UNIVERSITAT POLITÈCNICA DE VALÈNCIA

DEPARTAMENTO DE BIOTECNOLOGÍA

DETERMINACIÓN DEL UMBRAL DE DETECCIÓN DE VARIACIONES DEL GENOMA DEL TIPO SNVS Y PEQUEÑAS INDELS EN BASE A LOS VALORES DE LA RELACIÓN VAR/DEPTH EN MUESTRAS CLÍNICAS ANALIZADAS MEDIANTE TECNOLOGÍA NGS POR EXOMA

TRABAJO FIN DE MÁSTER EN BIOTECNOLOGÍA BIOMÉDICA

ALUMNO/A: JAIME OZÁEZ MARTÍNEZ

TUTOR/A: JOSÉ GADEA VACAS

COTUTOR/A: ALEJANDRO ROMERA LÓPEZ

Curso Académico: 2015 - 2016

VALENCIA, 08 de Julio de 2016

Resumen

El uso de las tecnologías de secuenciación masiva (NGS) para el diagnóstico de enfermedades de origen genético es cada vez más frecuente.

Tras el proceso de secuenciación masiva, cada nucleótido que no coincide con el genoma de referencia para esa posición genómica es considerado una variante. Con cierta frecuencia, aparecen falsos positivos que no representan la secuencia real presente en la muestra. El procedimiento habitual para las variantes reportadas con posible implicación diagnóstica consiste en la verificación de la presencia de dicha variante mediante secuenciación Sanger, que sigue siendo el *Gold Standard* de secuenciación en la mayoría de laboratorios. Esta verificación supone un gasto relevante, incrementándose de forma exponencial con un mayor número de genes analizados, como es el caso del exoma diagnóstico.

Existen ciertos parámetros que aportan información sobre la calidad de la secuenciación, la calidad del alineamiento, características de la región, etc. y que pueden estar relacionados con la aparición de falsos positivos. En este contexto, el objetivo del presente trabajo consiste en el estudio de dichos parámetros, junto con las características de las regiones genómicas afectadas, con la intención de identificar tendencias que expliquen la aparición de los falsos positivos. De este modo, se espera poder reducir el número de verificaciones por Sanger realizadas, aumentando así el beneficio coste-eficiencia del diagnóstico genético mediante NGS.

El resultado del estudio concluye con la revelación del *Strand Bias* como parámetro capaz de predecir, con un grado elevado de confianza, la presencia de un falso positivo en una variante reportada. Además, se identificaron características típicas de las regiones afectadas, frecuentemente relacionadas con errores durante la identificación de las variantes.

Palabras clave: NGS, secuenciación Sanger, confirmación, exoma, falso positivo.

Abstract

The use of NGS technologies for the genetic diagnosis of hereditary diseases is becoming increasingly common.

After sequencing, discordant nucleotides from the reference genome are considered as variants, and false positives (not representing the real sequence in the sample) are relatively common. The usual way to proceed with variants of probable clinical relevance is to check them by Sanger sequencing, the *gold standard* for sequencing in most laboratories. Such verification represents an extra cost that can be a significant expense depending both on the number of genes analyzed (as with exome analyses) and on the number of variants detected by NGS.

There are diverse parameters such as those which provide information on sequencing and alignment quality, or those that inform on genomic region features, and others, which could be related with the presence of false positives. Through the analysis of that kind of parameters, the aim of this study is to identify factors able to describe, explain, or infer the presence of false positives in NGS data, thus decreasing the need for Sanger sequencing verifications and increasing the benefit-efficiency ratio of the genetic diagnosis using NGS.

The results of this study indicate that the parameter *Strand Bias* is able to predict the presence of a false positive on a variant with high level of confidence. In addition, we have identified interesting features in the genomic regions where false positives tend to occur during the variants identification.

Key words: NGS, Sanger sequencing, confirmation, exome, false positive.

Índice

Abreviaturas

ADN: Ácido desoxirribonucleico

BAM: *Binary Alignment/Map*

CNV: *Copy number variant*

DE: Desviación estándar

EDTA: Ácido etilendiaminotetraacético

Fig.: Figura

FN: Falso negativo

FP: Falso positivo

GC%: Porcentaje de guaninas/citosinas

Indel: *Insertion/deletion polymorphism*

M_e: Mediana

Min: Minutos

NGS: *Next-Generation Sequencing*

Pb: Pares de bases

PCR: *Polymerase Chain Reaction*

SB: *Strand Bias*

SCA: *Spinocerebellar ataxia*

Seg: Segundos

SNP: *Single nucleotide polymorphism*

SNV: *Single nucleotide variant*

TD: *Touchdown PCR*

Tm: Temperatura de fusión

V/D: Var/Depth

VN: Verdadero negativo

VP: Verdadero positivo

$\overline{X}$: Media aritmética

1 Introducción

En la actualidad se conocen aproximadamente 8000 enfermedades genéticas, la mayoría de las de tipo monogénico, aunque únicamente en 4423 se conoce su base genética, lo que supone que en más de la mitad no se ha establecido una relación genotipo-fenotipo[1]. La mayoría de estas 4423 enfermedades están causadas por mutaciones localizadas en las regiones codificantes del genoma y en las zonas intrónicas adyacentes, produciendo modificaciones la estructura y función de las proteínas codificadas, y produciendo el fenotipo patológico[1].

Algunas de estas enfermedades son genéticamente heterogéneas, lo que conlleva, en muchos casos, la necesidad de analizar decenas o cientos de genes para poder llegar a un diagnóstico definitivo. Tradicionalmente, la estrategia para la detección de este tipo de enfermedades ha sido el uso de la secuenciación Sanger de los genes más frecuentemente mutados. Si se identifica una mutación en el gen secuenciado, se confirma el diagnóstico. En caso contrario, se sigue analizando uno a uno los genes candidatos hasta encontrar, en el mejor de los casos, la mutación causante de la enfermedad. Esta estrategia, suponía una inversión de tiempo y dinero elevada y derivando en un tiempo de diagnóstico que podía extenderse entre 1-10 años[1]. Esta demora en la obtención del diagnóstico puede impedir la aplicación de soluciones terapéuticas a corto plazo y puede privar o retrasar un adecuado asesoramiento genético para el paciente y sus familiares. Además, muchos pacientes presentan síntomas clínicos poco específicos o solapantes entre diversas patologías lo que puede hacer difícil el análisis dirigido a un grupo de genes determinado[2].

Aunque la secuenciación Sanger sigue siendo el método de referencia en secuenciación, el uso de esta tecnología para el análisis de genes de gran tamaño o ante la necesidad de testar varios genes candidatos, supone una limitación. El desarrollo de nuevas tecnologías de secuenciación capaces de solventar estas limitaciones, como la tecnología de *Next Generation Sequencing (NGS)*, ha sido de gran ayuda y ha suscitado un gran interés de cara a su inclusión en la práctica clínica rutinaria para ayudar en el diagnóstico de enfermedades genéticas. La NGS permite la secuenciación en paralelo de miles a millones de secuencias, reduciendo el coste y el tiempo global empleado y mejorando la eficiencia diagnóstica del análisis genético[3].

1.1 Secuenciación masiva (NGS, *Next-generation Sequencing*)

La secuenciación masiva está cambiando el modelo de diagnóstico molecular de los pacientes afectos de patología genética. Se estima que aproximadamente el 85% de las variantes patogénicas se encuentran regiones codificantes o en regiones de *splicing* adyacente. En base a esta información, la aplicación de la *NGS* al diagnóstico genético se está abordando siguiendo la siguiente estrategia:

- Resecuenciación dirigida al estudio de un solo gen: se recomienda en situaciones en las que la enfermedad se produce por mutaciones producidas en un solo gen o cuando la heterogeneidad en la enfermedad es baja. El estudio de un genes grandes o de genes para los que hay un gran número de muestras a analizar, son situaciones ideales para la aplicación de la NGS.
- Resecuenciación dirigida a panel de genes: se recomienda en situaciones en las que la enfermedad presenta heterogeneidad genética o en el caso de que se quieran analizar simultáneamente trastornos con clínicas solapantes.
- Exoma: permite estudiar todas (o la gran mayoría) de las regiones codificantes y zonas de splicing adyacente de los genes presentes en el genoma. Se utiliza frente a una patología con extrema heterogeneidad, en pacientes con dos o más fenotipos no relacionados o en ausencia de características clínicas claves al momento del diagnóstico.

Nuevos desarrollos de esta tecnología, están permitiendo además la identificación de Variaciones en el Número de Copias (CNVs) de los genes estudiados o incluso la detección de expansiones de regiones repetitivas (ej: tripletes expandidos en el caso de las SCAs).

En la actualidad, aunque técnicamente es posible y el precio de secuenciar un genoma humano completo sea una ínfima parte de lo que costó la obtención del primer borrador en 2001, la secuenciación de rutina de genomas completos todavía sigue siendo económicamente inabordable para la mayoría de las instituciones y su uso en la clínica está muy limitado.

Desde un punto de vista técnico y aunque existen distintos instrumentos y métodos de secuenciación masiva, la mayoría de las técnicas empleadas siguen un mismo esquema de trabajo que se puede resumir en:

- Fragmentación de ADN obtenido de cada paciente, generando fragmentos de un tamaño adecuado para cada tipo de secuenciación.
- Captura de las regiones ADN de interés para la preparación de una librería de fragmentos. En el caso de los paneles de resecuenciación y en exoma, estas regiones incluyen regiones codificantes y zonas de *splicing* adyacentes.
- Amplicación clonal de las secuencias capturadas.
- Secuenciación. Existen numerosos métodos de secuenciación. En el caso de la plataforma de Illumina, la secuenciación se lleva a cabo alternando ciclos de amplificación con terminadores reversibles (en cada ciclo, el nucleótido complementario al molde se une emitiendo fluorescencia en una longitud de onda específica), y ciclos de toma de imágenes con un sistema óptico para la identificación de la base nitrogenada incorporada a la cadena nucleotídica creciente, dando lugar a lo que se conoce como lecturas[4]. Los fragmentos de ADN son secuenciados desde los

32 extremos, creando lecturas apareadas conocidas como *paired-ends*, generando insertos de tamaño conocido que facilitan el correcto mapeo de las cortas secuencias[5].

Una vez realizada la secuenciación y utilizando diversos algoritmos bioinformáticos, las lecturas generadas son revisadas descartando aquellas con baja calidad. Las restantes son ensambladas y alineadas contra el genoma de referencia y se producen la identificación y la anotación de las variantes presentes con la información presente en las bases de datos.

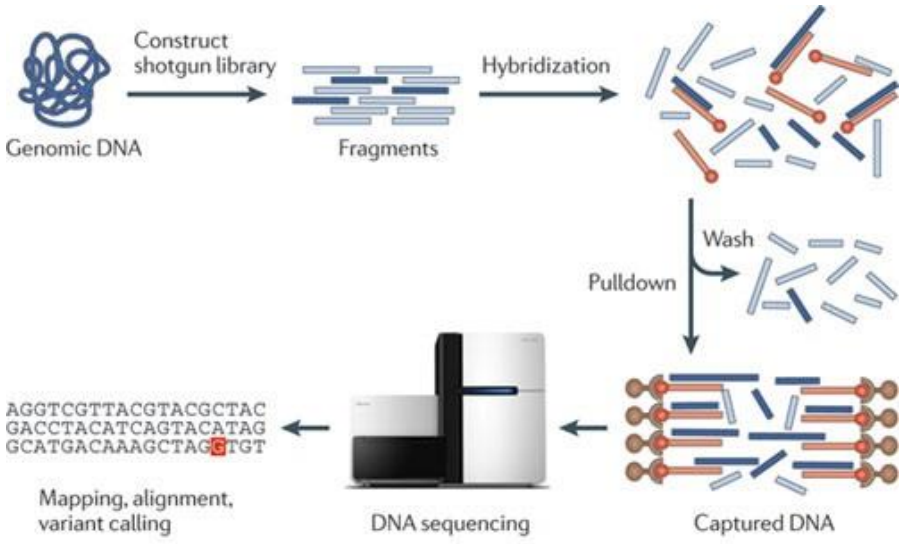


Figura 1. Flujo de pasos en el desarrollo de la llamada de secuenciación y llamada de variantes en resecuenciación dirigida[6].

1.2 Detección de variantes mediante NGS y verificación Sanger.

La detección de las variantes se realiza mediante comparación de las secuencias obtenidas en NGS frente a la secuencia del genoma de referencia utilizado (hg19 o hg38). Sin embargo, durante el proceso que incluye tanto la secuenciación NGS como la identificación y llamada de variantes, existen factores que pueden conducir a la aparición de falsos positivos o falsos negativos[7]. Factores como el propio proceso de PCR, el método de captura y secuenciación utilizados, problemas de alineamiento debido a la presencia de regiones complejas o con homología, así como la presencia de regiones con muy baja cobertura, pueden dificultar la correcta identificación de las variantes presentes en la muestra estudiada.

Debido a la posibilidad de que se presenten estos errores, y hasta la fecha, las variantes identificadas mediante NGS con posible relevancia clínica son verificadas por secuenciación Sanger, que hoy en día sigue siendo el *Gold Estandard* en secuenciación en la mayoría de laboratorios[8].

Por todo ello el proceso de selección de las variantes de interés, la exclusión de variantes sin relevancia clínica así como la identificación de falsos positivos, supone un gran esfuerzo y coste para los laboratorios. La identificación de nuevos parámetros que ayuden a agilizar y facilitar la discriminación de estos tres tipos de variantes y que eviten la necesidad de llevar a cabo una verificación por secuenciación Sanger, son esenciales para el futuro del diagnóstico genético[9].

La selección de variantes con relevancia clínica, así como la exclusión de variantes polimórficas sin implicación diagnóstica se ha visto facilitada por la aparición de bases de datos tanto específicas de patología como de bases de datos destinadas al tipaje de la población general (ej: dbSNP, ESP y ExAC).

El reto más importante pasa, por tanto, en intentar establecer parámetros que nos permitan identificar *a priori* la presencia de falsos positivos sin necesidad de llevar a cabo su verificación mediante secuenciación Sanger. Hasta ahora, los filtros establecidos para eliminación de variantes erróneas han consistido en la utilización de puntos de corte para ciertos parámetros, siguiendo unos criterios basados en la experiencia de cada laboratorio. Los parámetros más frecuentemente evaluados para este propósito han sido:

- Cobertura. Corresponde con el número de lecturas totales que cubren una posición determinada. Se ha observado que un incremento de la cobertura se correlaciona inversamente con el número de falsos positivos, ya que cuanto más veces se encuentra representada cierta posición, mayor es la precisión a la hora de determinar un valor de V/D, reduciendo la desviación generada por otros factores que interfieren durante el proceso de identificación de variantes. En un escenario ideal, la secuenciación por NGS debería presentar una cobertura homogénea por toda la zona secuenciada, sin que hubiese variaciones entre distintas regiones. Valores bajos de cobertura pueden conducir a la aparición de falsos positivos y negativos. Distintos factores pueden afectar a la cobertura de las zonas secuenciadas, como el procesado de la señal durante la secuenciación, el porcentaje de GC de la región secuenciada, el método de construcción de la librería, el procesado del pipeline informático, etc.[10].
- Var/Depth (V/D). Este parámetro indica el cociente entre el número de lecturas que reportan la variante y el número total de lecturas para esa posición, por lo que oscila entre valores de 0 y 1. El valor teórico de V/D para una variante presente en heterocigosis en la muestra (1 copia del alelo de referencia y 1 copia del alelo alternativo) es 0.5, ya que la mitad de las lecturas que cubren dicha posición deberían reportar la variante, mientras que la otra mitad deberían reportar el alelo de referencia. El valor teórico para una variante en homocigosis será de 1.
- Contenido de nucleótidos guanina y citosina de la región adyacente. Esto puede afectar tanto a la eficiencia del método de captura de las regiones de interés y como a la eficiencia del proceso de amplificación durante la PCR. El porcentaje de GC de la secuencia de ADN

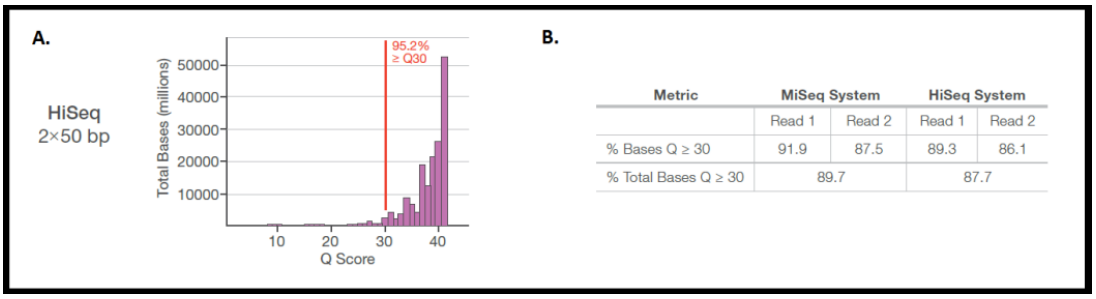
también se ha observado que puede alterar la eficiencia del proceso de fractura del DNA mediante sonicación[11].

- *Strand Bias* (SB). Representa la desviación que se produce cuando las variantes tienden a aparecer en las lecturas con una orientación determinada, por lo que indica diferencias entre las lecturas con sentido directo y sentido reverso, a la hora de reportar las variantes en cierta posición genómica. Idealmente, una variante con 12 lecturas, y con un V/D = 0.5, debería presentar 6 lecturas que reportasen la variante, 3 de ellas en las lecturas en sentido (+) y 3 de ellas con lecturas en sentido (-). En este caso el SB reportado sería cero[12]. En la realidad, además de estos valores, otros factores intrínsecos a la estructura de la región analizada y al procesado de las muestras afectan en el cálculo del valor de SB. En ciertas ocasiones las variantes tienden a aparecer reportadas preferentemente en lecturas de uno de los dos sentidos, ya sea por culpa del método de captura que propicia que las lecturas que cubren esta región tengan una orientación determinada, o por particularidades de las secuencias que influyen en el proceso de secuenciación, haciendo que únicamente las lecturas con un sentido determinado sean capaces de reportar la variante. Los estudios que existen acerca de este parámetro indican cierta tendencia de las variantes con elevado valor de SB a corresponderse con falsos positivos, ya que en ocasiones podrían estar indicando la presencia de fragmentos difíciles de secuenciar en uno de los 2 sentidos, como son los homopolímeros (repeticiones consecutivas de un mismo nucleótido) generando falsos positivos[8]. El rango en el que puede presentarse el SB comprende valores desde cero hasta infinito. Hasta la fecha este parámetro ha sido poco utilizado debido al desconocimiento sobre su significado y la manera de interpretarlo[12].
- Calidad de secuenciación. Indica, en cada base de cada lectura, cómo de nítida e intensa es la señal del nucleótido incorporado a la cadena creciente. Esta calidad corresponde con la probabilidad de que cierto nucleótido haya sido leído como un error de secuenciación y se representar mediante una escala logarítmica llamada escala *Phred*. Casas comerciales como Illumina indican que su tecnología de secuenciación consigue un elevado número de lecturas libres de errores, con una proporción superior al 80% de las bases secuenciadas con una calidad por encima de 30 en la escala Phred (Fig.2.)[13,14].

  Este valor se define como: Phred score = - 10 log (prob error).

Calidad Phred	Probabilidad de base incorrecta	Precisión de la secuenciación
10	1 entre 10	90.00%
20	1 entre 100	99.00%
30	1 entre 1000	99.90%
40	1 entre 10000	99.99%

Tabla 1. Probabilidades de error y precisión de la secuenciación en función de distintos valores de calidad Phred.



Metric	MiSeq System		HiSeq System	
	Read 1	Read 2	Read 1	Read 2
% Bases Q ≥ 30	91.9	87.5	89.3	86.1
% Total Bases Q ≥ 30	89.7		87.7	

Figura 2. Representación gráfica y tabla con los valores de calidad de secuenciación en los equipos de Illumina. A. Representación del número de bases secuenciadas en intervalos de calidad de secuenciación. B. Comparación entre el número de bases con una calidad superior a Q30 en lecturas secuenciadas en MiSeq y HiSeq[14].

2 Objetivos

- Estudio de los parámetros con posible implicación en la presencia de falsos positivos en diagnóstico por exoma.
- Identificación de factores técnicos y biológicos con influencia en la generación de error durante la identificación de variantes exoma diagnóstico.
- Análisis de las limitaciones de la identificación de variantes en exoma.
- Establecimiento de medidas correctoras y de medidas preventivas de aparición de falsos positivos.

3 Material y métodos

3.1 Selección de pacientes

Un total de 16 individuos fueron incluidos para la realización de este estudio. Estos individuos fueron pacientes referidos previamente a nuestro laboratorio para la realización de un estudio genético mediante exoma. De cada uno de los pacientes, se obtuvo muestra de sangre mediante punción venosa que fue almacenada a 4°C en tubos EDTA. Las muestras se registraron con un código para preservar el anonimato de los pacientes y mantener la trazabilidad de las muestras individuales.

3.2 Obtención del ADN de los pacientes

La extracción y purificación del ADN se realizó utilizando el sistema QIAcube (QIAGEN®, Hilden, Germany) y el kit comercial QIAamp® DNA Blood Mini Kit (QIAGEN®, Hilden, Germany), siguiendo el protocolo recomendado por el fabricante.

El ADN obtenido se cuantificó usando el espectrofotómetro NanoDrop™(Thermo Fisher Scientific Inc., Waltham, EEUU). El control de calidad del ADN extraído, se realizó analizando el ratio 260/280 ($r_{260/2680}$~1.8) y ratio 260/230 ($r_{260/2680}$~1.8), además del valor de turbidez (<0.1). Una vez extraído, los DNAs se conservaron a -20^{o}C hasta su utilización.

3.3 Secuenciación masiva y análisis bioinformático de las muestras

Los exones y regiones intrónicas adyacentes de más de 21000 genes presentes en el genoma fueron analizados a partir del ADN extraído de las muestras de sangre de cada uno de los pacientes.

La captura de las regiones de interés (librería de fragmentos) se llevó a cabo utilizando el método de captura *SureSelectXT Human AllExonV5 (Agilent Technologies, Santa Clara, EEUU)*. La secuenciación de las regiones seleccionadas se realizó utilizando la plataforma de secuenciación *IlluminaHiSeq2000 (Illumina, Inc., San Diego, EEUU)*, siguiendo la estrategia de *paired-ends*.

El análisis bioinformático de las secuencia de ADN obtenidas se llevó a cabo utilizando un algoritmo bioinformático (*pipeline)* propio. De manera resumida este algoritmo consistió en:

- Filtrado de secuencias de mala calidad y duplicados de PCR.
- Mapeo y alineamiento de las secuencias frente al genoma de referencia (GRCh37/hg19).
- Identificación y anotación de variantes.

3.4 Selección de variantes para estudio

Durante este estudio se llevó a cabo la selección de tres grupos de variantes:

- Para la estudio de posibles parámetros que permitiesen discriminar falsos positivos y falsos negativos, se analizaron los datos de 7 pacientes y se seleccionadas un número total de 79 variantes presentes en 51 genes. El criterio utilizado para la selección de estas variantes fue su posible implicación diagnóstica y por tanto las variantes polimórficas fueron descartadas de inicio. Teniendo en cuenta el objetivo principal del estudio, de entre las variantes con posible relevancia clínica se seleccionaron variantes en todos el rangos de Var/Depth.
- La selección de las variantes utilizadas para la identificación de factores que contribuyen a la presencia de falsos positivos se realizó a partir de los datos obtenidos en 6 muestras (2 tríos que consistían en un hijo/a afecto y los dos progenitores) (Fig. 3). De manera resumida los criterios de selección se muestran a continuación:
  - Selección de un tránscrito canónico (tránscrito más largo identificado sin codones de parada) para cada una de las variantes identificadas según la base de datos Ensembl.
  - Exclusión de variantes presentes en homocigosis en todas las muestras (V/D = 1).
  - Selección de las variantes con V/D comprendido entre los valores 0.1 y 0.5 en al menos la mitad de las muestras.
  - Exclusión de variantes con cobertura inferior a 10X en 2 o más muestras (de un total de 6) al considerarse que están por debajo del umbral mínimo de detección.
  - En el caso de inserciones/deleciones se excluyeron aquellas variantes localizadas en regiones de homopolímeros o zonas repetitivas.
  - Exclusión de variantes presentes en genes con pseudogenes descritos en las bases de datos Ensembl y GeneCards.

  De entre las variantes restantes se procedió a la selección manual de aquellas 33 (30 SNVs y 3 InDels) que presentaban un rango de V/D más amplio a lo largo de las 6 muestras.

- Para comprobar la capacidad de discriminación entre verdaderos y falsos positivos de los parámetros significativos, se utilizó un conjunto de datos externos en donde se seleccionaron 30 variantes al azar de 7 muestras analizadas mediante exoma diagnóstico.

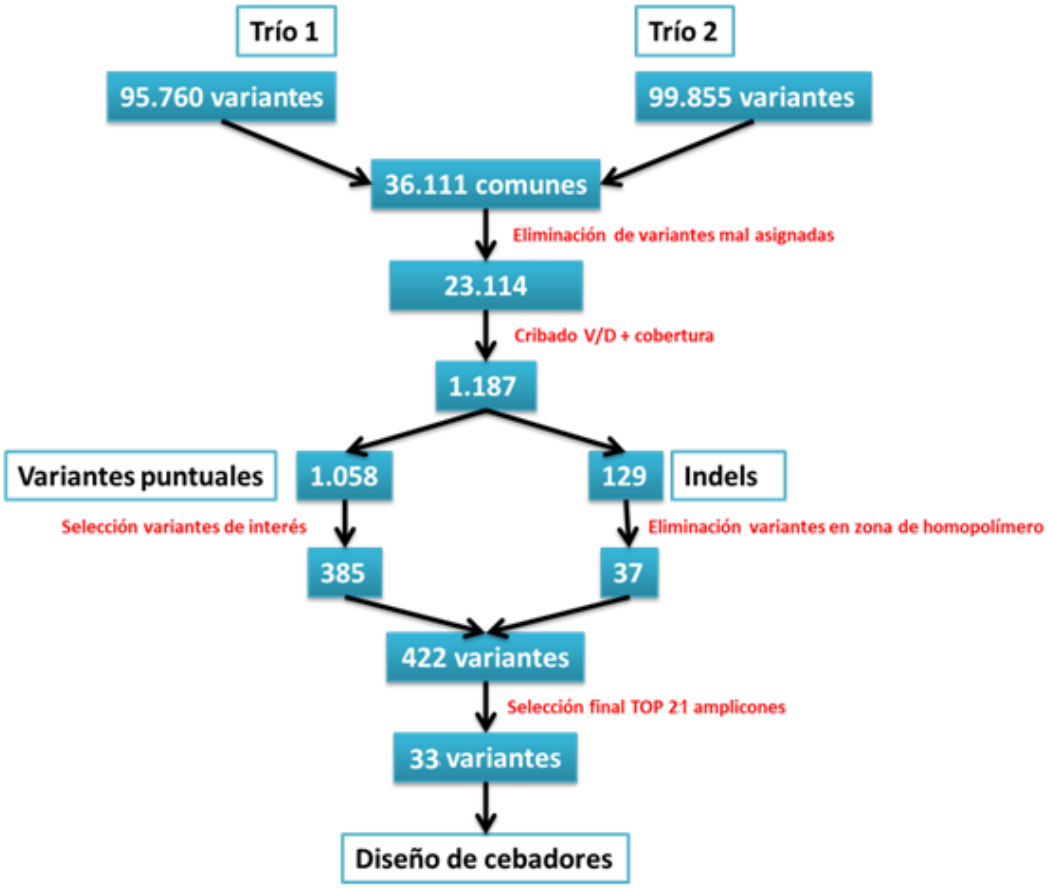


Figura 3. Algoritmo de selección de amplicones. Descripción esquemática del flujo de pasos seguidos desde la obtención de los datos iniciales, hasta la selección final de las variantes que van a ser verificadas por Sanger.

3.5 Verificación de las variantes mediante secuenciación Sanger

El proceso de validación de las 33 variantes seleccionadas para la identificación de factores que contribuyen a la presencia de falsos positivos consistió en el diseño de los cebadores de las regiones en las que se encuentran las variantes, su amplificación por PCR, secuenciación Sanger y lectura de cromatogramas. Las fases de este proceso se describen brevemente en los siguientes sub-apartados.

Este proceso para las 79 variantes incluidas en el proceso de identificación de posibles parámetros que permitiesen discriminar falsos positivos y falsos negativos, así como de las 30 variantes incluidas en la validación externa de dichos parámetros, se llevó a cabo previamente a la realización de este trabajo, en el ámbito del estudio diagnóstico.

3.5.1 Diseño de cebadores:

La secuencia de la región de interés para cada uno de los amplicones se obtuvo de la base de datos de Ensembl. El diseño de los cebadores se realizó utilizando la aplicación bioinformática Primer3Plus y teniendo en cuenta como criterios principales:

- Temperatura de fusión (Tm) = 57-70ºC;
- GC% = 20-80
- Longitud cebador = 18-30 pb
- Tamaño amplicón < 650 pb
- Distancia cebador-variante > 50 pb

La especificidad de los cebadores se analizó mediante:

- La herramienta informática *PCR in-silico* (UCSC). La presencia de un único amplicón era indicativo de especificidad.
- La herramienta informática BLAST (UCSC) nos permite determinar la homología de nuestros cebadores con el resto del genoma y evitar una disminución en la eficiencia de la PCR.

La ausencia de polimorfismos en la región de hibridación de los cebadores se analizó mediante la herramienta SNP Check. La presencia de variantes en estas regiones puede dar lugar a falsos negativos o a alteraciones artificiales en la cigosidad de las variantes.

Los cebadores diseñados se muestran a continuación:

Los cebadores diseñados se muestran a continuación:

Cebador	Secuencia	Tamaño del amplicón (pb)
AIM1L_E2.4_M_F AIM1L_E2.4_M_R	CCTAACTCACCCTGCAAAGC AACCTCTTTTGGGGTGGTG	688
LYZL2_E1_M_F LYZL2_E1_M_R	GGCCCAGAAGAGAGTTAGCA TTTCCATCCTCTGCCTTAGAA	512
SH3PXD2A_E5.1_M_F SH3PXD2A_E5.1_M_R	GAGCCTCTGGCTGACAGG GGTTTTGGGCCACTTACTCA	670
FIBP_I3_M_F FIBP_I3_M_R	CAGTGGGAGCCAATGAAAGT AGCTGCCGGAGACAGGTAT	592
DGKH_E6_M_F DGKH_E6_M_R	TCTTCTACTTCCTCCTTCCCTTT TCTCAGCTTCCACAAGAGCA	330
GAS6-AS1_E3_M_F GAS6-AS1_E3_M_R	GTCACACTGGGGTCCACAC CCCCGTTAGAATCTGCATGT	533
DNASE1L2_E2.2-5.1_M_F DNASE1L2_E2.2-5.1_M_R	CTTCGGTGACAGCAAAGTGT GTCAGAGCTCGGCGGGAGGG	644
MAP2K3_I1_M_F MAP2K3_I1_M_R	AGGTTTGTTTTCCGTTCTGC GGCCTCAGTTTACTCACTCCT	420
CBLB_E12_M_F CBLB_E12_M_R	AACACTTCCCAACCTTTTATGC GGCACAGGAATCTAGTAAAACTCC	622
TRGC2_E2_M_F TRGC2_E2_M_R	GCCACTTGTTCTGTTTATTTTTCC CAGTCATCTCTTACATGAAGCAGTAGT	693
CCDC120_E10.2_M_F CCDC120_E10.2_M_R	CCTGTGTGTGGAGACTTCCTC CCCATACCTGCCAGAGACAG	603
TTC39A_E1_M_F TTC39A_E1_M_R	CATCAGAGATCTGTTGGAGGTG CTCTCGTTTGGGGCTGCT	323
KRT4_E1.2_M_F KRT4_E1.2_M_R	TGTGGCAGGCTCAGGTCT CTACAACCTCAGGGGGAACA	425
ZNF354B_E3_M_F ZNF354B_E3_M_R	TCAGAAGCTTGAATAGTATGAGAAGTG AAACTGAGCTCAAATTTTTCAGTG	382
CWC27_E2_M_F CWC27_E2_M_R	GATTGCTACAAGTCAAATAGTTATTCA AGCAAAGAAGACCAAAGAGGA	467
SPTA1_E49_M_F SPTA1_E49_M_R	TTAAGATTCCACCAAGCCTCA AAAATTCCTCTCTCTTCCATGTG	282
LTBP1_E28_M_F LTBP1_E28_M_R	CAAAACTGAATTTGAAACATGC TCTAATTCTGTGAAACCCAAGC	286
CASP10_E9_M_F CASP10_E9_M_R	GGCCTTGTTTCAGTTCTTCA CACCCTTACCAAAGGTGTTGA	713
JMY_E8_M_F JMY_E8_M_R	CAGTGAATCATCCCCTTTGG CCCACAAAGAAATCCTACAGGT	329
LRRC1_E11_M_F LRRC1_E11_M_R	TGCATCCTTTTTATACTGTTTGG TTGAAGATGCACTGCCAAGA	325
MAP2K3_nE12_M_F MAP2K3_nE12_M_R	TCTCCTAGCCTGAGGAAGAAC TGCAGGAGCAAACAAATGAC	248

Tabla 2. Lista de cebadores diseñados para la verificación de las variantes por Sanger.

3.5.2 Amplificación por PCR

La amplificación de las regiones diana de interés se llevó a cabo mediante la técnica *Polymerase Chain Reaction* (PCR) utilizando el termociclador T100™ ThermalCycler (Bio-Rad Laboratories, Inc., Hercules, EEUU). De manera general los reactivos y las condiciones de temperatura utilizadas se muestran a continuación:

Temp (°C)	Tiempo	Ciclos
95	10 min	1
94	45 seg	14
60	45 seg	
72	1 min	
94	45 seg	9
60 (TD-0,5°C)	45 seg	
72	1 min	
94	45 seg	14
55	45 seg	
72	1 min	
72	10 min	1
4	∞	

Reactivos
Tp 5X (Promega)
$MgCl_2$ 25mM
dATP 100mM
dCTP 100mM
dGTP 100mM
dCTP 100mM
Oligo D (20 µM)
Oligo R (20 µM)
Taq Polimerasa
Agua (hasta 50µl)

Tabla 3. Programa de PCR y tabla de reactivos. Resumen de los ciclos, tiempo y temperatura de la reacción de PCR. Reactivos utilizados para la PCR.

Cuando fue necesario se modificaron las condiciones de temperatura (variación de la temperatura de hibridación) y reactivos (betaína) para aumentar la especificidad y eficiencia de la PCR. La especificad de los productos de PCR obtenidos se comprobó mediante electroforesis en un gel de agarosa al 2.5%.

3.5.3 Purificación y secuenciación Sanger

La purificación previa a la secuenciación de los productos de PCR obtenidos se llevó a cabo utilizando las placas ExcelaPure 96-Well UF Plate (Edge BioSystems, Gaithersburg, EEUU) mediante su acople a un sistema de vacío.

Tras la desnaturalización del ADN (5 minutos a 98 ^{o}C) utilizando el mix compuesto por un tampón de fabricación propia, se lleva a cabo la reacción de secuenciación (Tabla 4) el reactivo Sanger (ABI PRISM BigDye® Terminator V3.1 *Ready Reaction Cycle Sequencing Kit*, Applied Biosystems, California, EEUU).

Temp (°C)	Tiempo	Ciclos
96	1 min	1
96	10 seg	30
50	5 seg	
60	4 min	
4	∞	

Tabla 4. Programa de secuenciación. Resumen de los ciclos, tiempo y temperatura de la reacción de secuenciación.

Tras un paso de precipitación del ADN, utilizando las placas Optima DTR™ Ultra 96-Well Plate Kit (Edge BioSystems, Gaithersburg, EEUU), se introducen en el secuenciador (3730xl DNA Analyzer, Applied Biosystems,California, EEUU).

3.5.4 Comprobación de las variantes

Las secuencias obtenidas son analizadas utilizando mediante el uso de los programas Alamut ® Visual y Chromas 2.5.1.

3.6 Análisis de resultados

La distribución de las variantes en función de los parámetros objeto de estudio se analizó utilizando programa Microsoft Excel 2010.

Los estudios de significancia estadística se llevaron a cabo utilizando el test de Mann-Wilcoxon utilizando el programa R. La técnica de *Random Forest* fue utilizada para medir la asociación entre la variabilidad presente en los datos y cada uno de los parámetros analizados.

El análisis del alineamiento de las secuencias para cada muestra se realizó mediante la visualización de los archivos BAM en el programa Alamut® Visual.

3.7 Evaluación de los parámetros de interés identificados mediante el uso conjunto de validación externo.

Con la finalidad de evaluar la capacidad de discriminación de los parámetros seleccionados, se llevó a cabo una evaluación complementaria en un conjunto de datos externos que incluyeron 30 variantes en 7 muestras de exoma diagnóstico.

4 Resultados y discusión

4.1 Identificación de factores discriminantes de verdaderos/falsos positivos

Se seleccionaron 79 variantes durante el proceso de estudio genético mediante exoma realizado en 16 pacientes. Estas variantes se confirmaron mediante secuenciación Sanger y se analizó su correlación con diversos parámetros (V/D, cobertura, contenido GC de la región flanqueante y *Strand Bias*) con el fin de identificar parámetros/factores que puedan ayudar a discriminar los verdaderos positivos de los falsos positivos

4.1.1 Estudio independiente de parámetros

- Var/Depth

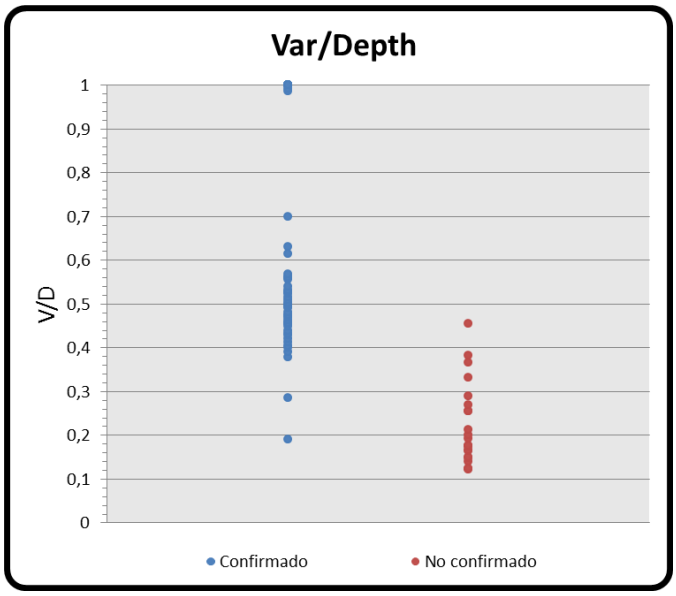


Figura 4. Representación gráfica de la distribución de las variantes en función del Var/Depth. Confirmados, N=57; No confirmados, N=22.

El análisis del Var/Depth o número de lecturas que presentan la variante respecto al número de lecturas totales para cada posición muestra que la mayoría de variantes confirmadas en heterocigosis presentan un rango de valores de V/D entre 0.191 y 0.700, y las variantes confirmadas en homocigosis presentan valores que van desde 0.986 hasta 1 ($\overline{X}$= 0.998; M_e= 1; DE = 0.003) (Fig.4.). En cambio, las variantes no confirmadas presentan valores de V/D comprendidos entre 0.122 y 0.455 ($\overline{X}$= 0.221; M_e = 0.184; DE = 0.093). Por tanto, existe un rango de valores de V/D en el cual se observan tanto variantes confirmadas como no confirmadas comprendido desde 0.191 hasta 0.455. Por debajo del V/D de 0.191 no se confirmó ninguna variante.

En la figura 4, se puede observar como aquellas variantes confirmadas se agrupan en torno a los valores de VD de 0.5 (valor teórico para la heterocigosis) o a V/D de 1 (valor teórico para homocigosis) mientras que los valores de V/D de las variantes no confirmadas quedan agrupados en valores más inferiores.

Sobre la totalidad de variantes descritas, se identificaron dos de ellas discordantes ya que se confirmaron por secuenciación Sanger con valores inferiores a 0.3 de V/D. La primera de estas se trata de una inserción TAGC> TAGCAGCAGC en el gen *NOTCH4* que presentó un valor de V/D de 0.191, lo cual se trata de una circunstancia poco habitual. Se han descrito pequeñas inserciones/deleciones (Indels) que presentan valores de V/D más reducidos que las variantes puntuales. Esto es debido a que es más complicado alinear las lecturas que contienen este tipo de variantes con la secuencia de referencia, y esto suele ser más acusado cuanto más grande sea el tamaño del indel[15], lo cual explicaría la confirmación por Sanger de esta variante. La segunda variante discordante presentó un V/D = 0.286 y característicamente un total de 14 lecturas, y porcentaje de GC 70.29%. La combinación del escaso número de lecturas obtenido junto con un % de GC de la región bastante elevado podría estar influyendo en la precisión del V/D, como se verá más adelante.

Sin embargo, el V/D ofrece una información parcial si se desconoce el número total de lecturas que han reportado la variante. Por ejemplo, para un V/D de 0.2 con una cobertura de 10X, la variante habrá sido reportada en 2 de las lecturas, en cambio, para una cobertura de 100X con el mismo V/D, la variante habrá sido reportada en 20 ocasiones, lo que hace menos probable que se trate de un falso positivo. Por este motivo, se realizó el mismo análisis normalizando el V/D frente el número de lecturas para cada variante (Fig.5).

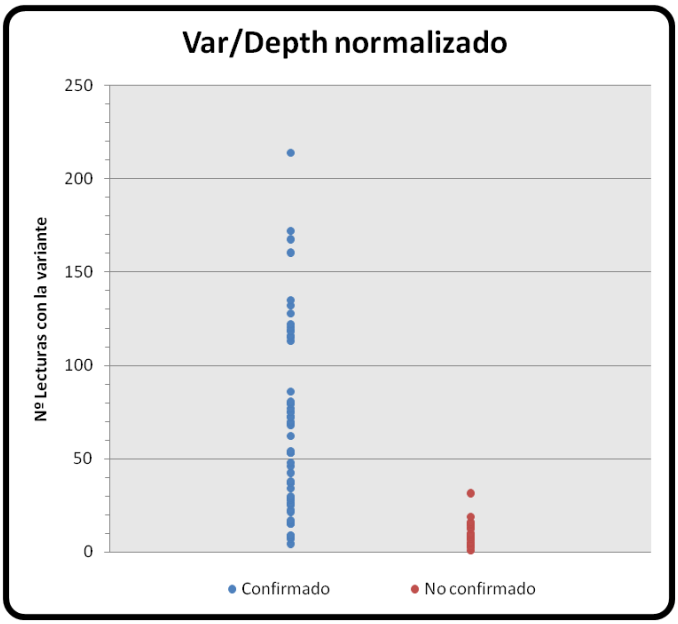


Figura 5. Representación gráfica de la distribución de las variantes en función del Var/Depth normalizado frente el número de lecturas. Confirmados, N=57; No confirmados, N=22.

En este caso lo que se analiza es el número de veces que ha sido identificada cada variante. Las variantes confirmadas se encuentran en un rango entre 4 y 214 lecturas con la variante ($\overline{X}$= 65.84; M_e = 53; DE = 48.82) y las variantes no confirmadas entre 1 y 30 lecturas con la variante ($\overline{X}$= 9.78; M_e = 8.5; DE = 6.96).

El resultado obtenido difiere ligeramente del anterior ya que con los datos normalizados se aprecian más claramente las diferencias entre los 2 grupos de variantes. Los falsos positivos se agrupan en valores inferiores de número de lecturas que reportan la variante, mientras que para las variantes confirmadas existe una dispersión más amplia del número de lecturas que presentan la variante que alcanzan valores muy superiores en comparación con las variantes no confirmadas.

Para comprobar que grado de significación estadística existía entre la distribución de las variantes atendiendo a los parámetros V/D y V/D normalizado, se realizó el test de Mann-Wilcoxon con cada grupo de datos. Los resultados obtenidos para cada uno de ellos se muestran a continuación:

- Distribución mediante V/D: $p = 5.22x10^{-11}$; W = 28.

- Distribución V/D normalizado: $p = 3.5x10^{-9}$; W = 86.5.

Ambos parámetros tienen un buen nivel de significación estadística. No obstante el valor W (Wilcoxon), que ofrece información sobre cómo de separadas están las medias de los 2 grupos, es superior el en test de V/D normalizado. Así pues, el número de veces que es reportada una variante aporta una información estadísticamente más útil que el V/D a la hora de predecir la presencia de un falso positivo en una variante reportada. Es por esto que parece intuirse una ligera correlación entre el V/D normalizado y la presencia o ausencia de falsos positivos.

- Cobertura

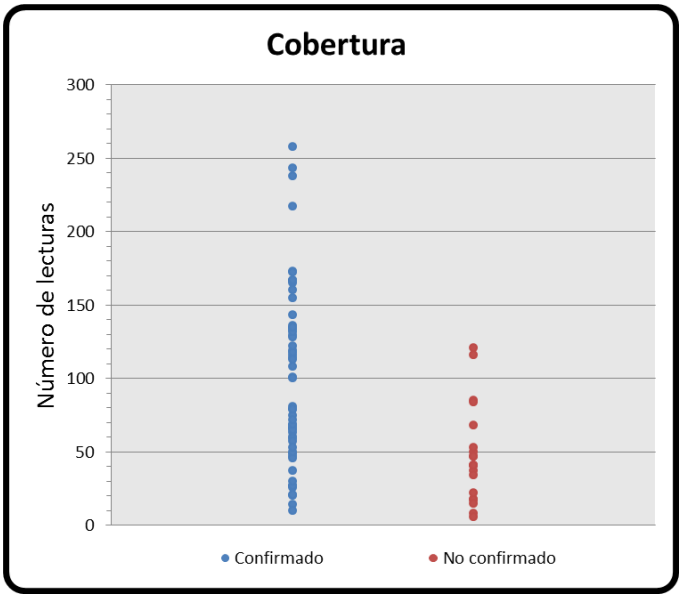


Figura 6. Representación gráfica de la distribución de las variantes en función de la cobertura. Confirmados, N=57; No confirmados, N=22.

En el análisis de la cobertura (Fig.6.), las variantes confirmadas presentan un número de lecturas entre 10 hasta 258 lecturas totales ($\overline{X}$= 101.36; M_e = 100; DE = 60.46). Las variantes no confirmadas presentan valores entre 6 y 121 lecturas ($\overline{X}$= 46.32; M_e= 41; DE = 31.51). De manera similar a lo observado para el V/D, valores más elevados de cobertura se asocian con variantes confirmadas, mientras que los falsos positivos presentan valores más reducidos de número de lecturas, observándose una menor proporción de estas variantes por encima de 55 lecturas totales.

Estos datos apuntarían hacia la posible existencia de una relación inversa entre el número de lecturas y la tasa de falsos positivos. Una menor cobertura favorecería la aparición de un valor de V/D más aleatorio, ya que variaciones pequeñas en el número de veces que ha sido leído el alelo alternativo afectan en mayor medida a la precisión del V/D. Por lo tanto, parece existir cierta dependencia entre la cobertura y la aparición de falsos positivos, por lo que es conveniente disponer de coberturas lo más elevadas que sean posibles de todas las regiones analizadas, para conseguir valores de V/D más representativos de la cigosidad real de las variantes.

- Porcentaje de GC

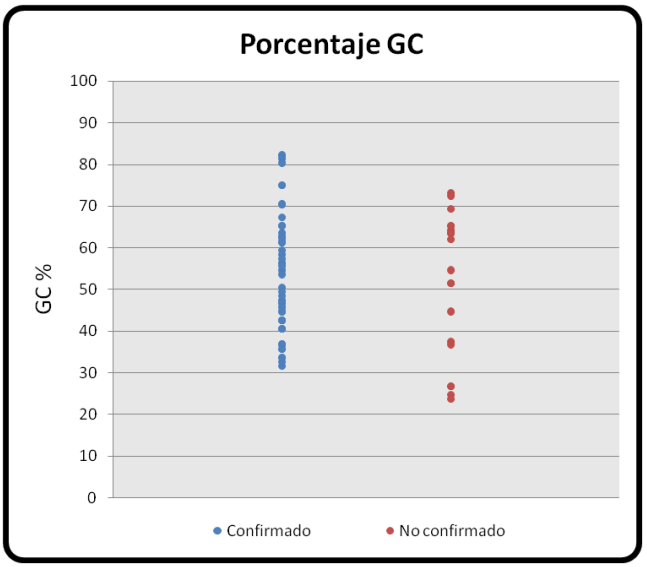


Figura 7. Representación gráfica de la distribución de las variantes en función del porcentaje de GC. Confirmados, N=57; No confirmados, N=22.

En lo que se refiere al porcentaje de GC (GC%) (Fig.7), se calculó el contenido de los nucleótidos guanina y citosina presentes en los 100 pares de bases flanqueantes a la posición de cada variante (50 pares de bases hacia cada sentido).

Para las variantes confirmadas, el rango de valores de GC% se encuentra entre 31.68% y 82.00% ($\overline{X}$= 53.42%; M_e= 53.46%, DE = 13.94). Para las variantes no confirmadas el rango de valores se encuentra entre 23.76% y 73.26% ($\overline{X}$ = 52.45%; M_e= 58.22%; DE = 16.34). No se observa ningún patrón diferencial claro en el porcentaje de GC entre las variantes confirmadas y las no confirmadas, no obstante parece que el % de GC se encuentra ligeramente desplazado hacia valores más elevados para aquellas confirmadas. Estos datos coinciden con los resultaos previamente publicados, en los que la llamada de variantes resultó independiente del porcentaje GC de la región flanqueante en muestras secuenciadas en la plataforma HiSeq 2000[16].

En nuestro caso, atendiendo a intervalos concretos de porcentajes de GC (Tabla 5), se observa como la mayoría de las variantes confirmadas se encuentran entre 40% y 60%, mientras que la mayor parte de los falsos positivos presentan porcentajes de GC incluidos entre 60% y 80%.

GC %	Confirmados	No confirmados
0% - 20%	0.00%	0.00%
20% - 40%	15.79%	27.73%
40% - 60%	47.37%	22.72%
60% - 80%	29.82%	50.00%
80% - 100%	7.02%	0.00%

Tabla 5. Proporción del número de variantes confirmadas y no confirmadas en función del contenido de GC.

Los valores más bajos de porcentaje de GC coinciden con la presencia de homopolímeros poli-T. En el presente trabajo las únicas 4 variantes que presentan homopolímeros Poli–T corresponden con los porcentajes de GC: 24.74%, 23.76% y 26.73% en 2 variantes. Ninguna de ellas se logró confirmar mediante secuenciación Sanger.

- *Strand Bias*

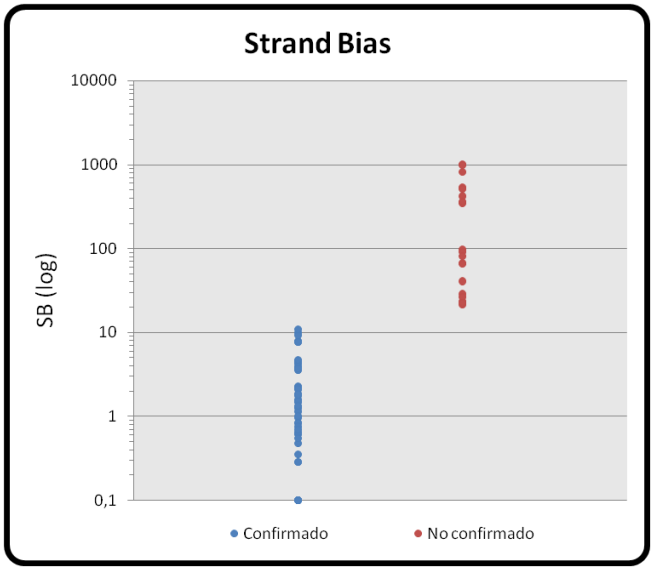


Figura 8. Representación gráfica de la distribución de las variantes en función del *Strand Bias*. Confirmados, N=57; No confirmados, N=22.

El análisis de las variantes en función de su valor de *Strand Bias* se muestra en la figura 8. Este parámetro aporta información sobre el número de lecturas que presentan la variante en un sentido y el número de lecturas que presentan la variante en el otro. Se observa que las variantes confirmadas presentan valores de SB comprendidos entre 0 y 10.887 ($\overline{X}$= 2.453; M_e= 1.505; DE = 2.77), mientras que lo falsos positivos presentan un rango de valores entre 22.683 y 1017.887 ($\overline{X}$ = 275.16; M_e= 91.33; DE = 322.06). En este caso se aprecia una diferencia clara entre los valores de SB que presentan las variantes confirmadas y las no confirmadas.

Para comprobar si existía significación estadística en esta distribución de valores, se realizó el test de Mann-Wilcoxon entre los 2 grupos de datos. Los resultados mostraron que las variantes confirmadas y las no confirmadas contienen valores de SB estadísticamente diferentes entre sí ($p = 7x10^{-12}$). Por lo tanto este parámetro presenta una potencial capacidad de separar los 2 grupos de datos.

4.1.2 Estudio comparado de parámetros

Con el objetivo de tratar de mejorar la discriminación entre verdaderos y falsos positivos, se llevaron a cabo combinaciones de los parámetros en parejas para combinar su potencial utilidad. En este apartado se representan gráficamente las distribuciones de todos los parámetros, comparándolos en gráficas de 2 dimensiones, en busca de tendencias compartidas capaces de explicar las diferencias entre verdaderos y falsos positivos.

- Var/Depth vs *Strand Bias*

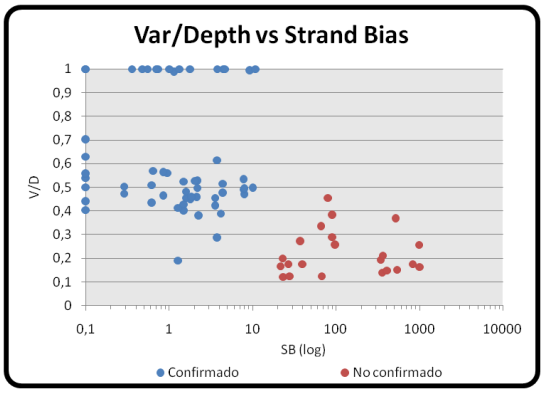


Figura 9. Representación gráfica de la distribución de las variantes en función del V/D y el SB. Confirmados, N=57; No confirmados, N=22.

En la figura 9 se representa la comparación del el efecto de los 2 parámetros que presentan una influencia individual mayor en la confirmación de las variantes: V/D y SB. En primer lugar se observa que, la mayoría de las variantes confirmadas en heterocigosis, presentan un V/D en torno a 0.4 y 0.6 y un SB entre 1 y 10. Las variantes no confirmadas presentan valores de SB superiores a 22.682 y valores de V/D por lo general inferiores a 0.4. Respecto a las variantes confirmadas en homocigosis, presentan cualquier valor de SB entre 0 y 10. No existen variante con valores de V/D cercanos a 1 y que no se hayan confirmado. Encontramos entonces un resultado lógico y esperable, que consiste en la aparición de los mayores valores de V/D coincidiendo con los menores valores de SB, y viceversa.

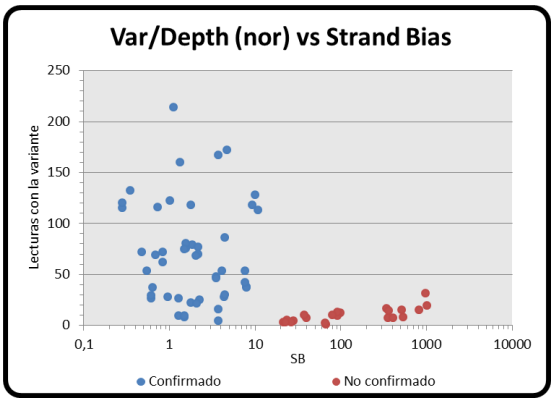


Figura 10. Representación gráfica de la distribución de las variantes en función del V/D normalizado por cobertura y el *Strand Bias*. Confirmados, N=57; No confirmados, N=22.

En la figura 10 observamos la distribución de las variantes al modificar el eje del V/D, utilizando los valores normalizados con el número de lecturas. En este caso los falsos positivos ocupan rangos menores de lecturas en comparación con los verdaderos positivos, creando una mayor separación entre ambos grupos de variantes. De este modo, las diferencias entre verdaderos y falsos positivos se acentúan cuando se comparan en función de estos 2 parámetros.

- Comparación de otros parámetros

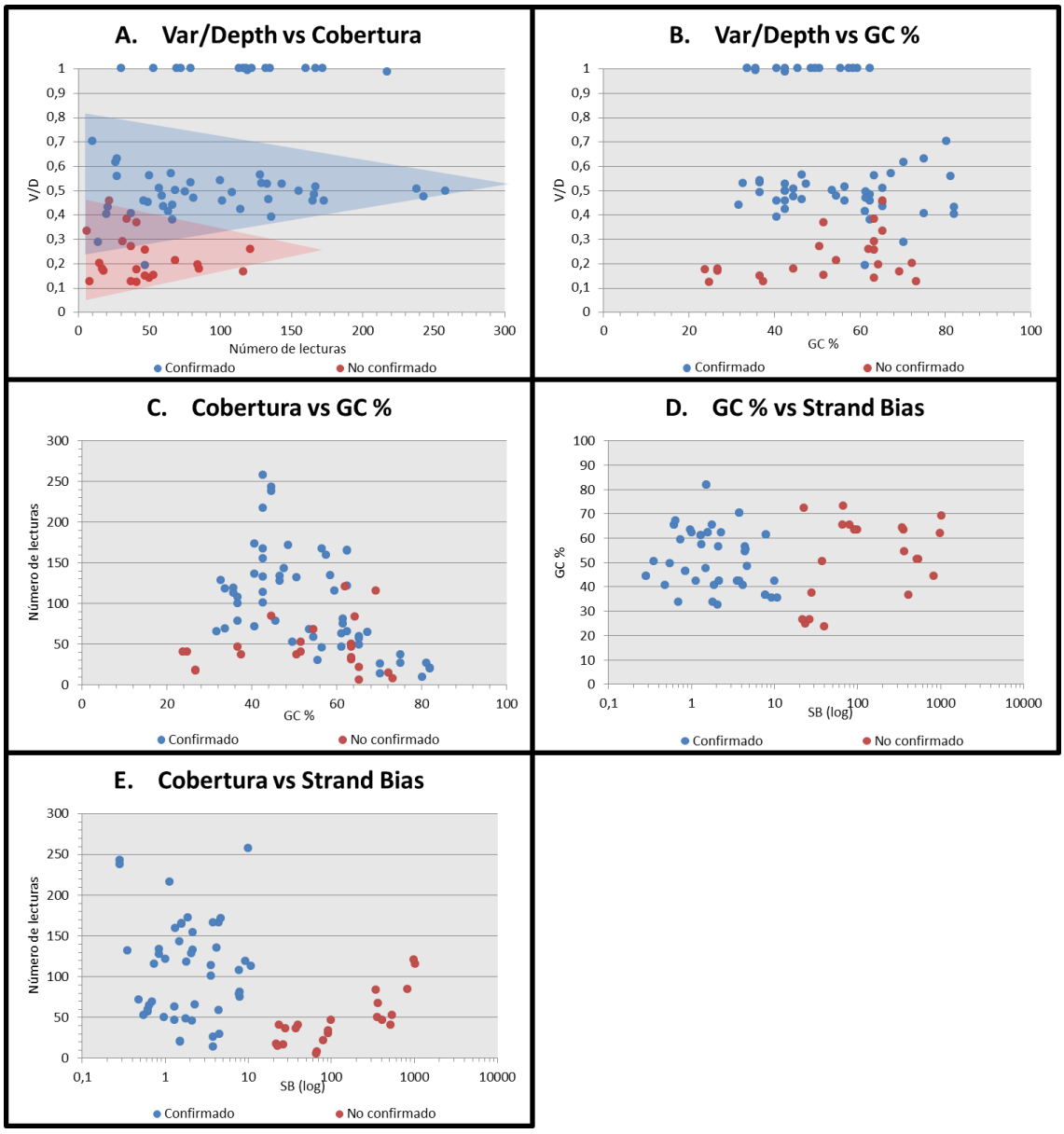


Figura 11. Representación gráfica de la distribución de las variantes en función de 2 parámetros comparados. Confirmados, N=57; No confirmados, N=22.

Al analizar conjuntamente el V/D y la cobertura (Fig.11 A), se observa un ajuste del valor de V/D hacia el teórico esperado cuando la cobertura va aumentando.

	Promedio V/D		Desviación Estándar V/D	
Nº de lecturas	**Confirmado**	**No confirmado**	**Confirmado**	**No confirmado**
1 - 51	0,473	0,233	0,147	0,105
51 - 100	0,479	0,183	0,055	0,026
101 - 150	0,535	0,21	0,056	0,065
> 150	0,54	-	0,021	-

Tabla 6. Promedio y desviación estándar de los valores de Var/Depth, medidos en intervalos de cobertura para variantes confirmadas y no confirmadas.

En la Tabla 6 se analiza la media y la desviación estándar de los valores de V/D contenidos en intervalos de cobertura de 50 en 50 lecturas. Se observa que el promedio se mantiene mientras que la dispersión disminuye al incrementar la cobertura.

Al comparar los parámetros de V/D y GC% (Fig.11 B) se observa una mayor dispersión de los valores de V/D al sobrepasar un GC % del 60%. Esta dispersión puede deberse a una mayor estabilidad de la doble hélice de ADN, que dificulta la secuenciación de las regiones con GC% extremo, dando lugar a una pérdida de reproducibilidad de los valores de V/D.

-Considerando los parámetros de cobertura y GC% (Fig.11 C), el dato más remarcable es la mayor cobertura para las variantes que presentan un porcentaje de GC más próximo al 50%, presentando un número de lecturas inferior cuando el porcentaje de GC se aleja de este valor. Este dato coincide con el obtenido resultado por Benjamini *et al.*[11], donde se analizó la cobertura de las regiones secuenciadas en función de su contenido en GC, obteniendo un pico máximo de cobertura en las regiones que contenían entre un 40% y un 55% de GC. La explicación reside en la mayor estabilidad del ADN con GC% elevados, ya que esto afecta al patrón de fragmentación durante la sonicación, y a desviaciones durante la PCR por alterar la temperatura de fusión de la doble cadena, lo que deriva en una menor captura de fragmentos de ADN de estas regiones, reduciendo la cobertura [17,18].

En el caso de la comparación entre GC% y SB (Fig.11 D) la única separación entre los grupos de variantes consiste en la influencia ejercida por los valores de SB, por lo que no se observa ninguna nueva tendencia relevante que no se haya visto con anterioridad.

Por último, en lo que se refiere a la relación entre cobertura y SB (Fig.11 E) parecen ser independientes el uno del otro, ya que no se observan tendencias diferentes a las establecidas por los parámetros individualmente.

4.1.3 Determinación de las variables de interés (*Random Forest*)

Con la intención de estudiar la capacidad individual de cada uno de los parámetros para discriminar entre falsos y verdaderos positivos, se utilizó la técnica de *Random Forest* con los datos obtenidos para las 79 variantes analizadas. Gráficamente, un mayor poder de discriminación se refleja en un mayor valor en el eje de coordenadas X (Fig.12).

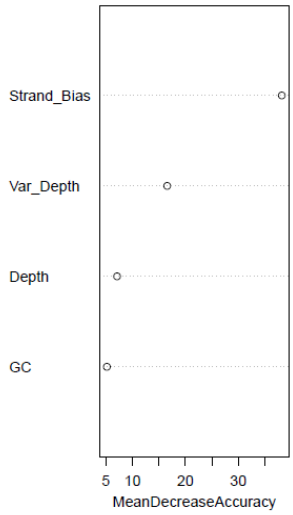


Figura 12. Representación gráfica del resultado de la técnica de *Random Forest* realizada para las 4 variables analizadas en las 79 variantes individuales.

En nuestro estudio, la variable con más peso es el *Strand Bias*, siendo prácticamente el único parámetro capaz de distinguir completamente entre los verdaderos y los falsos positivos. La siguiente variable con mayor peso es el Var/Depth, ya que sus valores consiguieron explicar en menos ocasiones las diferencias entre los dos grupos de muestras, por lo que los datos presentan una menor dependencia del V/D comparada con el SB. Tanto la cobertura como el GC % son poco o nada informativos respecto a la presencia de falsos positivos en la variantes reportadas por NGS.

4.1.4 Validación de los parámetros identificados mediante un conjunto de datos externo

Tras la obtención de estos resultados, quisimos comprobar la aplicabilidad de los mismos en un nuevo grupo de variantes, escogidas también de confirmaciones realizadas en otras muestras. Ya que el valor mínimo de SB para los falsos positivos se había observado en 22.683, se estableció un umbral con este valor de SB para clasificar las variantes en verdaderos positivos (las que presentasen un SB < 22.683) y falsos positivos (las que presentasen un SB >= 22.683).

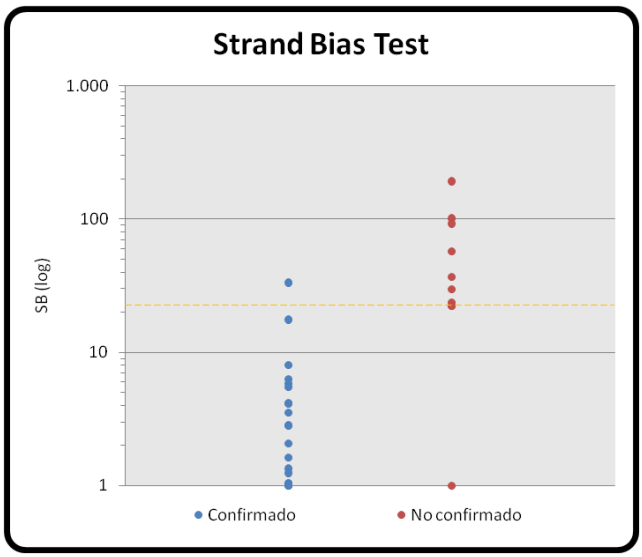


Figura 13. Representación gráfica de las variantes testeadas con umbral de *Strand Bias*. La línea discontinua naranja marca el punto de corte utilizado en este modelo para considerar los falsos positivos. Confirmados, N=20; No confirmados, N=10.

Para un total de 30 variantes, 22 de ellas presentaron un SB por debajo del umbral establecido, y 8 por encima (Fig.13.). Los resultados obtenidos fueron los siguientes:

	Confirmadas	**No confirmadas**
SB < 22,683	19 (63.3%)	3 (10%)
SB > 22,683	1 (3.3%)	7 (23.3%)

- 19 variantes NGS con un SB<22.683 fueron confirmadas por Sanger (VP) (63.33%)
- 7 variante NGS con SB>22.683 no fueron confirmadas por Sanger (VN) (23.33%)
- 1 variante NGS con SB > 22.683 fue confirmada por Sanger (FN) (3.3%)
- 3 variantes NGS con SB < 22.683 no fueron confirmadas por Sanger (FP) (10.00%)

Los valores de sensibilidad (VP / (VP + FN)) y especificidad (VN / VN + FP)) derivados de este modelo resultaron 95% y 70% respectivamente.

No obstante, y a pesar del haber obtenido un valor de especificada no muy elevado, los resultados obtenidos son prometedores. Hay que señalar que el punto de corte del SB ha sido establecido a partir de un pequeño tamaño muestral, por lo que sería necesario aumentarlo para ajustar más este valor y mejorar el valor de especificidad.

En un análisis más pormenorizado, llama la atención la presencia de un falso positivo con un valor de SB = 0. Este valor indica que no existe ninguna desviación entre las lecturas que reportan la variante en el sentido directo (+) y el sentido reverso (-). Sin embargo, al analizar el alineamiento de esta variante en el archivo BAM observamos que sí que existen diferencias entre las lecturas de ambos sentidos (Fig.14).

Figura 34. Visualización del alineamiento de la variante no confirmada con SB = 0 en Alamut Visual. Arriba se encuentra la secuencia de referencia, y cada una de las barras horizontales verdes y azules corresponde con las lecturas alineadas para esta región en la muestra seleccionada. Los cambios respecto al genoma de referencia aparecen representados en las lecturas con la letra del nucleótido de la variante.

Esta variante (V/D= 0.21 y cobertura =14x) se localiza adyacente a un homopolímero Poli-C ubicado en el gen *NPC1*, presentando la variante únicamente en las lecturas con orientación reversa (color azul, -). Normalmente, esto generaría un valor de SB elevado, pero el alineador contempla otra serie de parámetros (como la complejidad y características de la región) que puede modificar el valor final del SB. De la misma manera el SB calculado en el pipeline bioinformático también se ve afectado cuando varias muestras son secuenciadas simultáneamente en la misma carrera del secuenciador, generando un valor que corresponde a la media aritmética de los valores de SB de todas las muestras. Ambos factores podrían explicar el valor de SB obtenido para esta variante.

En el estudio pormenorizado también se observó la presencia de una variante confirmada (V/D=0.929 y cobertura =14x) con SB muy superior al umbral establecido. Dicha variante presentaba un SB = 32.926. No obstante, gráficamente observamos el cambio en 13 de las 14 lecturas, apareciendo en los 2 sentidos de las lecturas, y con una calidad de secuenciación superior a 30 en la escala Phred (Fig.15.). En este caso, la presencia de un número muy bajo de lecturas en la orientación directa (color verde, +) puede que esté afectando la estimación del cálculo del valor de SB.

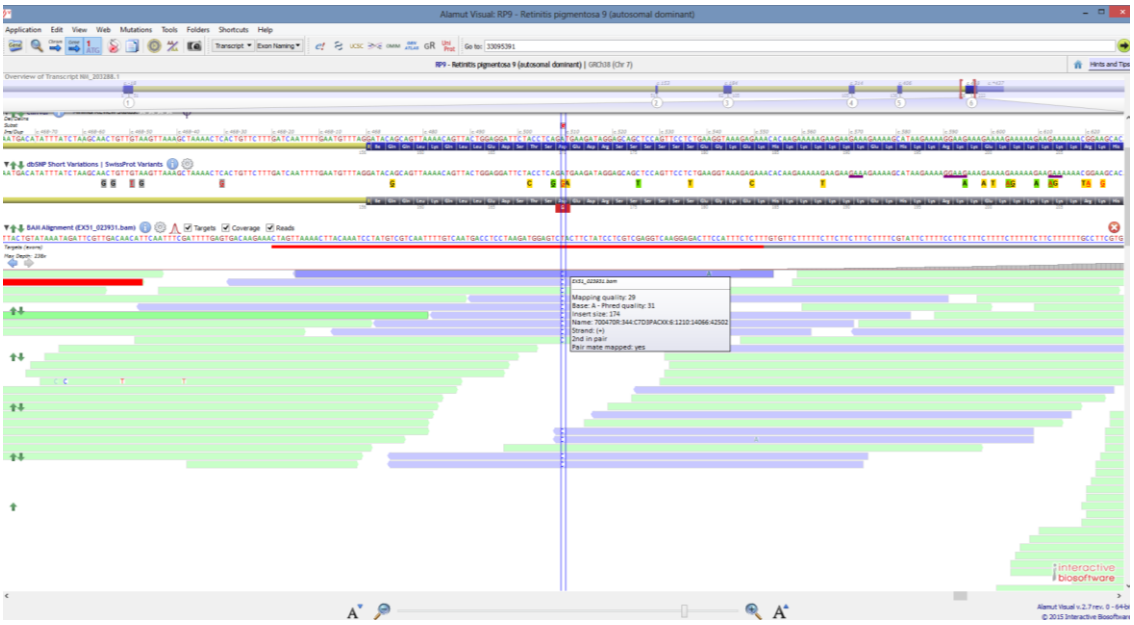


Figura 15. Visualización del alineamiento de la variante confirmada con SB = 32.926, en Alamut Visual.

4.2 Identificación de factores que contribuyen a la presencia de Falsos Positivos

Para estudiar los factores que contribuyen a la presencia de falsos positivos, seleccionamos y confirmamos por Sanger un total de 33 variantes presentes en al menos 3 de las 6 muestras (2 tríos), las cuales presentaban un V/D en el rango de 0.1-0.5 y podían ser candidatas a presentar falsos positivos.

Los resultados obtenidos para estas variantes se muestran en la tabla 7.

Un análisis de estos resultados junto el estudio detallado de los alineamientos de secuencia obtenidos para estas variantes nos ha permitido identificar diversos factores que parecen contribuir a la aparición de falsos positivos. La presencia de estos factores suele correlacionarse además con la presencia de un SB elevado.

Gen	Chr	Pos	Alelo Referencia	Alelo Variante	P1.1 (V/D)	P1.1 (Conf)	P1.2 (V/D)	P1.2 (Conf)	P1.3 (V/D)	P1.3 (Conf)	P2.1 (V/D)	P2.1 (Conf)	P2.2 (V/D)	P2.2 (Conf)	P2.3 (V/D)	P2.3 (Conf)
AIM1L	chr1	26671658	G	A	0.382	NO	0.304	NO	0.161	NO	0.420	NO	0.250	NO	0.181	NO
AIM1L	chr1	26671690	A	G	0.348	NO	0.281	NO	0.143	NO	0.400	NO	0.229	NO	0.164	NO
CASP10	chr2	202073781	G	T	0.200	NO	0.222	NO	0.312	NO	0.409	NO	0.136	NO	0.167	NO
CBLB	chr3	105421320	G	A	0.300	NO	0.211	NO	0.263	NO	0.143	NO	0.278	NO	0.421	NO
CCDC120	chrX	48925467	C	G	0.154	NO	0.308	NO	0.143	NO	0.556	NO	0.280	NO	0.115	NO
DGKH	chr13	42733400	A	T	0.250	NO	0.423	NO	0.267	NO	0.200	NO	0.281	NO	0.353	NO
DGKH	chr13	42733401	G	T	0.235	NO	0.423	NO	0.158	NO	0.158	NO	0.258	NO	0.314	NO
DGKH	chr13	42733403	G	T	0.167	NO	0.261	NO	0.111	NO	0.100	NO	0.188	NO	0.139	NO
DNASE1L2	chr16	2287020	A	C	0.484	NO	0.382	NO	0.536	NO	0.250	NO	0.306	NO	0.152	NO
FIBP	chr11	65654937	A	G	0.250	NO	0.333	NO	0.400	NO	0.444	NO	0.105	NO	0.188	NO
GAS6-AS1	chr13	114537583	T	G	0.156	NO	0.314	NO	0.438	NO	0.360	NO	0.195	NO	0.237	NO
JMY	chr5	78608354	C	T	0.167	NO	0.375	NO	0.087	NO	0.077	NO	0.304	NO	0.154	NO
KRT4	chr12	53207583	C	CCGG	0.389	NO	0.188	NO	0.225	NO	0.250	NO	0.250	NO	0.115	NO
LRRC1	chr6	53778661	T	G	0.156	—	0.385	NO	0.395	NO	0.304	NO	0.176	NO	0.400	NO
LTBP1	chr2	33586479	G	T	0.262	NO	0.409	NO	0.312	NO	0.269	NO	0.147	NO	0.125	NO
LYZL2	chr10	30918486	G	A	0.345	NO	0.000	NO	0.435	NO	0.252	NO	0.498	NO	0.285	NO
LYZL2	chr10	30918549	G	C	0.228	NO	0.000	NO	0.381	NO	0.190	NO	0.452	NO	0.262	NO
SH3PXD2A	chr10	105363565	C	A	0.533	NO	0.222	NO	0.333	NO	0.167	NO	0.318	NO	0.250	NO
TTC39A	chr1	51787862	A	C	0.211	NO	0.455	NO	0.375	NO	0.316	NO	0.227	NO	0.167	NO
CWC27	chr5	64070617	A	AAATT ATGAC	0.286	SI	0.146	SI	0.906	SI	0.267	SI	0.381	SI	0.429	SI
MAP2K3	chr17	21194769	C	T	0.241	SI	0.385	SI	0.236	SI	0.256	SI	0.333	SI	0.346	SI
MAP2K3	chr17	21194771	A	T	0.244	SI	0.385	SI	0.236	SI	0.253	SI	0.135	SI	0.338	SI
MAP2K3	chr17	21194785	T	C	0.282	SI	0.358	SI	0.239	SI	0.247	SI	0.360	SI	0.354	SI
MAP2K3	chr17	21194820	G	A	0.301	SI	0.315	SI	0.278	SI	0.225	SI	0.329	SI	0.282	SI
MAP2K3	chr17	21194843	T	C	0.274	SI	0.270	SI	0.246	SI	0.256	SI	0.292	SI	0.296	SI
MAP2K3	chr17	21194852	G	A	0.241	SI	0.286	SI	0.262	SI	0.267	SI	0.315	SI	0.299	SI
MAP2K3	chr17	21194859	G	A	0.259	SI	0.286	SI	0.283	SI	0.276	SI	0.321	SI	0.319	SI
MAP2K3	chr17	21216891	G	A	0.429	SI	0.167	SI	0.163	SI	0.000	NO	0.371	SI	0.226	SI
SPTA1	chr1	158584091	A	G	0.327	SI	0.222	SI	0.211	SI	0.188	SI	0.407	SI	0.439	SI
TRGC2	chr7	38284794	C	T	0.231	SI	0.592	SI	0.458	SI	0.320	SI	0.306	SI	0.400	SI
TRGC2	chr7	38284804	T	C	0.244	SI	0.587	SI	0.463	SI	0.288	SI	0.273	SI	0.390	SI
TRGC2	chr7	38284808	A	C	0.255	SI	0.567	SI	0.478	SI	0.273	SI	0.270	SI	0.397	SI
ZNF354B	chr5	178293175	C	CT	0.000	NO	0.330	SI	0.009	NO	0.008	NO	0.459	SI	0.217	SI

Tabla 7. Resumen de resultados de las 197 variantes verificadas por Sanger.

Los factores identificados como asociados a la presencia de falsos positivos se describen en los apartados siguientes.

4.2.1 Regiones homopoliméricas

Las regiones de homopolímero suelen ser propensas a la presencia de falsos positivos. En estos casos (Fig.15.) la variante está adyacente al homopolímero, el alelo variante coincide con la base de que se compone este homopolímero, está al final de la secuencia y solamente en un tipo de hebra (ej. tras una región poli-T, se reporta un cambio de cualquier base por una T). Estas variantes presentan un valor de SB muy elevado.

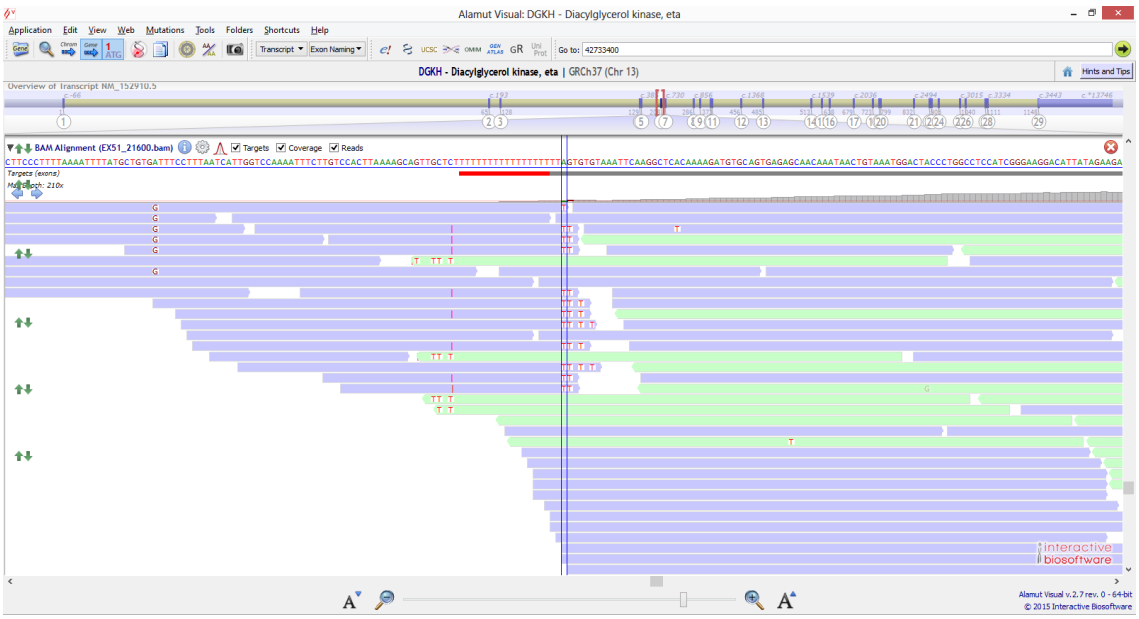


Figura 16. Visualización de archivo BAM en Alamut Visual. Representación visual del alineamiento en una región de homopolímero.

En la Figura se observa una variante del gen *DGKH* en la muestra P2.3, situada en la posición genómica 42733400, se presenta un cambio A>T con V/D = 0.353 y SB = 33.132. Adyacente a esta posición también se observan el mismo tipo de cambio. Está variante se presenta únicamente en las lecturas (azul, +) cuyo sentido hace que sea leído el homopolímero completo y cuyo final se encuentra cercano a la última base de la región homopolimérica. Esta variante no fue confirmada tras secuenciación Sanger. De la misma manera observamos que en el extremo opuesto del homopolímero se reportan solamente cambios puntuales en las lecturas con sentido reverso (verdes, -). Las lecturas que tienen su inicio dentro de la región homopolimérica no presentan las variantes.

Una posible explicación a este fenómeno sería el *slippage* producido por la polimerasa durante el proceso de amplificación de PCR [19]. Este *patinazo* hace que se produzcan errores en las secuencias que contienen homopolímeros, adicionando nucleótidos extra downstream de la región homopolimérica. El patrón observado durante el alineamiento de secuencias podría explicarse teniendo en cuenta que:

- Cuando la región homopolimérica coincide con el extremo de la lectura, la polimerasa adicionará unas pocas bases extra incorrectas, y estás y las bases posteriores de la secuencia serán reportadas como variantes tras el alineamiento, pero no serán suficientes para descartar dicha lectura por exceso de *mismatches*.
- Cuando la región homopolimérica se localiza centrada en la lectura, la polimerasa adicionará unas pocas bases extra incorrectas, y estás y una gran cantidad de bases posteriores de la secuencia serán reportadas como variantes tras el alineamiento, superando en este caso el límite de errores que se establece para las lecturas de mala calidad siendo excluida del alineamiento.

4.2.2 Lecturas con final prematuro

Un efecto de alineamiento parecido sucede en algunas regiones que no presentan regiones de homopolímeros. Estas variantes también presentan elevados valores de SB, ya que dicha variante únicamente es reportada por lecturas en un único sentido. En este caso, la región analizada no presenta características fácilmente detectables que puedan estar afectando a la secuenciación. En este tipo de variantes, se observa que las lecturas que reportan la variante tienen una gran tendencia a finalizar su longitud justamente en el lugar donde se encuentra la misma o en posiciones adyacentes, observándose en el alineamiento unas estructuras características en forma de "huecos" con reducida cobertura en sentido downstream a la variante. En estas regiones aparecen lecturas de diversos tamaños.

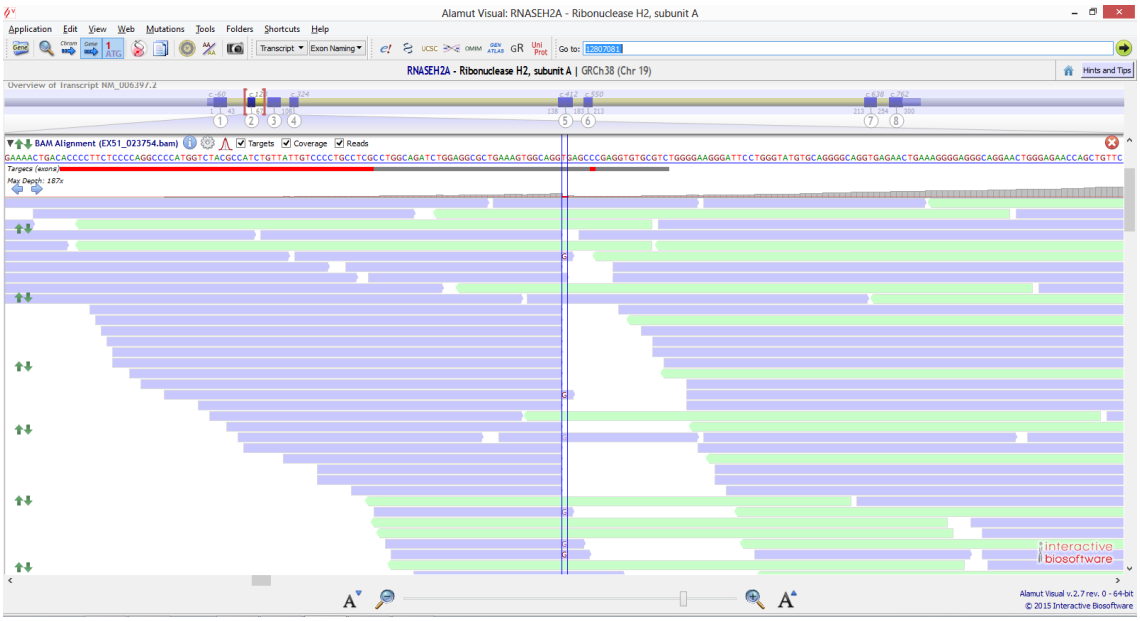


Figura 17. Visualización de archivo BAM en Alamut Visual. Representación visual de un alineamiento característico de variante no confirmada, que presenta SB elevado sin presencia de región homopolimérica.

En la figura 17, la variante del gen *RNASEH2A*, situada en la posición genómica 12807081, presenta un cambio T>G con un V/D = 0.382, y un SB = 91.333. En este caso se aprecia cómo

las lecturas de color azul presentan la variante y terminando abruptamente en dicha posicion o en posiciones contiguas. Esta variante no se observa cuando la secuencia tiene su inicio ligeramente upstream respecto a la posición de la variante, ni en las lecturas del sentido opuesto (verdes). Este tipo también se caracterizan por presentar una calidad baja de secuencia tanto en la variante como en las bases posteriores a la misma (dato no mostrado).

Una explicación para este fenómeno sería el contexto biológico en el que se encuentra la variante. La mayoría de estas variantes se encuentran en regiones intronicas, que son seleccionadas en el proceso de captura por encontrarse adyacentes a los exones. Estas zonas están muy poco conservadas, por lo que la secuencia que se puede encontrarse en las muestras puede ser totalmente diferente a la secuencia del genoma de referencia, haciendo muy dificil alinear ciertas lecturas, si las discordiancias son demasiado grandes. Estas regiones son conflictivas para el alineamiento de las lecturas, dando lugar a alinemientos subóptimos, que pueden generar este tipo de estructuras.

También, es posible que en estos casos, la secuencia upstream a la variante puede contener ciertas particularidades (no siempre la misma en cada caso) que generen problemas a la hora de amplificarse (ya sea formación de horquillas por presencia de palíndromos, elevado contenido en GC, etc) lo que puede producir una pérdida de procesividad de la polimerasa a partir de cierta posición genómica. Esto podría explicar el final prematuro en el que coinciden gran cantidad de lecturas. Las lecturas que consiguen sobrepasar esta región donde se reporta la variante, lo hace con bases de muy baja calidad de secuenciación, lo que da lugar a lecturas con una secuencia incorrecta que son descartadas por exceso de mismatches.

Este fenómeno fue reportado también por Nakamura *et al*.[20] al analizar los errores asociados a la secuenciación que eran carácterísticos de los equipos de Illumina. En este estudio observaron caidas drásticas en la calidad de las bases en puntos concretos de las lecturas, lo que derivaba en la aparición de falsos positivos en dichas regiones. Existía, además, una correlación entre la media de la calidad Phred de las bases secuenciadas y la porporción de variantes reportadas. En estos casos se observarón dos patrones característicos:

- En el 100% de los casos la variante identificada repetía la misma base que la que se encuentra en las posiciones precedentes (-1 y -2).
- Además, observaba la presencia del triplete GGC en la región upstream cercana a la variante. Estos resultados se reproducen en el 100% de las variantes en las que hemos encontrado la presencia de este fenómeno.

4.2.3 Lecturas mal alineadas

En algunas situaciones la presencia de falsos positivos parece deberse a un alineamiento erróneo de la secuencia. El método de secuenciación masiva mediante la estrategia de *paired-ends* utilizado en este trabajo permite generar lecturas apareadas separadas por un tamaño de inserto que en nuestro caso es aproximadamente de 100 pb. En algunos casos, el alineador indica que el tamaño de inserto entre ambas secuencias está fuera de rango, o que la pareja de lecturas del *paired-end* no ha mapeado correctamente. Este tipo de lecturas son marcadas con un mensaje que identifica el error cuando se utiliza visualiza el alineamiento.

Mientras que las lecturas correctamente alineadas son útiles para la detección de variantes de nucleótido único o pequeñas indels, este tipo de lecturas que son marcadas como erroneas (mostrando parejas de lecturas con un tamaño de inserto anómalo o con una orientación distinta a la esperada), pueden utilizarse como potenciales indicadores de variantes estructurales[5].

En la figura 18 se observa la aparición de estas lecturas de color rojo, en la muestra P1.1, en la posición 30918486 dentro del gen *LYZL2*. Se observa la aparición de un cambio G>A, con un V/D = 0.345 y en la mayoría de los casos es reportado por las lecturas marcadas en rojo, que indican un tamaño de inserto de -1340146 pb. Se observa como en las mismas lecturas que reportan la variante, reporta siempre variantes cercanas; en cambio, las lecturas que no reportan la variante, no reportan ninguna otra. La variante en cuestión no se confirma, por lo que parece que se ha producido un alineamiento incorrecto de las secuencias que ha sido la causa de la aparición de los falsos positivos de la zona. Este hecho podría deberse a la presencia de pseudogenes o zonas de homología, donde alinean las lecturas incorrectamente, reportando falsos positivos.

La presencia de lecturas con el mensaje "*Pair mate mapped: no*" es indicativo de que el alienador ha conseguido mapear solamente una de las lecturas del *pair end*, pero no su pareja, por lo que también podríamos sospechar de un alineamiento incorrecto.

De este modo, las lecturas pueden aparecer marcadas en rojo al visualizar el archivo BAM en Alamut Visual, junto con el mensaje *Insert size* = X, refiriéndose a que la distancia a la que se encuentran las lecturas de la pareja que forman en ese inserto, está fuera de lo normal. Si el tamaño es demasiado grande, implica una distancia muy superior a lo esperado en la secuenciación por *paired-ends* realizada en el secuenciador de Illumina (~100 pb), lo que podría indicar la presencia de una inserción (tamaño de inserto inferior al esperado) o deleción (tamaño de inserto mayor al esperado). De la misma manera, puede aparecer el mensaje mostrando un tamaño de inserto con valor negativo, lo que indicaría que la pareja de lecturas se encuentra superpuesta, y valores elevados de este tamaño de inserto sugieren separaciones grandes de las lecturas, en sentidos opuestos. En estos casos, podría sospecharse de reordenamientos (inversiones, translocaciones).

La aparición de este tipo de lecturas aberrantes puede deberse a la presencia de *paired-ends* quiméricos, donde la pareja de lecturas contiene secuencias que no se encuentran cercanos en el genoma, aunque la frecuencia con la que se generan estas quimeras es reducida. Por lo tanto, para que este tipo de alineamientos sean representativos de una variante estructural deben ocurrir con una frecuencia elevada en la misma región[5]. Ahora bien, el tamaño de los *paired-ends* es limitado, por lo que la detección de variantes estrcuturales se restringe a aquellas que presentan un tamaño limitado.

En cualquier caso, la opcion más probable en la mayoría de los casos de estos alineamientos aberrantes es la presencia de regiones conflictivas (pseudogenes, regiones muy variables, regiones de homología), ya que estas lecturas suelen reportar gran cantidad de variantes cada una de ellas, y nunca se confirman tras la secuenciación Sanger.

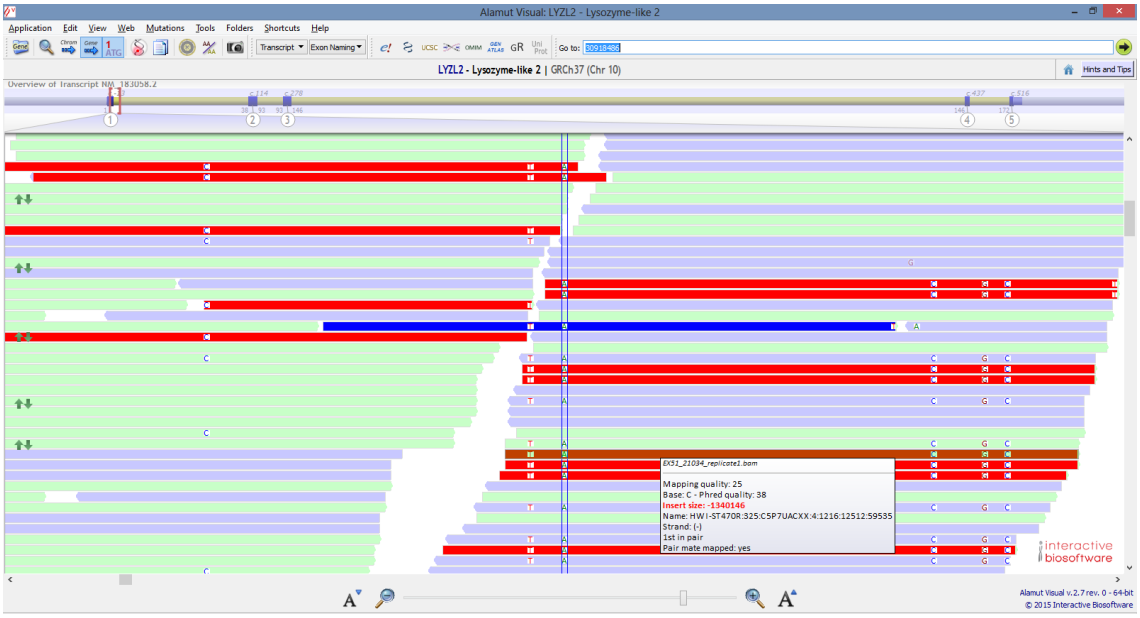


Figura 18. Representación del alineamiento de lecturas de la variante del gen *LYZL2* en Alamut Visual. Puede observarse el alineamiento que presenta gran cantidad de lecturas con un tamaño incorrecto, lo que desemboca en la aparición de gran cantidad de variantes que no se confirman.

4.2.4 Correlación del V/D con la altura de pico en Sanger

Para algunas de las variantes analizadas se ha observado que existe una correlación entre el valor de V/D obtenido en NGS y el tamaño del pico obtenido durante la secuenciación Sanger.

A modo de ejemplo se presentan 3 variantes analizadas en el gen *TRGC2*. En la muestra P1.2 estas variantes fueron reportadas con un V/D cercano a 0.5 (0.592, 0.587 y 0.567) y se confirmaron mediante secuenciación Sanger con valores tipicos de la heterocigosis (aproximadamente la mitad que la altura normal de los picos en homocigosis) (Fig.19).

Figura 4. Cromatograma de la secuenciación Sanger de las variantes del gen *TRGC2* en la muestra P2.2. Se observa en primer plano el cromatograma en el visualizador Chromas donde pueden verse la altura de los picos de fluorescencia para cada base secuenciada, mediante un código de colores. Al fondo se encuentra Alamut Visual como guía para encontrar la posición de la variante dentro de la secuencia en el cromatograma. Las flechas indican las posiciones de las variantes en cuestión.

En la muestra P2.2, estas variantes presentaron un V/D más bajo del esperado (0.306, 0.273 y 0.27) y en la confirmación por Sanger, se observó para las tres variantes un pico con una altura del 50% con respecto a lo esperado para una variante en heterocigosis (Fig.20.).

Figura 20. Cromatograma de la secuenciación Sanger de las variantes del gen *TRGC2* en la muestra P2.2.Se observa en primer plano el cromatograma en el visualizador Chromas donde pueden verse la altura de los picos de fluorescencia para cada base secuenciada, mediante un código de colores. Al fondo se encuentra Alamut Visual como guía para encontrar la posición de la variante dentro de la secuencia en el cromatograma. Las flechas indican las posiciones de las variantes en cuestión.

Este fenómeno podría explicarse mediante la presencia de pseudogenes (estos alterarían la dosis génica para el gen modificando el V/D obtenido en NGS y el tamaño del pico obtenido en Sanger). Aunque este gen había sido analizado en búsqueda de pseudogenes descritos, no es posible excluir que existan pseudogenes no reportados en las bases de datos o regiones de homología que puedan estar afectando el resultado.

La presencia de mosaicismo somático para la variante, es decir, que dentro del mismo individuo coexistan 2 o más poblaciones de células con distinto genotipo y en proporción variable, también podría explicar este tipo de resultados.

5 Conclusiones

- Existen parámetros representativos de la probabilidad de la presencia de falsos positivos en las variantes reportadas en NGS.
- El Strand Bias ha resultado ser el parámetro con mayor poder de predicción de la presencia de falsos positivos en la secuenciación masiva por exoma.
- El modelo diseñado utilizando un umbral de Strand Bias tiene una eficacia limitada, y sería preciso ampliar el estudio para establecer un punto de corte más preciso.
- El Var/Depht aporta cierta información, aunque esta es limitada si se desconoce la cobertura de la variante.
- Tanto la cobertura como el porcentaje de GC ofrecen una información individual reducida para la predicción de la presencia de falsos positivos.
- Existen otros parámetros poco estudiados en cuanto a su posible implicación en la aparición de falsos positivos en NGS, como son la calidad Phred de secuenciación de las variantes, y la posición de la variante dentro de la lectura.
- En el diagnóstico genético mediante secuenciación masiva por exoma, actualmente existen ciertas limitaciónes técnicas que delimitan su uso para detección de SNVs y pequeñas indels, ya que las CNVs son todavía difíciles de estudiar.

6 Referencias bibliográficas

1. Md., S. S.-G. *et al.* Diagnóstico Molecular De Enfermedades Genéticas: Del Diagnóstico Genético Al Diagnóstico Genómico Con La Secuenciación Masiva. *Rev. Médica Clínica Las Condes* **26,** 458–469 (2015).

2. Need, A. C. *et al.* Clinical application of exome sequencing in undiagnosed genetic conditions. *J. Med. Genet.* **49,** 353–61 (2012).

3. Jiménez-Escrig, A., Gobernado, I. & Sánchez-Herranz, A. Secuenciación de genoma completo: Un salto cualitativo en los estudios genéticos. *Rev. Neurol.* **54,** 692–698 (2012).

4. Rodríguez-Santiago, B. & Armengol, L. Tecnologías de secuenciación de nueva generación en diagnóstico genético pre- y postnatal. *Diagnostico Prenat.* **23,** 56–66 (2012).

5. Dalca, A. V. & Brudno, M. Genome variation discovery with high-throughput sequencing data. *Brief. Bioinform.* **11,** 3–14 (2010).

6. Bamshad, M. J. *et al.* Exome sequencing as a tool for Mendelian disease gene discovery. TL - 12. *Nat. Rev. Genet.* **12 VN - r,** 745–755 (2011).

7. Linderman, M. D. *et al.* Analytical validation of whole exome and whole genome sequencing for clinical applications. *BMC Med. Genomics* **7,** 20 (2014).

8. Nelson, A. C. *et al.* Criteria for Clinical Reporting of Variants from a Broad Target Capture NGS Assay without Sanger Verification. *JSM Biomarkers* **2,** 2–7 (2015).

9. Lim, E. C. P. *et al.* Next-generation sequencing using a pre-designed gene panel for the molecular diagnosis of congenital disorders in pediatric patients. *Hum. Genomics* **9,** 33 (2015).

10. Ross, M. G. *et al.* Characterizing and measuring bias in sequence data. *Genome Biol.* **14,** R51 (2013).

11. Benjamini, Y. & Speed, T. P. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res.* **40,** 1–14 (2012).

12. Guo, Y. *et al.* The effect of strand bias in Illumina short-read sequencing data. *BMC Genomics* **13,** 666 (2012).

13. Minoche, A. E., Dohm, J. C. & Himmelbauer, H. Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and Genome Analyzer systems. *Genome Biol.* **12,** R112 (2011).

14. Illumina. Quality Scores for Next-Generation Sequencing. *Http://Res.Illumina.Com/Documents/Products/Technotes/Technote_Q-Scores.Pdf* 1–2 (2011).

15. Rehm, H. L. *et al.* ACMG clinical laboratory standards for next-generation sequencing. *Genet. Med.* **15,** 733–47 (2013).

16. Motoike, I. *et al.* Validation of multiple single nucleotide variation calls by additional

exome analysis with a semiconductor sequencer to supplement data of whole-genome sequencing of a human population. *BMC Genomics* **15,** 673 (2014).

17. Ku, C. S. *et al.* Exome sequencing: Dual role as a discovery and diagnostic tool. *Ann. Neurol.* **71,** 5–14 (2012).

18. Dohm, J. C., Lottaz, C., Borodina, T. & Himmelbauer, H. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res.* **36,** (2008).

19. Viguera, E., Canceill, D. & Ehrlich, S. D. In vitro replication slippage by DNA polymerases from thermophilic organisms. *J. Mol. Biol.* **312,** 323–333 (2001).

20. Nakamura, K. *et al.* Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Res.* **39,** (2011).