Escola Tècnica Superior d'Enginyeria ETSE-UV

MÁSTER UNIVERSITARIO EN BIOINFORMÁTICA

TRABAJO DE FIN DE MÁSTER

ANÁLISIS DE EXOMAS PARA EL DIAGNÓSTICO DE NEUROPATÍAS PERIFÉRICAS HEREDITARIAS

AUTOR/A:
ANA SÁNCHEZ MONTEAGUDO

DIRECTORES:
CARMEN ESPINÓS ARMERO
FRANCISCO GARCÍA GARCÍA

TUTOR:
MIGUEL ÁNGEL GARCÍA PÉREZ

JULIO 2016

VNIVERSITAT DE VALÈNCIA

Escola Tècnica Superior d'Enginyeria ETSE-UV

MÁSTER UNIVERSITARIO EN BIOINFORMÁTICA

TRABAJO DE FIN DE MÁSTER

ANÁLISIS DE EXOMAS PARA EL DIAGNÓSTICO DE NEUROPATÍAS PERIFÉRICAS HEREDITARIAS

AUTOR/A:
ANA SÁNCHEZ MONTEAGUDO

DIRECTORES:
CARMEN ESPINÓS ARMERO
FRANCISCO GARCÍA GARCÍA

TUTOR:
MIGUEL ÁNGEL GARCÍA PÉREZ

TRIBUNAL:

PRESIDENTE/A:

VOCAL 1:

VOCAL 2:

FECHA DE DEFENSA:

CALIFICACIÓN:

RESUMEN

Las neuropatías periféricas hereditarias abarcan un gran grupo de formas clínicas, con un amplio espectro de fenotipos y una alta heterogeneidad genética. Destaca en este grupo la enfermedad de Charcot-Marie-Tooth (CMT) o neuropatía hereditaria sensitivo-motora, trastorno neurológico hereditario más frecuente que se relaciona con formas de neuropatía hereditaria motora distal (dHMN), que sólo implican afectación motora. En ocasiones, los pacientes presentan además signos clínicos propios de esclerosis lateral amiotrófica (ALS) o atrofia muscular espinal (SMA), lo que dificulta todavía más el diagnóstico molecular. En los últimos años, la secuenciación de exomas o WES (*whole-exome sequencing*) se ha convertido en una herramienta coste-efectiva para examinar simultáneamente todos los genes implicados en neuropatías con fines diagnósticos.

En este trabajo se ha llevado a cabo la secuenciación de exoma, mediante la tecnología Illumina, en ocho casos no relacionados con un diagnóstico clínico inicial de CMT o dHMN. Los datos obtenidos en la secuenciación fueron procesados por los *pipelines* establecidos en el Centro Nacional de Análisis Genómico (CNAG) y en la Plataforma de Bioinformática para las Enfermedades Raras (BiER), y se ha realizado un estudio comparativo de la cobertura y las variantes obtenidas. Se han analizado las variantes obtenidas a partir de los dos procesados en una selección de genes asociados a neuropatías. En dos casos, han sido identificadas y validadas mutaciones previamente descritas en los genes *FIG4* (p.I41T/c.446+5G>C) y *SOD1* (p.E22G), y en otros dos, cambios noveles en los genes *BICD2* (p.S5W) y *DAO* (p.W320R). Las variantes p.I41T y p.E22G en los genes *FIG4* y *SOD1*, respectivamente, han sido previamente descritas como patológicas. En el resto, los estudios *in silico* indican que probablemente se trate de mutaciones patológicas. Además, el análisis funcional basado en la utilización de construcciones de minigenes ha mostrado que la variante c.446+5G>C identificada en *FIG4* alteraría el correcto procesado del mRNA. En cuatro casos no se ha logrado identificar y validar posibles variantes candidatas a ser causantes de la enfermedad con esta estrategia. Los datos obtenidos en esta investigación son útiles para la búsqueda de nuevos genes implicados en neuropatías, lo que permitirá avanzar en el conocimiento de las bases moleculares de estas enfermedades.

Palabras clave: secuenciación de exoma, neuropatías periféricas hereditarias, diagnóstico genético, minigenes.

Financiación: Fundació Per Amor a l'Art (para A.S.M.), PI12/0453 (ISCIII).

ABSTRACT

Inherited peripheral neuropathies are a group of disorders with genetic and clinical heterogeneity. Charcot-Marie-Tooth disease (CMT), also known as hereditary sensory and motor neuropathy, is closely related to distal hereditary motor neuropathies (dHMN), which have only motor involvement. Patients occasionally show additional disorders-related signs of amyotrophic lateral sclerosis (ALS) or spinal muscular atrophy (SMA), and it complicates genetic diagnosis. Whole-exome sequencing (WES) has become a cost-effective method to examine all known genes implicated in these neuropathies for diagnostic purposes.

In this work, WES was applied using Illumina technology to eight unrelated cases with predominant signs of CMT or dHMN. Sequencing reads were processed using pipelines developed by Centro Nacional de Análisis Genómico (CNAG) and the Plataforma de Bioinformática para las Enfermedades Raras (BiER), and a comparative study of coverage and variants obtained was performed. Data from both pipelines were examined for variants in genes previously associated with neuropathies. In two cases, known variants were identified and validated in *FIG4* (p.I41T/c.446+5G>C) and *SOD1* (p.E22G), as well as novel variants in other two cases in *BICD2* (p.S5W) and *DAO* (p.W320R). *FIG4* p.I41T and *SOD1* p.E22G were previously reported as pathogenic. For the remaining variants, in silico predictions suggested that they are probably damaging. Functional minigene analysis of *FIG4* c.446+5G>C showed that it alters mRNA splicing. In four cases, no candidate variants in known neuropathy-related genes were identified and validated following this strategy. This study has provided useful data for further studies to identify novel neuropathy genes and to improve the understanding of molecular bases underlying these diseases.

Keywords: whole-exome sequencing, inherited peripheral neuropathies, genetic diagnosis, minigenes.

Funding: Fundació Per Amor a l'Art (to ASM), PI12/0453 (ISCIII).

ÍNDICE

ABREVIATURAS

5'UTR: *5-prime untranslated region*

ALS: esclerosis lateral amiotrófica (*amyotrophic lateral sclerosis*)

ARCA: ataxia cerebelar autosómica recesiva (*autosomic recessive cerebellar ataxia*)

BAM: *binary alignment format*

BWA: *Burrows-Wheeler aligner*

CMT: neuropatía de Charcot-Marie-Tooth

CNV: variante del número de copias (*copy-number variation*)

dHMN: neuropatía hereditaria motora o atrofia espinal distal (*distal hereditary motor neuropathy*)

DNA: Ácido desoxirribonucleico (*deoxyribonucleic acid*)

GATK: *Genome Analysis Toolkit*

GRCh37: Genoma Humano de Referencia versión 37 (*Human Genome Reference Consortium version 37*)

HGMD: *Human Genome Mutation Database*

HSMN: neuropatía hereditaria sensitivo-motora (*hereditary motor and sensitive neuropathy*)

HSAN: neuropatía hereditaria sensitiva y autonómica (*hereditary sensitive and autonomic neuropathy*)

HSP: paraparesia espástica hereditaria (*hereditary spastic paraparesis*)

Indels: inserciones/deleciones

MAPQ: calidad de mapeo en escala Phred (*Phred-scaled mapping quality*)

mRNA: Ácido ribonucleico mensajero (*messenger ribonucleic acid*)

OMIM: *Online Mendelian Inheritance in Man*

RT-PCR*:* Reacción en cadena de la polimerasa con transcriptasa inversa *(Reverse transcription polymerase chain reaction)*

SCA: ataxia espinocerebelosa (*spinocerebellar ataxia*)

SMA: atrofia muscular espinal (*spinal muscular atrophy*)

SNV: variantes de una sola base (*single nucleotide variants*)

VCF: *variant calling format*

WES: secuenciación de exoma (*whole-exome sequencing*)

WGS: secuenciación de genoma (*whole-genome sequencing*)

INTRODUCCIÓN

1. Neuropatías periféricas hereditarias

El sistema nervioso periférico, constituido por nervios motores y sensitivos, conecta y permite el intercambio de información entre el sistema nervioso central (cerebro y médula espinal) y el resto de órganos y extremidades. Las neuropatías periféricas hereditarias comprenden un amplio grupo de trastornos que producen una degeneración progresiva de los nervios periféricos. Se caracterizan por presentar una alta heterogeneidad clínico-genética, y constituyen el grupo de enfermedades neurológicas hereditarias más frecuentes (Baets *et al*., 2014).

Se distinguen distintos tipos de trastornos en función de la fisiología de los nervios afectados: neuropatías hereditarias sensitivo-motoras (HSMN) o enfermedad de Charcot-Marie-Tooth (CMT); neuropatía hereditaria motora distal (dHMN) cuando sólo hay afectación de nervios motores; y neuropatía hereditaria sensitiva y autonómica (HSAN) si sólo los nervios sensitivos están alterados. Los fenotipos clínicos de estos tres subgrupos solapan, de forma que mutaciones en un único gen pueden dar lugar a CMT, dHMN o HSAN (Weis y Senderek 2014; Echaniz-Laguna 2015).

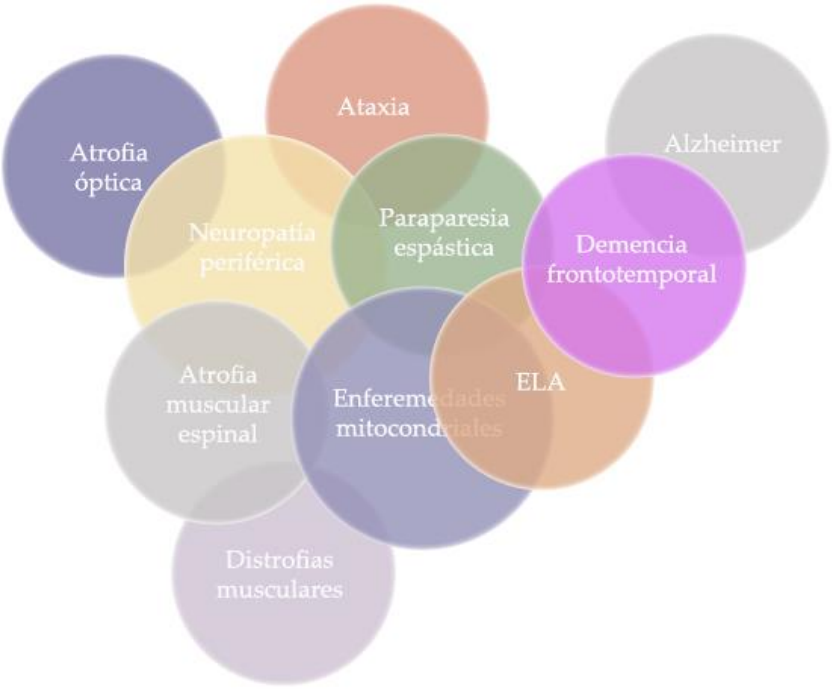

Figura I1: Solapamiento fenotípico y genético de diferentes grupos de enfermedades neurodegenerativas. (ELA: esclerosis lateral amiotrófica)

1.1 Neuropatía de Charcot-Marie-Tooth

La neuropatía de Charcot-Marie-Tooth (CMT), considerada una enfermedad rara, es el trastorno neurológico hereditario más frecuente, con una prevalencia de 28 casos por cada 100.000 habitantes (Combarros *et al.*, 1987; Skre, 1974). Fue descrita por primera vez en 1886 por Howard Henry Tooth (Inglaterra) y Jean-Marie Charcot y Pierre Marie (Francia) como "atrofia muscular peroneal". La clínica de la enfermedad se caracteriza por una debilidad muscular distal progresiva y atrofia, con pérdida de sensibilidad asociada, comenzando en las extremidades inferiores y alcanzando finalmente las superiores.

Siguiendo criterios electrofisiológicos e histopatológicos, las formas de CMT se clasifican en desmielinizantes (CMT1) o axonales (CMT2). Desde el punto de vista genético, las distintas formas de CMT se clasifican según el patrón de herencia mendeliano que sigan, ya sea autosómico dominante, recesivo o ligado a X, y en función de los genes o *loci* causantes de la enfermedad. Además, la genética de esta enfermedad resulta más compleja, dado que mutaciones patológicas en un mismo gen pueden ser transmitidas de forma recesiva o dominante.

Las distintas mutaciones causales de CMT tienen lugar en genes implicados en la biología y desarrollo de nervios periféricos, que codifican para proteínas con distintas localizaciones entre las que se incluyen mielina compactada y no compactada, células de Schwann y axones, y que están involucradas en diversas funciones, desde la compactación y mantenimiento de la mielina hasta la formación del citoesqueleto, el transporte axonal, la transducción de señales, el metabolismo mitocondrial o la reparación del DNA (Rossor *et al.* 2013).

El estudio de las bases genéticas de CMT comenzó en 1991 con la identificación de la duplicación en tándem en el cromosoma 17p11.2-12, región en la que se localiza el gen *PMP22,* como causante del subtipo CMT1A. Desde la publicación del primer borrador del Genoma Humano en 2001, el desarrollo de las técnicas de secuenciación masiva ha acelerado el descubrimiento de nuevos genes implicados en CMT. Actualmente, más de 70 genes han sido asociados a CMT, siendo aún desconocidos muchos de ellos, y el número de nuevas formas poco frecuentes causadas por mutaciones *de novo* en casos esporádicos no cesa de aumentar (Baets *et al.* 2014; Timmerman *et al.* 2014)

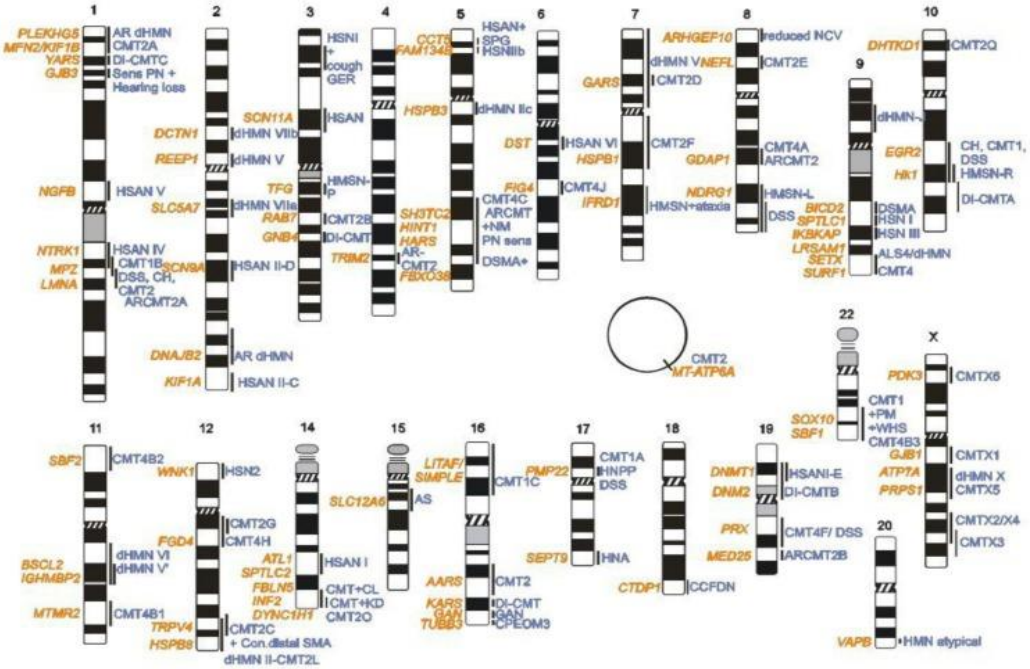

Figura I2: Genes y *loci* implicados en la enfermedad de Charcot-Marie-Tooth y otras neuropatías periféricas hereditarias relacionadas. En naranja se indica el nombre de los genes y su correspondiente *locus* cromosómico en barras verticales. El fenotipo resultante para cada gen se indica en azul (adaptado de Timmerman *et al*., 2014).

1.2 Neuropatías hereditarias motoras

Los casos de neuropatía hereditaria motora distal (dHMN) comprenden formas con degeneración de neuronas motoras de la médula espinal, en las que los pacientes no presentan afectación sensitiva. Son descritas como formas "espinales" de CMT, en concreto solapan clínica y genéticamente con formas de CMT2 (Mathis *et al.*, 2015). Desde 2001, se han identificado mutaciones en al menos 19 genes asociados a dHMN, y 8 de ellas causan también CMT2. Muchos de estos genes codifican para proteínas implicadas en tráfico de vesículas, transporte axonal, procesamiento de RNA, síntesis de proteínas, respuesta a estrés oxidativo y apoptosis (Rossor *et al.* 2012; Farrar y Kiernan 2014).

Estudios de correlación genotipo-fenotipo en amplias cohortes de familias y casos esporádicos han resultado en un bajo porcentaje de mutaciones identificadas en los genes asociados a dHMN, lo que sugiere que deben existir otros genes aún no descritos implicados en la fisiopatología de las neuropatías motoras (Dierick *et al.* 2008; Rossor *et*

al. 2012). Otros trastornos en los que también predomina la afectación de nervios motores son la atrofia muscular espinal (SMA) o la esclerosis lateral amiotrófica (ALS).

Las formas más comunes de SMA, que se manifiestan en la infancia, se refieren a alteraciones en el gen *SMN1*, ya sean mutaciones puntuales o deleciones, que resultan en una insuficiencia de la proteína para la que codifica en las neuronas motoras. La aplicación de técnicas de secuenciación masiva para el diagnóstico de esta enfermedad ha permitido ampliar las bases genéticas de esta enfermedad, habiéndose identificado mutaciones que sigue patrones de herencia tanto autosómico dominantes o recesivos y ligado a X en al menos 13 genes que han permitido explicar casos de SMA tanto en la infancia como en edad adulta que no se deben a alteraciones en *SMN1*. Mutaciones en algunos de ellos también han sido asociadas a formas de CMT (Farrar y Kiernan 2014).

ALS es la enfermedad de la motoneurona más común (Renton *et al.* 2013). Este trastorno afecta tanto a las neuronas motoras superiores e inferiores de la médula espinal, a diferencia de la SMA o la dHMN en las que únicamente las neuronas motoras inferiores se encuentran afectadas. Cursa con una parálisis progresiva que avanza rápidamente desembocando finalmente en una muerte por fallo respiratorio, típicamente 2 o 3 años después del debut de los síntomas.

Se estima que el 60% de casos familiares y un 10% de casos esporádicos de ALS son de origen genético (Renton *et al.*, 2013), habiéndose descrito todos los patrones mendelianos de herencia posibles. Así, en los casos familiares se asume un origen genético de la enfermedad, pero en los casos esporádicos se desconoce si realmente se deben a causas genéticas o a otros factores no heredables. Es de especial interés que muchos de los casos descritos como esporádicos realmente son casos familiares en los que, debido a la penetrancia incompleta y/o expresividad variable de la mutación, algunos de sus miembros son portadores asintomáticos de la enfermedad (Andersen y Al-Chalabi 2011). Hasta ahora, se han identificado mutaciones causantes de ALS en más de 25 genes, algunos de ellos mediante técnicas de análisis de ligamiento y clonación posicional, y el resto gracias a la aplicación de técnicas de secuenciación masiva, que además ha ayudado a resolver el diagnóstico genético de muchos de los casos esporádicos (Marangi y Traynor 2015).

1.3 Otros trastornos del sistema nervioso relacionados con neuropatías

Existen otros trastornos del sistema nervioso en los que la neuropatía no es la causa principal y forma parte de un amplio espectro de síntomas, resultando en un cuadro clínico más complejo. Es el caso de las paraparesias espásticas hereditarias (HSP) y las ataxias hereditarias, que afectan principalmente al sistema nervioso central.

Las HSPs son trastornos degenerativos de la médula espinal que se caracterizan clínicamente por la presencia de espasticidad y debilidad progresiva de los miembros inferiores. Existen formas "puras" de HSP, que comienzan en la infancia y el signo predominante es la espasticidad, que se distinguen de las formas "complicadas" en las que existe mayor implicación de otras partes del sistema nervioso como el cerebro, cerebelo, circuitos extrapiramidales, nervios periféricos y craneales. Esto provoca síntomas adicionales como retraso mental, distonía, parkinsonismo, sordera o atrofia óptica, entre otros (Schneider y Brás, 2015).

Al igual que en CMT, en las HSPs se observa una notable heterogeneidad genética habiéndose descrito mutaciones causales en más de 60 genes que siguen todos los patrones de herencia mendelianos posibles, implicados en procesos celulares similares a los descritos en CMT (Schneider y Brás, 2015). Como se ha indicado anteriormente, la neuropatía puede ser un síntoma adicional en este trastorno, por ello mutaciones en diversos genes han sido asociadas a ambos trastornos, al haber sido identificadas en pacientes que presentan cuadros clínicos difícilmente distinguibles entre HSP y CMT, como por ejemplo sucede con los genes *BSCL2* y *KIF5A* (Timmerman *et al.*, 2013). El descubrimiento de nuevos genes implicados en CMT y HSP aumentará el conocimiento de la fisiopatología de ambos trastornos, y con ello mejorarán las correlaciones genotipo-fenotipo.

Las ataxias hereditarias comprenden una serie de manifestaciones neurológicas que implican disfunción en aquellas partes del sistema nervioso que coordinan el movimiento y que conllevan degeneración de distintas partes del cerebro. Por ello, en los pacientes se observa descoordinación en extremidades inferiores y superiores o incluso en los movimientos oculares (Coutelier *et al*., 2015). Las ataxias son clínicamente heterogéneas debido a que frecuentemente se encuentran implicadas otras partes no cerebrales del

sistema nervioso, ejemplo de ello es el hecho de que la neuropatía periférica está presente en muchos casos (Linnemann et al., 2015). Genéticamente, se han descrito todos los modos mendelianos de herencia posibles. Las formas con herencia autosómica dominante, denominadas ataxias espinocerebelosas (SCA) se deben principalmente a expansiones del motivo CAG, así como otros tipos de expansiones en regiones no codificantes. En las formas de herencia autosómica recesiva (ARCA), existe una mayor variabilidad clínica y genética, con diversos genes implicados en fenotipos complejos (Coutelier *et al*. 2015). Se han identificado mutaciones causales en *SETX* en pacientes que presentan ataxia además de signos de neuropatía periférica (Nanetti *et al.*, 2013), así como en genes principalmente asociados a CMT como *GJB1* (Liu *et al.*, 2015).

2. Aplicaciones de la secuenciación masiva para el diagnóstico genético y su contribución a la investigación de enfermedades raras

Las enfermedades raras, minoritarias o poco frecuentes son aquellas cuya prevalencia en la población es inferior a 1/2000 habitantes en Europa (Boycott *et al.*, 2013). La identificación de genes responsables de estas enfermedades supone una importante mejora en el diagnóstico de los pacientes y el manejo de la enfermedad así como en el asesoramiento genético a familiares. El conocimiento de las bases genéticas permite esclarecer los mecanismos fisiopatológicos de la enfermedad y es un punto de partida para el desarrollo de nuevas terapias.

Tradicionalmente, la caracterización de nuevos genes implicados en enfermedades se ha llevado a cabo mediante secuenciación por Sanger de genes candidatos, seleccionados en base a estudios complementarios de cartografiado, análisis de ligamiento y clonación posicional (Gilissen *et al.*, 2012). Dichas estrategias se aplican en familias con varios individuos afectados o grupos de pacientes con un cuadro clínico similar con el fin de acotar el *locus* asociado con la enfermedad. Sin embargo, si la región identificada es grande y contiene muchos genes, identificar la mutación clínica responsable de la enfermedad secuenciando gen a gen por Sanger es una tarea tediosa y costosa, tanto económicamente como en cuanto a tiempo requerido. El cartografiado genómico no resulta exitoso cuando se aplica en familias pequeñas o en una cohorte de casos esporádicos sin una historia familiar definida (Ku *et al.*, 2011). Todo ello hace evidente la

necesidad de diseñar nuevas aproximaciones que puedan ser aplicadas con éxito en pequeños núcleos familiares, e incluso sobre un único paciente, y que no consuman tanto tiempo y recursos.

En los últimos años, el número de investigaciones en marcha para la identificación de genes causantes de patologías poco frecuentes ha aumentado gracias al desarrollo de las tecnologías de secuenciación masiva o de nueva generación, comúnmente conocidas por su acrónimo en inglés NGS (*next-generation sequencing*). Las mejoras técnicas conseguidas y su abaratamiento han permitido la aplicación de la secuenciación masiva en el ámbito clínico, y han resultado particularmente ventajosas para el diagnóstico de trastornos con una alta heterogeneidad clínica y genética al permitir el estudio simultáneo de un número considerable de genes (Liu *et al.*, 2015). Por ello, muchas enfermedades como las neuropatías periféricas, en las que el número de genes continúa aumentando y es difícil disponer de muestras de familiares habiendo de ser estudiados como casos esporádicos, son ideales para ser estudiadas mediante NGS.

Existen tres estrategias basadas en NGS para el diagnóstico genético (Figura I3): secuenciación de exoma completo o WES (*whole-exome sequencing*), de genoma completo o WGS (*whole-genome sequencing*) y secuenciación de paneles de genes o *targeted sequencing*. Desde sus inicios, la estrategia de NGS más popular para el descubrimiento de nuevos genes causantes de enfermedades ha sido la WES. Ésta consiste en el enriquecimiento mediante diferentes sistemas de captura y posterior secuenciación de regiones codificantes e intrónicas flanqueantes del genoma (Figura I3). La WGS permite además estudiar las regiones no codificantes del genoma (Figura I3). A diferencia de estas dos estrategias, en los paneles de genes se estudian únicamente exones, intrones y regiones promotoras de genes específicos asociados a ciertas enfermedades, con el diseño de sondas de captura de éstas y otras regiones concretas en las que con frecuencia se detectan mutaciones asociadas a un determinado trastorno (Figura I3). Este tipo de secuenciación dirigida resulta económica y de fácil aplicación diagnóstica para el cribado genético de amplias cohortes de pacientes con un fenotipo similar, sin la necesidad de disponer de familiares afectados, y para la detección de variantes en genes que, debido a su tamaño, resulta costoso su estudio mediante secuenciación por Sanger.

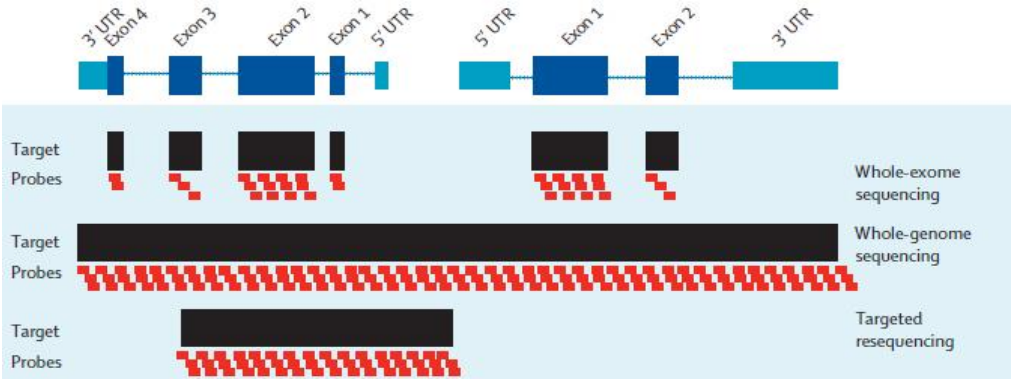

Figura I3: Representación de las regiones del genoma en las que se centra el diseño de las estrategias basadas en NGS para el diagnóstico genético que se utilizan en la actualidad (WES/WGS: secuenciación de exoma/genoma; *targeted sequencing*: secuenciación dirigida, panel de genes) (adaptado de Bettens *et al*. 2013).

En la actualidad, la WGS no es una técnica ampliamente utilizada, ya que en comparación con la WES, supone un mayor coste económico y el análisis de los datos obtenidos entraña una mayor dificultad debido a su tamaño. Sin embargo, existe una proporción considerable de regiones no codificantes potencialmente funcionales que podrían albergar alteraciones causantes de enfermedades. Además, la cobertura obtenida en WGS es más uniforme que en WES, por lo que los datos obtenidos son más adecuados para la detección de variantes del número de copias (CNVs) (Sims *et al.,* 2014).

La aplicación de NGS ha cambiado el panorama en la investigación de enfermedades raras, al haber permitido la caracterización de un número significativo de nuevos genes implicados en patologías en poco tiempo. El síndrome de Miller fue la primera enfermedad mendeliana cuya causa genética fue identificada con la aplicación de WES, al detectarse mutaciones causales en el gen *DHODH* en 4 individuos afectados pertenecientes a tres familias (Ng *et al*., 2010). El exoma comprende el 1% del genoma humano, lo que supone aproximadamente 30 Mb de tamaño, y se estima que el 85% de las mutaciones patológicas se encuentran en dichas regiones (Liu *et al.*, 2015). La aplicación de WES para resolver el origen genético de patologías poco frecuentes ha resultado en la identificación de más de 500 genes noveles asociados a enfermedades entre los años 2010 y 2015 (Chong *et al*., 2015). Además, ha dado lugar a la detección de mutaciones causantes de fenotipos nuevos en genes previamente asociados a otras patologías (Gilissen *et al.,* 2011), lo que indicaría que probablemente diferentes enfermedades compartan mecanismos fisiopatológicos comunes. De la misma forma, se

ha demostrado la variabilidad fenotípica existente causada por diferentes mutaciones en un mismo gen, lo que ha permitido definir nuevos mecanismos de la enfermedad.

El acceso a técnicas de NGS para el diagnóstico genético de pacientes con enfermedades raras es crucial para completar el conocimiento de las bases genéticas de patologías humanas. En aquellas cuya causa genética sea conocida, los pacientes podrán obtener un diagnóstico rápido, y en aquellas otras en las que se desconozca, será posible la búsqueda de nuevos genes causantes de la patología.

3. Análisis de variantes genéticas: manejando una gran cantidad de datos

La aplicación de WES o WGS elimina la necesidad de priorizar genes candidatos a secuenciar. En cambio, el reto reside en distinguir la posible variante causal del resto de polimorfismos no patológicos y errores de secuenciación de forma eficiente. La priorización e interpretación de variantes se ha convertido un paso crucial en la identificación de nuevos genes implicados en la fisiopatología de enfermedades mendelianas.

De forma rutinaria, tras la secuenciación de exoma o genoma, las lecturas obtenidas son alineadas frente a un genoma de referencia y posteriormente se identifican las variantes presentes en la muestra. Diferentes estudios han demostrado que, en promedio, en un experimento de WES de una muestra de origen europeo-americano se llegan a identificar alrededor de 20.000 variantes (Bamshad *et al.* 2011), siendo obviamente mucho mayor este número en un experimento de WGS. Esta cifra puede variar dependiendo del procedimiento utilizado para el enriquecimiento y captura de regiones codificantes del genoma, de la plataforma de secuenciación y de los algoritmos de procesamiento de secuencias implementados para el alineamiento frente al genoma de referencia e identificación de variantes (Rabbani *et al.*, 2012). A este respecto, la elección de herramientas bioinformáticas a utilizar influye sobre los resultados que se pueden obtener de un mismo *set* de datos. Se han publicado diferentes trabajos en los que tratan de establecer la concordancia y precisión de diferentes *pipelines* de procesado de datos de secuenciación masiva para la identificación de variantes (Liu *et al.*, 2013; O'Rawe *et al.*, 2013; Yu y Sun, 2013; Pirooznia *et al.*, 2014). En ellos, demuestran que es posible obtener diferentes resultados en función del alineamiento de las lecturas, los

tratamientos post-alineamiento, los parámetros configurados en los diferentes programas utilizados y los modelos matemáticos en los que se basan los programas para la identificación de variantes. Todo ello hace que la concordancia entre *pipelines* sea baja, siendo posible detectar variantes con un *pipeline* que otros quizás no detecten. En la búsqueda de mutaciones causantes de enfermedades mendelianas, es conveniente no dejar escapar ninguna posibilidad, por lo que es recomendable procesar los datos con diferentes herramientas y combinar los resultados obtenidos en todas ellas.

Con el fin de reducir el número de falsos positivos de entre las variantes detectadas en un *pipeline*, éstas pueden ser filtradas en función de ciertos parámetros de calidad, como el número de lecturas en las que aparece (al menos 5 lecturas independientes) o el porcentaje de lecturas (al menos un 20% para variantes en heterocigosis y un 80% para variantes en homocigosis) (Gilissen *et al.* 2012). A continuación, es conveniente filtrar las variantes frente a bases de datos de frecuencias poblacionales. Para la identificación de mutaciones causales en enfermedades raras, se sigue la suposición de que la variante ha de ser novel (no descrita en ningún individuo) o estar presente en la población con una frecuencia inferior al 1%. Con este paso, se consigue reducir el número de variantes entre 150 y 500 cuando se trabaja con variantes obtenidas mediante WES, siendo necesario por tanto aplicar más criterios de priorización como el efecto que puede causar sobre la función de la proteína (Rabbani *et al.*, 2012). Aplicando este criterio, se descartarían las variantes sinónimas al ser consideradas no deletéreas, quedando en su mayoría variantes no sinónimas, que afectan al *splicing* del mRNA o que resultan en una proteína truncada (cambios de la pauta de lectura y/o generación de un codón de parada). De forma complementaria, se aplican métodos bioinformáticos, consistentes en predicciones *in silico*, para valorar el impacto funcional de estas variantes, que permiten saber si se localizan en regiones conservadas evolutivamente o que miden la alteración que podrían causar sobre la estructura de la proteína, entre otros. Adicionalmente, mediante WGS va a ser posible obtener información sobre regiones no codificantes conservadas cuya funcionalidad en la mayoría de casos aún se desconoce, lo que complica su análisis y filtrado.

En un experimento de WES, la identificación e interpretación de la posible variante causal es sencilla cuando ha sido previamente descrita en otros individuos que presentan

el mismo fenotipo, o si de forma clara se puede demostrar el efecto deletéreo sobre la proteína. Para el resto de cambios, su interpretación se complica al no contar con este tipo de información, sobre todo cuando se trata de genes nuevos de los que no se conocen muchos pacientes afectados por mutaciones en ellos ni tampoco el papel que juegan en la fisiopatología de la enfermedad.

Aun así, para una correcta interpretación siempre hay que tener en cuenta el modo de herencia. La tasa de éxito en la resolución del diagnóstico genético mediante aplicación de WES es mayor si se estudian tríos (padre, madre e hijo afectado) en lugar de únicamente el paciente, y se mejora si se estudian más miembros afectados de la familia. El estudio incluyendo a familiares resulta de gran utilidad, porque permite combinar distintas estrategias tradicionales como el análisis de ligamiento o el cartografiado por homocigosidad para la identificación de la variante causal, según si la enfermedad segrega en la familia con un patrón de herencia autosómico dominante o recesivo respectivamente (Figura I4 A y B). Sin embargo, esto encarece el coste del estudio, además de que no siempre es posible contar con muestras de familiares enfermos. En el caso de estudios individuales o casos esporádicos, se pueden plantear igualmente ambos tipos de herencia, según si los progenitores son portadores de mutaciones patológicas que confluyen en el caso índice (Figura I4 C), o por el contrario se trata de una mutación *de novo* ocurrida en el paciente y no presente en los progenitores (Figura I4 E). Otras estrategias en las que tiene cabida cualquier modo de herencia son frecuentemente aplicadas en casos esporádicos. Se pueden estudiar conjuntamente individuos no relacionados cuyos fenotipos solapan (Figura I4 D), suponiéndose que no existe heterogeneidad genética y que el origen de la patología es común Por último, cuando no se dispone de una historia familiar claramente definida, puede resultar interesante seguir una aproximación basada en el estudio de genes candidatos (Figura I4 F), centrándose el análisis en las variantes identificadas en genes cuya implicación en el mecanismo de la enfermedad en cuestión sea conocido como método para resolver de forma sencilla el diagnóstico.

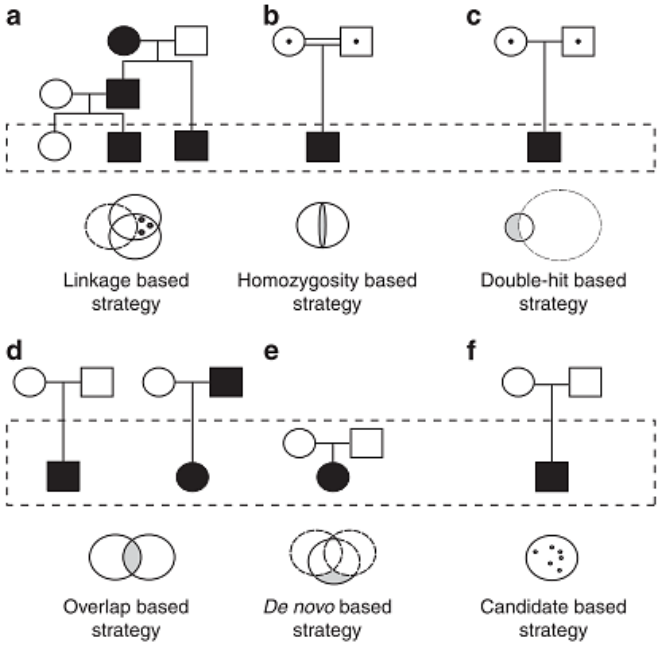

Figura I4: Estrategias e hipótesis para la identificación de variantes causales mediante la aplicación de WES y el modo de herencia: A) análisis de ligamiento, B) cartografiado por homocigosidad, C) heterocigosis compuesta, D) solapamiento de fenotipos, E) mutaciones *de novo* y F) análisis de genes candidatos. En cada pedigree, los símbolos negros representan individuos afectados, indicándose con un punto en su interior a los portadores sanos de la enfermedad. Los rectángulos distinguen a los individuos estudiados mediante WES, mientras que los círculos muestran las variantes que quedarían acotadas con cada estrategia (adaptado de Gilissen *et al.,* 2012).

HIPÓTESIS Y OBJETIVOS

HIPÓTESIS

Las neuropatías periféricas hereditarias comprenden un grupo heterogéneo de trastornos neurológicos que presentan una alta heterogeneidad tanto clínica como genética. La secuenciación del exoma, una estrategia interesante porque es relativamente económica y abordable, ha hecho posible el descubrimiento de cientos de genes implicados en enfermedades mendelianas. Es recomendable aplicar esta metodología principalmente cuando se trata del estudio de pacientes bien caracterizados clínicamente en quienes se han descartado mutaciones en genes frecuentes y/o genes candidatos. Planteamos aquí mejorar el desarrollo de la **secuenciación de exoma** para el diagnóstico genético **mediante un estudio comparativo de dos *pipelines* informáticos** con el fin de mejorar la tasa de éxito de pacientes diagnosticados genéticamente. Esta aproximación permitirá la **identificación de nuevas mutaciones** lo que repercutirá favorablemente en el consejo genético que estos enfermos reciben y mejorará nuestra comprensión sobre la fisiopatología de estas enfermedades.

OBJETIVOS

El objetivo general es la mejora del diagnóstico genético en pacientes con CMT/dHMN mediante el estudio comparativo de dos *pipelines* informáticos a partir de datos de secuenciación de exoma.

Objetivos específicos:

1.1.- Secuenciación de exoma en ocho pacientes afectados por CMT/dHMN.

1.2.- Análisis informático de los resultados mediante dos *pipelines* CNAG y BiER, y selección de variantes de interés.

1.3.- Estudios *in silico* y genéticos para el cribado de cambios.

1.4.- Investigación funcional de cambios candidatos a ser la mutación causal.

MATERIAL Y MÉTODOS

1. Selección de pacientes y secuenciación de exoma

Para este estudio, se seleccionaron 8 casos esporádicos no relacionados supervisados por el Servicio de Neurología del Hospital Universitari i Politècnic La Fe, con diagnóstico clínico inicial de CMT/dHMN en los que la búsqueda de mutaciones patológicas en los genes más frecuentes de CMT y dHMN mediante otras técnicas había resultado negativa. Los pacientes dieron su consentimiento para participar en este estudio.

Para la secuenciación de exoma, las muestras de DNA genómico se obtuvieron en nuestro laboratorio a partir de sangre periférica de los individuos seleccionados. Posteriormente, tanto la generación da las librerías como su secuenciación mediante la tecnología Illumina, se llevaron a cabo en el Centro Nacional de Análisis Genómico (CNAG). El kit de captura empleado fue *SureSelect All-Exon V5 (50 Mb)* (Agilent Technologies, USA).

2. Genes estudiados implicados en neuropatías periféricas hereditarias

Con el fin de acotar la búsqueda de mutaciones causales en los casos estudiados, se confeccionó una lista de genes previamente asociados a neuropatías, que comparten signos clínicos similares con los observados en los pacientes estudiados, que se recoge en el Anexo I. Para ello, se consultaron revisiones bibliográficas recientes (Chen *et al.* 2013; Rossor *et al.* 2013; Coutelier *et al.* 2015) y las bases de datos OMIM y HGMD, con información sobre mutaciones en enfermedades mendelianas, además de otras especializadas en neuropatías (*Neuromuscular Disease Center, University of Washington*).

Los genes han sido clasificados en diferentes categorías según el tipo de neuropatía hereditaria en la que están comúnmente implicados: HSMN, HSAN, dHMN, SMA, ALS, HSP y ataxia cerebelosa. Se obtuvieron las coordenadas cromosómicas en el GRCh37 a través de la herramienta Ensembl BioMart (Kinsella *et al.*, 2011) para cada uno de los genes con el fin de generar un fichero en formato BED con el que se ha trabajado posteriormente para analizar la cobertura obtenida en estas regiones y las variantes localizadas en ellas.

3. Análisis bioinformático

3.1 *Pipelines* de procesamiento de datos de secuenciación de exoma

Las lecturas o *reads* resultantes de la secuenciación de exoma en el CNAG de las muestras seleccionadas, recogidas en ficheros en formato FASTQ, fueron procesadas simultáneamente por los *pipelines* de procesado de datos de secuenciación de exoma implementados en el CNAG y en la Plataforma de Bioinformática para las Enfermedades Raras (BiER). En la Figura M1, se recoge el esquema que resume las fases en las que consiste el procesado de datos de secuenciación de exoma, indicando en cada una de ellas los programas utilizados en cada *pipeline.*

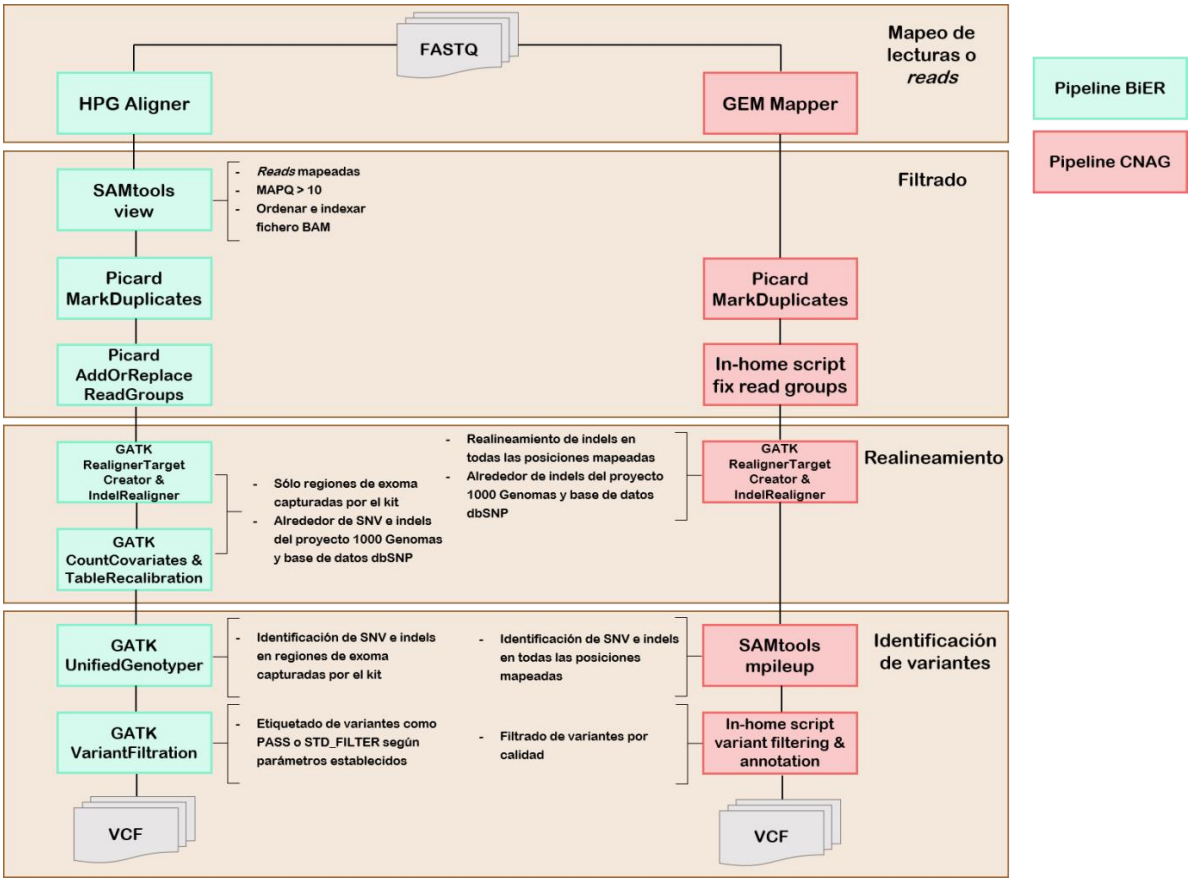

Figura M1: Esquema de los *pipelines* de procesado de datos de secuenciación de exoma implementados en el CNAG y en el BiER. Para cada una de las fases de procesado, se indican los programas utilizados, así como las diferencias en cuanto a la forma de uso entre ambos.

Por una parte, en el *pipeline* implementado en el BiER (Figura M1) las lecturas fueron alineadas frente al GRCh37 utilizando el programa HPG Aligner (Tárraga *et al.*, 2014), el cual se basa en la utilización de arreglos de sufijos (*suffix arrays*) para la mejora en el rendimiento de este proceso. Los ficheros BAM, que contienen toda la información relativa al alineamiento de las lecturas frente al genoma de referencia, fueron filtrados con el programa SAMtools (Li *et al*., 2009) según la calidad de mapeo (MAPQ) resultante (superior a 10 en la escala Phred), y se marcaron los alineamientos que implicaban lecturas duplicadas utilizando la herramienta MarkDuplicates incluida en Picard (http://picard.sourceforge.net). A continuación, se pasó a trabajar únicamente con las regiones exónicas del GRCh37, cuyas coordenadas cromosómicas son proporcionadas por el fabricante del kit de captura empleado en este estudio en un fichero en formato BED. Para el realineamiento de indels (inserciones/deleciones), recalibración de calidades de las bases e identificación de variantes se utilizaron las herramientas de GATK disponibles para ello (McKenna *et al*., 2009), siguiendo las guías de buenas prácticas recomendadas por sus desarrolladores (DePristo *et al*., 2011; Van der Auwera *et al*., 2013).

Por otra parte, en el *pipeline* implementado por el CNAG (Figura M1), para el alineamiento de las lecturas frente al GRCh37 se utilizó el programa GEM Mapper (Marco-Sola *et al.*, 2012), y para la identificación de variantes se utilizó el algoritmo implementado en SAMtools para ello (Li, 2011a), descartando finalmente variantes obtenidas de baja calidad. A diferencia del *pipeline* seguido en el BiER, en éste se trabaja con todas las posiciones del genoma de referencia sobre las que se han alineado lecturas, y sólo se realizó el realineamiento de indels con las herramientas de GATK previo a la identificación de variantes.

3.2 Análisis primario

Se realizó un control de calidad de los datos generados por ambos *pipelines* utilizando el programa FastQC (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc) y SAMtools en varios puntos: sobre los ficheros en formato FASTQ iniciales y en los diferentes ficheros BAM generados tras el mapeo, tras la eliminación de lecturas duplicadas y tras el realineamiento de SNV e indels. También se llevó a cabo un control de calidad de las variantes obtenidas utilizando el programa HPG Variant (Medina *et al*., 2012) y una

comparativa de las mismas manejando los ficheros en formato VCF generados por ambos *pipelines* con comandos de la *shell* de Linux.

3.3 Análisis de la cobertura

Se evaluó la cobertura obtenida en cada muestra tras el procesado de cada *pipeline* de los datos de secuenciación, estudiándose con mayor detalle la cobertura obtenida en las regiones del genoma en las que se localizan los genes implicados en neuropatías de interés. Para ello, se utilizó la herramienta DepthOfCoverage incluida en GATK, los ficheros BAM finales a partir de los cuales cada *pipeline* realizó la identificación de variantes, y las coordenadas cromosómicas de las regiones exónicas del GRCh37 capturadas por el kit empleado que proporciona el fabricante en un fichero en formato BED. Tras la medición de la cobertura, se generaron diferentes *scripts* en R para el manejo de las tablas obtenidas, y para su representación gráfica se utilizó el paquete ggplot2 (Wickham, 2009).

3.4 Anotación y priorización de variantes

Los ficheros en formato VCF que recogen las variantes obtenidas para cada muestra por cada *pipeline* utilizado, fueron anotados de diferente forma. Los ficheros VCF proporcionados por el CNAG habían sido anotados con el programa SnpEff (Cingolani *et al*., 2012). Para trabajar con las variantes detectadas por el *pipeline* del BiER, se combinó la anotación proporcionada por las herramientas web BiERapp (Alemán *et al.*, 2014) y Ensembl Variant Effect Predictor (McLaren *et al.*, 2010). Los ficheros VCF anotados fueron filtrados con la herramienta Tabix (Li, 2011b) para examinar las variantes localizadas en los genes previamente asociados a neuropatías de interés, utilizando para ello el fichero BED generado anteriormente con las coordenadas cromosómicas en el GRCh37 de dichos genes. Se trabajó en paralelo con las variantes obtenidas en cada *pipeline* con el fin de identificar la posible variante causal de la enfermedad en cada uno de los casos estudiados. Se desarrollaron distintos scripts en Bash, Python y R para filtrar y combinar la información de los diferentes ficheros generados en este proceso. Para reducir el número de variantes y determinar las candidatas en cada caso estudiado, se establecieron los siguientes criterios para facilitar su priorización:

a) Se clasificaron las variantes por el efecto que causan en la proteína, seleccionando aquellas que resultan potencialmente deletéreas (cambios no sinónimos, de pauta de lectura y/o generación de un codón de parada, y que alterarían el *splicing* del mRNA).

b) Las variantes seleccionadas se filtraron según su frecuencia poblacional o MAF (*Minor Allele Frequency*), conservando aquellas que presentan una frecuencia inferior al 1% en las diferentes bases de datos de frecuencias poblacionales consultadas: 1000 Genomas, Exome Variant Server (EVS), CIBERER Spanish Variant Server (CSVS) y Exome Aggregation Consortium (ExAC).

c) Se realizaron búsquedas bibliográficas y en bases de datos que recogen y clasifican variantes causantes de enfermedades (HGMD y ClinVar) con el fin de conocer si alguna de las variantes había sido previamente descrita como patológica en otros individuos con neuropatías. De ser así, la probabilidad de que la variante fuera la causante de la enfermedad era alta, y por tanto quedaba priorizada frente al resto.

d) Sobre aquellas variantes cuya frecuencia poblacional era inferior al 1% y no habían sido descritas previamente como patológicas, se realizaron diferentes estudios *in-silico* para evaluar el posible impacto que podría tener sobre la función de la proteína:

    i. Para determinar si el aminoácido afectado por el cambio y su entorno forman parte de un clúster conservado en la proteína, se seleccionaron proteínas homólogas en especies cercanas en las bases de datos UniProtKB y Ensembl para realizar un alineamiento múltiple de sus secuencias aminoacídicas con la herramienta web Clustal Omega (http://www.ebi.ac.uk/Tools/msa/clustalo/).

    ii. Para completar la evaluación de la conservación entre especies de la posición en la que se encuentra la variante, se consultaron en la base de datos UCSC Genome Browser las puntuaciones o *scores* calculados para dichas posiciones nucleotídicas por las herramientas de predicción de conservación GERP (Cooper *et al.*, 2010) y PhyloP (Pollard *et al.*, 2010). Estos métodos de predicción se basan en algoritmos que evalúan la probabilidad de que la

posición sea neutral o por el contrario pertenezca a una región conservada en la que el cambio evolutivo está restringido. Para ello, a partir del alineamiento múltiple de secuencias nucleotídicas entre mamíferos (GERP) o en más de 100 vertebrados (PhyloP) comparan el número de sustituciones esperadas y observadas bajo hipótesis de neutralidad.

iii. Para las variantes no sinónimas, se utilizaron las herramientas de predicción de patogenicidad SIFT (Ng y Henikoff, 2001), PolyPhen-2 (Adzhubei *et al.*, 2000) y PROVEAN (Choi y Chan, 2015). En ellas se han implementado algoritmos que, basándose principalmente en la homología de secuencia aminoacídica con otras especies y la conservación de la estructura de la proteína, calculan una puntuación que se interpreta como la probabilidad de que la variante sea patológica.

iv. Para aquellas variantes que pudiesen alterar el *splicing* del mRNA, se utilizaron diferentes programas de predicción de alteración de sitios de reconocimiento (secuencias consenso de los extremos 5' y 3' de exones e intrones) de la maquinaria que participa en este proceso (espliceosoma). Se utilizó: NNSplice (http://www.fruitfly.org/seq_tools/splice.html), Human Splicing Finder (http://www.umd.be/HSF3/), NetGene2 (http://www.cbs.dtu.dk/services/NetGene2/), y SpliceView (http://bioinfo.itb.cnr.it/~webgene/wwwspliceview.tml).

Finalmente, en aquellas muestras en las que se logró identificar variantes que podrían ser las causantes de la patología y cuyo modo de herencia era compatible con lo descrito en la bibliografía para el fenotipo asociado, se confirmó mediante secuenciación por Sanger la presencia de la mutación. Se realizaron estudios de segregación en los casos en los que se disponía de muestra de DNA de familiares sanos o enfermos.

4. Análisis funcional de variantes de *splicing* mediante minigenes

Para investigar el efecto que podría causar la variante c.446+5G>C identificada en el gen *FIG4* en el procesado del mRNA, se realizó un ensayo funcional basado en la obtención de construcciones de minigenes a partir del vector pSPL3. Para ello, se siguieron procedimientos ampliamente descritos en la bibliografía, que se resumen en la Figura M2 (Bottillo *et al.*, 2007; Desviat *et al.*, 2012).

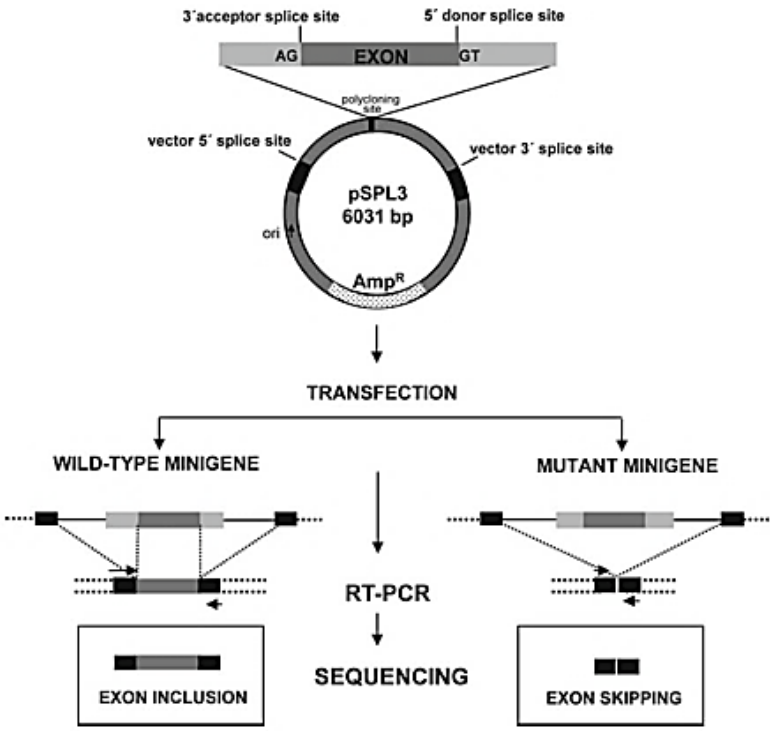

Figura M2: Esquema del protocolo para el estudio de variantes de *splicing* mediante la obtención de construcciones de minigenes a partir del vector pSPL3. Las cajas negras representan los exones del vector, que se denominan SD6 y SA2. Las cajas en gris oscuro representan el exon a estudiar, que se subclona en el vector junto con sus secuencias intrónicas flanqueantes (gris claro) (adaptado de Desviat *et al.*, 2012).

La variante c.446+5G>C se sitúa en la secuencia consenso 5' (*splicing donor site*) del intrón 4 del gen *FIG4*. Por ello, se diseñaron *primers* para amplificar con la enzima de alta fidelidad *PfuTurbo® DNA Polymerase* (Agilent Technologies, Santa Clara, California, USA) a partir de DNA genómico del paciente el exón 4 junto con 100-150 pb de sus secuencias intrónicas flanqueantes. Los sitios de corte de las enzimas *Xho*I y *Nhe*I se incluyeron en los *primers* para subclonar el fragmento amplificado entre dichos sitios en el vector pSPL3, como se indica en la Figura M2. Al ser el paciente portador en heterocigosis del cambio,

teóricamente se obtendrían indistintamente construcciones con el cambio (mutantes) o con la secuencia de referencia (*wild-type*) (Figura M2). Se digirió vector e inserto con ambas enzimas, para posteriormente proceder con la ligación con la enzima *T4 DNA Ligase* (Promega, Fitchburg, Wisconsin, USA) y transformación en bacterias electro-competentes de la cepa DH5α de *E.Coli*. Las colonias obtenidas fueron testadas por PCR para comprobar la presencia de la construcción, y se confirmó mediante secuenciación por Sanger la direccionalidad de los insertos así como la presencia o no del cambio. Como todas las colonias testadas resultaron positivas para la construcción *wild-type* (pSPL3::FIG4-E4-WT), por mutagénesis dirigida a partir de uno de ellos se obtuvo la construcción con la mutación c.446+5G>C (pSPL3::FIG4-E4-Mut). Se sembraron células HeLa que fueron transfectadas con ambas construcciones, y se extrajo RNA con TRIzol® Reagent (Life Technologies, Carlsbad, California, USA) de los cultivos a las 24 y 48 horas post-transfección. Desde el RNA extraído, se realizó una RT-PCR para obtener cDNA y amplificar los transcritos utilizando *primers* específicos de los exones SD6 y SA2 del vector (Figura M2). Por último, se llevó a cabo la separación de los productos de PCR mediante electroforesis en gel de agarosa al 2% teñido con GelRed™. Se extrajeron y purificaron las bandas de gel detectadas correspondientes a cada transcrito, para corroborar mediante secuenciación por Sanger el procesamiento de las construcciones llevado a cabo por parte de la maquinaria de *splicing* celular.

Los *primers* utilizados para la amplificación, subclonación de los insertos en el vector y secuenciación de las construcciones se incluyen en el Anexo II.

RESULTADOS

1. Análisis primario

1.1 Control de calidad inicial de las secuencias

La secuenciación de exoma llevada a cabo en las ocho muestras de pacientes no relacionados generó un total de 60 Gb de datos. Como en los ficheros FASTQ enviados desde el CNAG las secuencias ya no contenían restos de adaptadores, no fue necesario realizar el paso de eliminación de este tipo de secuencias. Igualmente, se evaluó la calidad de las lecturas obtenidas.

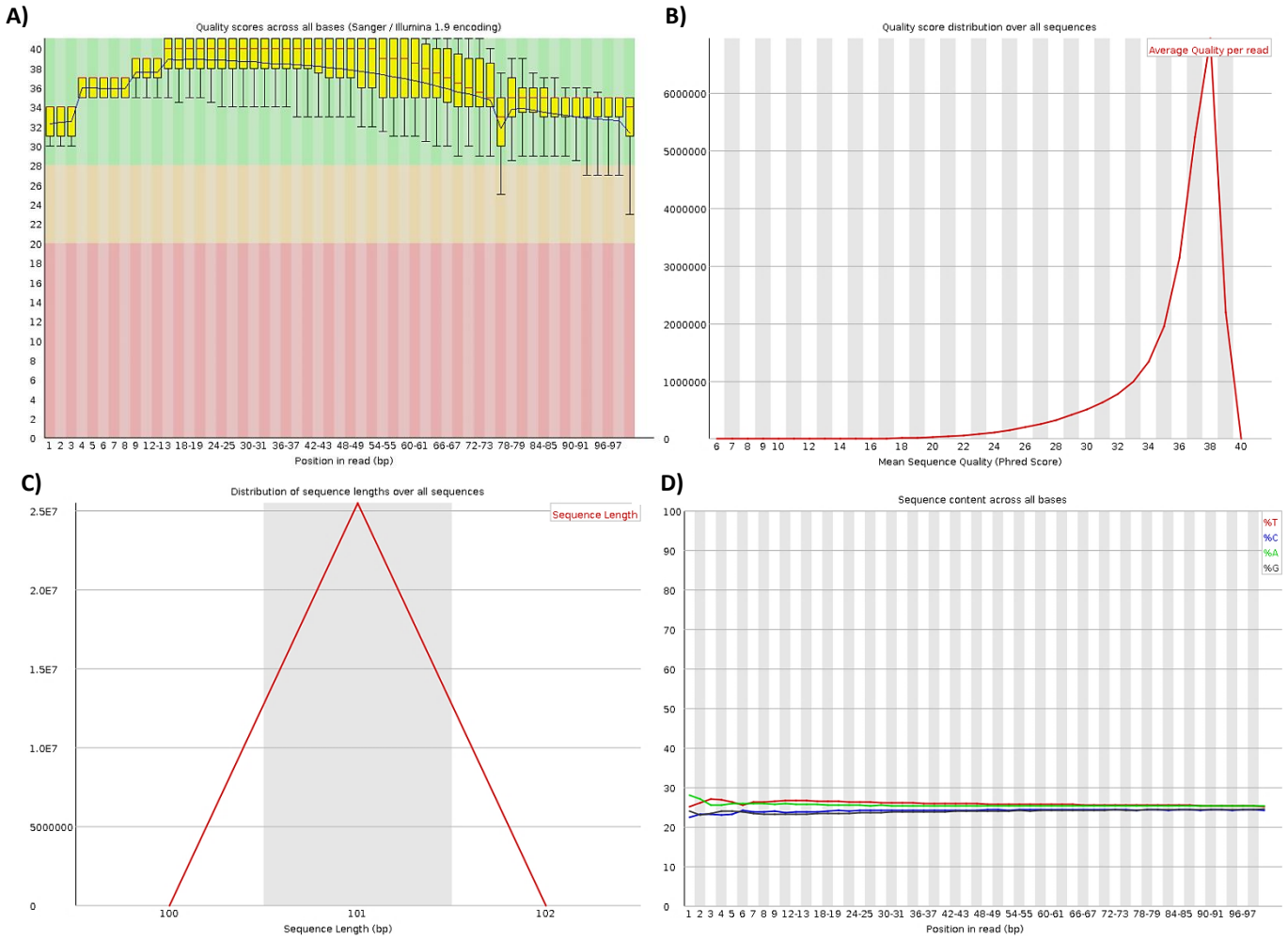

Figura R1: Gráficos representativos de A) la calidad por base, B) la calidad media de las lecturas, C) la distribución de tamaños de las lecturas y D) la distribución de porcentajes de cada nucleótido por base para todas las muestras generados con el programa FastQC.

En general, se observó que la calidad media por base se situaba por encima de 30 a lo largo de todas las lecturas. De la misma forma, la calidad media de secuencia se situaba por encima de este valor en la mayor parte de los casos (Figura R1 A y B). Se confirmó que todas las lecturas presentaban un tamaño en torno a 100 pb (Figura R1 C), y que los porcentajes relativos de cada uno de los 4 nucleótidos posibles (adenina, timina, guanina y citosina) eran similares para cada una de las posiciones de las lecturas, siguiendo distribuciones paralelas (Figura R1 D). Puesto que todos los parámetros evaluados se encontraban dentro de la normalidad, no se consideró que fuese necesario realizar ningún tipo de limpieza en las lecturas antes de proceder con el mapeo.

1.2 Control del procesado de lecturas de cada *pipeline*

En ambos *pipelines*, tras el alineamiento de las lecturas frente al GRCh37, tienen lugar varios pasos de filtrado en el que el número de lecturas mapeadas disminuye hasta la identificación de variantes. Se ha descrito este descenso para cada uno de los procesados llevados a cabo.

En el *pipeline* establecido en el BiER, hay cuatro pasos en los que el número de lecturas inicial disminuye (mapeo, filtrado según calidad de mapeo o MAPQ, eliminación de duplicados y realineamiento en las regiones de exoma capturadas por el kit empleado). Como se observa en la Tabla R1, más del 99% de las parejas de lecturas o *read pairs* obtenidas en la secuenciación pudieron ser alineadas frente al GRCh37 en todas las muestras con el programa HPG Aligner utilizado para ello en este *pipeline.* A continuación, fueron filtradas individualmente las lecturas por MAPQ, y resultó que alrededor del 50% presentaban una MAPQ inferior a 10 en la escala Phred, por ello el número total de lecturas iniciales resultó reducido a la mitad. Finalmente, tras las etapas de eliminación de lecturas duplicadas y realineamiento local, para la identificación de variantes se contaba con alrededor del 36% de las lecturas totales iniciales. Puesto que en este *pipeline* se trabaja únicamente con las regiones de exoma capturadas por el kit empleado a partir de la etapa de realineamiento local, en el recuento final sólo hay lecturas localizadas en dichas regiones.

Tabla R1: Recuento de lecturas en cada una de las muestras secuenciadas tras las etapas de filtrado establecidas en el *pipeline* del BiER.

	SGT038	**SGT077**	**SGT161**	**SGT187**
N lecturas directo	31471997	27308034	27566691	29730857
N lecturas reverso	31471997	27308034	27566691	29730857
N parejas lecturas mapeadas	31304563	27166800	27429566	29593092
% parejas lecturas mapeadas	99,47	99,48	99,50	99,54
N lecturas MAPQ > 10	32844926	28443473	28775620	31001122
% lecturas MAPQ >10	52,18	52,08	52,19	52,14
N lecturas *single hit*	28115216	24680760	24997317	26579044
% lecturas *single hit*	44,67	45,19	45,34	44,70
N lecturas *single hit* realineadas	22724178	20024427	20213242	21530503
% lecturas *single hit* realineadas	36,10	36,66	36,66	36,21
	SGT230	**SGT238**	**SGT241**	**SGT274**
N lecturas directo	30415770	29472514	29513223	30394832
N lecturas reverso	30415770	29472514	29513223	30394832
N parejas lecturas mapeadas	30296046	29333890	29365739	30268625
% parejas lecturas mapeadas	99,61	99,53	99,50	99,58
N lecturas MAPQ > 10	31712462	30770469	30803708	31699515
% lecturas MAPQ >10	52,13	52,20	52,19	52,15
N lecturas *single hit*	27084895	26406037	26174909	27122226
% lecturas *single hit*	44,52	44,80	44,34	44,62
N lecturas *single hit* realineadas	22377265	21543680	21579194	22387721
% lecturas *single hit* realineadas	36,79	36,55	36,56	36,83

N: número; %: porcentaje; MAPQ: calidad de mapeo en escala Phred; *single hit*: lecturas no duplicadas alineadas de forma única frente al genoma de referencia.

En contraste con el *pipeline* del BiER, en el procesado llevado a cabo en el CNAG sólo tienen lugar dos pasos que conllevan una reducción del número de lecturas inicial (mapeo y eliminación de duplicados). Como se detalla en la Tabla R2, con el programa GEM Mapper utilizado en el CNAG para el alineamiento de lecturas frente al GRCh37, sólo se lograron alinear alrededor del 85% de las parejas de lecturas iniciales en todas las muestras. Al no realizarse ningún tipo de filtrado adicional por calidad además de la eliminación de lecturas duplicadas, se contaba con algo más del 80% de parejas de lecturas iniciales para la identificación de variantes, localizadas a lo largo de todas las regiones del genoma humano de referencia mapeadas.

Tabla R2: Recuento de lecturas en cada una de las muestras secuenciadas tras las etapas de filtrado establecidas en el *pipeline* del CNAG

	SGT038	**SGT077**	**SGT161**	**SGT187**
N lecturas directo	31471997	27308034	27566691	29730857
N lecturas reverso	31471997	27308034	27566691	29730857
N parejas lecturas mapeadas	27098380	23450991	23668265	25554609
% parejas lecturas mapeadas	86,10	85,88	85,86	85,95
N parejas lecturas *single hit* realineadas	26533415	22904031	23170780	24894597
% parejas lecturas *single hit* realineadas	84,31	83,87	84,05	83,73
	SGT230	**SGT238**	**SGT241**	**SGT274**
N lecturas directo	30415770	29472514	29513223	30394832
N lecturas reverso	30415770	29472514	29513223	30394832
N parejas lecturas mapeadas	26257368	25386147	25365246	26242677
% parejas lecturas mapeadas	86,33	86,13	85,95	86,34
N parejas lecturas *single hit* realineadas	25639411	24773292	24539542	25693406
% parejas lecturas *single hit* realineadas	84,30	84,06	83,15	84,53

N: número; %: porcentaje; MAPQ: calidad de mapeo en escala Phred; *single hit*: lecturas no duplicadas alineadas de forma única frente al genoma de referencia.

1.3 Comparativa de variantes obtenidas en cada *pipeline*

A partir de los ficheros VCF generados en cada uno de los *pipelines* de procesado de datos de secuenciación masiva utilizados, se realizó el recuento de variantes identificadas en cada muestra (Tabla R3), tanto en el exoma completo como en los genes relacionados con neuropatías de interés, así como el número de variantes obtenidas en común por ambos *pipelines*.

Tabla R3: Recuento de variantes obtenidas para cada muestra en cada uno de los pipelines

	SGT038	**SGT077**	**SGT161**	**SGT187**
N variantes *pipeline* BiER	53851	53230	52870	52320
N variantes *pipeline* BiER genes neuropatías	506	487	489	539
N variantes *pipeline* CNAG	85942	83416	83533	84015
N variantes *pipeline* CNAG genes neuropatías	902	830	866	920
N variantes intersección	42631	42191	42016	41339
N variantes intersección genes neuropatías	442	415	424	459
	SGT230	**SGT238**	**SGT241**	**SGT274**
N variantes *pipeline* BiER	52009	52960	52484	54182
N variantes *pipeline* BiER genes neuropatías	513	526	507	473
N variantes *pipeline* CNAG	80023	82898	79876	84176
N variantes *pipeline* CNAG genes neuropatías	854	903	863	817
N variantes intersección	41064	42230	41801	42885
N variantes intersección genes neuropatías	433	466	451	410

N: número.

Como se observa en la Tabla R3, el número de variantes obtenidas a partir del *pipeline* del CNAG para cada muestra ha sido superior, ya que se buscaban variantes en todas las regiones del GRCh37 en las que se habían alineado lecturas, independientemente de que se trataran o no de regiones exónicas.

El estudio a nivel de variantes únicas obtenidas entre las 8 muestras ha revelado que existe una concordancia del 42% entre ambos *pipelines* para todas las regiones del exoma secuenciadas (Figura R2 A). En cuanto a las variantes localizadas en los genes relacionados con neuropatías (Figura R2 B), la concordancia también se encuentra en torno a ese valor. A pesar de que en el *pipeline* del BiER la identificación de variantes se reduce únicamente a las regiones de exoma capturadas por el kit empleado, con la optimización del alineamiento realizada por el programa GATK en torno a esas regiones y su algoritmo para identificar variantes, se ha obtenido un porcentaje significativo de variantes en exoma completo (11,3%) no detectadas por el *pipeline* del CNAG. Comparando únicamente las variantes localizadas en los genes relacionados con neuropatías de interés, se observa la misma tendencia, habiéndose obtenido un 9% de variantes exclusivamente en el procesado del BiER.

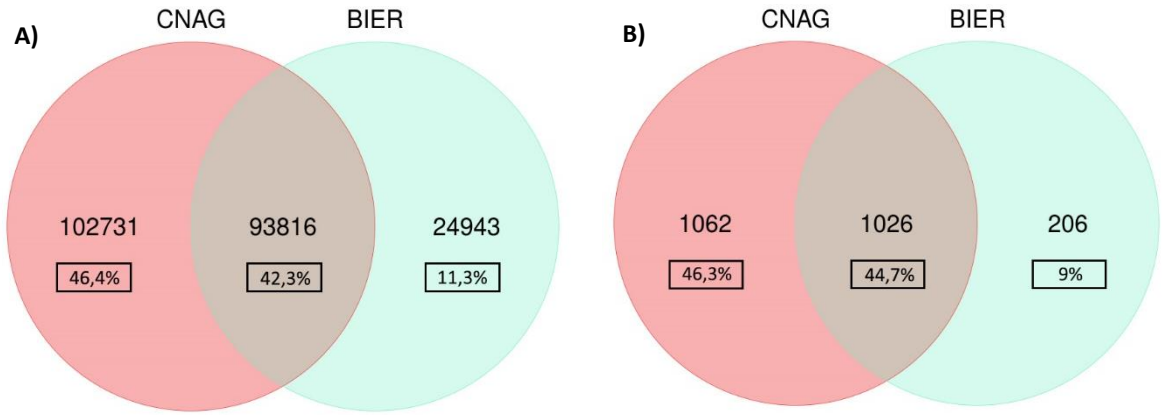

Figura R2: Concordancia de variantes únicas identificadas entre las 8 muestras entre ambos *pipelines*, CNAG y BiER, a nivel de A) exoma completo y B) genes relacionados con neuropatías. En cada caso se indica el número de variantes y el porcentaje sobre el total que dicho valor representa.

Adicionalmente, se realizó un control de calidad sobre las variantes obtenidas en cada *pipeline*, cuyo resultado se recoge en la Tabla R4. Los valores del ratio de Ts/Tv (transiciones/transversiones) registrados se sitúan en el rango entre 2,3 y 2,5. De nuevo destaca el hecho de que, aunque el número de variantes obtenidas en el *pipeline* del BiER es inferior, el porcentaje que cumplen los criterios de calidad ha sido superior con respecto al CNAG.

Tabla R4: Evaluación de las variantes obtenidas para cada muestra.

***Pipeline* BiER**

Muestra	N variantes detectadas	N variantes en homocigosis	N variantes en heterocigosis	N SNV	N indels	Ratio Ts/Tv	% PASS
SGT038	53851	19109	34686	50431	3420	2,34	94,58
SGT077	53270	19025	34208	50049	3221	2,34	95,21
SGT161	52870	19235	33592	49720	3150	2,34	94,92
SGT187	52320	19779	32500	49080	3240	2,33	94,7
SGT230	52009	20555	31414	48668	3341	2,30	94,41
SGT238	52960	19927	32995	49652	3308	2,36	94,95
SGT241	52484	19679	32767	49234	3250	2,36	94,74
SGT274	54182	18994	35141	50730	3452	2,32	94,54

***Pipeline* CNAG**

Muestra	N variantes detectadas	N variantes en homocigosis	N variantes en heterocigosis	N SNV	N indels	Ratio Ts/Tv	% PASS
SGT038	85942	33194	52748	80248	5694	2,51	82,27
SGT077	83416	31780	51636	78111	5305	2,48	80,70
SGT161	83533	32702	50831	78187	5346	2,47	81,22
SGT187	84015	34875	49140	78448	5567	2,49	81,50
SGT230	80023	34525	45498	74659	5364	2,46	82,54
SGT238	82898	33457	49441	77418	5480	2,49	81,94
SGT241	79876	32252	47624	74681	5195	2,51	82,12
SGT274	84176	31984	52192	78619	5557	2,47	82,21

N: número; SNV: *single nucleotide variant*; indels: inserciones/deleciones, Ratio Ts/Tv: Ratio transiciones/trasversiones; % PASS: porcentaje de variantes que cumplen los criterios de calidad establecidos en cada *pipeline*.

2. Análisis de la cobertura

La profundidad de secuenciación o cobertura alcanzada tras alinear las lecturas frente al genoma de referencia, eliminar duplicados y realizar otros filtrados adicionales es una medida de la calidad y sensibilidad para detectar variantes. Por ello, se evaluó la cobertura alcanzada en cada uno de los *pipelines* utilizados con la que se ejecutó la identificación de variantes en los datos de secuenciación de exoma obtenidos. La evaluación de la cobertura se realizó tanto a nivel de exoma completo como en los genes asociados a neuropatías de interés.

Tabla R5: Evaluación de la cobertura alcanzada en el exoma completo

Cobertura exoma completo BiER						
Muestra	**Media**	**Mediana**	**% bases > 10**	**% bases > 15**	**% bases > 30**	**% bases > 50**
SGT038	36,52	35	94,9	88,2	58,2	24,0
SGT077	32,12	30	93,1	84,2	49,6	16,6
SGT161	32,45	31	93,2	84,5	50,6	17,1
SGT187	34,53	33	94,3	86,8	55,0	20,4
SGT230	36,23	35	94,9	88,2	58,1	23,5
SGT238	34,76	33	94,3	86,8	55,0	21,0
SGT241	34,99	33	94,5	87,2	55,9	21,3
SGT274	36,18	34	94,8	88,0	58,1	23,4
Total	34,70	33	94,3	86,7	55	20,7
Cobertura exoma completo CNAG						
Muestra	**Media**	**Mediana**	**% bases > 10**	**% bases > 15**	**% bases > 30**	**% bases > 50**
SGT038	75,63	68	97,2	95,6	85,7	66,1
SGT077	64,74	58	96,7	94,4	81,1	57,8
SGT161	65,62	59	96,8	94,5	81,4	58,9
SGT187	70,59	64	97,1	95,2	84,2	63,2
SGT230	74,84	67	97,2	95,5	85,6	66,1
SGT238	71,33	64	97,1	95,2	84	63,2
SGT241	71,81	65	97,1	95,3	84,5	64
SGT274	74,85	68	97,1	95,4	85,5	66,1
Total	71,07	64	97	95,1	84	63,1

Tabla R6: Evaluación de la cobertura alcanzada en los genes asociados a neuropatías

Cobertura genes neuropatías BiER						
Muestra	**Media**	**Mediana**	**% bases > 10**	**% bases > 15**	**% bases > 30**	**% bases > 50**
SGT038	35,33	33	94,9	87,7	55,8	21,7
SGT077	31,55	30	93,5	84,4	48,1	15,2
SGT161	31,43	30	93,3	84,0	48,3	15,2
SGT187	33,80	32	94,8	86,9	53,6	18,7
SGT230	35,58	34	95,4	88,4	56,9	21,9
SGT238	33,74	32	94,2	86,3	52,9	19,0
SGT241	34,34	33	94,7	87,1	54,6	19,8
SGT274	35,43	34	95,1	88,2	56,8	21,7
Total	33,90	32,25	94,5	86,6	53,4	19,2
Cobertura genes neuropatías CNAG						
Muestra	**Media**	**Mediana**	**% bases > 10**	**% bases > 15**	**% bases > 30**	**% bases > 50**
SGT038	73,71	65	98,2	96,7	85,8	65,1
SGT077	64,04	57	97,8	95,6	82,1	57,3
SGT161	64,08	58	97,8	95,6	81,8	57,7
SGT187	69,64	63	98,1	96,5	85,0	62,7
SGT230	74,13	67	98,3	96,7	86,7	66,1
SGT238	69,62	62	98,1	96,2	84,4	61,9
SGT241	71,10	64	98,2	96,5	85,3	64,0
SGT274	73,92	67	98,2	96,5	86,8	65,8
Total	70,03	62,88	98,1	96,3	84,7	62,6

El estudio de la cobertura a nivel de exoma completo (Tabla R5) muestra que más del 90% de las bases capturadas por el kit han sido leídas al menos 10 veces (10x), habiéndose alcanzado una cobertura media superior a 30x con el procesado realizado en el BiER, mientras que con el procesado del CNAG, la cobertura media ha sido superior a 60x. En los genes asociados a neuropatías, se observó la misma tendencia en la cobertura; en todas las muestras se ha alcanzado una cobertura media en cada *pipeline* equivalente a la obtenida para el mismo a nivel de exoma completo (Tabla R6). Estas mediciones son coherentes con los resultados mostrados anteriormente del control del procesado de lecturas (Tablas R1 y R2), en los que debido a que en el *pipeline* del BiER se aplican filtros adicionales, el número de lecturas con las que se realizó la identificación de variantes es inferior con respecto al CNAG. El porcentaje de bases leídas disminuye drásticamente a partir de una cobertura 30x en el *pipeline* del BiER, en el que esta profundidad de secuenciación sólo ha sido conseguida en un 50% de las bases.

Se analizó con mayor detalle la cobertura en cada uno de los genes asociados a neuropatías de interés en ambos *pipelines*. En la Figura R3 se muestran *boxplots* que permiten comparar la cobertura media obtenida entre las 8 muestras en un conjunto representativo del total de genes asociados a neuropatías de interés. En ellos, se observa que la distribución de coberturas por gen es similar en ambos *pipelines*, siendo el rango de valores entre los que se sitúan las coberturas medias en el *pipeline* del CNAG mayor que en el BiER. Además, se detectan amplias diferencias de cobertura media entre las muestras en aquellos genes localizados en el cromosoma X, como *GJB1* o *ASH1*, debido a la diferencia de dosis entre hombres y mujeres. Se generaron tablas y representaciones gráficas de la cobertura alcanzada en cada intervalo así como *boxplots* con el resto de genes analizados, las cuales pueden ser consultadas en la siguiente web: http://bioinfo.cipf.es/bierwiki/espinos/exome_analysis.

3. Variantes candidatas identificadas en genes asociados a neuropatías

De las variantes obtenidas en cada *pipeline* de procesado de datos de secuenciación de exoma utilizado, se rescataron en cada muestra las variantes que se localizaban en los genes asociados a neuropatías de interés. Se aplicaron distintos criterios de filtrado y priorización para intentar identificar posibles mutaciones causantes de la neuropatía, y se lograron identificar 11 cambios en 6 genes (Tabla R7).

Seis variantes localizadas en los genes *PLEKHG5* y *PNPLA6* se excluyeron por tratarse de falsos positivos ya que no superaron la validación mediante secuenciación por Sanger.

En cuatro de los ocho casos, se identificaron y validaron variantes localizadas en los genes *FIG4*, *BICD2*, *SOD1* y *DAO* (Tabla R7). Estas mutaciones resultan en un cambio de aminoácido en la proteína para la que codifican o afectan al *splicing* del mRNA, y fueron obtenidas en ambos *pipelines*.

Por último, en dos pacientes no se detectaron cambios en los genes relacionados con neuropatías que cumplieran los criterios de filtrado y priorización establecidos.

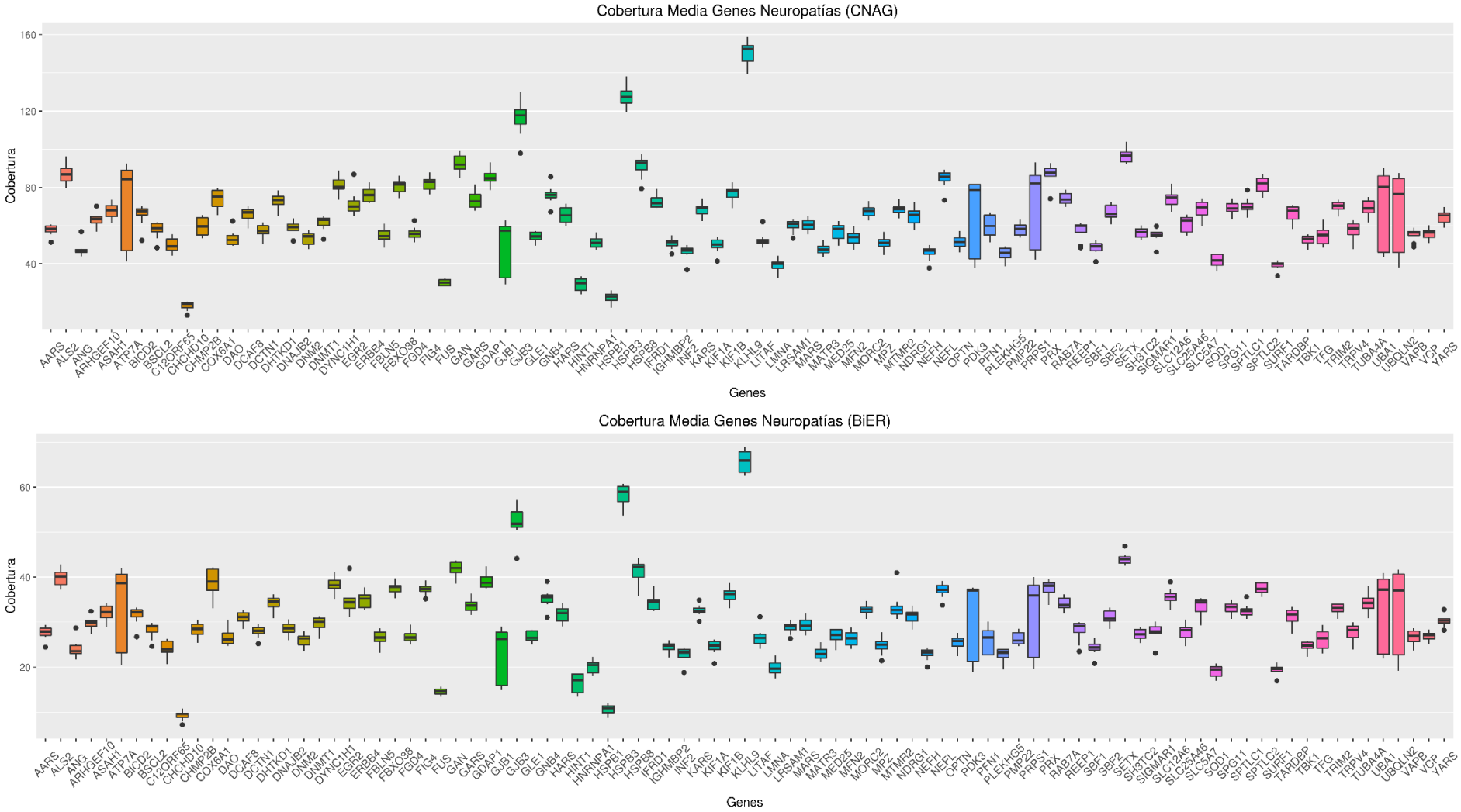

Figura R3: Distribución de la cobertura media obtenida en los *pipelines* del CNAG y BiER en un grupo representativo del total de genes asociados a neuropatías de interés.

Tabla R7: Variantes candidatas a ser la mutación causal identificadas en genes asociados a neuropatías.

Muestra	Gen	Herencia	cDNA	Proteína	*Pipeline* BiER	*Pipeline* CNAG	Validación Sanger
SGT038	*FIG4*	Recesiva	c.122T>C	p.I41T	Sí	Sí	Positivo
			c.446+5G>C	-	Sí	Sí	Positivo
SGT077	*PLEKHG5*	Recesiva	c.1853A>C	p.V618G	Sí	No	Negativo
			c.1856A>C	p.V619G	Sí	No	Negativo
			c.1880A>C	p.V627G	Sí	No	Negativo
			c.1883A>C	p.V628G	Sí	No	Negativo
SGT161	No se identificaron variantes candidatas en genes relacionados con neuropatías						
SGT187	No se identificaron variantes candidatas en genes relacionados con neuropatías						
SGT230	*BICD2*	Dominante	c.14G>C	p.S5W	Sí	Sí	Positivo
SGT238	*SOD1*	Dominante	c.65A>G	p.E22G	Sí	Sí	Positivo
SGT241	*DAO*	Dominante	c.958T>C	p.W320R	Sí	Sí	Positivo
SGT274	*PNPLA6*	Recesiva	c.2098-2A>G	-	Sí	No	Negativo
			c.2099T>G	p.V700G	Sí	No	Negativo

3.1 Variantes validadas previamente descritas

Tres cambios validados mediante secuenciación por Sanger habían sido previamente descritos en la bibliografía y/o en bases de datos (Tabla R8).

Tabla R8: Variantes validadas previamente descritas en la bibliografía y/o bases de datos de variantes.

Muestra	Clasificación clínica	Gen	Herencia	HGMD/dbSNP	Variante
SGT038	CMT	*FIG4*	Recesiva	CM073163	c.122T>C (p.I41T)
				rs370857139	c.446+5G>C
SGT238	dHMN	*SOD1*	Dominante	CM961342	c.65A>G (p.E22G)

El paciente SGT038 resultó ser heterocigoto compuesto para c.122T>C y c.446+5G>C, cambios que se encuentran en el exón 2 e intrón 4, respectivamente, del gen *FIG4* (Tabla R8). c.122T>C conduciría a un cambio aminoacídico de una isoleucina a una timina en la posición 41 de la proteína (p.I41T), y está anotada como patológica en la base de datos HGMD. c.446+5G>C se localiza en la secuencia consenso 5' de *splicing* del intrón 4 (*splicing donor site*), y está anotada en la base de datos dbSNP, sin indicación de asociación alguna con fenotipo clínico.

Al estar localizada esta segunda variante en una región que participa en el *splicing* del mRNA, se realizaron estudios *in silico* para evaluar las posibles consecuencias de este cambio (Tabla R9). De los cuatro programas utilizados, dos de ellos (NetGene2 y Human Splicing Finder) no reconocen el sitio de *splicing* en presencia del cambio, mientras que los otros dos (NNSplice y SpliceView) siguen reconociendo el sitio en presencia del cambio pero con una menor confianza.

Tabla R9: Estudio *in silico* de *FIG4* c.446+5G>C identificado en el paciente SGT038.

	NetGene2	**Human Splicing Finder**	**NNSplice**	**SpliceView**
c.446+5G>C *(Splicing donor site)*	*Donor site* original no reconocido	*Donor site* original no reconocido	*Score donor site* original disminuye	*Score donor site* original disminuye

Splicing donor site: secuencia consenso 5' del intrón, *Score*: puntuación.

Para confirmar los resultados de estas predicciones, se realizó un ensayo *in vitro* basado en la utilización de construcciones de minigenes, que permitió estudiar cómo se comporta la maquinaria de *splicing* celular en presencia y ausencia de este cambio. Los resultados de este ensayo, recogidos en la Figura R3, han mostrado que del transcrito generado a partir de la construcción con la secuencia de referencia para el *splicing donor site* del exón 4 (pSPL3::FIG4-E4-WT), se obtiene una única banda que incluye este exón. Sin embargo, del transcrito generado a partir de la construcción con el cambio c.446+5G>C (pSPL3::FIG4-E4-Mut), se obtiene una banda de menor tamaño correspondiente con un procesado incorrecto del mRNA que no incluye el exón 4 (Figura R3 A y B). La secuenciación mediante Sanger de ambas bandas de cDNA ha confirmado el *skipping* del exón 4 en presencia del cambio (Figura R3 B). Este incorrecto procesamiento del mRNA de *FIG4* conduciría a un desplazamiento de la pauta de lectura y a la generación de un codón de parada temprano, por lo que la traducción de este transcrito resultaría en una proteína truncada (p.F98IfsX16) (Figura R3 C).

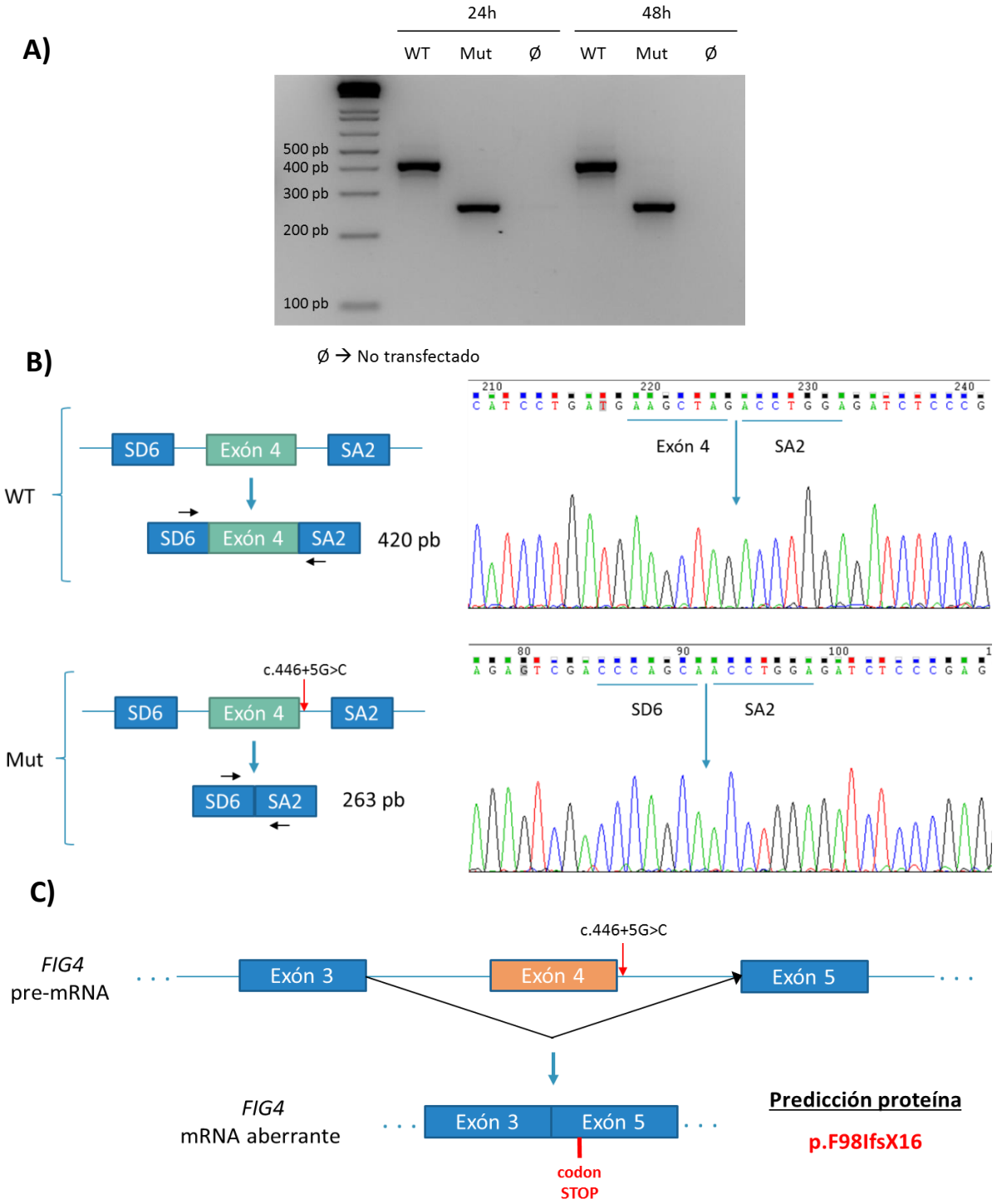

Figura R4: Ensayo *in vitro* de la variante de *splicing* c.446+5G>C identificada en *FIG4* en el paciente SGT038. A) Productos de la amplificación por PCR con los *primers* SD6 y SA2 de los transcritos generados tras la expresión de las construcciones de minigenes pSPL3::FIG4-E4-WT y Mut, a las 24 y 48 horas post-transfección, en células HeLa. El producto de 420 pb corresponde al correcto procesamiento del mRNA que incluye el exón 4, mientras que el producto de 263 pb corresponde a un *skipping* del exón 4 como consecuencia de la mutación c.446+5G>C. **B)** Esquema del procesado del mRNA de las construcciones de minigenes y electroferogramas de la secuenciación mediante Sanger de las bandas detectadas en A). **C)** Predicción del efecto que tendría el *skipping* del exón 4 a nivel del procesado del mRNA y de la proteína

Por último, en el paciente SGT238 se identificó en heterocigosis la variante c.65A>G, que se encuentra en el exón 1 del gen *SOD1* (Tabla R8). Esta sustitución conduciría a un cambio aminoacídico de una glutamina a una glicina en la posición 22 de la proteína (p.E22G), y está anotada como patológica en la base de datos HGMD. Se confirmó mediante secuenciación por Sanger que su hermana enferma era también portadora en heterocigosis de este cambio.

3.2 Variantes validadas noveles

Se identificaron y validaron dos variantes noveles mediante secuenciación por Sanger (Tabla R10).

Tabla R10: Variantes validadas noveles.

Muestra	Clasificación clínica	Gen	Herencia	Variante	Polyphen-2	SIFT	PROVEAN	PhyloP	GERP
SGT230	dHMN	*BICD2*	Dominante	c.14G>C (p.S5W)	0,972	0,009	-0,671	1,05	4,35
SGT241	dHMN	*DAO*	Dominante	c.958T>C (p.W320R)	0,992	0,26	-7,766	1,92	4,87

Polyphen-2: 1 (probablemente deletérea), 0 (benigna); SIFT: 0 (deletérea), > 0.02 (tolerada); PROVEAN: < -2,5 (deletérea); PhyloP: > 1 (región conservada), < 1 (región de rápida evolución); GERP: > 2 (región conservada).

En el paciente SGT230 se identificó en heterocigosis la variante c.14G>C en el gen *BICD2* (Tabla R10), que conduciría a un cambio aminoacídico de una serina a un triptófano en la posición 5 de la proteína (p.S5W). El alineamiento múltiple de las secuencias aminoacídicas de homólogos en mamíferos cercanos que se recoge en la Figura R3, demuestra que esta posición se encuentra conservada en dichas especies así como en el organismo modelo *Drosophila melanogaster*.

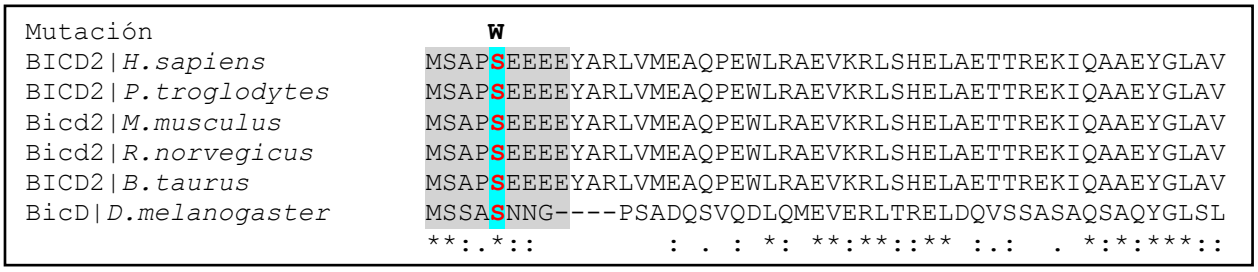

Figura R5: Fragmento del alineamiento múltiple de la secuencia aminoacídica de homólogos de la proteína BICD2 en diferentes organismos. Se muestra en azul y rojo la conservación de la posición en la que se encuentra el cambio c.14G>C (p.S5W) detectado en el paciente SGT230, y en gris los aminoácidos que lo rodean.

En el paciente SGT241 se identificó en heterocigosis la variante c.958T>C en el gen *DAO,* que conduciría a un cambio aminoacídico de un triptófano a una arginina en la posición 320 de la proteína. El alineamiento múltiple de las secuencias aminoacídicas de homólogos en mamíferos cercanos así como en los organismos modelo *Caenorhabditis elegans* y *Drosophila melanogaster* muestra que la posición en la que se localiza el cambio y su entorno se encuentran conservados (Figura R4).

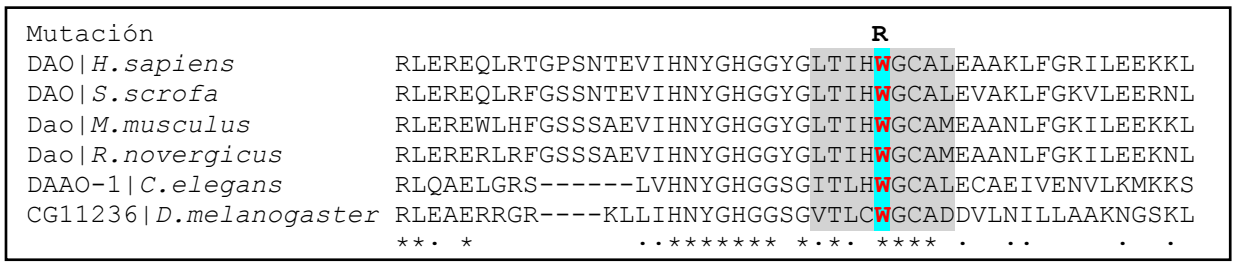

Figura R6: Fragmento del alineamiento múltiple de la secuencia aminoacídica de homólogos de la proteína DAO en diferentes organismos. Se muestra en azul y rojo la conservación de la posición en la que se encuentra el cambio c.958T>C (p.W320R) detectado en el paciente SGT238, y en gris los aminoácidos que lo rodean.

Al no contar con evidencias previas sobre el efecto que las mutaciones c.14G>C (p.S5W) y c.958T>C (p.W320R) identificadas en *BICD2* y *DAO* respectivamente podrían causar sobre la función de la proteína, se realizaron estudios *in silico* para evaluar su posible patogenicidad, cuyos resultados se indican en la Tabla R10. Los predictores de conservación consultados concluyen que son cambios localizados en regiones de evolución restringida (puntuación en GERP superior a 2), y existe cierta conservación de estos residuos en vertebrados (puntuación en PhyloP superior a 1). La evaluación de patogenicidad realizada con los predictores PolyPhen-2 y SIFT sugieren que la variante identificada en *BICD2* tendría un efecto deletéreo sobre la proteína, mientras que para el cambio identificado en *DAO* son las predicciones de PolyPhen-2 y PROVEAN las que estiman este mismo efecto.

DISCUSIÓN

En este trabajo, se ha aplicado WES en ocho pacientes no relacionados diagnosticados de padecer CMT/dHMN, con el fin de resolver el diagnóstico genético en ellos. Los datos generados en la secuenciación han sido procesados de forma simultánea por los *pipelines* implementados en el CNAG y en el BiER. Ambos *pipelines* utilizan diferentes programas para alinear las lecturas frente al genoma de referencia, y siguen diferentes planteamientos para realizar la identificación de variantes. En el *pipeline* del BiER se utiliza la herramienta UnifiedGenotyper de GATK para centrar la búsqueda de variantes en las regiones de exoma capturadas por el kit empleado. En cambio, en el *pipeline* del CNAG, se detectan variantes utilizando el programa SAMtools en todas aquellas regiones del genoma de referencia sobre las que se han conseguido alinear lecturas. La concordancia de variantes obtenidas entre ambas aproximaciones se encuentra alrededor del 45%, que es un porcentaje relativamente bajo, probablemente debido a las diferentes aproximaciones que siguen ambos *pipelines* para el mapeo de lecturas e identificación de variantes. En estudios realizados por otros autores en los que comparan resultados obtenidos por diferentes *pipelines,* en los que utilizan estos mismos programas para la identificación de variantes*,* la concordancia era inferior al 60% (Liu *et al*., 2013; O'Rawe *et al.,* 2013; Yu y Sun, 2013).

La cobertura alcanzada en un experimento de secuenciación de exoma es crucial para la determinación de variantes, especialmente para la correcta detección de cambios en heterocigosis, ya que pueden ser causantes de enfermedades que siguen un patrón de herencia autosómico dominante así como recesivo cuando están presentes en heterocigosis compuesta. La elección de algoritmo para el alineamiento de lecturas frente al genoma de referencia puede influir sobre la cobertura con la que finalmente se realizará la identificación de variantes. Esto es debido a que existen algoritmos que resultan más eficientes para alinear lecturas en regiones complejas (ricas en motivos de repetición, que presentan inserciones y/o deleciones u otro tipo de reordenamiento) o que presentan un número alto de desapareamientos, en su mayoría debido a errores de secuenciación (Sims *et al.*, 2014). El porcentaje de lecturas mapeadas en el *pipeline* del BiER ha sido mayor, pero debido a que se realizaron filtrados adicionales además del habitual eliminado de lecturas duplicadas, la cobertura media resultante en este *pipeline* con la que se ha realizado la identificación de variantes es inferior (<30x) con respecto al

pipeline del CNAG (> 60x). Son cientos los estudios de secuenciación de exoma con fines diagnósticos publicados que establecen una cobertura media 30x con mínimo un 80% de las bases leídas 10 veces como el estándar para la detección de variantes (Zhou *et al*., 2012; Thauvin-Robinet *et al.*, 2013; Yu *et al.,* 2013). Con ambos procesados se cumplirían estos estándares, siendo además el porcentaje de bases con cobertura 10x superior al 90%. Es posible que la aplicación de filtrados adicionales sobre las lecturas mapeadas en el *pipeline* del BiER, que suponen una bajada de la cobertura, resulte en una pérdida de secuencias que potencialmente podrían aportar información para la identificación de variantes. Dado que se realizan realineamientos locales para optimizar la identificación de variantes en las regiones de exoma capturadas por el kit empleado, sería conveniente valorar si compensa la bajada de cobertura, principalmente debida a filtrar por calidad de mapeo, frente a la información que se podría obtener de las lecturas descartadas.

A partir de las variantes obtenidas en los dos *pipelines*, se han investigado únicamente aquellas localizadas en genes implicados en el desarrollo de neuropatías como aproximación inicial para identificar, de forma efectiva, la mutación causal de la enfermedad. De esta forma, tras aplicar criterios de filtrado y priorización, de los ocho casos estudiados en este trabajo, se han logrado identificar en ambos *pipelines* variantes patológicas ya descritas en dos de ellos así como cambios noveles en otros dos.

En los pacientes SGT038 y SGT238 se han identificado mutaciones clínicas ya conocidas en los genes *FIG4* y *SOD1,* respectivamente. El cambio c.122T>C (p.I41T) detectado en *FIG4* ha sido ampliamente descrito como mutación asociada a CMT4J, que se caracteriza por presentar un patrón de herencia autosómico recesivo y un cuadro clínico severo, con un debut temprano de los síntomas (Chow *et al*., 2007; Nicholson *et al*., 2011; Menezes *et al*., 2014). En todos los casos publicados, los pacientes presentan este cambio en heterocigosis compuesta con otro cambio que resulta en un alelo nulo, bien porque da lugar a una proteína truncada o porque se trata de una deleción de parte del gen. En el paciente SGT038, la segunda variante identificada en *FIG4,* c.446+5G>C, se localiza en la secuencia consenso 5' de *splicing* del intrón 4 (*splicing donor site*). Los estudios *in silico* realizados han revelado que la maquinaria de *splicing* podría no ser capaz de reconocer este sitio en presencia del cambio. Para corroborar estas predicciones, se ha realizado un

ensayo *in vitro* utilizando construcciones de minigenes, el cual ha demostrado que en presencia del cambio c.446+5G>C tendría lugar una deleción (*skipping*) del exón 4 en el mRNA del gen *FIG4*. Esto implicaría el desplazamiento de la pauta de lectura y la aparición de un codón de parada temprano, por tanto, se generaría una proteína truncada (p.F98IfsX16). Consecuentemente, se trataría de una variante deletérea que modificaría el *splicing* del mRNA. La variante c.65A>G (p.E22G) detectada en *SOD1* ha sido descrita en una extensa familia española como causante de ALS, y pertenece al grupo de mutaciones descritas en este gen que se asocian a un fenotipo clínico más leve, con un debut tardío de los síntomas y progresión lenta de la enfermedad (Syriani *et al.*, 2009; Zou *et al*., 2016). Mutaciones en *SOD1* causantes de esta patología presentan tanto un patrón de herencia autosómico dominante como recesivo, y explican el 20% de casos de ALS familiares y el 4% de casos esporádicos (Chen *et al.,* 2013). Con el estudio llevado a cabo en este trabajo en el paciente SGT238, ha sido posible resolver el diagnóstico genético de la familia, ya que su hermana enferma también es portadora de la misma mutación. La patogenicidad de las variantes c.122T>C (p.I41T) en *FIG4* y c.65A>G (p.E22G) en *SOD1* está suficientemente demostrada al haber sido descritas en otros pacientes y haber sido realizados diversos estudios funcionales para comprender el mecanismo de la enfermedad (Chow *et al.,* 2007; Syriani *et al.,* 2009; Nicholson *et al.,* 2011; Menezes *et al.,* 2014). En cambio, no se tienen evidencias clínicas sobre el efecto deletéreo que podría causar la variante c.446+5G>C identificada en *FIG4.* Por ello, sería necesario analizar mRNA del paciente para confirmar los resultados obtenidos en el ensayo *in vitro* realizado.

En los pacientes SGT230 y SGT241 se identificaron variantes noveles en heterocigosis en los genes *BICD2* y *DAO*, respectivamente. Mutaciones con patrón de herencia autosómico dominante en *BICD2* se han asociado a formas de SMA, caracterizadas por presentar un comienzo temprano de la enfermedad y debilidad muscular predominante en las extremidades inferiores (Rossor *et al.*, 2015). *BICD2* codifica para una proteína que funciona como adaptador del complejo motor del citoesqueleto constituido por la proteína dineína que está implicado en el transporte axonal (Jerath y Shy, 2014). Presenta tres dominios *coiled-coil* altamente conservados al que se unen otras proteínas así como moléculas que son transportadas por este complejo. Las mutaciones descritas hasta

ahora en BICD2 se concentran en dichos dominios conservados (Rossor *et al.*, 2015), sin embargo, la variante c.14G>C (p.S5W) identificada en el paciente SGT230 se encuentra localizada en el extremo N-terminal. En cuanto al gen *DAO*, únicamente ha sido descrito un caso de ALS familiar causado por una mutación con patrón de herencia autosómico dominante, en la que manifestaciones de la enfermedad habían dado comienzo en la edad adulta (Mitchell *et al*., 2010). La proteína DAO actúa como una oxidasa dependiente de FAD encargada de la degradación del aminoácido D-serina, y se ha demostrado que este aminoácido se encuentra en exceso en la médula espinal en pacientes con ALS y en un modelo de ratón de la enfermedad, lo que provoca la degeneración de las motoneuronas (Sasabe *et al*., 2012). Esta proteína presenta distintos dominios de unión a FAD; la mutación identificada en el paciente SGT241 se localiza en el dominio C-terminal, mientras que la posición de la mutación descrita por Mitchell y colaboradores (2010) es colindante al dominio central. Se han realizado estudios *in silico* para evaluar el daño que podrían causar sobre la función de la proteína las variantes identificadas en *BICD2* y *DAO*, cuyos resultados sugieren que puedan causar un efecto deletéreo. En ambos casos, para corroborar que se trata de mutaciones clínicas, sería necesario realizar estudios de segregación en familiares sanos o enfermos para determinar si las variantes cosegregan con la enfermedad así como posteriormente y en la medida de lo posible, estudios funcionales que esclarezcan el mecanismo de enfermedad.

Independientemente del planteamiento y programas utilizados en los dos *pipelines*, siguiendo las mismas pautas para el filtrado y priorización, las mutaciones candidatas identificadas y validadas localizadas en genes asociados a neuropatías de interés han sido detectadas igualmente por ambos. Por una parte, respecto al *pipeline* del BiER, el hecho de haber obtenido una cobertura inferior no ha sido un problema para identificar estos cambios nucleotídicos como verdaderos positivos. En cambio, se ha apreciado una mayor tendencia en la detección de falsos positivos, ya que las variantes candidatas detectadas exclusivamente con el *pipeline* del BIER tras filtrar y priorizar eran de baja calidad y no se validaron finalmente mediante Sanger. Por otra parte, en cuanto al *pipeline* del CNAG, aunque la cobertura alcanzada y el número de variantes detectadas ha sido superior, no ha aportado variantes distintas a las obtenidas en el *pipeline* del BiER que cumplieran los criterios de filtrado y priorización establecidos para seleccionar posibles variantes

causales. Consecuentemente, pese a las diferencias existentes entre ambos *pipelines* en los procesos de filtrado, los resultados válidos se han logrado con los dos sistemas.

En los casos en los que no se ha logrado identificar posibles variantes causales en genes asociados previamente a neuropatías, puesto que se ha llevado a cabo un análisis más detallado de la cobertura obtenida en estas regiones, sería interesante examinar cuántas de ellas presentan una cobertura inferior a 10x. Podría ser que la mutación causal se localice en alguna de esas regiones y no ha sido posible su correcta detección. Hay que tener en cuenta que la cobertura en secuenciación de exoma es irregular, debido a que hay regiones que resultan más enriquecidas que otras durante la captura y preparación de la librería. Este proceso se encuentra sesgado por diferentes factores por los que todas las regiones no resultan igualmente representadas, como el contenido en GC o la presencia de repeticiones (Ku *et al*, 2011; Rehm, 2013; Sims *et al*., 2014). El perfil de coberturas varía entre kits de captura de exoma, siendo los diseños que presentan una mayor densidad de sondas de captura los que resultan en una cobertura más uniforme y, por tanto, mayor sensibilidad para la detección de variantes. Con el diseño de paneles de genes en los que se incluyeran genes implicados en neuropatías de interés, se podría optimizar el diseño de sondas para las regiones complejas, de forma que se podría estudiar con una mejor cobertura si en estos pacientes realmente la mutación clínica se localiza en ellos. Suponiendo que en dichos casos la causa de la enfermedad no se debe a mutación en genes conocidos, los datos obtenidos en este trabajo son un punto de partida para la caracterización de nuevos genes implicados en neuropatías. Para ello, se tendrían que reanalizar los datos de WES conjuntamente con los de otros casos, familiares o no, con el mismo o parecido cuadro clínico y que, consecuentemente, podrían compartir un origen genético común. Además, estos datos pueden volver a ser reanalizados conforme sean descritos nuevos genes con el propósito de averiguar si presentan posibles variantes patológicas en estos genes noveles. Sin embargo, el WES tiene también limitaciones en su diseño y por tanto, si la causa genética de la enfermedad se localiza en una región reguladora y/o intrónica profunda, sería preciso llevar a cabo un estudio por WGS para su detección.

CONCLUSIONES

1. El procesamiento de datos de secuenciación de exoma, con los *pipelines* implementados en el BiER y en el CNAG, de ocho pacientes no relacionados diagnosticados de CMT/dHMN y posterior análisis de variantes localizadas en genes implicados en neuropatías, ha permitido identificar y validar cambios candidatos a ser la mutación causal en el 50% de los casos.

2. A pesar de las diferencias de cobertura obtenidas en los *pipelines* implementados en el BiER y en el CNAG así como los diferentes procedimientos que utilizan para la identificación de variantes, las mutaciones candidatas validadas localizadas en genes implicados en los grupos de neuropatías objeto de estudio han sido detectadas por ambos.

3. En el paciente SGT038 se han identificado en heterocigosis compuesta las variantes c.122T>C (p.I41T) y c.446+5G>C en el gen *FIG4* como candidatas a ser las mutaciones patológicas. La mutación p.I41T ha sido ampliamente descrita como causante de CMT4J, y los estudios *in silico* realizados predicen que la variante c.446+5G>C conduciría a una alteración en el *splicing* del mRNA.

4. El ensayo *in vitro* basado en el análisis de construcciones de minigenes ha confirmado que, en presencia de la variante de *splicing* c.446+5G>C en *FIG4,* tiene lugar un procesado incorrecto del mRNA en el que no se incluiría el exón 4, lo que resultaría en una proteína truncada (p.F98IfsX16). Estos resultados corroboran las predicciones *in silico* obtenidas para esta variante, pero sería conveniente analizar el mRNA del paciente para validarlo.

5. En el paciente SGT238 se ha identificado en heterocigosis la variante c.65A>G (p.E22G) en el gen *SOD1*. Esta mutación ha sido previamente descrita como causante de un fenotipo más leve de ALS. Se ha confirmado que su hermana enferma es también portadora de este cambio.

6. En el paciente SGT230 se ha identificado en heterocigosis la variante c.14G>C (p.S5W) en el gen *BICD2,* y en el paciente SGT241 en heterocigosis la variante c.958T>C (p.W320R) en el gen *DAO*. Ambos cambios son noveles. Aunque los estudios *in silico* sugieren que podrían ser variantes patológicas, es necesario

realizar estudios de segregación así como ensayos funcionales para concluir que realmente son la causa de la neuropatía.

7. En cuatro de los ocho casos estudiados no ha sido posible identificar y/o validar posibles variantes causales en genes implicados en CMT/dHMN y otras neuropatías relacionadas. Los datos obtenidos en dichos casos resultan útiles para la búsqueda de nuevos genes asociados a neuropatías, así como para la mejora de herramientas diagnósticas basadas en panel de genes, lo que permitiría profundizar en el conocimiento de las bases moleculares de estos trastornos y ampliaría el espectro de las manifestaciones clínicas asociadas a estos genes.

BIBLIOGRAFÍA

Adzhubei, I., Schmidt, S., Peshkin, L., Ramensky, V., Gerasimova, A., Bork, P., … Kondrashov, A. S. (2000). A method and server for predicting damaging missense mutations. *Nature Methods*, *37*(16), 3133–3164.

Aleman, A., Garcia-Garcia, F., Salavert, F., Medina, I., & Dopazo, J. (2014). A web-based interactive framework to assist in the prioritization of disease candidate genes in whole-exome sequencing studies. *Nucleic Acids Research*, *42*(W1), W88–W93.

Andersen, P. M., & Al-Chalabi, A. (2011). Clinical genetics of amyotrophic lateral sclerosis: what do we really know? *Nature Reviews Neurology*, *7*(11), 603–615.

Baets, J., De Jonghe, P., & Timmerman, V. (2014). Recent advances in Charcot-Marie-Tooth disease. *Current Opinion in Neurology*, *27*(5), 532–40.

Bamshad, M. J., Ng, S. B., Bigham, A. W., Tabor, H. K., Emond, M. J., Nickerson, D. a., & Shendure, J. (2011). Exome sequencing as a tool for Mendelian disease gene discovery. *Nature Reviews Genetics*, *12*(11), 745–755.

Bettens, K., Sleegers, K., & Van Broeckhoven, C. (2013). Genetic insights in Alzheimer's disease. *The Lancet Neurology*, *12*(1), 92–104.

Bottillo, I., De Luca, A., Schirinzi, A., Guida, V., Torrente, I., Calvieri, S., … Dallapiccola, B. (2007). Functional analysis of splicing mutations in exon 7 of NF1 gene. *BMC Medical Genetics*, *8*, 4.

Boycott, K. M., Vanstone, M. R., Bulman, D. E., & MacKenzie, A. E. (2013). Rare-disease genetics in the era of next-generation sequencing: discovery to translation. *Nature Reviews Genetics*, *14*(10), 681–691.

Chen, S., Sayana, P., Zhang, X., & Le, W. (2013). Genetics of amyotrophic lateral sclerosis: an update. *Molecular Neurodegeneration*, *8*(1), 28.

Choi, Y., & Chan, A. P. (2015). PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. *Bioinformatics*, *31*(16), 2745–2747.

Chong, J. X., Buckingham, K. J., Jhangiani, S. N., Boehm, C., Sobreira, N., Smith, J. D., … Bamshad, M. J. (2015). The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities. *The American Journal of Human Genetics*, *97*(2), 199–215.

Chow, C. Y., Zhang, Y., Dowling, J. J., Jin, N., Adamska, M., Shiga, K., … Meisler, M. H. (2007). Mutation of FIG4 causes neurodegeneration in the pale tremor mouse and patients with CMT4J. *Nature*, *448*(7149), 68–72.

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., … Lu, X. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly*, *6*(2), 1–13.

Combarros, O., Calleja, J., Polo, J., & Berciano, J. (1987). Prevalence of hereditary motor and sensory neuropathy in Cantabria. *Acta Neurol Scand*, *75*, 9–12.

Cooper, G. M., Goode, D. L., Ng, S. B., Sidow, A., Bamshad, M. J., Shendure, J., & Nickerson, D. a. (2010). Single-nucleotide evolutionary constraint scores highlight disease-causing mutations. *Nature Methods*, *7*(4), 250–251.

Coutelier, M., Stevanin, G., & Brice, A. (2015). Genetic landscape remodelling in spinocerebellar ataxias: the influence of next-generation sequencing. *Journal of Neurology*, *262*(10), 2382–2395.

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V, Maguire, J. R., Hartl, C., … Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, *43*(5), 491–8.

Desviat, L. R., Pérez, B., & Ugarte, M. (2012). Minigenes to Confirm Exon Skipping Mutations. In *Exon Skipping: Methods and Protocols* (Vol. 531, pp. 37–47).

Dierick, I., Baets, J., Irobi, J., Jacobs, A., De Vriendt, E., Deconinck, T., … Timmerman, V. (2007). Relative contribution of mutations in genes for autosomal dominant distal hereditary motor neuropathies: a genotype-phenotype correlation study. *Brain*, *131*(5), 1217–1227.

Echaniz-Laguna, a. (2015). The shifting paradigm of Charcot-Marie-Tooth disease. *Revue Neurologique*, 3–9.

Farrar, M. a, & Kiernan, M. C. (2014). The Genetics of Spinal Muscular Atrophy: Progress and Challenges. *Neurotherapeutics : The Journal of the American Society for Experimental NeuroTherapeutics*, *12*(2), 290–302.

Gilissen, C., Hoischen, A., Brunner, H. G., & Veltman, J. a. (2011). Unlocking Mendelian disease using exome sequencing. *Genome Biology*, *12*(9), 228.

Gilissen, C., Hoischen, A., Brunner, H. G., & Veltman, J. a. (2012). Disease gene identification strategies for exome sequencing. *European Journal of Human Genetics*, *20*(5), 490–497.

Jerath, N. U., & Shy, M. E. (2014). Hereditary motor and sensory neuropathies: Understanding molecular pathogenesis could lead to future treatment strategies. *Biochimica et Biophysica Acta*, *1852*(4), 667–678.

Kinsella, R. J., Kähäri, A., Haider, S., Zamora, J., Proctor, G., Spudich, G., … Flicek, P. (2011). Ensembl BioMarts: A hub for data retrieval across taxonomic space. *Database*, *2011*, 1–9.

Ku, C. S., Naidoo, N., & Pawitan, Y. (2011). Revisiting Mendelian disorders through exome sequencing. *Human Genetics*, *129*(4), 351–370.

Li, H. (2011a). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, *27*(21), 2987–2993.

Li, H. (2011b). Tabix: Fast retrieval of sequence features from generic TAB-delimited files. *Bioinformatics*, *27*(5), 718–719.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079.

Linnemann, C., Tezenas du Montcel, S., Rakowicz, M., Schmitz-Hubsch, T., Szymanski, S., Berciano, J., … Schols, L. (2015). Peripheral Neuropathy in Spinocerebellar Ataxia Type 1, 2, 3, and 6. *Cerebellum*.

Liu, X., Han, S., Wang, Z., Gelernter, J., & Yang, B. Z. (2013). Variant Callers for Next-Generation Sequencing Data: A Comparison Study. *PLoS ONE*, *8*(9), 1–11.

Liu, Y.-T., Lee, Y.-C., & Soong, B.-W. (2015). What we have learned from the next generation sequencing: contributions to the genetic diagnoses and understanding of pathomechanisms of neurodegenerative diseases. *Journal of Neurogenetics*, *7063*, 1–26.

Marangi, G., & Traynor, B. J. (2015). Genetic causes of amyotrophic lateral sclerosis: New genetic analysis methodologies entailing new opportunities and challenges. *Brain Research*, *1607*, 75–93.

Marco-Sola, S., Sammeth, M., Guigó, R., & Ribeca, P. (2012). The GEM mapper: fast, accurate and versatile alignment by filtration. *Nature Methods*, *9*(12), 1185–8.

Mathis, S., Goizet, C., Tazir, M., Magdelaine, C., Lia, A.-S., Magy, L., & Vallat, J.-M. (2015). Charcot–Marie–Tooth diseases: an update and some new proposals for the classification. *Journal of Medical Genetics*, jmedgenet–2015–103272.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., … DePristo, M. A. (2009). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*, 254–260.

McLaren, W., Pritchard, B., Rios, D., Chen, Y., Flicek, P., & Cunningham, F. (2010). Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics*, *26*(16), 2069–2070.

Medina, I., De Maria, A., Bleda, M., Salavert, F., Alonso, R., Gonzalez, C. Y., & Dopazo, J. (2012). VARIANT: Command Line, Web service and Web interface for fast and accurate functional characterization of variants found by Next-Generation Sequencing. *Nucleic Acids Research*, *40*(W1), 54–58.

Menezes, M. P., Waddell, L., Lenk, G. M., Kaur, S., MacArthur, D. G., Meisler, M. H., & Clarke, N. F. (2014). Whole exome sequencing identifies three recessive FIG4 mutations in an apparently dominant pedigree with Charcot–Marie–Tooth disease. *Neuromuscular Disorders*, *24*(8), 666–670.

Mitchell, J., Paul, P., Chen, H.-J., Morris, A., Payling, M., Falchi, M., … de Belleroche, J. (2010). Familial amyotrophic lateral sclerosis is associated with a mutation in D-amino acid oxidase. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(16), 7556–61.

Nanetti, L., Cavalieri, S., Pensato, V., Erbetta, A., Pareyson, D., Panzeri, M., … Mariotti, C. (2013). SETX mutations are a frequent genetic cause of juvenile and adult onset cerebellar ataxia with neuropathy and elevated serum alpha-fetoprotein. *Orphanet Journal of Rare Diseases*, *8*(1), 1.

Ng, P. C., & Henikoff, S. (2001). Predicting Deleterious Amino Acid Substitutions Predicting Deleterious Amino Acid Substitutions, 863–874.

Ng, S. B., Buckingham, K. J., Lee, C., Bigham, A. W., Tabor, H. K., Dent, K. M., … Bamshad, M. J. (2010). Exome sequencing identifies the cause of a mendelian disorder. *Nature Genetics*, *42*(1), 30–35.

Nicholson, G., Lenk, G. M., Reddel, S. W., Grant, A. E., Towne, C. F., Ferguson, C. J., … Meisler, M. H. (2011). Distinctive genetic and clinical features of CMT4J: A severe neuropathy caused by mutations in the PI(3,5)P2 phosphatase FIG4. *Brain*, *134*(7), 1959–1971.

O'Rawe, J., Jiang, T., Sun, G., Wu, Y., Wang, W., Hu, J., … Lyon, G. J. (2013). Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. *Genome Medicine*, *5*(3), 28.

Pirooznia, M., Kramer, M., Parla, J., Goes, F. S., Potash, J. B., McCombie, W., & Zandi, P. P. (2014). Validation and assessment of variant calling pipelines for next-generation sequencing. *Human Genomics*, *8*(1), 14.

Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R., & Siepel, A. (2010). Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Research*, *20*(1), 110–121.

Rabbani, B., Mahdieh, N., Hosomichi, K., Nakaoka, H., & Inoue, I. (2012). Next-generation sequencing: impact of exome sequencing in characterizing Mendelian disorders. *Journal of Human Genetics*, *57*(10), 621–632.

Rehm, H. L. (2013). Disease-targeted sequencing: a cornerstone in the clinic. *Nature Reviews. Genetics*, *14*(4), 295–300.

Renton, A. E., Chiò, A., & Traynor, B. J. (2013). State of play in amyotrophic lateral sclerosis genetics. *Nature Neuroscience*, *17*(1), 17–23.

Rossor, a. M., Kalmar, B., Greensmith, L., & Reilly, M. M. (2012). The distal hereditary motor neuropathies. *Journal of Neurology, Neurosurgery & Psychiatry*, *83*(1), 6–14.

Rossor, A. M., Oates, E. C., Salter, H. K., Liu, Y., Murphy, S. M., Schule, R., … North, K. N. (2015). Phenotypic and molecular insights into spinal muscular atrophy due to mutations in BICD2. *Brain: A Journal of Neurology*, *138*(Pt 2), 293–310.

Rossor, A. M., Polke, J. M., Houlden, H., & Reilly, M. M. (2013). Clinical implications of genetic advances in Charcot-Marie-Tooth disease. *Nature Reviews. Neurology*, *9*(10), 562–71.

Sasabe, J., Miyoshi, Y., Suzuki, M., Mita, M., Konno, R., Matsuoka, M., … Aiso, S. (2012). D-Amino acid oxidase controls motoneuron degeneration through D-serine. *Proceedings of the National Academy of Sciences*, *109*(2), 627–632.

Schneider, S., & Tomás Brás, J. M. (2015). *Movement Disorder Genetics*. (S. A. Schneider & J. M. T. Brás, Eds.). Cham: Springer International Publishing.

Sims, D., Sudbery, I., Ilott, N. E., Heger, A., & Ponting, C. P. (2014). Sequencing depth and coverage: key considerations in genomic analyses. *Nature Reviews Genetics*, *15*(2), 121–32.

Skre, H. (1974). Genetic and clinical aspects of Charcot Marie Tooth's disease. *Clin Genet*, *6*, 98–118.

Syriani, E., Morales, M., & Gamez, J. (2009). The p.E22G mutation in the Cu/Zn superoxide-dismutase gene predicts a long survival time. Clinical and genetic characterization of a seven-generation ALS1 Spanish pedigree. *Journal of the Neurological Sciences*, *285*(1-2), 46–53.

Tarraga, J., Arnau, V., Martinez, H., Moreno, R., Cazorla, D., Salavert-Torres, J., … Medina, I. (2014). Acceleration of short and long DNA read mapping without loss of accuracy using suffix array. *Bioinformatics*, *30*(23), 3396–3398.

Thauvin-Robinet, C., Auclair, M., Duplomb, L., Caron-Debarle, M., Avila, M., St-Onge, J., … Rivière, J. B. (2013). PIK3R1 mutations cause syndromic insulin resistance with lipoatrophy. *American Journal of Human Genetics*, *93*(1), 141–149.

Timmerman, V., Clowes, V. E., & Reid, E. (2013). Overlapping molecular pathological themes link Charcot-Marie-Tooth neuropathies and hereditary spastic paraplegias. *Experimental Neurology*, *246*, 14–25.

Timmerman, V., Strickland, A., & Züchner, S. (2014). Genetics of Charcot-Marie-Tooth (CMT) Disease within the Frame of the Human Genome Project Success. *Genes*, *5*(1), 13–32.

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., … DePristo, M. A. (2013). *From fastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline*. *Current Protocols in Bioinformatics*.

Weis, J., & Senderek, J. (2014). Introduction to the hereditary neuropathies. *Peripheral Nerve Disorders*, 59–61.

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. Springer International Publishing.

Yu, T. W., Chahrour, M. H., Coulter, M. E., Jiralerspong, S., Okamura-Ikeda, K., Ataman, B., … Walsh, C. A. (2013). Using Whole-Exome Sequencing to Identify Inherited Causes of Autism. *Neuron*, *77*(2), 259–273.

Yu, X., & Sun, S. (2013). Comparing a few SNP calling algorithms using low-coverage sequencing data. *BMC Bioinformatics*, *14*(1), 274.

Zhou, Q., Lee, G. S., Brady, J., Datta, S., Katan, M., Sheikh, A., … Aksentijevich, I. (2012). A hypermorphic missense mutation in PLCG2, encoding phospholipase Cγ2, causes a dominantly inherited autoinflammatory disease with immunodeficiency. *American Journal of Human Genetics*, *91*(4), 713–720.

Zou, Z.-Y., Liu, M.-S., Li, X.-G., & Cui, L.-Y. (2016). H46R SOD1 mutation is consistently associated with a relatively benign form of amyotrophic lateral sclerosis with slow progression. *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration*, *8421*(June), 1–4.

ANEXO I

Listado de genes asociados a neuropatías de interés agrupados según el tipo de neuropatía en la que se encuentran comúnmente implicados

Neuropatía Hereditaria Sensitivo-Motora (HSMN)	*AARS*	*GALC*	*KIF5A*	*PEX1*	*SLC12A6*
	AIFM1	*GAN*	*LITAF*	*PEX7*	*SOX10*
	ARHGEF10	*GARS*	*LMNA*	*PHYH*	*SURF1*
	ARSA	*GDAP1*	*LRSAM1*	*PLA2G6*	*TFG*
	COX6A1	*GJB1*	*MARS*	*PLEKHG5*	*TRIM2*
	CTDP1	*GNB4*	*MED25*	*PMM2*	*TRPV4*
	DCAF8	*HARS*	*MFN2*	*PMP22*	*TUBB3*
	DNAJC3	*HINT1*	*MORC2*	*PRPS1*	*VCP*
	DHTKD1	*HK1*	*MPZ*	*PRX*	*YARS*
	DNM2	*IFRD1*	*MTMR2*	*RAB7A*	
	EGR2	*INF2*	*NDRG1*	*SBF1*	
	FGD4	*KARS*	*NEFL*	*SBF2*	
	FIG4	*KIF1B*	*PDK3*	*SH3TC2*	
Neuropatía Hereditaria Sensitiva y Autonómica (HSAN)	*ATL1*	*GJB3*	*KIF1A*	*SPTLC1*	*SPTLC2*
Neuropatía Hereditaria Motora distal (dHMN)	*ATP7A*	*DYNC1H1*	*HSPB3*	*LAS1L*	*SLC5A7*
	BSCL2	*FBLN5*	*HSPB8*	*MYH14*	
	DCTN1	*FBXO38*	*IGHMBP2*	*REEP1*	
	DNAJB2	*HSPB1*	*KLHL9*	*SIGMAR1*	
Atrofia Muscular Espinal (SMA)	*ASAH1*	*CHCHD10*	*EXOSC8*	*SMN2*	*VAPB*
	BICD2	*EXOSC3*	*SMN1*	*UBA1*	*VRK1*
Paraparesia Espástica Hereditaria (HSP)	*AMPD2*	*CYP2U1*	*HSPD1*	*PNPLA6*	*TECPR2*
	AP4B1	*CYP7B1*	*IFIH1*	*RAB3GAP2*	*USP8*
	AP4E1	*DDHD1*	*KIAA0196*	*REEP2*	*VPS37A*
	AP4M1	*DDHD2*	*KIF1C*	*RTN2*	*WDR48*
	AP4S1	*ENTPD1*	*L1CAM*	*SLC16A2*	*ZFR*
	AP5Z1	*ERLIN1*	*LYST*	*SLC33A1*	*ZFYVE26*
	ARL6IP1	*ERLIN2*	*MAG*	*SPAST*	*ZFYVE27*
	ARSI	*FA2H*	*NIPA1*	*SPG11*	
	B4GALNT1	*FLRT1*	*NT5C2*	*SPG20*	
	C12orf65	*GBA2*	*PGAP1*	*SPG21*	
	C19orf12	*GJC2*	*PLP1*	*SPG7*	
Esclerosis Lateral Amiotrófica (ALS)	*ANG*	*DAO*	*GRN*	*OPTN*	*TUBA4A*
	ALS2	*ERBB4*	*HNRNPA1*	*PFN1*	*UBQLN2*
	C9orf72	*FUS*	*MATR3*	*SOD1*	
	CHMP2B	*GLE1*	*NEFH*	*TARDBP*	
Ataxia Cerebelosa	*ABHD12*	*ATXN2*	*MRE11A*	*PNKP*	*SETX*
	ATM	*ERCC8*	*PIK3R5*	*POLR3A*	*TDP1*

ANEXO II

***Primers* utilizados para la subclonación del exón 4 y 100-150 pb de secuencias intrónicas flanqueantes del gen *FIG4* en el vector pSPL3** (los sitios de corte aparecen subrayados).

FIG4-E4-XhoI_F	5'-ggaaCTCGAGAGAGAATAATCCTAAGACACC-3'
FIG4-E4-NheI_R	5'-gcgcGCTAGCTTTCTACCTGATCTACGACA-3'

***Primers* utilizados para la secuenciación mediante Sanger de las construcciones pSPL3::FIG4-E4-WT y Mut.**

pSPL3seq-F	5'-CATGCTCCTTGGGATGTTGATG-3'
pSPL3seq-R	5'-ACTGTGCGTTACAATTTCTGG-3'

***Primers* específicos de los exones SD6 y SA2 del vector pSPL3 utilizados para la amplificación por PCR y secuenciación mediante Sanger de los transcritos generados tras la expresión de las construcciones pSPL3::FIG4-E4-WT y Mut.**

SD6-F	5'-TCTGAGTCACCTGGACAACC-3'
SA2-R	5'-ATCTCAGTGGTATTTGTGAGC-3'