Figura 4-6.** Representación de una lectura del secuenciador en formato FastQ. La primera línea muestra tras la @ diversa información acerca de la máquina, el ciclo o la posición de la lectura. La segunda línea es la secuencia leída. La última línea informa de la calidad de la lectura en formato ASCII (la probabilidad de haber realizado una lectura errónea por cada base o error estimado por base).

Los parámetros de calidad de los datos se exploraron con el programa FastQC (www.bioinformatics.babraham.ac.uk/projects/), que informa de los siguientes parámetros:

- Estadísticas básicas. Tabla con el resumen de los datos (figura 4-7). Contiene:
    - Nombre del archivo.
    - Tipo de archivo. Refiere si el archivo contiene lecturas de bases (A, T, C, G) o datos de color que posteriormente deben transformarse en bases.

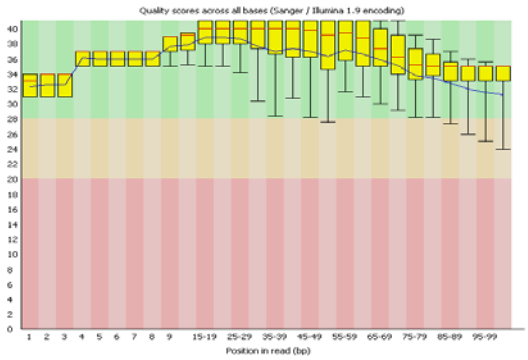
- Codificación. Describe la codificación ASCII de calidad del archivo.
- Número total de secuencias procesadas (real y estimado). Se pueden analizar la totalidad de los datos o seleccionar parte.
- Secuencias filtradas. Indica el número de secuencias que se han eliminado si se ha elegido el modo de filtrado.
- Longitud de las secuencias. Muestra las longitudes de la secuencia más corta y la más larga. Si todas miden lo mismo sólo mostrará un valor.
- El % de GC total en los datos. Las zonas ricas en GC corresponden a áreas de codificación (exomas).

| Basic sequence stats | |
|---|---|
| Measure | Value |
| Filename | 141.fq.gz |
| File type | Conventional base calls |
| Encoding | Sanger / Illumina 1.9 |
| Total Sequences | 28170365 |
| Filtered Sequences | 0 |
| Sequence length | 100 |
| %GC | 45 |

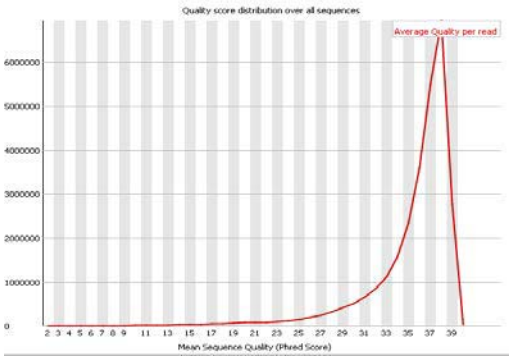
**Figura 4-7.** Estadísticas básicas en el programa FastQC.

- Calidad de la secuencia por base. Gráfica que muestra la calidad de las bases según la posición de lectura. Los datos son buenos si la calidad de las lecturas en las distintas posiciones es alta y homogénea (las barras amarillas, que muestran el rango intercuartílico 25-75%, son pequeñas, de tamaño semejante en todas las posiciones de lectura y de alta puntuación). Son malos datos si la calidad de las lecturas largas es mala, observándose un aumento de variabilidad en las lecturas de las últimas posiciones. La línea roja representa el valor de la mediana; la línea azul la calidad media (gráfica 4-2).

- Puntuaciones de calidad por secuencia. Gráfica que representa la distribución de la calidad de las lecturas. Los datos serán mejores cuantas más lecturas tengan de buena calidad (gráfica 4-3). Si una gran cantidad de secuencias en un ciclo determinado son de baja calidad, puede deberse a un problema sistemático del proceso de secuenciación.



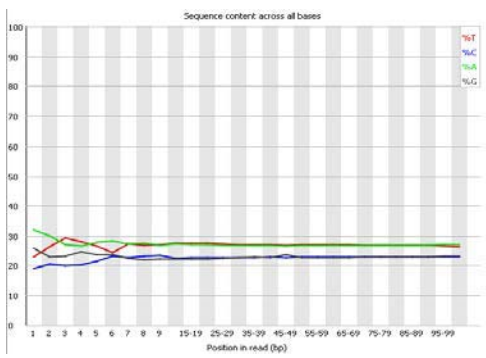

**Gráfica 4-2.** Calidad de la secuencia por base. Ejemplo de datos de buena calidad.



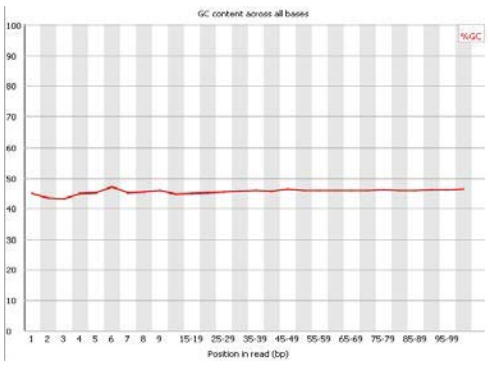

**Gráfica 4-3.** Distribución de la calidad de las lecturas.

- Contenido de bases en la secuencia. Muestra la frecuencia de cada base en cada posición de lectura (gráfica 4-4). En el caso de los datos obtenidos del secuenciador Illumina, que utiliza bibliotecas de fragmentos realizados de forma aleatoria, no deberían aparecer diferencias significativas en las frecuencias de las distintas bases en cada posición de lectura. Si se observaran claras diferencias, podría indicar la presencia de una secuencia sobrerrepresentada contaminando la biblioteca. El programa muestra un error cuando la diferencia entre A y T o G y C es mayor del 20% en alguna de las posiciones de lectura.

- Contenido de GC por base. Muestra la distribución de GC por posición de lectura (gráfica 4-5). En una biblioteca aleatoria debería ser homogéneo, siendo la línea roja prácticamente horizontal. La aparición de desviaciones puede deberse a la presencia de alguna secuencia sobrerrepresentada en la biblioteca.



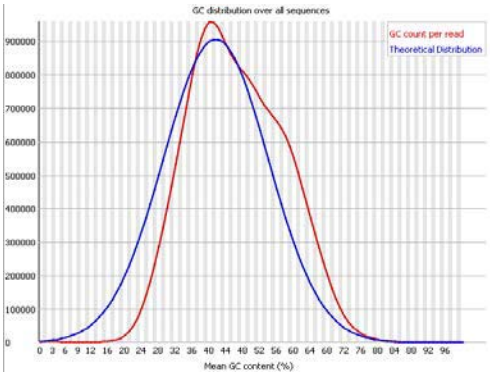

**Gráfica 4-4.** Frecuencia de cada base por posición de lectura.



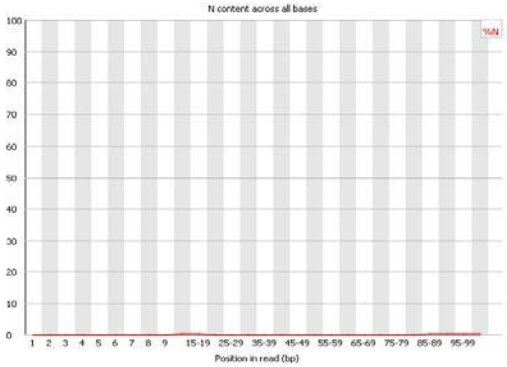

**Gráfica 4-5.** Contenido de GC por posición de lectura.

- Contenido de GC por secuencia. Muestra la distribución de GC por secuencia. La línea azul muestra la distribución esperada, y la roja la real. Cuanto más parecidas sean ambas, de mejor calidad son los datos (gráfica 4-6).

- Contenido de Ns por base. Se utiliza N para marcar las bases de las que no se ha obtenido lectura. Cuantas más Ns contengan los datos su calidad será menor (gráfica 4-7).



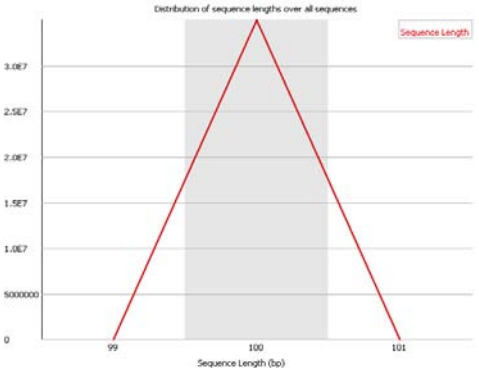

**Gráfica 4-6.** Distribución del contenido de GC.



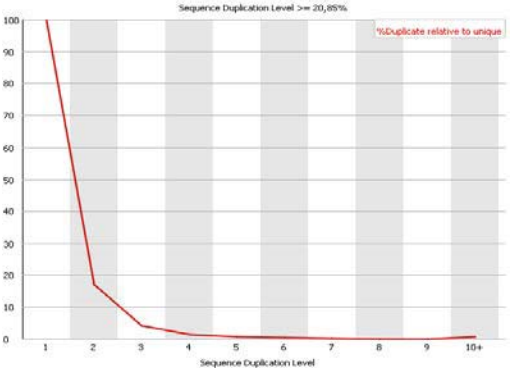

**Gráfica 4-7.** Cantidad de Ns por posición de lectura.

- Distribución de la longitud de secuencias. La gráfica muestra las longitudes de las lecturas en pares de bases (bp) (gráfica 4-8).

- Niveles de duplicación de secuencias. Muestra el grado de duplicación de cada secuencia. Cuanto menor es este número, mejor es la calidad de los datos, y suele traducir un adecuado nivel de cobertura de las secuencias diana (gráfica 4-9).



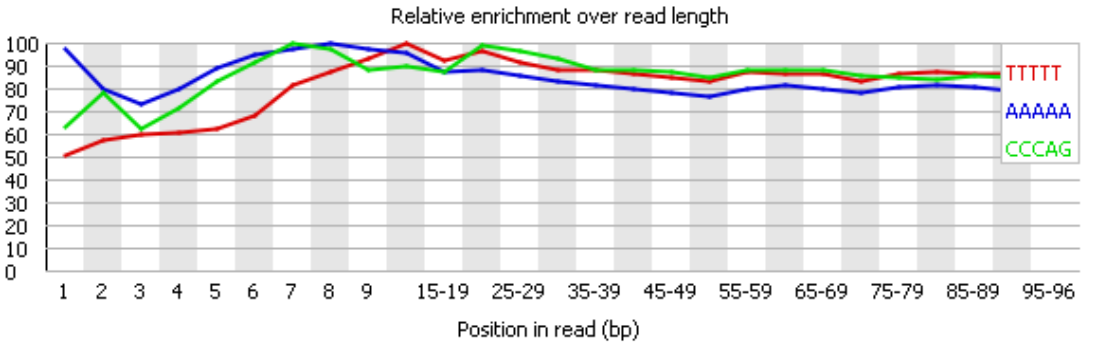

**Gráfica 4-8.** Distribución de la longitud de las secuencias.




**Gráfica 4-9.** Distribución de lecturas duplicadas.

- Secuencias sobrerrepresentadas. Informa de la presencia de secuencias que aparecen en mayor cantidad de la esperada. Además, evalúa cada una de estas secuencias y sugiere su posible significado (habitualmente secuencias que contaminan la biblioteca). En el caso de la plataforma Illumina suelen corresponder a las secuencias de los cebadores utilizados para la PCR.

- Contenido en K-mer. K-mer se refiere a oligómeros específicos (de longitud k) que se pueden utilizar para localizar regiones de interés. Esta gráfica muestra la cantidad de veces que esas secuencias aparecen en la muestra. Por defecto, muestra los pentámeros que aparecen en mayor frecuencia de la esperada. Puede identificar problemas como, por ejemplo, si se encuentran las secuencias que corresponden a los adaptadores en posiciones intermedias de las lecturas. Igualmente, puede mostrar si el enriquecimiento de todas las secuencias ha sido homogéneo o algunas están sobre-enriquecidas (gráfica 4-10).




| Sequence | Count | Obs/Exp Overall | Obs/Exp Max | Max Obs/Exp Position |
|---|---|---|---|---|
| TTTTT | 14413220 | 3,176 | 3,65 | 10-14 |
| AAAAA | 14999545 | 3,151 | 3,777 | 8 |
| CCCAG | 7796865 | 3,045 | 3,447 | 7 |

**Gráfica 4-10.** Este caso muestra un sobre-enriquecimiento de algo más de 3 veces lo esperado de los pentámeros AAAA, TTTTT y CCCAG.

Para comprobar la cobertura se utilizó el programa bedtools (http://bedtools.readthdocs.org), y para representarla gráficamente se utilizaron el script que se especifica en el Anexo II y el programa estadístico gratuito R (http://cran.r-project.org/).

**Alineación de las lecturas con un genoma de referencia.**

El secuenciador Illumina produce millones de lecturas cortas que posteriormente hay que alinear con un genoma de referencia para obtener la secuencia de ADN original. Existen varios algoritmos de alineación (MAQ, BWA, Bowtie). En este caso se ha utilizado el algoritmo BWA (*Burrows-Wheeler Aligner*), por su rapidez, precisión y acceso gratuito (http://bio-bwa.sourceforge.net/). El genoma de referencia se descargó del UCSC Genome Browser website con la orden `bwa index -a bwtsw -p hg19M/hg19M hg19M/hg19M.fa` en Linux.

**Transformación del archivo a formato BAM.**

El alineamiento de los dos archivos .fq con BWA genera dos archivos en formato .sai, que transformaremos en un único archivo en formato .sam (*Sequence Alignment/Map*), y finalmente, utilizando el programa Picard (http://picard.sourceforge.net), en un formato .bam (forma binaria de los archivos .sam), que permite utilizar estos datos en otros programas y se está confirmando como estándar en los estudios de secuenciación.

Se ha optado por utilizar la opción LENIENT para VALIDATION_STRINGENCY con el objetivo de que el programa no se pare si encuentra un error en los datos, pero lo marque como advertencia. Utilizamos la herramienta `MarkDuplicates` para localizar las secuencias duplicadas.

Los archivos .marked.bam y metrics resultantes contienen todas las lecturas e identifican de qué tipo son, incluyendo:

- Lecturas no alineadas (marcadas como *unmapped*).
- Lecturas duplicadas (marcadas como *duplicates*).
- Lecturas "*non-PF*", aquellas que no pasan el filtro de calidad (Li et al., 2009).

**Localización de las variantes presentes en los sujetos del estudio.**

El siguiente paso a realizar es la identificación de todas las variaciones presentes en cada exoma. Para ello se ha utilizado el programa informático Genome Analysis Tool Kit o GATK (desarrollado por el Broad Institute, Cambridge,

Massachusetts http://www.broadinstitute.org/gatk/), disponible de forma libre en la red.

El programa realiza los siguientes pasos:

1. Mapeo inicial.

Los resultados del alineamiento se encuentran en el archivo .bam. En este primer paso, en el que se utiliza la herramienta `RealignerTargetCreator`, se identifican regiones que requieren un re-alineamiento, habitualmente por presencia de indeles que no existen en el genoma de referencia. El archivo resultante lo denominamos .bam.list.

2. Realineamiento alrededor de los indeles.

El programa re-alinea las lecturas alrededor de estos indeles, lo que minimiza el riesgo de falsos positivos al buscar posteriormente variantes. Se lleva a cabo con la herramienta `IndelRealigner`. El archivo resultante lo denominamos .marked.realigned.bam. El archivo .bam.list recoge las áreas de realineamiento.

3. Mejora de las lecturas iniciales.

El error estimado por base (puntuación de calidad de base) es en lo que se basan todos los algoritmos estadísticos de localización de variaciones. Habitualmente, las puntuaciones asignadas por los secuenciadores son inexactas, por lo que previo a la búsqueda de las variaciones se realiza la recalibración de sus valores de calidad utilizando covariantes específicas y enmascarando los SNPs ya conocidos (los que aparecen en las bases de datos públicas como la dbSNP). El proceso se muestra en la figura 4-6. La dos covariantes que se utilizan por defecto en la recalibración son: Puntuación de Calidad (*QualityScoreCovariate*) y Grupo de lectura (*ReadGroupCovariate*). Otras (Dinucleótido (*DinucCovariate*), Homopolímero, Ciclo de lectura (*CycleCovariate*), o Contexto del nucleótido) son opcionales. Para este paso se utilizan las herramientas `CountCovariates`, que genera una tabla con las covariantes seleccionadas, y `TableRecalibration`, que recalibra las puntuaciones (figura 4-8).

Da como resultado un archivo BAM recalibrado (.recal.bam) y las siguientes tablas de datos en formato .csv:

- Lista de variaciones encontradas.
- Tabla de calidades cuantificadas.
- Tabla de recalibración por grupo de lectura.
- Tabla de recalibración por puntuación de calidad.
- Tabla de cuantificación con las covariantes opcionales.

El formato .csv se puede pasar a Excel para facilitar su lectura (figura 4-9).

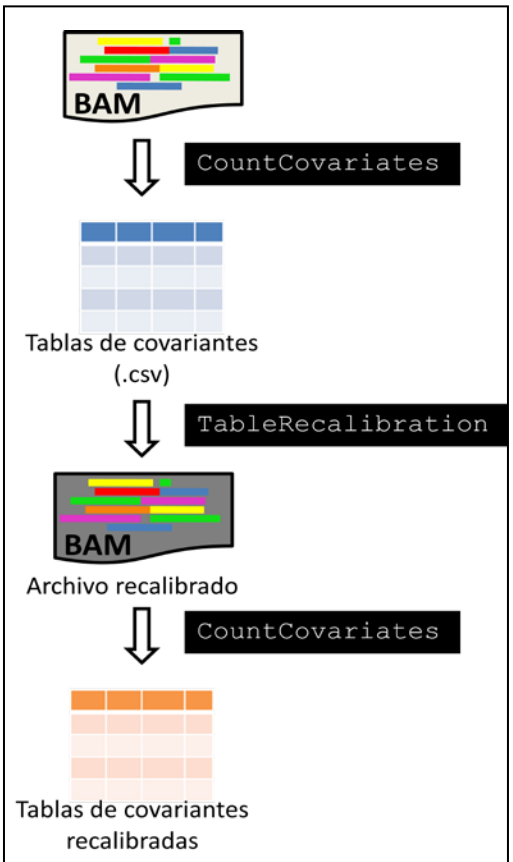



**Figura 4-8.** Proceso de recalibración de los datos iniciales. Se obtienen el archivo BAM recalibrado y las tablas con las covariantes.

| | A | B | C | D | E | F | G | H | I | J | K | L |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | ##fileformat=VCFv4.1 | | | | | | | | | | | |
| 2 | #CHROM | POS | ID | REF | ALT | QUAL | FILTER | INFO | | | | |
| 3 | chr1 | 787399 | . | N | <VQSR> | . | . | END=787399;VQSLOD=4.2560;culprit=QD | | | | |
| 4 | chr1 | 879676 | . | N | <VQSR> | . | . | END=879676;VQSLOD=1.5797;culprit=FS | | | | |
| 5 | chr1 | 881627 | . | N | <VQSR> | . | . | END=881627;VQSLOD=1.6795;culprit=ReadPosRankSum | | | | |
| 6 | chr1 | 883625 | . | N | <VQSR> | . | . | END=883625;VQSLOD=5.7015;culprit=FS | | | | |
| 7 | chr1 | 887801 | . | N | <VQSR> | . | . | END=887801;VQSLOD=6.4644;culprit=FS | | | | |
| 8 | chr1 | 888639 | . | N | <VQSR> | . | . | END=888639;VQSLOD=-0.1246;culprit=MQ | | | | |
| 9 | chr1 | 888659 | . | N | <VQSR> | . | . | END=888659;VQSLOD=-0.0734;culprit=FS | | | | |
| 10 | chr1 | 889158 | . | N | <VQSR> | . | . | END=889158;VQSLOD=4.9954;culprit=FS | | | | |
| 11 | chr1 | 889159 | . | N | <VQSR> | . | . | END=889159;VQSLOD=5.6730;culprit=FS | | | | |
| 12 | chr1 | 897325 | . | N | <VQSR> | . | . | END=897325;VQSLOD=6.1049;culprit=FS | | | | |
| 13 | chr1 | 900505 | . | N | <VQSR> | . | . | END=900505;VQSLOD=2.8746;culprit=QD | | | | |
| 14 | chr1 | 900730 | . | N | <VQSR> | . | . | END=900730;VQSLOD=-0.7701;culprit=MQ | | | | |
| 15 | chr1 | 909238 | . | N | <VQSR> | . | . | END=909238;VQSLOD=6.1700;culprit=FS | | | | |
| 16 | chr1 | 909309 | . | N | <VQSR> | . | . | END=909309;VQSLOD=4.9366;culprit=FS | | | | |
| 17 | chr1 | 910438 | . | N | <VQSR> | . | . | END=910438;VQSLOD=4.3081;culprit=QD | | | | |
| 18 | chr1 | 915227 | . | N | <VQSR> | . | . | END=915227;VQSLOD=6.3432;culprit=FS | | | | |
| 19 | chr1 | 948870 | . | N | <VQSR> | . | . | END=948870;VQSLOD=5.8225;culprit=FS | | | | |
| 20 | chr1 | 948921 | . | N | <VQSR> | . | . | END=948921;VQSLOD=5.5373;culprit=FS | | | | |
| 21 | chr1 | 949608 | . | N | <VQSR> | . | . | END=949608;VQSLOD=1.5185;culprit=MQRankSum | | | | |
| 22 | chr1 | 949654 | . | N | <VQSR> | . | . | END=949654;VQSLOD=6.0026;culprit=FS | | | | |
| 23 | chr1 | 949925 | . | N | <VQSR> | . | . | END=949925;VQSLOD=4.7833;culprit=FS | | | | |
| 24 | chr1 | 977330 | . | N | <VQSR> | . | . | END=977330;VQSLOD=0.9851;culprit=FS | | | | |
| 25 | chr1 | 981931 | . | N | <VQSR> | . | . | END=981931;VQSLOD=3.8563;culprit=FS | | | | |
| 26 | chr1 | 982941 | . | N | <VQSR> | . | . | END=982941;VQSLOD=5.6179;culprit=FS | | | | |

**Figura 4-9**. Ejemplo de tabla de recalibración.

4. Localización de variaciones.

Todas las variaciones presentes en cada individuo se extraen del archivo anterior utilizando la herramienta Unified Genotyper. Utiliza un archivo de referencia como comparación (en este caso hemos utilizado el dbSNP versión 135) y da lugar a un archivo .vcf (*Variant Call Format*).

Permite seleccionar el valor de varios parámetros:

`-stand_call_conf.` Umbral mínimo de calidad a partir del cual las variantes se consideran de alta calidad y se tienen en cuenta. En este estudio se ha seleccionado una Q de 50.0, es decir, se seleccionan como válidas sólo aquellas variantes con una posibilidad de que sean un error de $10^{-5}$. El valor por defecto del programa es de 30.0.

`-stand_emit_conf.` Umbral de calidad a partir del cual las variantes se recogen en el informe, aunque marcan como de baja calidad (LowQual). En este caso se ha seleccionado un valor de 10.

`-dcov` (`downsample_to_coverage`). Selecciona de forma aleatoria exclusivamente una cantidad determinada de lecturas de cada zona alineada, lo que limita problemas en áreas de excesiva cobertura. En este caso limitamos el número de lecturas a 1000.

5. Recalibración de la puntuación de la calidad de las variantes.

Tras la localización de las variantes, el programa genera una tabla con la puntuación de calidad de cada una de ellas. Para ello compara los resultados con las bases de datos HapMap 3 y dbSNP, y las áreas polimórficas presentes en el chip Omni del Proyecto de los 1000 genomas. Estas bases de datos codifican sus variantes como:

- Sitios conocidos. El estatus de variante conocida o nueva no es utilizado por el algoritmo, siendo utilizado exclusivamente con un propósito informativo.
- Sitios de entrenamiento. Las variantes de entrada que se superponen a esos sitios de entrenamiento se utilizan para construir el modelo Gaussiano.
- Sitios verdaderos. Se utiliza para decidir el corte de sensibilidad de la VQSLOD en estos sitios, generalmente con la idea de recabar el 99% de los sitios de HapMap, por ejemplo.

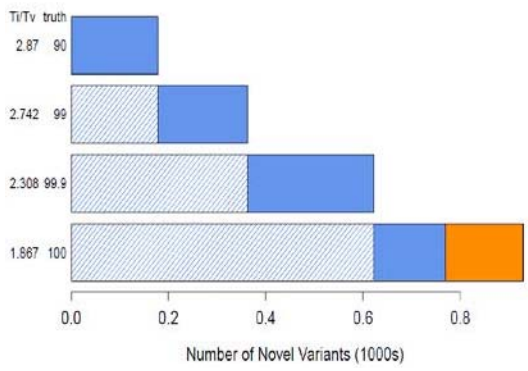
Igualmente, la probabilidad de esos sitios de ser verdaderas variaciones se especifica en escala Phred.

Las variaciones no presentes en esas bases son evaluadas y se les asigna un log OR (VQSLOD), clasificándolas como verdadero vs. falso positivo. El objetivo es permitir al investigador diferenciar si cada variación encontrada obedece a una mutación auténtica o no es más que un error de secuenciación o de procesamiento de los datos.

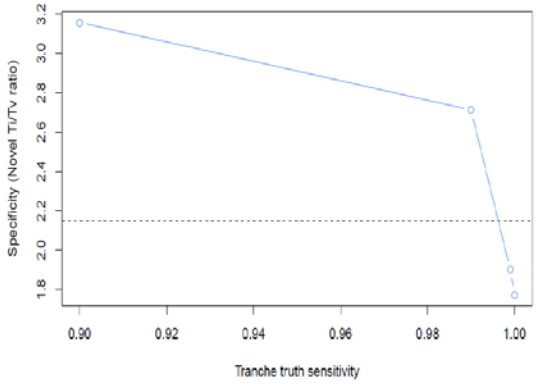
Para ello se utilizan las herramientas `VariantRecalibration` (genera el modelo Gaussiano para evaluar la calidad de las variantes) y `ApplyRecalibration` (que aplica el modelo generado a las variantes halladas), y se obtiene un archivo snp.vcf.recalibrated. En este archivo constan las variantes encontradas junto con su VQSLOD score). Especifica además de cada una de ellas numerosas características (tabla 4-1).

Para una revisión de los argumentos específicos que puede utilizar esta aplicación, visitar las páginas web:
http://www.broadinstitute.org/gatk/gatkdocs/org_broadinstitute_sting_gatk_walkers_variantrecalibration_VariantRecalibrator.html
http://www.broadinstitute.org/gatk/gatkdocs/org_broadinstitute_sting_gatk_walkers_variantrecalibration_ApplyRecalibration.html

Unos de los archivos generados por la recalibración, el .tranches, contiene gráficas que muestran datos de sensibilidad y especificidad respecto a la localización de las variantes (gráficas 4.11 y 4.12).




**Gráfica 4.11.** Verdaderos y falsos positivos en función de la sensibilidad.




**Gráfica 4.12.** Relación sensibilidad/especificidad.

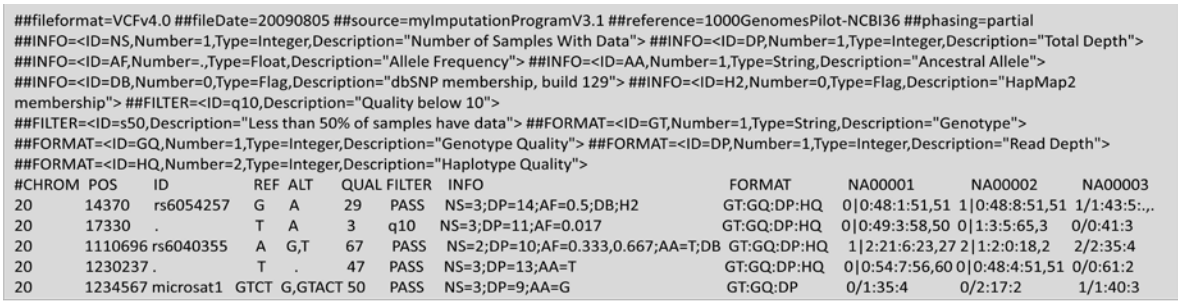
6. Filtrado de las variantes encontradas.

Permite la eliminación de lo que se pueden considerar falsos positivos (defectos de la técnica) filtrando según una serie de parámetros:

- `clusterWindowSize` o filtro de agrupación de SNPs. Si se encuentran 3 o más variaciones en un grupo de X pares de bases (X es el número de bases que se especifique), el programa lo marca como SnpCluster, y habitualmente corresponden a falsos positivos. En este caso se ha mantenido el valor por defecto, que es 10 pares de bases.
- `"HARD_TO_VALIDATE"` o variaciones "difíciles de validar". Si un fragmento contiene una secuencia con una MQ (*MappingQuality*) de 0 (es decir, que se podría alinear en varios sitios distintos con la misma concordancia) y el resto del fragmento muestra un 10% de no concordancia, es difícil valorar si es o no un artefacto.
- `"LowCoverage"` o filtro de baja cobertura. Filtra variaciones con coberturas menores de un número a seleccionar de lecturas, ya que son potenciales artefactos. Mantenemos el valor por defecto, de 5 lecturas.
- `"VeryLowQual"` o filtro de muy baja calidad. Elimina aquellas variaciones con una puntuación de calidad de menos de 30, que suelen ser artefactos.
- `"LowQual"` o filtro de baja calidad. Elimina aquellas variaciones con puntuaciones de calidad entre 30 y 50, que pueden ser artefactos.
- `"LowQD"` o filtro de baja QD (confianza de la variante/profundidad no filtrada). Puntuaciones bajas del parámetro QD suelen representar falsos positivos. En este caso, eliminamos las variantes con puntuaciones QD por debajo de 1.5.
- `"StrandBias"`. Variaciones que solo aparecen en las lecturas de la misma dirección son habitualmente artefactos, por lo que se filtran.

Igualmente, el programa nomina como AF 1.00 las mutaciones homocigotas y como 0.50 las heterocigotas.

El archivo VCF definitivo contiene las variantes de interés junto con información sobre cada una de ellas (figura 4-10). Los campos que contienen estos archivos se explican en la tabla 4-1.



```
##fileformat=VCFv4.0 ##fileDate=20090805 ##source=myImputationProgramV3.1 ##reference=1000GenomesPilot-NCBI36 ##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data"> ##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=.,Type=Float,Description="Allele Frequency"> ##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129"> ##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2
membership"> ##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data"> ##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality"> ##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM  POS      ID          REF  ALT     QUAL FILTER  INFO                              FORMAT      NA00001        NA00002        NA00003
20      14370    rs6054257   G    A       29   PASS    NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:.,.
20      17330    .           T    A       3    q10     NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3   0/0:41:3
20      1110696  rs6040355   A    G,T     67   PASS    NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2  2/2:35:4
20      1230237  .           T    .       47   PASS    NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 0/0:61:2
20      1234567  microsat1   GTCT G,GTACT 50   PASS    NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
```

**Figura 4-10.** Archivo VCF.

**Tabla 4-1. Campos del archivo VCF.**

| Nombre | Significado |
|---|---|
| **CHROM** | Cromosoma. |
| **POS** | Posición de referencia. |
| **ID** | Identificador en la base dbSNP (si está presente). |
| **REF** | Base de referencia. En las inserciones, señala la base previa. |
| **ALT** | Variación encontrada (alelo alternativo). |
| **QUAL** | Puntuación de calidad de la variación ALT. |
| **FILTER** | Filtros. PASS indica que ha pasado todos los filtros; si no pasa alguno, lo muestra. |
| **INFO** | Información adicional. Tiene muchas posibilidades:<br>AA. Alelo ancestral.<br>AF. Frecuencia del alelo.<br>DB. Descrito en la dbSNP.<br>DP. Profundidad de cobertura combinada de todas la muestras.<br>END. Posición final de la variante descrita (para CNVs).<br>H2. Presente en la base hapmap2.<br>MQ (*MappingQuality*). Posibilidad de una secuencia determinada de alinear en varios sitios distintos con la misma concordancia. Una puntuación de 0 supone que se podría alinear en muchos lugares distintos.<br>MQ RMS. (*MappingQualityRankSumTest*). Aproximación z de la U de Mann-Whitney del Rank Sum Test de las calidades de mapeo (lecturas con la base de referencia vs. lecturas con el alelo alternativo).<br>MQRankSum ReadPosRankSum. Aproximación z de la U de Mann-Whitney del Rank Sum Test para la distancia desde el final de la lectura para las lecturas con el alelo alternativo. Si un alelo sólo aparece cerca de los finales de las lecturas, es más probable que sea un error.<br>NS. Número de muestras con datos de esa variación.<br>SB (StranBias). Sesgo de cadena en esta posición (variaciones que sólo se observan en una dirección de la lectura). Un mayor sesgo indica falsos positivos.<br>FS (FisherStrand). Valor de la p con el test de Fisher para detectar sesgos de cadena.<br>SOMATIC. Indica que es una mutación somática (genómica del cáncer). |
| **FORMAT** | GT. Genotipo. Muestra 2 valores separados por una barra (alelos diploides). Para los alelos de los cromosomas X o Y sólo se da un valor.<br>• 0 para el alelo de referencia.<br>• 1 para el primer alelo en la lista ALT (alelo alternativo).<br>• 2 para el segundo alelo en la lista. Y así sucesivamente.<br>GQ. Calidad del genotipo (también en base logarítmica).<br>DP. Profundidad de lectura en esa posición en esa muestra.<br>HaplotypeScore. Consistencia del sitio con dos haplotipos segregados. Valores altos reflejan regiones mal alineadas, generalmente por indeles u otros artefactos.<br>QD (QualByDepth). Confianza de la variante (según la puntuación de calidad)/profundidad no filtrada. Valores bajos reflejan falsos positivos.<br>FT. Resultado de la aplicación de los distintos filtros al genotipo. |
| **NA** | Datos FORMAT para cada una de las muestras, con su número de identificación. |

**Filtrado de las variantes encontradas.**

Como se describió en el capítulo 1, cuando se compara cada exoma individual con un genoma de referencia se encuentran entre 15000 y 25000 variaciones. Muchas de ellas son SNPs en secuencias polimórficas habituales en la población, sin significado patológico. Otras variaciones (SNPs no frecuentes, variaciones estructurales, inserciones o delecciones) son de significado incierto. Determinar cuáles son las responsables de un fenotipo concreto (el TBP, en el caso que nos ocupa) representa una tarea compleja. Para ello se utilizan distintos filtros (Tucker et al., 2009; Ng et al., 2009; Stitziel, Kiezun, & Sunyaev, 2011; Robinson et al., 2011b; Jimenez-Escrig, Gobernado, & Sanchez-Herranz, 2012), que se expondrán a continuación.

Para facilitar el filtrado se ha utilizado el programa ANNOVAR (www.openbioinformatics.org/annovar/). El archivo resultante se puede leer en formato excel e incluye los siguientes datos, entre otros, de cada variación encontrada (figura 4-11): Función (exón o *splicing*), gen, función del exón, cambio de amino-ácido, puntuación de conservación, duplicaciones de segmentos, frecuencia en la que el alelo aparece en la base de datos de los 1000 genomas, referencia dbSNP y puntuaciones de predicción (según programas como el PolyPhen y el AVSIFT).

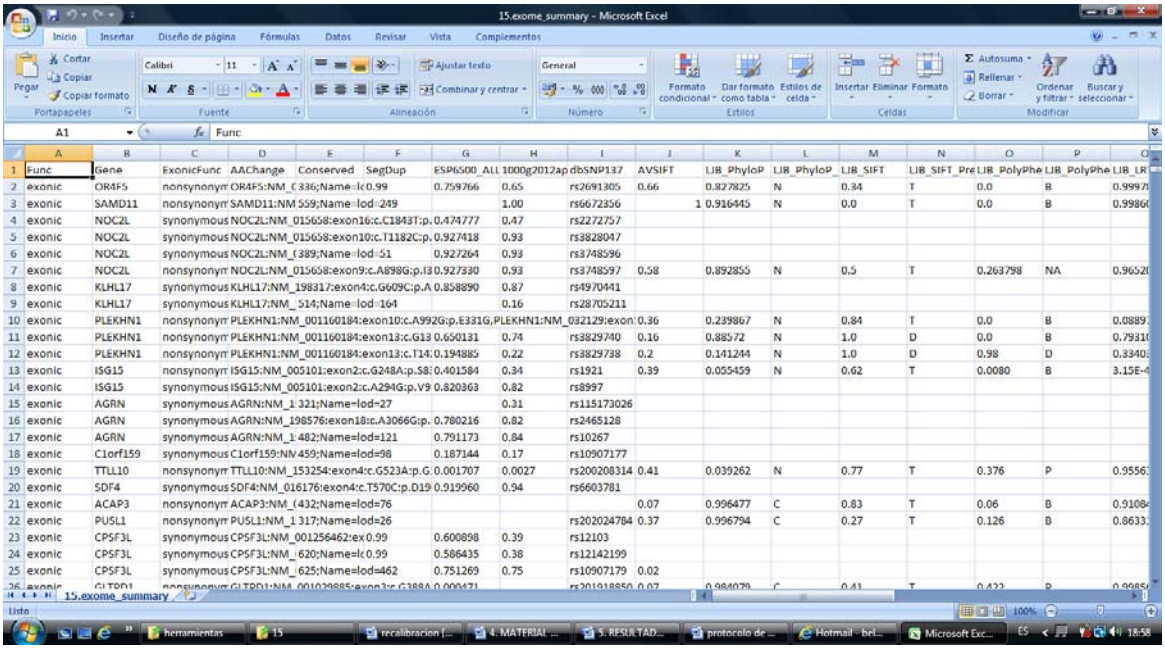



**Figura 4-11.** Archivo resultante del programa ANNOVAR abierto con Excel.

1. Filtrado por variantes raras.

Las enfermedades hereditarias son poco frecuentes, por lo que se espera que las mutaciones que las causan sean igualmente infrecuentes en la población. Por tanto, se pueden descartar con cierta seguridad aquellas presentes en las bases de datos de los 1000 genomas y de los 6500 exomas (Via, Gignoux, & Burchard, 2010; Mu, Lu, Kong, Lam, & Gerstein, 2011). Este filtro puede eliminar hasta el 90% de las variaciones. Sin embargo, dado que la información fenotípica en esas bases de datos es incompleta, y asumiendo que algún individuo presente en ellas pudiera padecer la enfermedad y no haberse detectado o manifestado por ser una patología de inicio tardío, se puede elegir filtrar por variaciones que aparecen en menos de un determinado tanto por ciento de los individuos.

2. Filtrado basado en la función.

Las mutaciones en áreas codificantes o de splicing tienen más probabilidades de ser patogénicas que las situadas en zonas no codificantes. No obstante, se sabe que mutaciones en áreas intrónicas reguladoras pueden determinar patologías. Igualmente, las mutaciones sinónimas son menos frecuentemente patogénicas que aquellas que modifican la estructura de la proteína (*missense*, *nonsense*, inserciones/delecciones).

3. Filtrado por predicción.

Se puede asumir que áreas del ADN que están conservadas en distintas especies tienen funciones de importancia, por lo que mutaciones en estos loci tienen más probabilidades de ser patogénicas. La puntuación se realiza con herramientas como AVSIFT, CDPred, PholyPhen o GERP. Este filtro debe utilizarse con cautela, nunca como dato exclusivo (Cooper et al., 2010).

4. Eliminación de los segmentos duplicados y selección de aquellas mutaciones marcadas como PASS.

Se descartan mutaciones presentes en segmentos duplicados, con grandes posibilidades de ser falsos positivos, y se seleccionan aquellas que el programa GATK

marca como PASS, es decir, de buena calidad, y que con seguridad son verdaderas variaciones.

5. Comparación de los exomas de estudio.

El programa ANNOVAR contiene una función para comparar los exomas seleccionados entre sí y mostrar sólo aquellas variaciones presentes en todos ellos.

6. Filtro manual.

Se descartan de forma manual algunos genes, ya sea porque su función no es compatible en absoluto con la enfermedad o porque son genes con alta variabilidad (lo que se comprueba también en controles) debido, por ejemplo, a una gran longitud, o asociados a pseudogenes que alteran el resultado de la secuenciación o la alineación.

7. Otros.

-Filtrado por mutaciones en homocigosis en el caso de enfermedades de herencia recesiva. Para ellos se utiliza el valor AF descrito previamente.

-Selección de áreas de interés. Puedes ser regiones señaladas por estudios de ligamiento (Smith et al., 2011c), o utilizando información familiar. Si, por ejemplo, dos hermanos están afectados por una misma enfermedad, se buscará la mutación sólo en aquellas áreas del genoma que comparten.

**Filtrado con la plataforma KGGSeq.**

Para comprobar los resultados se realizó un segundo análisis de las variantes utilizando la plataforma KGGSeq (Li, Gui, Kwan, Bao, & Sham, 2012), conjunto de herramientas informáticas y estadísticas destinadas a filtrar y priorizar variaciones del exoma en las enfermedades hereditarias (figura 4-12).

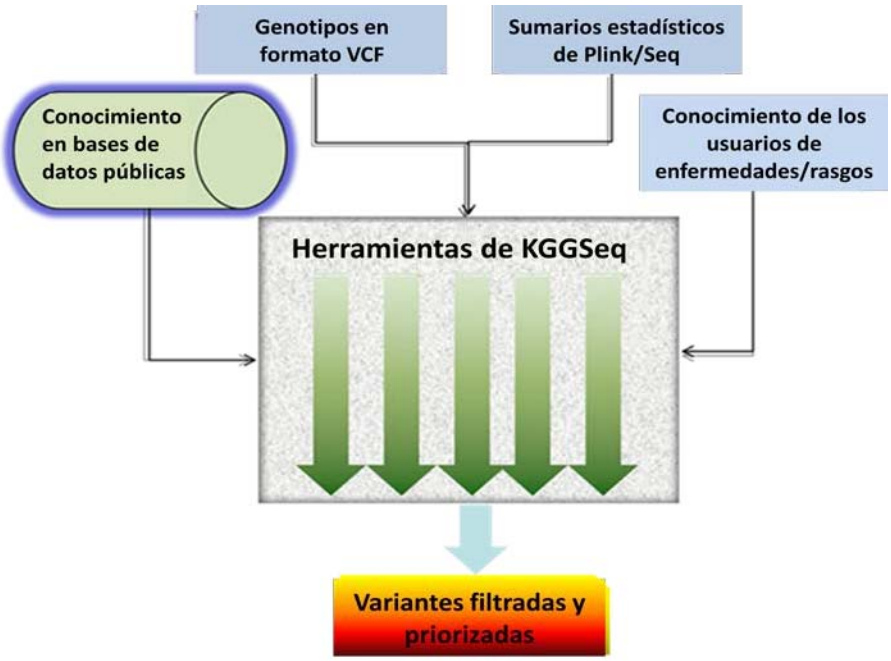



**Figura 4-12.** Diagrama de funcionamiento de KGGSeq.

Esta plataforma utiliza archivos en formato vcf para realizar el análisis. Es necesario construir un archivo vcf con las variaciones de todos los individuos que se pretenden analizar y comparar. Igualmente, precisa de información sobre la familia. Para esto último necesita un fichero ped con los siguientes datos:

- Familia (valor numérico).
- Identidad del individuo (valor numérico).
- Identidad del padre (valor numérico; en caso de ser desconocido, 0).
- Identidad de la madre (valor numérico; en caso de ser desconocido, 0).
- Sexo (1 varón; 2 hembra).
- Estatus (0 desconocido, 1 no afectado, 2 afectado).

En nuestro caso, disponemos de 3 individuos sin relación de primer grado, por lo que los padres son desconocidos. El fichero ped quedaría de la siguiente manera:

1 1 0 0 1 2

1 2 0 0 1 2

1 3 0 0 1 2

El análisis da como resultado un archivo flt.txt que se puede abrir con el programa Excel. Muestra una tabla con todas las variantes presentes en los individuos de estudio que cumplen los criterios de calidad establecidos por el programa.

El script para que funcione el programa (ver Anexo I) se compone de cinco partes:

I. Especificar el entorno para que funcione el programa. Incluye el genoma de referencia (hg19) y la localización de los programas en el ordenador.

II. Especificar los archivos de entrada (el archivo vcf con los datos de todos los sujetos de estudio y el archivo ped que hemos construido).

III. Especificar los archivos de salida. Nombre y formato.

IV. Especificar los valores de calidad. Tanto del genotipo como de las variantes.

Ejemplos de las variables utilizadas para el genotipo se exponen en la tabla 4-2; las variables utilizadas para las variantes se exponen en la tabla 4-3.

**Tabla 4-2. Variables utilizadas por KGGSeq para el genotipo.**

| Variable | Significado |
|---|---|
| --gty-qual | Calidad mínima del genotipo (Phred Quality Score). El programa utiliza el valor 10 por defecto. |
| --gty-dp | Profundidad mínima de cobertura. El programa utiliza 4x por defecto. |
| --gty-af-ref | Establece la fracción máxima del alelo alternativo (AF). El valor por defecto es 0.05. |
| --gty-af-alt | Establece la fracción mínima del alelo alternativo (AF). El valor por defecto es 0.25. |

**Tabla 4-3. Variables utilizadas por KGGSeq para las variantes.**

| Variable | Significado |
|---|---|
| --disable-vcf-filter | Orden para no eliminar las variantes no marcadas como PASS en el archivo VCF. |
| --seq-qual | Valor mínimo de calidad (*Phred Quality Score*) de la variante. El programa utiliza el valor 50 por defecto. |
| --seq-mq | Valor mínimo de calidad de mapeado. 20 es el valor por defecto. |
| --seq-sb | Establece el sesgo de cadena máximo de la variante. Utiliza por defecto el valor 10. |

V. Filtrado y priorización. Determina los filtros a utilizar para la selección de las variantes de interés.

*Filtrado por tipo de herencia.*

Se especifica el tipo de variantes que selecciona según el tipo de herencia esperada:

1. Modelo recesivo. Excluye las variantes para las que uno o más sujetos afectos son heterocigotos.
2. Mutación causal recesiva con penetrancia completa. Excluye las variantes en las que tanto los afectados como los no afectados tienen los mismos genotipos homocigotos.
3. Dominante poco común o heterocigosidad compuesta. Excluye las variantes en las que uno o más de los sujetos afectos tiene genotipos homocigotos de referencia.
4. Mutación causal dominante con penetrancia completa. Excluye las variantes presentes en heterocigosis que comparten los sujetos afectados y no afectados.
5. Dominante poco común o heterocigosidad compuesta. Excluye las variantes en las que uno o más de los sujetos afectos tiene genotipos homocigotos alternativos.
6. Mutación causal de total penetrancia. Excluye las variantes para las que los sujetos afectos no comparten alelos.
7. Mutaciones *de novo* con total penetrancia. Sólo incluye las variantes para las que los sujetos no afectos tienen genotipo homocigoto para el mismo alelo y todos los sujetos afectos tiene genotipos heterocigotos.
8. Mutaciones somáticas en los tumores.

En este caso se excluyeron los modelos recesivos, las mutaciones *de novo* y las somáticas.

*Filtrado por tipo de mutación.*

0. *Frameshift.* Indel que resulta en un cambio total de la secuencia original.
1. *Nonframeshift.* Indel que produce una pérdida de un aminoácido en la proteína resultante.

2. *Stoploss.* Pérdida de un codón de parada.
3. *Stopgain.* Indel que resulta en la aparición de un codón de parada que trunca la proteína.
4. *Missense.* Cambio de una base que produce un cambio del codón, que codifica para un aminoácido diferente.
5. *Splicing.* Variante situada en los 2 pares de bases correspondientes a un área de splicing (el número de pb se puede modificar).
6. Sinónima. Mutaciones que no producen variación en la proteína.
7. Exónica. Sólo localiza variantes en las áreas exónicas.
8. UTR5. Variante situada en una región 5' no traducida.
9. UTR3. Variante situada en una región 3' no traducida.
10. Intrónica. Sólo localiza variantes en áreas intrónicas.
11. *Upstream.* La variante se sitúa en el área de 1 Kb cadena arriba de un área de inicio de transcripción (la distancia se puede modificar).
12. *Downstream.* La variante se sitúa en el área de 1 Kb cadena abajo de un área de fin de transcripción (la distancia se puede modificar).
13. ncRNA. La variante se sitúa en un área transcrita sin código de anotación en la definición de gen.
14. *Intergenic.* Variantes en áreas inter-genes.
15. *Unknown.* Variantes que el programa no ha logrado situar.

Para el análisis seleccionamos 0, 1, 2, 3, 4 y 7.

*Filtro por frecuencia de los alelos en las bases de datos públicas.*

Filtra con los datos de bases de datos como la de los 1000 genomas, la de los 6500 exomas, o la dbSNP.

*Filtro por tipo de variante.*

Se pueden seleccionar, incluir o quitar de los resultados los indeles, mutaciones de nucleótido único, algunas regiones del genoma...

Para hacer más eficiente nuestro análisis hemos excluido las variaciones en los cromosomas X e Y por no ser compatible una mutación en ellos con el tipo de herencia en esta familia.

*Filtro por función.*

Permite filtrar las variaciones según estén clasificadas como pseudogenes, áreas de unión de factores de transcripción, activadores y otras características definidas en UniProt. Algunos ejemplos se recogen en la tabla 4.4.

**Tabla 4-4. Tipos de función del área codificada.**

| Función | Definición |
|---|---|
| Área de interés | Región de interés en la secuencia |
| Sitio activo | Amino-ácido(s) implicados en la actividad de un enzima |
| Región de unión de calcio | Área de unión a moléculas de calcio |
| Sitio de glicosilación | Idem |
| Cadena | Área de cadena polipeptídica en la proteína madura |
| *Coiled-coil region* | Área que forma un ovillo |
| Secuencia conflicto | Secuencias diferentes en distintos artículos |
| Puente disulfuro | Idem |
| *DNA-binding region* | Región de union al ADN |
| Hélice | Estructura secundaria de hélice |
| Región intra-membrana | Idem |
| Metionina de inicio | Idem |
| Residuo modificado | Idem |
| *Metal ion-binding site* | Sitio de union de iones metálicos |
| *Short sequence motif* | Secuencia corta (hasta 20 aminoácidos) de interés biológico. |
| *Mutagenesis site* | Sitio que ha sido alterado a nivel experimental |
| Aminoácido no estándar | Idem |
| Región de unión a fosfato | Idem |
| Péptido | Se transcribe en un péptido activo. |
| *Repeat* | Repetición de secuencia interna. |
| Región transmembrana | Idem |
| Residuo inseguro | Dudas en la secuencia |
| Variante de secuencia | Los autores aseguran que hay variantes en la secuencia |

*Filtro por predicción.*

Muestra las puntuaciones de predicción según múltiples algoritmos: SIFT (utiliza el algoritmo "*Sorting Tolerant From Intolerant*"), PolyPhen2, LRT (*likelihood ratio test*; con valores de 0 a 1, teniendo mayor riesgo de ser patógena cuanto más cerca de 1), *Mutation-Taster* (con puntuaciones de 0 a 1, siendo mayor el riesgo de patogenicidad

cuanto más alta), *MutationAssessor* (que define el impacto funcional de la mutación) (Reva, Antipin, & Sander, 2011) o FATHMM (que clasifica las variantes como D, patógena, o T, tolerada) (Shihab et al., 2013).

VI. Anotación. Permite seleccionar la manera de mostrar las variantes seleccionadas.

**Visualización de los resultados.**

Se ha utilizado el programa IGV (*Integrative Genomics Viewer*), (Robinson et al., 2011a) que muestra de forma gráfica los datos obtenidos (figura 4-13). Permite diferenciar con mayor facilidad verdaderas variaciones frente a artefactos. Igualmente, facilita la comparación entre varias muestras.

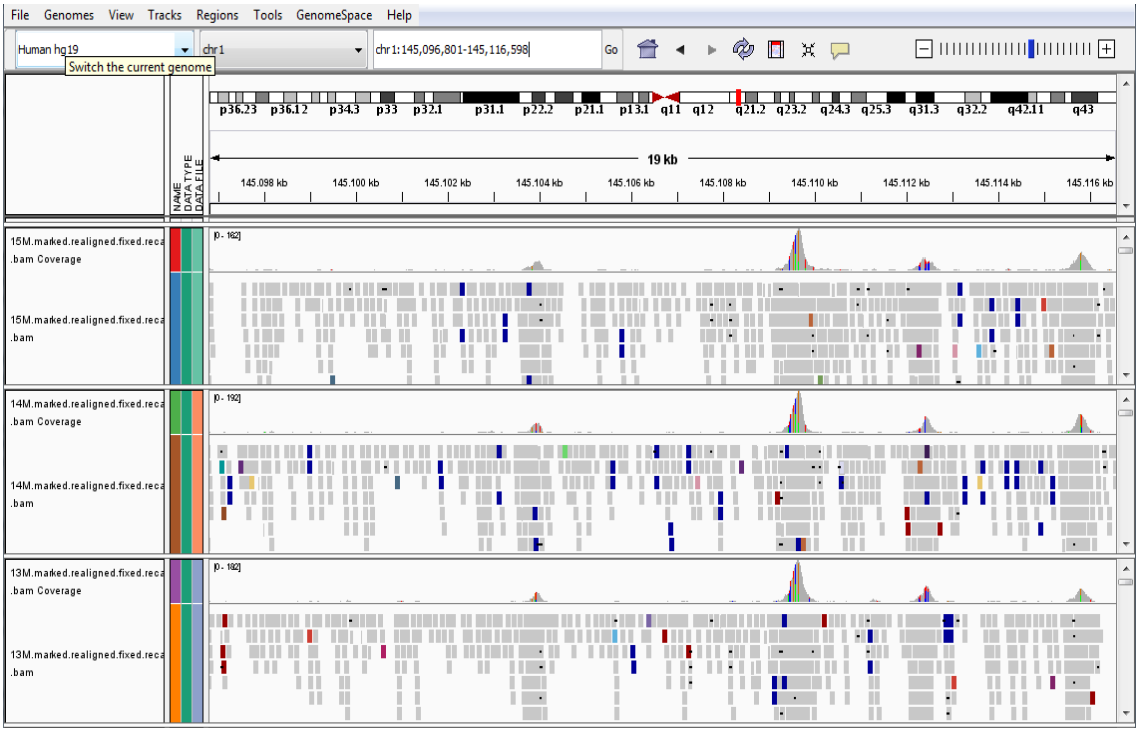



**Figura 4-13.** Ejemplo de visualización de los datos con el programa IGV. Se muestra el área cromosómica correspondiente, y en la parte superior muestra un histograma con la profundidad de cobertura de las lecturas para cada posición.

**Confirmación de las mutaciones con la técnica de Sanger.**

El método de secuenciación de Sanger se basa en la utilización de cuatro tubos distintos en los que se incluye la hebra de ADN a secuenciar, ADN polimerasa, un cebador, deoxiribonucleótidos correspondientes a las cuatro bases (A, G, T y C) y una cantidad de dideoxiribonucleótido de un tipo concreto en cada uno de los tubos. Los dideoxiribonucleótidos, al carecer de grupo hidroxilo en el carbono 3', finalizan el crecimiento de la cadena. Por tanto, en cada tubo se producen cadenas de distintas logitudes que, tras ser sometidos a electroforesis, generan un patrón de bandas del que se puede deducir la secuencia de ADN.

El proceso consta de tres fases: la amplificación del fragmento de ADN a secuenciar a través de PCR, el análisis con electroforesis de los fragmentos obtenidos y, finalmente, la secuenciación.

Los materiales utilizados para llevar a cabo esta técnica se recogen en el cuadro 5.1.

1. Amplificación del fragmento con PCR (reacción en cadena de la polimerasa).

El proceso comienza con la amplificación del fragmento de ADN a secuenciar. Primero se somete a una temperatura superior a 90 ºC para que se desnaturalice y se separen las dos hebras de la cadena. Posteriormente se añaden cebadores específicos, que se combinan con las hebras independientes a una temperatura de unos 45-65ºC. Finalmente, se añaden ADN polimerasa y deoxiribonucleótidos (dNTPs), y se inicia la formación de las cadenas complementarias a una temperatura de 72ºC. Cada doble cadena recién formada inicia de nuevo el proceso de desnaturalización-unión a cebadores-elongación, sirviendo las nuevas copias de molde para hacer más copias, creciendo el número de forma exponencial en cada ciclo. Al final de n ciclos, el número de copias de cada hebra será de $2^n$. El número de ciclos es habitualmente de 28 a 40, obteniéndose al menos un millón de copias.

Actualmente el proceso se lleva a cabo de forma totalmente automática gracias a la utilización de polimerasas termoestables junto con el diseño de termocicladores, aparatos que permiten llevar a cabo los ciclos de tiempo y temperatura necesarios de un modo rápido.

El programa utilizado se expone en la tabla 5-4. Se llevaron a cabo 35 ciclos.

**Tabla 4-5. Programa para la amplificación (PCR).**

| Proceso | Temperatura (ºC) | Tiempo |
|---|---|---|
| Desnaturalización | 94 | 30 segundos |
| Anillamiento | 55 | 30 segundos |
| Extensión | 72 | 30 segundos |
| Extensión final | 72 | 3 minutos |

**Cuadro 4-3. Materiales utilizados para la secuenciación con el método de Sanger.**

**Amplificación PCR:**

-74 µL de H2O.

-10 µL de buffer 100 mM Tris-HCl, pH 8,3, 500 mm KCl, 1,5 mM $MgCl_2$; concentrado 10 veces.

-6 µL de *primer* o cebador (seleccionados con el software Primer3).

-4 µL de dNTPs.

-6 µL de ADN.

-1 µL de Taq polimerasa.

-Magnesio: El magnesio es necesario para que actúen las polimerasas, siendo la concentración habitual de 1,5 mEq/L. En determinadas reacciones fue necesario emplear concentraciones mayores para aumentar el rendimiento.

**Análisis electroforético:**

-Buffer: 250 ml de agua destilada + 50 ml de TAE50X.

-Geles: Disolución de 0,6g de agarosa en 30 ml de agua destilada + 600 µl de tampón TAE50X en un matraz Erlenmeyer. Se calentó la mezcla hasta ebullición en un horno microondas.

**Secuenciación:**

*Filtrado:*

-Columnas Wizard® SV Gel and PCR Clean-Up System (Promega, Madison, USA).

-Solución *Membrane Binding.*

-Tubo de Eppendorf.

-700 + 500µl de solución *Membrane Wash.*

-Tubo limpio de microcentrifugado de 1,5 ml.

-50 ml de *Nuclease-Free Water.*

*Secuenciación:*

-2 µl de *primer* único (*forward* o *reverse*) a una concentración de 5 pmol/µl.

-3 µl de agua estéril.

-dNTPs, buffer, ADN polimerasa.

-4 µl de ddNTPs.

2. Análisis electroforético del producto de PCR.

Los fragmentos obtenidos con la amplificación se detectaron por electroforesis en geles de agarosa con tinción de bromuro de etidio. El bromuro de etidio, sustancia que emite fluorescencia sometida a luz ultravioleta, tiene la propiedad de intercalarse entre los pares de bases adyacentes del ADN.

Tras incluir el producto de PCR en el gel de agarosa y aplicar una corriente eléctrica de 130mV durante unos 30-60 minutos, se separó lo suficiente para poder visualizar las diferentes moléculas de ADN presentes en el medio. Posteriormente el gel se sumergió en una solución de bromuro de etidio con una concentración de 0,5µg/ml. Tras 10 minutos incluído en la sustancia fluorescente se procedió a limpiar el exceso de bromuro de etidio con agua destilada. Finalmente, se realizó la lectura, registrando la imagen obtenida con un sistema de fotografiado digital de geles.

3. Secuenciación.

Para realizar la secuenciación el primer paso es purificar los productos de la PCR. Para ello se mezclaron con un volumen equivalente de solución *Membrane Binding* y se filtró la mezcla en la columna, facilitando el proceso con centrifugación a 10000 revoluciones durante 1 minuto. Tras retirar el sobrandante, se añadió la solución *Membrane Wash* (que contiene etanol) y se centrifugó de nuevo. Este proceso se repite, realizando una segunda centrifugación de 5 minutos. Posteriormente, se transfirió la columna a un tubo limpio y se añadió *Nuclease-Free Water.* Tras incubar un minuto a temperatura ambiente se centrifugó de nuevo 1 minuto a 10000 revoluciones. Finalmente, se retiró la columna y se mantuvo el ADN refrigerado hasta la secuenciación.

La secuenciación de los productos de PCR amplificados se realizó en un secuenciador automatizado, por secuenciación cíclica (*Amplicycle TM Sequencing Kit,* Perkin Elmer). Finalmente, la secuencia fue leída con los programas informáticos:

-Chromas v 2.0 (Technelysium Pty Ltd, Queenland, Australia).

-Generunner v 3.05 (Hasting Software Inc.)

# 5. RESULTADOS.

La secuenciación del exoma se computó en seis ficheros (dos de cada sujeto) con formato .fq.gz (formato comprimido de FastQ) correspondientes a los sujetos II, III y IV (tabla 5-1).

**Tabla 5-1. Ficheros .fq.gz.**

| Sujetos | Tamaño fichero | Cobertura solicitada | Nº de secuencias |
|---|---|---|---|
| II | 3,16 GB | 50x | 34952624 |
| | 3,19 GB | 50x | 34952624 |
| III | 2,55 GB | 50x | 28170365 |
| | 2,57 GB | 50x | 28170365 |
| IV | 2,79GB | 50x | 30727880 |
| | 2,80 GB | 50x | 30727880 |

## A. CALIDAD DE LOS DATOS

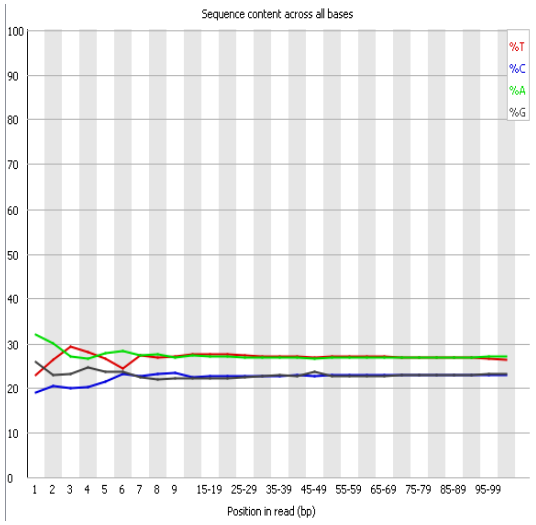
Las estadísticas del programa FastQC mostraron desviaciones semejantes en todas las muestras:

- Contenido de bases en la secuencia.

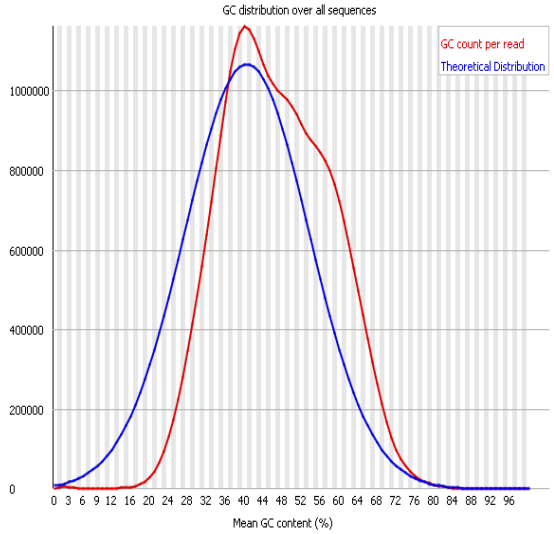
  En todas ellas se observa un enriquecimiento de adenina en las primeras posiciones de lectura. Esto probablemente esté en relación con los adaptadores que se utilizan para crear las bibliotecas para el secuenciador de Illumina, que son secuencias poliA (gráfica 5-1).

- Contenido de GC por secuencia.

  En todas las muestras obtenemos un contenido en GC mayor del esperado. Esto se debe a que el patrón que muestra el programa está hecho sobre un genoma tipo, y nuestras muestras corresponden exclusivamente a secuencias exómicas, habitualmente más ricas en GC (gráfica 5-2).

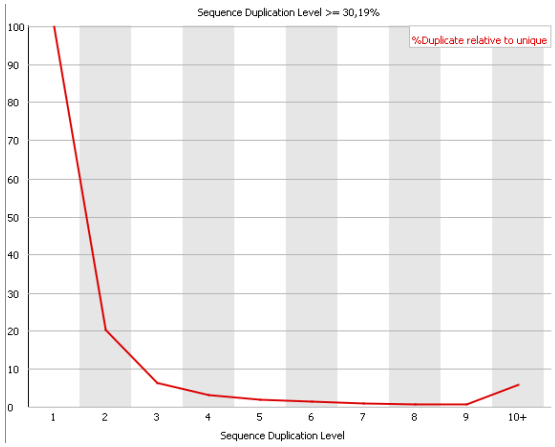



**Gráfica 5-1.** Contenido de bases en la secuencia.




**Gráfica 5-2.** Contenido de GC por secuencia.

- Niveles de duplicación de secuencias.

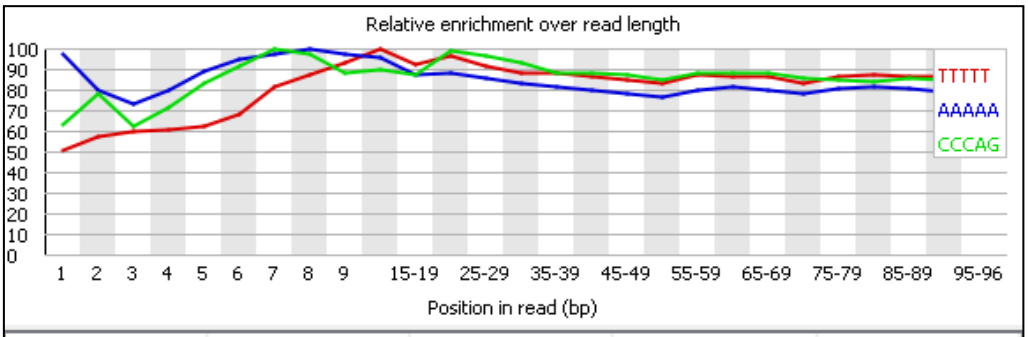



Se encuentran niveles levemente por encima del límite aceptable (20%). En dos de las muestras los valores no llegan al 22%. En la tercera, rondan el 30% (gráfica 5-3).

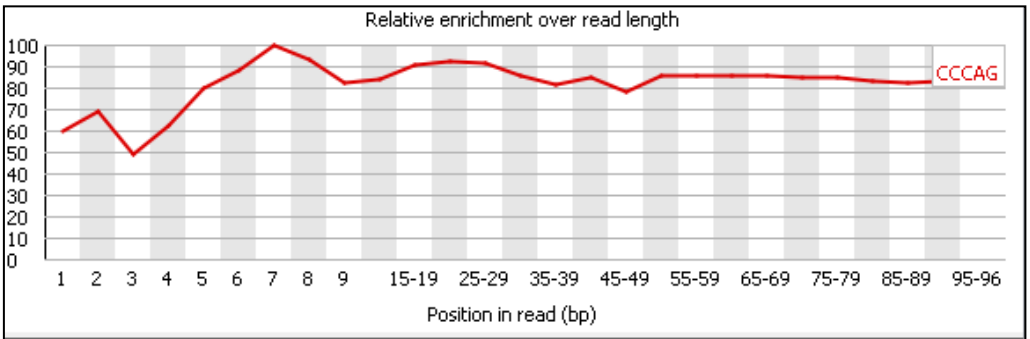
**Gráfica 5-3.** Niveles de duplicación de secuencias.

- Contenido en k-mer.

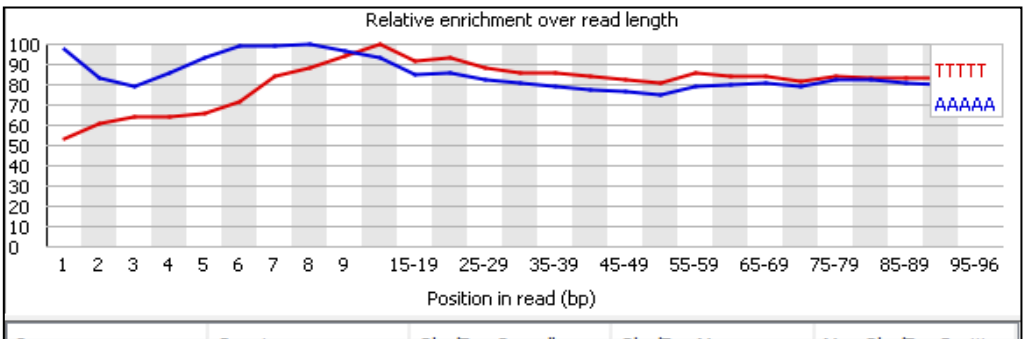
  Se encuentran en las muestras mayor contenido del esperado en poliT, poliA y el pentámero CCCAG en las tres muestras, probablemente en relación con leves defectos de la técnica.




| Sequence | Count | Obs/Exp Overall | Obs/Exp Max | Max Obs/Exp Position |
|---|---|---|---|---|
| TTTTT | 14413220 | 3,176 | 3,65 | 10-14 |
| AAAAA | 14999545 | 3,151 | 3,777 | 8 |
| CCCAG | 7796865 | 3,045 | 3,447 | 7 |

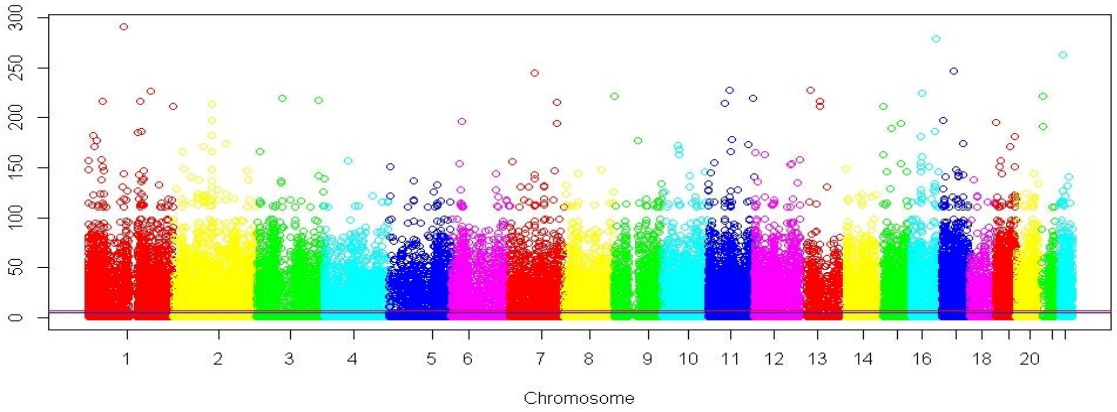



| Sequence | Count | Obs/Exp Overall | Obs/Exp Max | Max Obs/Exp Position |
|---|---|---|---|---|
| CCCAG | 6020810 | 3,005 | 3,536 | 7 |

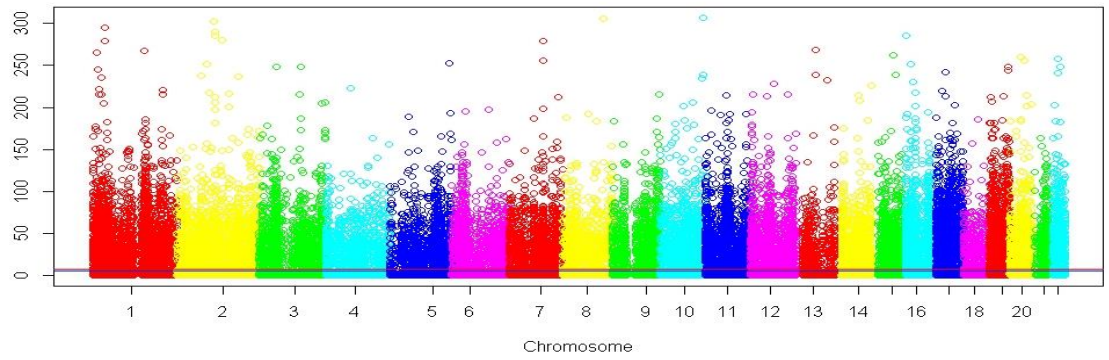



| Sequence | Count | Obs/Exp Overall | Obs/Exp Max | Max Obs/Exp Position |
|---|---|---|---|---|
| TTTTT | 11661460 | 3,13 | 3,683 | 10-14 |
| AAAAA | 12360200 | 3,114 | 3,767 | 8 |

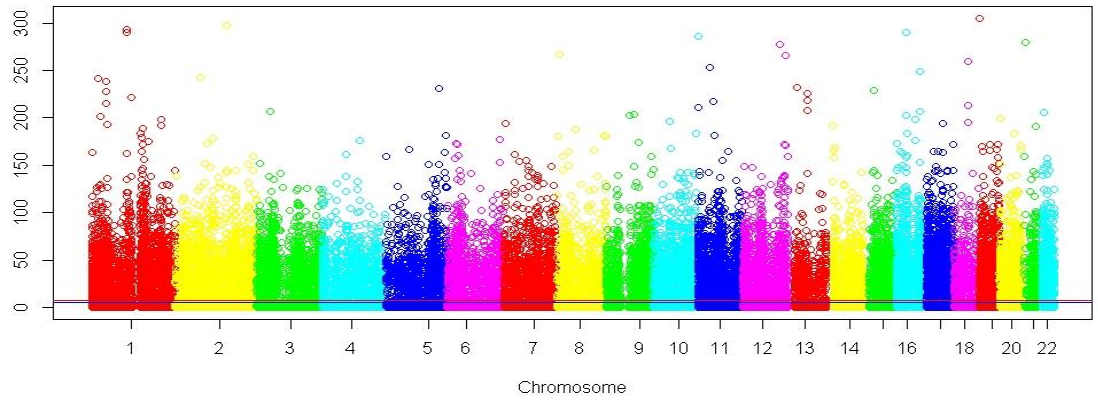
El análisis de profundidad de cobertura de las muestras con bedtools se muestra en las siguientes gráficas tipo Manhattan 5-4, 5-5 y 5-6, correspondientes a la secuenciación de los sujetos II, III y IV respectivamente. Como se puede observar, la cobertura de la mayoría de las secuencias supera el 50x, salvo las áreas correspondientes a los centrómeros. Algunas secuencias alcanzan el 300x.




**Gráfica 5-4.** Profundidad de cobertura de las lecturas del sujeto II.




**Gráfica 5-5.** Profundidad de cobertura de las lecturas del sujeto III.



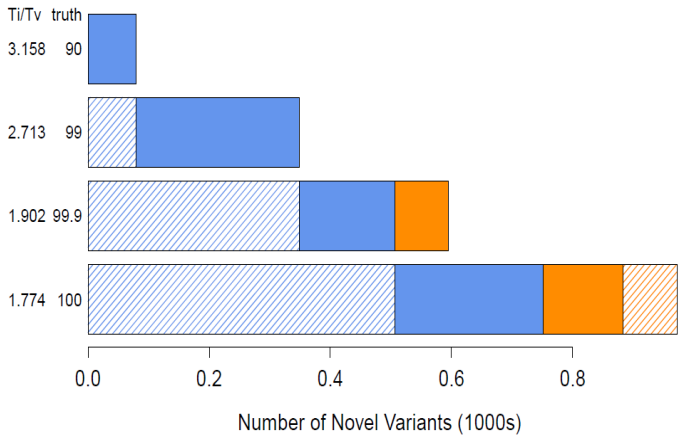

**Gráfica 5-6.** Profundidad de cobertura de las lecturas del sujeto IV.

## B. PROCESO DE ANÁLISIS.

### Localización de las variantes y recalibración.

Recalibración de la puntuación de calidad de las variantes.

Como se explica en el capítulo 4, la recalibración de la puntuación de las variantes se lleva a cabo con un modelo gaussiano, que clasifica las variantes encontradas en verdaderos o falsos positivos según una serie de parámetros. El programa aporta unas gráficas que informan sobre la sensibilidad y especificidad en la localización de dichas variaciones.




**Gráfica 5-7.** Parámetros de sensibilidad-especificidad. A medida que aumenta la sensibilidad en la localización de nuevas variantes van aumentando los verdaderos positivos, pero también los falsos positivos. En este caso, perteneciente al sujeto IV, con una sensibilidad de 90 y Ti/Tv de 3.158 encontramos unas 100 variaciones, todas ellas verdaderos positivos, pero sólo localizamos una pequeña parte de las que realmente hay. Al aumentar la sensibilidad a 99 y bajar la Ti/Tv a 2.713 (segunda barra) añadimos unas 250 más (azul), junto con las 100 (barra azul rallada) que ya aparecían con una sensibilidad menor. En el tercer tramo (tercera barra) hallaremos unos 500 verdaderos positivos, 150 más que con la anterior (azul), pero también seleccionamos unos 100 falsos positivos (naranja). Con una sensibilidad de 100 encontraremos unas 750 variaciones que son verdaderos positivos (azul + azul rallado) pero también unas 250 que son falsos postivos.

Eje x. Número de variantes nuevas.

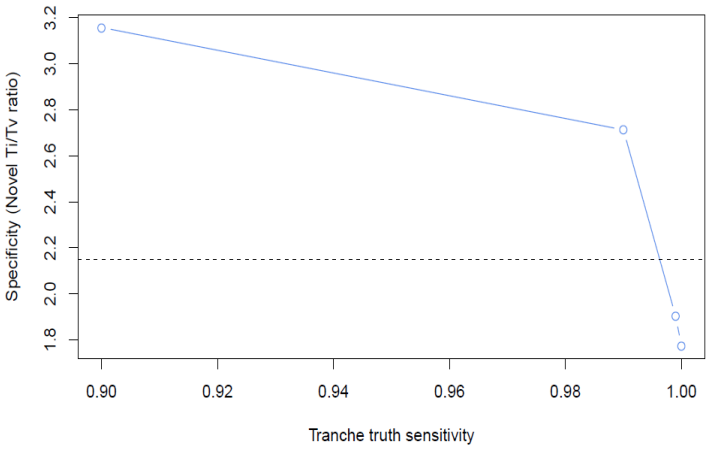
Eje y. *Novel transition to tranversion ratio* (equivalente de especificidad) y sensibilidad verdadera global (*truth*).

Azul. Verdaderos positivos obtenidos al aumentar la sensibilidad.

Azul rallado. Verdaderos positivos acumulados.

Naranja. Falsos positivos obtenidos al aumentar la sensibilidad.

Naranja rallado. Falsos positivos acumulados.




**Gráfica 5-8.** En esta gráfica, del mismo sujeto, se observa la relación entre una medida de especificidad (Ti/Tv) y la sensibilidad.

**Filtrado de las variantes encontradas.**

Tras la localización de las variantes encontramos:

- Sujeto II: 23843 variantes.
- Sujeto III: 23283 variantes.
- Sujeto IV: 23791 variantes.

Método GATK-ANNOVAR.

1. Filtrado por variantes raras.

Elegimos exclusivamente aquellas mutaciones que aparecen en la base de los 1000 genomas y la de los 6500 exomas con una frecuencia menor de 0.0005. Pese a que se espera que pueda haber pacientes con trastorno bipolar en estas bases de datos, con una frecuencia semejante a la población general (estas bases carecen de información fenotípica), las posibilidades de que sean además portadores de una mutación asociada a una herencia autosómica dominante son mucho más bajas, por lo que 0.0005 nos parece un valor adecuado.

2. Filtrado basado en la función.

Se eliminan las mutaciones sinónimas, dada su benignidad.

3. Filtrado por predicción.

Se eligen valores de AVSIFT menores de 0.05 y valores de PolyPhen entre 0.95 y 1. Se incluyen aquellas mutaciones para las que no se puede calcular el valor de predicción.

4. Eliminación de los segmentos duplicados y selección de aquellas mutaciones marcadas como PASS (criterio de calidad).

| | Total | Variantes raras | Función | Predicción | Seg Dup | PASS | Todos los filtros |
|---|---|---|---|---|---|---|---|
| IV | 23791 | 2653 | 11972 | 13096 | 21997 | 19012 | 198 |
| III | 23283 | 2492 | 11799 | 13725 | 21561 | 18476 | 179 |
| II | 23843 | 2700 | 12006 | 14155 | 22067 | 18881 | 179 |

**Tabla 5-2.** Número de mutaciones encontradas en cada sujeto y número de ellas que pasan los distintos filtros (GATK).

5. Comparación de los exomas de estudio.

Utilizando el programa ANNOVAR se seleccionaron sólo las variaciones que estuvieran presentes en los tres sujetos del estudio, quedando sólo 63 que pasaran todos los filtros descritos (figura 5-1).

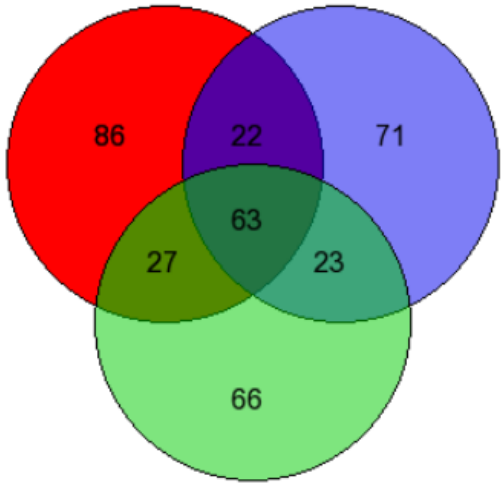



**Figura 5-1.** Diagrama Venn (http://genevenn.sourceforge.net) en el que se muestran las mutaciones halladas en cada individuo y las que comparten entre ellos.

6. Filtrado final.

De las 63 variantes encontradas, se descartaron 50 por tratarse de áreas de splicing o indeles en homocigosis, lo que no es compatible con el modelo de herencia del caso que nos ocupa, y otras 7 por estar presentes en otros sujetos de nuestra base de datos, bien por defectos de la técnica o bien por ser variaciones comunes en nuestra población. Otras 4 fueron eliminadas por no tener sus productos una función que tenga relación con el sistema nervioso central.

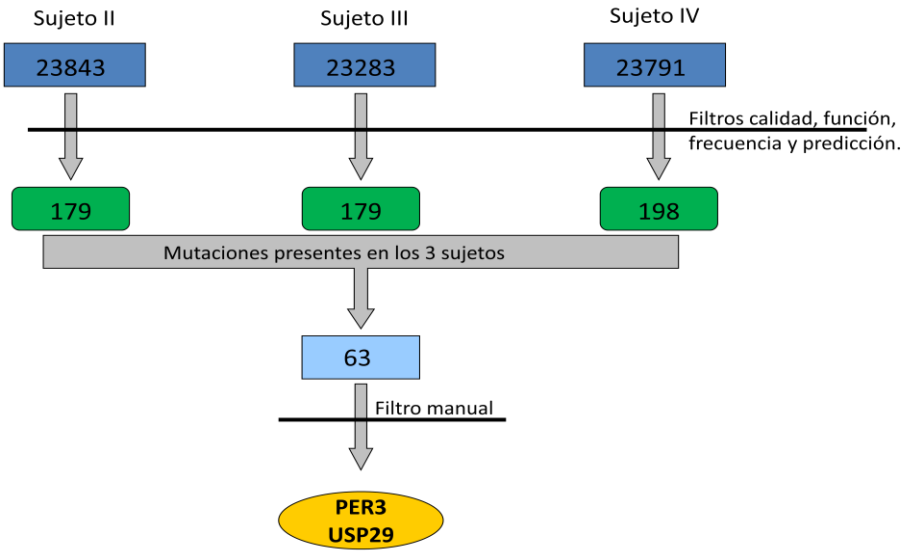



**Figura 5-2.** Proceso de filtrado.

Tras el proceso de filtrado nos quedaron las siguientes mutaciones:

**-Gen PERIOD3 (PER3).**

Cromosoma 1, exón 3. Mutación no sinónima de nucleótido único. Cambio de adenina por guanina en la posición 7846853. rs201111117 (figura 5-3).

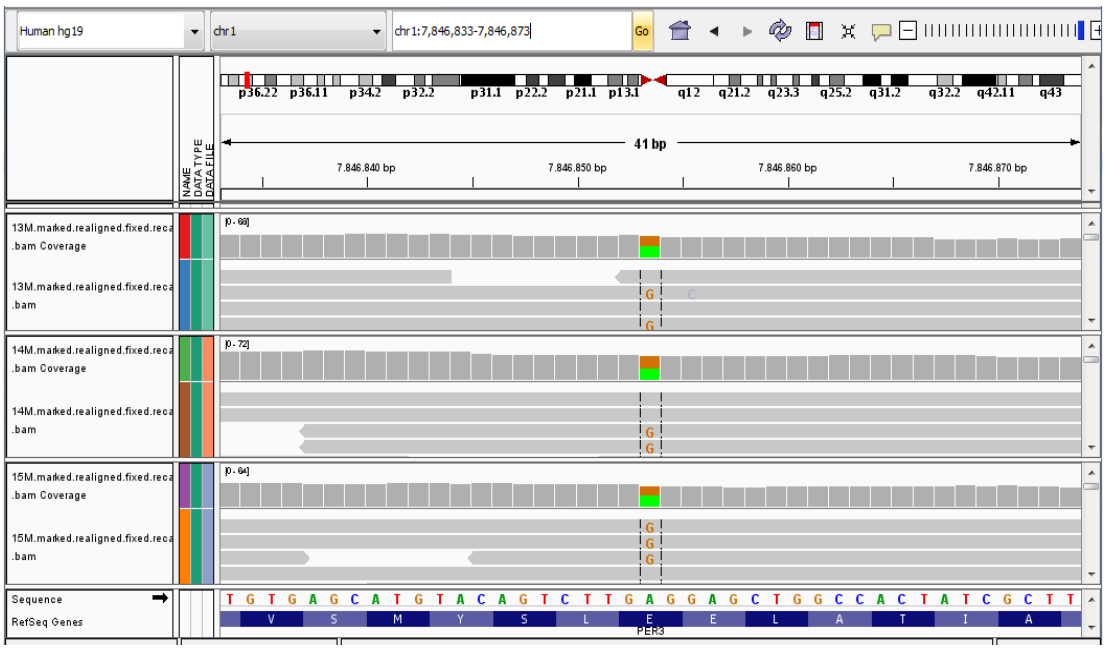



**Figura 5-3.** Captura de pantalla del programa IGV que muestra la mutación en el gen PER3, presente en los 3 sujetos de estudio.

El gen PERIOD3 está situado en el brazo corto del cromosoma 1 (1p36.23), entre los genes VAMP3 y UTS2. Con una longitud de 60.86 Kb, está compuesto por 21 exones y las regiones intrónicas intermedias (figuras 5-4 y 5-5).

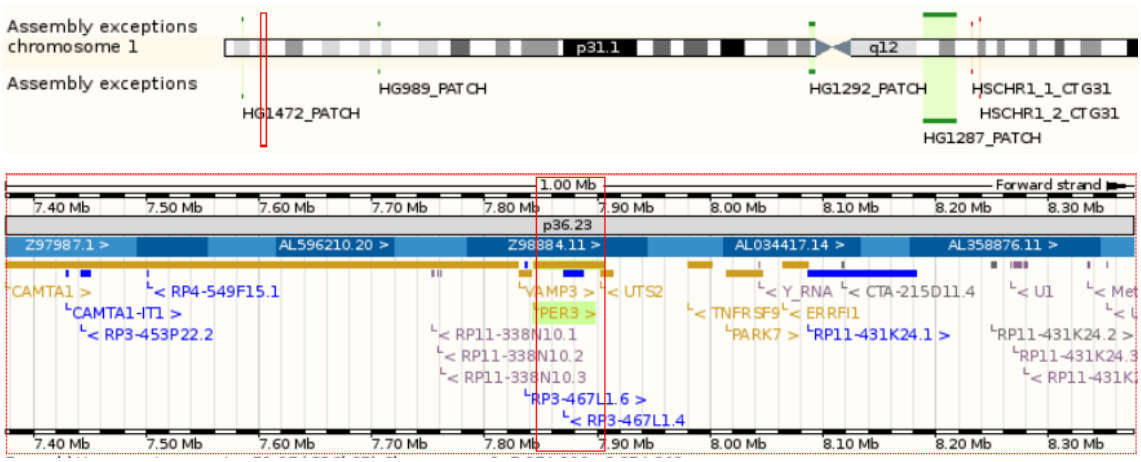



**Figura 5-4.** Gen PERIOD3. Tomado de la base Ensembl (http://www.ensembl.org/).

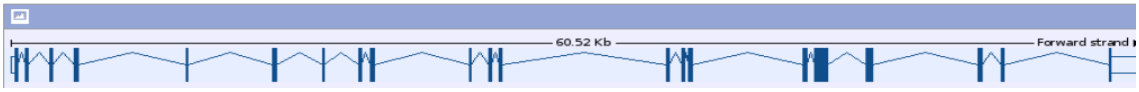



**Figura 5-5.** Gen PERIOD3. Tomado de Vega Genoma Browser (http://vega.sanger.ac.uk/Homo_sapiens).

**-Gen *Ubiquitin Specific Peptidase* 29 (USP29).**

Cromosoma 19, exón 4. Delección con cambio en los tripletes de lectura. Pérdida de una citosina en la posición 57640904 (figura 5-6).

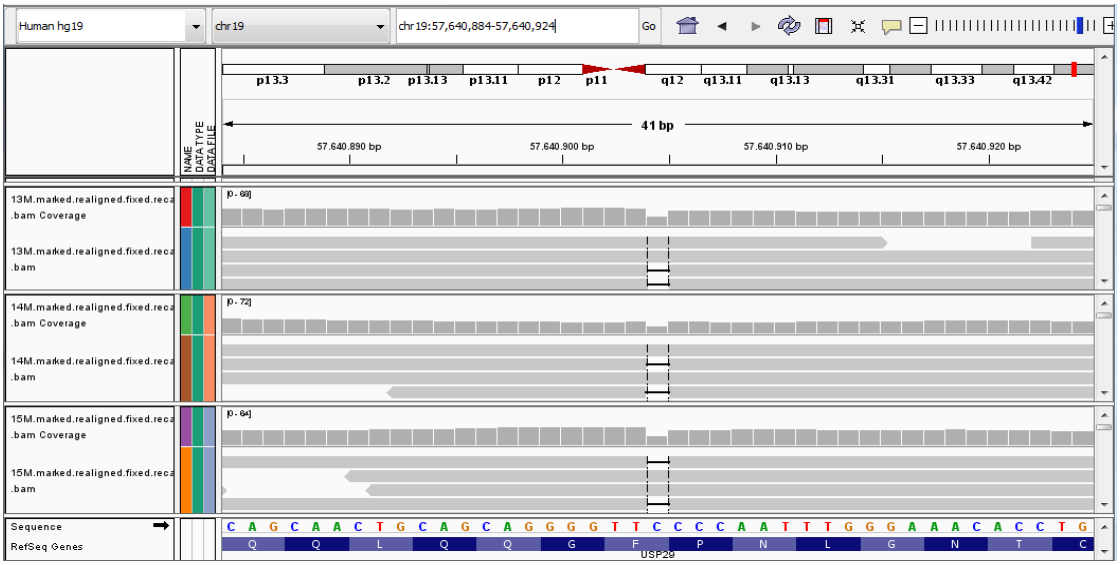



**Figura 5-6.** Captura de pantalla del programa IGV que muestra la delección en el gen USP29, presente en los 3 sujetos de estudio.

El gen USP29 se sitúa en el brazo largo del cromosoma 19 (19q13.43), junto al gen U3 (figura 5-7). Tiene una longitud de 11.88 Kb y se compone de 4 regiones exómicas (figura 5-8 y tabla 5-3).

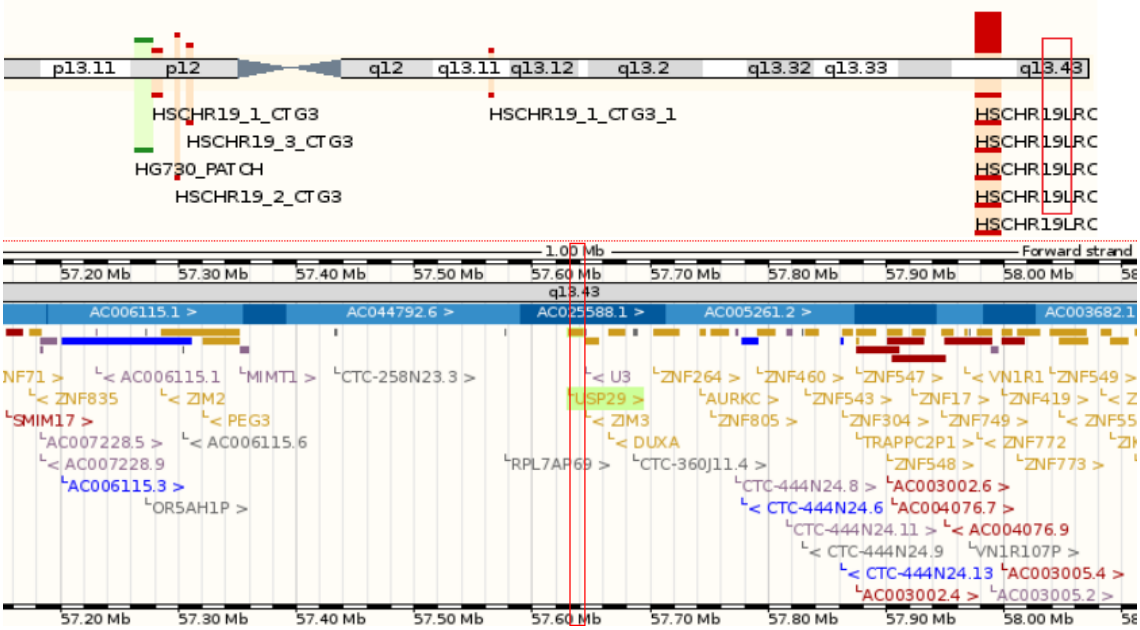



**Figura 5-7.** Gen USP29. Tomado de la base Ensembl (http://www.ensembl.org/).

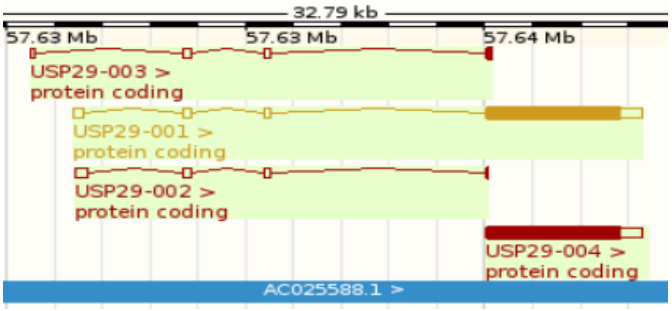



**Figura 5-8.** Estructura del gen USP29 (gráfico). Tomado de la base Ensembl (http://www.ensembl.org/).

**Tabla 5-3. Estructura del gen USP29.** Tomado de la base Ensembl (http://www.ensembl.org/).

| Nº exón | Comienzo | Final | Longitud |
|---|---|---|---|
| **Secuencia 5' upstream** | | | |
| **1** | 57,631,411 | 57,631,597 | 187 |
| **Intrón 1-2** | 57,631,598 | 57,633,692 | 2095 |
| **2** | 57,633,693 | 57,633,842 | 150 |
| **Intrón 2-3** | 57,633,843 | 57,635,406 | 1564 |
| **3** | 57,635,470 | 57,635,507 | 101 |
| **Intrón 3-4** | 57,635,508 | 57,640,027 | 4520 |
| **4** | 57,640,028 | 57,643,294 | 3267 |

MÉTODO KGGSeq.

Con el análisis con el programa KGGSeq se encontraron 10 mutaciones que satisfacían los criterios impuestos, situadas en los siguientes genes: PER3, RCC1, KIT, TMEM155, ANKRD31, AK8, THBS1, PAPD5, OR7C1 y USP29.

Se realizó igualmente un filtrado manual de las mutaciones encontradas:

- Basándonos en las puntuaciones de predicción, el programa clasifica las mutaciones en RCC1, AK8 y THBS1 como de menor riesgo de ser patogénicas.
- OR7C1 codifica para un receptor olfatorio, por lo que la función es poco compatible con la patología que nos ocupa.
- La mutación en KIT, además de tener valores menores en riesgo de ser patogénica, se sitúa en un área de splicing, afectando la mutación exclusivamene a una de las isoformas, estando, además, en heterocigosis.
- PAPD5 afecta también a un área de splicing.

Por tanto, además de las mutaciones halladas con el análisis previo, se añaden a la lista de mutaciones candidatas:

**-Gen *Transmembrane protein 155* (TMEM155).**

Cromosoma 4, exón 5. Mutación de nucleótido único. Sustitución de A por G en la posición 122682805 que genera la aparición de un codón de parada (figura 5-9).

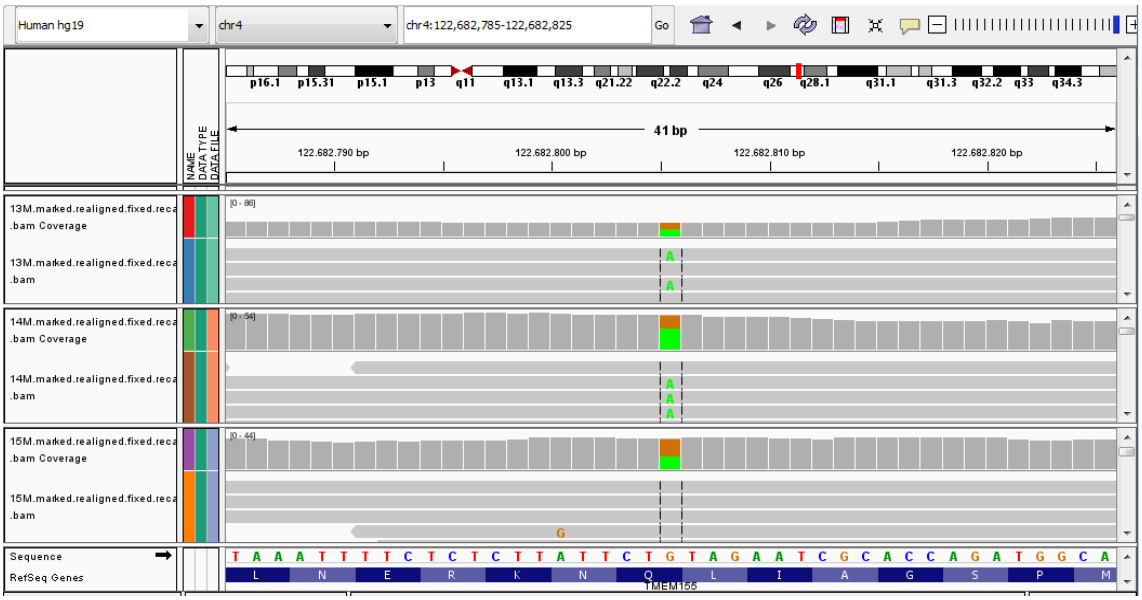



**Figura 5-9.** Captura de pantalla del programa IGV que muestra la mutación en el gen TMEM155, presente en los 3 sujetos de estudio.

El gen TMEM155 se sitúa en el brazo largo del cromosoma 4 (4q27). Tiene una longitud de 6,49 Kb y posee 6 exones (figura 5-10).


**Figura 5-10.** Gen TMEM155. Tomado de la base Ensembl (http://www.ensembl.org/).

**-Gen *Ankyrin repeat domain-containing protein 31* (ANKRD31).**

Cromosoma 5, exón 7. Consiste en la delección de tres bases (TCA) en la posición 74491715 sin cambio en los codones de lectura (figura 5-11).

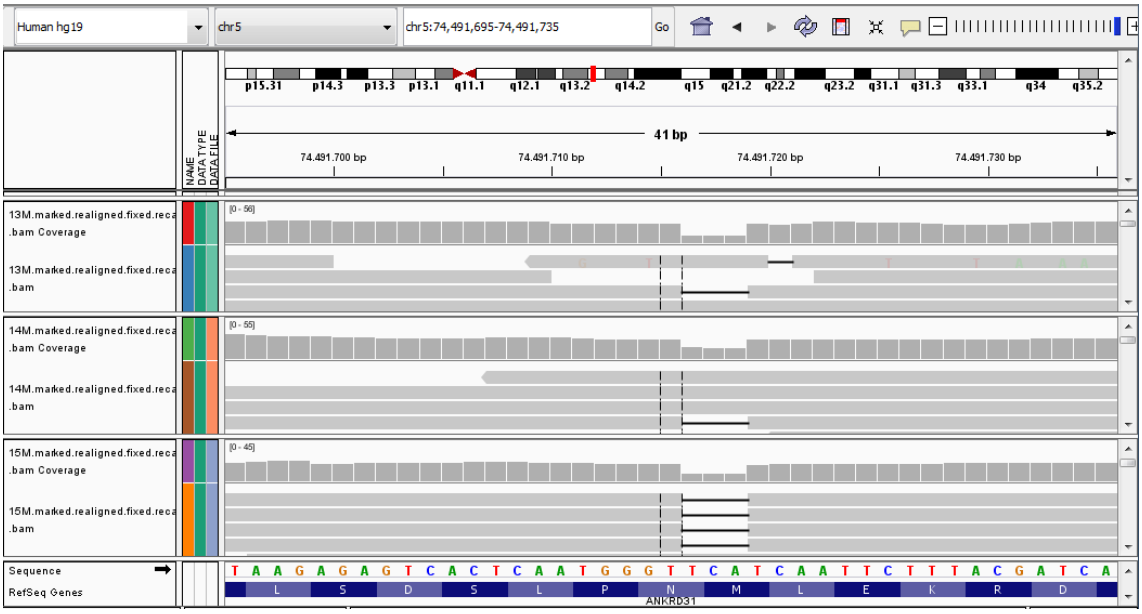



**Figura 5-11.** Captura de pantalla del programa IGV que muestra la delección en el gen ANKRD31, presente en los 3 sujetos de estudio.

El gen ANKRD31 se sitúa en el brazo largo del cromosoma 5 (5q13.3). Tiene una longitud de 168,60 Kb y posee 25 exones (figura 5-12).

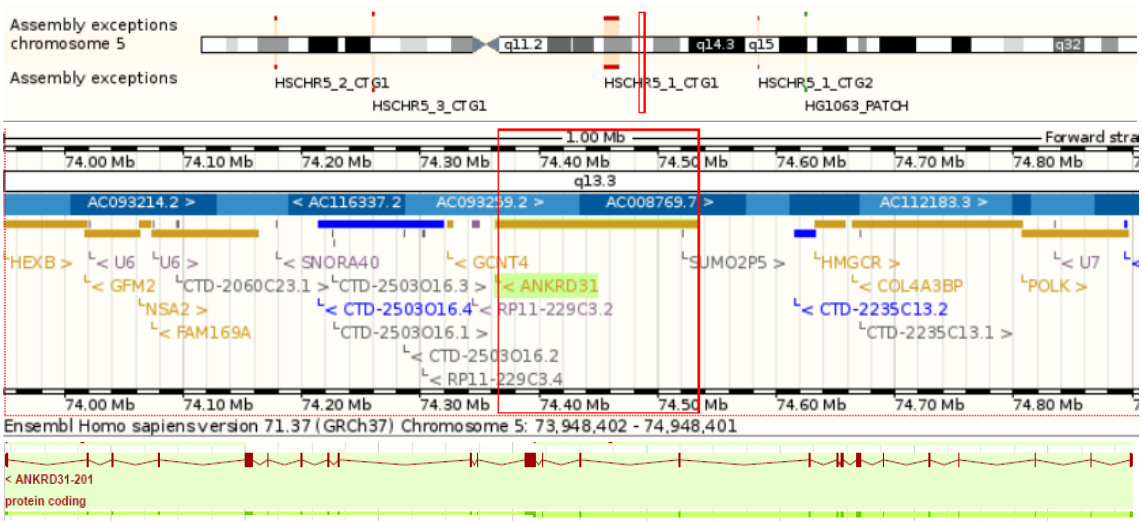



**Figura 5-12.** Gen ANKRD31. Tomado de la base Ensembl (http://www.ensembl.org/).

## C. COMPROBACIÓN DE LA PRESENCIA DE LAS MUTACIONES.

Se comprobó la presencia de la mutación de nucleótido único en el gen PERIOD3 con la técnica de Sanger en los tres sujetos de estudio para descartar que se trate de un defecto de la técnica de secuenciación con las herramientas descritas en el capítulo 4. Para ello se secuenció el exón 3 del gen PER3 (figura 5-13).

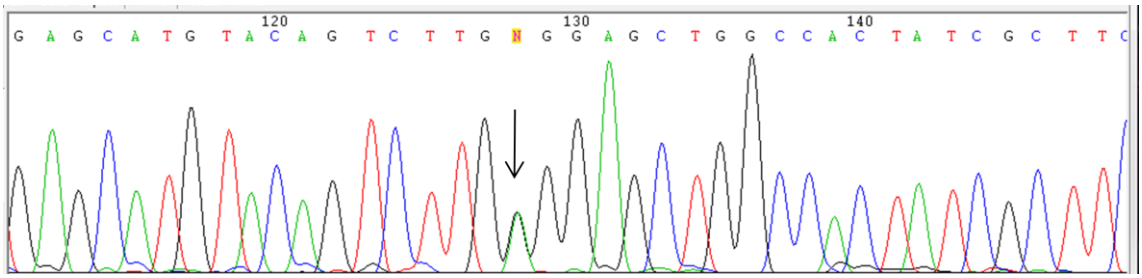



**Figura 5-13.** Visualización de la presencia de la mutación c.A347G en el exón 3 del gen PER3 (flecha).

# 6. DISCUSIÓN.

El filtrado realizado deja las siguientes mutaciones como posibles responsables del trastorno bipolar de herencia autosómica dominante presente en la familia de estudio:

### A. Gen PERIOD3 (PER3).

PER3:NM_016831:exon3:c.A347G:p.E116G. Cromosoma 1, exón 3. Mutación no sinónima de nucleótido único. Cambio de adenina por guanina. rs201111117. En la proteína resultante, se sustituye un ácido glutámico en la posición 116 por una glicina.

**Patogenicidad de la mutación.**

Si se observa la puntuación de conservación con PolyPhen (http://genetics.bwh.harvard.edu/ggi/pph2/) (figura 6-1), se comprueba que la mutación modifica en gran medida la proteína resultante, siendo por tanto muy probablemente patogénica.

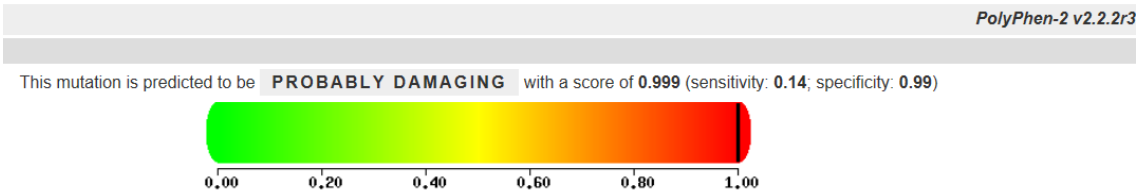



**Figura 6-1.** Representación gráfica de la puntuación de la mutación según PolyPhen.

Otra evidencia que apoya la patogenicidad de la mutación es que no sólo el residuo mutado en PER3 está muy conservado inter-especies (figura 6-2), sino que, además, lo está en las otras proteínas PER presentes en el ser humano. Alineando las 3 proteínas PERIOD con Uniprot (http://www.uniprot.org) se observa conservación de ese residuo en todos ellos.

```
95    EFFQILSQNG--APQADVSMYSLEELATIASEHTSKNTDTFVAVFSFLSG 142   PER3_HUMAN
180   EYYQQWSLEEGEPCSMDMSTYTLEELEHITSEYTLQNQDTFSVAVSFLTG 229   PER1_HUMAN
153   EYYQLLMSSEGHPCGADVPSYTVEEMESVTSEHIVKNADMFAVAVSLVSG 202   PER2_HUMAN
```

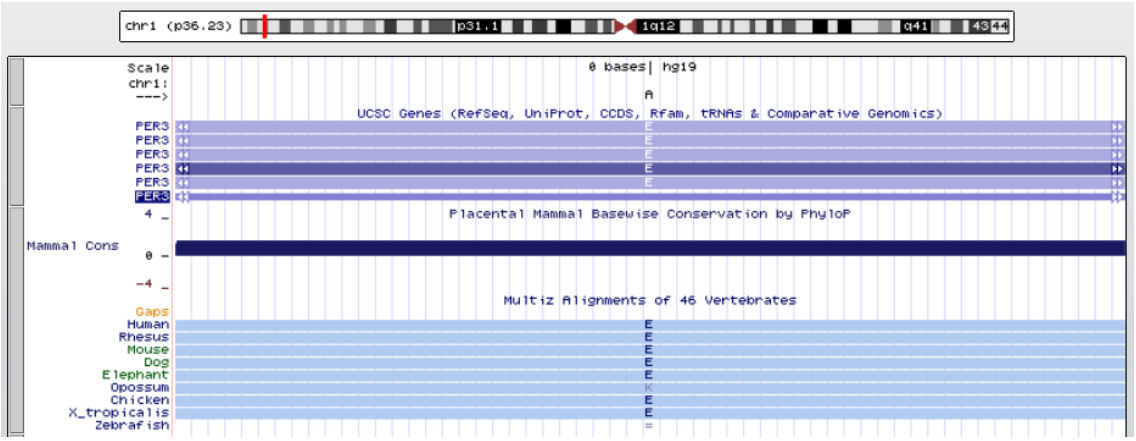



**Figura 6-2.** Captura de pantalla del UCSC Genome Browser (http://genome.ucsc.edu) que muestra la conservación del residuo de ácido glutámico en las distintas especies.

**Función del gen.**

El gen PERIOD3 codifica para la proteína PERIOD3. La función de esta proteína tiene relación con la regulación del ritmo circadiano. Éste determina, además de los patrones de sueño-vigilia, variaciones de presión arterial, temperatura corporal y movimientos intestinales. Igualmente, condiciona otros aspectos como variaciones en la actividad física, los procesos cognitivos o la conducta alimentaria.


**Figura 6-3.** Representación esquemática de los ritmos circadianos.

En los mamíferos, el reloj circadiano se sitúa en el núcleo supraquiasmático (NSQ) del hipotálamo, y utiliza la luz para su regulación. Comprende la secreción de diversas hormonas, siendo las más importantes la melatonina y el cortisol. Las neuronas del área dorsomedial del núcleo supraquiasmático utilizan como principales neurotransmisores el ácido gamma-aminobutírico (GABA) y AVP (arginina-vasopresina). La aferencia principal proviene de la retina, a través del tracto retino-hipotalámico (Moore & Lenn, 1972; Brown et al., 2008), utilizando principalmente glutamato. Otras conexiones del NSQ son el tálamo, a través del tracto geniculo-hipotalámico, que utiliza péptido Y, y los núcleos del rafe por neurotransmisión serotoninérgica. A través de sus conexiones con otros núcleos hipotalámicos regula la función del sistema nervioso autónomo y la secreción endocrina (Hastings, O'Neill, & Maywood, 2007).

Sin embargo, en ausencia de luz externa, el reloj interno permite mantener los patrones previamente descritos, pudiendo funcionar de forma autónoma. Otros estímulos, como la presencia de alimento, también pueden regular los ritmos. Estos estímulos no lumínicos modulan relojes circadianos externos al NSQ, situados en otras áreas del sistema nervioso central u otros órganos.

A nivel molecular el reloj circadiano se basa en una serie de bucles interrelacionados que se regulan unos a otros por mecanismos de feed-back (Ueda et al., 2005; Cardoso, de, Silva, & Cortez, 2009). El mecanismo molecular implicado se simplifica en la figura 6-4. Comienza con la síntesis de las proteínas CLOCK (CLK) y BMAL1 (o ARNTL, *aryl hydrocarbon receptor nuclear translocator-like*). Penetran en el núcleo y se unen formando un dímero CLK-BMAL1. NPAS2 (*neuronal PAS domain-containing protein 2*), análogo de CLK, puede realizar también una función semejante a éste. Este dímero activa, a través de la interacción con las E-box de los genes correspondientes, la síntesis de CRY (Cryptochrome) y los genes PERIOD 1, 2 y 3 (PER1, PER2 y PER3). Las cuatro proteínas salen al citoplasma e interaccionan unas con otras formando los dímeros CRY-PER2 y los trímeros CRY-PER1-PER3. Estos dímeros y trímeros penetran de nuevo en el núcleo inhibiendo la transcripción de CLK y BMAL1, formándose menos dímeros CLK-BMAL1, lo que inhibe a su vez la transcripción de CRY y PER. Los picos de RNA mensajero (es decir, la cantidad que se transcribe) de CLK y BMAL1 están desfasados unas 12 horas de los picos de RNA de CRY y PER, lo que hace que el ciclo completo dure 24 horas aproximadamente (Yamamoto et al., 2004; Lowrey & Takahashi, 2011; Mohawk, Green, & Takahashi, 2012).

Como mecanismos de control extra, el dímero CLK-BMAL1 activa la trascripción de los genes RORA y Rev-ErbA (*retinoic acid-related orphan nuclear receptors*). Estas moléculas se unen a sus receptores nucleares presentes en el promotor del gen BMAL1, regulando la transcripción de éste. RORA la activa mientras que Rev-ErbA la inhibe. Finalmente, el dímero CLK-BMAL1 activa también la trascripción de DEC. La acción que éste ejerce sobre BMAL1 todavía no está clara (Li et al., 2004; Bode, Shahmoradi, Taneja, Rossner, & Oster, 2011; Tsang et al., 2012).

Los procesos de fosforilación a través de kinasas (casein-kinasa 1 y GSK3beta) son los que determinan tanto la capacidad las moléculas de trasladarse dentro y fuera del núcleo como su degradación por ubiquitinización en el proteosoma (Harms, Young, & Saez, 2003).

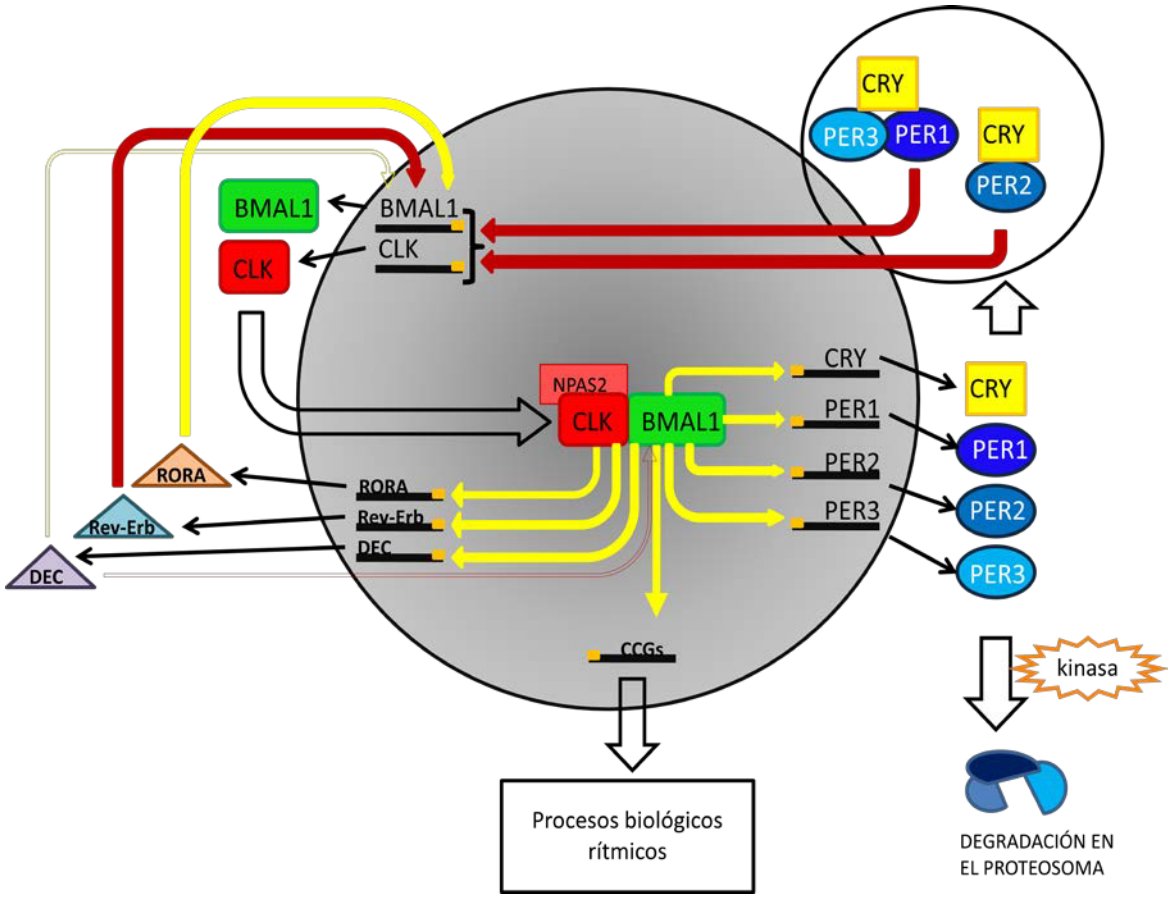



**Figura 6-4.** Reloj circadiano, bases moleculares. Flechas amarillas: activación; flechas rojas: inhibición; líneas negras: genes; cuadradillos naranjas: E-boxes. CCGs: genes controlados por el reloj circadiano (*clock controlled genes*).

**Ritmo circadiano y trastorno bipolar.**

La relación entre el ritmo circadiano y los trastornos del ánimo es sujeto de debate desde hace mucho tiempo, dado que síntomas como la alteración de los patrones de sueño, actividad y apetito son nucleares en dichos trastornos. En los años 60 comenzaron a estudiarse variaciones en el patrón circadiano de secreción de diversas hormonas y electrolitos en los trastornos afectivos (Knapp, Keane, & Wright, 1967; Lohrenz, Fullerton, Fahs, & Wenzel, 1968; Moody & Allsopp, 1969). En los 70 comenzaron a aparecer las primeras teorías al respecto (Kripke, Mullaney, Atkinson, & Wolf, 1978; von et al., 1985), progresivamente mejoradas y completadas con la caracterización de las bases moleculares del reloj circadiano a finales de la década de los 90 (Partonen, 1998; Solberg, Horton, & Turek, 1999; Bunney & Bunney, 2000). Se

postula que la capacidad del reloj circadiano de adaptarse a distintos estímulos externos es básica en la regulación del humor en respuesta a los cambios de estación, ritmos de sueño, niveles de estrés, etc. Una incapacidad para llevar a cabo esta regulación podría desencadenar la aparición de trastornos afectivos (Grandin, Alloy, & Abramson, 2006).

Estas teorías se ve apoyadas por hallazgos genéticos, que relacionan variaciones en los genes implicados en la regulación del ritmo circadiano y los trastornos del ánimo (McClung, 2007a; Mendlewicz, 2009; Kennaway, 2010; Etain et al., 2011). Igualmente, la existencia del trastorno depresivo estacional (episodios depresivos de aparición exclusiva en las épocas del año con menos horas de luz solar) y la eficacia de terapias como la privación de sueño (Vogel, Traub, Ben-Horin, & Meyers, 1968; Pflug & Tolle, 1971; Bunney & Bunney, 2012) (aunque de efecto breve) o la terapia lumínica (Lewy, Kern, Rosenthal, & Wehr, 1982; Pail et al., 2011) en el tratamiento de algunos trastornos afectivos señalan que existe una relación entre ambos fenómenos.

El mecanismo biológico por el que los ritmos circadianos interfieren en la regulación del humor todavía no está claro. Se ha objetivado que neurotransmisores como la serotonina, dopamina, noradrenalina o glutamato, y los enzimas implicados en su metabolismo, siguen patrones cíclicos de síntesis, secreción y eliminación, al igual que muchos de sus receptores. Por ejemplo, en ratones mutantes para el gen CLOCK se observa un aumento de activación de las neuronas dopaminérgicas del área tegmental ventral, generando unas alteraciones conductuales compatibles con un modelo de manía (McClung, 2007a). No es difícil plantear que alteraciones en los ritmos normales de estos sistemas pueden producir síntomas afectivos (Carlsson, Svennerholm, & Winblad, 1980; Siegel & Rogawski, 1988; Morin, 1999; Lambert, Reid, Kaye, Jennings, & Esler, 2002; Malek, Dardente, Pevet, & Raison, 2005).

El litio y el ácido valproico (principales tratamientos para el TBP) tienen importantes efectos sobre el ritmo circadiano, lo que podría explicar su eficacia en este trastorno. El litio altera los ritmos de sueño-vigilia, la latencia del sueño REM, y

otros parámetros como la temperatura corporal. A nivel molecular se ha observado que alarga el ciclo, produciendo un retardo de fase. Se cree que su acción tiene que ver con la actividad de GSK3beta (*glycogen synthase kinase 3 beta*), aunque se desconoce el mecanismo exacto, ya que el litio inhibe el enzima, lo que, de forma aislada, acorta el ciclo (Hirota et al., 2008). A nivel molecular, GSK3beta es una serina-treonina kinasa que regula los patrones de fosforilación de parte de los componentes del reloj circadiano, como PER2, CRY o Rev-Erb (Hirota et al., 2008). Paul y cols. han mostrado alteraciones en los ritmos circadianos y la conducta de ratones *knock-out* para este enzima (Paul, Johnson, Jope, & Gamble, 2012). Igualmente, Sahar y cols. observaron en animales que el estrés crónico aumenta la fosforilación de este enzima, alterando la expresión de PER2 (Sahar, Zocchi, Kinoshita, Borrelli, & Sassone-Corsi, 2010), lo que relacionaría el estrés con la aparición de trastornos afectivos. También se conoce que GSK3beta juega un papel esencial en la vía de señalización Wnt-beta catenina, que regula la expresión genética, y procesos de metabolismo, supervivencia y adhesión celular (Sun, Rodriguez, & Kim, 2009; Hur & Zhou, 2010; Amar, Belmaker, & Agam, 2011) (figura 6-5).

Otro fármaco, la ketamina (agonista glutamatérgico), ha mostrado un rápido e importante efecto antidepresivo, con aparición de respuesta en unas 24 horas (Diazgranados et al., 2010; Ibrahim et al., 2011). Parece que la ketamina bloquea el estímulo que supone el complejo CLOCK-BMAL1 para la transcripción de PER y CRY. El bloqueo de la vía Akt/GSK3beta hace desaparecer este efecto (Bellet, Vawter, Bunney, Bunney, & Sassone-Corsi, 2011), lo que apoya la participación de estos enzimas en el mecanismo de acción del fármaco.

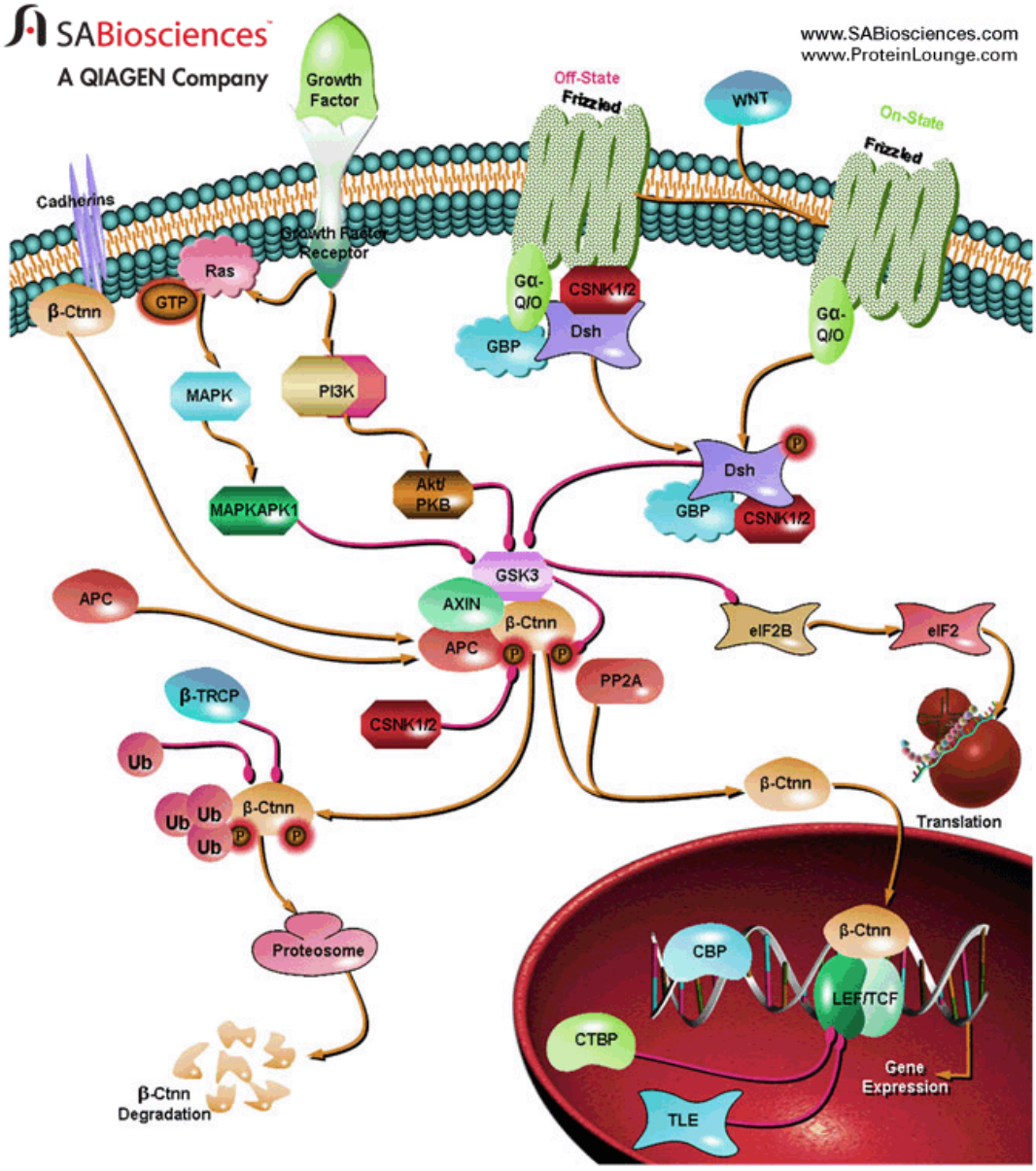



**Figura 6-5.** Exposición de las múltiples vías de señalización en las que está implicada GSK3. La imagen está tomada de la página web de Pathway Central (http://www.sabiosciences.com/).

**Estudios publicados.**

En 2001 Ebisawa y cols. secuenciaron el gen PERIOD3 encontrando 13 polimorfismos de nucleótido único, una región en el exoma 18 con una variante de número de copia de 54 pares de bases (los sujetos presentaban 4 o 5 copias del fragmento), y 4 indeles. Encontraron que el alelo corto (4 copias del polimorfismo) era más frecuente en los sujetos con un trastorno de retardo de fase del sueño. Posteriormente, el hallazgo fue replicado (Ebisawa et al., 2001; Archer et al., 2003). En

2008, Groeger y cols. estudiaron el efecto de ese polimorfismo sobre las alteraciones en distintas funciones cognitivas (atención, memoria y función ejecutiva) y motoras tras someter a los sujetos a una privación de sueño (40 horas seguidas de vigilia). Objetivaron que los sujetos homocigotos para el alelo largo (de 5 repeticiones) se desenvolvían significativamente peor en las tareas que requerían utilización de las funciones ejecutivas en las primeras horas de la mañana tras una noche sin dormir (Groeger et al., 2008). Ese mismo grupo comprobó que sus resultados se reflejaban en un patrón diferente de activación cerebral utilizando resonancia magnética funcional dependiendo del genotipo (Vandewalle et al., 2009). Los sujetos homocigotos para el alelo largo parecen tener una predominancia de la actividad simpática respecto de la parasimpática durante el sueño. Esta diferencia es más marcada durante la fase no REM y corresponde con variaciones en las medidas electroencefalográficas del patrón de sueño (Viola et al., 2007; Dijk & Archer, 2010). En un estudio reciente se ha relacionado este mismo polimorfismo con una diferente secreción de cortisol en un grupo de individuos sanos (Wirth et al., 2013).

Johansson y cols. estudiaron el polimorfismo 647 Val/Gly del gen PER3, encontrando una mayor preferencia por un patrón diurno de actividad según el *Home-Östberg morningness-eveningness questionnaire* en los sujetos portadores de, al menos, un alelo con Gly (Johansson et al., 2003).

El ratón *knock-out* para este gen (Per3-/-) (Shearman, Jin, Lee, Reppert, & Weaver, 2000) presenta una alteración de fase en algunos tejidos periféricos, aunque no en el núcleo supraquiasmático (Pendergast, Friday, & Yamazaki, 2010; Pendergast, Niswender, & Yamazaki, 2012) y unas variaciones de actividad alteradas respecto del salvaje según el patrón de luz-oscuridad (Van der Veen & Archer, 2010; Hasan, van, V, Winsky-Sommerer, Dijk, & Archer, 2011).

Centrándonos en el TBP, son numerosos los estudios que lo relacionan con alteraciones en los genes implicados en la regulación del ritmo circadiano (Roybal et al., 2007; McClung, 2007b; Murray & Harvey, 2010; McCarthy, Nievergelt, Kelsoe, & Welsh, 2012). Así, el gen PERIOD3 ha sido señalado previamente como gen candidato. Sin embargo, la evidencia respecto a PER3 en concreto es todavía bastante escasa.

Nievergelt y cols. (Nievergelt et al., 2006) encuentran evidencia de ligamiento del TBP con el gen PERIOD3 entre otros de los genes implicados en el reloj circadiano. En un estudio de Benedetti y cols. encuentran una relación con la homocigosis para el alelo largo de PER3 y una edad de aparición más precoz del TBP (Benedetti et al., 2008). Ese mismo grupo encuentra un predominio de debut de TBP como depresión post-parto en los homocigotos para el alelo corto (Dallaspezia et al., 2011). Finalmente, Rocha y cols. hallaron una relación entre una peor calidad de sueño en pacientes con TBP y el polimorfismo rs228727 del gen PER3 (Rocha et al., 2010).

### B. Gen Ubiquitin Specific Peptidase 29 (USP29).

USP29:NM_020903:exon4:c.861delC:p.F287fs. Cromosoma 19, exón 4. Delección con cambio en los tripletes de lectura. Pérdida de una citosina en la posición 57640904. La consecuencia en la proteína resultante (Q9HBJ7) es un cambio de secuencia proteica desde el aminoácido 288, manteniéndose en la posición 287 una fenilalanina.

**Patogenicidad de la mutación.**

A diferencia del caso anterior, en este caso el resultado de la mutación del gen es una proteína USP29 truncada, por lo que no se puede calcular la puntuación PolyPhen. Sin embargo, dada la importante alteración que esto supone en la proteína, será muy probablemente patogénica.

El área mutada está altamente conservada inter-especies, lo que apoya la importancia de su integridad (figura 6-6).

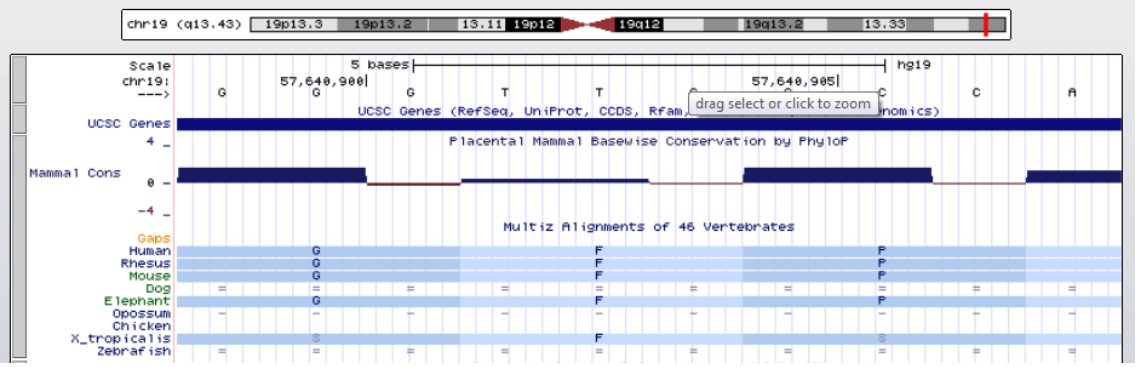



**Figura 6-6.** Captura de pantalla del UCSC Genome Browser (http://genome.ucsc.edu) que muestra la conservación de los residuos en este área en las distintas especies.

**Función del gen.**

El gen USP29 codifica para una hidrolasa ubiquitin carboxi-terminal tipo 2 que contiene dos dominios de proteasa de procesamiento específico de ubiquitina. Los procesos de ubiquitinización están muy extendidos en las distintas células, encargándose habitualmente de "marcar" moléculas que ya no son útiles para que sean posteriormente degradadas y eliminadas por el proteasoma.

En el caso de USP29, se ha visto que desubiquitina P53, estabilizándolo e inhibiendo su degradación, lo que desencadena en la célula un proceso de apoptosis (figura 6-7). La trascripción de USP29 se promueve por la asociación JTV1-FBP, que se activa frente al estrés oxidativo (Liu et al., 2011). Esta respuesta se ha observado, por ejemplo, en la respuesta al estrés asociada con el metabolismo de la dopamina (Ko et al., 2005).

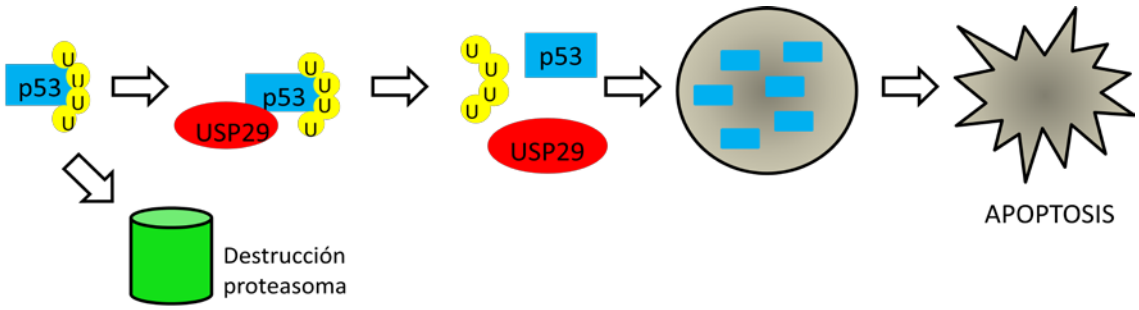



**Figura 6-7.** Representación esquemática de la función de USP29. La molécula p53 ubiquitinazada se degrada por el proteasoma. USP29 elimina las ubiquitinas estabilizando p53, por lo que se acumula en la célula, produciendo su apoptosis.

Es muy poco probable que una mutación en heterocigosis que produce una proteína truncada produzca una patología con una frecuencia tan escasa como la que encontramos en las formas familiares del TBP. No obstante, el gen USP29 sufre un fenómeno de *imprinting* por el que sólo se expresa la copia heredada del padre, lo que explicaría por qué aparecería un fenotipo patológico en los sujetos del presente estudio pese a tener sólo un alelo mutado (Buettner, Walker, & Singer-Sam, 2005; Huang & Kim, 2009; Kang et al., 2011).

**C. OTROS.**

Los otros dos genes hallados son de función aún no conocida. Las probablidades de que sea alguno de los dos el causante de la patología en esta familia son escasas por diversos motivos:

- El gen TMEM155 es muy variable, habiéndose descrito en él la presencia de múltiples SNPs. Hay descritas al menos 12 mutaciones con resultado de aparición de un codón de parada, por lo que si la proteína truncada resultante causara en heterocigosis un TBP familiar, la prevalencia de este trastorno sería mucho más elevada que la existente.
- En gen ANKRD31 es un gen extremadamente largo, con 21 exones, y muy variable. El resultado de la mutación que nos ocupa (delección en el exón 7) es una proteína no funcionante en heterocigosis. Siguiendo el mismo argumento que para el caso anterior, si este tipo de mutaciones produjera un TBP familiar, la frecuencia de éste sería mucho más elevada.

**D. ASPECTOS ÉTICOS** (Parker, 2002; Shendure et al., 2004; McGuire, Caulfield, & Cho, 2008; Tucker et al., 2009; Kaye, Boddington, de, Hawkins, & Melham, 2010; Singleton, 2011; Ong, Grody, & Deignan, 2011; Green et al., 2012; Cassa et al., 2012)

Desde el informe Belmont (http://www.hhs.gov/ohrp/humansubjects/guidance/belmont.html), los principios éticos de los estudios médicos se basan en la triada de respeto, beneficio y justicia. Las guías éticas y la legislación posterior recogen estos principios mediante la obligación de la participación voluntaria, el consentimiento informado, el derecho a la privacidad, la confidencialidad, la minimización del riesgo y la ausencia de discriminación. La adhesión a estos patrones tiene que ser monitorizada por los comités éticos de hospitales o agencias institucionales.

Los actuales estudios genéticos, dadas sus peculiaridades, plantean diversos dilemas éticos más allá de las normas aceptadas hasta ahora.

**Selección de la muestra. Investigación en familias.**

Cuando se plantea un estudio familiar es necesario inicialmente identificar aquellas familias más apropiadas para el estudio, generalmente aquellas con varios miembros afectados por la misma enfermedad. Habitualmente la información sobre la patología de los familiares parte de un sujeto inicial. Esto plantea el problema de si es o no ética la utilización de esta información personal de los familiares sin su consentimiento expreso para la investigación o, incluso, para contactar con ellos de cara a su posible inclusión. A este respecto, se recomienda que sea el propio sujeto inicial el que pregunte a los miembros de su familia si desean que se contacte con ellos para la realización del estudio. Esto no evita, en cualquier caso, la aparición de presiones familiares para que den el consentimiento.

**Consentimiento informado.**

Desde mediados del siglo XX se utiliza el consentimiento informado (CI) en la investigación biomédica para minimizar los riesgos de abuso o perjuicio del participante y asegurar sus derechos. Este documento debe recoger la participación voluntaria y en qué términos, y los derechos del paciente, tanto de los relativos al uso de sus datos personales como el de retirar su consentimiento en cualquier momento.

El CI en los estudios genéticos actuales, debido a su complejidad, presenta diversas peculiaridades.

1_ Se desconoce actualmente mucha de la información que aporta el genoma, y lo esperable es que en los próximos años ese conocimiento aumente. Esto abre la posibilidad de volver a analizar los datos en el futuro para obtener mayor información. ¿Habría que solicitar un nuevo consentimiento informado para hacerlo? ¿Se puede o se debe contactar de nuevo con el individuo para aportarle esa nueva información?

2_ La creación de bases de datos abiertas a los investigadores es una herramienta muy deseable para progresar en el conocimiento del genoma, pero abre la puerta a que esa

información pueda ser utilizada en nuevos estudios con fines distintos de aquellos para los que el paciente dio su consentimiento inicial. Para solucionar este problema se han propuesto diversas opciones:

-Volver a contactar con los individuos para que firmen nuevos consentimientos informados cada vez que se proponga utilizar sus muestras en estudios nuevos. Nadie duda de la complejidad de esta opción, siendo inviable en la mayoría de los casos.

-Creación de un consentimiento amplio o global con el que el sujeto consiente para múltiples usos en el contexto de la investigación, siendo necesaria una aprobación previa del comité de ética correspondiente. Es la solución más práctica, pero surgen dudas acerca de si es ético que el sujeto consienta sobre cosas que desconoce, ya que el principio fundamental del CI es que el sujeto conozca de antemano y pueda elegir cómo va a ser utilizada su información personal.

En cualquier caso, parece deseable que las bases de datos tengan medios para no permitir usos para los que no hubo consentimiento, y que los investigadores mantengan una integridad profesional asegurando que su investigación no sobrepasa los términos del consentimiento inicial.

Otro problema en relación con estas bases de datos es el de asegurar los derechos de los participantes a retirar el consentimiento en cualquier momento y que sus datos sean eliminados. Esto se vuelve muy complicado si estos datos se encuentran en bases abiertas a otros investigadores y están siendo utilizados en otros estudios.

3_ Es controvertida también la realización de estos estudios en determinados sujetos, como en los menores o en sujetos incapacitados. En el caso de los menores, si la patología fuera de inicio en la edad adulta, se podría demorar la decisión y el acceso a los resultados hasta la mayoría de edad. En el caso de enfermos mentales graves en dudosa capacidad de consentir, se debe consultar con el comité de ética.

En lo que hay un acuerdo general es en que habría que desarrollar un CI estándar para este tipo de estudios, compatible con las legislaciones existentes, que deberá recoger de forma clara los deseos del sujeto respecto a temas como:

-La información que desea o no recibir actualmente y en el futuro a medida se amplíen los conocimientos sobre el genoma humano.

-Los fines para los que consiente el uso de sus muestras y en qué términos, y la posibilidad de contactar de nuevo en el caso de necesitarse nuevo consentimiento.
-La posibilidad de que sus muestras e información genética sean almacenadas o compartidas, bajo qué términos y con quién.
Igualmente, se propone plantear los estudios para analizar el menor volumen de genoma necesario para obtener los resultados que se buscan.

**Devolución de la información obtenida a los participantes.**

Es indiscutible que los participantes en estos estudios tienen derecho a acceder a sus datos personales si así lo desean. Igualmente, tienen derecho a ser informados de los resultados obtenidos. Sin embargo, determinar cuál es la información concreta que hay que transmitir, de qué manera y quién es el responsable resulta todavía controvertido.

1_ Qué información transmitir y cómo hacerlo.

Los estudios de secuenciación genómica encuentran numerosas mutaciones en el genoma de cada individuo. La mayoría de ellas no tienen que ver con la enfermedad en estudio, representando hallazgos incidentales o imprevistos. El conocimiento actual no nos permite diferenciar en la mayoría de los casos cuáles de ellas son realmente patogénicas o de riesgo. Incluso en algunos casos en los que se ha encontrado una asociación válida entre una variante y una enfermedad se desconoce cómo interpretarla, dada la ausencia de conocimiento de la herencia, la penetrancia o el significado del estatus de portador. Hace falta experiencia para interpretar y comprender de forma adecuada las implicaciones sanitarias y sociales de las mutaciones identificadas.

Los estudios de secuenciación deberían realizarse con un protocolo de investigación formal que incluyera la forma de en la que se devuelven los resultados y la forma de realizar consejo genético. Hasta el momento actual no existen mecanismos estándar para la discusión y transmisión de los resultados de estos estudios. Numerosas instituciones han publicado recomendaciones sobre este tema, como la Comisión Nacional Asesora en Bioética de EEUU, el Centro de Control y Prevención de

Enfermedades, el grupo de trabajo del *National Heart Lung and Blood Institute* sobre la comunicación de resultados en los estudios genéticos (http://www.nhlbi.nih.gov), o el grupo Eurogentest (www.eurogentest.org), entre otras. Estas recomendaciones varían, pero todas ellas insisten en que la información transmitida tenga validez científica, significación clínica y la existencia de posibilidad de una intervención médica beneficiosa. En el caso de variantes de riesgo conocido, asociadas a enfermedades graves para las que existe tratamiento, los investigadores deben informar al participante. En el caso de variantes de riesgo dudoso, o que asocien patologías de escasa gravedad o sin tratamiento, el beneficio de informar a los participantes tiene que ser sopesado frente al derecho de éstos de no ser informados.

2_ Transmisión de información sensible a los familiares de los participantes.

En los estudios de secuenciación se obtiene información que tiene implicaciones no sólo para el propio sujeto del estudio sino también para otros miembros de su familia, y la transmisión de esta información puede generar un conflicto ético.

Se propone que durante el proceso de consentimiento informado se explique a los participantes esta posibilidad y que se les anime a incluir a los familiares cercanos en la decisión de participar en estos estudios, no siendo en principio necesario realizar un consentimiento informado a éstos. Pero puede suceder que el paciente se niegue a que la información sea trasmitida a sus familiares. La Sociedad Americana de Genética Humana recoge que la información no autorizada del riesgo genético está éticamente justificada si:

-Los intentos para convencer al paciente han sido infructuosos.

-El daño potencial en los familiares es muy probable, inminente, previsible y grave.

-El sujeto en riesgo es identificable.

-La enfermedad es tratable, prevenible o los estándares de cuidado aceptan que su detección precoz reduce el riesgo o mejora el pronóstico.

**Privacidad y confidencialidad de los datos.**

El mantenimiento de la confidencialidad en los estudios de investigación se ha estado asegurando hasta el momento actual con herramientas como la anonimización de los datos y la seguridad informática, lo que ha permitido el acceso a los datos exclusivamente a las personas autorizadas. Actualmente, estos medios se han quedado insuficientes en relación con:

-La extensa información que aportan los estudios de secuenciación.

-La creación de bases de datos abiertas a numerosos investigadores, necesarias para progresar en el estudio del genoma.

Ambas circunstancias facilitan que esa información, pese a estar anonimizada, pueda ser relacionada con el sujeto, aumentando las posibilidades de re-identificación de los individuos, además de aportar información sobre sus familiares. Para proteger la privacidad de los sujetos participantes se han propuesto nuevas estrategias como la limitación del volumen de genoma que se comparte, la degradación de los datos o la de-identificación con códigos.

Además de información relativa a enfermedades, en los estudios genéticos se puede obtener otro tipo de información que puede ser de interés y podría tener importantes repercusiones legales, como la relativa a paternidades. ¿Qué hacer con esta información?

Igualmente, se plantea si terceras personas ajenas a la investigación podrían acceder a la información disponible para fines concretos y, a priori, lícitos, como la investigación de delitos o la identificación de víctimas de grandes desastres.

**Control institucional.**

Además de por los consensos internacionales (informe Belmont, declaración de Helsinki), la investigación biomédica está regulada por legislaciones específicas de cada país. En el caso de España, está sujeta a la Ley 14/2007 de Investigación Biomédica, que regula todo aquello relacionado con la disciplina, incluyendo normativa acerca de biobancos. Respecto a la protección de datos, está regulada por la Ley Orgánica 15/1999 de Protección de Datos de carácter personal. Esta normativa se completa con

guías de práctica y requerimientos de las propias instituciones, principalmente los comités de ética en la investigación clínica (CEIC).

En el momento actual, y dada la generalización de las bases de datos compartidas discutida previamente, cabe reflexionar sobre si es necesaria una legislación internacional que regule estos temas de forma que sea semejante en todos los países. Igualmente se plantea si los CEIC tal y como han funcionado hasta este momento son suficientes para realizar un adecuado seguimiento y control del respeto de los derechos de los participantes en estos estudios. En algunos de los grandes proyectos internacionales se han creado Comités de acceso específicos, que supervisan quién accede a los datos y bajo qué condiciones. Todavía, sin embargo, no hay consenso claro sobre qué criterios seguir a la hora de decidir cuáles deben ser estos requisitos.

## E. CONSIDERACIONES FINALES

La secuenciación de exoma completo ha demostrado ser una herramienta útil y eficiente en el descubrimiento de los genes responsables de enfermedades con herencia mendeliana. Sin embargo, hasta ahora no se ha utilizado de forma generalizada para el estudio de enfermedades de herencia compleja. La existencia de familias con patologías de herencia compleja que presentan una aparente herencia mendeliana (como la que se muestra en este estudio) y el hallazgo a través de esta técnica de un gen responsable, apoya la teoría de que una parte de la heredabilidad de estas enfermedades se puede explicar por mutaciones de muy escasa frecuencia pero gran penetrancia, de herencia familiar o debido a mutaciones esporádicas. El hallazgo de estas variantes familiares puede tener escasa utilidad diagnóstica a corto plazo, pero es muy importante para la elaboración de hipótesis fisiopatológicas, permitiendo el desarrollo de modelos biológicos y facilitando la investigación de posibles tratamientos.

El haber logrado limitar a dos las posibles mutaciones responsables del trastorno es un resultado extraordinario, dada la multitud de mutaciones que aparecen en el exoma de cualquier ser humano. Este resultado, sin embargo, nos obliga a continuar la investigación hasta poder discriminar cuál de las dos mutaciones es realmente la causante de la enfermedad. Esto requiere de posteriores estudios celulares y con animales transgénicos de los que pueden derivarse nuevos modelos biológicos de TBP.

# 7. CONCLUSIONES.

1. La secuenciación de exoma completo se dibuja como una herramienta útil para el estudio de las enfermedades de herencia compleja, lo que abre una nueva vía de investigación para aclarar las bases genéticas de las enfermedades psiquiátricas.

2. Los resultados de este estudio apoyan la hipótesis de que una parte de la heredabilidad del trastorno bipolar se puede explicar a través de mutaciones de escasa frecuencia y alta penetrancia, heredadas o *de novo*.

3. El trastorno bipolar de herencia autosómica dominante presente en la familia de estudio está probablemente relacionado con la mutación del gen PERIOD3, aunque no puede descartarse que la responsable sea la mutación del gen USP29 o que sea necesaria la presencia de ambas mutaciones al tiempo.

4. El hecho de que una de las dos mutaciones candidatas se sitúe en el gen PERIOD3, siendo este un gen señalado como candidato en algunos estudios de trastorno bipolar, sustentaría las teorías que relacionan esta enfermedad con una alteración en los ritmos circadianos.

# REFERENCIAS BIBLIOGRÁFICAS

Akhras, M. S., Unemo, M., Thiyagarajan, S., Nyren, P., Davis, R. W., Fire, A. Z. et al. (2007). Connector inversion probe technology: a powerful one-primer multiplex DNA amplification system for numerous scientific applications. *PLoS.One., 2,* e915.

Akiskal, H. S., Bourgeois, M. L., Angst, J., Post, R., Moller, H., & Hirschfeld, R. (2000). Re-evaluating the prevalence of and diagnostic composition within the broad clinical spectrum of bipolar disorders. *J Affect.Disord, 59 Suppl 1,* S5-S30.

Albert, T. J., Molla, M. N., Muzny, D. M., Nazareth, L., Wheeler, D., Song, X. et al. (2007). Direct selection of human genomic loci by microarray hybridization. *Nat Methods, 4,* 903-905.

Alkan, C., Kidd, J. M., Marques-Bonet, T., Aksay, G., Antonacci, F., Hormozdiari, F. et al. (2009). Personalized copy number and segmental duplication maps using next-generation sequencing. *Nat Genet, 41,* 1061-1067.

Amar, S., Belmaker, R. H., & Agam, G. (2011). The possible involvement of glycogen synthase kinase-3 (GSK-3) in diabetes, cancer and central nervous system diseases. *Curr.Pharm.Des, 17,* 2264-2277.

American Psychiatric Association (2000). *Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition: DSM-IV-TR®*. American Psychiatric Pub.

Archer, S. N., Robilliard, D. L., Skene, D. J., Smits, M., Williams, A., Arendt, J. et al. (2003). A length polymorphism in the circadian clock gene Per3 is linked to delayed sleep phase syndrome and extreme diurnal preference. *Sleep, 26,* 413-415.

Arnone, D., Cavanagh, J., Gerber, D., Lawrie, S. M., Ebmeier, K. P., & McIntosh, A. M. (2009). Magnetic resonance imaging studies in bipolar disorder and schizophrenia: meta-analysis. *Br.J Psychiatry, 195,* 194-201.

Badner, J. A. & Gershon, E. S. (2002). Meta-analysis of whole-genome linkage scans of bipolar disorder and schizophrenia. *Mol.Psychiatry, 7,* 405-411.

Balanza-Martinez, V., Rubio, C., Selva-Vera, G., Martinez-Aran, A., Sanchez-Moreno, J., Salazar-Fraile, J. et al. (2008). Neurocognitive endophenotypes (endophenocognitypes) from studies of relatives of bipolar disorder subjects: a systematic review. *Neurosci.Biobehav.Rev., 32,* 1426-1438.

Barnett, J. H. & Smoller, J. W. (2009). The genetics of bipolar disorder. *Neuroscience, 164,* 331-343.

Baum, A. E., Akula, N., Cabanero, M., Cardona, I., Corona, W., Klemens, B. et al. (2008a). A genome-wide association study implicates diacylglycerol kinase eta (DGKH) and several other genes in the etiology of bipolar disorder. *Mol.Psychiatry, 13,* 197-207.

Baum, A. E., Hamshere, M., Green, E., Cichon, S., Rietschel, M., Noethen, M. M. et al. (2008b). Meta-analysis of two genome-wide association studies of bipolar disorder reveals important points of agreement. *Mol.Psychiatry, 13,* 466-467.

Bellet, M. M., Vawter, M. P., Bunney, B. G., Bunney, W. E., & Sassone-Corsi, P. (2011). Ketamine influences CLOCK:BMAL1 function leading to altered circadian gene expression. *PLoS.One., 6,* e23982.

Benedetti, F., Dallaspezia, S., Colombo, C., Pirovano, A., Marino, E., & Smeraldi, E. (2008). A length polymorphism in the circadian clock gene Per3 influences age at onset of bipolar disorder. *Neurosci.Lett., 445,* 184-187.

Benitez, B. A., Alvarado, D., Cai, Y., Mayo, K., Chakraverty, S., Norton, J. et al. (2011). Exome-sequencing confirms DNAJC5 mutations as cause of adult neuronal ceroid-lipofuscinosis. *PLoS.One., 6,* e26741.

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G. et al. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature, 456,* 53-59.

Bergen, S. E., O'Dushlaine, C. T., Ripke, S., Lee, P. H., Ruderfer, D. M., Akterin, S. et al. (2012). Genome-wide association study in a Swedish population yields support for greater CNV and MHC involvement in schizophrenia compared with bipolar disorder. *Mol.Psychiatry, 17,* 880-886.

Bertelsen, A., Harvald, B., & Hauge, M. (1977). A Danish twin study of manic-depressive disorders. *Br.J Psychiatry, 130,* 330-351.

Beyer, J. L., Young, R., Kuchibhatla, M., & Krishnan, K. R. (2009). Hyperintense MRI lesions in bipolar disorder: A meta-analysis and review. *Int.Rev.Psychiatry, 21,* 394-409.

Biesecker, L. G., Shianna, K. V., & Mullikin, J. C. (2011). Exome sequencing: the expert view. *Genome Biol., 12,* 128.

Bigos, K. L., Mattay, V. S., Callicott, J. H., Straub, R. E., Vakkalanka, R., Kolachana, B. et al. (2010). Genetic variation in CACNA1C affects brain circuitries related to mental illness. *Arch Gen.Psychiatry, 67,* 939-945.

Bilguvar, K., Ozturk, A. K., Louvi, A., Kwan, K. Y., Choi, M., Tatli, B. et al. (2010). Whole-exome sequencing identifies recessive WDR62 mutations in severe brain malformations. *Nature, 467,* 207-210.

Birney, E., Stamatoyannopoulos, J. A., Dutta, A., Guigo, R., Gingeras, T. R., Margulies, E. H. et al. (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature, 447,* 799-816.

Bode, B., Shahmoradi, A., Taneja, R., Rossner, M. J., & Oster, H. (2011). Genetic interaction of Per1 and Dec1/2 in the regulation of circadian locomotor activity. *J Biol.Rhythms, 26,* 530-540.

Bolze, A., Byun, M., McDonald, D., Morgan, N. V., Abhyankar, A., Premkumar, L. et al. (2010). Whole-exome-sequencing-based discovery of human FADD deficiency. *Am.J Hum.Genet, 87,* 873-881.

Bonnefond, A., Durand, E., Sand, O., De, G. F., Gallina, S., Busiah, K. et al. (2010). Molecular diagnosis of neonatal diabetes mellitus using next-generation sequencing of the whole exome. *PLoS.One., 5,* e13630.

Branton, D., Deamer, D. W., Marziali, A., Bayley, H., Benner, S. A., Butler, T. et al. (2008). The potential and challenges of nanopore sequencing. *Nat Biotechnol., 26,* 1146-1153.

Braslavsky, I., Hebert, B., Kartalov, E., & Quake, S. R. (2003). Sequence information can be obtained from single DNA molecules. *Proc.Natl.Acad.Sci U.S.A, 100,* 3960-3964.

Bredy, T. W., Grant, R. J., Champagne, D. L., & Meaney, M. J. (2003). Maternal care influences neuronal survival in the hippocampus of the rat. *Eur.J Neurosci., 18,* 2903-2909.

Brown, S. A., Kunz, D., Dumas, A., Westermark, P. O., Vanselow, K., Tilmann-Wahnschaffe, A. et al. (2008). Molecular insights into human daily behavior. *Proc.Natl.Acad.Sci U.S.A, 105,* 1602-1607.

Buettner, V. L., Walker, A. M., & Singer-Sam, J. (2005). Novel paternally expressed intergenic transcripts at the mouse Prader-Willi/Angelman Syndrome locus. *Mamm.Genome, 16,* 219-227.

Bunney, B. G. & Bunney, W. E. (2012). Mechanisms of Rapid Antidepressant Effects of Sleep Deprivation Therapy: Clock Genes and Circadian Rhythms. *Biol.Psychiatry*.

Bunney, W. E. & Bunney, B. G. (2000). Molecular clock genes in man and lower animals: possible implications for circadian abnormalities in depression. *Neuropsychopharmacology, 22,* 335-345.

Cardno, A. G., Marshall, E. J., Coid, B., Macdonald, A. M., Ribchester, T. R., Davies, N. J. et al. (1999). Heritability estimates for psychotic disorders: the Maudsley twin psychosis series. *Arch Gen.Psychiatry, 56,* 162-168.

Cardoso, F. R., de, O. C. F., Silva, D., & Cortez, C. M. (2009). A simple model for circadian timing by mammals. *Braz.J Med.Biol.Res., 42,* 122-127.

Carlsson, A., Svennerholm, L., & Winblad, B. (1980). Seasonal and circadian monoamine variations in human brains examined post mortem. *Acta Psychiatr.Scand.Suppl, 280,* 75-85.

Cassa, C. A., Savage, S. K., Taylor, P. L., Green, R. C., McGuire, A. L., & Mandl, K. D. (2012). Disclosing pathogenic genetic variants to research participants: quantifying an emerging ethical responsibility. *Genome Res., 22,* 421-428.

Champagne, F. & Meaney, M. J. (2001). Like mother, like daughter: evidence for non-genomic transmission of parental behavior and stress responsivity. *Prog.Brain Res., 133,* 287-302.

Chanock, S. J., Manolio, T., Boehnke, M., Boerwinkle, E., Hunter, D. J., Thomas, G. et al. (2007). Replicating genotype-phenotype associations. *Nature, 447,* 655-660.

Chen, D. T., Jiang, X., Akula, N., Shugart, Y. Y., Wendland, J. R., Steele, C. J. et al. (2011a). Genome-wide association study meta-analysis of European and Asian-ancestry samples identifies three novel loci associated with bipolar disorder. *Mol.Psychiatry*.

Chen, W. J., Lin, Y., Xiong, Z. Q., Wei, W., Ni, W., Tan, G. H. et al. (2011b). Exome sequencing identifies truncating mutations in PRRT2 that cause paroxysmal kinesigenic dyskinesia. *Nat Genet, 43,* 1252-1255.

Chen, Z., Cui, L., Li, M., Jiang, L., Deng, W., Ma, X. et al. (2012). Voxel based morphometric and diffusion tensor imaging analysis in male bipolar patients with first-episode mania. *Prog.Neuropsychopharmacol.Biol.Psychiatry, 36,* 231-238.

Choi, M., Scholl, U. I., Ji, W., Liu, T., Tikhonova, I. R., Zumbo, P. et al. (2009). Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proc.Natl.Acad.Sci U.S.A, 106,* 19096-19101.

Chubb, J. E., Bradshaw, N. J., Soares, D. C., Porteous, D. J., & Millar, J. K. (2008). The DISC locus in psychiatric illness. *Mol.Psychiatry, 13,* 36-64.

Cichon, S., Craddock, N., Daly, M., Faraone, S. V., Gejman, P. V., Kelsoe, J. et al. (2009). Genomewide association studies: history, rationale, and prospects for psychiatric disorders. *Am.J Psychiatry, 166,* 540-556.

Cichon, S., Muhleisen, T. W., Degenhardt, F. A., Mattheisen, M., Miro, X., Strohmaier, J. et al. (2011). Genome-wide association study identifies genetic variation in neurocan as a susceptibility factor for bipolar disorder. *Am.J Hum.Genet, 88,* 372-381.

Clark, L., Kempton, M. J., Scarna, A., Grasby, P. M., & Goodwin, G. M. (2005). Sustained attention-deficit confirmed in euthymic bipolar disorder but not in first-degree relatives of bipolar patients or euthymic unipolar depression. *Biol.Psychiatry, 57,* 183-187.

Clark, L., Sarna, A., & Goodwin, G. M. (2005). Impairment of executive function but not memory in first-degree relatives of patients with bipolar I disorder and in euthymic patients with unipolar depression. *Am.J Psychiatry, 162,* 1980-1982.

Clayton-Smith, J., O'Sullivan, J., Daly, S., Bhaskar, S., Day, R., Anderson, B. et al. (2011). Whole-exome-sequencing identifies mutations in histone acetyltransferase gene KAT6B in individuals with the Say-Barber-Biesecker variant of Ohdo syndrome. *Am.J Hum.Genet, 89,* 675-681.

Coffey, A. J., Kokocinski, F., Calafato, M. S., Scott, C. E., Palta, P., Drury, E. et al. (2011). The GENCODE exome: sequencing the complete human exome. *Eur.J Hum.Genet, 19,* 827-831.

Collins, F. S., Morgan, M., & Patrinos, A. (2003). The Human Genome Project: lessons from large-scale biology. *Science, 300,* 286-290.

Cooper, G. M., Goode, D. L., Ng, S. B., Sidow, A., Bamshad, M. J., Shendure, J. et al. (2010). Single-nucleotide evolutionary constraint scores highlight disease-causing mutations. *Nat Methods, 7,* 250-251.

Craddock, N. & Jones, I. (1999). Genetics of bipolar disorder. *J Med.Genet, 36,* 585-594.

Craddock, N., Khodel, V., Van, E. P., & Reich, T. (1995). Mathematical limits of multilocus models: the genetic transmission of bipolar disorder. *Am.J Hum.Genet, 57,* 690-702.

Craddock, N., O'Donovan, M. C., & Owen, M. J. (2005). The genetics of schizophrenia and bipolar disorder: dissecting psychosis. *J Med.Genet, 42,* 193-204.

Craddock, N. & Sklar, P. (2009). Genetics of bipolar disorder: successful start to a long journey. *Trends Genet, 25,* 99-105.

Cui, L., Chen, Z., Deng, W., Huang, X., Li, M., Ma, X. et al. (2011). Assessment of white matter abnormalities in paranoid schizophrenia and bipolar mania patients. *Psychiatry Res., 194,* 347-353.

Dahl, F., Gullberg, M., Stenberg, J., Landegren, U., & Nilsson, M. (2005). Multiplex amplification enabled by selective circularization of large sets of genomic DNA fragments. *Nucleic Acids Res., 33,* e71.

Dallaspezia, S., Lorenzi, C., Pirovano, A., Colombo, C., Smeraldi, E., & Benedetti, F. (2011). Circadian clock gene Per3 variants influence the postpartum onset of bipolar disorder. *Eur.Psychiatry, 26,* 138-140.

Daoud, H., Zhou, S., Noreau, A., Sabbagh, M., Belzil, V., onne-Laporte, A. et al. (2012). Exome sequencing reveals SPG11 mutations causing juvenile ALS. *Neurobiol.Aging, 33,* 839.

Davis, B. D. (1990). The human genome and other initiatives. *Science, 249,* 342-343.

De, P. L., Crescini, A., Deste, G., Fusar-Poli, P., Sacchetti, E., & Vita, A. (2012). Brain structural abnormalities at the onset of schizophrenia and bipolar disorder: a meta-analysis of controlled magnetic resonance imaging studies. *Curr.Pharm.Des, 18,* 486-494.

Delvecchio, G., Sugranyes, G., & Frangou, S. (2012). Evidence of diagnostic specificity in the neural correlates of facial affect processing in bipolar disorder and schizophrenia: a meta-analysis of functional imaging studies. *Psychol.Med.,* 1-17.

Diazgranados, N., Ibrahim, L., Brutsche, N. E., Newberg, A., Kronstein, P., Khalife, S. et al. (2010). A Randomized Add-on Trial of an N-methyl-d-aspartate Antagonist in Treatment-Resistant Bipolar Depression. *Arch Gen.Psychiatry, 67,* 793-802.

Dijk, D. J. & Archer, S. N. (2010). PERIOD3, circadian phenotypes, and sleep homeostasis. *Sleep Med.Rev., 14,* 151-160.

Djurovic, S., Gustafsson, O., Mattingsdal, M., Athanasiu, L., Bjella, T., Tesli, M. et al. (2010). A genome-wide association study of bipolar disorder in Norwegian individuals, followed by replication in Icelandic sample. *J Affect.Disord, 126,* 312-316.

Doi, H., Yoshida, K., Yasuda, T., Fukuda, M., Fukuda, Y., Morita, H. et al. (2011). Exome sequencing reveals a homozygous SYT14 mutation in adult-onset, autosomal-recessive spinocerebellar ataxia with psychomotor retardation. *Am.J Hum.Genet, 89,* 320-327.

Ebisawa, T., Uchiyama, M., Kajimura, N., Mishima, K., Kamei, Y., Katoh, M. et al. (2001). Association of structural polymorphisms in the human period3 gene with delayed sleep phase syndrome. *EMBO Rep., 2,* 342-346.

Escamilla, M. A. & Zavala, J. M. (2008). Genetics of bipolar disorder. *Dialogues.Clin.Neurosci., 10,* 141-152.

Etain, B., Milhiet, V., Bellivier, F., & Leboyer, M. (2011). Genetics of circadian rhythms and mood spectrum disorders. *Eur.Neuropsychopharmacol., 21 Suppl 4,* S676-S682.

Fañañas Saura, L. & Sáiz Ruiz, J. (2000). *Manual de introducción a la genética en psiquiatría*. Masson.

Ferreira, M. A., O'Donovan, M. C., Meng, Y. A., Jones, I. R., Ruderfer, D. M., Jones, L. et al. (2008). Collaborative genome-wide association analysis supports a role for ANK3 and CACNA1C in bipolar disorder. *Nat Genet, 40,* 1056-1058.

Flicek, P., Amode, M. R., Barrell, D., Beal, K., Brent, S., Carvalho-Silva, D. et al. (2012). Ensembl 2012. *Nucleic Acids Res., 40,* D84-D90.

Garrett, A. S., Reiss, A. L., Howe, M. E., Kelley, R. G., Singh, M. K., Adleman, N. E. et al. (2012). Abnormal amygdala and prefrontal cortex activation to facial expressions in pediatric bipolar disorder. *J Am.Acad.Child Adolesc.Psychiatry, 51,* 821-831.

Gershon, E. S. & Goldin, L. R. (1986). Clinical methods in psychiatric genetics. I. Robustness of genetic marker investigative strategies. *Acta Psychiatr.Scand., 74,* 113-118.

Gilissen, C., Arts, H. H., Hoischen, A., Spruijt, L., Mans, D. A., Arts, P. et al. (2010). Exome sequencing identifies WDR35 variants involved in Sensenbrenner syndrome. *Am.J Hum.Genet, 87,* 418-423.

Girirajan, S., Campbell, C. D., & Eichler, E. E. (2011). Human copy number variation and complex genetic disease. *Annu.Rev.Genet, 45,* 203-226.

Glassner, B. & Haldipur, C. V. (1983). Life events and early and late onset of bipolar disorder. *Am.J Psychiatry, 140,* 215-217.

Gnirke, A., Melnikov, A., Maguire, J., Rogov, P., LeProust, E. M., Brockman, W. et al. (2009). Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat Biotechnol., 27,* 182-189.

Gottesman, I. I. & Shields, J. (1973). Genetic theorizing and schizophrenia. *Br.J Psychiatry, 122,* 15-30.

Grandin, L. D., Alloy, L. B., & Abramson, L. Y. (2006). The social zeitgeber theory, circadian rhythms, and mood disorders: review and evaluation. *Clin.Psychol.Rev., 26,* 679-694.

Green, R. C., Berg, J. S., Berry, G. T., Biesecker, L. G., Dimmock, D. P., Evans, J. P. et al. (2012). Exploring concordance and discordance for return of incidental findings from clinical sequencing. *Genet Med., 14,* 405-410.

Groeger, J. A., Viola, A. U., Lo, J. C., von, S. M., Archer, S. N., & Dijk, D. J. (2008). Early morning executive functioning during sleep deprivation is compromised by a PERIOD3 polymorphism. *Sleep, 31,* 1159-1167.

Grof, P., Duffy, A., Alda, M., & Hajek, T. (2009). Lithium response across generations. *Acta Psychiatr.Scand., 120,* 378-385.

Grozeva, D., Kirov, G., Ivanov, D., Jones, I. R., Jones, L., Green, E. K. et al. (2010). Rare copy number variants: a point of rarity in genetic risk for bipolar disorder and schizophrenia. *Arch Gen.Psychiatry, 67,* 318-327.

Haack, T. B., Danhauser, K., Haberberger, B., Hoser, J., Strecker, V., Boehm, D. et al. (2010). Exome sequencing identifies ACAD9 mutations as a cause of complex I deficiency. *Nat Genet, 42,* 1131-1134.

Hales RE, Yudofsky SC, & Gabbard GO (2008). *The American Psychiatric Publishing Textbook of Psychiatry*. (5th ed.).

Hallahan, B., Newell, J., Soares, J. C., Brambilla, P., Strakowski, S. M., Fleck, D. E. et al. (2011). Structural magnetic resonance imaging in bipolar disorder: an international collaborative mega-analysis of individual adult patient data. *Biol.Psychiatry, 69,* 326-335.

Harismendy, O., Ng, P. C., Strausberg, R. L., Wang, X., Stockwell, T. B., Beeson, K. Y. et al. (2009). Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biol., 10,* R32.

Harms, E., Young, M. W., & Saez, L. (2003). CK1 and GSK3 in the Drosophila and mammalian circadian clock. *Novartis.Found.Symp., 253,* 267-277.

Harris, T. D., Buzby, P. R., Babcock, H., Beer, E., Bowers, J., Braslavsky, I. et al. (2008). Single-molecule DNA sequencing of a viral genome. *Science, 320,* 106-109.

Hasan, S., van, d., V, Winsky-Sommerer, R., Dijk, D. J., & Archer, S. N. (2011). Altered sleep and behavioral activity phenotypes in PER3-deficient mice. *Am.J Physiol Regul.Integr.Comp Physiol, 301,* R1821-R1830.

Hastings, M., O'Neill, J. S., & Maywood, E. S. (2007). Circadian clocks: regulators of endocrine and metabolic rhythms. *J Endocrinol., 195,* 187-198.

Hattori, E., Toyota, T., Ishitsuka, Y., Iwayama, Y., Yamada, K., Ujike, H. et al. (2009). Preliminary genome-wide association study of bipolar disorder in the Japanese population. *Am.J Med.Genet B Neuropsychiatr.Genet, 150B,* 1110-1117.

Hayden, E. P. & Nurnberger, J. I., Jr. (2006). Molecular genetics of bipolar disorder. *Genes Brain Behav., 5,* 85-95.

Hill, S. K., Harris, M. S., Herbener, E. S., Pavuluri, M., & Sweeney, J. A. (2008). Neurocognitive allied phenotypes for schizophrenia and bipolar disorder. *Schizophr.Bull., 34,* 743-759.

Hirota, T., Lewis, W. G., Liu, A. C., Lee, J. W., Schultz, P. G., & Kay, S. A. (2008). A chemical biology approach reveals period shortening of the mammalian circadian clock by specific inhibition of GSK-3beta. *Proc.Natl.Acad.Sci U.S.A, 105,* 20746-20751.

Hodges, E., Xuan, Z., Balija, V., Kramer, M., Molla, M. N., Smith, S. W. et al. (2007). Genome-wide in situ exon capture for selective resequencing. *Nat Genet, 39,* 1522-1527.

Huang, J. M. & Kim, J. (2009). DNA methylation analysis of the mammalian PEG3 imprinted domain. *Gene, 442,* 18-25.

Hummer, T. A., Hulvershorn, L. A., Karne, H. S., Gunn, A. D., Wang, Y., & Anand, A. (2012). Emotional Response Inhibition in Bipolar Disorder: A Functional Magnetic Resonance Imaging Study of Trait- and State-Related Abnormalities. *Biol.Psychiatry*.

Hur, E. M. & Zhou, F. Q. (2010). GSK3 signalling in neural development. *Nat Rev.Neurosci., 11,* 539-551.

Hyman, E. D. (1988). A new method of sequencing DNA. *Anal.Biochem., 174,* 423-436.

Ibrahim, L., Diazgranados, N., Luckenbaugh, D. A., hado-Vieira, R., Baumann, J., Mallinger, A. G. et al. (2011). Rapid Decrease in Depressive Symptoms with an N-methyl-D-aspartate Antagonist in ECT-Resistant Major Depression. *Prog.Neuropsychopharmacol.Biol.Psychiatry, 35,* 1155-1159.

International Human Genome Sequencing Consortium (2004). Finishing the euchromatic sequence of the human genome. *Nature, 431,* 931-945.

Jimenez Escrig, A. (2007). *Textbook of Neurogenetics*. (1 ed.) Díaz Santos, S. A.

Jimenez-Escrig, A., Gobernado, I., Garcia-Villanueva, M., & Sanchez-Herranz, A. (2012). Autosomal recessive Emery-Dreifuss muscular dystrophy caused by a novel mutation (R225Q) in the lamin A/C gene identified by exome sequencing. *Muscle Nerve, 45,* 605-610.

Jimenez-Escrig, A., Gobernado, I., & Sanchez-Herranz, A. (2012). [Whole genome sequencing: a qualitative leap forward in genetic studies]. *Rev.Neurol, 54,* 692-698.

Jogia, J., Ruberto, G., Lelli-Chiesa, G., Vassos, E., Maieru, M., Tatarelli, R. et al. (2011). The impact of the CACNA1C gene polymorphism on frontolimbic function in bipolar disorder. *Mol.Psychiatry, 16,* 1070-1071.

Johansson, C., Willeit, M., Smedh, C., Ekholm, J., Paunio, T., Kieseppa, T. et al. (2003). Circadian clock-related polymorphisms in seasonal affective disorder and their relevance to diurnal preference. *Neuropsychopharmacology, 28,* 734-739.

Johnson, J. O., Gibbs, J. R., Van, M. L., Houlden, H., & Singleton, A. B. (2010). Exome sequencing in Brown-Vialetto-van Laere syndrome. *Am.J Hum.Genet, 87,* 567-569.

Johnson, J. O., Mandrioli, J., Benatar, M., Abramzon, Y., Van, D., V, Trojanowski, J. Q. et al. (2010). Exome sequencing reveals VCP mutations as a cause of familial ALS. *Neuron, 68,* 857-864.

Jones, I. & Craddock, N. (2001). Candidate gene studies of bipolar disorder. *Ann.Med., 33,* 248-256.

Kang, E. R., Iqbal, K., Tran, D. A., Rivas, G. E., Singh, P., Pfeifer, G. P. et al. (2011). Effects of endocrine disruptors on imprinted gene expression in the mouse embryo. *Epigenetics., 6,* 937-950.

Kaye, J., Boddington, P., de, V. J., Hawkins, N., & Melham, K. (2010). Ethical implications of the use of whole genome methods in medical research. *Eur.J Hum.Genet, 18,* 398-403.

Kempton, M. J., Geddes, J. R., Ettinger, U., Williams, S. C., & Grasby, P. M. (2008). Meta-analysis, database, and meta-regression of 98 structural imaging studies in bipolar disorder. *Arch Gen.Psychiatry, 65,* 1017-1032.

Kendler, K. S., Pedersen, N. L., Neale, M. C., & Mathe, A. A. (1995). A pilot Swedish twin study of affective illness including hospital- and population-ascertained subsamples: results of model fitting. *Behav.Genet, 25,* 217-232.

Kennaway, D. J. (2010). Clock genes at the heart of depression. *J Psychopharmacol., 24,* 5-14.

Kerner, B., Lambert, C. G., & Muthen, B. O. (2011). Genome-wide association study in bipolar patients stratified by co-morbidity. *PLoS.One., 6,* e28477.

Kessler, R. C., Akiskal, H. S., Ames, M., Birnbaum, H., Greenberg, P., Hirschfeld, R. M. et al. (2006). Prevalence and effects of mood disorders on work performance in a nationally representative sample of U.S. workers. *Am.J Psychiatry, 163,* 1561-1568.

Kessler, R. C., Berglund, P., Demler, O., Jin, R., Merikangas, K. R., & Walters, E. E. (2005). Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the National Comorbidity Survey Replication. *Arch Gen.Psychiatry, 62,* 593-602.

Kieseppa, T., Partonen, T., Haukka, J., Kaprio, J., & Lonnqvist, J. (2004). High concordance of bipolar I disorder in a nationwide sample of twins. *Am.J Psychiatry, 161,* 1814-1821.

Kim, P., Thomas, L. A., Rosen, B. H., Moscicki, A. M., Brotman, M. A., Zarate, C. A., Jr. et al. (2012). Differing amygdala responses to facial expressions in children and adults with bipolar disorder. *Am.J Psychiatry, 169,* 642-649.

Klimes-Dougan, B., Ronsaville, D., Wiggs, E. A., & Martinez, P. E. (2006). Neuropsychological functioning in adolescent children of mothers with a history of bipolar or major depressive disorders. *Biol.Psychiatry, 60,* 957-965.

Knapp, M. S., Keane, P. M., & Wright, J. G. (1967). Circadian rhythm of plasma 11-hydroxycorticosteroids in depressive illness, congestive heart failure, and Cushing's syndrome. *Br.Med.J, 2,* 27-30.

Ko, H. S., von, C. R., Sriram, S. R., Kim, S. W., Chung, K. K., Pletnikova, O. et al. (2005). Accumulation of the authentic parkin substrate aminoacyl-tRNA synthetase cofactor, p38/JTV-1, leads to catecholaminergic cell death. *J Neurosci., 25,* 7968-7978.

Kripke, D. F., Mullaney, D. J., Atkinson, M., & Wolf, S. (1978). Circadian rhythm disorders in manic-depressives. *Biol.Psychiatry, 13,* 335-351.

Kuhn, R. M., Haussler, D., & Kent, W. J. (2012). The UCSC genome browser and associated tools. *Brief.Bioinform.*.

Kulkarni, S., Jain, S., Janardhan Reddy, Y. C., Kumar, K. J., & Kandavel, T. (2010). Impairment of verbal learning and memory and executive function in unaffected siblings of probands with bipolar disorder. *Bipolar.Disord, 12,* 647-656.

Lachman, H. M., Pedrosa, E., Petruolo, O. A., Cockerham, M., Papolos, A., Novak, T. et al. (2007). Increase in GSK3beta gene copy number variation in bipolar disorder. *Am.J Med.Genet B Neuropsychiatr.Genet, 144B,* 259-265.

Lalonde, E., Albrecht, S., Ha, K. C., Jacob, K., Bolduc, N., Polychronakos, C. et al. (2010). Unexpected allelic heterogeneity and spectrum of mutations in Fowler syndrome revealed by next-generation exome sequencing. *Hum.Mutat., 31,* 918-923.

Lam, K., Guo, H., Wilson, G. A., Kohl, S., & Wong, F. (2011). Identification of variants in CNGA3 as cause for achromatopsia by exome sequencing of a single patient. *Arch Ophthalmol., 129,* 1212-1217.

Lambert, G. W., Reid, C., Kaye, D. M., Jennings, G. L., & Esler, M. D. (2002). Effect of sunlight and season on serotonin turnover in the brain. *Lancet, 360,* 1840-1842.

Le-Niculescu, H., Patel, S. D., Bhat, M., Kuczenski, R., Faraone, S. V., Tsuang, M. T. et al. (2009). Convergent functional genomics of genome-wide association data for bipolar disorder: comprehensive identification of candidate genes, pathways and mechanisms. *Am.J Med.Genet B Neuropsychiatr.Genet, 150B,* 155-181.

Lee, M. T., Chen, C. H., Lee, C. S., Chen, C. C., Chong, M. Y., Ouyang, W. C. et al. (2011). Genome-wide association study of bipolar I disorder in the Han Chinese population. *Mol.Psychiatry, 16,* 548-556.

Lehne, B., Lewis, C. M., & Schlitt, T. (2011). Exome localization of complex disease association signals. *BMC.Genomics, 12,* 92.

Lewy, A. J., Kern, H. A., Rosenthal, N. E., & Wehr, T. A. (1982). Bright artificial light treatment of a manic-depressive patient with a seasonal mood cycle. *Am.J Psychiatry, 139,* 1496-1498.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N. et al. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics., 25,* 2078-2079.

Li, M. X., Gui, H. S., Kwan, J. S., Bao, S. Y., & Sham, P. C. (2012). A comprehensive framework for prioritizing variants in exome sequencing studies of Mendelian diseases. *Nucleic Acids Res., 40,* e53.

Li, Y., Song, X., Ma, Y., Liu, J., Yang, D., & Yan, B. (2004). DNA binding, but not interaction with Bmal1, is responsible for DEC1-mediated transcription regulation of the circadian gene mPer1. *Biochem.J, 382,* 895-904.

Linke, J., King, A. V., Rietschel, M., Strohmaier, J., Hennerici, M., Gass, A. et al. (2012a). Increased medial orbitofrontal and amygdala activation: evidence for a systems-level endophenotype of bipolar I disorder. *Am.J Psychiatry, 169,* 316-325.

Linke, J., Witt, S. H., King, A. V., Nieratschker, V., Poupon, C., Gass, A. et al. (2012b). Genome-wide supported risk variant for bipolar disorder alters anatomical connectivity in the human brain. *Neuroimage., 59,* 3288-3296.

Liu, J., Chung, H. J., Vogt, M., Jin, Y., Malide, D., He, L. et al. (2011). JTV1 co-activates FBP to induce USP29 transcription and stabilize p53 in response to oxidative stress. *EMBO J, 30,* 846-858.

Lohrenz, F. N., Fullerton, D. T., Fahs, H., & Wenzel, F. J. (1968). Adrenocortical function in depressive states--study of circadian variation in plasma and urinary steroids. *Int.J Neuropsychiatry, 4,* 21-25.

Lowrey, P. L. & Takahashi, J. S. (2011). Genetics of circadian rhythms in Mammalian model organisms. *Adv.Genet, 74,* 175-230.

Lydall, G. J., Bass, N. J., McQuillin, A., Lawrence, J., Anjorin, A., Kandaswamy, R. et al. (2011). Confirmation of prior evidence of genetic susceptibility to alcoholism in a genome-wide association study of comorbid alcoholism and bipolar disorder. *Psychiatr.Genet, 21,* 294-306.

Malek, Z. S., Dardente, H., Pevet, P., & Raison, S. (2005). Tissue-specific expression of tryptophan hydroxylase mRNAs in the rat midbrain: anatomical evidence and daily profiles. *Eur.J Neurosci., 22,* 895-901.

Malhotra, D., McCarthy, S., Michaelson, J. J., Vacic, V., Burdick, K. E., Yoon, S. et al. (2011). High frequencies of de novo CNVs in bipolar disorder and schizophrenia. *Neuron, 72,* 951-963.

Mamanova, L., Coffey, A. J., Scott, C. E., Kozarewa, I., Turner, E. H., Kumar, A. et al. (2010). Target-enrichment strategies for next-generation sequencing. *Nat Methods, 7,* 111-118.

Mardis, E. R. (2011). A decade's perspective on DNA sequencing technology. *Nature, 470,* 198-203.

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A. et al. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature, 437,* 376-380.

Marti-Masso, J. F., Ruiz-Martinez, J., Makarov, V., Lopez de, M. A., Gorostidi, A., Bergareche, A. et al. (2012). Exome sequencing identifies GCDH (glutaryl-CoA dehydrogenase) mutations as a cause of a progressive form of early-onset generalized dystonia. *Hum.Genet, 131,* 435-442.

Martinowich, K., Schloesser, R. J., & Manji, H. K. (2009). Bipolar disorder: from genes to behavior pathways. *J Clin.Invest, 119,* 726-736.

Marx, J. (2007). Behavioral genetics. Evidence linking DISC1 gene to mental illness builds. *Science, 318,* 1062-1063.

Maxam, A. M. & Gilbert, W. (1977). A new method for sequencing DNA. *Proc.Natl.Acad.Sci U.S.A, 74,* 560-564.

McCarthy, M. J., Nievergelt, C. M., Kelsoe, J. R., & Welsh, D. K. (2012). A survey of genomic studies supports association of circadian clock genes with bipolar disorder spectrum illnesses and lithium response. *PLoS.One., 7,* e32091.

McClung, C. A. (2007b). Clock genes and bipolar disorder: implications for therapy. *Pharmacogenomics., 8,* 1097-1100.

McClung, C. A. (2007a). Circadian genes, rhythms and the biology of mood disorders. *Pharmacol.Ther., 114,* 222-232.

McDonald, C., Zanelli, J., Rabe-Hesketh, S., Ellison-Wright, I., Sham, P., Kalidindi, S. et al. (2004). Meta-analysis of magnetic resonance imaging brain morphometry studies in bipolar disorder. *Biol.Psychiatry, 56,* 411-417.

McGuire, A. L., Caulfield, T., & Cho, M. K. (2008). Research ethics and the challenge of whole-genome sequencing. *Nat Rev.Genet, 9,* 152-156.

McQueen, M. B., Devlin, B., Faraone, S. V., Nimgaonkar, V. L., Sklar, P., Smoller, J. W. et al. (2005). Combined analysis from eleven linkage studies of bipolar disorder provides strong evidence of susceptibility loci on chromosomes 6q and 8q. *Am.J Hum.Genet, 77,* 582-595.

Mendlewicz, J. (2009). Disruption of the circadian timing systems: molecular mechanisms in mood disorders. *CNS.Drugs, 23 Suppl 2,* 15-26.

Mendlewicz, J. & Rainer, J. D. (1977). Adoption study supporting genetic transmission in manic--depressive illness. *Nature, 268,* 327-329.

Merikangas, K. R., Akiskal, H. S., Angst, J., Greenberg, P. E., Hirschfeld, R. M., Petukhova, M. et al. (2007). Lifetime and 12-month prevalence of bipolar spectrum disorder in the National Comorbidity Survey replication. *Arch Gen.Psychiatry, 64,* 543-552.

Merikangas, K. R., Chakravarti, A., Moldin, S. O., Araj, H., Blangero, J. C., Burmeister, M. et al. (2002). Future of genetics of mood disorders research. *Biol.Psychiatry, 52,* 457-477.

Merikangas, K. R. & Low, N. C. (2004). The epidemiology of mood disorders. *Curr.Psychiatry Rep., 6,* 411-421.

Metzker, M. L. (2005). Emerging technologies in DNA sequencing. *Genome Res., 15,* 1767-1776.

Metzker, M. L. (2010). Sequencing technologies - the next generation. *Nat Rev.Genet, 11,* 31-46.

Miklowitz, D. J., Goldstein, M. J., Nuechterlein, K. H., Snyder, K. S., & Mintz, J. (1988). Family factors and the course of bipolar affective disorder. *Arch Gen.Psychiatry, 45,* 225-231.

Min, B. J., Kim, N., Chung, T., Kim, O. H., Nishimura, G., Chung, C. Y. et al. (2011). Whole-exome sequencing identifies mutations of KIF22 in spondyloepimetaphyseal dysplasia with joint laxity, leptodactylic type. *Am.J Hum.Genet, 89,* 760-766.

Miro, X., Meier, S., Dreisow, M. L., Frank, J., Strohmaier, J., Breuer, R. et al. (2012). Studies in humans and mice implicate neurocan in the etiology of mania. *Am.J Psychiatry, 169,* 982-990.

Mohawk, J. A., Green, C. B., & Takahashi, J. S. (2012). Central and peripheral circadian clocks in mammals. *Annu.Rev.Neurosci., 35,* 445-462.

Moody, J. P. & Allsopp, M. N. (1969). Circadian rhythms of water and electrolyte excretion in manic-depressive psychosis. *Br.J Psychiatry, 115,* 923-928.

Moore, R. Y. & Lenn, N. J. (1972). A retinohypothalamic projection in the rat. *J Comp Neurol, 146,* 1-14.

Morin, L. P. (1999). Serotonin and the regulation of mammalian circadian rhythmicity. *Ann.Med., 31,* 12-33.

Mu, X. J., Lu, Z. J., Kong, Y., Lam, H. Y., & Gerstein, M. B. (2011). Analysis of genomic variation in non-coding elements using population-scale sequencing data from the 1000 Genomes Project. *Nucleic Acids Res., 39,* 7058-7076.

Murphy, K. C. (2002). Schizophrenia and velo-cardio-facial syndrome. *Lancet, 359,* 426-430.

Murray, G. & Harvey, A. (2010). Circadian rhythms and sleep in bipolar disorder. *Bipolar.Disord, 12,* 459-472.

Musunuru, K., Pirruccello, J. P., Do, R., Peloso, G. M., Guiducci, C., Sougnez, C. et al. (2010). Exome sequencing, ANGPTL3 mutations, and familial combined hypolipidemia. *N.Engl.J Med., 363,* 2220-2227.

Ng, S. B., Bigham, A. W., Buckingham, K. J., Hannibal, M. C., McMillin, M. J., Gildersleeve, H. I. et al. (2010). Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. *Nat Genet, 42,* 790-793.

Ng, S. B., Turner, E. H., Robertson, P. D., Flygare, S. D., Bigham, A. W., Lee, C. et al. (2009). Targeted capture and massively parallel sequencing of 12 human exomes. *Nature, 461,* 272-276.

Nievergelt, C. M., Kripke, D. F., Barrett, T. B., Burg, E., Remick, R. A., Sadovnick, A. D. et al. (2006). Suggestive evidence for association of the circadian genes PERIOD3 and ARNTL with bipolar disorder. *Am.J Med.Genet B Neuropsychiatr.Genet, 141B,* 234-241.

Nurnberger, J. I., Jr. (2012). General genetics of bipolar disorder. In S.M.Strakowski (Ed.), *The Bipolar Brain* (pp. 187-202). New York: Oxford University Press.

Nurnberger, J. I., Jr., Adkins, S., Lahiri, D. K., Mayeda, A., Hu, K., Lewy, A. et al. (2000). Melatonin suppression by light in euthymic bipolar and unipolar patients. *Arch Gen.Psychiatry, 57,* 572-579.

O'Roak, B. J., Deriziotis, P., Lee, C., Vives, L., Schwartz, J. J., Girirajan, S. et al. (2011). Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. *Nat Genet, 43,* 585-589.

Oedegaard, K. J., Greenwood, T. A., Johansson, S., Jacobsen, K. K., Halmoy, A., Fasmer, O. B. et al. (2010). A genome-wide association study of bipolar disorder and comorbid migraine. *Genes Brain Behav., 9,* 673-680.

Okou, D. T., Locke, A. E., Steinberg, K. M., Hagen, K., Athri, P., Shetty, A. C. et al. (2009). Combining microarray-based genomic selection (MGS) with the Illumina Genome Analyzer platform to sequence diploid target regions. *Ann.Hum.Genet, 73,* 502-513.

Okou, D. T., Steinberg, K. M., Middle, C., Cutler, D. J., Albert, T. J., & Zwick, M. E. (2007). Microarray-based genomic selection for high-throughput resequencing. *Nat Methods, 4,* 907-909.

Olsen, L., Hansen, T., Djurovic, S., Haastrup, E., Albrecthsen, A., Hoeffding, L. K. et al. (2011). Copy number variations in affective disorders and meta-analysis. *Psychiatr.Genet, 21,* 319-322.

Ong, F. S., Grody, W. W., & Deignan, J. L. (2011). Privacy and data management in the era of massively parallel next-generation sequencing. *Expert.Rev.Mol.Diagn., 11,* 457-459.

Owen, M. J., Williams, H. J., & O'Donovan, M. C. (2009). Schizophrenia genetics: advancing on two fronts. *Curr.Opin.Genet Dev., 19,* 266-270.

Ozgul, R. K., Siemiatkowska, A. M., Yucel, D., Myers, C. A., Collin, R. W., Zonneveld, M. N. et al. (2011). Exome sequencing and cis-regulatory mapping identify mutations in MAK, a gene encoding a regulator of ciliary length, as a cause of retinitis pigmentosa. *Am.J Hum.Genet, 89,* 253-264.

Pail, G., Huf, W., Pjrek, E., Winkler, D., Willeit, M., Praschak-Rieder, N. et al. (2011). Bright-light therapy in the treatment of mood disorders. *Neuropsychobiology, 64,* 152-162.

Pandey, A., Davis, N. A., White, B. C., Pajewski, N. M., Savitz, J., Drevets, W. C. et al. (2012). Epistasis network centrality analysis yields pathway replication across two GWAS cohorts for bipolar disorder. *Transl.Psychiatry, 2,* e154.

Papolos, D. F., Faedda, G. L., Veit, S., Goldberg, R., Morrow, B., Kucherlapati, R. et al. (1996). Bipolar spectrum disorders in patients diagnosed with velo-cardio-facial syndrome: does a hemizygous deletion of chromosome 22q11 result in bipolar affective disorder? *Am.J Psychiatry, 153,* 1541-1547.

Parker, L. S. (2002). Ethical issues in bipolar disorders pedigree research: privacy concerns, informed consent, and grounds for waiver. *Bipolar.Disord, 4,* 1-16.

Partonen, T. (1998). One pacemaker in seasonal affective disorder. *Med.Hypotheses, 51,* 297-298.

Paul, J. R., Johnson, R. L., Jope, R. S., & Gamble, K. L. (2012). Disruption of circadian rhythmicity and suprachiasmatic action potential frequency in a mouse model with constitutive activation of glycogen synthase kinase 3. *Neuroscience, 226,* 1-9.

Pendergast, J. S., Friday, R. C., & Yamazaki, S. (2010). Distinct functions of Period2 and Period3 in the mouse circadian system revealed by in vitro analysis. *PLoS.One., 5,* e8552.

Pendergast, J. S., Niswender, K. D., & Yamazaki, S. (2012). Tissue-specific function of Period3 in circadian rhythmicity. *PLoS.One., 7,* e30254.

Perez de Castro, I., Santos, J., Torres, P., Visedo, G., Saiz-Ruiz, J., Llinares, C. et al. (1995). A weak association between TH and DRD2 genes and bipolar affective disorder in a Spanish sample. *J Med.Genet, 32,* 131-134.

Perrier, E., Pompei, F., Ruberto, G., Vassos, E., Collier, D., & Frangou, S. (2011). Initial evidence for the role of CACNA1C on subcortical brain morphology in patients with bipolar disorder. *Eur.Psychiatry, 26,* 135-137.

Pflug, B. & Tolle, R. (1971). Disturbance of the 24-hour rhythm in endogenous depression and the treatment of endogenous depression by sleep deprivation. *Int.Pharmacopsychiatry, 6,* 187-196.

Pierson, T. M., Adams, D., Bonn, F., Martinelli, P., Cherukuri, P. F., Teer, J. K. et al. (2011). Whole-exome sequencing identifies homozygous AFG3L2 mutations in a spastic ataxia-neuropathy syndrome linked to mitochondrial m-AAA proteases. *PLoS.Genet, 7,* e1002325.

Porreca, G. J., Zhang, K., Li, J. B., Xie, B., Austin, D., Vassallo, S. L. et al. (2007). Multiplex amplification of large sets of human exons. *Nat Methods, 4,* 931-936.

Porteous, D. (2008). Genetic causality in schizophrenia and bipolar disorder: out with the old and in with the new. *Curr.Opin.Genet Dev., 18,* 229-234.

Priebe, L., Degenhardt, F. A., Herms, S., Haenisch, B., Mattheisen, M., Nieratschker, V. et al. (2012). Genome-wide survey implicates the influence of copy number variants (CNVs) in the development of early-onset bipolar disorder. *Mol.Psychiatry, 17,* 421-432.

Pruitt, K. D., Tatusova, T., Brown, G. R., & Maglott, D. R. (2012). NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res., 40,* D130-D135.

Purcell, S. M., Wray, N. R., Stone, J. L., Visscher, P. M., O'Donovan, M. C., Sullivan, P. F. et al. (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature, 460,* 748-752.

QIAGEN. (2012). QIAamp DNA Blood Midi/Maxi Handbook. Third Edition. Ref Type: Catalog

Quintin, P., Benkelfat, C., Launay, J. M., Arnulf, I., Pointereau-Bellenger, A., Barbault, S. et al. (2001). Clinical and neurochemical effect of acute tryptophan depletion in unaffected relatives of patients with bipolar affective disorder. *Biol.Psychiatry, 50,* 184-190.

Raffan, E., Hurst, L. A., Turki, S. A., Carpenter, G., Scott, C., Daly, A. et al. (2011). Early Diagnosis of Werner's Syndrome Using Exome-Wide Sequencing in a Single, Atypical Patient. *Front Endocrinol.(Lausanne), 2,* 8.

Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T. D. et al. (2006). Global variation in copy number in the human genome. *Nature, 444,* 444-454.

Reva, B., Antipin, Y., & Sander, C. (2011). Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res., 39,* e118.

Rios, J., Stein, E., Shendure, J., Hobbs, H. H., & Cohen, J. C. (2010). Identification by whole-genome resequencing of gene defect responsible for severe hypercholesterolemia. *Hum.Mol.Genet, 19,* 4313-4318.

Ripke, S., Sanders, A. R., Kendler, K. S., Levinson, D. F., Sklar, P., Holmans, P. A. et al. (2011). Genome-wide association study identifies five new schizophrenia loci. *Nat Genet, 43,* 969-976.

Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G. et al. (2011a). Integrative Genomics Viewer. *Nature Biotechnology, 29,* 24-26.

Robinson, P. N. (2010). Whole-exome sequencing for finding de novo mutations in sporadic mental retardation. *Genome Biol., 11,* 144.

Robinson, P. N., Krawitz, P., & Mundlos, S. (2011b). Strategies for exome and genome sequence data analysis in disease-gene discovery projects. *Clin.Genet, 80,* 127-132.

Rocha, P. M., Neves, F. S., Alvarenga, N. B., Hughet, R. B., Barbosa, I. G., & Correa, H. (2010). Association of Per3 gene with bipolar disorder: comment on "Association study of 21 circadian genes with bipolar I disorder, schizoaffective disorder, and schizophrenia". *Bipolar.Disord, 12,* 875-876.

Ronaghi, M., Karamohamed, S., Pettersson, B., Uhlen, M., & Nyren, P. (1996). Real-time DNA sequencing using detection of pyrophosphate release. *Anal.Biochem., 242,* 84-89.

Ronaghi, M., Nygren, M., Lundeberg, J., & Nyren, P. (1999). Analyses of secondary structures in DNA by pyrosequencing. *Anal.Biochem., 267,* 65-71.

Ronaghi, M., Pettersson, B., Uhlen, M., & Nyren, P. (1998). PCR-introduced loop structure as primer in DNA sequencing. *Biotechniques, 25,* 876-2, 884.

Roybal, K., Theobold, D., Graham, A., DiNieri, J. A., Russo, S. J., Krishnan, V. et al. (2007). Mania-like behavior induced by disruption of CLOCK. *Proc.Natl.Acad.Sci U.S.A, 104,* 6406-6411.

Sahar, S., Zocchi, L., Kinoshita, C., Borrelli, E., & Sassone-Corsi, P. (2010). Regulation of BMAL1 protein stability and circadian function by GSK3beta-mediated phosphorylation. *PLoS.One., 5,* e8561.

Sanders, S. J., Murtha, M. T., Gupta, A. R., Murdoch, J. D., Raubeson, M. J., Willsey, A. J. et al. (2012). De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature, 485,* 237-241.

Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proc.Natl.Acad.Sci U.S.A, 74,* 5463-5467.

Savitz, J., van der, M. L., & Ramesar, R. (2008). Personality endophenotypes for bipolar affective disorder: a family-based genetic association analysis. *Genes Brain Behav., 7,* 869-876.

Savitz, J. B. & Ramesar, R. S. (2006). Personality: is it a viable endophenotype for genetic studies of bipolar affective disorder? *Bipolar.Disord, 8,* 322-337.

Schadt, E. E., Turner, S., & Kasarskis, A. (2010). A window into third-generation sequencing. *Hum.Mol.Genet, 19,* R227-R240.

Schulze, K. K., Walshe, M., Stahl, D., Hall, M. H., Kravariti, E., Morris, R. et al. (2011). Executive functioning in familial bipolar I disorder patients and their unaffected relatives. *Bipolar.Disord, 13,* 208-216.

Schulze, T. G., tera-Wadleigh, S. D., Akula, N., Gupta, A., Kassem, L., Steele, J. et al. (2009). Two variants in Ankyrin 3 (ANK3) are independent genetic risk factors for bipolar disorder. *Mol.Psychiatry, 14,* 487-491.

Scott, L. J., Muglia, P., Kong, X. Q., Guan, W., Flickinger, M., Upmanyu, R. et al. (2009). Genome-wide association and meta-analysis of bipolar disorder in individuals of European ancestry. *Proc.Natl.Acad.Sci U.S.A, 106,* 7501-7506.

Segurado, R., tera-Wadleigh, S. D., Levinson, D. F., Lewis, C. M., Gill, M., Nurnberger, J. I., Jr. et al. (2003). Genome scan meta-analysis of schizophrenia and bipolar disorder, part III: Bipolar disorder. *Am.J Hum.Genet, 73,* 49-62.

Seidman, L. J., Kremen, W. S., Koren, D., Faraone, S. V., Goldstein, J. M., & Tsuang, M. T. (2002). A comparative profile analysis of neuropsychological functioning in patients with schizophrenia and bipolar psychoses. *Schizophr.Res., 53,* 31-44.

Seifuddin, F., Mahon, P. B., Judy, J., Pirooznia, M., Jancic, D., Taylor, J. et al. (2012). Meta-analysis of genetic association studies on bipolar disorder. *Am.J Med.Genet B Neuropsychiatr.Genet, 159B,* 508-518.

Serretti, A. & Mandelli, L. (2008). The genetics of bipolar disorder: genome 'hot regions,' genes, new potential candidates and future directions. *Mol.Psychiatry, 13,* 742-771.

Shearman, L. P., Jin, X., Lee, C., Reppert, S. M., & Weaver, D. R. (2000). Targeted disruption of the mPer3 gene: subtle effects on circadian clock function. *Mol.Cell Biol., 20,* 6269-6275.

Shendure, J., Mitra, R. D., Varma, C., & Church, G. M. (2004). Advanced sequencing technologies: methods and goals. *Nat Rev.Genet, 5,* 335-344.

Shendure, J., Porreca, G. J., Reppas, N. B., Lin, X., McCutcheon, J. P., Rosenbaum, A. M. et al. (2005). Accurate multiplex polony sequencing of an evolved bacterial genome. *Science, 309,* 1728-1732.

Shih, R. A., Belmonte, P. L., & Zandi, P. P. (2004). A review of the evidence from family, twin and adoption studies for a genetic contribution to adult psychiatric disorders. *Int.Rev.Psychiatry, 16,* 260-283.

Shihab, H. A., Gough, J., Cooper, D. N., Stenson, P. D., Barker, G. L., Edwards, K. J. et al. (2013). Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Hum.Mutat., 34,* 57-65.

Siegel, J. M. & Rogawski, M. A. (1988). A function for REM sleep: regulation of noradrenergic receptor sensitivity. *Brain Res., 472,* 213-233.

Singleton, A. B. (2011). Exome sequencing: a transformative technology. *Lancet Neurol, 10,* 942-946.

Sklar, P., Smoller, J. W., Fan, J., Ferreira, M. A., Perlis, R. H., Chambert, K. et al. (2008). Whole-genome association study of bipolar disorder. *Mol.Psychiatry, 13,* 558-569.

Smith, E. N., Bloss, C. S., Badner, J. A., Barrett, T., Belmonte, P. L., Berrettini, W. et al. (2009). Genome-wide association study of bipolar disorder in European American and African American individuals. *Mol.Psychiatry, 14,* 755-763.

Smith, E. N., Koller, D. L., Panganiban, C., Szelinger, S., Zhang, P., Badner, J. A. et al. (2011b). Genome-wide association of bipolar disorder suggests an enrichment of replicable associations in regions near genes. *PLoS.Genet, 7,* e1002134.

Smith, E. N., Koller, D. L., Panganiban, C., Szelinger, S., Zhang, P., Badner, J. A. et al. (2011a). Genome-wide association of bipolar disorder suggests an enrichment of replicable associations in regions near genes. *PLoS.Genet, 7,* e1002134.

Smith, K. R., Bromhead, C. J., Hildebrand, M. S., Shearer, A. E., Lockhart, P. J., Najmabadi, H. et al. (2011c). Reducing the exome search space for mendelian diseases using genetic linkage analysis of exome genotypes. *Genome Biol., 12,* R85.

Smoller, J. W. & Finn, C. T. (2003). Family, twin, and adoption studies of bipolar disorder. *Am.J Med.Genet C.Semin.Med.Genet, 123C,* 48-58.

Sobczak, S., Honig, A., Nicolson, N. A., & Riedel, W. J. (2002). Effects of acute tryptophan depletion on mood and cortisol release in first-degree relatives of type I and type II bipolar patients and healthy matched controls. *Neuropsychopharmacology, 27,* 834-842.

Solberg, L. C., Horton, T. H., & Turek, F. W. (1999). Circadian rhythms and depression: effects of exercise in an animal model. *Am.J Physiol, 276,* R152-R161.

Soni, S., Whittington, J., Holland, A. J., Webb, T., Maina, E. N., Boer, H. et al. (2008). The phenomenology and diagnosis of psychiatric illness in people with Prader-Willi syndrome. *Psychol.Med., 38,* 1505-1514.

St, C. D., Blackwood, D., Muir, W., Carothers, A., Walker, M., Spowart, G. et al. (1990). Association within a family of a balanced autosomal translocation with major mental illness. *Lancet, 336,* 13-16.

Stenberg, J., Dahl, F., Landegren, U., & Nilsson, M. (2005). PieceMaker: selection of DNA fragments for selector-guided multiplex amplification. *Nucleic Acids Res., 33,* e72.

Stitziel, N. O., Kiezun, A., & Sunyaev, S. (2011). Computational and statistical approaches to analyzing variants identified by exome sequencing. *Genome Biol., 12,* 227.

Sullivan, P. F. (2007). Spurious genetic associations. *Biol.Psychiatry, 61,* 1121-1126.

Sulonen, A. M., Ellonen, P., Almusa, H., Lepisto, M., Eldfors, S., Hannula, S. et al. (2011). Comparison of solution-based exome capture methods for next generation sequencing. *Genome Biol., 12,* R94.

Summerer, D. (2009). Enabling technologies of genomic-scale sequence enrichment for targeted high-throughput sequencing. *Genomics, 94,* 363-368.

Sun, T., Rodriguez, M., & Kim, L. (2009). Glycogen synthase kinase 3 in the world of cell migration. *Dev.Growth Differ., 51,* 735-742.

Sundaram, S. K., Huq, A. M., Sun, Z., Yu, W., Bennett, L., Wilson, B. J. et al. (2011). Exome sequencing of a pedigree with Tourette syndrome or chronic tic disorder. *Ann.Neurol, 69,* 901-904.

Swendsen, J., Hammen, C., Heller, T., & Gitlin, M. (1995). Correlates of stress reactivity in patients with bipolar disorder. *Am.J Psychiatry, 152,* 795-797.

Takata, A., Kato, M., Nakamura, M., Yoshikawa, T., Kanba, S., Sano, A. et al. (2011a). Exome sequencing identifies a novel missense variant in RRM2B associated with autosomal recessive progressive external ophthalmoplegia. *Genome Biol., 12,* R92.

Takata, A., Kim, S. H., Ozaki, N., Iwata, N., Kunugi, H., Inada, T. et al. (2011b). Association of ANK3 with bipolar disorder confirmed in East Asia. *Am.J Med.Genet B Neuropsychiatr.Genet, 156B,* 312-315.

ten, B., Jr. & Grody, W. W. (2008). Keeping up with the next generation: massively parallel sequencing in clinical diagnostics. *J Mol.Diagn., 10,* 484-492.

Tewhey, R., Warner, J. B., Nakano, M., Libby, B., Medkova, M., David, P. H. et al. (2009). Microdroplet-based PCR enrichment for large-scale targeted sequencing. *Nat Biotechnol., 27,* 1025-1031.

Theis, J. L., Sharpe, K. M., Matsumoto, M. E., Chai, H. S., Nair, A. A., Theis, J. D. et al. (2011). Homozygosity mapping and exome sequencing reveal GATAD1 mutation in autosomal recessive dilated cardiomyopathy. *Circ.Cardiovasc.Genet, 4,* 585-594.

Thum, T., Gross, C., Fiedler, J., Fischer, T., Kissler, S., Bussen, M. et al. (2008). MicroRNA-21 contributes to myocardial disease by stimulating MAP kinase signalling in fibroblasts. *Nature, 456,* 980-984.

Torkamani, A., Topol, E. J., & Schork, N. J. (2008). Pathway analysis of seven common diseases assessed by genome-wide association. *Genomics, 92,* 265-272.

Townsend, J. D., Torrisi, S. J., Lieberman, M. D., Sugar, C. A., Bookheimer, S. Y., & Altshuler, L. L. (2012). Frontal-Amygdala Connectivity Alterations During Emotion Downregulation in Bipolar I Disorder. *Biol.Psychiatry*.

Tsang, A. H., Sanchez-Moreno, C., Bode, B., Rossner, M. J., Garaulet, M., & Oster, H. (2012). Tissue-specific interaction of Per1/2 and Dec2 in the regulation of fibroblast circadian rhythms. *J Biol.Rhythms, 27,* 478-489.

Tsankova, N. M., Berton, O., Renthal, W., Kumar, A., Neve, R. L., & Nestler, E. J. (2006). Sustained hippocampal chromatin regulation in a mouse model of depression and antidepressant action. *Nat Neurosci., 9,* 519-525.

Tucker, T., Marra, M., & Friedman, J. M. (2009). Massively parallel sequencing: the next big thing in genetic medicine. *Am.J Hum.Genet, 85,* 142-154.

Turner, E. H., Lee, C., Ng, S. B., Nickerson, D. A., & Shendure, J. (2009). Massively parallel exon capture and library-free resequencing across 16 genomes. *Nat Methods, 6,* 315-316.

Ueda, H. R., Hayashi, S., Chen, W., Sano, M., Machida, M., Shigeyoshi, Y. et al. (2005). System-level identification of transcriptional circuits underlying mammalian circadian clocks. *Nat Genet, 37,* 187-192.

Valouev, A., Ichikawa, J., Tonthat, T., Stuart, J., Ranade, S., Peckham, H. et al. (2008). A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning. *Genome Res., 18,* 1051-1063.

Van der Veen & Archer, S. N. (2010). Light-dependent behavioral phenotypes in PER3-deficient mice. *J Biol.Rhythms, 25,* 3-8.

Vandewalle, G., Archer, S. N., Wuillaume, C., Balteau, E., Degueldre, C., Luxen, A. et al. (2009). Functional magnetic resonance imaging-assessed brain responses during an executive task depend on interaction of sleep homeostasis, circadian phase, and PER3 genotype. *J Neurosci., 29,* 7948-7956.

Vassos, E., Steinberg, S., Cichon, S., Breen, G., Sigurdsson, E., Andreassen, O. A. et al. (2012). Replication Study and Meta-Analysis in European Samples Supports Association of the 3p21.1 Locus with Bipolar Disorder. *Biol.Psychiatry*.

Vazquez, G. H., Kahn, C., Schiavo, C. E., Goldchluk, A., Herbst, L., Piccione, M. et al. (2008). Bipolar disorders and affective temperaments: a national family study testing the "endophenotype" and "subaffective" theses using the TEMPS-A Buenos Aires. *J Affect.Disord, 108,* 25-32.

Vederine, F. E., Wessa, M., Leboyer, M., & Houenou, J. (2011). A meta-analysis of whole-brain diffusion tensor imaging studies in bipolar disorder. *Prog.Neuropsychopharmacol.Biol.Psychiatry, 35,* 1820-1826.

Via, M., Gignoux, C., & Burchard, E. G. (2010). The 1000 Genomes Project: new opportunities for research and social challenges. *Genome Med., 2,* 3.

Viola, A. U., Archer, S. N., James, L. M., Groeger, J. A., Lo, J. C., Skene, D. J. et al. (2007). PER3 polymorphism predicts sleep structure and waking performance. *Curr.Biol., 17,* 613-618.

Vissers, L. E., Fano, V., Martinelli, D., Campos-Xavier, B., Barbuti, D., Cho, T. J. et al. (2011). Whole-exome sequencing detects somatic mutations of IDH1 in metaphyseal chondromatosis with D-2-hydroxyglutaric aciduria (MC-HGA). *Am.J Med.Genet A, 155A,* 2609-2616.

Voelkerding, K. V., Dames, S. A., & Durtschi, J. D. (2009). Next-generation sequencing: from basic research to diagnostics. *Clin.Chem., 55,* 641-658.

Vogel, G. W., Traub, A. C., Ben-Horin, P., & Meyers, G. M. (1968). REM deprivation. II. The effects on depressed patients. *Arch Gen.Psychiatry, 18,* 301-311.

von, Z. D., Dirlich, G., Doerr, P., Emrich, H. M., Lund, R., & Ploog, D. (1985). Are biological rhythms disturbed in depression? *Acta Psychiatr.Belg., 85,* 624-635.

Walsh, T., McClellan, J. M., McCarthy, S. E., Addington, A. M., Pierce, S. B., Cooper, G. M. et al. (2008). Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. *Science, 320,* 539-543.

Wang, F., Kalmar, J. H., He, Y., Jackowski, M., Chepenik, L. G., Edmiston, E. E. et al. (2009). Functional and structural connectivity between the perigenual anterior cingulate and amygdala in bipolar disorder. *Biol.Psychiatry, 66,* 516-521.

Wang, F., McIntosh, A. M., He, Y., Gelernter, J., & Blumberg, H. P. (2011). The association of genetic variation in CACNA1C with structure and function of a frontotemporal system. *Bipolar.Disord, 13,* 696-700.

Wang, H., Chen, X., Dudinsky, L., Patenia, C., Chen, Y., Li, Y. et al. (2011a). Exome capture sequencing identifies a novel mutation in BBS4. *Mol.Vis., 17,* 3529-3540.

Wang, J. L., Yang, X., Xia, K., Hu, Z. M., Weng, L., Jin, X. et al. (2010). TGM6 identified as a novel causative gene of spinocerebellar ataxias using exome sequencing. *Brain, 133,* 3510-3518.

Wang, K., Li, M., & Bucan, M. (2007). Pathway-based approaches for analysis of genomewide association studies. *Am.J Hum.Genet, 81,* 1278-1283.

Wang, X., Wang, H., Cao, M., Li, Z., Chen, X., Patenia, C. et al. (2011b). Whole-exome sequencing identifies ALMS1, IQCB1, CNGA3, and MYO7A mutations in patients with Leber congenital amaurosis. *Hum.Mutat., 32,* 1450-1459.

Watkins, D., Schwartzentruber, J. A., Ganesh, J., Orange, J. S., Kaplan, B. S., Nunez, L. D. et al. (2011). Novel inborn error of folate metabolism: identification by exome capture and sequencing of mutations in the MTHFD1 gene in a single proband. *J Med.Genet, 48,* 590-592.

Weaver, I. C., Cervoni, N., Champagne, F. A., D'Alessio, A. C., Sharma, S., Seckl, J. R. et al. (2004). Epigenetic programming by maternal behavior. *Nat Neurosci., 7,* 847-854.

Weedon, M. N., Hastings, R., Caswell, R., Xie, W., Paszkiewicz, K., Antoniadi, T. et al. (2011). Exome sequencing identifies a DYNC1H1 mutation in a large pedigree with dominant axonal Charcot-Marie-Tooth disease. *Am.J Hum.Genet, 89,* 308-312.

Wehr, T. A., Sack, D. A., & Rosenthal, N. E. (1987). Sleep reduction as a final common pathway in the genesis of mania. *Am.J Psychiatry, 144,* 201-204.

Wellcome Trust Case Control Consortium (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature, 447,* 661-678.

Wilson, G. M., Flibotte, S., Chopra, V., Melnyk, B. L., Honer, W. G., & Holt, R. A. (2006). DNA copy-number analysis in bipolar disorder and schizophrenia reveals aberrations in genes involved in glutamate signaling. *Hum.Mol.Genet, 15,* 743-749.

Wirth, M., Burch, J., Violanti, J., Burchfiel, C., Fekedulegn, D., Andrew, M. et al. (2013). Association of the Period3 clock gene length polymorphism with salivary cortisol secretion among police officers. *Neuro.Endocrinol.Lett., 34,* 27-37.

Wu, J., Jiao, Y., Dal, M. M., Maitra, A., de Wilde, R. F., Wood, L. D. et al. (2011). Whole-exome sequencing of neoplastic cysts of the pancreas reveals recurrent mutations in components of ubiquitin-dependent pathways. *Proc.Natl.Acad.Sci U.S.A, 108,* 21188-21193.

Wyatt, R. J. & Henter, I. (1995). An economic evaluation of manic-depressive illness--1991. *Soc.Psychiatry Psychiatr.Epidemiol., 30,* 213-219.

Xu, B., Roos, J. L., Dexheimer, P., Boone, B., Plummer, B., Levy, S. et al. (2011). Exome sequencing supports a de novo mutational paradigm for schizophrenia. *Nat Genet, 43,* 864-868.

Yamamoto, T., Nakahata, Y., Soma, H., Akashi, M., Mamine, T., & Takumi, T. (2004). Transcriptional oscillation of canonical clock genes in mouse peripheral tissues. *BMC.Mol.Biol., 5,* 18.

Yosifova, A., Mushiroda, T., Kubo, M., Takahashi, A., Kamatani, Y., Kamatani, N. et al. (2011). Genome-wide association study on bipolar disorder in the Bulgarian population. *Genes Brain Behav., 10,* 789-797.

Zhang, D., Cheng, L., Qian, Y., liey-Rodriguez, N., Kelsoe, J. R., Greenwood, T. et al. (2009). Singleton deletions throughout the genome increase risk of bipolar disorder. *Mol.Psychiatry, 14,* 376-380.

Zollner, S. & Pritchard, J. K. (2007). Overcoming the winner's curse: estimating penetrance parameters from case-control data. *Am.J Hum.Genet, 80,* 605-615.

# ANEXOS

# ANEXO I. PROTOCOLO ANÁLISIS EXOMA

1. **Descargar los datos de Internet.**
2. **Control de errores de la descarga.**

```
md5sum –c mds.txt
```

#El número obtenido tiene que coincidir con el que consta en el archivo md5.txt de la descarga.

3. **Control de calidad de los datos.**

#Analizar con el programa FastQC.

```
./fastqc
```

#Valorar la cobertura con el programa bedtools.

```
PATH=~/Public/bedtools/bin:$PATH

./bedtools/bin/coverageBed -abam bedtools/14M.bam -b
bedtools/exomeplus15.bed -d > bedtools/out.txt
```

#Para estudios de exoma completo, cortar las columnas 3-7 y quitar los 3 primeros caracteres (chr) del fichero out1.txt.

```
cut -f 1-2,8 bedtools/out.txt > bedtools/out1.txt
cat bedtools/out1.txt | sed 's/^...//' > bedtools/out2.txt
```

#Selección de muestra de datos para gráficos tipo Manhattan.

```
shuf bedtools/out2.txt > bedtools/out3.txt
head -n 40000 bedtools/out3.txt > bedtools/out4.txt
```

4. **Alinear con genoma de referencia.**

#Abrir el Linux (.sh). Señalar dónde se encuentran los distintos programas.

```
PATH=~/Public/gatk:$PATH
PATH=~/Public/picard:$PATH
PATH=~/Public/bwa:$PATH
PATH=~/Public/annovar:$PATH
```

#Bajar el genoma de referencia e indexarlo. Solo la primera vez que se utiliza.

```
bwa index -a bwtsw -p hg19M/hg19M hg19M/hg19M.fa
```

#Alinear el archivo .fq con el genoma de referencia. Da como resultado un archivo .sai.

```
bwa aln -t 4 -f 13/131M.sai hg19M/hg19M 13/131.fq.gz
```

#Quitar -I si no es Illumina 1.3. 13/131M.sai es la ruta y archivo de salida; 13/131.fq.gz el de entrada. Hacemos lo mismo con el segundo archivo (mate pair ends).

```
bwa aln -t 4 -f 13/132M.sai hg19M/hg19M 13/132.fq.gz
```

#Transformar los archivos .sai en un único archivo .sam

```
bwa sampe -f 13/13M.sam -
r"@RG\tID:13M\tLB:13M\tSM:13M\tPL:ILLUMINA" hg19M/hg19M
13/131M.sai 13/132M.sai 13/131.fq.gz 13/132.fq.gz
```

#Transformar el archivo .sam en .bam. #java –Xmx16G –jar. es una orden que amplía la memoria RAM que puede utilizar el java, haciendo que el programa pueda ir más rápido. Se utiliza LENIENT para que reporte los errores pero el programa no se pare cuando los encuentre.

```
java -Xmx16G -jar picard/SortSam.jar \SO=coordinate
\INPUT=13/13M.sam \OUTPUT=13/13M.bam
\VALIDATION_STRINGENCY=LENIENT \CREATE_INDEX=true
```

#Marcar duplicados. Los archivos .marked.bam y metrics resultantes contienen todas las lecturas e identifican de qué tipo son.

```
java -Xmx16G -jar picard/MarkDuplicates.jar \INPUT=13/13M.bam
\OUTPUT=13/13M.marked.bam \METRICS_FILE=13/metrics
\CREATE_INDEX=true \VALIDATION_STRINGENCY=LENIENT
```

5. **Localización de variantes.**

#Se identifican regiones que requieren un re-alineamiento, habitualmente por presencia de indeles que no existen en el genoma de referencia. El archivo resultante lo denominamos .bam.list.

```
java -Xmx16G -jar gatk/GenomeAnalysisTK.jar \-T
RealignerTargetCreator \-R hg19M/hg19M.fa \-o 13/13M.bam.list \-
I 13/13M.marked.bam
```

#El programa re-alinea las lecturas alrededor de estos indeles, lo que minimiza el riesgo de falsos positivos al buscar posteriormente variantes.

```
java -Xmx16G -jar gatk/GenomeAnalysisTK.jar \ -I
13/13M.marked.bam \ -R hg19M/hg19M.fa \-T IndelRealigner \-
targetIntervals 13/13M.bam.list \ -o 13/13M.marked.realigned.bam
```

#Comprobación de que la información está sincronizada entre cada lectura y su pareada (*mate pair ends*).

```
java -Xmx16G -jar picard/FixMateInformation.jar
\INPUT=13/13M.marked.realigned.bam
\OUTPUT=13/13M.marked.realigned.fixed.bam \SO=coordinate
\VALIDATION_STRINGENCY=LENIENT \CREATE_INDEX=true
```

#Corrección de la calidad de las lecturas. Generación de la tabla de covariantes. Recoge como comparador la base de datos de SNPs (dbSNP), la versión 135, que es la compatible con el hg19.

```
java -Xmx16G -jar gatk/GenomeAnalysisTK.jar -I
13/13M.marked.realigned.fixed.bam -R hg19M/hg19M.fa -T
CountCovariates -cov ReadGroupCovariate -cov
QualityScoreCovariate -cov CycleCovariate -cov DinucCovariate -
recalFile 13/13M.recal_data.csv -knownSites:dbsnp,VCF
dbsnp135.hg19.vcf
```

#Corrección de la calidad de las lecturas. Recalibración de las puntuaciones.

```
java -Xmx16G -jar gatk/GenomeAnalysisTK.jar \-l INFO \-R
hg19M/hg19M.fa \-I 13/13M.marked.realigned.fixed.bam \-T
TableRecalibration \--out
13/13M.marked.realigned.fixed.recal.bam \-recalFile
13/13M.recal_data.csv
```

#Localización de las variaciones presentes en la muestra (SNPs e indels). Se establece a partir de qué Q se informan y se validan las variantes encontradas, se delimita un número máximo de lecturas por área a tener en cuenta y se solicita como parámetros de salida información de balance de alelos (alelo referencia/alelo ref+alternativo) y de profundidad total de cobertura.

```
java -Xmx16G -jar gatk/GenomeAnalysisTK.jar -glm BOTH -R
hg19M/hg19M.fa -T UnifiedGenotyper -I
13/13M.marked.realigned.fixed.recal.bam -D dbsnp135.hg19.vcf -o
13/13.snps.vcf -metrics 13/snps.metrics -stand_call_conf 50.0 -
stand_emit_conf 10.0 -dcov 1000 \-A DepthOfCoverage -A
AlleleBalance -L exomeplus15.bed
```

#Recalibración de la puntuación de la calidad de las variantes.

```
java -Xmx16G -jar gatk/GenomeAnalysisTK.jar \-T
VariantRecalibrator \-R hg19M/hg19M.fa \-input 13/13.snps.vcf \-
resource:hapmap,known=false,training=true,truth=true,prior=15.0
hapmap.vcf \-
resource:omni,known=false,training=true,truth=false,prior=12.0
1000G_omni2.5.hg19.sites.vcf \-
resource:dbsnp,known=true,training=false,truth=false,prior=8.0
dbsnp135.hg19.vcf \-an QD -an HaplotypeScore -an MQRankSum -an
ReadPosRankSum -an FS -an MQ \-recalFile 13/13M.recal \-
tranchesFile 13/13M.tranches \-rscriptFile 13/13M.plots.R \--
maxGaussians 4 --percentBadVariants 0.05
```

#Se aplica la recalibración.

```
java -Xmx16G -jar gatk/GenomeAnalysisTK.jar \-T
ApplyRecalibration \-R hg19M/hg19M.fa \-input 13/13.snps.vcf \--
ts_filter_level 99.0 \-tranchesFile 13/13M.tranches \-recalFile
13/13M.recal \-o 13/13M.snp.vcf.recalibrated
```

#Se filtran las variables encontradas según los parámetros seleccionados.

```
java -Xmx16G -jar gatk/GenomeAnalysisTK.jar \-R hg19M/hg19M.fa
\-T VariantFiltration \--variant 13/13M.snp.vcf.recalibrated \-o
13/13M.snp.recalibrated.filtered.vcf \--clusterWindowSize 10 \--
filterExpression "MQ0 >= 4 && ((MQ0 / (1.0 * DP)) > 0.1)" \--
filterName "HARD_TO_VALIDATE" \--filterExpression "DP < 5 " \--
filterName "LowCoverage" \--filterExpression "QUAL < 30.0 " \--
filterName "VeryLowQual" \--filterExpression "QUAL > 30.0 &&
QUAL < 50.0 " \--filterName "LowQual" \--filterExpression "QD <
1.5 " \--filterName "LowQD" \--filterExpression "SB > -10.0 " \-
-filterName "StrandBias"
```

#Abrir ANNOVAR y convertir el formato.

```
perl annovar/convert2annovar.pl --format vcf4 --includeinfo
13/13M.snp.recalibrated.filtered.vcf > 13/13M.snp.annovar

cd annovar
mkdir 13
cd ..
cp 13/13M.snp.annovar annovar/13/
cd annovar
cd ..
```

#Añade las puntuaciones de frecuencias en bases de datos.

```
./summarize_annovar.pl --buildver hg19 13/13M.snp.annovar  --
verdbsnp 137 --ver1000g 1000g2012apr --veresp 6500 ./humandb -
outfile 13/13M.snps   -remove -alltranscript
```

#Para pasar de un formato delimitado por comas a un formato tabulado. Transforma en txt.

```
perl -pe 'while (s/(,"[^"]+),/\1<COMMA>/g) Difu; s/"//g;
s/,/\t/g; s/<COMMA>/,/g' < 13/13M.snps.exome_summary.csv >
13/13.exome_summary.txt
```

6. **Análisis con KGGSeq.**

#Unificar los archivos .vcf de los 3 sujetos de estudio.

```
java -Xmx8g -jar GenomeAnalysisTK.jar \-R
~/Public/hg19M/hg19M.fa \-T CombineVariants \--variant
vcf/13.vcf \--variant vcf/14.vcf \--variant vcf/15.vcf -o
bipolar.vcf \-genotypeMergeOptions UNIQUIFY
```

#Localización del programa y genoma de referencia.

```
java -Xms256m -Xmx1300m -jar kggseq.jar --buildver hg19
```

#Archivos de entrada.

```
--no-resource-check --vcf-file bipolar.vcf --ped-file
bipolar.ped
```

#Bases de datos para el filtrado.

```
--db-gene refgene --regions-out chrX,chrY --db-filter
hg19_dbsnp131,hg19_1kg201202,hg19_ESP5400,hg19_dbsnp135
```

#Variables para el filtrado.

```
--rare-allele-freq 0.001 --genotype-filter 2,3,5 --ibs-check
--gene-feature-in 0,1,2,3,4,7 --db-score dbnsfp
```

#Órdenes para la realización de otros análisis. Se le añade una lista de genes candidatos.

```
--mendel-causing-predict--pubmed-mining searchTerm --candi-list
ABCB1,ACE,ADCY9,ADORA2A,ADRA1A,ADRA1B,ADRA2A,ADRA2B,ADRA2C,ADRB1
,AGTR1,AKT1,APOE,AR,AVPR1A,AVPR1B,BDNF,CACNA1C,CCK,CCKAR,CCKBR,C
CL2,CCND2,CD3E,CD47,CHRM2,CHRNA7,CLOCK,CNR1,CNTF,COMT,CREB1,CRHB
P,CRHR1,CRHR2,CYP2C9,DAOA,DISC1,DRD1,DRD2,DRD3,DRD4,DRD5,DTNBP1,
ESR1,ESR2,FGFR1,FGFR2,FGFR3,FGFR4,GABBR1,GABRA3,GABRA5,GABRA6,GA
D1,GNAL,GNAS,GNB3,GPR50,GRIA1,GRIA2,GRIA3,GRIA4,GRIK1,GRIK2,GRIK
3,GRIK4,GRIK5,GRIN1,GRIN2A,GRIN2B,GRIN2C,GRIN2D,GRIN3A,GRM7,GSK3
A,GSK3B,HTR1A,HTR1B,HTR2A,HTR2B,HTR3A,HTR3B,HTR4,HTR5A,HTR6,HTR7
,IL1B,IL6,ITPR1,KCNC2,LEP,LEPR,LRP1,MAOA,MAOB,MMP2,NFKB1,NGFR,NO
S1,NOS3,NPY,NR3C1,NTRK2,NTRK3,OLIG1,OLIG2,OLIG3,OPRD1,OPRK1,OPRM
1,P2RX4,P2RX7,PAM,PDE10A,PDE11A,PDE1A,PDE2A,PDE5A,PDE6C,PDE9A,PE
NK,PER1,PER2,PER3,PLA2G2A,PLA2G4A,POMC,PRKCH,SLC6A1,SLC6A2,SLC6A
3,SLC6A4,STAT3,SYN3,TAC1,TACR1,TH,TNF,TPH1,TPH2,ADRA1B,ADRA2C,AG
TR1,AKT1,CACNA1C,CHRM2,CRHR1,CRHR2,DISC1,DRD1,DRD5,ESR2,GABRA3,G
ABRA6,GNAS,GRIA1,GRIA3,GRIK1,GRIK5,GRIN1,GRIN2B,GRIN2C,GSK3A,HTR
2A,HTR2C,HTR3A,HTR3B,HTR5A,LEPR,M6PR,NOS1,NR3C1,NTRK2,OPRK1,OPRM
1,P2RX4,POMC,SYN3,TAC1,TACR1,TH
--pathway-annot cura
```

# ANEXO II.

## SCRIPT PARA LA GENERACIÓN DE LOS GRÁFICOS DE COBERTURA (TIPO MANHATTAN).

```
source("http://dl.dropbox.com/u/66281/0_Permanent/qqman.r")
gwahits <- read.table("C:/rfiles/out4p.txt", header=T)
P = as.numeric(1/10^gwahits$P)
CHR = gwahits$CHR
BP = as.integer(gwahits$BP)
gwa2 = unique(data.frame(CHR,BP,P))
manhattan(gwa2,colors=rainbow(6))
```

# ANEXO III. ABREVIATURAS MÁS FRECUENTES

| | |
|---|---|
| A | Adenina |
| ADN | Ácido desoxirribonucleico |
| ARN | Ácido ribonucleico |
| ANKRD31 | *Ankyrin repeat domain-containing protein* 31 |
| Bp o pb | *Base pairs* o pares de bases |
| BWA | *Burrows-Wheeler Aligner* |
| C | Citosina |
| CCD | *Coupled device* o dispositivo de carga acoplada |
| CEIC | Comité de Ética en la Investigación clínica |
| CI | Consentimiento informado |
| CLK | Clock |
| CIE10 | Clasificación internacional de enfermedades, 10ª edición |
| CRY | Cryptochrome |
| CNVs | *Copy number variants* |
| dNTP | Deoxiribonucleósido trifosfato |
| DSM-IV-TR | *Diagnostic and Statistical Manual of Mental Disorders, 4th Edition, Text Revision* |
| emPCR | PCR en emulsión |
| ENCODE | *Encyclopedia of DNA Elements* |
| G | Guanina |
| GABA | Ácido gamma-aminobutírico |
| GATK | *Genome Analysis Tool Kit* |
| Gly | Glicina |
| GSK3beta | *Glycogen synthase kinase 3 beta* |
| GWAS | *Genomewide association studies* |
| IGV | *Integrative Genomics Viewer* |
| KB | Kilobyte |
| mRNA | RNA mensajero |
| NSQ | Núcleo supraquiasmático |
| OD | Densidad óptica |
| OR | *Odds ratio* |
| PER3 | PERIOD3 |
| PCR | Reacción en cadena de la Polimerasa |
| SNP | *Single Nucleotide Polimorphism* o polimorfismo de nucleótido único |
| T | Timina |
| TBP | Trastorno bipolar |
| TMEM155 | *Transmembrane protein 155* |
| USP29 | *Ubiquitin Specific Peptidase* 29 |
| Val | Valina |

# PUBLICACIONES

Journal of Alzheimer's Disease 19 (2010) 873–884
DOI 10.3233/JAD-2010-1292
IOS Press



# Clinical-Genetic Correlations in Familial Alzheimer's Disease Caused by Presenilin 1 Mutations

Estrella Gómez-Tortosa[a,*], Sagrario Barquero[b], Manuel Barón[c], Eulogio Gil-Neciga[d], Fernando Castellanos[e], Martín Zurdo[e], Sagrario Manzano[b], David G. Muñoz[f], Adolfo Jiménez-Huete[g], Alberto Rábano[h], M. José Sainz[a], Rosa Guerrero[a], Isabel Gobernado[i], Julián Pérez-Pérez[j] and Adriano Jiménez-Escrig[i]

[a]*Department of Neurology, Fundación Jiménez Díaz, Madrid, Spain*
[b]*Department of Neurology, Hospital Clínico Universitario San Carlos, Madrid, Spain*
[c]*Department of Neurology, Fundación Hospital Alcorcón, Madrid, Spain*
[d]*Department of Neurology, Hospital Virgen del Rocio, Seville, Spain*
[e]*Department of Neurology, Hospital Virgen del Puerto, Plasencia, Spain*
[f]*Department of Neurology, S. Michael's Hospital, Toronto, Ontario, Canada*
[g]*Department of Neurology, Hospital Ruber Internacional, Madrid, Spain*
[h]*Department of Pathology, Fundación Hospital Alcorcón, Madrid, Spain*
[i]*Department of Neurology, Hospital Ramón y Cajal, Madrid, Spain*
[j]*Secugen SL, Madrid, Spain*

Accepted 15 September 2009

**Abstract**. We describe the clinical phenotype of nine kindred with presenile Alzheimer's disease (AD) caused by different presenilin 1 (PS1) point mutations, and compare them with reported families with mutations in the same codons. Mutations were in exon 4 (Phe105Val), exon 5 (Pro117Arg, Glu120Gly), exon 6 (His163Arg), exon 7 (Leu226Phe), exon 8 (Val261Leu, Val272Ala, Leu282Arg), and exon 12 (Ile439Ser). Three of these amino acid changes (Phe105Val, Glu120Gly, and Ile439Ser) had not been previously reported. Distinct clinical features, including age of onset, symptoms and signs associated with the cortical-type dementia and aggressiveness of the disease, characterized the different mutations and were quite homogeneous across family members. Age of onset fell within a consistent range: some mutations caused the disease in the thirties (P117R, L226F, V272A), other in the forties (E120G, H163R, V261L, L282R), and other in the fifties (F105V, I439S). Associated features also segregated with specific mutations: early epileptic activity (E120G), spastic paraparesis (V261L), subcortical dementia and parkinsonism (V272A), early language impairment, frontal signs, and myoclonus (L226F), and late myoclonus and seizures (H163R, L282R). Neurological deterioration was particularly aggressive in PS1 mutations with earlier age of onset such as P117R, L226F, and E120G. With few exceptions, a similar clinical phenotype was found in families reported to have either the same mutation or different amino acid changes in the same codons. This series points to a strong influence of the specific genetic defect in the development of the clinical phenotype.

Keywords: Early onset, familial Alzheimer's disease, mutations, presenilin 1

## INTRODUCTION

Mutations in the presenilin 1 (PS1) gene are the most common cause of autosomal dominant early-onset Alzheimer's disease (AD). Since the first PS1 mutation was reported in 1995 [1], 176 different muta-

*Correspondence to: Estrella Gómez-Tortosa, MD, PhD, Servicio de Neurología, Fundación Jiménez Díaz, 28040 Madrid, Spain. Tel.: +34 91 550 4913; Fax: +34 91 550 4882; E-mail: egomezt@fjd.es.

ISSN 1387-2877/10/$27.50 

tions have been listed in the AD Mutation Database (http://www.molgen.ua.ac.be/Admutations/). It is estimated that 11% of referral-based series with early onset AD (< 65 years of age) can be explained by PS1 mutations [2]. In different populations as much as 36% to 66% of early-onset autosomal dominant cases are attributable to PS1 mutations [3–5], with the percentage decreasing to 20% when there is an early onset with family history of AD but no clear autosomal dominant pattern [6].

The study of AD caused by these mutations is of extraordinary importance for several reasons. First, the study of PS1 protein dysfunction has provided important insights into the physiopathology of AD. PS1 is a transmembrane aspartyl protease which functions as a subunit of the $\gamma$-secretase complex involved in the processing of the amyloid-$\beta$ protein precursor (A$\beta$PP) to the amyloid-$\beta$ (A$\beta$) peptide. The underlying mechanism of most PS1 mutations to produce AD seems to be related to increased production of the A$\beta_{42}$ fragment of A$\beta$PP, which has a greater tendency to aggregate. Transfected cell lines expressing PS1 mutations cause increased production of A$\beta_{42}$ [7] and pathologic examination of brains with PS1 mutations show increased A$\beta_{42}$ deposition when compared to sporadic AD [8].

Secondly, AD caused by PS1 mutations serves as a good model for the careful analysis of clinical-pathological and genetic correlations. Consecutive descriptions have revealed a heterogeneous clinical and pathological phenotype among families. The range of age at onset is wide, and atypical features such as behavioral disturbances, movement disorders, pyramidalism, or seizures are more frequent than in late-onset AD. However, despite the plenitude of genetic screening studies identifying novel PS1 mutations, there are relatively few reports offering the type of extensive clinical information that would allow for genotype-phenotype correlations. For example, some PS1 mutations have been associated with AD with spastic paraparesis or early seizures, but it remains to be clarified what level of genotype alteration determines the association: amino acid change, position of the codon, or the exon involved. In addition, pathogenicity of some variations in the PS1 gene has not been completely cleared out, and this is a growing issue due to the increasing access to sequencing techniques.

Finally, the availability of genetic testing for presymptomatic diagnosis in the fully penetrant mutations provides a group of individuals in which to define the preclinical stages of AD with cognitive and neuroimaging techniques. Furthermore, thorough knowledge of age at onset ranges in particular mutations opens the possibility of administering very early treatments.

We describe the clinical phenotype of nine independent kindred with different PS1 point mutations recruited in the context of a clinical-genetic project of familial AD in Spain, and compare them with reported clinical data on families with mutations in the same codons. The series includes three novel PS1 mutations, because of the amino acid-type substitution, in codons with mutations already reported as linked to early-onset AD.

## MATERIALS AND METHODS

### *Subjects*

The GENODEM project is a prospective research program that collects clinical-genetic information and DNA samples of Spanish kindred with familial degenerative dementias, mostly of the Alzheimer type. The project involves 4 academic hospitals in Madrid (Fundación Jiménez Díaz, Hospital Clínico San Carlos, Hospital Ramón y Cajal, and Fundación Hospital Alcorcón) that use common standardized dementia protocols, and also have collaborations with neurologists leading Dementia Units in the southern half of Spain. The study was approved by the Research Ethics Committee at Fundación Jiménez Díaz.

All cases with standardized clinical criteria for AD, age at onset before 85 years of age, and with at least an affected first-degree sibling diagnosed with degenerative dementia and/or having a similar clinical dementia as reported by relatives, are included in the project and a blood sample for extracting DNA is taken after informed consent. Probands and available affected siblings undergo an extensive clinical and cognitive assessment, blood analysis, MRI and SPECT brain studies. Relatives of the probands aid in the construction of detailed family pedigrees and in the collection of information on affected cases in previous generations. In 25 presenile familial AD cases, a genetic study including sequencing of PS1, PS2, and A$\beta$PP genes was conducted. No mutations in the PS2 or A$\beta$PP gene were found. This study focused on describing the clinical phenotype of nine independent families in which nine different PSEN 1 mutations were found.

### *Molecular genetics*

DNA was extracted from peripheral blood leukocytes using QIAmp Mini-kit DNA blood (Qiagen,

Chatsworth, CA). All PS1 gene exons and intronic adjacent regions were amplified by polymerase chain reaction (PCR) with CertAmp (Biotools, Madrid, Spain). The purified PCR reactions were sequenced from both directions by way of an automated sequencer (AB 3730XL) using the Big Dye 3.1 Terminator Cycle sequencing kit. DNA sequences were analyzed using Seqscape software.

PolyPhen program was used to predict the impact of the novel variants on protein function (http://genetics.bwh.harvard.edu/pph). It predicts the functional effect of amino acid changes by considering evolutionary conservation, the physiochemical differences, and the proximity of the substitution to predicted functional domains and/or structural features. PolyPhen scores of > 2.0 are expected to be "probably damaging" to protein structure and function.

## RESULTS

In nine independent families, nine different point mutations were found in the PS1 gene. Mutations were in exon 4 (Phe105Val), exon 5 (Pro117Arg, Glu120Gly), exon 6 (His163Arg), exon 7 (Leu226Phe), exon 8 (Val261Leu, Val272Ala, Leu282Arg), and exon 12 (Ile439Ser) as represented in Fig. 1. Three of them (Phe105Val, E120Gly, and Ile439Ser) are novel mutations since these amino acid changes had not been previously reported in those positions. All families had a clear autosomal dominant pattern of inheritance of dementia, except for one case (Phe105Val) in which no family history was available. We describe the clinical phenotypes associated with each mutation as follows and review the clinical data available on families with the same mutations and with alternative mutations in the same residues. Table 1 summarizes significant clinical features of our cohort and of reported families with mutations in the same codons.

### *Phe105Val (F105V)*

This mutation was found in a 55-year-old woman who started with memory deficits at 52 years of age. The dementia was seemingly sporadic since she only had a brother who was cognitively normal at 55 years of age and both parents had died relatively young. Her cognitive decline was a typical aphaso-apraxo-amnestic syndrome with no motor or atypical features.

Pathogenicity of this novel mutation, in the absence of other affected cases in which to investigate cosegregation with the disease, is supported by a PolyPhen score of 2.37 predicting F105V to be "probably damaging", and by two previous changes described in this codon – to leucine [9] and to isoleucine [5] – associated with AD at similar age of onset. F105L was found in a female with onset of cognitive symptoms at 52 years of age. At age 60, she developed Parkinson-like symptoms and died fully demented and bedridden at age 63. Brain autopsy confirmed AD excluding concomitant Parkinson's disease. Clinical data of the family with the F105I mutations is restricted to age of onset range of three affected members, 53 to 58 years.

### *Pro117Arg (P117R)*

The proband is a 35-year-old woman originally from Colombia whose memory deficits had begun about eighteen months before. She had given up her job as housekeeper at age 34 because of increasing difficulties in organizing and recalling her duties. Her examination showed complete temporal disorientation, executive and verbal memory deficits, and visuoconstructive deficits. Her Mini Mental State Examination (MMSE) score was 19/30. There were no motor signs and her gait was normal. During the next two years, neurological deterioration was rapid and included the development of myoclonus, seizures, and gait impairment. She died completely dependent and bedridden at age 37. Her family history was remarkable for dementia in her father (beginning at age 37, died at 42), an older brother (starting at age 38, died at 41), and two older sisters (beginning at ages 36 and 37 years who are now 39 and 40 years respectively).

Besides this change to arginine [10], mutations in this codon position have been described to result in changes to alanine, P117A [11,12], leucine, P117L [13,14], or serine, P117S [15] in families with different European and South American origins. In all codon 117 mutations, age at onset is very early, in the twenties or early thirties and with a very rapid evolution to severe cognitive deterioration. Reports extending clinical descriptions point to the presence of early behavioral and neuropsychiatric symptoms, myoclonus, seizures, and extrapyramidal signs, in the first three years of evolution [10,14]. Interestingly, the same P117A mutation is described in two families with different clinical phenotypes, one with typical AD with no motor features [12] but the other with a prominent ataxic-spastic syndrome associated with the cognitive decline [11]. Most cases reported with P117 mutations died before 40 years of age.

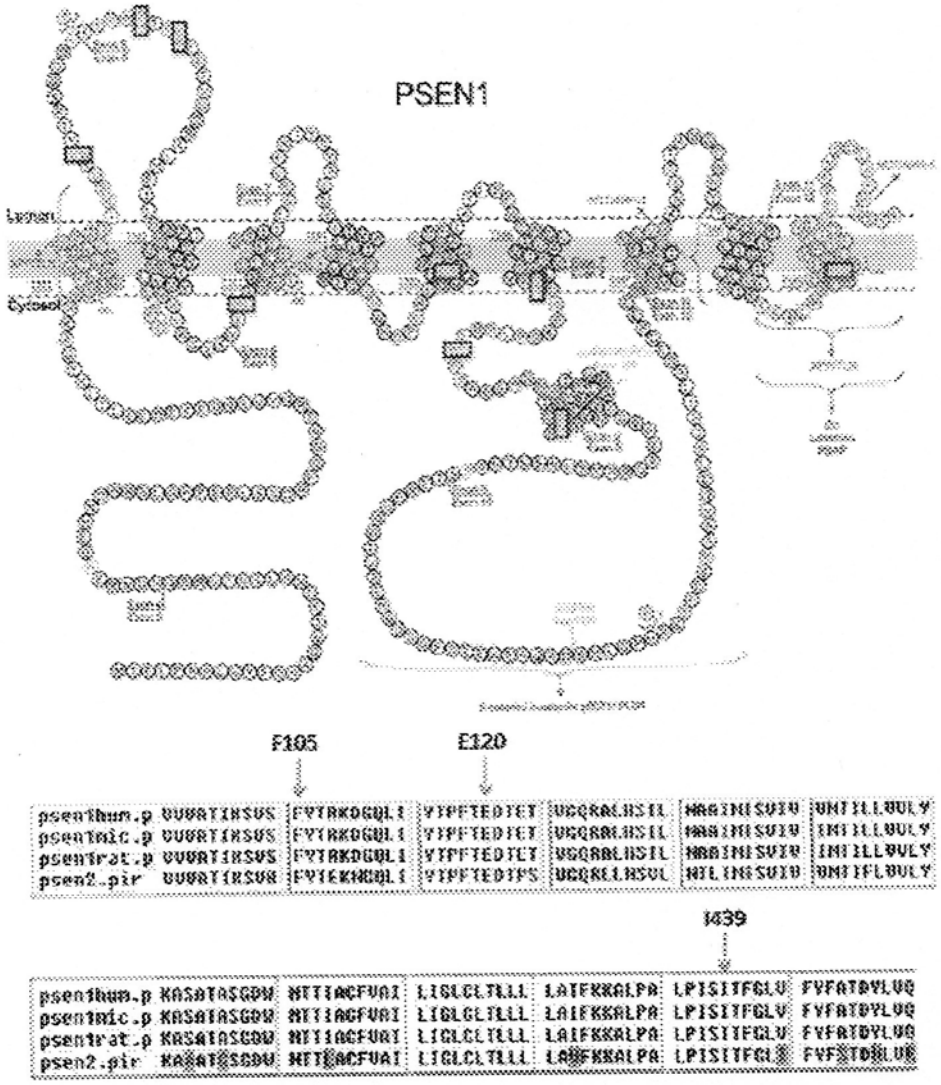



Fig. 1. Scheme of the PSEN 1 protein (taken from http://www.alzforum.org, reprinted from International Review of Cytology, Vol 254, Dillen K. & Annaert W, Copyright (2006), with permission from Elsevier) showing the aminoacid substitutions caused by the different mutations in red circles. The yellow squares highlight the location of the nine mutations described. Below, aminoacid residues in positions 105, 120 and 439 (where the three new mutations are reported) are conserved in most animal species PS1 and in PS2.

*Glu120Gly (E120G)*

The most important features in this family include a very homogeneous age at onset (around 39–40 years), epileptic activity, and very rapid cognitive and functional deterioration. Family tree includes the affected father, his homoczygotic twin, and two affected offspring, brother and sister.

The father started with behavioral disturbances and memory deficits at around 39 years of age, was committed to a psychiatric institution where he died at age 42. He had generalized seizures. His twin brother had a similar dementing illness in his late thirties and died at 42 years of age but clinical details were not available. The former's son had onset of recent memory deficits at 39 years of age, in addition to visual hallucinations and behavioral disturbances. Cerebral MRI showed diffuse cortical atrophy. He was diagnosed with hebephrenic schizophrenia and for more than a year was under psychiatric observation and treated with neuroleptic drugs and benzodiazepines. Cognitive deterioration was evident during this time and he was soon unable to resume his work as a sales representative. At 41 years of age, an EEG showed epileptic abnormalities as gener-

Table 1

Summary of significant clinical features of our cohort and of reported families with mutations in the same codons

| | GENODEM Families | | Review of families with mutations in the same codons | | | |
|---|---|---|---|---|---|---|
| | Age onset (yr) # | Distinct features | Mutations in same codon | Age at onset (yr) | Distinct features | References |
| **Exon 4** | | | | | | |
| *F105V* | 52 | Cortical dementia | F105L | 52 | Late parkinsonian features | [9] |
| | | | F105I | 53–58 | No clinical information | [5] |
| **Exon 5** | | | | | | |
| *P117R* | 33 | Cortical dementia. Rapid deterioration | P117R | 36 | Early myoclonus, seizures, parkinsonism | [10] |
| | (33–38) | with myoclonus, seizures and gait | P117A | 24–34 | Cortical dementia (a)/Ataxic-spastic variant (b) | [12] (a), [11] (b) |
| | | impairment | P117S | < 30 | Aggressive AD with rapid progression | [15] |
| | | AgeD:37–42 yrs | P117L | 23–32 | Early myoclonus, seizures, parkinsonism | [13], [14] |
| *E120G* | 39–40 | Early epilepsy, rapid deterioration | E120D | 41–53 | Early seizures | [3], [16], [17] |
| | | AgeD: 42–44 yrs | E120K | 32–39 | No clinical information | [18] |
| **Exon 6** | | | | | | |
| *H163R* | 46 | Cortical dementia. Late seizures and | H163R | 42–50 | Cortical dementia. Late seizures/myoclonus | [2–4], [6], [8], [10], |
| | (42–50) | myoclonus. | | | | [17], [19–21], [23] |
| | | AgeD: 47–65 yrs | H163Y | 44–65 | Cortical dementia. Late seizures/myoclonus | [22] |
| **Exon 7** | | | | | | |
| *L226F* | 33 | Early dysarthria, non fluent aphasia, | L226F | 33 | Early behavioral, frontal signs, mild parkinsonism | [23] |
| | (33–36) | frontal signs, myoclonus, | | | (FTD-like). Progress to mutism. | |
| | | mild parkinsonism. AgeD: 42 yrs | L226R | < 49 | No clinical description; one case diagnosed of Pick's disease | [24] |
| **Exon 8** | | | | | | |
| *V261L* | 40 | Spastic paraparesis, dystonia | V261L | 40 | (same family) | [25] |
| | | | V261F | 38 | Spastic paraparesis | [2], [26] |
| *V272A* | 26–36 | Subcortical cognitive deficits. | | | | |
| | | Parkinsonism [27]. AgeD: 36–46 yrs | V272A | 34 | Limited clinical information: myoclonus | [28] |
| *L282R* | 35–50 | Middle-late myoclonus, seizures, | L282R | 40–49 | No clinical information | [29] |
| | | parkinsonian gait. AgeD: 49–58 yrs | L282V | 41–52 | Middle-late myoclonus, seizures, parkinsonism / $2^{nd}$ proband with mild dysmetria | [30] |
| **Exon 12** | | | | | | |
| *I439S* | 55 (47–55) | Cortical dementia. AgeD: 58–70 yrs | I439V | > 55 | Asymptomatic at 55 yr of age | [2] |

#age of the proband/s (range for the family). AgeD: ages at death (proband and/or ancestors).
FTD: frontotemporal dementia.

alized spike-wave, although he has yet to show clinical seizures. He is now 42 year-old, he does not speak, and only on occasion obeys very simple commands. He has bilateral limb rigidity without tremor or myoclonus and slow gait in anteflexion posture, all these features likely related to a chronic effect of neuroleptics. Ocular movements are normal and there are no pyramidal signs. A recent cerebral MRI shows bilateral hippocampal atrophy (Scheltens grade 4), SPECT with FP-CIT shows normal dopamine transporter receptors density in striatum, and ApoE genotype is $\varepsilon 3/\varepsilon 3$.

His sister began having recurrent generalized seizures at age 39, with an EEG showing right frontal paroxysmal activity with bilateral diffusion. Lamotrigine was not effective and she was finally free of seizures with valproate. Memory deficits started also at this age and progressed to substantial cognitive deterioration in the two subsequent years. At age 42, she was already completely dependent for activities of daily living, spoke only single words, could not walk, and had double incontinence. She had a dystonic posture in both hands, predominantly in the right. Ocular movements were normal and there were no pyramidal or other extrapyramidal signs. Cerebral MRI showed cortico-subcortical atrophy. Brain SPECT with HMPAO showed bilateral temporoparietal hypoperfusion and in basal ganglia, predominantly on the left side. She is now 43 year-old, bedridden with decorticate posture in limbs, and only responds to pain. ApoE genotype is also $\varepsilon 3/\varepsilon 3$.

This is a novel mutation whose pathogenicity is supported by cosegregation with the disease in this family and a PolyPhen score of 2.34 predicting "probably damaging". Amino acid changes in this codon position had been reported only to aspartate, E120D [3,5,16,17] or lysine, E120K [18]. Age at onset range is similar for all the families with E120D mutations, 41 to 53 years. Clinical data are very limited but it is remarkable that the family reported by Reznik-Wolf et al. [16] was also characterized by seizures in the first two years of the disease. The family reported as having E120K mutation had an earlier age of onset, in their thirties, but there are no further clinical information.

### *His163Arg (H163R)*

Through a first proband (V-1, female) whose genetic analysis showed the PS1 H163R mutation, we recruited an extensive tree (Fig. 2) of related families clustered in a small town with another four affected cases (two females and two males). Review of death certificates allowed us to collect data on age at death of affected cases in the two previous generations and some clinical information was also available.

*Ancestors data*: Ages at onset in 10 ancestors show a short range (42 to 50, mean $45.7 \pm 3.6$ years) with mean time of evolution until death of 8 years (ages at death 47 to 65, mean $54.2 \pm 6.3$ years). The father of the proband (case IV-4) started with cognitive decline at around 46 years of age and died with severe dementia at 54 years. His clinical report, dated in the 1980s, stated the diagnosis of possible Pick's disease based on behavioral disturbances.

*Probands*: Clinical information of the five affected cases is summarized in Table 2. Briefly, the three females and one male (case V-5) had a very similar phenotype, starting at 46 years of age with progressive memory deficits and depression and evolving into a typical aphaso-apraxo-amnestic cortical syndrome. Two of them are now in more advanced stages and show hypertonia and multifocal myoclonus. The man is now 52 years old and has suffered three generalized seizures in the last year. However, the second man (case V-3) has developed, from age 42, an amnesic-aphasic syndrome associated with prominent behavioral disturbances (aggressive behavior, agitation, and visual hallucinations). ApoE genotype is $\varepsilon 3/\varepsilon 3$ in two cases and $\varepsilon 3/\varepsilon 4$ in the other two.

H163R is a frequent mutation, with multiple reports in different populations [2–4,6,8,10,17,19–21]. Remarkably, the onset age of all the families reported with this mutation is tightly clustered in the mid-forties (mean 46, range 42 to 50 years). Most of the cases developed typical AD, some with common neuropsychiatric symptoms such as anxiety, apathy, depression, or irritability. Myoclonus and/or seizures are frequent though in advanced stages of the disease. The other pathogenic mutation described in this codon position is a change to tyrosine, H163Y [22]. This family had a very homogeneous phenotype, basically the same as the one described in H163R families, except for a wider range of age at onset, 44 to 65 years.

### *Leu226Phe (L226F)*

These kindred include an affected father, who died at age 42 after 6 years of progressive cognitive deterioration, and the proband, a woman who at age 33 developed depression, dysarthria with non fluent speech, intentional tremor in right hand, and memory deficits. At that moment she was still able to perform household chores and daily activities quite independently. CT scan showed mild cortical atrophy,

Table 2
Clinical features of the five cases with H163R PS1 mutation

| Case Number | Age at Onset (yrs)/Gender | Sequential clinical features (yrs of evolution) | Neuroimaging | ApoE |
|---|---|---|---|---|
| V-1 | 46/Female | Depression. Anomia, recent memory deficits, executive deficits. (3 yrs) | CT scan: normal<br>SPECT: very slight left temporal hypoperfusion | 3/3 |
| V-2 | 46/Female | Depression. Temporal desorientation, anomia, recent memory deficits, visuoconstructional and executive deficits. Positive frontal reflexes. Mutism, hypertonia and multifocal myoclonus. No parkinsonism or seizures. (8 yrs) | MRI: slight parietal atrophy<br>SPECT: sligth left parietal hypoperfusion | 3/3 |
| V-3 | 42/Male | Aggressive behavior, apathy, recent memory deficits. Delirium, visual hallucinations, motor and verbal stereotypes, logopenic, positive frontal reflexes. (6 yrs) | CT scan: normal<br>SPECT: bilateral parieto-occipital hypoperfusion | 3/4 |
| V-4 | 46/Female | Depression. Anomia, recent memory deficits, visuoconstructional and executive deficits. (5 yrs) | MRI: mild small vessel disease<br>SPECT: left temporal hypoperfusion | 3/4 |
| V-5 | 46/Male | Depression, irritability, anomia and memory deficits. Mutism, hypertonia and gait difficulties, myoclonus, three seizures. (5 yrs) | CT scan: bitemporal atrophy | – |

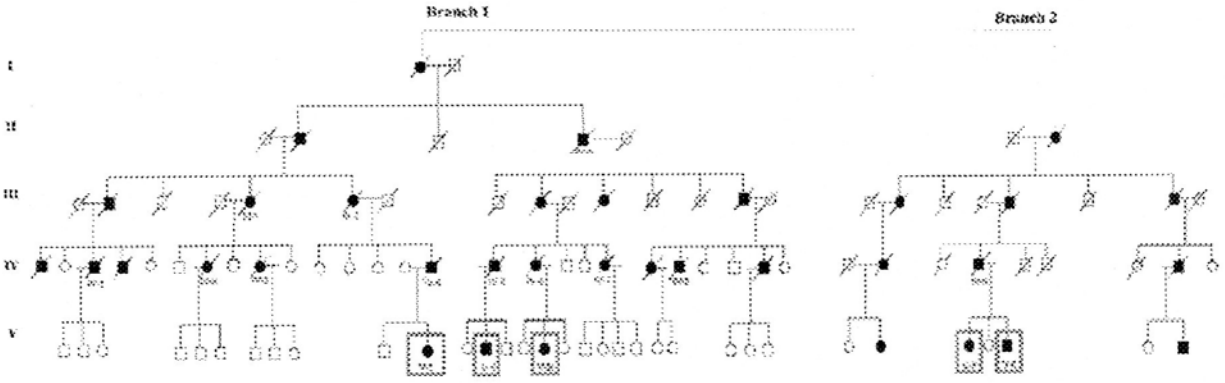



Fig. 2. Family pedigree with the H163R PS1 mutation including five generations. Squares represent the cases clinically studied and followed-up.

thyroid antibodies were positive, and she was treated for thyrotoxic state. When examined again at age 36, she had clearly deteriorated with MMSE score of 18/30. Language was dysarthric and aphasic, nonfluent with some paraphasias, and there were significant deficits in verbal memory, executive and visuoconstructive tests. She had rest and kinetic bilateral tremor, mild limb bradykinesia and decreased arm swing, reflex myoclonus, grasping (right predominant all) and release of other frontal reflexes (glabellar, sucking, jaw). Brain MRI showed diffuse cortical atrophy, predominantly biparietal. Brain SPECT with FP-CIT was normal, and an EEG showed diffuse slowness.

L226F has been described [23] in a patient clinically diagnosed of frontotemporal dementia at age 33, according to behavioral symptoms, frontal atrophy in CT scan and severe hypoperfusion in frontal areas by SPECT. However, a severe deterioration in short-term memory subsequently developed, along with bilateral primitive reflexes and slight parkinsonian symptoms. He died mute at the age of 38 years and postmortem diagnosis was AD. His mother had begun with cognitive deterioration at age 33 and died at the age of 44. Another mutation in this position, L226R [24], has been described in a family whose affected members died in their early fifties. The proband's clinical phenotype and age of onset is obscured by the association of childhood delay in motor skills and two decades of various substances abuse. It is of note that the proband's affected cousin was clinically diagnosed with Pick's disease, though clinical description is not provided.

### *Val261Leu (V261L)*

This family is characterized by spastic paraparesis associated with the dementia syndrome. Little information is available from the proband's mother (she died before 60 years of age with gait disorder and dementia) and brother (48 years old with cognitive deterio-

ration and gait difficulties). The proband is a female who started at age 40 years with progressive attentional and memory deficits. She was examined by our group when her family consulted for a second opinion, but was first studied and published by Jimenez Caballero et al. [25]. Frontal cortical reflexes were negative but she had very brisk limb reflexes, bilateral Babinski sign and dystonic posture in right arm. Parkinsonian signs were non-significant except for mild hypomimic face and gait with decreased arm-swing. Two years later, tone was increased in legs and gait was slightly spastic. Cognitive assessment showed a non-fluent language, with circumlocutions, and deficits in visual and verbal memory tests, executive and visuoconstructive functions. Brain MRI showed cortico-subcortical atrophy predominant in frontal lobes. Brain SPECT showed bilateral temporal hypoperfusion.

There are no other families reported with V261L mutation. In this codon, another mutation, V261F [26], also presented with spastic paraparesis in the four affected cases, beginning at age 38 in the proband case, with death occurring in the forties in the three relatives.

*Val272Ala (V272A)*

This family, previously described by our group [27], is characterized by a very early age of onset (range 26 to 36 years), subcortical-type cognitive deterioration, and parkinsonism. The kindred include four affected members: two sisters were studied clinically, and retrospective information was obtained about the mother (hospital records) and grandmother (verbal information from relatives revealed that she died in her forties of puerperal infection after several years of cognitive deterioration).

When she was 30, the mother had developed apathy and inhibition that was first interpreted as postpartum depression, but three years later were associated with parkinsonian gait, dysarthria, and cognitive deterioration. Motor symptoms were improved with levodopa. She died at 44 years of age following progressive cognitive and motor decline.

At 26 years of age, one daughter became forgetful, slow of thought, and apathetic and was first treated for depression. Three years later her neuropsychological assessment was in the lower limits of normality, except for clearly impaired score in learning tasks. Brain MRI and EEG were normal. One year later (age 30), a PET with $^{18}$FDG showed hypometabolism of bilateral frontal cortex and ventral cinguli. In the following four years, a marked cognitive disorder progressed bringing about severe deficits in attentional and executive tasks and recent memory. She died at age 36 and her brain study revealed neurofibrillary tangles as a Braak stage VI and neuritic plaques fulfilling CERAD criteria for AD plus Lewy bodies and Lewy neurites in substantia nigra, nucleus basalis of Meynert, parahippocampus, and amygdala. Her sister began to experience cognitive deterioration at age 31, and at 33 years her MMSE score was 19/30. Both sisters had an ApoE genotype $\varepsilon 3/\varepsilon 3$.

There is another Spanish family recently described with this same mutation [28] with very limited clinical information. The proband case began with dementia and myoclonus at age 34 and died at age 42.

*Leu282Arg (L282R)*

The proband is a woman who at age 35 began complaining of memory and concentration deficits in addition to dyscalculia. A neuropsychological exam one year later revealed a MMSE of 24/30, naming deficits, constructional and limb dyspraxia, recent verbal memory deficits and mild simultanagnosia. Cerebral MRI showed bilateral parietotemporal atrophy and brain SPECT revealed biparietal hypoperfusion. Four years later she had non-fluent aphasia, bilateral myoclonus, and parkinsonian gait. She is now 42 years old and last year began with generalized seizures.

There was a family history of dementia in her father (starting at age 42, death at 49), two aunts (starting at age 48 and 50 with death at 54 and 58, respectively) as well as her paternal grandmother, who died at 58 years of age. One brother of the proband is a presymptomatic carrier at age 37. A brain $^{18}$FDG-PET study in this case shows already slight hypometabolism in parietotemporal areas.

The same mutation was previously described in another Spanish family [29] with age at onset of affected cases 43 $\pm$ 7 years (mean $\pm$ SD) and age at death 56 $\pm$ 3 years. No further clinical description is available. Brain autopsy of the proband confirmed AD. Another mutation in this codon, L282V, has been described by Dermaut et al. [30]. One of the probands had a similar phenotype to our case, developing myoclonus, extrapyramidal signs, and seizures in middle-advanced stages, although age at onset was later (45 years). The second proband (cousin) developed a cortical-type dementia at age 41 with mild dysmetria appearing in middle stages of the disease. Ages at death in this family range from 49 to 57 years.

*Ile439Ser (I439S)*

The proband case is a man who started with recent memory deficits at around 55 years of age. After two years the clinical picture remains only cognitive without other neurological signs. Neuropsychological assessment shows deficits in executive functions, praxis, and verbal memory. Neuropsychiatric Inventory reveals high scores for apathy and depression and middle-low range values for anxiety and irritability. Cerebral MRI shows diffuse cortical atrophy. The same mutation has been found in an affected second-cousin with onset of dementia in his fifties. There is information about four affected women in previous generations who died between 58 and 70 years of age after about a decade of cognitive deterioration.

This is a novel mutation whose pathogenicity is supported by cosegregation with the disease in this family. PolyPhen analysis predicted I439S to be "possibly damaging" to PS1 function. Mutations in this codon position, the most distal carboxiterminal PS1 mutation identified to date, had been reported only as resulting in a change to valine [2]. Information on clinical correlates of this mutation is limited, since it was found in two sisters with a double PS1 mutation in which the mother, carrying the I439V mutation, was asymptomatic at 55 years of age.

## DISCUSSION

The review of this series of families with different PS1 mutations emphasizes the clinical variability of presenile AD and points to a strong influence of the specific genetic defect in the development of the clinical phenotype.

Distinct clinical features, including age at onset, symptoms and signs associated with the cortical-type dementia and aggressiveness of the disease, characterized the different mutations and were quite homogeneous across family members. Age of onset, under 56 years in all families, fell within a consistent range: some mutations causing the disease in the thirties (P117R, L226F, V272A), other in the forties (E120G, H163R, V261L, L282R), and other in the seizures (H163R, L282R). Evolution of the disease was also particularly aggressive in PS1 mutations with earlier age at onset such as P117R, L226F, and E120G. Families with several affected members who were personally examined by our group showed no remarkable clinical differences among them.

Despite its paucity of some clinical descriptions, the review of the literature for families with the same mutations also revealed a quite concordant clinical phenotype with similarities found in age of onset and "core" clinical features. For example, almost all European and Asian families with H163R mutation had an age of onset in their mid-forties, with myoclonus and seizures in advanced stages of the disease. Spastic paraparesis was a constant in families with V261L mutations, as was a very aggressive AD with early myoclonus and seizures in P117R mutations, or prominent frontal syndrome in L226F. In their extensive review, Larner and Doran [31] have emphasized the clinical variability of AD caused by PS1 mutations, pointing to both interfamilial and intrafamilial heterogeneity. Within the spectrum of almost two hundred different PS1 mutations, there are examples of individuals with the same mutations and different phenotypes. However, comparison of phenotypes may be variable depending on whether clinical information is more or less complete, how descriptions cover different stages of the disease, and also which clinical symptoms/signs are considered as core features. Mild behavioral symptoms, extrapyramidal signs, and myoclonus are frequent in presenile AD and present in many PS1 mutations at some stages of the disease. In comparison, features such as associated spastic paraparesis, early epilepsy, fronto-temporal syndrome, cerebellar signs, significant parkinsonism or very aggressive course, appear associated only with specific mutations, with clinical differences among affected family members being only mild. Our detailed review of this group of PS1 mutations suggests that there is usually a consistent clinical profile for each mutation.

The strong influence of the specific mutation on age of onset is supported by clinical observations [32], studies in transgenic models of Drosophila as well as cell cultures expressing the mutations. Seidner and colleagues [33] have shown that the capacity of a panel of

to restore normal function also correlated with their ability to support Notch cleavage and signaling, which was found as a good correlate of $\gamma$- secretase activity. This $\gamma$-secretase activity seems to finally determine the amount of increased A$\beta_{42}$. In cell cultures expressing different PS1 mutations, the degree of increased levels of A$\beta_{42}$ strongly correlates with age of onset [34]. Cell lines expressing three different P117 mutations, which produce a particularly aggressive early onset AD, show increased A$\beta_{42}$, decreased neurite outgrowth, and altered endoproteolytic cleavage of PS1 [15]. Moreover, the severity of this clinical phenotype is closely correlated with the abundance of plaques and tangles in affected brains. *In vitro* studies of S170F, another mutation with a very aggressive phenotype and age of onset in the late twenties, also show a 3-fold increase of A$\beta_{42}$ [35]. Applying all these data to AD caused by PS1 mutations supports the notion that, rather than other genetic or environmental factors, age of onset is primarily determined by the mutations themselves, by their effects on $\gamma$- secretase function and amount of amyloid deposition.

The correlation between mutation site across the protein domain and phenotype does not allow definition of a clear pattern, but our study suggests that the codon affected has an important influence in the clinical phenotype, while the amino acid-type substitution is less relevant. As represented in Fig. 1, a majority of the pathogenic PS1 mutations are clustered in transmembrane domains (TMD), and probably impair PS1 function by disrupting the $\alpha$-helical face [36,37]. Mutations in particular codons are especially aggressive, but ages at onset show a wide range independently of whether the mutation is located in TMD, hydrophilic or hydrophobic loops. In this respect, it was interesting to compare phenotypes in families with different point mutations in the same codons. Most variant amino acid changes in a given residue did not make much difference in the clinical picture, as evidenced by H163R or T, E120G or D, P117R, L or S. The importance of the region is suggested by mutations in the second hydrophilic loop (as seen in residues 113, 115, 117, 120) which are characterized by early epileptic activity and fast cognitive deterioration [38], while spastic paraparesis is frequently associated with mutations in exons 8 and 9. However, very close mutations such as those found in residues 269 and 272 may cause very different phenotypes, the first being a late-onset AD [39] and the second, a subcortical type of dementia with parkinsonism starting in the early thirties [27]. On the contrary, PS1 mutations characterized by a fronto-temporal phenotype (such as L113P, G183V, L226F, insR352) are spread throughout the PS1 protein.

Altogether, interpreting these data is a difficult task. Some authors have suggested that PS1 mutations may have a more dramatic impact when amino acid changes are non-conservative substitutions than when they are semi-conservative, or when a negatively-charged amino acid is substituted by a positively-charged one [3,17]. Miklossy and colleagues [40] have pointed to the fact that several mutations associated with very early onset involved a proline residue suggesting that these mutations may drastically alter the conformation of the PS1 protein by the substitution of hydrophobic and hydrophilic amino acids. Tridimensional models of PS1 suggest that the pathological effect of mutations may be related to their position pointing outward on interfacial PS1 surface and changing its shape or electrostatic properties [23]. Using fluorescence lifetime imaging microscopy, Berezovska et al. [41] have shown that several PS1 mutations increase proximity of PS1 N and C epitopes and change the configuration of the PS1-A$\beta$PP complex. In any event, it seems that genetic differences have a clear influence in the clinical phenotype.

Finally, pathogenicity of the three new mutations is supported by both PolyPhen prediction scores and the algorithm of Guerreiro et al. [28]. Two of the mutations are definitive pathogenic (E120G and I439S, with 2 cases affected, other mutations reported in the same residue, residues conserved in PS1 and PS2, plus I439 fits the helix or TMD rule) and one is probable pathogenic (F105V with 1 case affected, not present in controls, other mutations reported in the same residue and residues conserved in PS1 and PS2).

In summary, the analysis of clinical-genetic correlations in this series of families along with the review of families with mutations in the same residues suggests that clinical phenotype is strongly influenced by the specific genetic defect in question.

## DISCLOSURE STATEMENT

Authors' disclosures available online (http://www.j-alz.com/disclosures/view.php?id=157).

## REFERENCES

[1] Sherrington R, Rogaev EI, Liang Y, Rogaeva EA, Levesque G, Ikeda M, Chi H, Lin C, Li G, Holman K, et al. (1995) Cloning of a gene bearing missense mutations in early-onset familial Alzheimer's disease. *Nature* **375**, 754-760.

[2] Rogaeva EA, Fafel KC, Song YQ, Medeiros H, Sato C, Liang Y, Richard E, Rogaev EI, Frommelt P, Sadovnick AD, Meschino W, Rockwood K, Boss MA, Mayeux R, St George-Hyslop P (2001) Screening for PS1 mutations in a referral-based series of AD cases: 21 novel mutations. *Neurology* **57**, 621-625.

[3] Campion D, Dumanchin C, Hannequin D, Dubois B, Belliard S, Puel M, Thomas-Anterion C, Michon A, Martin C, Charbonnier F, Raux G, Camuzat A, Penet C, Mesnage V, Martinez M, Clerget-Darpoux F, Brice A, Frebourg T (1999) Early-onset autosomal dominant Alzheimer disease: prevalence, genetic heterogeneity, and mutation spectrum. *Am J Hum Genet* **65**, 664-670.

[4] Lleó A, Blesa R, Queralt R, Ezquerra M, Molinuevo JL, Peña-Casanova J, Rojo A, Oliva R (2002) Frequency of mutations in the presenilin and amyloid precursor protein genes in early-onset Alzheimer disease in Spain. *Arch Neurol* **59**, 1759-1763.

[5] Raux G, Guyant-Marechal L, Martin C, Bou J, Penet C, Brice A, Hannequin D, Frebourg T, Campion D (2005) Molecular diagnosis of autosomal dominant early onset Alzheimer's disease: an update. *J Med Genet* **42**, 793-795.

[6] Kamimura K, Tanahashi H, Yamanaka H, Takahashi K, Asada T, Tabira T (1998) Familial Alzheimer's disease genes in Japan. *J Neurol Sci* **160**, 76-81.

[7] Mehta ND, Refolo LM, Eckman C, Sanders S, Yager D, Perez-Tur J, Younkin S, Duff K, Hardy J, Hutton M (1998) Increased Aβ42(43) from cell lines expressing presenilin 1 mutations. *Ann Neurol* **43**, 256-258.

[8] Ishii K, Ii K, Hasegawa T, Shoji S, Doi A, Mori H (1997) Increased Aβ42(43)-plaque deposition in early-onset familial Alzheimer's disease brains with the deletion of exon 9 and the missense point mutation (H163R) in the PS-1 gene. *Neurosci Lett* **228**, 17-20.

[9] Finckh U, Muller-Thomsen T, Mann U, Eggers C, Marksteiner J, Meins W, Binetti G, Alberici A, Hock C, Nitsch RM, Gal A (2000) High prevalence of pathogenic mutations in patients with early-onset dementia detected by sequence analyses of four different genes. *Am J Hum Genet* **66**, 110-117.

[10] Zekanowski C, Styczynska M, Peplonska B, Gabryelewicz T, Religa D, Ilkowski J, Kijanowska-Haladyna B, Kotapka-Minc S, Mikkelsen S, Pfeffer A, Barczak A, Luczywek E, Wasiak B, Chodakowska-Zebrowska M, Gustaw K, Laczkowski J, Sobow T, Kuznicki J, Barcikowska M (2003) Mutations in presenilin 1, presenilin 2 and amyloid precursor protein genes in patients with early-onset Alzheimer's disease in Poland. *Exp Neurol* **184**, 991-996.

[11] Anheim M, Hannequin D, Boulay C, Martin C, Campion D, Tranchant C (2007) Ataxic variant of Alzheimer's disease caused by Pro117Ala PS1 mutation. *J Neurol Neurosurg Psychiatry* **78**, 1414-1415.

[12] Kauwe JS, Wang J, Chakraverty S, Goate AM, Henao-Martinez AF (2008) Novel presenilin 1 variant (P117A) causing Alzheimer's disease in the fourth decade of life. *Neurosci Lett* **438**, 257-259.

[13] Wisniewski T, Dowjat WK, Buxbaum JD, Khorkova O, Efthimiopoulos S, Kulczycki J, Lojkowska W, Wegiel J, Wisniewski HM, Frangione B (1998) A novel Polish presenilin-1 mutation (P117L) is associated with familial Alzheimer's disease and leads to death as early as the age of 28 years. *NeuroReport* **9**, 217-221.

[14] Alberici A, Bonato C, Borroni B, Cotelli M, Mattioli F, Binetti G, Gennarelli M, Luca MD, Simonati A, Perani D, Rossini P, Padovani A (2007) Dementia, delusions and seizures: storage disease or genetic AD. *Eur J Neurol* **14**, 1057-1059.

[15] Dowjat WK, Kuchna I, Wisniewski T, Wegiel J (2004) A novel highly pathogenic Alzheimer presenilin-1 mutation in codon 117 (Pro117Ser): Comparison of clinical, neuropathological and cell culture phenotypes of Pro117Leu and Pro117Ser mutations. *J Alzheimers Dis* **6**, 31-43.

[16] Reznik-Wolf H, Treves TA, Davidson M, Aharon-Peretz J, St George Hyslop PH, Chapman J, Korczyn AD, Goldman B, Friedman E (1996) A novel mutation of presenilin 1 in familial Alzheimer's disease in Israel detected by denaturing gradient gel electrophoresis. *Hum Genet* **98**, 700-702.

[17] Poorkaj P, Sharma V, Anderson L, Nemens E, Alonso ME, Orr H, White J, Heston L, Bird TD, Schellenberg GD (1998) Missense mutations in the chromosome 14 familial Alzheimer's disease presenilin 1 gene. *Hum Mutation* **11**, 216-221.

[18] Hutton M, Busfield F, Wragg M, Crook R, Perez-Tur J, Clark RF, Prihar G, Talbot C, Phillips H, Wright K, Baker M, Lendon C, Duff K, Martinez A et al. (1996) Complete analysis of the presenilin 1 gene in early onset Alzheimer's disease. *Neuro report* **7**, 801-805.

[19] Tanahashi H, Kawakatsu S, Kaneko M, Yamanaka H, Takahashi K, Tabira T (1996) Sequence analysis of presenilin 1 gene mutation in Japanese Alzheimer's disease patients. *Neurosci Lett* **218**, 139-141.

[20] Kamino K, Sato S, Sakaki Y, Yoshiiwa A, Nishiwaki Y, Takeda M, Tanabe H, Nishimura T, Ii K, St George-Hyslop PH, Miki T, Ogihara T (1996) Three different mutations of presenilin 1 gene in early-onset Alzheimer's disease families. *Neurosci Lett* **208**, 195-198.

[21] Poduslo SE, Herring K, Neal M (1996) A presenilin 1 mutation in an early onset Alzheimer's family: no association with presenilin 2. *Neuro report* **7**, 2018-2020.

[22] Axelman K, Basun H, Lannfelt L (1998) Wide range of disease onset in a family with Alzheimer disease and a His163Tyr mutation in the presenilin-1 gene. *Arch Neurol* **55**, 698-702.

[23] Zekanowski C, Golan MP, Krzysko KA, Lipczynska-Lojkowska W, Filipek S, Kowalska A, Rossa G, Peplonska B, Styczynska M, Maruszak A, Religa D, Wender M, Kulczycki J, Barcikowska M, Kuznicki J (2006) Two novel presenilin 1 gene mutations connected with frontotemporal dementia-like clinical phenotype: genetic and bioinformatic assessment. *Exp Neurol* **200**, 82-88.

[24] Coleman P, Kurlan R, Crook R, Werner J, Hardy J (2004) A new presenilin Alzheimer's disease case confirms the helical alignment of pathogenic mutations in transmembrane domain 5. *Neurosci Lett* **364**, 139-140.

[25] Jiménez Caballero PE, de Diego Boguna C, Martin Correa E, Serviá Candela M, Marsal Alonso C (2008) A novel presenilin 1 mutation (V261L) associated with presenile Alzheimer's disease and spastic paraparesis. *Eur J Neurol* **15**, 991-994.

[26] Farlow MR, Murrell J, Unverzagt FW, Phillips M, Takao M, Ghetti B, Hulette C (2001) Familial Alzheimer's disease with spastic paraparesis associated with a mutation at codon 261 of the presenilin 1 gene. In *Alzheimer's disease: advances in etiology, pathogenesis and therapeutics*, Iqbal K, Sisodia SS, Winblad B, eds. Chichester, John Wiley, pp 53-60.

[27] Jimenez-Escrig A, Rabano A, Guerrero C, Simon J, Barquero MS, Guell I, Ginestal RC, Montero T, Orensanz L (2004) New V272A presenilin 1 mutation with very early onset subcortical dementia and parkinsonism. *Eur J Neurol* **11**, 663-669.

[28] Guerreiro RJ, Baquero M, Blesa R, Boada M, Brás JM, Bullido MJ, Calado A, Crook R, Ferreira C, Frank A, Gómez-Isla T, Hernández I, Lleó A, Machado A, Martínez-Lage P, Masdeu J, Molina-Porcel L, Molinuevo JL, Pastor P, Pérez-Tur J, Relvas R, Oliveira CR, Ribeiro MH, Rogaeva E, Sa A, Samaranch L, Sánchez-Valle R, Santana I, Tárraga L, Valdivieso

F, Singleton A, Hardy J, Clarimón J (2008) Genetic screening of Alzheimer's disease genes in Iberian and African samples yields novel mutations in presenilins and APP. *Neurobiol Aging*, in press.

[29] Aldudo J, Bullido MJ, Arbizu T, Oliva R, Valdivieso F (1998) Identification of a novel mutation (Leu282Arg) of the human presenilin 1 gene in Alzheimer's disease. *Neurosci Lett* **240**, 174-176.

[30] Dermaut B, Kumar-Singh S, De Jonghe C, Cruts M, Löfgren A, Lübke U, Cras P, Dom R, De Deyn PP, Martin JJ, Van Broeckhoven C (2001) Cerebral amyloid angiopathy is a pathogenic lesion in Alzheimer's disease due to a novel presenilin 1 mutation. *Brain* **124**, 2383-2392.

[31] Larner AJ, Doran M (2006) Clinical phenotypic heterogeneity of Alzheimer's disease associated with mutations of the presenilin-1 gene. *J Neurol* **253**, 139-158.

[32] Lippa CF, Swearer JM, Kane KJ, Nochlin D, Bird TD, Ghetti B, Nee LE, St. George-Hyslop P, Pollen DA, Drachman DA (2000) Familial Alzheimer's disease site of mutation influences clinical phenotype. *Ann Neurol* **48**, 376-379.

[33] Seidner GA, Ye Y, Faraday MM, Alvord WG, Fortini ME (2006) Modeling clinically heterogeneous presenilin mutations with transgenic Drosophila. *Curr Biol* **16**, 1026-1033.

[34] Duering M, Grimm MO, Grimm HS, Schröder J, Hartmann T (2005) Mean age of onset in familial Alzheimer's disease is determined by amyloid beta 42. *Neurobiol Aging* **26**, 785-788.

[35] Piccini A, Zanusso G, Borgui R, Noviello C, Monaco S, Russo R, Damonte G, Armirotti A, Gelati M, Giordano R, Zambenedetti P, Russo C, Ghetti B, Tabaton M (2007) Association of a presenilin 1 S170F mutation with a novel Alzheimer disease molecular phenotype. *Arch Neurol* **64**, 738-745.

[36] Alzheimer's Disease Collaborative Group (1995) The structure of the presenilin 1 (S182) gene and identification of six novel mutations in early onset AD. *Nat Genet* **11**, 219-222.

[37] Hardy J, Crook R (2001) Presenilin mutations line up along transmembrane α-helices. *Neurosci Lett* **306**, 203-205.

[38] Finckh U, Kuschel C, Anagnostouli M, Patsouris E, Pantes GV, Gatzonis S, Kapaki E, Davaki P, Lamszus K, Stavrou D, Gal A (2005) Novel mutation and repeated findings of mutations in familial Alzheimer disease. *Neurogenetics* **6**, 85-89.

[39] Larner AJ, Ray PS, Doran M (2007) The R269H mutation in presenilin-1 presenting as late-onset autosomal dominant Alzheimer's disease. *J Neurol Sci* **252**, 173-176.

[40] Miklossy J, Taddei K, Suva D, Verdile G, Fonte J, Fisher C, Gnjec A, Ghika J, Suard F, Mehta PD, McLean CA, Masters CL, Brooks WS, Martins RN (2003) Two novel presenilin-1 mutations (Y256S and Q222H) are associated with early-onset Azheimer's disease. *Neurobiol Aging* **24**, 655-662.

[41] Berezovska O, Lleo A, Herl LD, Frosch MP, Stern EA, Bacskai BJ, Hyman BT (2005) Familial Alzheimer's disease presenilin 1 mutations cause alterations in the conformation of presenilin and interactions with amyloid precursor protein. *J Neurosci* **25**, 3009-3017.

# Extended Kindred With Recessive Late-Onset Alzheimer Disease Maps to Locus 8p22-p21.2

## *A Genome-wide Linkage Analysis*

*Manuel Baron, MD PhD,* Estrella Gomez-Tortosa, MD PhD,† Zoltan Bochdanovits, PhD,‡ Isabel Gobernado, MD,§ Alberto Rabano, MD PhD,‖ David G. Munoz, MD PhD,¶ Peter Heutink, PhD,‡ and Adriano Jimenez-Escrig, MD Ph.D#*

**Abstract:** Late-onset Alzheimer disease (LOAD) is a complex genetic disorder. Although genes involved in early-onset forms were discovered more than a decade ago, LOAD research has only been able to point out small effect loci, with the exception of APOE. We mapped the gene predisposing to LOAD in an extended inbred family coming from a genetically isolated region (24 sampled individuals, 12 of whom are affected), completing a genome-wide screen with an Affymetrix10 K single nucleotide polymorphism microarray. Genotyping results were evaluated under model-dependent (dominant and recessive) and model-free analysis. We obtained a maximum nonparametric linkage score of 3.24 ($P = 0.00006$) on chromosome 8p22-p21.2. The same genomic position also yielded the highest multipoint heterogeneity LOD (HLOD) under a recessive model (HLOD = 3.04). When we compared the results of the model-dependent analysis, a higher score was obtained in the recessive model (3.04) than in the dominant model (1.0). This is a new locus identified in LOAD, in chromosome 8p22-p21.2 and encompassing several candidate genes, among them CLU and PPP3CC that were excluded by sequencing. The finding of a recessive model of inheritance, consistent with the assumption of inbreeding as a morbidity factor in this population, supports the notion of a role of recessive genes in LOAD.

**Key Words:** Alzheimer disease, genetics, recessive transmission, PPP3CC, clusterin

(*Alzheimer Dis Assoc Disord* 2012;26:91–95)

Late-onset Alzheimer disease (LOAD), the most common neurodegenerative disease, is a complex genetic disorder. Genes involved in early-onset forms were identified more than a decade ago and efforts in LOAD research have pointed out several loci, although only the role of the apolipoprotein E (APOE) gene has been clearly established. Of the 664 genes listed in the AlzGene database as of the end of September, 2010[1] only a few have been consistently replicated.[2] At present, a small fraction of Alzheimer cases can be explained by the genes identified thus far: the *APP, PSEN1,* and *PSEN2*. These genes cover the early-onset cases, and no clear evidence of a specific gene mutation for familial LOAD has been discovered. Contrariwise to what happens in other neurodegenerative diseases, awareness of recessive transmission in Alzheimer disease (AD) has rambled from relative obscurity to total oblivion.

The identification of specific genes contributing to LOAD is made challenging by the fact that the condition appears in the late stage of life; in elderly patients, the genetic effect is obscured by environmental factors, comorbidity, or phenocopies. Extensive families with LOAD are uncommon, so family-based linkage studies—the most effective method for identifying causative genes—are difficult to apply to the study of the disease. In 2005, we reported an extensive family with pathology-confirmed LOAD without mutations in the genes currently associated with AD.[3] The age of onset in these kindred ranged from the sixth to the eighth decade. This age of onset is similar to sporadic AD unlike most autosomal dominant familial AD, in which an earlier age at onset is frequent. Moreover, they were originally from an isolated population area. Therefore, their genetic homogeneity provided increased power to identify loci with moderate effect, as the potential number of susceptibility genes that contribute to the LOAD should be reduced.

We report the genome-wide screen that was carried out in these kindred using a high-density single nucleotide polymorphism (SNP) microarray. This study is the first genome-wide screen of its type using the GeneChip Human Mapping 10 K 2.0 assay conducted on a single extensive family with LOAD and points out to a recessive transmission of the trait.

## PATIENTS AND METHODS

### Family Description

A detailed description of the LOAD family is contained in reference.[3] Briefly, the family comprises 3 extensive kindred from a genetically isolated population from Spain. Twelve affected and 16 unaffected members of these kindred were examined clinically and a postmortem brain study was carried out in 3 affected cases, which rendered a pathological diagnosis of AD (Braak stage VI). Dementia has been recorded in 6 generations of ancestors of the cases examined. A review of death certificates allowed all individuals to be linked across 3 extensive pedigrees. Judging by surname and geographic location, we

Receive for publication November 18, 2010; Accepted February 3, 2011.

From the *Fundacion Hospital Alcorcon; †Fundación Jiménez Díaz; #Hospital Ramon y Cajal, S. de Neurologia; §Hospital Ramon y Cajal, S. de Psiquiatría; ‖C.Reina Sofia, Red CIEN, Madrid, Spain; ‡VUMC, Department of Human Genetics; Amsterdam, The Netherlands; and ¶St. Michael Hospital, Pathology Department, Toronto, Canada.

This work received a research grant from the Fundacion Areces.

The authors declare no conflicts of interest.

Reprints: Adriano Jimenez-Escrig, MD, PhD, Hospital Ramon y Cajal, S. de Neurologia, 28034 Madrid, Spain (e-mail: adriano.jimenez@hrc.es).

strongly suspect the 3 kindred have a common founder. The examined affected individuals had progressive memory loss with onset between 57 and 74 years of age, along with seizures, myoclonus, and Parkinsonism in advanced stages. Brains examined in postmortem studies showed widespread neocortical neuritic plaques and neurofibrillary tangles (Braak stage VI), amyloid angiopathy, and Lewy bodies restricted to limbic areas.

Sequencing exons 16 and 17 of the APP gene, and exons 4 to 12 of the PSEN1 and PSEN2 genes did not disclose any mutations. Genotyping with markers located 1 to 3 cM from the aforementioned genes further excluded linkage to these genes. In these kindred, APOE4 is not likely to be causing LOAD because most individuals were APOE 3/3.[3] Complex segregation analysis showed that the best model to fit the data was that of a major dominant gene with a gene frequency close to 3% in this population. Simulation analysis predicted an average logarithm of odds (LOD) score of 2.2 at $\theta = 0.05$.[3] The pedigree of the whole family is shown in Figure 1 with the genotyped family members indicated by dots.

Informed consent was obtained from all participants, and the study was approved by the Ethics Committee of the Hospital Ramon y Cajal, Madrid.

## Genotyping

Blood samples were taken from the 35 living members of the kindred. A genome-wide search was undertaken, genotyping 24 individuals (12 affected cases and 12 nonaffected) with the GeneChip Mapping 10 K 2.0 Xba Array containing 10,204 SNP markers. The average genotype call rate obtained from the 24 samples was 91.7% (range, 82.7-98.4), providing data on 10,067 genotypes per individual. We verified sample sexes by counting heterozygous SNPs on the X chromosome. The sex of all studied samples was confirmed and none of the 6 males typed were assigned a heterozygous state for any of the 194 to 257 (SNPs) mapping to the X chromosome. SNP genotypes were obtained by following the Affymetrix protocol for the GeneChip Human Mapping 10 K 2.0 Array and Assay kit according to manufacturer's instructions. The arrays were hybridized, washed, and scanned in the MRC Gene Service.

## Statistical Analysis

Genotypes were called by GDAS 2.0 software and exported in text file form to the free software tool ALOHOMORA[4] available at http://gmc.mdc-berlin.de/alohomora/ to import the data into the linkage program and also for quality control routines. PedCheck was used for detection of Mendelian errors.[5] SNPs with Mendelian

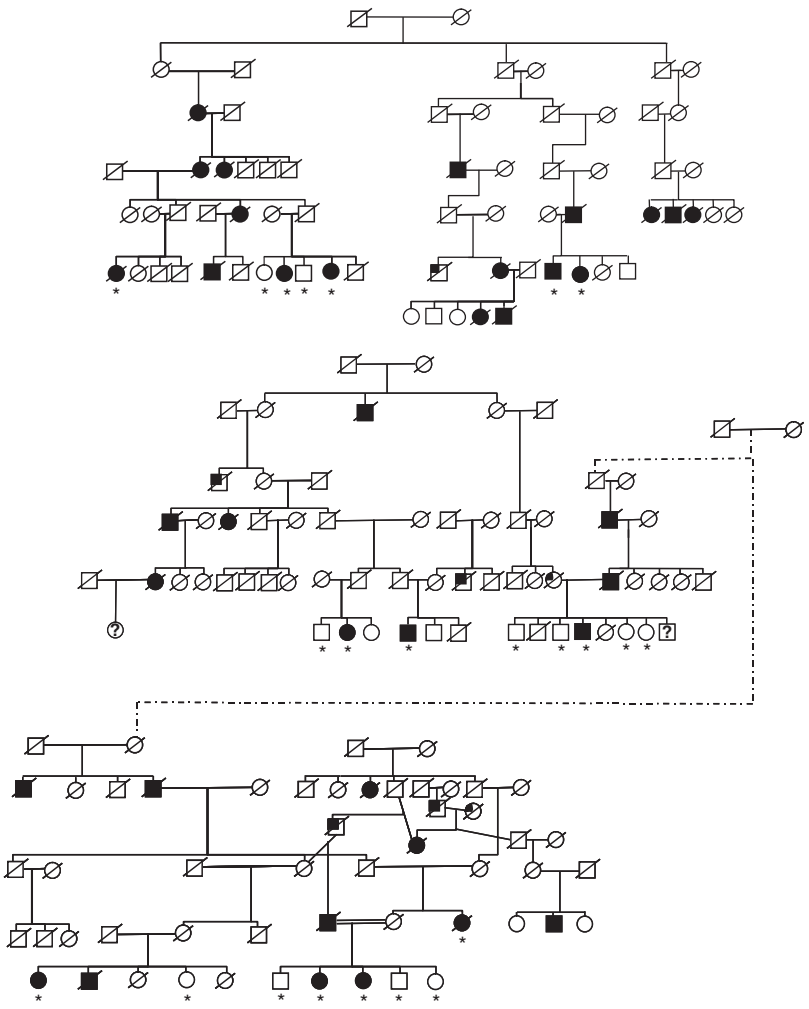


**FIGURE 1.** Pedigree diagram of the kindred under study. Only dotted individuals were included in the genome-wide screen.

errors and SNPs that were not informative for any individual were selectively removed. Non-Mendelian errors were identified by the Merlin option "error"[6] and the unlikely genotypes deleted in the individuals in which they occur.

Parametric autosomal dominant and recessive model assuming a reduced penetrance of the disease and nonparametric LOD score calculations were carried out with Merlin v1.1.2 (Multipoint Engine for Rapid Likelihood Inference),[6] chromosome by chromosome, using all SNPs in a chromosome simultaneously for a multipoint analysis. Family 1 was a 32-byte pedigree, too large for Merlin computation that skips the pedigree, so we had to drop the less informative individuals of the pedigree to get an appropriate size to run it in Merlin. As the underlying genetic model for this family was unknown, both parametric and nonparametric analyses were performed. In the parametric model, we selected a gene frequency of 0.04 and a penetrance of 80% for the gene carriers and 0.1% for the nongene carriers. Nonparametric Kong and Cox LOD (NPL) score was calculated using the Whittemore and Halpern NPL pairs statistic.[7] As a large increase in allele sharing among affected individuals in such a small number of families is expected, we chose the Kong and Cox (1997)[8] exponential model, which provides a better linkage test in this regard.

To generate empirical *P* values, we used Merlin to simulate genotype data with the same structures as the family data sets. A hundred genome-wide simulations were performed, and the resulting simulated data were analyzed with the simulate option; the highest NPL Z and LOD scores were recorded for each chromosome and for each simulation. For these kindred, the empirical limits for genome-wide significance were established at 2.91 and 2.4 for parametric LOD and NPL values, respectively. These genome-wide significance thresholds represent the NPL and LOD values that could be achieved by chance in our data set at a probability of 0.01.

### Analysis of Candidate Genes on Chromosome 8p22-p21.2

A bioinformatics search for candidate genes located at the minimal region on chromosome 8p22-p21.2 was conducted using the *Ensembl Genome Browser* (http://www.ensembl.org) database (build 34b). We selected the genes CLU and PPP3CC as candidate genes in this region and a search by direct sequencing for mutations in coding region of these genes and flanking intron regions was undertaken in 3 affected members of these kindred (individuals I5, II4, and III4).

## RESULTS

In the screening of the whole-genome LOD scores obtained through this approach for both the parametric and nonparametric criteria, no region achieved an LOD score greater than 3.2 except the SNP markers rs1390943, rs1390940, and rs898249 on chromosome 8p22-21.2, 34.56 to 36.71 Mb. The highest NPL score, 3.24 ($P = 0.00006$), was achieved by marker rs898249. The same genomic position also yielded the highest multipoint heterogeneity LOD (HLOD) score under a common recessive model of disease susceptibility (HLOD = 3.04). When comparing the results of the model-dependent analysis at this locus, a higher score was obtained in the recessive model (3.04) than in the dominant model (1.0).

Exclusion maps for the different chromosomes were compiled by combining the exclusion regions of the individual markers. Chromosomes 2, 6 to 7, and 12 to 16 were completely excluded. Assuming a total genome length of 3699 cM, 2855 cM could be excluded. In addition, an NPL score of 1.31 ($P = 0.007$) was obtained by SNP rs3810261 on region 19q13.33 only 5 cM apart from the APOE loci, which was among the 5 highest NPL scores obtained in the genome-wide screen.

Merlin was also used to reconstruct the most likely haplotypes segregating in the pedigrees corresponding to the most likely pattern of gene flow.[6] None of the 3 families shared a haplotype pattern spanning more than 2 markers. Because we could not restrict the linked region by the shared haplotype, we considered as the linked region a 14.5 Mb region (delimited by markers rs967326 and rs1446687) that had a LOD score $> 2.0$ in the nonparametric analysis. According to the Ensembl Genome Browser database, this genomic region contains more than 50 genes or ORFs. One of these genes close to the maximum LOD score found is PPP3CC, which codifies for the protein phosphatase 3 (formerly 2B) catalytic subunit, gamma isoform, a calcium-dependent calmodulin-stimulated protein phosphatase that may have a role in the calmodulin activation of calcineurin and is expressed in brain cortex and cerebellum.[9,10] As it has been associated with schizophrenia in several studies,[11,12] we sequenced its coding region and flanking intron regions of the PPP3CC gene without finding any mutation. Clusterin gene lies on the centromeric border of this region and had an NPL LOD score of 2.1. Despite this low LOD score, we sequenced it because it is a strong candidate gene, as it has been found associated to LOAD in the 2 most recent LOAD Genome Wide Association Studies (GWAS).[13,14]

## DISCUSSION

We report the first genome-wide screen conducted in a LOAD extensive family, with the result pointing to a LOAD locus on chromosome 8p22-p21.2. Furthermore, a direct comparison of the recessive and dominant model has determined a recessive model of transmission in these kindred as the optimal choice, which may have important implications to explain apparently sporadic cases of LOAD.

These kindred was thus a rather unique candidate in revealing new genetic causes of AD because of the number of affected and nonaffected DNA samples available and their origin in a genetic isolate. Genetic isolation increases the chances of finding a causative gene by decreasing the genetic complexity of the disease as it leads to lower heterogeneity and a monogenic or oligogenic disorder. Furthermore, a region in a state of linkage disequilibrium, which contains the responsible gene could be expected. According to the genealogical study, this genetic isolate could be considered a recent isolate (fewer than 20 generations),[3] so a long disequilibrium region ($> 1$ cM) is likely. GWAS have already been proven useful for AD when applied to isolated populations. In a Finnish population, a total of 8 chromosomal regions were identified.[15] In a tribal Arab Palestinian population, the most significant evidence for allelic association was observed on chromosomes 2, 9, and 10, with some evidence for association on chromosome 12 in the region implicated by outbreed LOAD populations.[16] However, extensive families with LOAD, which are suitable for genome-wide screens are rare. It is worth noting that only one genome

linkage analysis has been conducted with cases from two extended pedigrees and which points to the implication of gene TRPC4AP in chromosome 20q11.22,[17] whereas all other studies have included large samples of sporadic or mixed familial AD cases.

Dementia was present in these kindred in several generations, which at first pointed to an autosomal dominant pattern of transmission. Transmission through affected males made mitochondrial inheritance unlikely, and the occurrence of several instances of male-to-male transmissions excluded an X-linked inheritance pattern. The complex segregation analysis performed revealed the autosomal dominant pattern as the best model.[3] However, when we genotyped the family, the parametric model which obtained the highest LOD score was the autosomal recessive one. Our interpretation of this discrepancy is that the high inbreeding present in these families may lead to a pseudodominant pattern of transmission of a recessive trait, which might be very common in this population. We should keep in mind the possible presence of recessive forms of AD given that they have been reported in other neurodegenerative diseases such as Parkinson disease. Several recent findings also support the involvement of recessive genes as another important cause of AD. In the Wadi Ara study, which screened 821 elderly residents of a rural community in northern Israel with high rate of consanguinity, 20% of residents over 65 years of age (twice the usual rate), and 60% of those over 85 (compared with 40% in the general population) had AD.[18] A previous study indicated that AD occurs at a higher rate in the Saguenay region of Quebec, a Canadian community with a high incidence of intermarriage.[19] Recently, a mutation on *APP* with recessive transmission has been reported[20] and a genome-wide screen searching for extended homozygosity (shared regions of more than 1 Mb) carried out in 837 cases with LOAD has identified a homozygous region,[21] incidentally on chromosome 8p11.23, only 17 centimorgans apart from the loci we have reported.

Our screen has identified a LOAD locus on chromosome 8p22-p21.2, which has been reported previously in another genome-wide screen using approximately 6000 SNP markers at an average intermarker distance of 0.65 cM; this screen was carried out in 1902 individuals from 328 families with LOAD and 236 unrelated controls.[22] In this study, SNP rs4427168 at 8p21.3 showed significant values for linkage and association. This SNP is flanked by the SNPs rs720266 and rs952299, which have achieved an NPL LOD score of 3.02 ($P = 0.001$) and 3.19 ($P = 0.0006$) in our study, thus making it very likely that these SNPs share the same linkage signal for LOAD. Moreover, the locus of clusterin gene, which is less than 5 cM apart from the maximum NPL of our study, has recently been reported as associated with AD in 2 extensive genome-wide linkage studies.[13,14]

We previously established a higher frequency of the APOE-4 allele in these patients,[3] and APOE is located in the region of the peak linkage we detected on chromosome 19. This is consistent with other genome-wide screens on LOAD that had made a similar observation.[23–25] It is important to remark that there were no other significant loci in this study, unlike most GWAS performed in LOAD in the last couple of years, which have analyzed hundreds of cases and controls. These studies have included great genetic heterogeneity and have found many significant loci.[13,14,17,24–28]

Several putative candidate genes for AD could be considered in this region 8p22-p21.2. Among such genes are lipoprotein lipase and amino-acetyl transferases 1 and 2, which have been previously considered in association studies for AD and whose results have been controversial. A candidate gene in this region that we took into consideration is PPP3CC. Thought to be involved in hippocampal-dependent synaptic plasticity and memory storage,[9] this gene has been associated with schizophrenia in several studies.[11,12] As it is located very close to the maximum LOD score found, we carried on a sequence analysis of the PPP3CC but did not detect any mutations. Another gene close to this region is the CLU gene, which codifies clusterin (also known as apolipoprotein J) a chaperone protein that regulates amyloid formation and clearance. We sequenced its coding region without finding any mutation. Nevertheless, there are more than 50 known genes and ORFs in the region of linkage that merit consideration for future analysis.

In summary, these genetically informative kindred make the case for a new locus in 8p22-p21.2 associated with a recessive form of LOAD. As two other studies point to the same region or one very close to it, this genome spot merits an exhaustive search for candidate genes. These kindred further supports the role of recessive genes linked to LOAD, a phenomenon that may also play a role in explaining some apparently sporadic cases.

## REFERENCES


1. Bertram L, McQueen MB, Mullin K, et al. Systematic meta-analyses of Alzheimer disease genetic association studies: the AlzGene database. *Nat Genet*. 2007;39:17–23.
2. Bertram L, Tanzi RE. The genetic epidemiology of neurodegenerative disease. *J Clin Invest*. 2005;115:1449–1457.
3. Jimenez-Escrig A, Gomez-Tortosa E, Baron M, et al. A multigenerational pedigree of late-onset Alzheimer's disease implies new genetic causes. *Brain*. 2005;128:1707–1715.
4. Ruschendorf F, Nurnberg P. ALOHOMORA: a tool for linkage analysis using 10 K SNP array data. *Bioinformatics*. 2005;21:2123–2125.
5. O'Connell JR, Weeks DE. PedCheck: a program for identification of genotype incompatibilities in linkage analysis. *Am J Hum Genet*. 1998;63:259–266.
6. Abecasis GR, Cherny SS, Cookson WO, et al. Merlin–rapid analysis of dense genetic maps using sparse gene flow trees. *Nat Genet*. 2002;30:97–101.
7. Whittemore AS, Halpern J. A class of tests for linkage using affected pedigree members. *Biometrics*. 1994;50:118–127.
8. Kong A, Cox NJ. Allele-sharing models: LOD scores and accurate linkage tests. *Am J Hum Genet*. 1997;61:1179–1188.
9. Malleret G, Haditsch U, Genoux D, et al. Inducible and reversible enhancement of learning, memory, and long-term potentiation by genetic inhibition of calcineurin. *Cell*. 2001; 104:675–686.
10. Yamada K, Gerber DJ, Iwayama Y, et al. Genetic analysis of the calcineurin pathway identifies members of the EGR gene family, specifically EGR3, as potential susceptibility candidates in schizophrenia. *Proc Natl Acad Sci U S A*. 2007;104: 2815–2820.
11. Horiuchi Y, Ishiguro H, Koga M, et al. Support for association of the PPP3CC gene with schizophrenia. *Mol Psychiatry*. 2007;12:891–893.
12. Tabares-Seisdedos R, Rubenstein JL. Chromosome 8p as a potential hub for developmental neuropsychiatric disorders: implications for schizophrenia, autism and cancer. *Mol Psychiatry*. 2009;14:563–589.
13. Harold D, Abraham R, Hollingworth P, et al. Genome-wide association study identifies variants at CLU and PICALM associated with Alzheimer's disease. *Nat Genet*. 2009;41: 1088–1093.

14. Lambert JC, Heath S, Even G, et al. Genome-wide association study identifies variants at CLU and CR1 associated with Alzheimer's disease. *Nat Genet*. 2009;41: 1094–1099.
15. Hiltunen M, Mannermaa A, Thompson D, et al. Genome-wide linkage disequilibrium mapping of late-onset Alzheimer's disease in Finland. *Neurology*. 2001;57:1663–1668.
16. Farrer LA, Friedland RP, Bowirrat A, et al. Genetic and environmental epidemiology of Alzheimer's disease in arabs residing in Israel. *J Mol Neurosci*. 2003;20:207–212.
17. Poduslo SE, Huang R, Huang J, et al. Genome screen of late-onset Alzheimer's extended pedigrees identifies TRPC4AP by haplotype analysis. *Am J Med Genet B Neuropsychiatr Genet*. 2009;150B:50–55.
18. Bowirrat A, Oscar-Berman M, Logroscino G. Association of depression with Alzheimer's disease and vascular dementia in an elderly Arab population of Wadi-Ara, Israel. *Int J Geriatr Psychiatry*. 2006;21:246–251.
19. Vezina H, Heyer E, Fortier I, et al. A genealogical study of Alzheimer disease in the Saguenay region of Quebec. *Genet Epidemiol*. 1999;16:412–425.
20. Di FG, Catania M, Morbin M, et al. A recessive mutation in the APP gene with dominant-negative effect on amyloidogenesis. *Science*. 2009;323:1473–1477.
21. Nalls MA, Guerreiro RJ, Simon-Sanchez J, et al. Extended tracts of homozygosity identify novel candidate genes associated with late-onset Alzheimer's disease. *Neurogenetics*. 2009;10: 183–190.
22. Lee JH, Cheng R, Graff-Radford N, et al. Analyses of the National Institute on Aging Late-Onset Alzheimer's Disease Family Study: implication of additional loci. *Arch Neurol*. 2008;65:1518–1526.
23. Reiman EM, Webster JA, Myers AJ, et al. GAB2 alleles modify Alzheimer's risk in APOE epsilon4 carriers. *Neuron*. 2007;54:713–720.
24. Grupe A, Abraham R, Li Y, et al. Evidence for novel susceptibility genes for late-onset Alzheimer's disease from a genome-wide association study of putative functional variants. *Hum Mol Genet*. 2007;16:865–873.
25. Li H, Wetten S, Li L, et al. Candidate single-nucleotide polymorphisms from a genomewide association study of Alzheimer disease. *Arch Neurol*. 2008;65:45–53.
26. Potkin SG, Guffanti G, Lakatos A, et al. Hippocampal atrophy as a quantitative trait in a genome-wide association study identifying novel susceptibility genes for Alzheimer's disease. *PLoS One*. 2009;4:e6501.
27. Beecham GW, Schnetz-Boutaud N, Haines JL, et al. CALHM1 polymorphism is not associated with late-onset Alzheimer disease. *Ann Hum Genet*. 2009;73:379–381.
28. Bertram L, Lange C, Mullin K, et al. Genome-wide association analysis reveals putative Alzheimer's disease susceptibility loci in addition to APOE. *Am J Hum Genet*. 2008;83:623–632.

CASE OF THE MONTH

# AUTOSOMAL RECESSIVE EMERY–DREIFUSS MUSCULAR DYSTROPHY CAUSED BY A NOVEL MUTATION (R225Q) IN THE LAMIN A/C GENE IDENTIFIED BY EXOME SEQUENCING

**ADRIANO JIMENEZ-ESCRIG, MD, PhD,[1,2] ISABEL GOBERNADO, MD,[2,3] MERCEDES GARCIA-VILLANUEVA, MD, PhD,[4] and ANTONIO SANCHEZ-HERRANZ, BS, PhD[2,5]**

[1] Servicio de Neurologia, Hospital Ramon y Cajal and Universidad de Alcala, 28034 Madrid, Spain
[2] Unidad Central de Apoyo a Estudios Genomicos, IRYCIS, Madrid, Spain
[3] Servicio de Psiquiatria, Hospital Ramon y Cajal, Madrid, Spain
[4] Servicio de Anatomia Patologica, Hospital Ramon y Cajal, Madrid, Spain
[5] Servicio de Neurobiologia, Hospital Ramon y Cajal, Madrid, Spain

*Accepted 10 October 2011*

ABSTRACT: *Introduction:* The aim of this study is to describe a new mutation in the *LMNA* gene diagnosed by whole exome sequencing. *Methods:* A two-generation kindred with recessive limb-girdle muscular dystrophy was evaluated by exome sequencing of the proband's DNA. *Results:* Exome sequencing disclosed 194,618 variants (170,196 SNPs, 8482 MNPs, 7466 insertions, 8307 deletions, and 167 mixed combinations); 71,328 were homozygotic and 123,290 were heterozygotic, with 11,753 non-synonymous, stop-gain, stop-loss, or frameshift mutations occurring in the coding region or nearby intronic region. The cross-referencing of these mutations in candidate genes for muscular dystrophy showed a homozygote mutation c.G674A in exon 4 of *LMNA* causing a protein change R225Q in an arginine conserved from human to *Xenopus tropicalis* and in lamin B1. *Conclusions:* This technique will be preferred for studying patients with muscular dystrophy in the coming years.
*Muscle Nerve* **45:** 605–610, 2012

The clinical evaluation of patients with muscular dystrophy begins with an assessment of the pattern of transmission and the distribution of the muscle weakness followed by a number of ancillary tests, including serum creatine kinase (CK) measurement, electrocardiography (ECG), computerized tomography (CT) scanning or magnetic resonance imaging (MRI), and muscle biopsy, aimed to identify the involved gene. A final diagnosis is made when a confirmatory DNA mutation is discovered in any of the muscular dystrophy–causing genes.[1,2] Depending on the intricacy of the phenotype, it may take months or even years to ascertain the causal mutation. This is not always feasible, causing a significant loss in clinical and prognostic information. Furthermore, genetic diagnosis is critical if genetic counseling is desired and, in the near future, for inclusion of these patients in a gene therapy program.[3,4]

Over the last year, development of the next generation of sequencing technology has progressively facilitated sequencing of the whole genome, or only the coding region (exome), which harbors 85% of disease mutations,[5] in a time/cost frame suitable for application in the clinical setting.[6,7] However, data on the use of this technique for clinical diagnosis are scarce, although recently a few reports have highlighted its potential impact.[8,9] Herein we report the clinical and genetic data on a family with autosomal recessive Emery–Dreifuss muscular dystrophy caused by a novel mutation (R225Q) in the lamin A/C (*LMNA*) gene identified by exome sequencing.

## CASE REPORTS

The kindred being studied included 6 siblings whose parents were second cousins. Figure 1 shows the pedigree, and their clinical data are summarized in Table 1. Four had a limb-girdle progressive muscular dystrophy with onset in the first to third decade of life. Severity was variable among the siblings.

**Patient III-1 (Proband).** This 50-year-old woman began to have difficulty running at 14 years of age, with progressive proximal arm and leg muscle weakness, toe walking, heel-cord contractures, and loss of independent ambulation at 35 years of age. At age 28, a nerve conduction study of the fibular nerve was normal. At the time of this study she had normal cognition, cranial nerves, eye movements, coordination, and sensation. She had mild facial weakness that did not disturb whistling, smiling, or eye closure. Severe wasting was present in the shoulder and pelvic girdles, deltoid, and quadriceps muscles. In addition, neck contractures limited neck flexion, but elbow contractures were not present. Muscle power was Medical Research Council (MRC) grade 4 in the neck flexors and

**Abbreviations:** A, adenine; C, cytosine; CK, creatine kinase; CSF, cerebrospinal fluid; CT, computed tomography scan; ECG, electrocardiogram; p.H222Y, protein.histidine222tyrosine; *LMNA*, lamin A/C; *LMNB1*, lamin B1; MNP, multiple nucleotide polymorphism; MRC, Medical Research Council; MRI, magnetic resonance image; PCR, polymerase chain reaction; p.R225Q, protein.arginine225glutamine; SNP, single nucleotide polymorphism

**Key words:** Emery–Dreifuss; exome; LMNA; muscular dystrophy; next-generation sequencing

**Correspondence to:** A. Jimenez-Escrig; e-mail: adriano.jimenez@hrc.es



Published online 17 October 2011 in Wiley Online Library (wileyonlinelibrary.com). DOI 10.1002/mus.22324

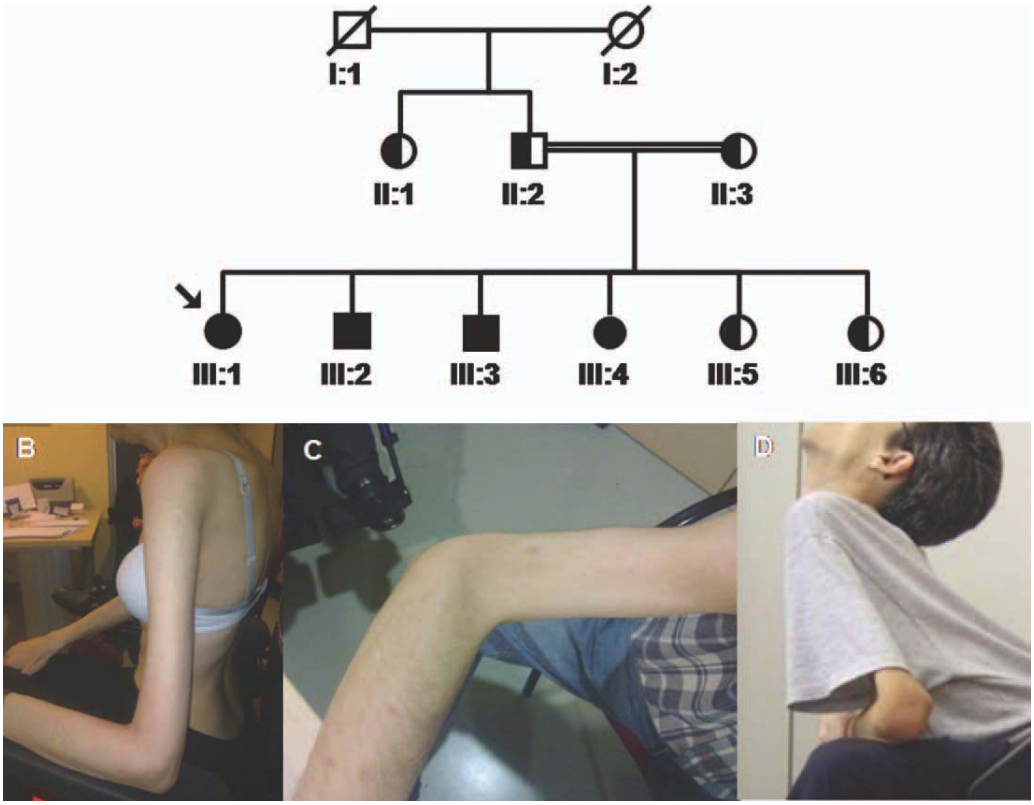



**FIGURE 1.** Pedigree of the kindred showing genotyping status. (**B**–**D**) Muscular dystrophic pattern in patients III-1, III-2, and III-3. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

extensors; 3 in the distal arm muscles; and 2 or 3 in the scapular, pelvic, and proximal limb muscles. She underwent ECGs at 26 and 32 years of age, both normal. After the genetic diagnosis, an ECG showed frequent supraventricular premature beats. In addition, she also manifested von Willebrand disease with bleeding during ovarian surgery.

**Patient III-2.** This 46-year-old man was first evaluated at 12 years of age for clumsy gait. At 16, he underwent surgery for Achilles tendon shortening. The muscular dystrophy progressed with shoulder and pelvic girdle involvement, but gait and proximal arm function were still preserved at the time of this evaluation. Elbow contractures since adolescence prevented full extension. At 40 years of age, he complained of palpitations and had Holter monitoring that showed episodes of atrial tachycardia and frequent supraventricular premature contractions as well as some blocked P waves and ventricular premature contractions. An echocardiogram obtained at that time was normal.

**Patient III-3.** This 43-year-old man first experienced difficulties arising from the floor at 4 years of age and developed a rapidly progressive gait disturbance; at age 5, a muscle biopsy from the right quadriceps revealed variation in fiber size and mild interstitial infiltrates. At age 12, Achilles lengthening was done, and elbow contractures were reported. At age 21, he began using a wheelchair. At age 40, he had severe and diffuse muscle weakness that was more pronounced in the neck flexors. Contractures were noted in the elbows and

**Table 1.** Clinical and genetic characteristics.

| Case | Age/age at onset/ age in wheelchair (years) | Age of cardiac involvement | Serum CK | Cardiac manifestations | c.G674A mutation |
|---|---|---|---|---|---|
| III-1 | 50/14/35 | At genetic diagnosis | 2× | Supraventricular premature beats | Homozygous |
| III-2 | 46/14/still ambulant | 41 y | 2–3× | Supraventricular and ventricular premature beats | Homozygous |
| III-3 | 43/4/25 | At genetic diagnosis | 1–3× | Supraventricular and ventricular premature beats | Homozygous |
| III-4 | 41/third decade/still ambulant | 39 y | 2–3× | Supraventricular premature beats | Homozygous |
| III-5 | 39/–/– | – | 1 | Normal | Heterozygous |
| III-6 | 37/–/– | – | 1 | Normal | Heterozygous |

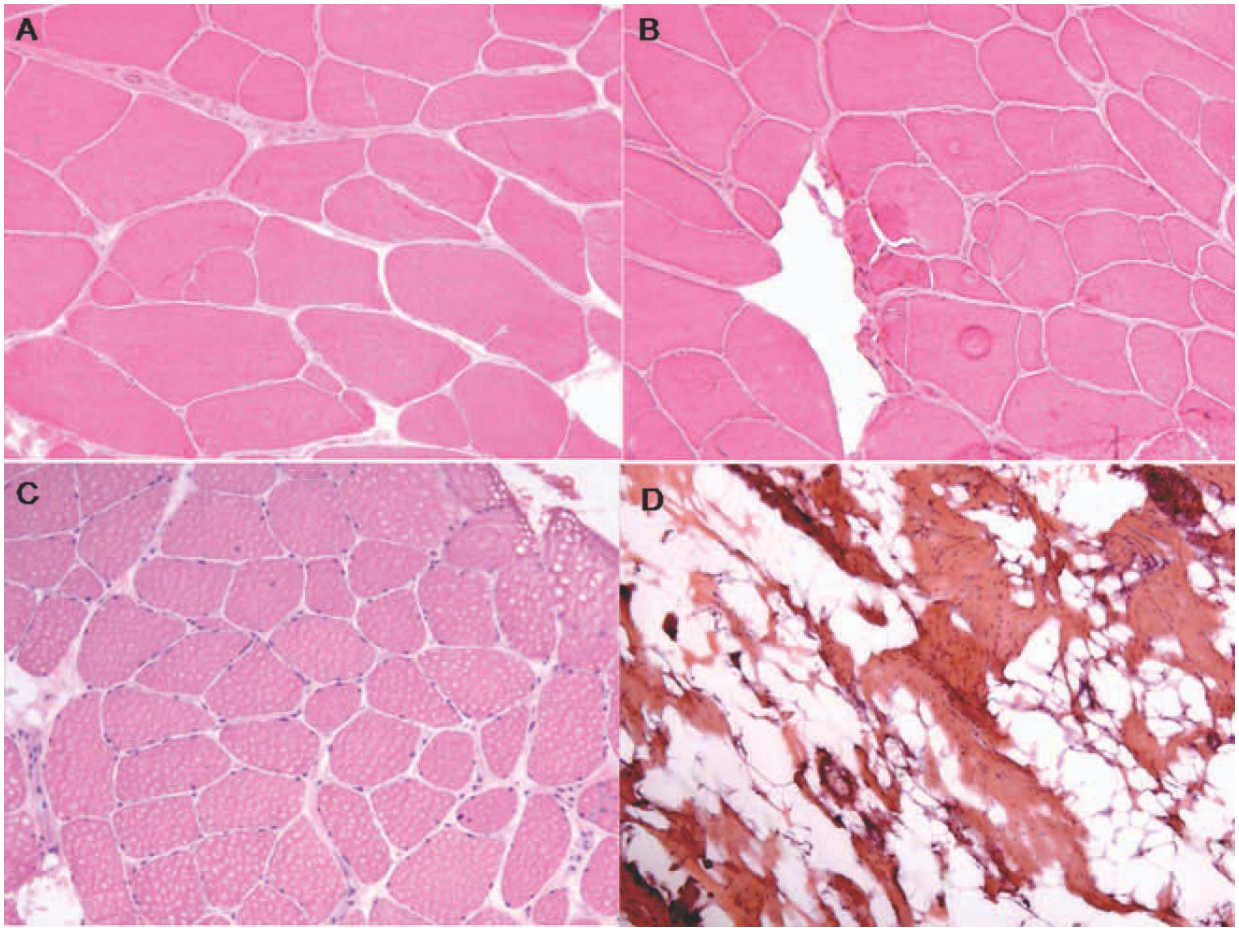



**FIGURE 2.** Muscle findings in patients III-1 and III-2 (hematoxylin and eosin stain). Muscle biopsies from deltoid (**A**, **C**) and quadriceps muscles (**B**, **D**) showing very moderate dystrophic changes with increased variability in fiber size (**A**–**C**). (**D**) Severe muscular dystrophy with replacement of muscle fibers by fat and connective tissue. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

hips. His intelligence was normal; he had never complained of cardiac disturbances, but once the genetic diagnosis of Emery–Dreifuss muscular dystrophy was made, a cardiac evaluation showed dense supraventricular and ventricular premature contractions.

**Patient III-4.** This 41-year-old woman was first evaluated at 24 years of age after presenting with right-sided horizontal diplopia and dizziness. She underwent a cranial MRI and serologic and CSF studies and was diagnosed with multiple sclerosis. Since that time, she has had several episodes of dizziness, paresthesia, or motor discoordination in the upper limbs, along with progressive arreflexic lower limb weakness. At 39, she complained of tachyarrhythmia, and ECG revealed premature supraventricular beats. Once the *LMNA* R225Q mutation was identified in this kindred she was found to be homozygous for the mutation. At this age, she had lower and upper limb proximal weakness (MRC $4^{+}/5$) and muscle atrophy in the pelvic and shoulder girdles. Walking was possible with a cane.

Patient II-2, the proband's father, was asymptomatic until 60 years of age, when he was diagnosed with myasthenia gravis. Until that time, he had regularly practiced sports and never had muscular complaints. At 80, he developed syncope and had a pacemaker implanted. His sister, patient II-1, had normal strength but was dependent upon a pacemaker since age 78, when she developed atrial ventricular block.

**Histopathological Studies.** Patients underwent five muscle biopsies throughout their evolution. Patient III-1 had a deltoid muscle biopsy that showed dystrophic changes, including variability of fiber size, increased endomysial connective tissue, and signs of necrosis and regeneration (Fig. 2). Patient III-2 had three muscle biopsies; the first two were of quadriceps and biceps and had dystrophic features, which were more marked in the quadriceps than in biceps. The third biopsy was in

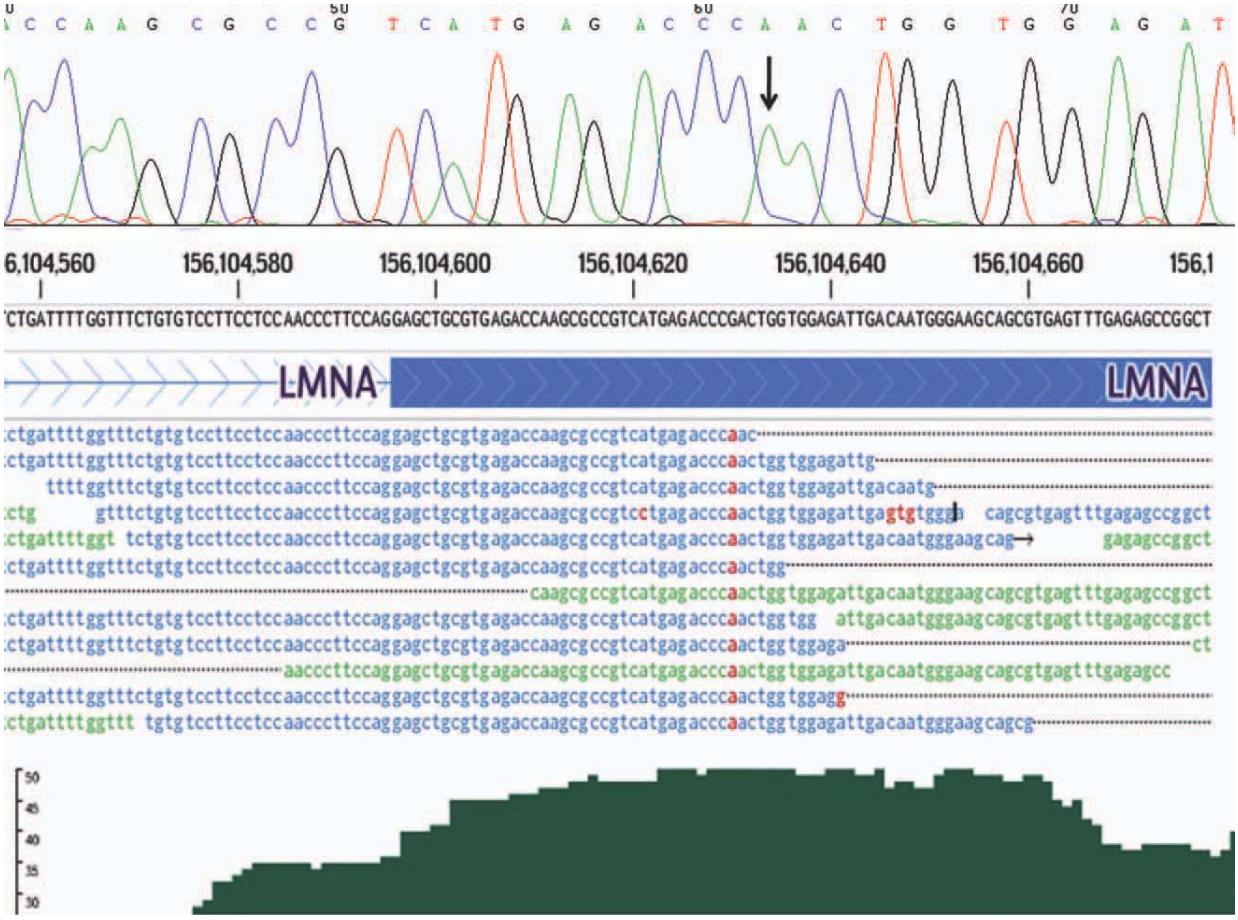



**FIGURE 3.** Next-generation sequencing multiple alignments at *LMNA* gene exon 4, showing the G→A mutation; and Sanger sequencing at this locus (top). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

the left quadriceps and showed extensive replacement of muscle fibers by fat and connective tissue. Patient III-3 was biopsied at 5 years of age. This biopsy was reported as suggestive of Duchenne muscular dystrophy.

**Genetic Diagnosis.** Genetic study previously done in this kindred had ruled out mutations in the calpain, dysferlin, and sarcoglycan-alpha, -beta, and -gamma genes. Because it was not possible to perform a targeted gene search we decided to undertake whole exome sequencing to identify the causal mutation. DNA from Patient III-1 was extracted from blood lymphocytes using a QIAamp DNA Blood Maxi Kit (Qiagen, Valencia, California). Whole exome sequencing was done at Otogenetics, Inc., using 10 $\mu$g of DNA, with an Illumina library preparation, exome capture, and next-generation sequencing by HiSeq2000 with a paired-end (2 × 100) protocol using a capture kit (SeqCap EZ Exome; NimbleGen) and exome enrichment kit (TruSeq; Illumina) for exome capture.

A total of 63,952,106 sequencing reads were produced, comprising 5.755 billion bases. From these, 91.34% were aligned with the human reference genome (hg19); thus, from a total of 194,954 exons, 187,444 (96.1%) were completely sequenced and 6260 partially sequenced, with an average coverage of 42×, thus leaving 1250 unsequenced exons. Exome sequencing found 194,618 variants [170,196 SNPs, 8482 MNPs (single and multiple nucleotide polymorphisms, respectively), 7466 insertions, 8307 deletions, and 167 mixed combinations] and 71,328 homozygotes and 123,290 heterozygotes, with 11,753 non-synonymous, stop-gain, stop-loss, or frameshift mutations occurring in the coding region or nearby intronic region. Cross-referencing of these mutations in candidate genes for muscular dystrophy showed a homozygote mutation c.G674A in exon 4 of *LMNA* causing the protein change R225Q (Fig. 3) in an arginine conserved from human to *Xenopus tropicalis* and in lamin B1 (Fig. 4). Sequence analysis of the rest of the *LMNA* gene in the patient demonstrated that the c.G674A was the only mutation.

We obtained informed consent from the family members who were receiving genetic counseling, clinical screening, and peripheral blood sampling

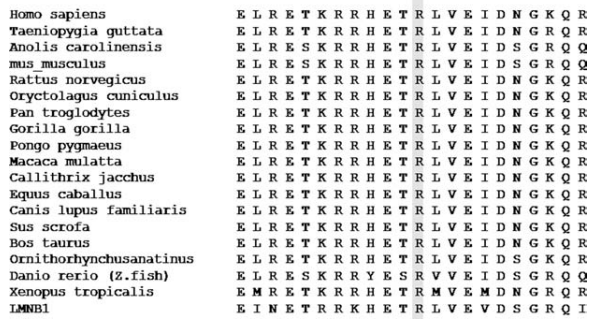



**FIGURE 4.** A multiple sequence alignment of the region of the *LMNA* protein segment containing the variant. The R225 is visible in bold. This arginine is conserved in all species identified and in the *LMNB1* gene.

for genetic testing. We confirmed the mutation by Sanger sequencing after polymerase chain reaction (PCR) amplification of exon 4 using primers and protocols as previously reported,[10] and used this method to test the rest of the kindred. All members of the last generation were clinically examined and genotyped, and their data are shown in Table 1. No other causative mutation was found. To exclude the possibility that mutation R225Q was a polymorphism, 200 chromosomes of unaffected individuals were analyzed.

## DISCUSSION

The patients studied showed characteristic features of Emery–Dreifuss muscular dystrophy (MIM 310300 and 310200), with early-onset contractures of elbows, ankles, and neck and progressive muscle weakness of the shoulder and pelvic girdle muscles in adulthood. Three modes of inheritance exist in this condition: X-linked, autosomal dominant, and autosomal recessive. X-linked Emery–Dreifuss muscular dystrophy is the most common form, affecting 1 in 1,000,000 people.[11] The incidence of the autosomal dominant form is unknown. The autosomal recessive type appears to be very rare; only a few cases have been reported.[12,13]

A number of factors led to a delay in genetic diagnosis in this kindred for several years: (1) contractures were underestimated due to the main complaint of muscle weakness; (2) the cardiomyopathy had a late onset and was nearly missed; and (3) recessive forms of *LMNA* mutations are very uncommon. The diagnosis was only possible once the capacity to do whole exome sequencing was developed. In particular, the cardiologic manifestations, which are the main hazardous condition in this disease, were underdiagnosed in this kindred until the *LMNA* gene mutation was discovered.

Cardiac involvement in Emery–Dreifuss includes atrial and ventricular arrhythmias, disorders of atrioventricular conduction, dilated cardiomyopathy, and sudden death.[14] Cardiac symptoms can appear later in the evolution of the disease, with a time lag reported between the onset of the muscle disease and cardiac disease ranging from 7 to 35 years and milder cardiac involvement in recessive forms.[14] *LMNA* mutations can present with supraventricular arrhythmias (atrial premature contractions, atrial tachycardia, atypical atrial flutter, atrial fibrillation, and the uncommon condition of atrial paralysis), disorders of atrioventricular conduction (any degree of atrioventricular block), ventricular arrhythmias (ventricular premature beats, non-sustained ventricular tachycardia), and impairment of left ventricular systolic function in the absence of left ventricular dilation (non-dilated cardiomyopathy).[15,16] Moreover, *LMNA* gene mutations are the most frequent genetic cause of dilated cardiomyopathy, accounting for 6–8% of all primary dilated cardiomyopathies and up to 40% when conduction disorders are present.[17–19] *LMNA* mutations have been shown to be associated with a very poor prognosis due to a high rate of sudden cardiac death and severe forms of heart failure requiring heart transplantation.[20] Dilated cardiomyopathies caused by *LMNA* gene defects are highly penetrant and can be malignant conditions characterized by a high rate of severe left ventricular dysfunction and life-threatening arrhythmias, which should lead to consideration of special indications for intracardiac defibrillator implantation.[21] Because heterozygote carriers of *LMNA* mutations can have no neuromuscular symptoms, heterozygote mutations might be responsible for cardiomyopathy or sudden death in the general population.[22,23] These patients will need Holter ECG as a screening method to rule out arrhythmias, as routine ECG can be less sensitive, and there could be treatment and prognosis implications.

The *LMNA* gene encodes two lamins, A and C, by differential maturation of the 3′ end of the mRNA. In addition to autosomal dominant Emery–Dreifuss, the *LMNA* gene is responsible for the autosomal recessive and a semidominant form of the disorder.[12] The p.R225Q variant has not been reported so far, but we ascribe pathogenic significance to it because the mutation segregated with the trait in this pedigree and the R225 residue is highly conserved in the phylogeny (Fig. 4). A very similar clinical profile was found in a patient homozygous for the c.C664T mutation, causing the amino acid change p.H222Y, affected by an autosomal recessive form of the disease.[12] The variable intrafamilial range of age of onset and severity has been reported in other missense mutations in *LMNA*.[24]

Whole exome sequencing is minimally invasive, operative (results can be obtained in <4 weeks),

and has a cost of ∽$2000, which is in the range of other tests used in clinical settings. In this regard, it can now be considered in the diagnostic work-up of patients with heterogenic disorders. This study has demonstrated the usefulness of the new-generation sequencing in the evaluation of patients with genetic myopathies. The kindred investigated showed characteristic features of Emery–Dreifuss muscular dystrophy, but other genes, apart from emerin that is an X-linked disorder, were considered first. As noted earlier in this study, the genetic diagnosis of patients with muscular dystrophies is arduous and can sometimes be elusive. There have been some reports of genetic diagnoses of neurological disease using exome sequencing, but they were made by exome sequencing of several members of an involved kindred.[8,9] In this report, exome sequencing was done only in the proband, and it determined the causal mutation.

An additional benefit of exome sequencing is that it provides a full picture of all the genes intrinsically involved in myopathies. It is therefore possible to get supplementary information on other genes that might have a modulator effect that can explain differences in severity among siblings. Besides, information on other genes can depict incidental present or future disorders such as in our patient with von Willebrand disease. It can also the detect carrier status of diseases present in other relatives. This technique will very likely become the preferred approach for studying patients with muscular dystrophy.

In conclusion, we have reported the usefulness of exome sequencing in the detection of the causal mutation in this kindred. Exome sequencing determined the causal mutation and was able to detect the polymorphisms presented in muscular dystrophy genes previously evaluated. In addition, we have reported a novel p.R225Q mutation in the *LMNA* gene that causes Emery–Dreifuss muscular dystrophy in homozygotic individuals and late-onset cardiomyopathy in heterozygote carriers. The diagnosis of autosomal recessive Emery–Dreifuss muscular dystrophy in this family allowed for an accurate identification of carrier status. This may prevent sudden death in individuals who lack muscle symptoms.

## REFERENCES


1. Emery AE. The muscular dystrophies. Lancet 2002;359:687–695.
2. Rocha CT, Hoffman EP. Limb-girdle and congenital muscular dystrophies: current diagnostics, management, and emerging technologies. Curr Neurol Neurosci Rep 2010;10:267–276.
3. Burgunder JM, Schols L, Baets J, Andersen P, Gasser T, Szolnoki Z, et al. EFNS guidelines for the molecular diagnosis of neurogenetic disorders: motoneuron, peripheral nerve and muscle disorders. Eur J Neurol 2011;18:207–217.
4. Grosse SD, Kalman L, Khoury MJ. Evaluation of the validity and utility of genetic testing for rare diseases. Adv Exp Med Biol 2010;686: 115–131.
5. Choi M, Scholl UI, Ji W, Liu T, Tikhonova IR, Zumbo P, et al. Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. Proc Natl Acad Sci USA 2009;106:19096–19101.
6. Mardis ER. The impact of next-generation sequencing technology on genetics. Trends Genet 2008;24:133–141.
7. ten Bosch JR, Grody WW. Keeping up with the next generation: massively parallel sequencing in clinical diagnostics. J Mol Diagn 2008; 10:484–492.
8. Lupski JR, Reid JG, Gonzaga-Jauregui C, Rio DD, Chen DC, Nazareth L, et al. Whole-genome sequencing in a patient with Charcot–Marie–Tooth neuropathy. N Engl J Med 2010;362:1181–1191.
9. Montenegro G, Powell E, Huang J, Speziani F, Edwards YJ, Beecham G, et al. Exome sequencing allows for rapid gene identification in a Charcot–Marie–Tooth family. Ann Neurol 2011;69:464–470.
10. Perrot A, Hussein S, Ruppert V, Schmidt HH, Wehnert MS, Duong NT, et al. Identification of mutational hot spots in LMNA encoding lamin A/C in patients with familial dilated cardiomyopathy. Basic Res Cardiol 2009;104:90–99.
11. Norwood FL, Harling C, Chinnery PF, Eagle M, Bushby K, Straub V. Prevalence of genetic muscle disease in Northern England: in-depth analysis of a muscle clinic population. Brain 2009;132:3175–3186.
12. Di Raffaele BM, Ricci E, Galluzzi G, Tonali P, Mora M, Morandi L, et al. Different mutations in the LMNA gene cause autosomal dominant and autosomal recessive Emery–Dreifuss muscular dystrophy. Am J Hum Genet 2000;66:1407–1412.
13. Zirn B, Kress W, Grimm T, Berthold LD, Neubauer B, Kuchelmeister K, et al. Association of homozygous LMNA mutation R471C with new phenotype: mandibuloacral dysplasia, progeria, and rigid spine muscular dystrophy. Am J Med Genet A 2008;146A:1049–1054.
14. Sanna T, Dello Russo A, Toniolo D, Vytopil M, Pelargonio G, De Martino G, Ricci E, et al. Cardiac features of Emery–Dreifuss muscular dystrophy caused by lamin A/C gene mutations. Eur Heart J 2003;24:2227–2236.
15. Malek LA, Labib S, Mazurkiewicz L, Saj M, Ploski R, Tesson F, et al. A new c.1621 C > G, p.R541G lamin A/C mutation in a family with DCM and regional wall motion abnormalities (akinesis/dyskinesis): genotype–phenotype correlation. J Hum Genet 2011;56:83–86.
16. Hookana E, Junttila MJ, Sarkioja T, Sormunen R, Niemela M, Raatikainen MJ, et al. Cardiac arrest and left ventricular fibrosis in a Finnish family with the lamin A/C mutation. J Cardiovasc Electrophysiol 2008;19:743–747.
17. Arbustini E, Pilotto A, Repetto A, Grasso M, Negri A, et al. Autosomal dominant dilated cardiomyopathy with atrioventricular block: a lamin A/C defect-related disease. J Am Coll Cardiol 2002;39: 981–990.
18. Hermans MC, Pinto YM, Merkies IS, de Die-Smulders CE, Crijns HJ, Faber CG. Hereditary muscular dystrophies and the heart. Neuromuscul Disord 2010;20:479–492.
19. Vytopil M, Benedetti S, Ricci E, Galluzzi G, Dello Russo A, Merlini L, et al. Mutation analysis of the lamin A/C gene (LMNA) among patients with different cardiomuscular phenotypes. J Med Genet 2003;40:e132.
20. van Berlo JH, de Voogt WG, van der Kooi AJ, van Tintelen JP, Bonne G, Yaou RB, et al. Meta-analysis of clinical characteristics of 299 carriers of LMNA gene mutations: do lamin A/C mutations portend a high risk of sudden death? J Mol Med 2005;83:79–83.
21. Antoniades L, Eftychiou C, Kyriakides T, Christodoulou K, Katritsis DG. Malignant mutation in the lamin A/C gene causing progressive conduction system disease and early sudden death in a family with mild form of limb-girdle muscular dystrophy. J Interv Card Electrophysiol 2007;19:1–7.
22. Fishbein MC, Siegel RJ, Thompson CE, Hopkins LC. Sudden death of a carrier of X-linked Emery–Dreifuss muscular dystrophy. Ann Intern Med 1993;119:900–905.
23. Vytopil M, Ricci E, Dello RA, Hanisch F, Neudecker S, Zierz S, et al. Frequent low penetrance mutations in the Lamin A/C gene, causing Emery Dreifuss muscular dystrophy. Neuromuscul Disord 2002;12: 958–963.
24. Kim HY, Ki CS, Kang SJ, Khang SK, Koh SH, Kim DW, et al. A novel LMNA gene mutation Leu162Pro and the associated clinical characteristics in a family with autosomal-dominant Emery–Dreifuss muscular dystrophy. Muscle Nerve 2008;38:1336–1339.

# Secuenciación de genoma completo: un salto cualitativo en los estudios genéticos

Adriano Jiménez-Escrig, Isabel Gobernado, Antonio Sánchez-Herranz

**Resumen.** En estos momentos se encuentra en plena expansión la llamada secuenciación paralela o de siguiente generación –*next generation sequencing* (NGS)–, que establece un salto de varios órdenes de magnitud en la longitud de los fragmentos secuenciados y la rapidez de su secuenciación. La NGS permite la secuenciación de un genoma humano completo en el tiempo y el coste económico de secuenciar dos o tres genes grandes con la técnica de Sanger. Mediante la NGS se pasa de examinar genes específicos seleccionados mediante estudio del fenotipo a explorar genomas enteros de grupos humanos o de otras especies. Esto está permitiendo conocer no sólo cómo es un genoma individual, sino cómo cambia el genoma humano de persona a persona, cómo difieren los genomas entre diferentes grupos humanos, e incluso cómo difiere el genoma de un tumor respecto del genoma sano del huésped.

**Palabras clave.** Exoma. Genoma. *Next generation sequencing*. Secuenciación.

Servicio de Neurología (A. Jiménez-Escrig); Servicio de Psiquiatría (I. Gobernado); Servicio de Neurobiología (A. Sánchez-Herranz); Hospital Universitario Ramón y Cajal. Unidad Central de Apoyo a Estudios Genómicos; IRYCIS (A. Jiménez-Escrig, I. Gobernado, A. Sánchez-Herranz). Madrid, España.

Correspondencia:
Dr. Adriano Jiménez Escrig.
Servicio de Neurología. Hospital Universitario Ramón y Cajal.
Ctra. Colmenar Viejo, km 9,1.
E-28034 Madrid.

E-mail:
adriano.jimenez@hrc.es

Aceptado tras revisión externa:
27.01.12.

Cómo citar este artículo:
Jiménez-Escrig A, Gobernado I, Sánchez-Herranz A. Secuenciación de genoma completo: un salto cualitativo en los estudios genéticos. Rev Neurol 2012; XX: XXX-XXX.

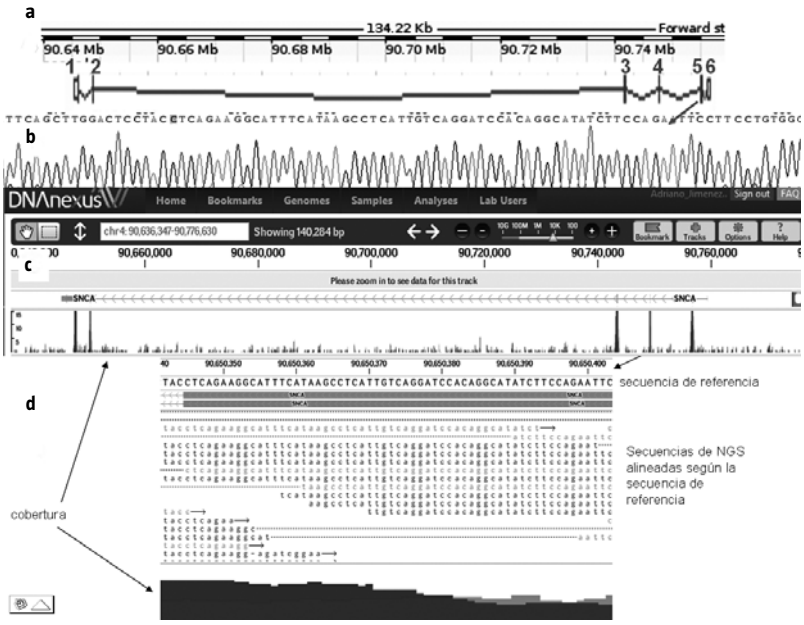


## Introducción

La secuenciación del ADN consiste en determinar el orden de las bases A, C, G y T en un fragmento de ADN. La secuenciación utilizada hasta la fecha se realiza por el método descrito por Sanger et al en 1977 [1], que permite obtener la secuencia de un fragmento determinado de ADN, un gen o parte de éste, como, por ejemplo, uno o varios exones (Figura, a). Con esta técnica se obtienen secuencias de hasta 500 bases aproximadamente (Figura, b). Sin embargo, la alta demanda de secuenciación ha llevado al desarrollo de tecnologías de secuenciación masiva basadas en realizar múltiples secuencias cortas (de alrededor de 100 pares de bases) de un modo paralelo, produciendo millones de secuencias al mismo tiempo y a un coste muy bajo. Una vez ensambladas estas secuencias a un genoma de referencia, se puede secuenciar, en lugar de un gen, múltiples genes o incluso un genoma completo. Estas secuencias individuales son más cortas y contienen más errores que la secuencia obtenida por Sanger, pero, como pueden repetirse múltiples veces (lo que se conoce como cobertura), se llega a conocer con exactitud la secuencia de millones de pares de bases (Figura, c). Es la llamada secuenciación paralela o de siguiente generación –*next generation sequencing* (NGS)–, que establece un salto de varios órdenes de magnitud en cuanto a la longitud de los fragmentos secuenciados y a la rapidez en su secuenciación [2-4]. Mediante la NGS es posible la secuenciación del genoma humano completo de un individuo en el mismo tiempo y coste económico que la secuenciación de dos o tres genes grandes con la técnica de Sanger [5].

La secuenciación del ADN ha producido un cambio radical en la manera de entender la genética, por la cual se ha pasado de estudiar la herencia basándose en patrones de transmisión o datos probabilísticas a conocer cuáles son las causas reales de esta herencia. Sin embargo, sus limitaciones tecnológicas han centrado los estudios genéticos a individuos con un fenotipo definido y enfermedades de herencia mendeliana producida por genes conocidos o a la búsqueda de genes causales en regiones del genoma previamente determinadas mediante análisis de ligamiento o mapeo por homocigosidad. Los estudios diagnósticos genéticos se realizan evaluando, de la forma más concienzuda posible, el fenotipo de un paciente y analizando mediante secuenciación por la técnica de Sanger el gen que se considera que puede estar afectado en función del estudio del fenotipo. Este método tiene una alta sensibilidad para detectar la mutación causante cuando existe una mutación en el gen señalado por el estudio previo del fenotipo, pero su rendimiento diagnóstico, es decir, el número de casos en los que está presente la mutación, es bajo, ya que en la mayor parte de los casos el estudio fenotípico es incompleto o insuficiente. Por el contrario, la NGS tiene menor sensibilidad para detectar mutaciones

**Figura.** Aspectos generales de la secuenciación. a) Estructura del gen *SNCA* (α-sinucleína) formado por seis exones; b) El método de Sanger secuencia generalmente un gen, exón a exón. Se muestra parte de la secuencia del exón 5 del gen *SNCA* por el método de Sanger; c) La secuenciación paralela o de siguiente generación (NGS) secuencia todo el genoma o el exoma. Se muestra la secuenciación exómica con NGS en la región del gen *SNCA,* en la que puede verse cómo la cobertura es por encima de 25 veces en la zona de los exones; d) Detalle de la NGS en la región mostrada en b, en la que se puede ver algún nucleótido en rojo (error de secuenciación) y cómo la cobertura de más de 30 secuencias permite eliminar estos errores.




cuando están presentes, es decir, puede presentar más falsos negativos, pero su rendimiento diagnóstico es mucho mayor, dado que se estudian de forma simultánea todos los genes del genoma [6]. Mediante esta técnica saltamos de examinar genes específicos relacionados con el fenotipo a estudiar genomas enteros de grupos humanos u otras especies. Esto está permitiendo conocer no sólo cómo es un genoma individual, sino también cómo cambia el genoma humano de persona a persona, cómo difieren los genomas entre diferentes grupos humanos, e incluso cómo difiere el genoma de un tumor respecto del genoma sano del huésped.

La capacidad de secuenciación que tiene la NGS es tal que teóricamente es posible secuenciar otros genomas presentes en el organismo de un individuo, como genomas tumorales o de microorganismos. Del mismo modo que se obtiene la secuencia el genoma mitocondrial en un sujeto al que se le realice un estudio de exoma completo, es posible detectar genomas virales, como los del virus de la inmunodeficiencia humana, hepatitis o cualquier otro virus presente en este individuo, simplemente añadiendo los genomas de estos virus al genoma de referencia utilizado [7]. Esto abre un campo inmenso de posibilidades diagnósticas, pero también plantea nuevos dilemas éticos no imaginados hasta el momento.

Recientemente, ha comenzado el empleo de las técnicas de secuenciación masiva para el diagnóstico de enfermedades mendelianas, bien secuenciando el genoma completo o sólo la región codificante (el exoma), que únicamente es un 1% del genoma, pero en el que está el 85% de las enfermedades hereditarias [8,9]. En el último año, se han publicado varios ejemplos en los que se ha empleado la NGS para estudiar enfermedades que pueden ser causadas por un alto número de genes, como la enfermedad de Charcot-Marie-Tooth (CMT) [10,11], y nosotros mismos la hemos empleado con éxito para el estudio de los genes en un caso de distrofia muscular de cinturas [12].

## Consideraciones técnicas

Aunque las plataformas difieren en sus configuraciones internas y en el tipo de reacciones químicas, el mecanismo de generación de la reacción es común y se basa en la secuenciación masiva de moléculas de ADN amplificadas de forma paralela. A través de ciclos de reacción en cadena de la polimerasa o de sucesivos ligamientos de oligonucleótidos, se obtiene la secuencia de cientos de millones de bases. Utilizando la tecnología clásica de Sanger, se pudo secuenciar un genoma humano en 13 años y con un coste estimado de $2{,}7 \times 10^9$ dólares [13]. Por comparación, en el año 2008, un genoma humano se secuenciaba en 5 meses y con un coste de $1{,}5 \times 10^6$ dólares, y en la actualidad se secuencia en semanas y el coste oscila entre $0{,}5\text{-}1 \times 10^5$ dólares. En este momento, existen cuatro plataformas comerciales de secuenciación y varios prototipos en experimentación, con una previsión de conseguir en poco tiempo una secuenciación genómica por 1.000 dólares [2].

Desde el punto de vista molecular, los dispositivos actuales para NGS utilizan unas rutinas previas muy similares. Si se quiere realizar una secuenciación genómica o exómica, se precisan aproximadamente 10 μg de ADN, que no debe estar fragmentado ni contaminado con sustancias orgánicas. Este ADN de alto peso molecular se somete a varios procesos antes de ser secuenciado, que incluyen la fragmentación en cadenas de ADN de tamaño corto, la captura, cuando se quiere secuenciar solamente las regiones del genoma que nos interesen,

como, por ejemplo, los exones, mediante array o en disolución, y su amplificación [14]. Una vez realizados estos procesos, se inicia la secuenciación, en la que se van a producir secuencias de tamaño corto (35-400 pares de bases, según el dispositivo utilizado) y, posteriormente, las secuencias obtenidas son alineadas a un genoma de referencia o, en el caso de los genomas *de novo,* ensambladas mediante regiones solapadas que compartan [15].

## Diagnóstico genético molecular mediante NGS

El desarrollo de la genética en años anteriores ha aumentado enormemente nuestros conocimientos de las enfermedades hereditarias mendelianas, lo que, a su vez, ha hecho mucho más complejo su estudio genético molecular. Ante un paciente con polineuropatía hereditaria sensitivomotora (CMT), la evaluación de la mutación causal es sencilla, si se limita al estudio de la duplicación en el gen *PMP22* y, en caso de ser negativa, al examen de otros genes involucrados (secuenciación de *PMP22*, *PMZ* y *Connexina32*). Cuando estos estudios sean negativos, si queremos un estudio genético completo que excluya todos los genes causantes de esta enfermedad conocidos hasta la fecha, debemos secuenciar más de 40 genes. Los trabajos de Lupski et al [10] y Montenegro et al [11] muestran cómo la secuenciación del exoma es una aproximación alternativa para la detección del gen causal en un individuo con CMT. En el trabajo de Lupski se estudió una familia con cuatro hermanos afectados de polineuropatía desmielinizante sensitivomotora y cuatro hermanos y padres sanos (CMT, o forma desmielinizante recesiva), en el cual se realizó una secuenciación completa del exoma del probando, detectándose una mutación en el gen *SH3TC2* (CMT4C). En el probando, la secuenciación del exoma encontró 1.165.204 mutaciones intragénicas, 54 de ellas cambios codificantes. Al examinar las variantes localizadas en los 40 genes potencialmente causales de CMT, dos de ellas, R954X, previamente descrita, e Y169H, una mutación nueva, se encontraban en el gen *SH3TC2,* explicando una transmisión recesiva. Este gen había sido descrito previamente como causante de CMT en familias de origen de Europa del este, turco o español [16,17].

El coste de la secuenciación del exoma completo es inferior al de secuenciar los más de 40 genes implicados en la CMT y tiene la ventaja de ser una técnica general, es decir, secuenciar el exoma es útil para cualquier enfermedad, mientras que secuenciar todos los genes involucrados en la CMT es una técnica de aplicación muy reducida, dado lo infrecuente de estos casos. Este mismo protocolo de estudio puede aplicarse a otras enfermedades hereditarias causadas por un alto número de genes, como miopatías, esclerosis lateral amiotrófica, enfermedades vasculares hereditarias, enfermedad de Parkinson y demencias (Tabla I) [18]. Además, esta técnica tiene la ventaja de poder detectar mutaciones causantes de otras enfermedades, no sólo de la enfermedad objeto del estudio, e incluso en varios genes causantes, lo que puede explicar diferencias en la penetrancia de algunas mutaciones. Es bien conocida la existencia de casos con CMT en individuos que presentan mutaciones en dos genes causales que cursan con una afectación mucho más grave [19,20]. Otras posibles indicaciones son su uso como técnica exploratoria en aquellos casos en los cuales hay una enfermedad grave y no existe un diagnóstico conocido, al ser una herramienta diagnóstica menos cruenta que biopsias u otras exploraciones, en especial en pacientes con deterioro cognitivo o motor progresivo. En estos casos, el estudio genómico pue-

**Tabla I.** Indicaciones de estudios de exoma.

| | |
|---|---|
| Indicaciones | Enfermedades hereditarias con amplia lista de genes candidatos (>3) |
| | Enfermedad de Charcot-Marie-Tooth (CMT) |
| | Miopatías |
| | Enfermedad de Parkinson familiar, distonías y otros movimientos anormales |
| | Epilepsia |
| | Leucodistrofias: CADASIL y otros |
| | Esclerosis lateral amiotrófica |
| | Demencias hereditarias |
| Otras indicaciones | Enfermedades progresivas de causa no aclarada |
| | Procesos largos y costosos |
| | Esclerosis múltiple |
| | Polineuropatía crónica inflamatoria desmielinizante |
| No indicado | Si existe un fuerte gen candidato (enfermedad de Huntington, CMT tipo 1) |
| | Enfermedades producidas por expansiones (por ejemplo, heredoataxias) |
| | Enfermedades producidas por reordenamientos del genoma (por ejemplo, distrofia facioescapulohumeral) |
| | ¿Variaciones en el número de copias? |

CADASIL: arteriopatía cerebral autosómica dominante con infartos subcorticales y leucoencefalopatía.

de detectar una enfermedad genética infrecuente. Finalmente, en procesos largos y costosos, como la esclerosis múltiple o la polineuropatía crónica inflamatoria desmielinizante, en la que a veces enfermedades hereditarias, por ejemplo, la arteriopatía cerebral autosómica dominante con infartos subcorticales y leucoencefalopatía, pueden imitar su clínica, el hallazgo de una causa genética evita al paciente tratamientos molestos y costosos.

No debe usarse la secuenciación exómica cuando exista un fuerte gen candidato (por ejemplo, enfermedad de Huntington) o ya se ha identificado la mutación en otros casos de la misma familia. Hay que tener en cuenta que la secuenciación exómica, no la genómica, detecta, sobre todo, mutaciones puntuales, es decir, mutaciones *missense*, *nonsense*, *indels* y mutaciones en el *splicing,* pero no sirve para detectar reordenamientos genómicos, como, por ejemplo, en la enfermedad de CMT por duplicación en el gen *PMP22* o en la miopatía facioescapulohumeral. Por último, una limitación a ambas técnicas son las alteraciones debidas a secuencias repetidas, dado que son técnicas basadas en generar secuencias cortas, por lo que algunas secuencias repetidas son de un tamaño mayor que los fragmentos secuenciados por NGS y, por lo tanto, no pueden ser alineadas correctamente.

La identificación rápida y completa de mutaciones causales y la posibilidad de examinar regiones intragénicas y promotoras no sólo va a permitir explicar con una profundidad impensable hasta ahora fenómenos como la diferente penetrancia de las enfermedades, sino también adscribir con exactitud cada uno de los síntomas a la causa real. En el síndrome de Miller (disóstosis acrofacial postaxial), trastorno autosómico recesivo caracterizado por anomalías esqueléticas en los huesos maxilares y, en ocasiones, en las extremidades, un subgrupo de pacientes presentaba también diarrea. El estudio de la búsqueda del gen causal de esta enfermedad se llevó a cabo secuenciando el exoma completo de dos hermanos y los padres, y adscribió la enfermedad al gen *DHODH.* Adicionalmente, se encontró en los pacientes con diarrea una mutación en el gen *DNAH5,* que previamente se había implicado en la discinesia ciliar primaria, un trastorno semejante a la fibrosis quística, por lo que la diarrea no es propia del síndrome de Miller, sino una superposición de otra enfermedad en este subgrupo [21].

Otra de las ventajas de la secuenciación genómica completa sobre los estudios genéticos clásicos es que no es necesario conocer la estructura hereditaria de la enfermedad para encontrar la mutación causal. Un ejemplo de esta utilidad es el estudio de Ng et al [22] en el síndrome de Freeman-Sheldon, trastorno con artrogriposis distal, anomalías faciales y sordera progresiva, en el que se secuenció el exoma de cuatro individuos con esta enfermedad no emparentados. El examen de aquellos genes que presentaban una mutación detectó que el único común a los cuatro era el gen *MYHY3,* un gen previamente descrito en este síndrome [22]. Este hallazgo no supuso, por tanto, encontrar un nuevo gen causal, pero el trabajo nos muestra cómo es posible encontrar genes causales sin tener información previa sobre el árbol genealógico, ligamiento o mecanismo patogénico de la enfermedad. Esta capacidad resulta especialmente útil cuando enfermedades hereditarias aparecen *de novo* al no existir en estos casos ningún otro familiar informativo. Utilizando información familiar, es posible encontrar nuevos genes con NGS examinando muy pocos casos, mientras que con los clásicos análisis de ligamiento es necesario recurrir a un alto número de individuos. En general, un análisis de ligamiento clásico precisa estudiar para trastornos autosómicos dominantes 20-30 sujetos entre afectados y no afectados, dependiendo de la estructura del árbol genealógico, siendo menor este número en el caso de enfermedades de transmisión recesiva. Por el contrario, secuenciando el genoma o el exoma completo, basta estudiar tres o cuatro casos para encontrar el gen causal. En el trabajo de Ng et al, se examinaron dos hermanos afectados, junto con sus padres, mediante secuenciación del exoma completo, encontrándose la variante causal [21]. Incluso en familias en las que el ligamiento clásico ha detectado una región candidata, hoy en día la secuenciación genómica o exómica completa es preferible al examen de la región candidata.

## Análisis de los datos

La NGS genera un enorme volumen de datos, lo que representa un reto mayor el análisis y la interpretación de los resultados que la obtención de la propia secuencia. La estrategia para examinar los resultados de la secuenciación genómica o exómica completa a fin de encontrar la mutación causante está todavía en un estadio primitivo, debido a su aparición tan reciente y a la escasa experiencia en el tema. Debido al enorme tamaño del genoma humano, 3.000 millones de pares de bases, y a la variabilidad de éste, el número de variantes que vamos a encontrar en un estudio genómico va a ser muy alto. El resultado obtenido en la NGS se compara con un genoma de referencia, y de esta comparación surgen

las posibles variantes genéticas, entre las que estarán las que causan la enfermedad. En nuestro estudio de exoma completo para evaluar una paciente con distrofia muscular de cinturas, encontramos 194.618 variantes –170.196 polimorfismos mononucleótidos (SNP), 8.482 polimorfismos multinucleótidos, 7.466 inserciones, 8.307 deleciones y 167 combinaciones mixtas–: 71.328 homocigotas y 123.290 heterocigotas [12]. Es preciso, por lo tanto, filtrar este resultado para poder conocer cuál es la mutación causante en el proceso que nos interesa. Para ello se efectúa un triple filtrado, primero se eliminan todas las variantes que se encuentran presentes en la población, generalmente SNP presentes en la 1000 Genomes database o en el dbSNP [23,24], puesto que cualquier variante causante debe ser de muy baja frecuencia. Después se consideran aquellos genes de interés, bien por coincidencia en otros familiares o bien por estar presentes en una región seleccionada por análisis de ligamiento, y finalmente se determina su causalidad según su función y nivel de conservación [25].

**Tabla II.** Diferentes procesos en los que se ha encontrado un gen asociado a una enfermedad neurológica mediante secuenciación genómica/exómica.

| | Enfermedad | Gen |
|---|---|---|
| Wang et al [32] | Ataxia espinocerebelosa | *TGM6* |
| Wang et al [33] | Discinesia paroxística cinesogénica | *PRRT2* |
| Logan et al [34] | Miopatía de inicio precoz | *MEGF10* |
| Aldahmesh et al [35] | Retraso mental y cuadriplejía espástica | *ELOVL4* |
| Martí-Massó et al [36] | Distonía de inicio precoz | *GCDH* |
| Doi et al [37] | Ataxia del adulto autosómica recesiva con retraso mental | *SYT14* |
| Murdock et al [38] | Polimicrogiria | *WDR62* |
| Weedon et al [39] | Charcot-Marie-Tooth autosómica dominante axonal | *DYNC1H1* |
| Noskova et al [40] | Ceroidolipofuscinosis del adulto | *DNAJC5* |
| Zimprich et al [41] | Enfermedad de Parkinson de inicio tardío | *VPS35* |
| Barak et al [42] | Malformación del desarrollo occipital | *LAMC3* |
| Klein et al [43] | Neuropatía sensitiva con demencia | *DNMT1* |
| Erlich et al [44] | Paraparesia espástica | *KIF1A* |
| Rafiq et al [45] | Retraso mental autosómico recesivo | *MAN1B1* |

## Filtrado por variantes raras

Las enfermedades hereditarias son de baja prevalencia y existen muy pocos individuos que compartan una misma mutación. En algunos casos, las mutaciones sólo existen en una misma familia afectada (mutaciones privadas). Por lo tanto, uno de los métodos para valorar si estas variantes son causantes de enfermedad es examinar si no existen como descritas en las bases de datos de polimorfismo más usuales, 1000 Genome database y dbSNP [23,24]. Un posible fallo de esta estrategia es que la información sobre el fenotipo en estas bases de datos es bastante incompleta, y alguno de los polimorfismos incluidos en ellas puede ser patogénico.

## Filtrado por función

El primer filtro que la mayoría de los investigadores utiliza para atribuir causalidad a una mutación es si esa variante tiene una repercusión funcional, es decir, si está en una región que codifica (mutaciones *missense, nonsense, Indels* y mutaciones en el aparato de *splicing*) o en otras regiones no codificantes del ARN mensajero. La principal razón de esto es que estas variantes suelen tener un efecto mayor que mutaciones en regiones no codificantes, y también porque es mucho más difícil conocer el efecto de mutaciones no codificantes y mutaciones sinónimas. En los estudios listados en la tabla II este tipo de filtrado fue eficaz, por cuanto en todos los casos se trató de una mutación en regiones codificantes. Sin embargo, esto puede no ser siempre el caso, pues existen ejemplos en clínica humana de patología por mutaciones en regiones intrónicas, regulatorias y mutaciones sinónimas [26,27].

## Filtrado por conservación

En algunas ocasiones, cuando existan varias mutaciones codificantes, puede ser difícil conocer cuál de todas ellas es la responsable del cuadro. En este caso es posible considerar el potencial efecto de la mutación en la estructura de la proteína y en su función. Pueden ayudar a valorar este efecto las puntuaciones de conservación, para lo cual existen herramientas de programas como SIFT, PolyPhen, CDPred, PhyloP y GERP, que generalmente asignan una mayor puntuación a la mutación en función de lo anteriormente señalado [28,29]. Sin embargo, aunque este dato puede ser informativo, debe utilizarse con cautela y en conjunción con otras estrategias de filtrado, nunca como un dato exclusivo.

## Aspectos éticos

El salto tan tremendo que la NGS ofrece en el estudio del ADN ha generado nuevos aspectos éticos y van a ser necesarias nuevas guías que aconsejen un uso racional de esta técnica en el campo clínico. Los principios éticos de los estudios médicos se basan en la tríada de respeto, beneficio y justicia para la persona según el informe Belmont (http://www.hhs.gov/ohrp/humansubjects/guidance/belmont.html). Las guías legales o éticas posteriores a este informe recogen estos principios mediante la obligación de la participación voluntaria, el consentimiento informado, el derecho a la privacidad, la confidencialidad, la minimalización del riesgo y la ausencia de discriminación. La adhesión a estos patrones tiene que ser monitorizada por los comités éticos de hospitales o agencias institucionales [30].

En general, los estudios genéticos se consideran algo más que una simple prueba diagnóstica, por su potencial repercusión familiar y social. En el caso de la NGS, son de aplicación las guías establecidas previamente para los estudios genéticos (Eurogenetest: www.eurogenetest.org), pero surgen nuevos aspectos derivados de la enorme cantidad de datos que se pueden obtener de estos estudios. La utilización de los de la NGS como una rutina de estudio clínico hace necesario familiarizarse con los aspectos éticos de su uso.

Los nuevos aspectos éticos a considerar por la aparición de la NGS son la obtención de resultados incidentales y la responsabilidad de buscar y comunicar o no estos resultados, el depósito de la información en bases de datos y la posible identificación de los individuos de estas bases de datos [31].

Es aconsejable aclarar, previamente a la realización del estudio, si se va a analizar sólo parte del estudio de secuenciación, por ejemplo, los genes para los cuales se indicó el estudio, o si, además, se examinarán aquellos genes que tienen especial repercusión, como pueden ser genes tumorales, y si este examen se reducirá sólo a aquellos genes sobre los cuales hay una posible intervención o incluirá genes de enfermedades sin capacidad de prevención o tratamiento.

Especialmente sensible es el estudio genómico de menores de edad o de pacientes con minusvalías psíquicas. En el caso de los menores de edad, se contempla la posibilidad de evitar el acceso a resultados de enfermedades de inicio en la edad adulta, pudiendo demorarse la comunicación de estos resultados hasta que el sujeto alcance la mayoría de edad. Además, hay que plantearse la posibilidad de que los familiares del paciente tengan que ser informados de resultados genéticos encontrados en el estudio, y de que el riesgo de algunas enfermedades puede cambiar a medida que tengamos nuevas informaciones sobre la función o causalidad de algunos genes. Por último, los resultados genéticos obtenidos por NGS generan un volumen de información muy elevado, y, además, la tendencia es a que esta información se deposite en bases de datos que ayuden a comprender mejor el efecto de variaciones presentes en el genoma. Esto debe tenerse en cuenta en el consentimiento informado, que debe recoger la forma de codificación, acceso a los datos, retirada de éstos y disponibilidad a suministrar estos resultados a otras bases de datos en caso de fusiones de éstas.

**Bibliografía**


1. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A 1977; 74: 5463-7.
2. Ansorge WJ. Next-generation DNA sequencing techniques. N Biotechnol 2009; 25: 195-203.
3. Diamandis EP. Next-generation sequencing: a new revolution in molecular diagnostics? Clin Chem 2009; 55: 2088-92.
4. Zhang J, Chiodini R, Badr A, Zhang G. The impact of next-generation sequencing on genomics. J Genet Genomics 2011; 38: 95-109.
5. Haas J, Katus HA, Meder B. Next-generation sequencing entering the clinical arena. Mol Cell Probes 2011; 25: 206-11.
6. Chan EY. Next-generation sequencing methods: impact of sequencing accuracy on SNP discovery. Methods Mol Biol 2009; 578: 95-111.
7. Bibby K, Viau E, Peccia J. Viral metagenome analysis to guide human pathogen monitoring in environmental samples. Lett Appl Microbiol 2011; 52: 386-92.
8. Bick D, Dimmock D. Whole exome and whole genome sequencing. Curr Opin Pediatr 2011; 23: 594-600.
9. Biesecker LG, Shianna KV, Mullikin JC. Exome sequencing: the expert view. Genome Biol 2011; 12: 128.
10. Lupski JR, Reid JG, Gonzaga-Jáuregui C, Rio DD, Chen DC, Nazareth L, et al. Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy. N Engl J Med 2010; 362: 1181-91.
11. Montenegro G, Powell E, Huang J, Speziani F, Edwards YJ, Beecham G, et al. Exome sequencing allows for rapid gene identification in a Charcot-Marie-Tooth family. Ann Neurol 2011; 69: 464-70.
12. Jiménez-Escrig A, Gobernado I, García-Villanueva M, Sánchez-Herranz A. Autosomal recessive Emery-Dreifuss muscular dystrophy caused by a novel mutation (R225Q) in the laminA/C gene identified by exome sequencing. **Muscle Nerve 2011. COMPLETAR!!!**
13. Stevens H. Dr. Sanger, meet Mr. Moore: next-generation sequencing is driving new questions and new modes of research. Bioessays 2012; 34: 103-5.
14. Bainbridge MN, Wang M, Wu Y, Newsham I, Muzny DM, Jefferies JL et al. Targeted enrichment beyond the consensus coding DNA sequence exome reveals exons with higher variant densities. Genome Biol 2011; 12: R68.
15. Cantacessi C, Jex AR, Hall RS, Young ND, Campbell BE, Joachim A, et al. A practical, bioinformatic workflow system for large data sets generated by next-generation sequencing. Nucleic Acids Res 2010; 38: e171.
16. Lupo V, Galindo MI, Martínez-Rubio D, Sevilla T, Vílchez JJ, Palau F, et al. Missense mutations in the SH3TC2 protein causing Charcot-Marie-Tooth disease type 4C affect its localization in the plasma membrane and endocytic pathway. Hum Mol Genet 2009; 18: 4603-14.

17. Lassuthova P, Mazanec R, Vondracek P, Siskova D, Haberlova J, Sabova J, et al. High frequency of SH3TC2 mutations in Czech HMSN I patients. Clin Genet 2011; Feb 3. [Epub ahead of print].
18. Day-Williams AG, Zeggini E. The effect of next-generation sequencing technology on complex trait research. Eur J Clin Invest 2011; 41: 561-7.
19. Al-Thihli K, Rudkin T, Carson N, Poulin C, Melancon S, Der Kaloustian VM. Compound heterozygous deletions of PMP22 causing severe Charcot-Marie-Tooth disease of the Dejerine-Sottas disease phenotype. Am J Med Genet A 2008; 146A: 2412-6.
20. Russo M, Laura M, Polke JM, Davis MB, Blake J, Brandner S, et al. Variable phenotypes are associated with PMP22 missense mutations. Neuromuscul Disord 2011; 21: 106-14.
21. Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, Dent KM, et al. Exome sequencing identifies the cause of a mendelian disorder. Nat Genet 2010; 42: 30-5.
22. Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C, et al. Targeted capture and massively parallel sequencing of 12 human exomes. Nature 2009; 461: 272-6.
23. Cochrane G, Karsch-Mizrachi I, Nakamura Y. The International Nucleotide Sequence Database Collaboration. Nucleic Acids Res 2011; 39: D15-8.
24. Mu XJ, Lu ZJ, Kong Y, Lam HY, Gerstein MB. Analysis of genomic variation in non-coding elements using population-scale sequencing data from the 1000 Genomes Project. Nucleic Acids Res 2011; 39: 7058-76.
25. Borsani G, Ballabio A, Banfi S. A practical guide to orient yourself in the labyrinth of genome databases. Hum Mol Genet 1998; 7: 1641-8.
26. Kurowska M, Szkowska-Golec A, Gruszka D, Marzec M, Szurman M, Szarejko I, et al. TILLING: a shortcut in functional genomics. J Appl Genet 2011; 52: 371-90.
27. Zhao Y, Clark WT, Mort M, Cooper DN, Radivojac P, Mooney SD. Prediction of functional regulatory SNPs in monogenic and complex disease. Hum Mutat 2011; 32: 1183-90.
28. Zou M, Baitei EY, Alzahrani AS, Parhar RS, Al-Mohanna FA, Meyer BF, et al. Mutation prediction by PolyPhen or functional assay, a detailed comparison of CYP27B1 missense mutations. Endocrine 2011; 40: 14-20.
29. Flanagan SE, Patch AM, Ellard S. Using SIFT and PolyPhen to predict loss-of-function and gain-of-function mutations. Genet Test Mol Biomarkers 2010; 14: 533-7.
30. Sijmons RH, Van Langen IM, Sijmons JG. A clinical perspective on ethical issues in genetic testing. Account Res 2011; 18: 148-62.
31. Kaye J, Boddington P, De Vries J, Hawkins N, Melham K. Ethical implications of the use of whole genome methods in medical research. Eur J Hum Genet 2010; 18: 398-403.
32. Wang JL, Yang X, Xia K, Hu ZM, Weng L, Jin X, et al. TGM6 identified as a novel causative gene of spinocerebellar ataxias using exome sequencing. Brain 2010; 133: 3510-8.
33. Wang JL, Cao L, Li XH, Hu ZM, Li JD, Zhang JG, et al. Identification of PRRT2 as the causative gene of paroxysmal kinesigenic dyskinesias. Brain 2011; 134: 3493-3501.
34. Tondeur S, Pangault C, Le CT, Lannay Y, Benmahdi R, Cubizolle A, et al. Expression map of the human exome in CD34+ cells and blood cells: increased alternative splicing in cell motility and immune response genes. PLoS One 2010; 5: e8990.
35. Aldahmesh MA, Mohamed JY, Alkuraya HS, Verma IC, Puri RD, Alaiya AA, et al. Recessive mutations in ELOVL4 cause ichthyosis, intellectual disability, and spastic quadriplegia. Am J Hum Genet 2011; 89: 745-50.
36. Martí-Massó JF, Ruiz-Martínez J, Makarov V, López de Munain A, Gorostidi A, Bergareche A, et al. Exome sequencing identifies GCDH (glutaryl-CoA dehydrogenase) mutations as a cause of a progressive form of early-onset generalized dystonia. Hum Genet 2012; 131: 435-42.
37. Doi H, Yoshida K, Yasuda T, Fukuda M, Fukuda Y, Morita H, et al. Exome sequencing reveals a homozygous SYT14 mutation in adult-onset, autosomal-recessive spinocerebellar ataxia with psychomotor retardation. Am J Hum Genet 2011; 89: 320-7.
38. Murdock DR, Clark GD, Bainbridge MN, Newsham I, Wu YQ, Muzny DM, et al. Whole-exome sequencing identifies compound heterozygous mutations in WDR62 in siblings with recurrent polymicrogyria. Am J Med Genet A 2011; 155A: 2071-7.
39. Weedon MN, Hastings R, Caswell R, Xie W, Paszkiewicz K, Antoniadi T, et al. Exome sequencing identifies a DYNC1H1 mutation in a large pedigree with dominant axonal Charcot-Marie-Tooth disease. Am J Hum Genet 2011; 89: 308-12.
40. Benítez BA, Alvarado D, Cai Y, Mayo K, Chakraverty S, Norton J, et al. Exome-Sequencing confirms DNAJC5 mutations as cause of adult neuronal ceroid-lipofuscinosis. PLoS One 2011; 6: e26741.
41. Zimprich A, Benet-Pagès A, Struhal W, Graf E, Eck SH, Offman MN, et al. A mutation in VPS35, encoding a subunit of the retromer complex, causes late-onset Parkinson disease. Am J Hum Genet 2011; 89: 168-75.
42. Barak T, Kwan KY, Louvi A, Demirbilek V, Saygi S, Tuysuz B, et al. Recessive LAMC3 mutations cause malformations of occipital cortical development. Nat Genet 2011; 43: 590-4.
43. Klein CJ, Botuyan MV, Wu Y, Ward CJ, Nicholson GA, Hammans S, et al. Mutations in DNMT1 cause hereditary sensory neuropathy with dementia and hearing loss. Nat Genet 2011; 43: 595-600.
44. Erlich Y, Edvardson S, Hodges E, Zenvirt S, Thekkat P, Shaag A, et al. Exome sequencing and disease-network analysis of a single family implicate a mutation in KIF1A in hereditary spastic paraparesis. Genome Res 2011; 21: 658-64.
45. Rafiq MA, Kuss AW, Puettmann L, Noor A, Ramiah A, Ali G, et al. Mutations in the alpha 1,2-mannosidase gene, MAN1B1, cause autosomal-recessive intellectual disability. Am J Hum Genet 2011; 89: 176-82.

## Title

**Summary.**

**Key words.**

# OTRA DOCUMENTACIÓN

# WHOLE EXOME SEQUENCING OF A PEDIGREE WITH AN AUTOSOMAL DOMINANT BIPOLAR DISORDER FOUND CAUSATIVE MUTATION IN PERIOD3-CIRCADIAN RHYTHM.

**Introduction.**

Bipolar disorder (BPD) was first described by Emil Kraepelin more than a century ago. Patients with BPD usually develop depression and manic episodes alternating with periods of normal mood.

BPD prevalence is 1% in common population, although some authors rise it up to 4-5% using attenuated phenotypes known as “bipolar spectrum”. The illness is characterized by high rates of recurrence, persistence of residual symptoms, cognitive impairment and worse quality of life, and means an important economic burden.

Over the last years we have been witnessing an important growth in the human genome knowledge and its importance in the development of diseases. Following the fast advances achieved at the beginning with the localization of the genes involved in diseases with mendelian heritability, scientific community crashed into complex heritability. Nowadays investigation in genetics of complex heritability is focused in two hypotheses:

- Common disease-common variants. Complex diseases are caused by the accumulated effect of multiple common genetic variants widespread in population, each one with a small effect.
- Common disease-multiple rare variants. Complex diseases are caused by uncommon genetic variants with high penetrance. Those variants can be single nucleotide variants (SNVs), copy number variants (CNVs) or insertion/deletions.

Sequencing is nowadays the best tool to test the second hypothesis. It makes possible to identify all the variations present in an individual genome and relate them to diseases with no need of prior hypothesis or big samples. The problem of this

technique has been its prohibitive cost, what has limited its use to small parts of the genome marked by linkage analysis, what introduces bias and makes it less efficient.

Even though current techniques of massive parallel or "new generation" sequencing make it faster and cheaper, whole genome sequencing has still a high cost. That is why some strategies have raised to minimize costs keeping efficiency. Ng and cols. published in 2009 the first study showing the usefulness of sequencing exome, about 30 megabases (1% of total genome). Given that the most part of disease-causing mutations known to date are in that coding part, it means an important economic and time saving approach.

There is nowadays a large number of articles that use whole exome sequencing to find mutations related to different diseases. It has been used in Psychiatry to study de novo mutations in Tourette's syndrome, schizophrenia and autism, being their conclusions waiting for replication.

Regarding BPD, it has not been found yet any gene related to this condition firmly. There are many explanations, such as, for example, the use of small and heterogeneous samples or the need for prior hypothesis that we lack of. Whole exome sequencing may overcome these setbacks. Assuming the hypothesis of common disease-multiple rare variants we sequenced the exome of three subjects of a family with a dominant autosomal BPD to find their causative mutation.

**Patients and methods.**

Sample

This study was done in a kindred whose pedigree is shown in figure 1. For exome study we chose subjects II, III and IV, dismissing subject I because of its genetic proximity to subject III.

All the subjects had been attended in different resources of Mental Health and diagnosed of BPD type I with DSM IV-TR criteria. We used the MINI INTERNATIONAL NEUROPSYCHIATRIC INTERVIEW, Spanish version 5.0, to confirm the diagnosis and look out for co-morbidities, followed by an interview carried out by a psychiatrist.

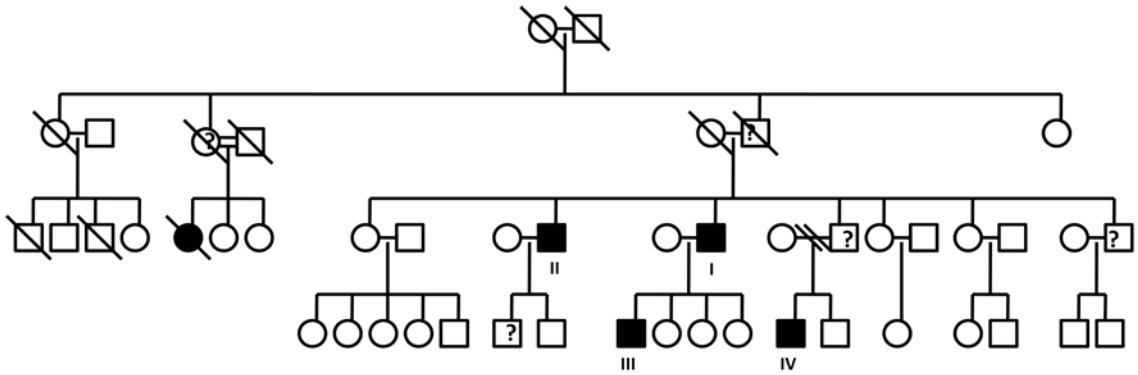



Figure 1. Pedigree. Question marks (?) tag probably affected subjects.

Informed consent for the collection of blood samples and medical history was obtained from participants. The study was approved by the Ethical Committee of the Hospital Ramón y Cajal.

Clinical information can be summarized as:

*Case II.* 60 year-old man, diagnosed of BPD at the age of 35 years during an admission to the hospital with a manic episode with psychotic symptoms. He has three other admissions with manic episodes. Throughout this time he has got also several depressive episodes. Currently, he is on treatment with lithium.

*Case III.* 34 year-old man, was diagnosed of BPD at the age of 27 years, when he was admitted to the hospital with a manic episode with psychotic symptoms. He does not have later admissions. He had also depressive episodes and is currently on valproic acid therapy.

*Case IV.* This man was diagnosed of BPD at the age of 22 years. He was admitted to the hospital with a manic episode with psychotic symptoms and treated with lithium. He has no further follow up and nowadays is without any pharmacologic treatment. There are some evidences of further hypomanic and depressive episodes that have not motivated the search for psychological or psychiatric evaluation.

Exome analysis.

For exome analysis, DNA was obtained from blood lymphocytes using QIAamp DNA Blood Maxi Kit (Quiagen, Valencia, California, USA). Whole exome sequencing was carried on in an external service (Otogenetics Inc.) using the SeqCap EZ Exome of

Nimblegen V2.0 and TruSeq of Illumina kits for the capture and enrichment of exome. The sequencing was done with HiSeq2000 platform (Illumina) at a coverage 50x. The raw sequences were first analyzed using a pipeline that included check data quality with FastQ, alignment to hg19 reference genome using BWA (*Burrows-Wheeler Aligner*) algorithm, and variant calling with the Genome Analysis Tool Kit (Broad Institute, Cambridge, Massachusetts, http://www.broadinstitute.org/gatk/). Variants on segmental duplication of not passing the GATK quality filtering were dismissed. To select meaningful variants we filtered the heterozygote variants against a population frequency of 1/10,000 in 1,000 genomes and 6,500 exomes databases, a functional AVSIFT criteria of less than 0.05 and Polyphen2 of 0.95-1 (very pathogenic mutations) and a later filtering that excluded variants in 25 house exomes. This was done with ANNOVAR (www.openbioinformatics.org/annovar/) and a confirmatory second analysis of variants was performed using KGGSeq tools (http://statgenpro.psychiatry.hku.hk/limx/kggseq/). Mutations were confirmed later by Sanger sequencing.

**Results:**

GATK-ANNOVAR

After the variant calling we localized ~23,000 mutations in each subject that were reduced to 173-190 after applying the heterozygous transmission model and frequency and functional filtration. From these, only 60 mutations were shared from the 3 cases, and after the manual filtering this figure was reduced to only two.

The first one was a c.A347G missense mutation in the exon 3 of PER3 gene that generates a p.E116G change. The second one was a c.861delC deletion of USP29 gene that generates a frame-shift.

KGGSeq analysis.

With this approach we obtained ten mutation that satisfied quality requirements, located within the following genes: PER3, RCC1, KIT, TMEM155, ANKRD31, AK8, THBS1, PAPD5, OR7C1 y USP29.

Mutations in RCC1, AK8, THBS1 and KIT where discarded for their lesser probability of being pathogenic based on prediction scores. Mutation in PAPD5 is located in a splicing

area, affecting only one of the translated protein isoforms. That makes it less probable to be the causing mutation. Finally, OR7C1 codifies for an olfactory receptor, molecule without a very much compatible function with BPD.

So, with this new analysis we confirmed the two previous mutations and added two more: a c.C100T mutation in exon 5 of TMEM155 gene, that produces a p.Q34X change, generating a stopgain, and a c.755_757 non-frameshift deletion of three nucleotides in exon 7 of ANKRD31 gene.

**Discussion.**

PERIOD3

In this kindred, the most likely causative mutation of the autosomal dominant BPD is the one located in PERIOD3 gene (PER3), that codifies for PERIOD3 protein, which plays part in the regulation of circadian rhythms. Those rhythms include sleep-wake pattern, variations in blood pressure and body temperature and bowel movements. It also determines variations in physical activity, cognitive processes or feeding behavior. The resultant protein of PER· c.A347G mutation has a glutamic acid replaced by glycine in position 116. Evidence supporting the importance of this mutation is that this base is not only conserved between species but also in the rest of PER proteins present in humans. Aligning the 3 PERIOD proteins according to Uniprot (http://www.uniprot.org) we can observe the conservation of the base.

```
95   EFFQILSQNG--APQADVSMYSLEELATIASEHTSKNTDTFVAVFSFLSG 142 P56645  PER3_HUMAN
180  EYYQQWSLEEGEPCSMDMSTYTLEELEHITSEYTLQNQDTFSVAVSFLTG 229 O15534 PER1_HUMAN
153  EYYQLLMSSEGHPCGADVPSYTVEEMESVTSEHIVKNADMFAVAVSLVSG 202 O15055 PER2_HUMAN
```

The circadian clock in mammals is located in the suprachiasmatic nucleus of hypothalamus and it's regulated by light. It regulates the secretion of some hormones, such as melatonin or cortisol. In absence of light the internal clock keeps the described patterns working in an autonomous way. It is based in a molecular level on interrelated loops regulated by feed-back mechanisms. It starts with the synthesis of CLOCK (CLK) and BMAL1 (or ARNTL) proteins. They penetrate in the nucleus forming a CLK-BMAL1

dimer. NPAS2 (*neuronal PAS domain-containing protein 2*) is a CLK analog and is able to develop the same function. The dimers interact with the E-boxes of CRY (Cryptochrome) and PERIOD 1, 2 and 3 genes. The new synthesized proteins go out to cytoplasm and form CRY-PER2 dimers and CRY-PER1-PER3 trimers. Those dimers and trimers go into the nucleus and inhibit CLK and BMAL1 transcription what, in turn, inhibit CRY and PER transcription. The messenger RNA peaks of CLK and BMAL1 are 12 hours dephased in relation to CRY and PER peaks, what makes the complete cycle last for 24 hours.

As an extra control mechanism, CLK-BMAL1 dimer activates transcription of RORA and Rev-ErbA genes (*retinoic acid-related orphan nuclear receptors*). Those molecules interact with their nuclear receptor situated in the promoter of BMAL1 gene, regulating its transcription, either activating (RORA) or inhibiting it (Rev-ErbA). Finally, CLK-BMAL1 activates also DEC transcription, with an unclear function. Phosphorilation processes through kinases (casein kinases, GSK3beta) determinate the in-out movements of molecules from the nucleus and also its degradation in proteasome.

Variations in PER3 gene have been related to sleep phase disorders, higher sensibility to sleep deprivation, variability in autonomous nervous system activity and other measures in relation with circadian rhythms.

Knockout mouse to this gene (Per3-/-) shows an alteration of phase in some tissues and a different pattern of activity in relation with light-dark cycles compared to the wild type.

One of the current theories of BPD physiopathology postulates that it's caused by an alteration in circadian rhythms. It's been found relation between variations in some genes involved in its regulation and mood disorders. So, PER3 gene has been previously pointed as a candidate gene. However, evidence about its connection with BPD was limited to non replicated studies.

USP29

The second putative gene in this study was USP29, were there is a frame-shift mutation that changes the amino acid sequence from position 288.

USP29 gene codifies an ubiquitin carboxi-terminal hydrolase type 2 that contains two protease domains for specific ubiquitin processing. Ubiquitin processes are highly extended in different cells, marking useless molecules to allow them to be degraded and removed by proteasome.

USP29 breaks the bonds between ubiquitin and p53 protein, stabilizing it and inhibiting its degradation, starting the apoptosis process of the cell. USP29 transcription is promoted by JTV1-FBP association, activated by oxidative stress related to, for example, dopamine metabolism.

TMEM155 and ANKRD31

TMEM155 and ANKRD31 are highly variable genes. If a deletion or any other changes in these genes that produce a non-functional protein in heterozygosis could cause an autosomic dominant BPD, then we would observe a very much higher prevalence of this type of disorder in population. Also, these types of mutations (with loss of function) very rarely produce disorders in heterozygosis. The same can be said about the USP29 mutation, but in this case, USP29 gene is imprinted, and only paternal copy is expressed. That could explain the development of the pathologic phenotype in spite of being mutated in heterocygosity.

Even though we cannot discard the rest, the mutation in PERIOD3 seems to be the most probable to be related to the autosomic dominant BPD in this family.

**Conclusion.**

Whole exome sequencing had proved itself a useful tool in discovering genes related to diseases with mendelian heritability. However there is still little evidence of its utility in complex diseases.

The existence of pedigrees with complex disorders like BPD with an apparent mendelian heritability, and the discovery of a likely responsible gene through this technique, supports the theory of common diseases-multiple uncommon variants.

That is, there is a part of the heritability of these diseases that can be explained by very infrequent mutations of high penetrance that can be inherited or de novo.

The discovery of these familiar variants may have little diagnostic utility in the short-term, but it is important for the development of pathway hypothesis, allowing the development of biological models and facilitating the finding of potential treatments.

The limitation of the possible causative mutations to two is an extraordinary result, given the number of mutations present in the exome of any subject. Although the functional pathway supports the mutation in PERIOD3 gene, this result, however, should be confirmed by exploring new affected cases in this kindred or in other familial cases of BPD.

The fact that one of the found mutations is located in PERIOD3 gene, being a candidate gene to BPD in some studies, supports the theories that relate this disease with an alteration in circadian rhythms.

As far as we know, this is the first time that whole exome sequencing in used in the study of BPD, and one of the few carried out in psychiatric pathologies. Its results, if replicated, can encourage us to think that we have reached the end of the empty-of-results search that has been investigation in genetics of psychiatric diseases till now. Nevertheless we don't forget that this result, even stimulating, may not be replicated for being an exceptional case.

Dra. **ITZIAR DE PABLO LÓPEZ DE ABECHUCO,** Secretaria del Comité Etico de Investigación Clínica del Hospital Ramón y Cajal

**CERTIFICA**

Que el Comité Etico de Investigación Clínica, ha evaluado el **PROYECTO DE INVESTIGACIÓN:**

Titulo:

**Estudio de trastornos neuropsiquiátricos mediante secuenciación de exoma completo**.

Investigador Principal: **Dra. Isabel Gobernado Ferrando**

Servicio: **Psiquiatría**

Y ha decidido su **APROBACIÓN.**

Lo que firmo en Madrid a 16 de abril de 2012

Hospital Universitario Ramón y Cajal
SaludMadrid
COMITÉ ETICO DE INVESTIGACIÓN CLÍNICA

Fdo.: Dra. Itziar de Pablo López de Abechuco
Secretaria del CEIC