Programa de doctorado en Biotecnología

Desarrollo de herramientas bioinformáticas aplicadas al Diagnóstico Genético mediante Secuenciación Masiva

Tesis doctoral

Sheila Zúñiga Trejos

Valencia, 2014

El Dr. Javier Benítez Ortiz, Director del Programa de Genética del Cáncer Humano del Centro Nacional de Investigaciones Oncológicas, en carácter de director de tesis, CERTIFICA que Dª. Sheila Zúñiga Trejos ha realizado bajo su supervisión el trabajo de investigación recogido en esta memoria que lleva por título ”Desarrollo de herramientas bioinformáticas aplicadas al Diagnóstico Genético mediante Secuenciación Masiva”.

Y para que así conste, en cumplimiento de la legislación vigente, firma la presente

En Valencia, a de de 2014

Dr. Javier Benítez Ortíz

A mi madre

AGRADECIMIENTOS

Esta tesis ha sido financiada íntegramente por la empresa Sistemas Genómicos S.L. a cuya gerente, Lorena Saus, estoy tremendamente agradecida por su gran apoyo en este largo camino recorrido hasta la finalización de este trabajo.

A Javier Benítez, por guiarme en el trabajo que presento en esta memoria. Gracias por tu dedicación, tu paciencia y tus consejos.

Al gran equipo de profesionales que comparten conmigo el día a día en la empresa y de los cuales he aprendido tantas cosas sin las cuales la realización de esta tesis hubiera sido imposible.

Quiero dar las gracias al Dr. Surralles por darnos la oportunidad de colaborar en su extraordinario trabajo sobre enfermedades raras y por su valioso feedback en el inicio del camino hacia la secuenciación de exomas.

Me gustaría disculparme con mi familia, especialmente contigo mamá, y amigos por no dedicarles más tiempo durante estos últimos años y agradecerles su cariño incondicional.

Por último, me gustaría darle las gracias a Paco, a mi Paco, por sus sonrisas, sus abrazos y su amor. Gracias por ayudarme a sobrellevar esta etapa tan difícil y exigente en mi carrera profesional.

ÍNDICE

INTRODUCCIÓN

En 1977 Fred Sanger y sus colaboradores publicaron dos artículos en los que describían un nuevo método de secuenciación del DNA que ha transformado la Biología de nuestros días [1]. Desde su publicación, la secuenciación Sanger ha evolucionado de forma asombrosa permitiendo alcanzar hitos tan relevantes para la Genética Humana como la secuenciación del primer genoma humano en el año 2003 [2], o la caracterización del primer haplotipo humano por el consorcio HapMap [3]. Sin embargo, la magnitud económica alcanzada por cada uno de estos hitos así como su dilatada duración en el tiempo dieron buena cuenta de la urgente necesidad de desarrollar nuevas tecnologías de secuenciación más baratas y eficientes. Como respuesta a este reclamo surgieron la primeras plataformas de secuenciación masiva abriendo una nueva era en las tecnologías de secuenciación y planteando un nuevo paradigma que se ha convertido en la consigna de este nuevo período de la secuenciación en el que actualmente nos encontramos inmersos, la secuenciación de un genoma por $1000 [4, 5].

1 Next-Generation Sequencing

La nueva generación de plataformas de secuenciación masiva o 'Next-Generation Sequencing' (NGS), lanzada al mercado en el año 2005, ha provocado una auténtica revolución en la investigación aplicada a ámbitos muy diversos de la Biología. Si bien es cierto que la secuenciación tradicional mediante la tecnología Sanger ha dominado el mercado durante al menos dos décadas, la vertiginosa evolución de estos nuevos secuenciadores, tanto en precisión como en rendimiento, así como el abaratamiento del coste por base han permitido la rápida expansión de su uso en la comunidad científica ofreciendo nuevas alternativas para la secuenciación de genomas completos, resecuenciación dirigida de zonas concretas del genoma, secuenciación de transcriptomas completo (RNA-Seq), identificación de "smallRNAs", estudios de interacción proteína-DNA (ChIP-seq) o estudios de metilación entre otros [6, 7].

A diferencia de la secuenciación Sanger, las plataformas de NGS permiten la obtención de miles o millones de fragmentos de DNA en un único proceso haciendo posible el desarrollo de complejos proyectos de secuenciación en cuestión de semanas. A pesar de las diferencias en su química todas estas plataformas de secuenciación masiva comparten las siguientes características:

- El DNA se fragmenta al azar y se ligan adaptadores específicos a ambos lados de cada molécula directamente generándose así una librería sin necesidad de clonar .

- La amplificación de la librería se produce mediante el anclaje del fragmento de DNA, a través de sus adaptadores, a una superficie sólida bien una esfera sintética o bien directamente a la placa de secuenciación.

- La secuenciación y detección de las bases ocurren al mismo tiempo en todas las moléculas de DNA (secuenciación masiva y paralela).

- Las lecturas generadas son cortas (150-400 pb). El ruido producido por estas tecnologías en relación con la señal que generan limita la obtención de lecturas de mayor longitud.

- Las plataformas NGS permiten realizar secuenciación de tipo “single-end” si solamente se lee uno de los extremos del los fragmentos de DNA o de tipo “paired-end” si se leen ambos extremos del mismo fragmento. La secuenciación tipo “paired-end” no solo facilita el posicionamiento de aquellas lecturas que pueden mapear en múltiples sitios sino que adem posibilita la identificación de variantes estructurales.

A pesar de todas estas particularidades comunes, cada plataforma de secuenciación masiva se basa en principios químicos distintos que generan diferencian notables entre ellas tanto cualitativa como cuantitativamente [8].

454-Roche fue la primera empresa en lanzar al mercado una plataforma de secuenciación masiva, cuya química está basada en la pirosecuenciación [9, 10]. La pirosecuenciación consiste en la detección de señales luminosas generadas a partir de grupos pirofosfato liberados tras la polimerización de un nuevo nucleótido complementario a una hebra de DNA molde. En sus inicios, pese a presentar grandes ventajas respecto a la secuenciación Sanger como el gran abaratamiento del coste de secuenciación por base (1/6 de lo que suponía en aquel entonces la secuenciación Sanger) y el gran aumento en la cantidad de información generada, equivalente a 50 secuenciadores Sanger (~25Mb), este primer secuenciador despertó muchas dudas en la comunidad científica respecto a la fiabilidad de las bases generadas, los problemas del análisis derivados de su corta longitud de lectura

(en aquel momento ~100nt), la dificultad que implicaba el manejo de un volumen de datos tan grande o incluso el elevado precio de su compra [11]. De nuevo, un gran salto tanto cualitativo como cuantitativo sorprendía a la comunidad científica al igual que ocurría con la aparición de las primeras plataformas de secuenciación capilar automatizadas basadas en el método Sanger unos años antes. Los secuenciadores de 454-Roche han evolucionado muy aprisa y actualmente la cantidad de información que generan es muy superior a su primera versión llegando a alcanzar 2Gb y una longitud de lectura de hasta 700pb dependiendo de la versión. Con el paso de los años otra serie de plataformas de secuenciación masiva han salido al mercado, entre ellas destacan las plataformas Illumina [12] y SOLiD [13]. La plataforma Illumina se basa en la incorporación de nucleótidos marcados con terminadores reversibles de manera que en cada ciclo de ligación solamente uno de los cuatro nucleótidos posibles se une de forma complementaria al DNA molde emitiendo una señal luminosa que es captada por un sistema óptico altamente sensible. Posteriormente, el terminador es eliminado para permitir la incorporación del siguiente nucleótido en ciclos sucesivos de secuenciación. La química empleada por Illumina permite generar lecturas de hasta 150pb llegando a producir hasta 600Gb de datos. El secuenciador SOLiD emplea una tecnología de ligación de oligonucleótidos marcados capaz de interrogar dos bases al mismo tiempo de manera que tras varias rondas de ligación y detección del fluoróforo cada nucleótido, excepto el primero y el último, es leído dos veces dando lugar a un nuevo tipo de codificación, que se denomina 'double-encode' o 'código de colores', en el que cada color identifica dos bases consecutivas. La longitud de las lecturas en la plataforma SOLiD alcanza los 75pb y puede llegar a generar hasta 200Gb de datos (Tabla 1).

La dificultad de reunir un mínimo número de muestras que pudiera justificar la puesta en marcha de un run completo unido al alto coste de adquisición de estos equipos impulsaron el desarrollo de una nueva serie de secuenciadores más económicos, capaces de generar una menor cantidad de datos con la misma precisión que las plataformas de secuenciación masiva más grandes y en menor tiempo. Roche lanzó el primer mini-secuenciador con tecnología NGS en el año 2010, el secuenciador GS Junior, que fue rápidamente seguido por el lanzamiento de MiSeq, de la compañía Illumina. Desde su lanzamiento, estas plataformas se han extendido rápidamente en la comunidad científica dotando a pequeños laboratorios de la tecnología de secuenciación más avanzada [14].

Tabla 1. Comparación entre las plataformas de secuenciación masiva más extendidas.

Proveedor	Plataforma y versión	Longitud de lectura máxima (nt)	Cantidad total de datos (Gb)	Tiempo de secuenciación	Tipos de librerías	Amplificación	Secuenciación	Modelo de error
Roche	454 - GS FLX+	700	1	18-20h	Fragmentos, paired-end y mate-pair	PCR en emulsión	Pirosecuenciación	Error aumenta en zonas de homopolímero y hacia el final de la lectura
	454 - GS FLX Titanium	400	0,5	10h				
	454 - GS Junior	400	0,05	10h				
LifeTech	SOLiD - 5500	75	200	8 días	Fragmentos, paired-end y mate-pair	PCR en emulsión	Ligación de oligonucleótidos marcados	Mayor error en zonas ricas en GC y hacia el final de las lecturas
	SOLiD - 4	50	100	12 días				
Illumina	Illumina - HiSeq2000	100	600	11 días	Fragmentos, paired-end y mate-pair	PCR en puente	Nucleótidos marcados con fluoróforos y con terminadores reversibles	Mayor error en zonas ricas en GC y hacia el final de las lecturas
	Illumina - HiSeq1000	100	300	8,5 días				
	Illumina - MiSeq	250	8,5	1 día				

En paralelo al desarrollo de esta nueva generación de plataformas, ha surgido un nuevo tipo de secuenciadores, todavía en sus fases iniciales de comercialización, que algunos denominan ya como la tercera generación [15, 16], que permite llevar a cabo la secuenciación de una única molécula de DNA (single-molecule sequencing) evitando la amplificación de los fragmentos de DNA mediante PCR. De esta forma, estas plataformas evitan las desviaciones generadas durante este proceso (generación de lecturas duplicadas e introducción de errores) al tiempo que reducen el precio total de secuenciación en gran medida. Un ejemplo de estas plataformas es el secuenciador de la empresa Pacific Biosciences, disponible comercialmente desde abril de 2011, basado en el uso de chips que contienen miles de pocillos en cuyo fondo se encuentra anclada una única proteína polimerasa que permite llevar a cabo la identificación de nucleótidos marcados a la velocidad de una enzima polimerasa en tiempo real (tecnología Single Molecule, Real Time o SMRT) [17]. Otro de los secuenciadores de tercera generación, aun no comercializado, es el de Oxford Nanopores [18] capaz de detectar nucleótidos individuales a su paso a través

de un nanoporo. Esta última generación de secuenciadores promete nuevas alternativas aún más baratas con las cuales poder resolver algunos de los hándicaps más limitantes asociados a los secuenciadores que producen lecturas más cortas como la posibilidad de identificar haplotipos, la capacidad de resolver zonas donde existen un gran número de repeticiones consecutivas o incluso la mejora en el posicionamiento de lecturas en zonas del genoma que presentan una alta homología con otras zonas como es el caso de genes que presentan pseudogenes o en familias génicas.

2 El impacto de la secuenciación masiva en la comunidad científica

Los avances intelectuales y tecnológicos tras la secuenciación del genoma humano han producido un profundo impacto en el progreso científico [19-22]. El éxito del proyecto original de secuenciación del genoma humano dio lugar al lanzamientos de mucho otros esfuerzos internacionales cuyo objetivo común ha sido el profundizar en el conocimiento de la variabilidad genética entre individuos [23]. Unos de los proyectos más importantes de la última década ha sido el proyecto de HapMap, iniciado en 2002, cuyo objetivo principal era mapear la diversidad haplotípica en el genoma humano mediante la determinación del genotipo de millones de variantes, su frecuencia y el grado de asociación entre ellas en muestras con distintos orígenes poblacionales. El proyecto de HapMap finalizó con la identificación de más de 8 millones de variantes comunes a lo largo del genoma, la mayoría de ellas localizadas mediante secuenciación Sanger o a través de plataformas de genotipado masivo [3, 24-27].

La llegada de las nuevas tecnologías de secuenciación ha aumentado generosamente la cantidad de datos y la velocidad de producción de los mismos. El proyecto “1000 genomes” fue pionero en emplear la secuenciación masiva de miles de individuos mediante tecnología NGS [28]. Desde su inicio en 2007, este proyecto ha logrado determinar la localización y frecuencias alélicas de más de 15 millones de SNVs, 1 millón de inserciones y deleciones y 20.000 variantes estructurales, la mayoría de ellas no descritas previamente [29]. Los resultados de este consorcio estiman que el 95% de la variabilidad de un individuo se encuentra presente en su base de datos y que cada persona lleva en su genoma entre 250-300 variantes de pérdida de función en genes anotados de las cuales, entre 50 y 100 variantes, se han asociado previamente a enfermedades hereditarias.

Durante los últimos años, han surgido otras muchas iniciativas internacionales que hacen uso de las nuevas plataformas de secuenciación como el “International Cancer Genome Consortium” (ICGC) [30], el proyecto “Cancer Genome Atlas o TCGA” [31], o el proyecto de ENCODE [32].

La gran cantidad de información generada por estos y otros consorcios de investigación ha sido depositada de forma paulatina en bases de datos públicas como dbSNP [33], que han visto cómo el número de variantes reportadas durante los últimos años aumentaba de forma exponencial con la llegada de las nuevas tecnologías de secuenciación (Figura 1). De la misma forma, el gran impacto provocado por la disponibilidad de esta valiosa información en bases de datos públicas así como la expansión del uso de estas plataformas gracias a la caída en el precio de secuenciación por base, se ve claramente reflejado en el enorme numero de publicaciones de diversa índole basadas en el uso de NGS durante los últimos años (Figura 2).

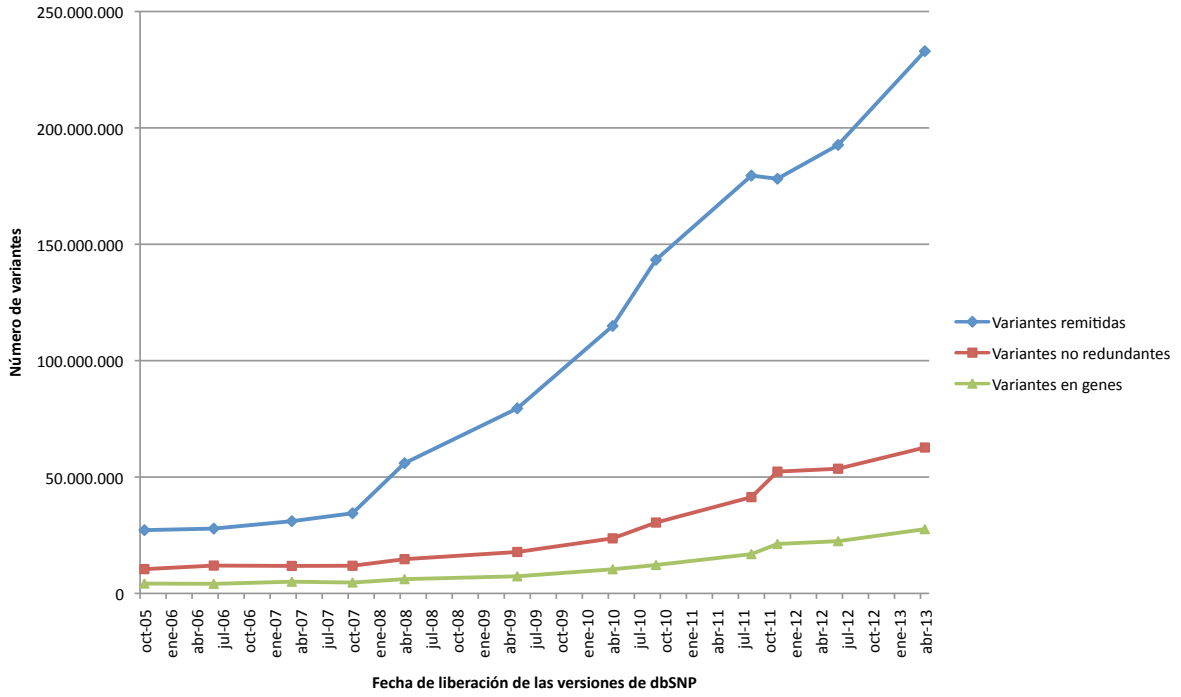


Figura 1. Crecimiento de la base de datos dbSNP a lo largo de los últimos 8 años. Desde la llegada de las plataformas de secuenciación masiva el número de variantes no redundantes y el número de variantes en genes se han multiplicado por 6 órdenes de magnitud.

El estudio y la integración de los datos 'ómicos' disponibles actualmente en diferentes bases de datos de dominio público han servido para la identificación de patrones genéticos comunes implicados en el desarrollo de enfermedades frecuentes en la población así como en la respuesta diferencial a fármacos o a determinados factores medioambientales [27]. La disponibilidad de la información generada por la comunidad científica de una forma organizada y estandarizada, es y será de vital importancia para poder realizar un pronóstico

precoz, un diagnóstico más preciso o incluso un tratamiento personalizado si bien antes será necesaria la secuenciación de un mayor número de individuos que abarquen un abanico poblacional más amplio y el desarrollo de nuevos métodos de análisis que permitan explotar la información obtenida de forma exitosa [34-36].

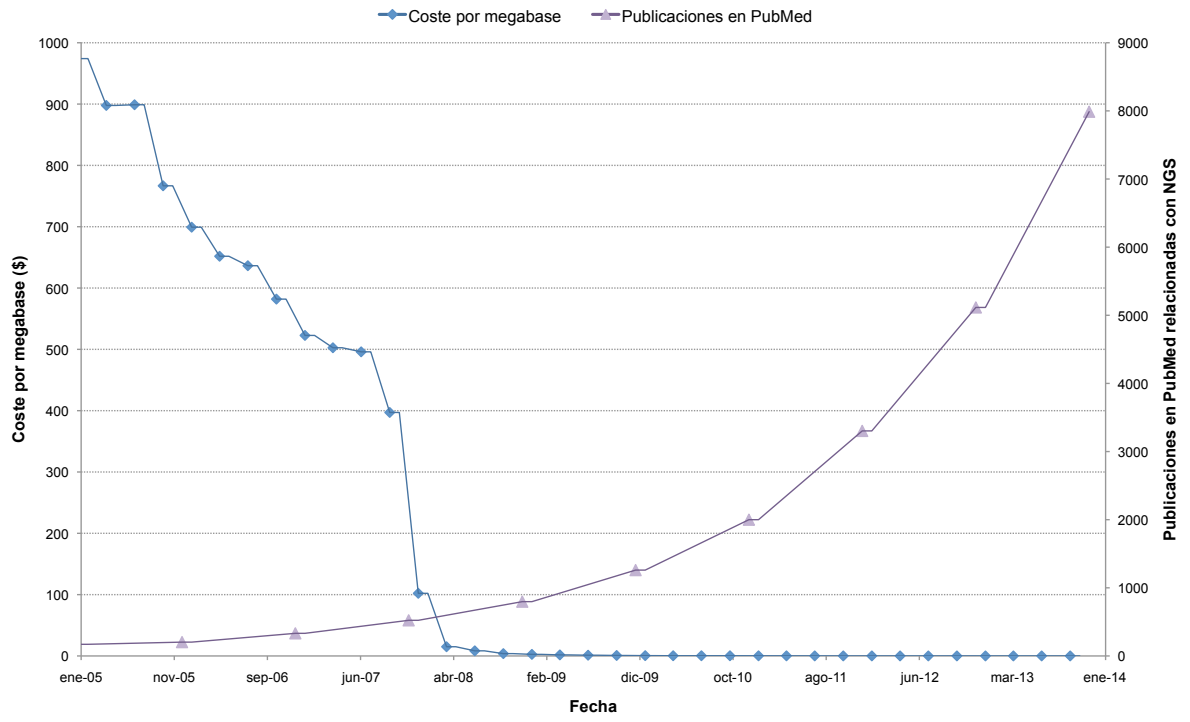


Figura 2. Precio de secuenciación por megabase y su repercusión en la comunidad científica. Los datos referentes al precio de secuenciación se muestran en dólares americanos y fueron tomados del National Human Genome Research Institute de los Estados Unidos. El número de publicaciones científicas reflejadas en el gráfico fue calculado atendiendo a la base de datos Pubmed y seccionando aquellos artículos relacionados con estudios en Humano que contuvieran alguna de las siguientes palabras: “Next-generation sequencing”, “high-throughput sequencing”, “new sequencing technologies”, “massive parallel sequencing” o “NGS”. Desde la llegada de las plataformas de secuenciación masiva en 2005, el precio de secuenciación por megabase ha disminuido de forma muy acusada permitiendo así expandir su uso en la comunidad científica. La gran repercusión que estos nuevos secuenciadores han tenido en campos muy diversos de la Genómica, Transcriptómica y Epigenómica se ve claramente reflejada en el número de publicaciones registradas en PubMed desde su aparición.

3 Resecuenciación dirigida y Diagnóstico Genético

La continua mejora de la química de secuenciación así como de los protocolos de trabajo y la automatización de los mismos han permitido perfeccionar la precisión y el rendimiento de las plataformas de secuenciación masiva dando lugar a un aumento en la cantidad de datos generados y a un descenso vertiginoso del coste de secuenciación por base. Sin embargo, pese a que actualmente el precio de secuenciar un genoma humano completo sea una ínfima parte de lo que costó la obtención del primer borrador en 2003, la secuenciación de rutina de genomas completos y el coste posterior de su análisis todavía sigue siendo económicamente inabordable para la mayoría de las instituciones [37]. Uno de los desarrollos que mayor impacto está teniendo en el estudio de genomas a gran escala mediante secuenciación masiva es la posibilidad de capturar zonas concretas del genoma, técnica conocida como resecuenciación dirigida o TargetSeq. Durante los últimos años se han desarrollado distintas técnicas de captura de ADN basadas en la hibridación de sondas o en sistemas de PCR multiplex [38-43] que permiten abarcar desde unas pocas kilobases hasta la captura del exoma completo, y que combinados con la secuenciación masiva hacen posible realizar un screening mutacional en cientos de genes a la vez, proporcionando una cantidad de información muy extensa sobre la genética de un individuo.

Antes de la llegada de las plataformas de secuenciación masiva, la investigación en enfermedades mendelianas, pese a ser una fuente extraordinaria de conocimiento había quedado relegada a un segundo plano debido a la falta de métodos de estudio más resolutivos [44]. Las técnicas empleadas tradicionalmente en el estudio de este tipo de enfermedades cuando la causa es desconocida como los análisis de ligamiento, el mapeo genético mediante homocigosidad o el cariotipado, han proporcionan una gran cantidad de información, sin embargo, cada una de estas técnicas individualmente no son capaces de identificar todos los tipos de variantes genómicas de manera que el éxito diagnóstico de la técnica elegida depende directamente del tipo de variante que provoca la enfermedad, parámetro que ‘a priori’ es difícilmente predecible [45]. Asimismo, la baja resolución de algunas de estas aproximaciones solamente permite acotar grandes regiones genómicas que en muchos casos incluyen un alto número de potenciales genes candidatos cuyo coste de validación mediante secuenciación Sanger, con el objeto de detectar la ausencia o presencia de mutaciones causales, es económicamente impracticable. A diferencia de los métodos tradicionales aplicados al estudio de genes candidatos en enfermedades mendelianas, la secuenciación del exoma completo mediante NGS ofrece la posibilidad de identificar un gran abanico de variantes de todo tipo distribuidas a lo largo de las zonas codificantes del genoma de un individuo de forma precisa y en un único ensayo. Durante los

últimos tres años, esta nueva técnica de screening del exoma ha sido aplicada con éxito en la identificación de nuevos genes candidatos principalmente en la investigación de desórdenes genéticos mendelianos y en especial en enfermedades raras, enfermedades aptas para estudios clásicos ya que solamente se dispone de un pequeño número de individuos afectos [46-50]. Otro de los campos de mayor aplicación de la secuenciación del exoma completo es en la investigación contra el cáncer donde se comparan muestras de tejido normal frente a muestras de tejido tumoral para la identificación de mutaciones somática [51-54].

Actualmente, el diagnóstico molecular en enfermedades mendelianas de origen conocido se realiza mediante la secuenciación bidireccional de exones y zonas de splicing mediante el método Sanger. Sin embargo, este procedimiento, ampliamente aceptado tanto por clínicos como por investigadores en genética, conlleva una serie de limitaciones en relación coste-eficiencia cuando se aplica a enfermedades con heterogeneidad genética donde no existe un único gen candidato sino un grupo de genes potencialmente candidatos. La posibilidad de diseñar sistemas de captura 'a medida' dirigidos a la selección de regiones específicas del genoma ha abierto una nueva perspectiva en el diagnóstico genético en este tipo de enfermedades [7, 55-58]. En contraposición a la secuenciación del exoma completo, la resecuenciación dirigida de un pequeño número de genes implica una reducción significativa en la cantidad de lecturas necesarias para la identificación de las variantes presentes con el consecuente abaratamiento en el coste por muestra. La llegada de mini-secuenciadores con tecnología NGS así como el continuo descenso de los costes de secuenciación por base forman la combinación ideal para la adopción de esta tecnología como método de preferencia en el diagnóstico genético de rutina en un futuro cercano.

4 Next-Generation bioinformatics

La transición desde la secuenciación tradicional automatizada de Sanger a plataformas con una mayor capacidad de producción de datos ha forzado el desarrollo de nuevos algoritmos y métodos de análisis necesarios para la correcta manipulación e interpretación de una nueva generación de datos de gran volumen y de una longitud muy inferior a la ofrecida por los secuenciadores tradicionales. Los continuos aumentos en los datos resultates de una carrera de secuenciación de estas máquinas, que se duplica aproximadamente cada 5 meses, ha sobrepasado con creces los recursos bioinformáticos disponibles hasta la fecha, suceso que algunos autores han denominado como "Next-Generation gap" [59]. Sin lugar a dudas, el desarrollo de las plataformas de Next-Generation Sequencing ha desencadenado en paralelo el desarrollo de un "Next-Generation Bioinformatics" que se enfrenta no solo a

un gran reto a nivel computacional sino también biológico ya que por primera vez es posible disponer de la información genética completa de un individuo. El desarrollo de herramientas bioinformáticas es crítico para la aplicación con éxito de la secuenciación masiva al diagnóstico genético. La nueva generación de secuenciadores requiere el desarrollo continuo de nuevas herramientas de análisis y de nuevas soluciones computacionales aplicadas tanto para el procesamiento de los datos como para su almacenamiento [60] [61]. Del mismo modo que la generación de datos es un proceso protocolizado y moderadamente estable para cada plataforma, el análisis de los datos no sigue unos estándares definidos y la combinación de diferentes piezas de software así como su integración con diferentes bases de datos proporcionan resultados muy distintos. La correcta identificación y caracterización de variantes es un proceso muy sensible y complejo no solo debido a la enorme cantidad de datos que han de manipularse al mismo tiempo sino también por la gran complejidad biológica que plantean este tipo de tecnologías. Actualmente, existen diversas herramientas de libre acceso que abordan pasos concretos del análisis global necesario para este tipo de datos. Todas estas herramientas pueden concatenarse entre sí a modo de tubería traduciendo el enorme flujo de datos de entrada en listados de variantes o genes candidatos mediante la generación de diferentes rutinas de trabajo o 'pipelines'. Sin embargo, la combinación de las distintas herramientas así como la elección de los parámetros adecuados en cada uno de los pasos del análisis producen resultados muy dispares. El desarrollo y establecimiento de pipelines eficientes, precisos y validados que permitan establecer y garantizar unos estándares de calidad mediante los cuales obtener una información depurada y fiable, y en un tiempo prudente, es una tarea compleja. El análisis de los datos plantea distintos retos que necesitan un exhaustivo conocimiento del funcionamiento de las plataformas de NGS así como de la biología del sistema en estudio si se quiere dar un respuesta adecuada.

La sorprendente aceleración en la generación de información genómica que proporcionan los instrumentos de secuenciación masivos no ha hecho más que empezar. Estas plataformas continúan evolucionando muy aprisa y la cantidad de información disponible se duplica por momentos. Sin lugar a dudas, la enorme cantidad de información generadas por los nuevos secuenciadores necesita la estandarización de protocolos de análisis de datos que permitan explotar los datos producidos de una forma eficiente, solo así será posible avanzar en el conocimiento sobre la complejidad del genoma humano y progresar así hacia un pronóstico más precoz, un diagnóstico más preciso y un tratamiento personalizado.

5 Objetivos de esta tesis

En esta tesis se describe el desarrollo y la generación de distintos protocolos de trabajo orientados a la investigación y diagnóstico genético de enfermedades mendelianas a partir de datos procedentes de dos de las plataformas de secuenciación masiva de última generación más extendidas en la comunidad científica entre 2009 y 2013, SOLiD y 454-Roche. La búsqueda de nuevos algoritmos, su integración, testeo y validación son los principales objetivos de este trabajo. Asímismo, esta tesis pretende servir de guía para la implantación de servicios de análisis de datos de estudios de resecuenciación dirigida en los que se cumplan con los estándares tradicionales aplicados en el diagnóstico genético marcando una serie de puntos de control y unos límites mínimos necesarios para alcanzar este objetivo.

En total, se analizaron tres conjuntos de datos (datasets) diferentes para los cuales se desarrollaron protocolos de análisis específicos, que abarcan desde la evaluación inicial del dato bruto hasta la identificación de las mutaciones causantes de la enfermedad o la priorización de genes candidatos y que responden a una serie de objetivos iniciales planteados a continuación.

5.1 Dataset 1: Análisis del exoma completo y su aplicación en el estudio de enfermedades raras

En Europa, se conocen como “enfermedades raras” aquellas cuya incidencia en la población es de 1 por cada 2.000 habitantes, lo que supone un total de ~250.000 personas afectadas. Actualmente, se han identificado alrededor de 7.000 enfermedades raras de carácter muy diverso. Se especula que la mayor parte de ellas se deben a la alteración de un único gen, alteraciones que se transmitirían posteriormente de generación en generación. La mayoría de las enfermedades raras alteran la vida del individuo que la padece de forma severa, algunas llegan a ser incluso letales [62]. El alto coste que supondría llevar a cabo los estudios necesarios para poder esclarecer la causa de cada una de estas enfermedades unido al pequeño porcentaje poblacional que las padecen dificultan enormemente su investigación. En la actualidad, entre el 57-71% de las enfermedades raras no tienen tratamiento. Sin embargo, es importante destacar que la investigación en enfermedades raras proporciona información de gran relevancia acerca de los mecanismos moleculares que se desencadenan a partir de la modificación de un determinado gen, información que posteriormente puede ser aplicada al estudio de enfermedades más complejas y extendidas en la población así como en el desarrollo de nuevas dianas terapéuticas [44, 63].

La secuenciación del exoma completo se plantea como la alternativa más efectiva actualmente en el estudio de enfermedades mendelianas raras, no solo por motivos económicos sino principalmente por motivos de eficiencia ya que la mayor parte de desórdenes mendelianos conocidos son causados por la aparición de mutaciones en zonas codificantes del genoma.

Los objetivos planteados para este primer dataset son:

- Generación de un pipeline robusto para el análisis del exoma completo.
- Aplicación del pipeline para la identificación de nuevos genes candidatos en enfermedades raras.
- Identificación de las ventajas y limitaciones del uso del exoma completo aplicado al diagnóstico genético.

5.2 Dataset 2: Desarrollo y análisis de un panel de genes orientado al Diagnóstico Genético de enfermedades cardiovasculares

Según la Organización Mundial de la Salud (OMS), las enfermedades cardiovasculares son la principal causa de mortalidad en el mundo. En 2008, el 30% de las defunciones se debieron a alguna enfermedad cardiovascular. Un alto porcentaje de los individuos diagnosticados de enfermedades cardiacas fallecen de forma súbita. Sin embargo, existe un grupo de individuos no diagnosticados, y en apariencia sanos, en los que este colapso cardiovascular inesperado es la primera manifestación de la enfermedad cardiaca subyacente [64-66]. Algunas de estas patologías son de origen hereditario siendo de gran importancia la identificación de la mutación causal tanto en el paciente como en los familiares en riesgo con el fin de establecer medidas preventivas. Sin embargo, la heterogeneidad genética que presentan este tipo de patologías, definiéndose como tal cuando un mismo fenotipo o cuadro clínico puede ser producido por mutaciones en diferentes genes siguiendo modelos de transmisión monogénico, provocan que el éxito en el diagnóstico para algunas de ellas tras el screening inicial de uno o varios de los genes mutados con mayor frecuencia sea muy bajo. La llegada de las plataformas de secuenciación masiva ha abierto nuevas posibilidades en el diagnóstico de enfermedades con heterogeneidad genética permitiendo el cribado de un gran número de genes a la vez en un mismo individuo de forma integral, rápida, eficiente y económicamente viable [67-69].
Los objetivos prioritarios planteados en este dataset son:

- Generar y evaluar el diseño de sondas de un panel de genes dirigido al diagnóstico de enfermedades cardiovasculares heterogéneas con riesgo de muerte súbita.

- Desarrollar una estrategia de análisis altamente sensible y específica para valorar la posibilidad de trasladar esta tecnología a la práctica clínica de rutina.

- Determinar las limitaciones y ventajas tecnológicas de esta metodología de estudio respecto a la secuenciación Sanger y frente al análisis del exoma completo.

5.3 Dataset 3: análisis de datos de mini-secuenciadores con tecnología NGS

La presencia de mutaciones deletéreas en los genes *BRCA1* y *BRCA2* confieren un alto riesgo de padecer cáncer de mama y ovario. Se estima que aproximadamente el 12% de la población general desarrollará cáncer de mama en algún momento de la vida. Sin embargo, esta tasa de incidencia poblacional es significativamente mayor en los portadores de mutaciones en los genes BRCA alcanzando valores de entre 40-87%. De la misma forma, la presencia de mutaciones en estos dos genes elevan la probabilidad de padecer cáncer de ovario, que en la población general se sitúa en torno al 1,4%, llegando a alcanzar valores de entre 16-68% si las mutaciones se encuentra en el gen *BRCA1* y alrededor del 11-27% si las mutaciones se localizan en el gen *BRCA2* [70]. Entre un 5 y un 10% de los casos de tumores de mama y ovario presentan un claro componente hereditario, es decir, existe un gen que confiere una alta susceptibilidad al desarrollo del cáncer cuando se encuentra mutado, siendo mayor su incidencia en familias con casos múltiples de cáncer [71, 72]. La identificación de mutaciones patogénicas en el paciente así como en los familiares a riesgo es de gran importancia para poder prevenir el desarrollo de la enfermedad y mejorar su tratamiento.

Debido al alto coste de secuenciación por base en la tecnología de Sanger, es común realizar alguna técnica de cribado mutacional inicialmente con DHPLC (Denaturing High-Performance Liquid Chromatography) [73] o HRMCA (High-Resolution Melting Curve Analysis) [74] para localizar la mutación y posteriormente secuenciarla por Sanger. A menudo estas técnicas tradicionales son laboriosas y poco eficientes haciendo necesaria finalmente la secuenciación del gen completo. Los mini-secuenciadores con tecnología NGS ofrecen nuevas y excitantes oportunidades para el diagnóstico genético de rutina a través de la secuenciación de paneles de genes de pequeño tamaño de una forma mucho más rentable, y en ocasiones más eficiente, que las técnicas tradicionales de diagnóstico. Su menor coste de adquisición, en comparación con las plataformas NGS de mayor tamaño, unido a la reducción en el número de muestras necesarias para comenzar una carrera

completa dada su menor capacidad de producción de datos han favorecido la introducción de estas novedosas tecnologías de secuenciación en pequeños y medianos laboratorios [75].

Los objetivos específicos planteados en este dataset son:

- Desarrollar una estrategia de análisis adaptada al estudio de los genes BRCA a partir de datos de pirosecuenciación.
- Determinar las ventajas y limitaciones de esta tecnología respecto a la secuenciación tradicional.

Los datos utilizados para el desarrollo de los diferentes pipelines recogidos en esta memoria provienen de muestras cedidas para su análisis por la empresa Sistemas Genómicos, de la cual esta autora es parte desde el año 2006, para el desarrollo de esta tesis doctoral.

MATERIALES Y MÉTODOS

1 Análisis del exoma completo y su aplicación en el estudio de enfermedades mendelianas

El pipeline de análisis presentado a continuación se desarrolló en base a los resultados obtenidos en una línea controlada de HapMap. Posteriormente, este protocolo de análisis fue aplicado para la identificación de nuevos genes candidatos en un paciente con una enfermedad rara, Anemia de Fanconi (FA). Esta enfermedad se caracteriza principalmente por un fallo en la médula ósea (trombocitopenia, pancitopenia progresiva) acompañado de malformaciones congénitas de distintos tipo (esqueléticas, urogenitales, renales, cardiacas, hiperpigmentación), hipersensibilidad a agentes que producen "DNA cross-linking" y una mayor predisposición al cáncer. Actualmente, se conocen 15 genes causantes de la aparición de FA, sin embargo las bases genéticas en algunos pacientes con FA, como el paciente analizado en esta tesis, todavía son desconocidas.

1.1 Preparación de las muestras y secuenciación

La preparación de las muestras y posterior secuenciación fue llevado a cabo por el departamento de Nuevas Tecnologías de la empresa Sistemas Genómicos.

1.1.1 Obtención del material biológico

La línea celular de HapMap empleada para poner a punto el diseño inicial del pipeline para el exoma completo fue NA12144, con código en el repositorio celular de Coriell GM12144. Esta línea celular correspondía a un individuo varón de 71 años de la población CEPH (residente en Utah, Estados Unidos, con ancestros del norte y oeste de Europa), previamente caracterizado por el consorcio de HapMap mediante plataformas de genotipado masivo. El material de partida de esta línea celular fue 3μg de DNA.

Una vez realizado el desarrollo del pipeline en la muestra control, el protocolo de análisis se aplicó al exoma de un individuo español afecto de FA en el que se habían descartado previamente mutaciones patogénicas en los genes más comunes asociados con esta enfermedad. La secuenciación del exoma completo en este paciente se realizó a partir de DNA extraído de sangre periférica según métodos convencionales.

1.1.2 Construcción de librerías

La captura de zonas específicas del DNA requiere la construcción de librerías de fragmentos que consisten en la fragmentación del DNA en segmentos de entre 150 y 200pb de longitud a los que posteriormente se ligan adaptadores específicos a ambos extremos. Los adaptadores sirven de punto de inicio para la secuenciación de forma unidireccional o bidireccional (Figura 3). Asimismo, estos adaptadores son utilizados para anclar la molécula de DNA a una esfera sintética, a partir de la cual se amplificará el fragmento de DNA, y se fijará la molécula de DNA amplificada a la superficie de la placa de secuenciación.

Figura 3. Estructura de las librerías de fragmentos. Las librerías de fragmentos ofrecen dos posibilidades distintas para la secuenciación a partir de los adaptadores incorporados a ambos lados del DNA (Adaptador 1 y Adaptador 2). Si la molécula de DNA es leída solamente por el extremo más cercano a la esfera sintética se producirá una única lectura por fragmento original de DNA, este tipo de secuenciación se denomina 'single-end'. Si por el contrario, la molécula de DNA es leída por ambos extremos, se generarán dos lecturas en cadenas opuestas, obteniéndose lecturas emparejadas o 'paired-end'.

1.1.3 Captura de zonas específicas

La captura y enriquecimiento del exoma se realizó mediante el uso del kit comercial SureSelect All Exons para 38Mb de Agilent [76]. Este sistema permite la selección de aproximadamente un 1.22% del tamaño total del genoma a través de sondas de RNA biotiniladas de 120pb de longitud. El diseño de estas sondas se basó, según los datos proporcionados por el fabricanteel fabricante, en la información genómica disponible en la base de datos 'Collaborative Consensus Coding Sequence' o CCDS en septiembre del año 2008. Asimismo, el kit de captura incluía también más de 700 miRNAs de la base de datos de Sanger v13, además de otros 300 RNAs no codificantes como snoRNAs o scaRNAs. Mediante este sistema de captura, los híbridos formados por las sondas de RNA del kit y el DNA diana son capturados a partir del pool inicial de fragmentos de DNA mediante la incubación con esferas marcadas con estreptavidina, las cuales son posteriormente apresadas a su vez por un potente sistema magnético. Los híbridos esfera-DNA capturados

por el sistema magnético, son lavados para eliminar hibridaciones inespecíficas. Posteriormente, las sondas de RNA son digeridas de manera que solamente permanece el DNA correspondiente a las zonas de interés (Figura 4).

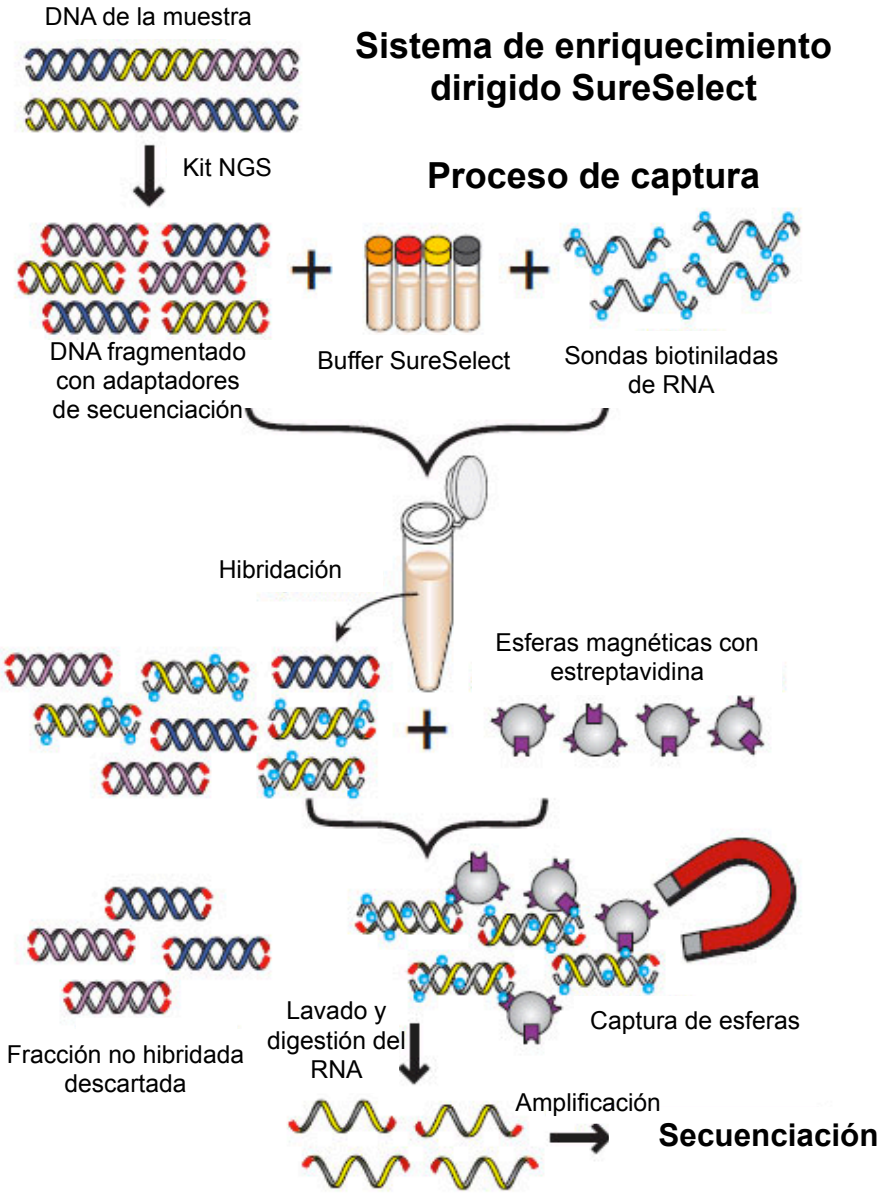


Figura 4. Protocolo de captura y enriquecimiento del exoma completo mediante el kit SureSelect (Imagen tomada de [76]).

1.1.4 Amplificación del DNA y enriquecimiento de esferas

Una vez capturado el DNA correspondiente a las zonas de interés, se lleva a cabo la amplificación de los fragmentos de DNA. En lugar de utilizar un método de clonación biológico como en la secuenciación Sanger, la amplificación del DNA en la plataforma SOLiD se lleva a cabo mediante PCR en emulsión de manera que se generan una serie de gotas o 'microrreactores' formados por una mezcla de aceite y una fase acuosa que contiene los reactivos de amplificación (primers, polimerasas, etc) [77]. En cada microrreactor tiene lugar un proceso de amplificación independiente a partir de una sola molécula de DNA de cadena simple, generándose miles de copias en cada microrreactor que permanecerán unidas a la esfera sintética de secuenciación a través de los adaptadores. Tras este proceso, se eliminan aquellas esferas en las que la amplificación no se ha llevado a cabo correctamente como es el caso de las esferas policlonales, que contienen más de un fragmento distinto de DNA, o aquellos microrreactores en los que no se ha producido la amplificación enriqueciendo así la muestra en esferas que contienen un único clon de DNA (esferas monoclonales). Terminada la fase de enriquecimiento, las esferas purificadas se unen de forma covalente a la placa de secuenciación a través de la modificación de los extremos 3' de las moléculas de DNA molde amplificadas.

1.1.5 Secuenciación masiva con SOLiD™

Las muestras presentadas en este estudio fueron secuenciadas con el sistema de secuenciación SOLiD™ (**S**equencing by **O**ligonucleotide **Li**gation and **D**etection). La plataforma SOLiD™ se basa en la ligación secuencial de oligonucleótidos marcados que permiten la interrogación de dos nucleótidos al mismo tiempo [19, 78]. El sistema utiliza cuatro fluoróforos distintos para codificar las dieciséis combinaciones posibles de dos bases dando lugar a un código de colores ('color-space') que caracteriza este tipo de secuenciación (Figura 5).

A)

	A	C	G	T
A	0=AA	1=AC	2=AG	3=AT
C	1=CA	0=CC	3=CG	2=CT
G	2=GA	3=GC	0=GG	1=GT
T	3=TA	2=TC	1=TG	0=TT

B)

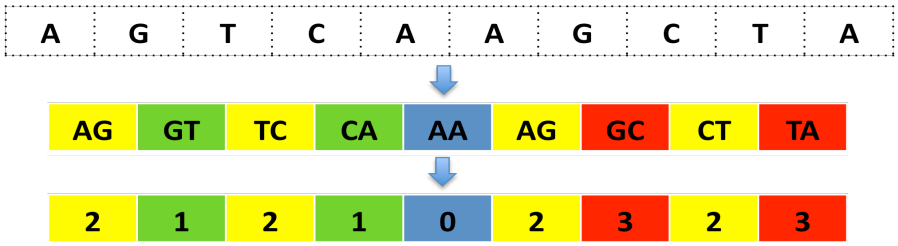


Figura 5. Secuenciación 'color-space'. A) Matriz color-space. El sistema de secuenciación emplea cuatro fluoróforos distintos que provienen de la combinación de dos bases de manera que un mismo fluoróforo pertenece a 4 dinucleótidos distintos. B) Decodificación de una secuencia en nucleótidos al sistema 'color-space'. La transformación de bases a colores es posible siguiendo la matriz de la figura 5A, de manera que la secuencia AGTCAAGCTA quedaría codificada en colores como 212102323.

- dinucleótidos A,G,T,C posibles entre las bases 1 y 5. La ligasa empleada por el sistema SOLiD necesita una longitud de oligonucleótido de al menos 8 bases, sin embargo, solo existe fidelidad de la enzima en las 5 primeras bases, de manera que solamente el octámero cuyas 5 primeras bases sean idénticas al DNA molde es ligado a la cadena en síntesis.
- Bloqueo (capping) de los extremos 5' fosfato que no han sido extendido y recogida de las señales de fluorescencia emitidas por el fluoróforo ligado al octámero (Figura 6C).
- Eliminación del fluoróforo mediante la eliminación de las tres últimas bases del extremo 5' (Figura 6D). Tras este proceso, las 5 primeras bases del oligonucléotido quedan

ancladas al DNA molde, 2 bases complementarias al DNA molde y 3 bases degeneradas.

- El proceso se repite un número determinado de veces, dependiendo de la longitud de lectura deseada, empleando distintos cebadores, cuya longitud es n-1, n-2, n-3 y n-4 respecto al extremo 3' del adaptador P1, y distintas sondas puente. Así, durante la primera ronda de primers, en el primer ciclo de ligación, se leen las bases 1 y 2 de todos los fragmentos de DNA en paralelo (Figura 6E y Figura 7).

La secuenciación en sentido reverso tiene lugar a partir del adaptador P2, ubicado en el extremo opuesto del primer P1. Las lecturas obtenidas desde este extremo del fragmento de DNA se denominan F5. Este tipo de secuenciación implica un paso adicional tras la eliminación del fluoróforo, la defosforilación del extremo 3' del oligonucleótido ligado para generar un extremo hidroxilo y proceder con el resto de pasos de la secuenciación (Figura 6 F-K).

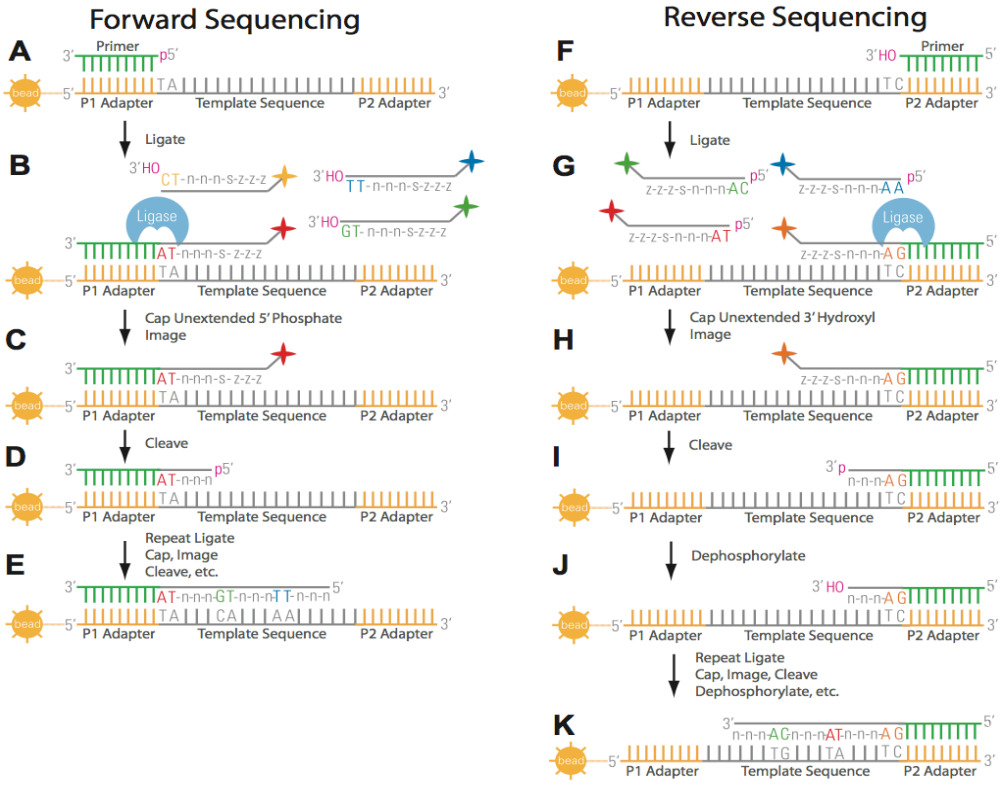


Figura 6. Secuenciación en el sistemas SOLiD [imagen tomada de [79]].

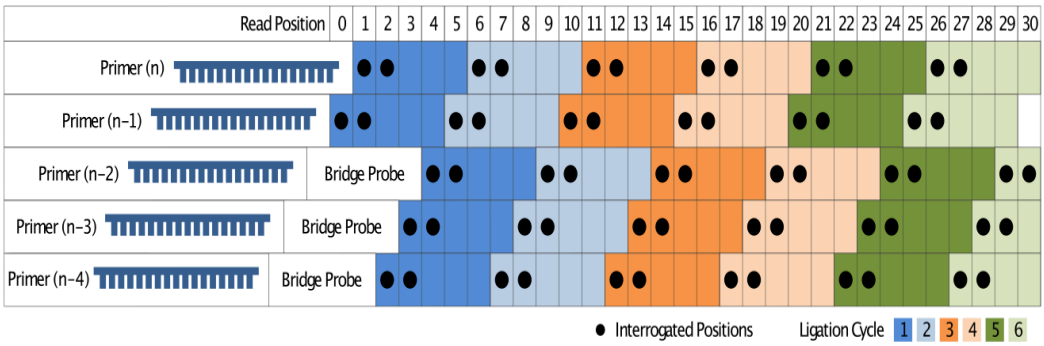


Figura 7. Bases interrogadas en cada ronda de primers y en cada uno de los ciclos de ligación [imagen tomada de [79]].

La secuenciación en "color-space" permite diferenciar entre errores de secuenciación y verdaderos polimorfismos. La Figura 8 muestra un claro ejemplo sobre el poder de detección de variantes de la codificación color-space exclusiva del sistema SOLiD.

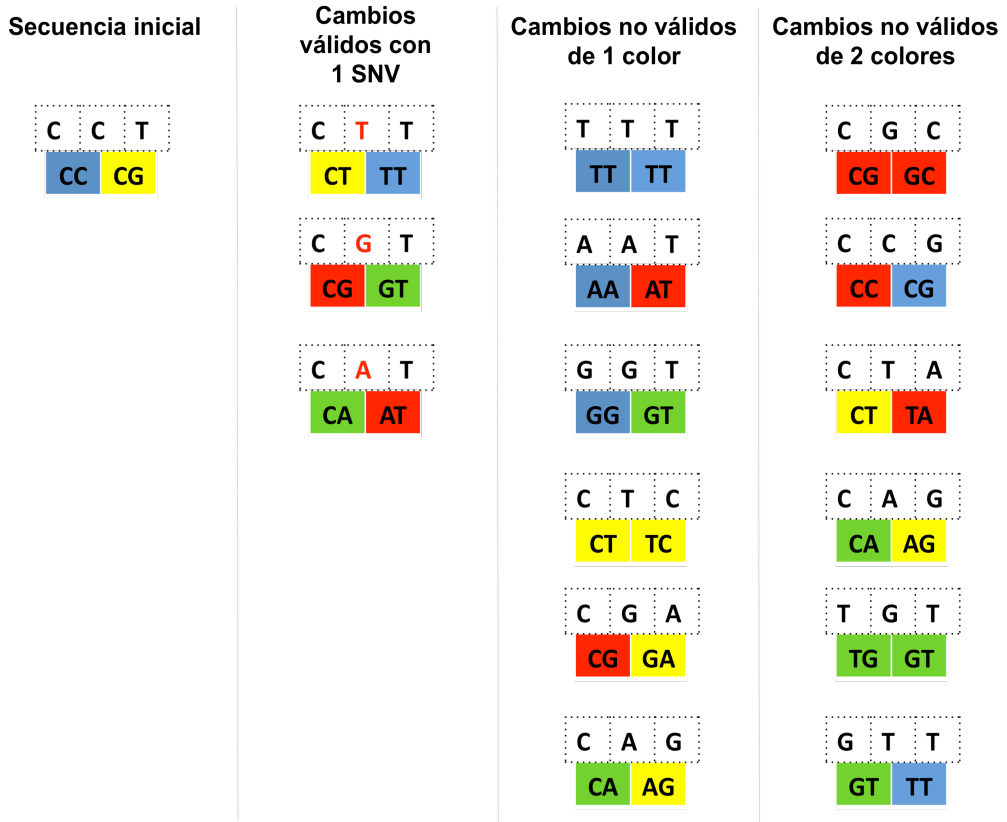


Figura 8. Precisión del sistema de secuenciación SOLiD en cuanto a la distinción entre verdaderas variantes y errores de secuenciación. Una modificación en la secuencia inicial CCT supondría un cambio de dos colores consecutivos de manera que solamente 3 de las 15 combinaciones restantes posibles podría ser válida.

El número de lecturas generadas para cada muestra depende directamente del número de esferas depositadas en la placa de secuenciación de manera que, partiendo del mismo tipo de librería, a mayor número de esferas depositadas mayor cantidad de lecturas se generarán. En el caso de la versión 4 de la plataforma de secuenciación SOLiD, existen dos placas de secuenciación, que pueden dividirse en espacios físicos independientes (cuartos u octavos) en los que se pueden depositar diferentes muestras, optimizándose así el espacio de secuenciación. Las muestras analizadas en este estudio se depositaron en un cuarto de placa de secuenciación cada una. La secuenciación se llevó a cabo siguiendo el primer protocolo de laboratorio de LifeTecnologies para la secuenciación de tipo 'paired-end' con SOLiD v4, obteniéndose varios millones de lecturas emparejadas por muestra. La longitud de las lecturas en 'forward' (lecturas F3) fue de 50pb mientras que la longitud de las lecturas leídas en 'reverse' (lecturas F5-P2) fue de 25nt.

1.2 Análisis de los datos

El análisis de datos para muestras procedentes de estudios de resecuenciación dirigida, y en particular para muestras de exoma, puede dividirse las siguientes fases: 1) control de calidad del dato bruto; 2) alineamiento; 3) procesado post-alineamiento; 4) identificación de variantes; 5) anotación de variantes; 6) priorización de genes candidatos (Figura 9). Cada una de estas etapas del protocolo de análisis conlleva el diseño y la validación de una serie de herramientas de código libre y la generación de distintos scripts propios a partir de los cuales se ejecutan los distintos pasos del análisis de manera secuencial.

Las siguientes secciones describen con detalle la generación del pipeline de análisis empleado para el estudio del exoma completo. El diagrama incluido en la última de las secciones de este apartado, muestra de manera global el pipeline desarrollado así como los diferentes programas, parámetros y criterios de filtrado empleados.

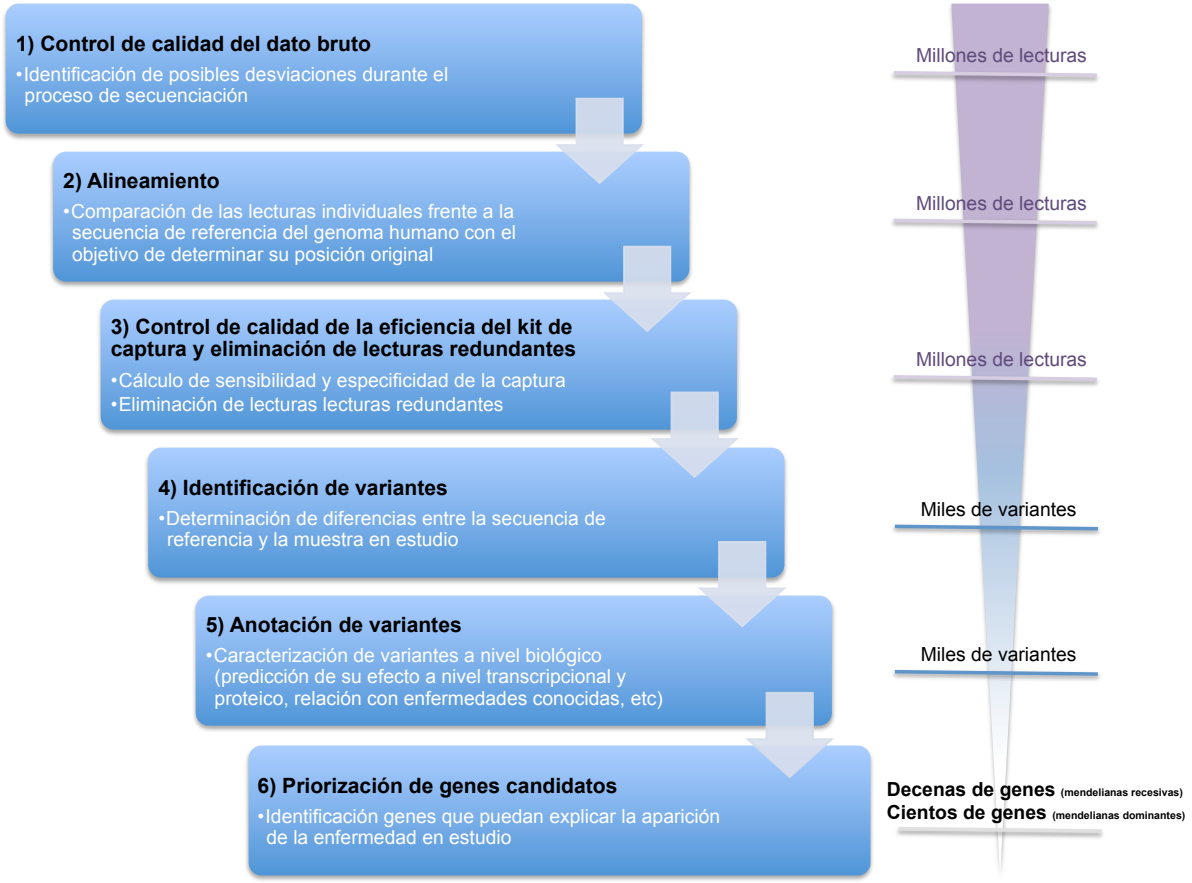


Figura 9. Pipeline de análisis de datos en muestras de exoma. A lo largo de las distintas etapas que componen el protocolo de análisis, la complejidad del dataset inicial va disminuyendo progresivamente hasta terminar en un listado final de genes candidatos priorizados en función del modelo de herencia vinculado a la enfermedad en estudio.

1.2.1 Control de calidad de los datos de secuenciación

De forma similar a la secuenciación tradicional mediante el método Sanger, en la secuenciación masiva se genera un valor de probabilidad de error para cada nucleótido secuenciado. El algoritmo Phred desarrollado por Edwin y Green para la secuenciación Sanger, ha sido un estándar hasta la fecha para establecer el grado de fiabilidad de las bases [80, 81] . Actualmente, todos los algoritmos empleados para la identificación de las bases ('base-calling') en las plataformas de secuenciación masiva emplean métodos de cálculo similares a Phred. Según este algoritmo, el valor de calidad vinculado a cada una de las bases generadas (Q) corresponde al cálculo de la probabilidad de error estimado para esa base (P_e) en escala logarítmica que obedece a la siguiente fórmula:

$$Q = -10 \log_{10}(P_e)$$

El análisis global del conjunto de valores de calidad en un dataset permite la rápida detección de desviaciones producidas durante el proceso de secuenciación facilitando así la toma de decisiones acerca de si una determinada muestra debe seguir el flujo de análisis de datos natural o debe volver a secuenciarse para garantizar unos estándares mínimos de calidad. Este proceso es de vital importancia en cualquier proyecto de secuenciación ya que el aumento del número de errores generados durante la secuenciación genera una mayor tasa de falsos positivos y negativos en la identificación de variantes.

El control de calidad en las muestras incluidas en este estudio contempló el estudio de los siguientes valores:

- Proporción de bases en función del valor de calidad.
- Proporción de bases respecto a los valores de calidad atendiendo al ciclo de ligación y primer empleado.

El control de calidad de los datos se basó en los resultados proporcionados por la herramienta SOLiD Experimental Tracking Software o SETs v4.0.1 [79], que permite visualizar de forma rápida desviaciones graves durante la secuenciación.

1.2.2 Alineamiento

El alineamiento o mapeo de las lecturas consiste en la comparación de cada una de ellas frente a una secuencia de referencia con el objetivo de identificar la posición genómica original de donde proviene ese fragmento de DNA secuenciado.
Los programas utilizados tradicionalmente para el alineamiento de lecturas obtenidas por secuenciación Sanger no permiten un buen rendimiento cuando se trata de datos de NGS. La gran cantidad de lecturas generadas por muestra unido a su corta longitud y las diferencias en su codificación, como es el caso de SOLiD, han impulsado el desarrollo de diferentes programas de alineamiento optimizados específicamente para las distintas plataformas de secuenciación masiva [82].
Desde el punto de vista computacional, el alineamiento es el paso del análisis más costoso ya que dependiendo del algoritmo de mapeo empleado, el tipo de secuenciación llevada a cabo y la cantidad de lecturas generadas por muestra se requiere una mayor o menor capacidad de computación [83]. Generalmente, este tipo de análisis se realiza en infraestructuras informática avanzadas tipo cluster, sistemas en los que distintos ordenadores se conectan entre sí para trabajar al unísono proporcionando los requerimientos computacionales necesarios para este tipo de trabajos [60, 84].

Los algoritmos de mapeo desarrollados para NGS comparten las siguientes características:

- Alinean de forma eficiente millones de lecturas cortas.
- Permiten el alineamiento de lecturas que incluyen errores de secuenciación o variantes y que por lo tanto no mapean de manera exacta frente a la referencia asumiendo un número de desapareamientos (mismatches) determinado.
- Permiten el alineamiento de secuencias que no son únicas en el genoma de referencia gracias al uso de lecturas paired-end.
- Han adoptado el concepto de "valor de calidad de mapeo" introducido por el algoritmo MAQ en el año 2008 [85]. De una forma similar al valor de calidad obtenido por base durante el proceso de "base calling" (ver sección 1.2.1), el valor de calidad de mapeo mide la probabilidad de error de que una lectura haya sido mapeada de forma incorrecta. El valor de calidad de mapeo estima la diferencia entre la primera y la segunda posición más probable de mapeo. Cuanto mayor sea la longitud del alineamiento y menor sea el número de desapareamientos frente a la secuencia de referencia, mayor es la probabilidad de que esa localización genómica sea correcta y por lo tanto mayor será el valor de calidad. El valor de mapeo se reporta en escala Phred siendo 0 el valor que corresponde a lecturas que mapean con la misma probabilidad en múltiples sitios o a lecturas con un gran número de desapareamientos frente a la referencia.
- Generan los resultados en un formato estándar denominado SAM, Sequence Alingment/Map format [86]. Este formato fue diseñado por el proyecto de los 1000Genomes para almacenar toda la información correspondiente al alineamiento de las lecturas de una forma estándar independientemente del tipo de secuenciación, tipo de librería construida o programa de alineamiento utilizado. El fichero en formato SAM reporta para cada lectura la siguiente información sobre su alineamiento:
  - QNAME: Nombre de la lectura o de la pareja de lecturas.
  - FLAG: indicador del estado de la lectura tras el alineamiento mediante el uso de caracteres ASCII con el fin de proporcionar la mayor cantidad de información posible ocupando el menor número de caracteres. El código incluido en esta columna describe si la lectura está mapeada, si está emparejada, la cadena en la que se encuentra, la cadena en la que se encuentra la lectura con la que forma pareja, etc.
  - RNAME: Nombre de la secuencia de referencia en la que mapea la lectura.
  - POS: Posición genómica menos una base a la izquierda en la que comienza a mapear esa lectura.

- MAPQ: Valor de probabilidad de mapeo en escala Phred.
- CIGAR: este campo describe el alineamiento generado proporcionando información acerca del número de mismatches, inserciones, deleciones, etc.
- RNEXT: Nombre de la secuencia de referencia en la que mapea la lectura con la que forma pareja. Si la secuencia donde mapean es la misma entonces este campo tiene el valor "=".
- PNEXT: Posición genómica menos una base a la izquierda donde mapea la lectura con la que forma pareja.
- TLEN: Distancia a la que se encuentran las parejas de lecturas. Si el orden y la orientación de las lecturas es correcto este valor reflejaría el tamaño de inserto real de la librería generada.
- SEQ: Secuencia de la lectura en nucleótidos.
- QUAL: Valores de calidad en escala Phred asociados a cada base de la lectura.

Bioscope es una suite de herramientas generada por LifeTechnologies (empresa que comercializa el secuenciador SOLiD) que incluye uno de los algoritmos de mapeo disponibles y más optimizados para lecturas en "color-space". Bioscope v1.2 realiza un mapeo local o parcial de las lecturas en lugar de pretender el mapeo global de las mismas de manera que facilita la exclusión de las bases con mayor error, que normalmente se localizan al final de las lecturas. Bioscope requiere la generación de un fichero de configuración inicial a partir del cual se lanza el programa de mapeo. Los requisitos de configuración para lanzar la aplicación requieren la optimización de ciertos parámetros siendo la definición de los esquemas de mapeo el parámetro más relevante. El algoritmo comienza a mapear las lecturas empleando diferentes esquemas que subdividen las lecturas en fragmentos de menor tamaño en los cuales se permite un número de desapareamientos o mismatches determinado. Así, el uso de un esquema 25.2.25 implicaría que: 1) la longitud del fragmento inicial que se va a mapear, o longitud del "seed", es de 25n; 2) el número máximo de mismatches permitidos respecto a la secuencia de referencia es de de un máximo de 2nt en este segmento de 25nt; 3) el fragmento de lectura seleccionado correspondería a los nucleótidos desde el 25 hasta el nucleótido 50 (Figura 10). El programa permite establecer un número de esquemas de mapeo personalizado si bien es cierto que cada uno de estos esquemas corresponde a una ronda de mapeo y que el incremento en su número, pese a generar un mayor número de lecturas mapeadas, incrementa el tiempo de computación tantas veces como número de esquemas de mapeo se empleen.

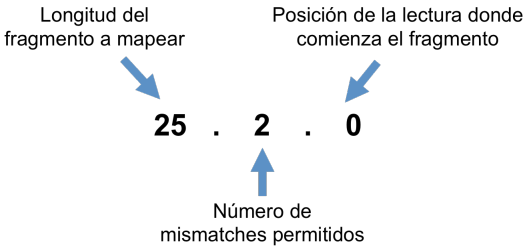


Figura 10. Descripción de los esquemas de mapeo empleados por Bioscope.

Según el algoritmo de Bioscope, una lectura se considera mapeada si existen menos de *n* posiciones distintas, definidas por el usuario, en el genoma usado como referencia donde la lectura pueda encajar con un mayor o menor número de desapareamientos establecidos en los esquemas de mapeo. Una vez identificadas las posibles posiciones de origen de esa lectura el alineamiento se extiende de forma local con el fin de alcanzar la mayor longitud de alineamiento posible respecto a la referencia, este proceso se denomina “seed and extend”. Para seleccionar la posición genómica más probable para cada lectura se utiliza el valor de calidad del mapeo. En el caso de lecturas emparejadas, el algoritmo alinea primero cada grupo de lecturas individualmente y posteriormente calcula un valor de probabilidad de que el mapeo sea correcto para la pareja de lecturas. De esta forma, si una lectura tiene 10 posibles zonas de alineamiento en la secuencia de referencia y su pareja también cuenta con 10 posibles lugares de mapeo se podrían generar hasta 100 alineamientos distintos combinando la misma pareja de lecturas. El valor de calidad de mapeo de la pareja de lecturas depende de los siguientes parámetros: longitud de alineamiento, número de mismatches, distancia entre ambas lecturas atendiendo al tamaño de inserto de la librería y número total de alineamientos posibles.

Para la correcta selección de los esquemas de mapeo es necesario tener en cuenta:

a) Los esquemas de mapeo con menor longitud en el seed generan un porcentaje de mapeo más alto pero al mismo tiempo implican una menor precisión en el alineamiento.
b) Permitir un mayor número de mismatches en el esquema de mapeo implica un mayor porcentaje de lecturas mapeadas al tiempo que incrementa la probabilidad de que una lectura sea posicionada incorrectamente.

Para optimizar el alineamiento en muestras de exoma se empleó el esquema mostrado en la Figura 11 en el que se usaron tres esquemas de mapeo diferentes para las lecturas de 50pb

y un esquema de mapeo para las lecturas de 25pb así como un número máximo de posiciones genómicas posibles para cada lectura de 20.

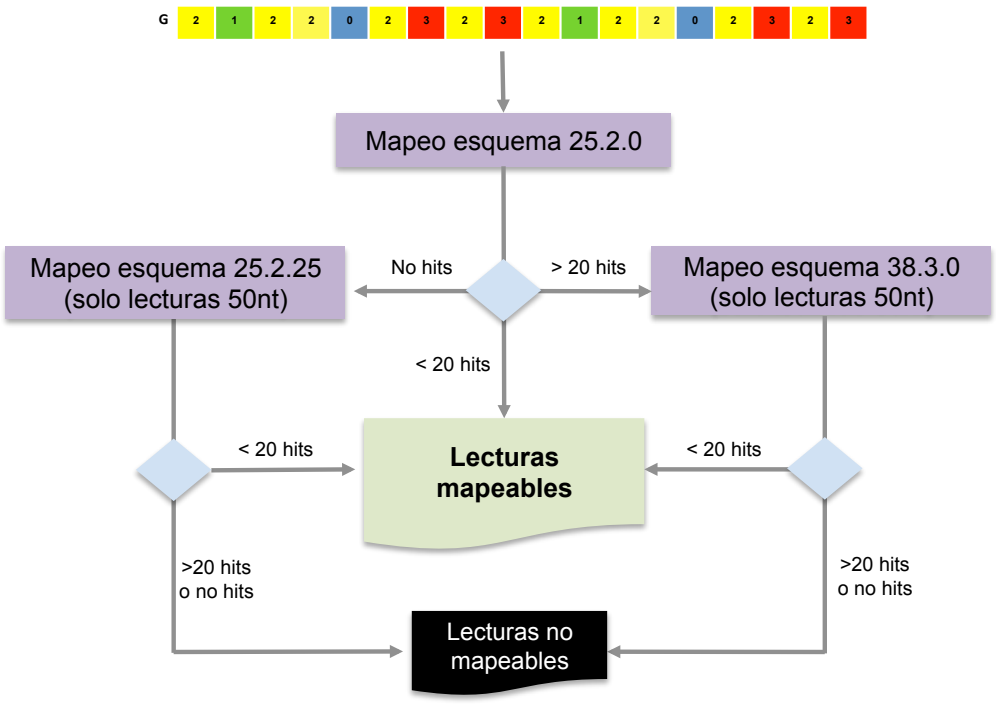


Figura 11. Uso de los distintos esquemas de mapeo en el software Bioscope.

Una de las mayores limitaciones de los algoritmos de alineamiento actuales es su capacidad de realizar alineamientos en las zonas donde existen indels ya que en lugar de permitir desapareamientos en una o varias bases contiguas, como ocurre en el caso de los errores de secuenciación o con la presencia variantes en un único nucleótido, deben permitir el salto de varias bases para poder alinear la lectura generándose así alineamientos con huecos (‘gapped alignment’).

La mayoría de las lecturas que contienen indels mapean frente al genoma con un gran número de mismatches en lugar de hacerlo mediante la generación de un hueco en el alineamiento de manera que en una primera ronda de alineamiento, no son posicionadas. La detección de indels es aun más compleja en zonas de homopolímero donde es complicado posicionar las lecturas de forma única. Para llevar a cabo el realineamiento de las lecturas en las zonas alrededor de los indels se utilizó el software ‘small-indel-tool’, que forma parte de la suite de herramientas de Bioscope, con los parámetros por defecto. Este software busca aquellas lecturas mapeadas parcialmente o aquellas lecturas cuya pareja no haya

sido mapeada e intenta realinearlas de nuevo permitiendo alineamientos con 'gaps' teniendo en cuenta el tamaño máximo y mínimo de inserto establecidos durante la generación de las librerías (en el caso de las lecturas paired-end) y un número de mismatches establecido para la ejecución del alineamiento con 'gaps'.

1.2.3 Control de calidad de la eficiencia del kit de captura y filtrado de lecturas útiles

El complejo protocolo de preparación de muestras y su posterior secuenciación por NGS, implica el uso de determinadas técnicas moleculares que van a introducir desviaciones en el dato final generado y que tendrán, dependiendo de su relevancia, un mayor o menor efecto en la identificación de variantes. Por tanto, es imprescindible en este punto analizar la robustez del sistema de captura evaluando una serie de parámetros fundamentales a partir de los datos del alineamiento:

- Sensibilidad del kit de captura: porcentaje de bases no cubiertas por el kit de captura a diferentes profundidades de lectura. Debido a la gran variabilidad en los porcentajes de cobertura que presentan los sistemas de captura mediante hibridación por sondas, es importante calcular el porcentaje de bases no cubiertas a diferentes valores con el fin de determinar si el número de lecturas mapeables y en rango son suficientes para llevar a cabo un análisis de variantes robusto.

- Especificidad del kit de captura: se consideran lecturas o bases 'on-target' aquellas dentro de las zonas diana. Para el cálculo de la especificidad del kit de captura se divide el número de lecturas 'on-target' respecto al total de lecturas o bases mapeables a lo largo del genoma (Figura 12). Una baja especificidad del kit podría indicar un problema de hibridación inespecífica, ya sea por problemas en el mismo kit de captura o por algún problema derivado de la muestra en sí, o una selección incorrecta del tamaño de los fragmentos de DNA siendo por tanto necesario su cálculo para descartar la presencia de desviaciones.

- Porcentaje de lecturas duplicadas. La probabilidad de que la fragmentación del DNA por sonicación rompa el DNA más de dos veces por el mismo sitio es muy baja de manera que si dos lecturas empiezan en la misma coordenada genómica, es muy probable que procedan del mismo fragmento de DNA. Asimismo, debido a la naturaleza del proceso de amplificación por PCR de las moléculas de DNA, es común la introducción de errores en este paso. El objetivo de este paso es

determinar el grado de eficiencia de la amplificación a la vez que se eliminan estas lecturas del fichero del alineamiento y reducir, por tanto, el número de variantes que pudieran ser identificadas de forma errónea en pasos posteriores. La secuenciación de tipo 'paired-end' es más mucho más eficiente que la secuenciación 'single-end' a la hora de eliminar las lecturas procedentes de duplicados de PCR ya que es necesario que ambos componentes de la pareja de lecturas coincidan en las mismas coordenadas genómicas que otra pareja de lecturas para ser consideradas como duplicados. El porcentaje de duplicados de PCR varía en gran medida en función de la cantidad inicial de DNA y el rendimiento proporcionado por el kit de captura pudiendo llegar a alcanzar un porcentaje muy elevando del total de las lecturas generadas, suceso que reduce en gran medida la profundidad de lectura media por base, parámetro directamente ligado con la tasa de detección de variantes.

- Tamaño medio del inserto: dado que la longitud de las sondas del kit SureSelect son de 120nt, si el tamaño de los fragmentos seleccionados es mucho mayor que el establecido durante la preparación de las librerías (150-180nt + los adaptadores específicos de la plataforma) el número de lecturas 'on target' será menor de manera que será necesario el incremento del número de lecturas iniciales por muestra para obtener una cobertura óptima en las zonas diana. El cálculo de este parámetro es por tanto muy importante para poder determinar la causa de una posible disminución en la eficiencia del kit de captura.

Para llevar a cabo el cálculo de la sensibilidad y especificidad del kit de captura así como la eliminación de lecturas con un bajo valor de mapeo se utilizó el software de Samtools v0.1.8 [86]. Para la eliminación de duplicados de PCR se utilizó el software de Picard Tools v1.31 [87]. Todos estos pasos se programaron de forma secuencial en un script en lenguaje Bash.

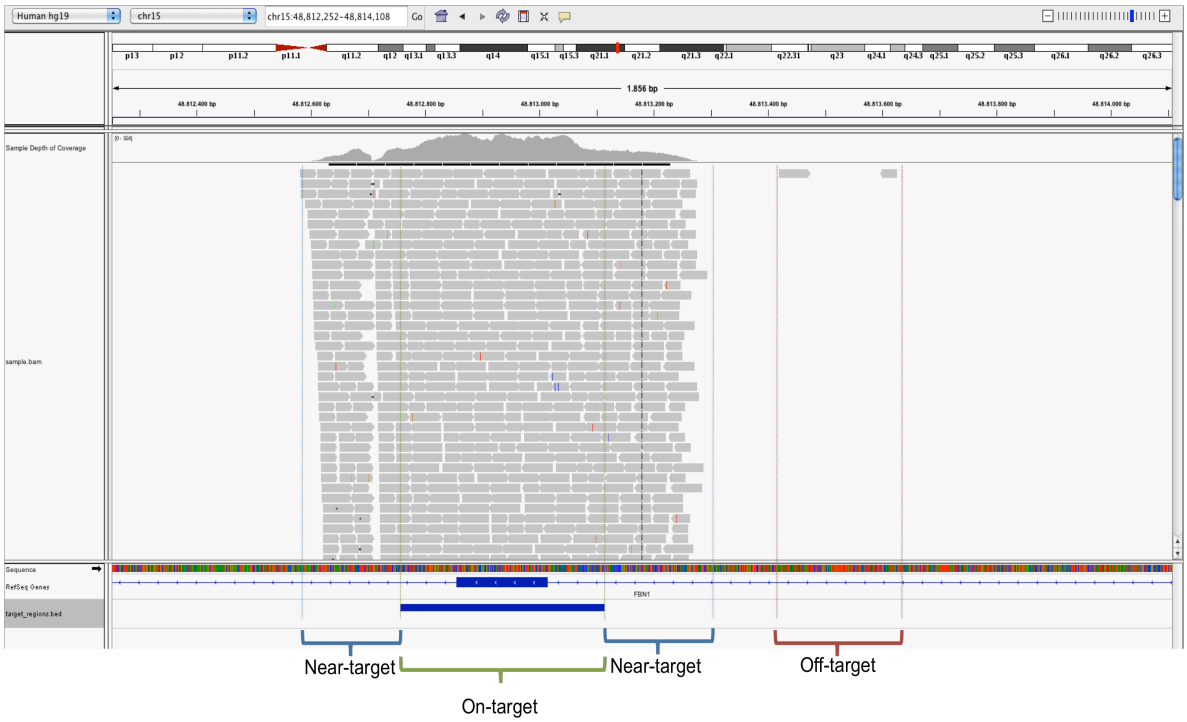


Figura 12. Lecturas “on target”, “near target” y “off-target”. La imagen corresponde al alineamiento de una muestra (en el gráfico marcada como ‘sample.bam’) a lo largo del exon 10 de la isoforma canónica del gen FBN1 visualizado a través del software Integrative Genomics Viewer [88]. El fichero mostrado en la imagen como ‘target_regiones.bed’ pertenece a las coordenadas de captura de las sondas diseñadas para el estudio del exón, cuyas coordenadas genómicas se muestran en el visualizador atendiendo a la información disponible en RefSeq (track correspondiente a ‘RefSeq genes’). Las lecturas localizadas dentro de las coordenadas de captura de las sondas se denominan ‘on-target’, aquellas que se encuentran a ambos lados de las zonas diana se denominan ‘near-target’ y las lecturas que se posicionan en zonas alejadas de las regiones diana se denominan ‘off-target’.

1.2.4 Identificación de variantes puntuales y de pequeñas inserciones y deleciones

La identificación de variantes consiste en la detección de zonas discordantes entre las lecturas alineadas y la secuencia de referencia a partir de las cuales se define un genotipo en función del número de lecturas que señalen uno u otro alelo. La credibilidad de una variante será mayor o menor en función de determinados parámetros siendo los principales la profundidad de lectura y la calidad de las bases que cubren esa posición. La llamada de variantes es el paso más complejo y sensible del pipeline global donde las variantes reales

se entremezclan con una mayor o menor tasa de error dependiendo de la muestra, la carrera (‘run’) y la tecnología de secuenciación.

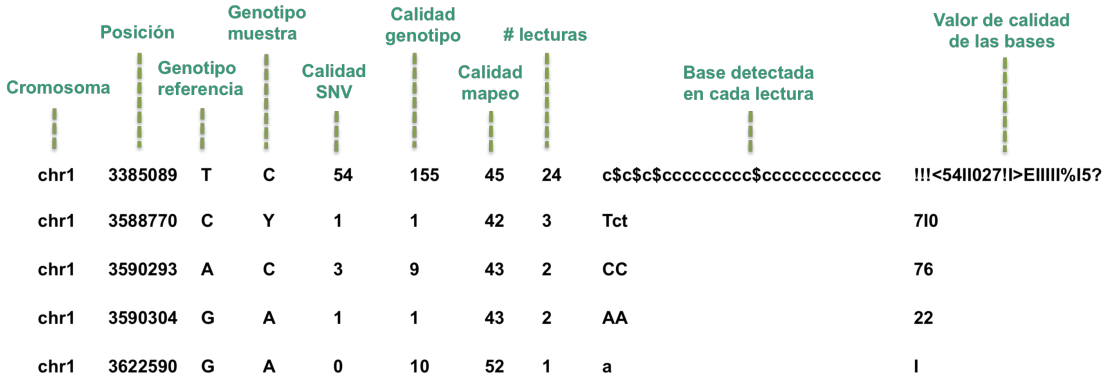
Una de la herramientas más completas para la identificación de variantes es Samtools v0.1.8. Este paquete de análisis no solo permite la identificación de variantes sino que también incluyen otra serie de comandos para la manipulación de los ficheros SAM tales como ordenar y fusionar ficheros de alineamiento, transformar los ficheros de SAM a BAM o viceversa, generar información resumida por posición sobre el alineamiento para la muestra en estudio (fichero en formato ‘PILEUP’) o visualizar el alineamiento en formato texto.

Para realizar la llamada de variantes, Samtools genera un fichero en formato ‘PILEUP’ a partir del fichero del alineamiento donde cada línea del fichero representa una posición genómica distinta en la cual se recogen una serie de parámetros que resumen la información proporcionada por todas las lecturas que cubren esa posición y que posibilitan de esta forma filtrar el listado resultante de variantes descartando aquellas cuya probabilidad de error es mayor (Figura 13).

Tal como se refleja en el apartado 1.2.2 de esta sección, para llevar a cabo la detección de indels con Bioscope, es imprescindible realizar una segunda ronda de alineamiento que permita huecos. Una vez realineadas las lecturas, el software ‘Small_indel_tool’ reporta todas aquellas posiciones donde existen inserciones o deleciones incorporando información adicional sobre la fiabilidad de estos indels que posteriormente es empleada para estimar la probabilidad de que ese indel y su genotipo asociado sean reales. La Figura 14 muestra un ejemplo de la identificación de una deleción y una inserción con este software.

Los programas mencionados en esta sección fueron incorporados a un script en Bash junto con una serie de scripts propios en Perl que permitieron la ejecución de este paso del pipeline de manera automática.

Figura 13. Formato "PILEUP" generados por Samtools v1.12 para la identificación de variantes.



chr1	3385089	T	C	54	155	45	24	ccc$ccccccccc$ccccccccccc	!!!<54II027!I>EIIIII%I5?
chr1	3588770	C	Y	1	1	42	3	Tct	7I0
chr1	3590293	A	C	3	9	43	2	CC	76
chr1	3590304	G	A	1	1	43	2	AA	22
chr1	3622590	G	A	0	10	52	1	a	I

Columna 1, nombre del cromosoma.
Columna 2, posición genómica de la variante.
Columna 3, genotipo de la referencia en esa posición.
Columna 4, genotipo identificado en la muestra, si la muestra es heterocigota para esa posición se emplea la nomenclatura internacional según la IUPAC (International Union of Pure and Applied Chemistry)
Columna 5, calidad del genotipo (GQ). Probabilidad de que el genotipo identificado sea erróneo. El valor se muestra en escala Phred.
Columna 6, valor de calidad de la variante (SNVQ). Este valor indica la probabilidad de que esa variante sea errónea. El valor se reporta en escala Phred.
Columna 7, valor de calidad máximo de mapeo (MAPQ) para esa posición.
Columna 8, número total de lecturas que leen esa posición (DP).
Columna 9, bases detectadas en cada una de las lecturas que cubren esa posición. Un punto o una coma simbolizan que la base identificada en una lectura en esa posición es igual a la referencia y que la lectura mapea en la cadena 'forward' o 'reverse' respectivamente. La aparición de alguno de los 4 nucleótidos en mayúsculas (A,G,T,C) simboliza un 'mismatch' en la cadena 'forward' mientras que si la lectura se encuentra en la cadena 'reverse' la codificación se realiza con minúsculas (a,g,t,c). Un patrón `\+[0-9]+[ACGTNacgtn]+' indica que existe una inserción entre esa posición de referencia y la siguiente. Un patrón `-[0-9]+[ACGTNacgtn]+' indica que esa base se ha delecionado en la muestra. El símbolo `^' marca el inicio de un segmento de secuencia y el símbolo '$' marca el final de manera que es posible reconstruir la secuencia de las lecturas a partir de este fichero.
Columna 10, valor de calidad en escala Phred asociado a las bases en esa posición en caracteres ASCII.

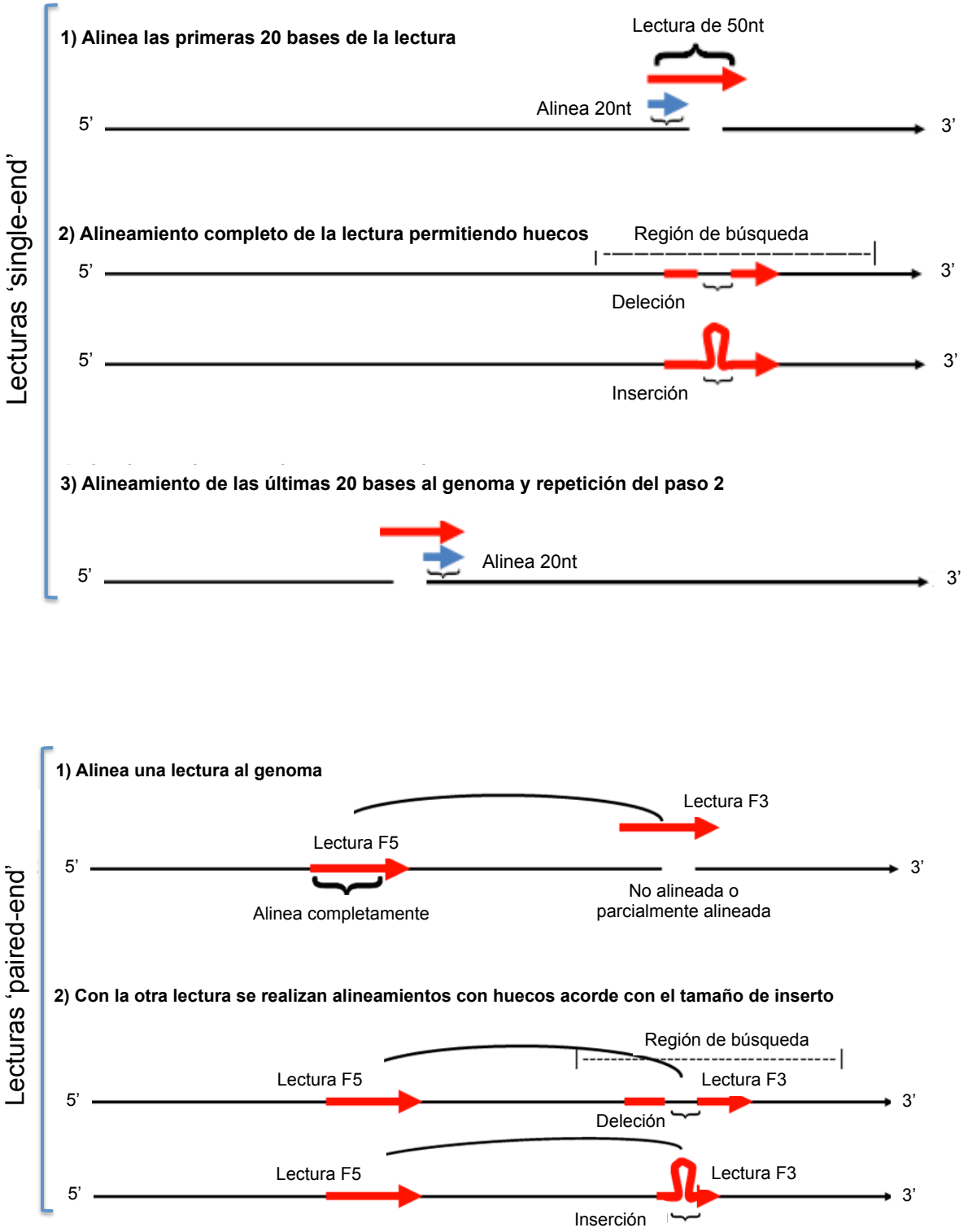


Figura 14. Identificación de inserciones y deleciones con la herramienta 'Small indel tool' de Bioscope en datasets con secuenciación tipo 'single-end' o 'paired-end'.

1.2.5 Anotación de variantes y priorización de genes candidatos

La finalidad en los estudios de resecuenciación dirigida es la determinación de la relación entre un genotipo y un fenotipo determinado. Para ello es necesario formatear y depurar el listado de variantes e integrar distintas fuentes de información biológica con la información técnica obtenida hasta este punto del análisis de manera que pueda facilitarse la interpretación de los resultados obtenidos.

Una de las mayores bases de datos de información biológica que existen en la actualidad es Ensembl [89, 90]. Ensembl alberga, entre otros muchos recursos, información sobre las principales bases de datos de variantes como dbSNP, HapMap, 1000Genomes o HGMD [91-95]. Durante los últimos años, Ensembl ha creado una serie de interfaces de programación o APIs (Application Programming Interfaces) [96, 97], a partir de los cuales es posible generar scripts a medida con los cuales extraer información biológica adicional de esta base de datos que permita estimar el daño potencial de cada una de las variantes identificadas. Tomando como base el script Variant Effect Predictor v.1.0 o VEP y las APIs de Ensembl para la versión 59 de la base de datos [98], se generó un script en Perl a partir del cual se obtuvo la información biológica disponible en Ensembl para cada variante identificada. Los resultados de la información descargada de Ensembl, los resultados obtenidos con SIFT [99] a través de la interfaz gráfica y la información técnica de la variante (posición genómica, valor de calidad, valor del genotipo, profundidad de lectura, etc) fueron integrados en un único fichero de resultados a través de scripts propios escritos en Perl y Bash. Los ficheros resultantes constaron de las siguientes columnas:

Chr: nombre del cromosoma

Start: posición genómica donde se inicia la variante.

End: posición genómica donde termina la variante.

Reference: alelo que aparece en la referencia.

Sample: alelos presentes en la muestra.

Type: tipo de variante (SNV, inserción o deleción).

Gene_ID: código de identificación del gen en Ensembl.

Transcript_ID: código de identificación del transcrito en Ensembl.

Ensembl_prediction: consecuencia la variante a nivel transcripcional.

Amino_acid_change: cambio aminoacídico producido y número de aminoácido (cDNA) en el que se produce.

Corresponding_variation: número de acceso de la variante en caso de que sea conocida.

SIFT_prediction: predicción del efecto de la variante a nivel proteico por el algoritmo SIFT.

SIFT_score: valor de predicción del software SIFT. Un valor por debajo de 0,05 se considera "damaging", es decir, el cambio produce un efecto deletéreo en la proteína.
Gene_name: nombre del gen al que afecta la variante.
Description: descripción del gen al que afecta la variante.
Genotypes: genotipos posibles identificados por las lecturas que cubren esas bases separados por '/' (solamente aplicable a indels).
Read_distribution: número de lecturas que marcan cada genotipo (solamente aplicable a indels).
Non-redundant_reads: lecturas no redundantes que cubren esa posición.
QV_SNV: valor de calidad de la variante.
SNV_Consensus_quality: valor de calidad del genotipo.
GERP_conservation_score: el valor "Genomic Evolutionary Rate Profiling" o GERP proporciona una estima del grado de conservación evolutivo de una base. El valor oscila entre -12,3 y 6,17 siendo 6,17 el valor más conservado [100].
OMIM_disease: enfermedad relacionada con el gen según OMIM.

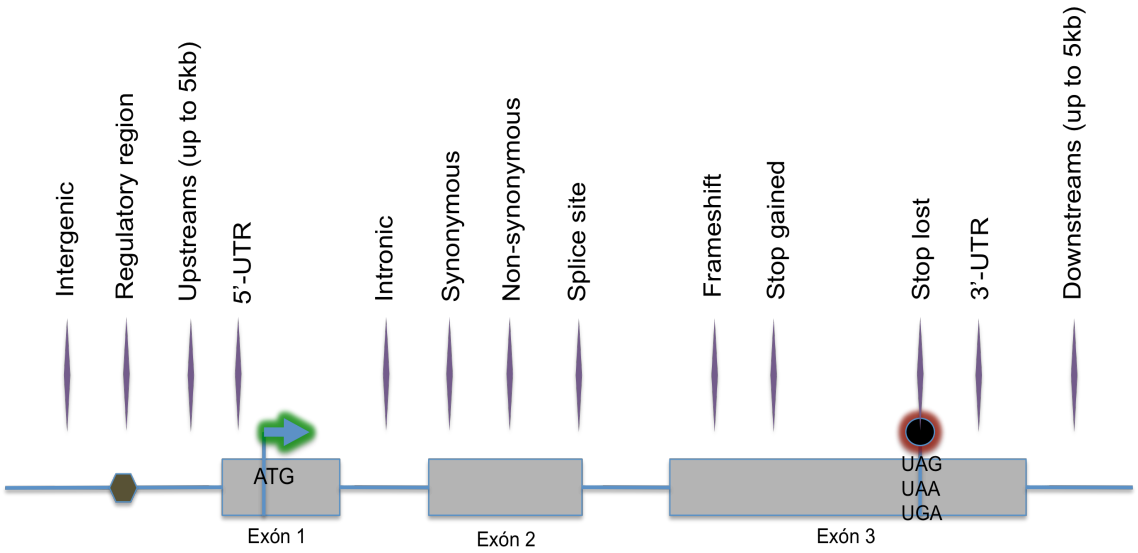
Para la identificación de nuevos genes candidatos, se llevaron a cabo los siguientes pasos:

a. Filtrado de variantes conocidas con la ayuda de la información depositada en grandes bases de datos tales como dbSNP, 1000Genomes y la base de datos de HapMap entre otras.
b. Clasificación de las variantes según su efecto a nivel transcripcional o proteico, o atendiendo a su localización genómica.
c. Determinación del grado de conservación evolutivo de la/las posiciones genómicas afectadas por el cambio.

Las variantes fueron clasificadas en función de su efecto a nivel transcripcional en las categorías descritas en la Figura 15. De manera paralela, las variantes fueron procesadas por el algoritmo SIFT con la finalidad de añadir información sobre su efecto a nivel proteico.
Para llevar a cabo la priorización de genes candidatos se tomó el conjunto de genes con variantes cuyo efecto a nivel transcripcional hubiera resultado potencialmente dañino (variantes clasificadas como "Non-synonymous", "Splice site", "Frameshift", "Stop gained" o "Stop lost"). En función del número de variantes por gen se generaron dos listas de genes candidatos, una para un modelo de herencia recesivo, siendo por tanto necesaria la presencia de al menos dos variantes en el mismo gen y en la misma posición (variante homocigota) o en posiciones distintas (asumiendo que los dos alelos estuvieran mutados y

el individuo fuera un heterocigoto compuesto), y otro listado para aquellos genes que cumplieran con los criterios genéticos de un modelo dominante, es decir, que existiera al menos una variante potencialmente patogénica en el gen.

Figura 15. Clasificación de las variantes dependiendo de su efecto a nivel transcripcional o de su localización genómica.



Intergenic – Variantes a más de 5Kb del extremo 5' o 3' de un transcrito
Regulatory region – Variantes en regiones reguladoras.
Upstream – Variantes dentro de las 5kb anteriores al extremo 5' del transcrito
5'-UTR – El cambio identificado se localiza en la región 5'-UTR.
Intronic – El cambio identificado se localiza en intrones.
Synonymous – Variantes localizadas en la zona codificante de un gen y cuya presencia no genera cambios en el aminoácido del que forman parte.
Non-synonymous – SNPs localizados en la secuencia codificante que provocan un cambio en el aminoácido que codifica la secuencia proteica.
Splice site – El cambio identificado se localiza a 1-3pb dentro del exón o a 3-8pb dentro de un intrón.
Frameshift – Indel que provoca un cambio en la pauta de lectura.
Stop gained – SNVs que provocan la aparición de un codón de parada.
Stop lost – SNVs que provocan la pérdida de un codón de parada.
3'-UTR – El cambio identificado se localiza en la región 3'-UTR.
Downstream – Variantes en las 5kb anteriores al extremo 3' del transcrito.
NMD transcript – La variación afecta a un transcrito que es degradado mediante el mecanismo celular nonsense-mediated decay (NMD).
Within non-coding gene – SNVs o indels localizados en un gen no codificante

1.2.6 Sensibilidad y Especificidad en la llamada de variantes

Las variantes identificadas en la muestra de exoma de la línea HapMap por NGS fueron contrastados con los datos de genotipado masivo mediante arrays del consorcio de HapMap incluidos en la Fase 3, correspondientes a 162 individuos, para la línea de HapMap NA12144 y la base de datos Ensembl v59. Los datos de genotipado fueron obtenidos a través del sitio FTP del consorcio de HapMap [101].

Para el cálculo de los parámetros necesarios, las variantes se clasificaron en:

- VPD=Verdaderos positivos cuyo genotipo identificado en la muestra resultó discordante respecto a los datos de HapMap para esa línea celular.
- VPC=Verdaderos positivos cuyo genotipo calculado fue el mismo presentado por el consorcio de HapMap para esa línea celular.
- FN=Falsos negativos, variantes no identificadas en la muestra y presentes en los datos de arrays de HapMap.
- VN=Verdaderos negativos, posiciones identificadas como homocigotas para la referencia tanto en la muestra como en los datos de arrays de HapMap.
- FPNV=Falso positivo y/o variantes noveles. Este número corresponde a las variantes que no obtienen un código de identificación tras anotarlas frente a Ensembl. *Téngase en cuenta que dada esta premisa, el número de falsos positivos puede estar ligeramente sobrestimado* ya que en esta categoría se incluyen también variantes noveles. No obstante, y dado que se trata de una línea celular estudiada por diferentes consorcios y cuya información ha sido depositadas en grandes bases de datos como dbSNP, el número de variantes noveles debe ser muy bajo alrededor de un 5% [102]
- VPANOT=variantes detectadas en la muestra que obtienen un código de identificación tras consultar Ensembl.

Se consideraron verdaderos positivos (VP): VPC+VPD+VPANOT
Se consideraron verdaderos negativos (VN): VN
Se consideraron falsos positivos (FP): FPNV
Se consideraron falsos negativos (FN): FN

La sensibilidad o tasa de verdaderos positivos se calculó en base a la fórmula:

$$\text{Sensibilidad} = VP/(VP+FN)$$

La especificidad se calculó en base a la fórmula:

Especificidad=VN/(VN+FP)

La tasa de falsos positivos se calculó siguiendo esta fórmula:

Tasa de falsos positivos = 1-Especificidad

Para facilitar la obtención de los datos de sensibilidad y especificidad en líneas HapMap se generó una herramienta en Bash y Perl que se describe a continuación.

$ sg_concordance_studies.sh -v <fichero.vcf> -s <lista_de_muestras> -d <panel> -h <HapMap>. El script es capaz de tomar el fichero en formato VCF con las variantes, seleccionar la muestra que ha de ser comparada frente a los datos de HapMap, obtener los datos de HapMap para la región concreta en estudio y comparar finalmente la muestra secuenciada por NGS con los datos de HapMap.

-v -- Variantes identificadas para las muestras en estudio en formato VCF.

-s -- Lista de muestras incluidas en el fichero VCF anterior que deben ser comparadas frente a los datos de HapMap.

-d -- Nombre de la aplicación para obtener las coordenadas de captura correctas.

-h -- Nombre de la línea celular HapMap.

Output – El resultado final es un fichero que contiene el número de variantes clasificadas por categoría a partir del cual se obtienen los valores de sensibilidad y especificidad clínicas.

1.2.7 Esquema general del pipeline de análisis desarrollado

El diagrama presentado en la Figura 16 muestra el esquema global del pipeline de análisis generado para el estudio de exomas completos. El esquema resalta los ficheros de entrada (INPUT) y de salida (OUTPUT) en cada uno de los pasos del pipeline (marcados con un círculo). A continuación, se detallan los programas 'open-source' y parámetros ejecutados así como la funcionalidad de los scripts propios desarrollados para cada uno de los pasos marcados en el diagrama:

Paso 1.- Control de calidad de los datos de secuenciación brutos. Revisión manual de los gráficos generados automáticamente por SETs.

Paso 2.- Alineamiento con Bioscope. De los 2 ficheros de configuración iniciales que toma Bioscope para el mapeo se modificaron los siguientes parámetros:

Lecturas F3: mapping.scheme.unmapped.50=25.2.0,25.2.15;
mapping.scheme.repetitive.50=38.3.0,25.2.0;
small.indel.frag.dependency=1;
ma.to.bam.output.filter=primary.

Lecturas F5: mapping.scheme.unmapped.25=25.2.0

El lanzamiento de la aplicación se realizó de forma manual mediante la interfaz gráfica proporcionada por Bioscope.

Paso 3.- pipe_exoma.sh: script en Bash que contiene los pasos que forman parte de la evaluación de la eficiencia del kit de captura junto con la plataforma de secuenciación, la llamada de variantes y parte de la anotación de las mismas. (pasos del 4 al 8). El fichero de entrada del script es el fichero BAM resultante del alineamiento con Bioscope. El fichero de salida es la información biológica obtenida de Ensembl junto con la información técnica generada en el proceso de llamada de variantes (profundidad de lectura, valor de calidad de la variante y valor de calidad del genotipo).

$ sg_pipe_exoma.sh

-Input 1 -- Fichero BAM resultante del alineamiento con Bioscope

Paso 4.- Control de calidad de la eficiencia del kit de captura.

4a) Eliminación y cálculo de lecturas duplicadas con MarkDuplicates de PicardTools.

4b) Cálculo del tamaño del inserto con CollectInsertSizeMetrics de PicardTools.

4c) Obtención de las bases cubiertas dentro de las zonas diana.

$ sg_unificar_intervalos_de_captura_solapantes.pl -- Genera intervalos de mayor tamaño a partir de las coordenadas de las sondas de captura en caso de que solapen.

-Input -- El fichero de entrada es un fichero en formato BED con las coordenadas de las sondas diseñadas. Cada línea pertenece a una sonda

Output -- El resultado del script es un fichero en formato BED con las coordenadas de las zonas capturadas (sondas solapadas)

$ samtools pileup –c -f referencia_hg19.fa muestra_sin_dup.bam > muestra_sin_dup.pileup - Genera el fichero con formato PILEUP para todas las posiciones cubiertas en la muestra a partir del fichero BAM en el que se han eliminado las lecturas duplicadas y la secuencia de referencia hg19.

$ sg_extraer_SNPs_fichero_tabulado.pl -- Obtiene las posiciones dentro de las zonas diana para cada cromosoma.

-Input1 -- Coordenadas de las zonas capturadas separadas por cromosoma

-Input2 -- Fichero en formato PILEUP separado por cromosoma

Output -- La salida del script es un fichero en formato PILEUP con las posiciones cubiertas dentro de las zonas diana

4d) Cálculo de la especificidad del kit de captura. La especificidad se calculó dividiendo la suma del número total de bases localizadas en las zonas diana respecto al total de bases mapeables.

4e) Cálculo de la sensibilidad del kit de captura. La sensibilidad se calculó dividiendo el número de bases cubiertas a una determinada profundidad de lectura frente al número total de bases diana.

4f) Cálculo de las estadísticas de mapeo con flagstat de Samtools.

Paso 5.- Llamada de SNVs.

$ samtools.pl varFilter -D 2000 muestra_sin_dup.pileup | awk '$6>=20' | awk '{if ($4!="*") print $0}' > snps_sin_duplicados_Q20 – Obtiene las variantes del fichero en formato PILEUP de Samtools con un valor de calidad del SNVs de al menos 20 y elimina las posiciones con indels.

$ sg_convertir_SNPs_de_heterocigotos_a_homocigotos.pl -- Samtools reporta las variantes heterocigotas con las letras R,Y,S,W,K,M según el código IUPAC, Este script modifica el genotipo del alelo alternativo cambiándolo por A,C,G o T según corresponda.

-Input -- Fichero con los SNVs obtenidos en el paso anterior

Output -- Fichero en formato PILEUP con los alelos alternativo heterocigotos transformados

Paso 6.- Llamada de indels con 'Small Indel Tool' de Bioscope con los parámetros por defecto a partir del BAM sin modificar. El lanzamiento de la herramienta se realizó a través de la interfaz gráfica proporcionada por Bioscope. La herramienta es capaz de generar el listado de indels teniendo en cuenta solamente las lecturas no redundantes y rescatando aquellas no mapeadas o con un valor de calidad de mapeo bajo.

Paso 7.- Anotación de SNVs e indels.

7a) Se aplicó una versión modificada del script VEP de Ensembl empleando como input el listado de variantes con el siguiente formato:

chromosoma<TAB>posición_inicio<TAB>posición_final<TAB>alelo_referencia/alelo_alternativo<TAB>cadena

7b) El efecto a nivel proteico de los SNVs no conocidos fue calculado con SIFT en paralelo a través de la interfaz de usuario proporcionada por este paquete y posteriormente añadido de forma automatizada al fichero final de anotación.

7c) La información técnica (profundidad de lectura, valor de calidad de la variante y valor de calidad del genotipo) fue combinada con la información biológica obtenida tras consultar Ensembl.

Paso 8.- Priorización de genes candidatos. Las variantes anotadas clasificadas como no conocidas y potencialmente patogénicas sirvieron de input para la determinación de los genes candidatos.

$ sg_candidate_genes.pl – Obtiene el listado de genes candidatos así como las variantes potencialmente patogénicas asociadas a estos.

-Input -- Fichero con las variantes potencialmente patogénicas en el formato obtenido tras la anotación (ver Tabla 7 en el apartado 1.2 de la sección Resultados).

Output1 -- El primero de los ficheros de salida es un listado de genes candidatos. Este fichero contiene tres columnas: código de identificación de Ensembl para el gen, el número de variantes por gen y la descripción del gen.

Output2 -- El segundo de los ficheros de salida es un listado de variantes potencialmente patogénicas incluidas en los genes candidatos y su anotación, tal como se reporta la información para todas las variantes una vez anotada.

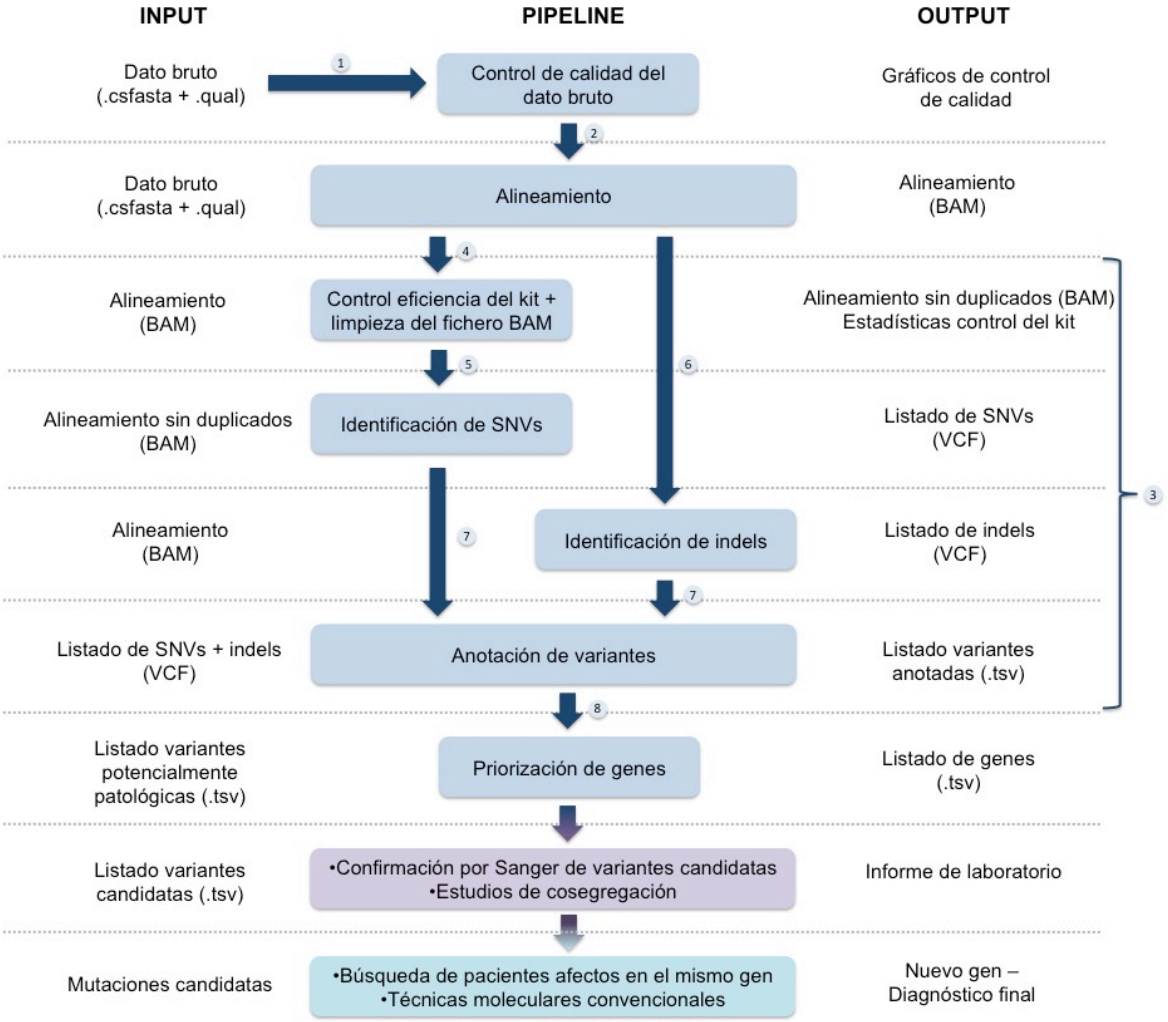


Figura 16. Definición del pipeline de análisis para la identificación de nuevos genes candidatos a partir de datos de exoma.

2 Desarrollo y análisis de paneles de genes orientado al diagnóstico genético en enfermedades cardiovasculares

Las siguientes secciones describen el estudio de validación llevado a cabo para el desarrollo de un sistema de análisis integral orientado al diagnóstico de enfermedades cardiovasculares con riesgo de muerte súbita a través de la secuenciación masiva. Inicialmente, el estudio se centra en el desarrollo de un sistema de captura de genes 'ad hoc' detallando la estrategia utilizada para su generación. Posteriormente, el estudio se focaliza en la optimización y protocolización clínica del pipeline implementado inicialmente en el apartado 1 de esta misma sección y su posterior validación mediante el uso de muestras control.

2.1 Generación de las sondas del sistema de captura

La tecnología de captura y enriquecimiento de zonas específicas del DNA SureSelect permite la generación de kits diseñados 'ad hoc' de tamaño muy variable, desde unas pocas kilobases hasta 3Mb. Para llevar a cabo el diseño del kit de captura de los genes se realizó una exhaustiva revisión bibliográfica por el grupo de expertos de la Unidad de Genética Médica de Sistemas Genómicos tras la cual se seleccionaron 72 genes implicados en distintas enfermedades cardiacas de origen genético mediante la consulta de diferentes recursos como OMIM [103], Pubmed [104], Gene [105], HGMD [93, 94, 106], Ensembl [89, 90, 107], GeneCards [108], Uniprot [109] y Genatlas Universite Paris Descartes [110]. La Tabla 2 resume los genes incluidos en el kit de captura y su relación con las distintas patologías cardiacas. Para llevar a cabo el diseño de las sondas en los genes seleccionados, se tomó la isoforma principal para cada gen según la base de datos HGMD professional release 2011.2. Las coordenadas genómicas de la isoforma se obtuvieron a través de la aplicación de Biomart [111] y la versión 59 de la base de datos de Ensembl. El diseño de las sondas se realizó a través del portal eArray [112] ampliando 100pb hacia dentro del intrón con el objetivo de asegurar la captura de las zonas de splicing. El diseño también incluyó las zonas 5' y 3' UTR para todas las isoformas. En el caso de aquellas isoformas en las que las zonas UTR no estaban descritas, se seleccionaron 200pb en la región 5'-UTR y 1000pb en la región 3'-UTR. Siguiendo las recomendaciones del fabricante, el diseño del sistema de enriquecimiento se realizó mediante sondas solapantes asegurando una cobertura por base de al menos 2x, es decir, cada base capturada por el kit personalizado se cubrió al menos por dos sondas diferentes. Para maximizar la sensibilidad

de la técnica se evitó el diseño de sondas en zonas repetitivas o de baja complejidad del genoma.

Tabla 2. Relación de genes y enfermedades incluidas en el kit de captura. El panel de muerte súbita se subdividió en 13 subpaneles para facilitar la revisión de los resultados dependiendo de la patología en estudio y la historia clínicas del paciente así como del árbol familiar en caso de estar disponible.

Patología	Número de genes	Nombre de los genes (HGNC) [127]
Displasia Arritmogénica del Ventrículo Derecho	8	DSP, DSG2, DSC2, JUP, PKP2, RYR2, TGFB3, TMEM43
Miocardiopatía Hipertrófica	24	ACTC1, ACTN2, CAV3, CSRP3, JPH2, LAMP2, LDB3, MYBPC3, MYH6, MYH7, MYL2, MYL3, MYLK2, MYOZ2, PLN, PRKAG2, SLC25A4, TCAP, TNNC1, TNNI3, TNNT2, TPM1, TTN, VCL
Miocardiopatía Dilatada	34	ABCC9, ACTC1, ACTN2, CALR3, CSRP3, DES, DSG2, DTNA, EYA4, FKTN, JPH2, LDB3, LMNA, MYBPC3, MYH6, MYH7, MYL2, MIOZ2, NEXN, PLN, PSEN1, PSEN2, RBM20, SCN5A, SGCD, TAZ, TCAP, TPMO, TNNC1, TNNI3, TNNT2, TPM1, TTN, VCL
Miocardiopatía no compactada del ventrículo izquierdo	6	DTNA, LDB3, ACTC1, MYH7, TNNT2, TAZ
Miocardiopatía Restrictiva Familiar	40	ABCC9, ACTC1, ACTN2, CALR3, CAV3, CSRP3, DES, DSG2, DTNA, EYA4, FKTN, JPH2, LAMP2, LDB3, LMNA, MYBPC3, MYH6, MYH7, MYL2, MYL3, MYLK2, MIOZ2, NEXN, PLN, PRKAG2, PSEN1, PSEN2, RBM20, SCN5A, SGCD, SLC25A4, TAZ, TCAP, TPMO, TNNC1, TNNI3, TNNT2, TPM1, TTN, VCL
Trastornos asociados a Aneurisma de Aorta	3	FBN1, FBN2, TGFBR2
Síndrome de Brugada	7	CACNA1C, CACNB2, GPD1L, KCNE3, SCN1B, SCN3B, SCN5A
Síndrome QT-largo	12	AKAP9, ANK2, CACNA1C, CAV3, KCNE1, KCNE2, KCNH2, KCNJ2, KCNQ, SCN5A, SCN4B, SNTA1
Síndrome QT-corto	5	CACNA1C, CACNA1B, KCNH2, KCNJ2, KCNQ1
Fibrilación Auricular Familiar	5	GJAS, KCNA5, KCNE2, KCNQ1, NPPA
Taquicardia Ventricular Polimórfica Catecolaminérgica	2	CASQ2, RYR2
Arritmia familiar	29	AKAP9, ANK2, CACNA1C, CACNB2, CASQ2, CAV3, DSC2, DSG2, DSP, GPD1L, JUP, KCNA5, KCNE1, KCNE2, KCNE3, KCNH2, KCNJ2, KCNQ1, NPPA, PKP2, PLN, RYR2, SCN1B, SCN3B, SCN4B, SCN5A, SNTA1, TGFB3, TMEM43
Muerte Súbita	72	ABCC9, ACTC1, ACTN2, AKAP9, ANK2, CACNA1B, CACNA1C, CACNB2, CALR3, CASQ2, CAV3, CSRP3, DES, DSC2, DSG2, DSP, DTNA, EYA4, FBN1, FBN2, FKTN, GJA5, GPD1L, JPH2, JUP, KCNA5, KCNE1, KCNE2, KCNE3, KCNH2, KCNJ2, KCNQ1, LAMP2, LDB3, LMNA, LRP6, MEF2A, MYBPC3, MYH6, MYH7, MYL2, MYL3, MYLK2, MYOZ2, NEXN, NPPA, PKP2, PLN, PRKAG2, PSEN1, PSEN2, RBM20, RYR2, SCN1B, SCN3B, SCN4B, SCN5A, SGCD, SLC25A4, SNTA1, TAZ, TCAP, TGFB3, TGFBR2, TMEM43, TMPO, TNNC1, TNNI3, TNNT2, TPM1, TTN, VCL

2.2 Preparación de las muestras y secuenciación

La preparación de las muestras y posterior secuenciación fueron llevadas a cabo por el departamento de Nuevas Tecnologías de la empresa Sistemas Genómicos.

2.2.1 Obtención del material biológico

La validación del estudio se realizó a través del análisis de 10 muestras de sangre periférica de pacientes que previamente habían sido diagnosticadas de la enfermedad de Marfan mediante secuenciación Sanger. Las muestras fueron seleccionadas en función del gen afectado y el tipo de mutación presentada de manera que se incluyeron variantes puntuales así como deleciones e inserciones pequeñas dentro de los rangos conocidos de detección (11pb para deleciones y 3pb en inserciones). La siguiente Tabla 3 resume el listado de muestras así como el gen y la mutación inicialmente detectada en cada individuo por secuenciación tradicional.

Tabla 3. Muestras utilizadas para la validación del sistema de diagnóstico y sus respectivas mutaciones.

Muestra	Gen estudiado	Mutación detectada por Sanger
BM3062	*FBN1*	NM_000138.3:c.8333T>G
BM3339	*FBN1*	NM_000138.3:c.1148-1G>A
BM3895	*FBN1*	NM_000138.3:c.2248del
BM4237	*KCNH2*	NM_000238.3:c.2464G>A
BM5307	*TGFBR2*	NM_003242.5:c.1314T>A
BM5357	*FBN1*	NM_000138.3:c.4326dupA
BM6091	*FBN1*	NM_000138.3:c.5076_5078delAAG
BM6092	*FBN1*	NM_000138.3:c.7039_7040del
BM6492	*FBN1*	NM_000138.3:c.3539G>T
BM6919	*MYBPC3*	NM_000256.3:c.2308+1G>A

Un año después de la validación inicial de este panel, se llevó a cabo un estudio retrospectivo más amplio en el que se analizaron los resultados en una cohorte de 163 pacientes con riesgo de muerte súbita, en total 71 mujeres y 92 hombres cuya media de edad oscilaba en torno a los 41 años. Las muestras de los pacientes fueron recibidas en Sistemas Genómicos en el período comprendido entre Enero 2011 y Noviembre 2012. La Figura 17 muestra el porcentaje de individuos por patología incluidos el estudio retrospectivo.

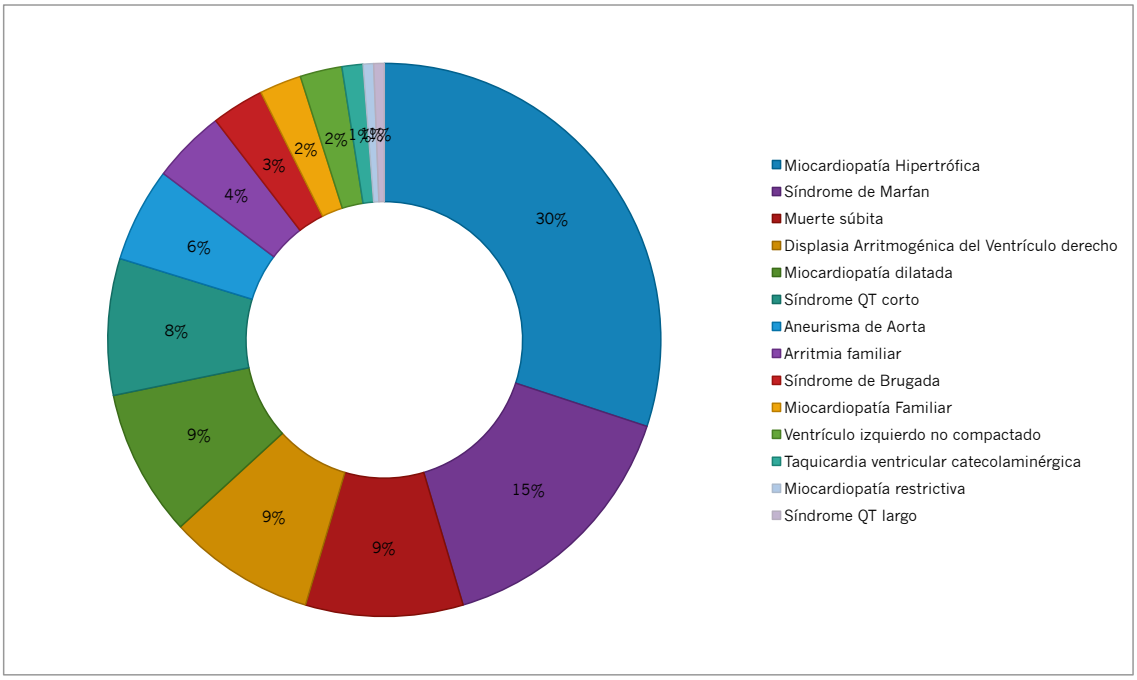


Figura 17. Distribución de los individuos incluidos en el estudio retrospectivo por patología.

2.2.2 Construcción de las librerías y captura de zonas específicas

Al igual que en las muestras de exoma, en este estudio se construyeron librerías de fragmentos (sección 1.1.2). Sin embargo, en estas librerías se incorporó a cada uno de los fragmentos de DNA de una misma muestra un "barcode" o etiqueta, una secuencia conocida de longitud fija (en este caso 5nt) única para cada muestra (Figura 18). Este tipo de librerías permite mezclar diferentes muestras y llevar a cabo la amplificación de las moléculas de DNA y su posterior secuenciación a la vez sin necesidad de separarlas físicamente optimizándose así el espacio de secuenciación y consiguiendo un ahorro considerable en los costes de reactivos.

La captura de los genes se realizó siguiendo el mismo protocolo expuesto en el apartado 1.1.3 de esta misma sección sustituyendo el kit comercial del exoma por las sondas diseñadas para el panel.

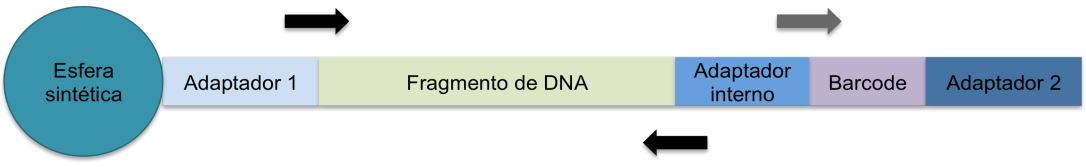


Figura 18. Estructura de una librería de fragmentos con barcodes. La secuenciación se inicia con la lectura del barcode (flecha gris), posteriormente se leen ambos extremos del fragmento del DNA. La lectura de la secuencia del barcode permite asignar cada una de las secuencias a una determinada muestra.

2.2.3 Amplificación del DNA y enriquecimiento de esferas

La amplificación del DNA y el enriquecimiento de las esferas se llevaron a cabo siguiendo el proceso descrito en el apartado 1.1.4 de esta sección tras mezclar las distintas muestras una vez etiquetadas.

2.2.4 Secuenciación

La secuenciación de las 10 muestras control se realizó en 1/4 de slide de la versión 4 de la plataforma SOLID tal como se describe en el apartado 1.1.5 de esta sección.

2.3 Análisis del panel

El protocolo de análisis de datos presentado en esta sección fue optimizado en base al protocolo generado para el análisis del exoma completo ya que el sistema de captura estaba basado en la misma tecnología y la plataforma de secuenciación era igualmente la misma, y por lo tanto, cualquier desviación o limitación grave tras el análisis se reportaría de igual manera en un caso u otro. El flujo de análisis de datos siguió el mismo esquema presentado en la Figura 16 (apartado 1.2.7 de esta misma sección). Sin embargo, para el análisis de estas muestras se introdujeron una serie de puntos de control más rigurosos así como mejoras en la llamada de variantes que permitieron reducir la tasa de falsos positivos y negativos. Adicionalmente, se añadieron nuevas fuentes de información clínica para facilitar la interpretación de las variantes.

2.3.1 Evaluación del diseño de sondas para la captura de los genes

Dada la complejidad del genoma humano, y tal como se citaba en el apartado 2.1 de esta sección, se evitó el diseño en aquellas regiones marcadas como repetitivas en el genoma según las bases de datos incluidas en eArray. Teniendo en cuenta esta premisa de partida, se generaron una serie de 'scripts' en Perl mediante los cuales fue posible el cálculo del porcentaje de zonas diana no cubiertas por las sondas del diseño. Adicionalmente, con el objetivo de indagar sobre la repercusión a nivel clínico de la imposibilidad de detectar variantes en estas zonas no cubiertas, se obtuvo un listado de variantes patogénicas o de susceptibilidad descritas en la base de datos HGMD Professional release 2011.2 incluidas en estas zonas.

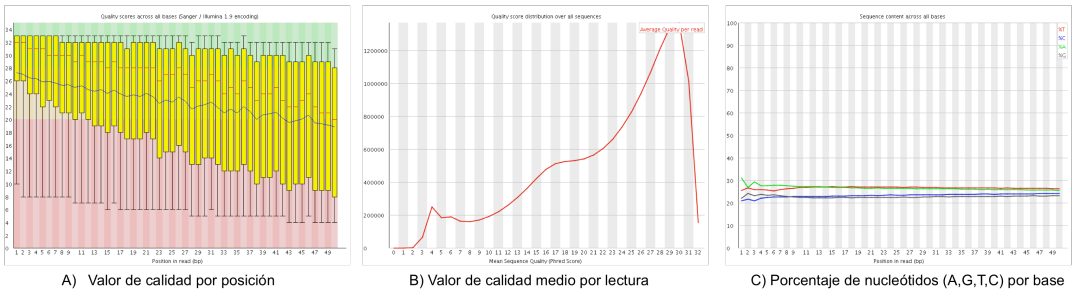
Por otro lado, se evaluó la homología de las sondas generadas con otras zonas del genoma no incluidas en el panel a fin de identificar sondas y/o regiones de conflicto donde la llamada de variantes pudiera verse comprometida. Según la empresa que comercializa los kit SureSelect, éstos tienen una especificidad media respecto al mapeo de entre el 50% y el

80%, es decir, entre el 50% y el 20% de las lecturas que mapean lo hacen fuera de los rangos del panel.

2.3.2 Control de calidad de los datos de secuenciación

El control de calidad de los datos en estas muestras se realizó con una nueva herramienta de análisis, FASTQC [113], que a diferencia del software utilizado en el estudio de exomas, permite obtener mucha más información acerca de la muestra así como la automatización de esta etapa del análisis. Los archivos .csfasta y .qual, obtenidos para cada muestra, se transformaron al formato FASTQ [114] mediante scripts propios para poder ser analizados por este software. Para cada muestra, el programa FASTQC generó una serie de gráficos (Figura 19) entre los que destacan:

1) Distribución de los valores de calidad por base. Para cada base, el programa genera un gráfico de cajas que permite visualizar la distribución de los valores de calidad en esa posición. Así, en el gráfico la barra amarilla representa los valores de los intercuartiles (25-75%), los brazos hacia arriba y hacia abajo de la caja amarilla representan el 90% y el 10% de los datos para esa posición respectivamente, la línea roja representa la mediana y la línea azul corresponde a la media.
2) Valor de calidad medio por lectura. Cuanto más alto y más estrecho sea el pico generado en este gráfico, mejor será la calidad global del dataset en estudio.
3) Contenido en colores por base. Porcentaje de cada uno de los 4 colores posibles por base siendo 0=A, 1=C, 2=G, 3=T y '.'=N. Mediante la visualización de este gráfico es posible identificar la presencia de adaptadores u otros artefactos en las lecturas.



A) Valor de calidad por posición B) Valor de calidad medio por lectura C) Porcentaje de nucleótidos (A,G,T,C) por base

Figura 19. Control de calidad de los datos brutos. Gráficas generadas por el software FASTQC.

2.3.3 Alineamiento

El alineamiento se realizó con una versión actualizada de Bioscope, versión 1.3, de la misma forma como se describe en el apartado 1.2.2 de esta misma sección.

2.3.4 Control de calidad de la eficiencia del kit de captura y filtrado de lecturas útiles

La eficiencia de la tecnología de captura se evaluó tal como se describe en el apartado 1.2.3 de esta misma sección. Adicionalmente, se calculó:

1. Porcentaje de lecturas con baja calidad de mapeo. El valor de calidad de mapeo estima la diferencia entre la primera y la segunda posición más probable de mapeo. Si el algoritmo de alineamiento no es capaz de determinar la mejor posición genómica entre un abanico reducido de posibilidades se asigna a la lectura un valor de mapeo de 0. Si la calidad del dato inicial es pobre el número de errores en las lecturas y por tanto el número de mismatches en el alineamiento será mayor y como consecuencia el valor de calidad de mapeo en un alto porcentaje de lecturas será bajo.
2. Regiones diana y variantes patogénicas o de susceptibilidad no cubiertas. Las tecnologías de captura mediante hibridación por sondas no aseguran la total cobertura de las zonas de captura de manera que, aparte de las zonas diana que no se han cubierto desde un inicio por las sondas diseñadas, se suman otras regiones donde la cobertura es nula o por debajo de los límites de detección más empleados (10x, 20x o 30x). La presencia de un mayor o menor número de zonas no cubiertas o zonas con baja profundidad de lectura pueden poner en riesgo la sensibilidad y especificidad del test diagnóstico y por lo tanto su evaluación supone un pilar básico a tener en cuenta con el uso de este tipo de metodologías. Para ello, se desarrolló una estrategia de análisis que tomara como fichero de entrada el fichero de alineamiento en formato BAM y que generara los siguientes ficheros a diferentes profundidades de lectura:
    - Bases no cubiertas en formato BED para poder ser visualizadas al mismo tiempo que el alineamiento mediante herramientas tipo Integrative Genomics Viewer o IGV.
    - Variantes patogénicas o de susceptibilidad descritas en la base de datos HGMD Professional release 2011.2 incluidas en estas zonas no cubiertas.

2.3.5 Identificación de variantes puntuales y de pequeñas inserciones y deleciones

El listado de variantes obtenido tras la utilización de diferentes 'callers' (programas para la llamada de variantes) puede resulta diferente en un porcentaje no despreciable de las variantes. Teniendo presente que la detección automatizada de variantes en NGS se realiza siempre buscando un equilibrio entre especificidad y sensibilidad se optó por la inclusión de un nuevo paquete Genome Analyzer Took Kit o GATK v.1.0.5777 [115], además de los dos softwares ya utilizados para la identificación de variantes en el exoma completo (ver Métodos 1.2.4). Al igual que Samtools, GATK está formado por un conjunto de herramientas de análisis que, aparte de realizar la llamada de variantes, permiten manipular un fichero BAM en múltiples aspectos siendo por tanto uno de los paquetes más completos y robustos de este tipo [116]. Una de las mayores diferencias entre ambos softwares a la hora de llamar variantes es que GATK es capaz de aprender de los datos que analiza para fijar diferentes límites mientras que Samtools necesita la definición de estos filtros por parte del usuario. Otra de las grandes diferencias entre ellos radica en el modelo utilizado para la detección de indels. Para la llamada de pequeñas inserciones y deleciones, Samtools utiliza la información que obtiene de un fichero en formato PILEUP. La versión GATK v.1.0.5777 por el contrario emplea un modelo basado en Dindel [117] que genera una serie de ventanas en las zonas donde existe una mayor tasa de error a partir de las cuales realinea las lecturas para redefinir los indels. Del mismo modo, las diferencias en el uso de distintos filtros internos como la profundidad mínima de lectura, el valor de calidad mínimo de las lecturas en la posición analizada, la existencia de variantes en ambas cadenas o la frecuencia mínima del alelo alternativo también provocan diferencias en los resultados.

Para realizar la llamada de variantes con GATK, se siguieron los estándares establecidos por sus desarrolladores [118], de manera que los ficheros BAM fueron realineados en zonas donde existía una alta proporción de variantes y/o una alta tasa de error para generar alineamientos con huecos redefiniendo así las zonas de indels y mejorando el valor de calidad de mapeo de estas lecturas tomando como input un fichero BAM y un dataset de variantes conocidas, en este caso dbSNP v133. Una vez realineados los ficheros, se procedió a realizar la llamada de variantes integrando tres herramientas distintas, Samtools v0.1.12 y GATK v.1.0.5777, para la llamada de SNVs y Small_indel_tool v1.2 junto a Dindel , incorporado como opción en GATK, para la llamada de indels. Para realizar la integración de estos tres paquetes se desarrolló un script en Bash que incorporaba, aparte de los software previamente mencionados, herramientas para el formateo de los ficheros intermedios como VCF Tools [119] y scripts propios escritos en Perl. El resultado final de la llamada de variantes fue un fichero en formato VCF4.0 estándar [120].

Parte de las muestras del estudio retrospectivo fue analizada con versiones posteriores de Samtools v.0.1.16 y GATK (v.1.5-28 y v.2.1-5) adoptando los mismos criterios de filtrado establecidos durante la puesta a punto del panel.

2.3.6 Anotación de variantes

La anotación de las variantes en este caso fue realizada empleando una versión más evolucionada del script "Variant Effect Predictor" de Ensembl frente a la versión 62 de esta base de datos [121]. Además de la información obtenida de Ensembl, en el análisis de estas muestras se incorporó una nueva fuente de información, la base de datos de Human Gene Mutation Database. Este recurso, tradicionalmente utilizado en el diagnóstico genético molecular, contiene información sobre la relación mutación-gen-enfermedad y por lo tanto puede facilitar e incluso generar un diagnóstico directo. La base de datos se presenta en dos versiones, una versión pública gratuita, incorporada en Ensembl, y una versión de pago más actualizada y más curada, HGMD Professional®, que fue incluida en el protocolo de anotación de las variantes mediante la generación de una serie de pequeños scripts en Perl agrupados dentro del script Bash general para la llamada y anotación de variantes. Los ficheros obtenidos tras el análisis bioinformático contenían, además de la información descrita en la sección 1.2.5, los siguientes campos:

Gene_name: Nombre HGNC del Gen.
Chr: cromosoma.
Position: posición genómica.
Reference: alelo de la referencia.
Sample: alelo presente en la muestra.
Variation_type: tipo de variante (SNV o indel).
Variation_length: longitud de la variante.
Genotype: genotipo de la muestra, heterocigoto u homocigoto.
Variant_ID: número de acceso de la variante en caso de que sea conocida.
HGMD: código de identificación de la base de datos HGMD Professional.
Ensembl_Gene_ID: código de identificación del gen en Ensembl.
Ensembl_Transcript_ID: código de identificación del transcrito en Ensembl.
Gene_description: descripción del gen al que afecta la variante.
Consequence_on_transcripts: efecto de la variante a nivel transcripcional.
aa_position: posición del aminoácido donde se produce el cambio.
aa_change: cambio aminoacídico producido
GERP_conservarion_score: valor de conservación GERP.

HGVS_ID: nomenclatura de la variante a nivel de cDNA siguiendo los estándares de la Human Genome Variation Society o HGVS [122] .
Condel_Prediction: predicción del algoritmo Condel [123].
Coverage: número de lecturas que cubren esa posición.
Genotype_quality: valor de calidad del genotipo.
Variant_quality: valor de calidad de la variante.
Interpro_ID: código de identificación del dominio proteico en Intepro.
Interpro_description: descripción del dominio proteico.
Disease: enfermedad relacionada con la variante según la base de datos HGMD.
PubMed_ID: publicaciones relacionadas con la variante según la base de datos HGMD.

Parte de las muestras del estudio retrospectivo se anotó frente a Ensembl 64 [124], en las que, además de generarse la información anteriormente descrita se añadieron los siguientes campos:
Freq_pop: frecuencia poblacional de la variante en distintas poblaciones.
SIFT_prediction: predicción del efecto de la variante a nivel proteico según el algoritmo SIFT.
PolyPhen_prediction: predicción del efecto de la variante a nivel proteico según el algoritmo PolyPhen [125].
Grantham_distance: distancia de Grantham entre aminoácidos [126].
Sequence: secuencia 50pb upstreams y downstreams de la variante.
Para llevar a cabo la priorización de genes candidatos se tomó el conjunto de genes con variantes cuyo efecto a nivel transcripcional hubiera resultado potencialmente dañino (variantes clasificadas como “Non-synonymous”, “Splice site”, “Frameshift”, “Stop gained” o “Stop lost”).

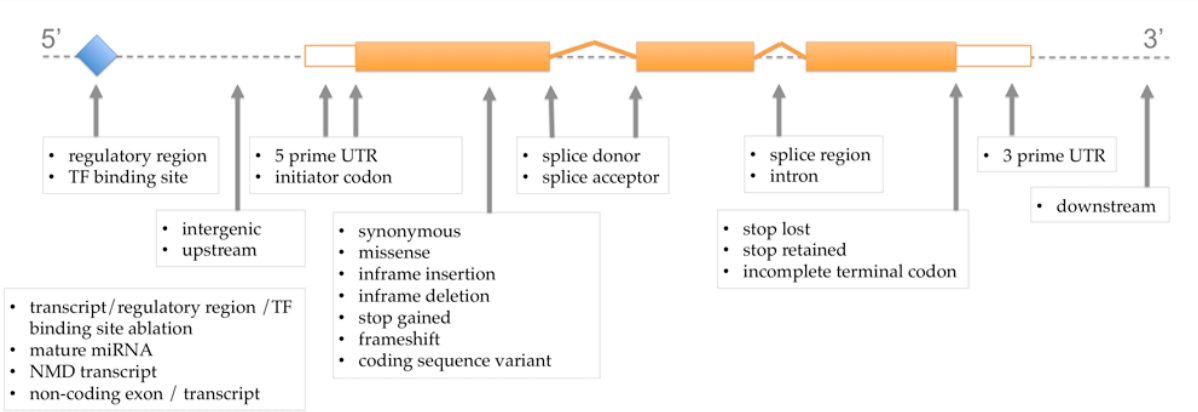


Figura 20. Clasificación de las variantes atendiendo a su efecto sobre el transcrito. Imagen tomada de Ensembl [127].

Se tomaron como variantes potencialmente patogénicas aquellas clasificadas como: 'missense', 'inframe insertion', 'inframe deletion', 'stop gained', 'frameshift', 'splice donor', 'splice aceptor', 'splice region', 'stop lost' o 'initiator codon variant'.

2.3.7 Sensibilidad y Especificidad en la llamada de variantes

La sensibilidad y especificidad de la llamada de variantes se calcularon siguiendo los pasos descritos en el apartado 1.2.6 de esta misma sección en base a los resultados obtenidos en el exoma completo tras aplicar los nuevos programas y criterios de filtrado para la llamada de variantes. Posteriormente, tras la validación del nuevo pipeline, se estimó la sensibilidad del panel en base a los resultados obtenidos en las 10 muestras control.

2.3.8 Esquema general del pipeline desarrollado

A continuación, se describen las principales etapas del análisis de las muestras pertenecientes al panel de genes (Figura 21).

Paso 1.- Evaluación del diseño de sondas para la captura de los genes diana. Paso exclusivo de este pipeline.

> $ sg_gene_panels_utilities.pl -- Proporciona los porcentajes de captura de los genes del panel atendiendo exclusivamente al diseño de las sondas. Adicionalmente, este script reporta el número de variantes registradas en la base de datos HGMD presentes en las zonas no cubiertas por el diseño realizado.
>
> -check -- Fichero en formato BED con las coordenadas de las sondas de captura generadas.
>
> -target -- Fichero en formato BED con las coordenadas empleadas para el diseño de las sondas.
>
> -prefix -- Prefijo que se añade a los ficheros de salida del script
>
> Output -- El fichero de salida consta de 11 columnas: 1) Nombre del gen=nombre HGNC del gen; 2) pb totales=pares de bases totales incluidas en el diseño; 3) %pb total no cubiertos=porcentaje de nucleótidos no cubiertos; 4) pb exones=bases pertenecientes a zonas de exones sin tener en cuenta UTRs; 5) %pb exones no cubiertos=porcentaje de bases correspondientes a zonas exónicas sin tener en cuenta las UTR que no se cubren; 6) pb splicing=nucleótidos correspondientes a zonas de splicing alternativo; 7) %pb splicing no cubiertos=porcentaje de nucleótidos en las zonas de splicing que no se cubren; 8) pb UTRs=nucleótidos en zonas UTR; 9) %pb UTRs no cubiertos=porcentaje de pares de bases en zonas UTR no cubiertas; 10) Vars. en HGMD=variantes totales descritas en la base de datos HGMD para un determinado gen dentro de las zonas diana; 11) % Vars. HGMD no cubiertas=porcentaje de variantes de la HGMD dentro de las zonas diana que no se cubren.

Paso 2.- Control de calidad de los datos de secuenciación brutos. Generación automática de los gráficos mediante FASTQC.

Paso 3.- Alineamiento de las lecturas con Bioscope (ver apartados 1.2.2 y 1.2.7 de esta misma sección).

Paso 4.- sg_pipe_Resecuenciacion_Variant_Calling.sh: script en Bash que contiene los pasos que forman parte de la evaluación de la eficiencia del kit de captura junto con la plataforma de secuenciación, la llamada de variantes y la anotación de variantes (pasos del 5 al 10). El script calcula los resultados asumiendo que el dataset procede de una secuenciación tipo 'paired-end' y los combina con los resultados obtenidos con el mismo dataset tratándolo como si procediera de una secuenciación tipo 'single-end' para minimizar el efecto que producen la menor calidad de las lecturas F5 así como las diferencias en la funcionalidad de los programas para la llamada de indels.

$ sg_pipe_resecuenciacion_dirigida.sh – Este script requiere 5 datos como input.

Input1 -- Fichero del alineamiento en formato BAM asumiendo secuenciación tipo 'paired-end'.

Input2 -- Fichero del alineamiento en formato BAM asumiendo secuenciación tipo 'single-end'.

Input3 -- Fichero de indels en formato GFF identificados siguiendo una estrategia para 'paired-end'.

Input4 -- Fichero de indels en formato GFF identificados siguiendo una estrategia para 'single-end'.

Input5 -- Nombre del panel. Dependiendo del panel (ejemplo: exoma o cardio1) las regiones diana son diferentes.

Paso 5.- Control de calidad de la eficiencia del kit de captura. Además de las herramientas expuestas en el Paso 4 del apartado 1.2.7 de esta misma sección, se añadieron tres puntos de control adicionales:

5a) Eliminación de lecturas con baja calidad de mapeo con Samtools view. El fichero de input es fichero BAM obtenido tras la eliminación de las lecturas duplicadas.

$ samtools view -b –q 1 muestra_sin_dup.bam > muestra_sin_dup_sinQ1.bam

5b) Identificación de posiciones no cubiertas

$ sg_panel_non_covered_stats.pl

-i -- Fichero BAM sin duplicados y sin lecturas con baja calidad de mapeo.

-prefix -- Prefijo añadido a los ficheros de salida

-o -- Directorio de salida

-p -- Nombreo del panel

-csv -- Nombre de los genes del panel

Output1 -- El primer fichero de salida es un listado con las bases no cubiertas así como el número de variantes de la HGMD incluidas en estas zonas.

Output2 -- El segundo de los ficheros tiene formato BED y contiene únicamente las regiones no cubiertas para poder ser visualizadas, junto con el alineamiento y las coordenadas del diseño, en la herramienta IGV para poder así evaluar de una forma visual su repercusión en un determinado gen.

5c) Reproducibilidad entre muestras atendiendo a la profundidad de lectura por base.

$ sg_pipe_coordenadas_comunes.sh – Genera una matriz con los valores de profundidad de lectura por muestra. El resultado de este fichero se cargó en R para calcular el coeficiente de correlación de Pearson entre muestras.

-Input -- El fichero de entrada es un fichero en formato PILEUP correspondientes a cada una de las muestras en estudio separados por espacios.

Output -- La salida del script es una matriz de valores donde cada línea representa una posición del diseño. El fichero contiene las siguientes columnas: cromosoma, posición genómica y profundidad de lectura por muestra en columnas independientes (tantas columnas como muestras haya).

Paso 6.- Realineamiento del fichero BAM alrededor de los indels con GATK (paso exclusivo de este pipeline). El programa emplea como input el fichero del alineamiento BAM sin duplicados y sin lecturas con baja calidad de mapeo, la base de datos dbSNP en formato VCF y las coordenadas de las zonas diana.

Paso 7.- Llamada de SNVs.

7a) Llamada de SNVs con Samtools – ver apartado 1.2.7 de esta misma sección. Los resultados se parsearon para obtener un fichero en formato VCF.

7b) Llamada de SNVs con GATK.

$ java -jar GenomeAnalysisTK.jar -T UnifiedGenotyper –R human_hg19.fa -glm SNP -stand_call_conf 20.0 -stand_emit_conf 20.0 -dcov 200 -I muestra_sin_dup_sinQ1_realineado.bam -o snps_GATK.vcf –L PanelCardio.bed – Identifica todos los SNVs posibles en el dataset a partir del fichero BAM realineado con una frecuencia alélica mayor a 0,2, parámetro por defecto.

$ java -jar GenomeAnalysisTK.jar -T VariantFiltration –R human_hg19.fa -filter "MQ0 >= 4 && ((MQ0 /(1.0*DP))> 0.1)" -filter "QUAL < 20.0" -filter "QD < 2.0" -filterName HARD -filterName QUAL_FILTER -filterName QD_FILTER -B:variant,VCF snps_GATK.vcf -o snps_GATK_filtrados.vcf – El comando filtra los SNVs añadiendo a la columna "FILTER" diferentes etiquetas: 1) "HARD" si la variante es identificada por 4 o más lecturas con un valor de calidad mapeo de 0 (MQ0) o si el valor ((MQ0 /(1.0*DP))> 0.1) siendo DP el número de lecturas que cubren una determinada posición, 2) "QUAL_FILTER" si el valor de calidad de la variante es menor de 20 y 3) "QD_FILTER" si el valor de "Quality by depth (QD)" es menor de 2. Además de eliminar las lecturas que no hubieran pasado alguno de estos 3 filtros, se eliminaron aquellas variantes con una profundidad de lecturas menor de 9x.

Paso 8.- Identificación de indels.

8a) Small Indel Tool – ver apartado 1.2.7 de esta misma sección. Los resultados se parsearon para generar un fichero en formato VCF.

8b) Llamada de indels con GATK.

$ java -jar GenomeAnalysisTK.jar -T UnifiedGenotyper –R human_hg19.fa -I muestra_sin_dup_sinQ1_realineado.bam -glm DINDEL –o indels_GATK.vcf –L PanelCardio1.bed – Identifica todos los indels posibles en el dataset a partir del fichero BAM realineado con una frecuencia alélica mayor a 0,2, parámetro por defecto.

$ java -jar GenomeAnalysisTK.jar -T VariantFiltration –R human_hg19.fa -filter "MQ0 >= 4 && ((MQ0 /(1.0*DP))> 0.1)" -filter "QUAL < 20.0" -filter "QD < 2.0" -filterName HARD -filterName QUAL_FILTER -filterName QD_FILTER -cluster 3 -window 10 -B:variant,VCF indels_GATK.vcf –o indels_GATK_filtrados.vcf – Se emplearon los mismos criterios de filtrado expuestos en el paso anterior.

Paso 9.- Anotación de SNVs e indels. Los ficheros de resultados de variantes se combinaron con los VCFTools y su módulo 'isec' para generar un único fichero de resultado para cada muestra priorizando los resultados obtenidos por GATK sobre Samtools. Este fichero en

formato VCF fue el input para el script VEP de Ensembl modificado para la versión 62 de esta base de datos. Adicionalmente, se añadieron los datos de la base HGMD Proffesional, diferente a la versión pública de la que hace uso Ensembl, a los resultados del script VEP modificado.

Los resultados del pipeline bioinformático fueron posteriormente analizados por los expertos en Genética Médica de Sistemas Genómicos iniciando el análisis por aquellas variantes con una profundidad de lectura de al menos 20x. En caso de no identificar ninguna variante candidata este límite fue reducido de forma progresiva. Además de los resultados proporcionados por el pipeline bioinformático desarrollado, el grupo de expertos en diagnóstico genético empleó las siguientes herramientas:

-Alamut [128] para confirmar y visualizar las variantes junto con información biológica adicional.

-IGV para visualizar el alineamiento donde se encuentra la variante en caso de que exista duda sobre la existencia de la misma debido, por ejemplo, a su baja profundidad de lectura o su cercanía con otras variantes.

-Revisión de la literatura científica incluyendo la existencia de estudios funcionales, modelos animales, coincidencia o no de fenotipo entre el paciente y los descritos en la literatura.

-Clasificación en bases de datos específicas de locus o enfermedad.

-Co-ocurrencia de la variante con una mutación patogénica en un paciente anterior o en uno descrito en la literatura.

-Estudios de cosegregación.

Tras realizar este análisis, las variantes se clasificaron en: 1) mutaciones patogénicas (mutaciones conocidas y mutaciones nuevas claramente patogénicas), 2) variantes de significado desconocido muy probablemente patogénicas, 3) variantes de significado desconocido probablemente no patogénicas y 4) variantes de significado desconocido inciertas. Todas las variantes clasificadas en las categorías 1 y 2 fueron confirmadas por Sanger.

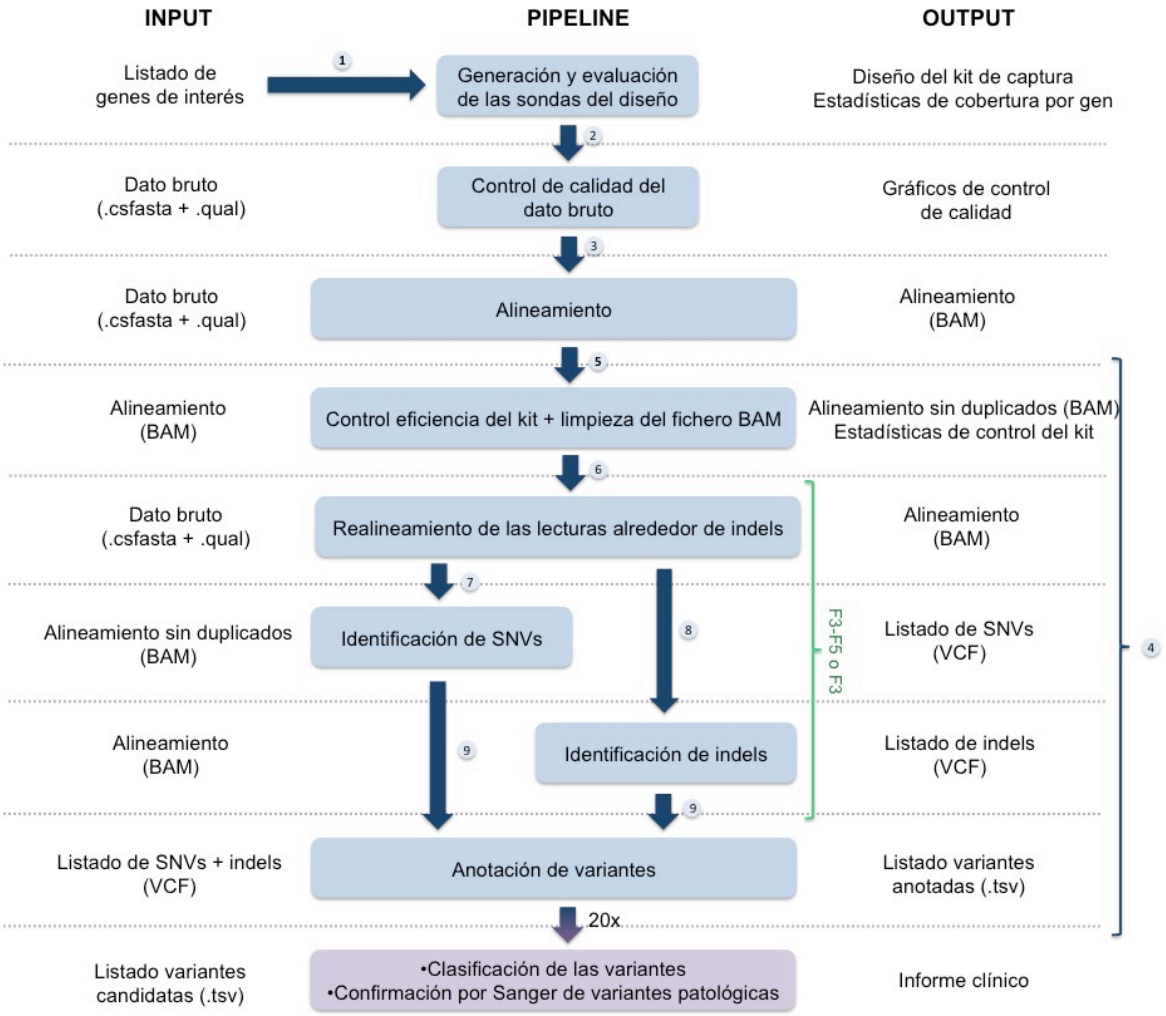


Figura 21. Pipeline aplicado al análisis del panel de enfermedades cardiacas.

3 Análisis de los genes BRCA1 y *BRCA2* en mini-secuenciadores con tecnología NGS

A continuación, se detalla el desarrollo de un protocolo de análisis para el estudio de los genes *BRCA1* y *BRCA2* en muestras procesadas en mini-secuenciadores con tecnología NGS para el diagnóstico molecular de cáncer de mama y ovario a partir de la modificación del pipeline establecido en el apartado 2 de esta misma sección.

3.1 Preparación de las muestras y secuenciación

La preparación de las muestras y posterior secuenciación fueron llevadas a cabo por el departamento de Nuevas Tecnologías de la empresa Sistemas Genómicos.

3.1.1 Obtención del material biológico

Para llevar a cabo la puesta a punto del pipeline de análisis se usaron 6 muestras control de sangre periférica de individuos que habían sido previamente diagnosticados mediante secuenciación tradicional por Sanger. Las muestras fueron seleccionadas en base a las diferentes mutaciones que presentaban los individuos. Las 6 muestras control, que presentaban un total de 32 variantes, se secuenciaron en el mismo run codificado como BF42 (Tabla 4). Asimismo, en un run posterior, se secuenció una línea celular de HapMap, BRCA12144, empleada anteriormente en la puesta a punto del exoma (ver sección 1.1.1), con el fin de confirmar la presencia de las variantes descritas para esta muestra en los genes BRCA por el consorcio de consorcio de HapMap mediante arrays de genotipado masivo. Adicionalmente, tal como refleja la Tabla 4, se añadieron una serie de muestras control en runs sucesivos a fin de establecer y verificar las limitaciones de la tecnología en estudio.

Tabla 4. Variantes identificadas en las muestras control. Las mutaciones causales se encuentran marcada en negrita y en cursiva.

Run	MID	Muestra	Gen	
			BRCA1	***BRCA2***
BF42	1	09S491	***NM_007294.3:c.815-824dup10***	-
	3	09S218	NM_007294.3:c.2311T>C NM_007294.3:c.4308T>C NM_007294.3:c.548-58_548-58delT NM_007294.3:c.4485-64C>G NM_007294.3:c.3548A>G	***NM_000059.3:c.9026_9030delATCAT*** NM_000059.3:c.7806-14T>C NM_000059.3:c.1114C>A
	5	10S068	***NM_007294.3:c.66_67delAG***	-
	6	10S1106	***NM_007294.3:c.211A>G*** NM_007294.3:c.2077G>A NM_007294.3:c.3113A>G NM_007294.3:c.2311T>C NM_007294.3:c.4308T>C NM_007294.3:c.2082C>T NM_007294.3:c.5075-53C>T NM_007294.3:c.3548A>G	NM_000059.3:c.-26G>A NM_000059.3:c.7806-14T>C NM_000059.3:c.3396A>G NM_000059.3:c.7796A>G
	7	09S432	-	***NM_000059.3:c.715dupA*** NM_000059.3:c.681+56C>T
	8	09S523	NM_007294.3:c.3548A>G NM_007294.3:c.2311T>C NM_007294.3:c.4308T>C NM_007294.3:c.4837A>G NM_007294.3:c.2082C>T NM_007294.3:c.4485-64C>G	***NM_000059.3:c.658_659delGT*** NM_000059.3:c.7806-14T>C
BF67	3	BRCA09880	***Deleción de los exones del 1 al 3***	-
	8	BRCA09992	***Deleción de los exones del 8 al 13***	-
BF79	1	BRCA10714	-	***NM_000059.3:c.6275_6276delTT***
	2	BRCA10726	-	***NM_000059.3:c.9310_9311delAA***
BF80	6	BRCA12144	NM_007294.3:c.442-34C>T + 28 posiciones igual a la referencia	NM_000059.3:c.-26G>A NM_000059.3:c.3396A>G NM_000059.3:c.4563A>G NM_000059.3:c.7397T>C NM_000059.3:c.7806-14T>C NM_000059.3:c.8755-66T>C NM_000059.3:c.*105A>C + 26 posiciones interrogadas igual a ref.
BF87	2	BRCA11314	-	***NM_000059.3:c.5720_5723del***
BF96	6	BRCA11928	-	***NM_000059.3:c.956dupA***
DB06	8	BRCA12836	-	***Deleción del inicio del exon 11***
	6	BRCA12773	-	***Deleción del exon 2***

3.1.2 Construcción de librerías y captura de zonas específicas

La captura de los genes *BRCA1* y *BRCA2* se realizó mediante un sistema comercial de amplificación por PCR en multiplex, BRCA MASTR® de Multiplicom, siguiendo el procedimiento estándar marcado por el fabricante [129]. Este sistema, implica la

construcción para cada muestra de una librería de fragmentos a partir de primers de fusión, propios de la plataforma 454, que incluyen primers específicos diseñados para amplificar las zonas de interés (Figura 22). Adicionalmente, con el propósito de optimizar la capacidad del secuenciador y abaratar los costes en la amplificación y secuenciación de las muestras, éstas fueron procesadas en pool empleando un sistema de etiquetas denominadas Multiplex Identifiers o MIDs, similar al sistema de etiquetas expuesto en la plataforma SOLiD en la sección 2.2.2.

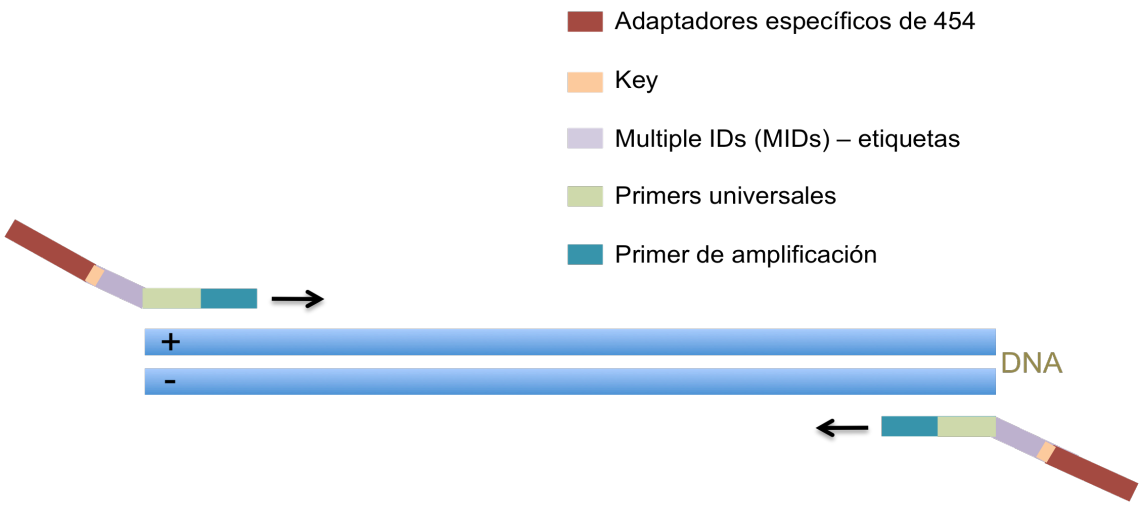


Figura 22. Preparación de librerías para la captura de los genes *BRCA1* y *BRCA2* en secuenciadores 454-Roche. La librería incluye primers específicos para la amplificación por PCR de las regiones de interés así como adaptadores específicos de la plataforma de secuenciación a través de los cuales el fragmento de DNA se une a la esfera de secuenciación, se amplifica y posteriormente se secuencia. Las librerías generadas para cada una de las muestras se marcan con diferentes etiquetas o MIDs, a través de las cuales es posible identificar las lecturas pertenecientes a esa muestra del resto de lecturas del pool una vez terminado el proceso de secuenciación.

A excepción de las muestras incluidas en los runs BF42 y BF67, que se capturaron con la versión 2.0 del kit BRCA MASTR, el resto de muestras empleadas en este estudio se capturaron con la versión 2.1 del kit. Las diferencias entre el kit 2.0 y 2.1 se centraron en una mejora del rendimiento de las PCRs del exón 7 y del exón 14.

3.1.3 Amplificación del DNA y enriquecimiento de esferas

De forma similar a la plataforma SOLiD, la amplificación y el enriquecimiento de esferas monoclonales en los secuenciadores 454-Roche también se realizan en emulsión tras la

unión del fragmento de DNA a una esfera sintética de agarosa (ver Métodos 1.1.4). A continuación, las esferas son depositadas sobre placas microperforadas (PicoTiterPlate^{TM}) que contienen millones de pocillos de un volumen de 75 picolitros que aseguran que solamente una esfera de agarosa, que soporta las miles de copias de un mismo fragmento de DNA, será depositada en un mismo pocillo permitiéndose así la optimización del espacio de secuenciación (Figura 20A).

3.1.4 Secuenciación

El método de secuenciación por síntesis incorporado en las plataformas 454-Roche se denomina pirosecuenciación [9, 10]. En la pirosecuenciación, tras la deposición de las esferas de agarosa en los micropocillos, se añaden unas esferas adicionales que contienen las enzimas necesarias para la síntesis de DNA (Figura 23A). A continuación, se añade uno de los 4 dNTPs posibles de manera que por cada incorporación de un dNTP a la cadena en síntesis se libera un grupo pirofosfato o PPi el cual desencadena una serie de reacciones que finalmente producen una señal luminosa que es captada por un sistema óptico altamente sensible (Figura 23B). La cantidad de luz emitida es proporcional al número de grupos pirofosfato liberados de manera que si en la cadena de ssDNA molde existen varios nucleótidos consecutivos que corresponden al mismo dNTP se genera una mayor intensidad de la señal. Posteriormente, se eliminan los dNTPs que no se hayan unido a la cadena en síntesis y se añade un nuevo dNTP. El proceso se repite hasta alcanzar una longitud media de 200-450pb de longitud, dependiendo de la aplicación y versión de la plataforma [130]. A partir de las señales luminosas emitidas en cada ciclo de primers se genera un diagrama por pocillo denominado fluograma, en el que cada pico representa el núcleotido identificado en cada posición y su altura representa el número de nucleótidos de ese tipo consecutivos (Figura 23C).

3.2 Análisis del panel

En las siguientes secciones se detalla el desarrollo de una serie de pipelines adaptados a la tecnología de secuenciación 454-Roche generado a partir del pipeline generado para el análisis del panel de genes presentado en el apartado 2 de esta sección. Cada una de las muestras control fue analizada dos veces, una primera vez haciendo uso de la primera versión del pipeline para 454-Roche, y una segunda vez ejecutando un pipeline más actualizado de forma retrospectiva.

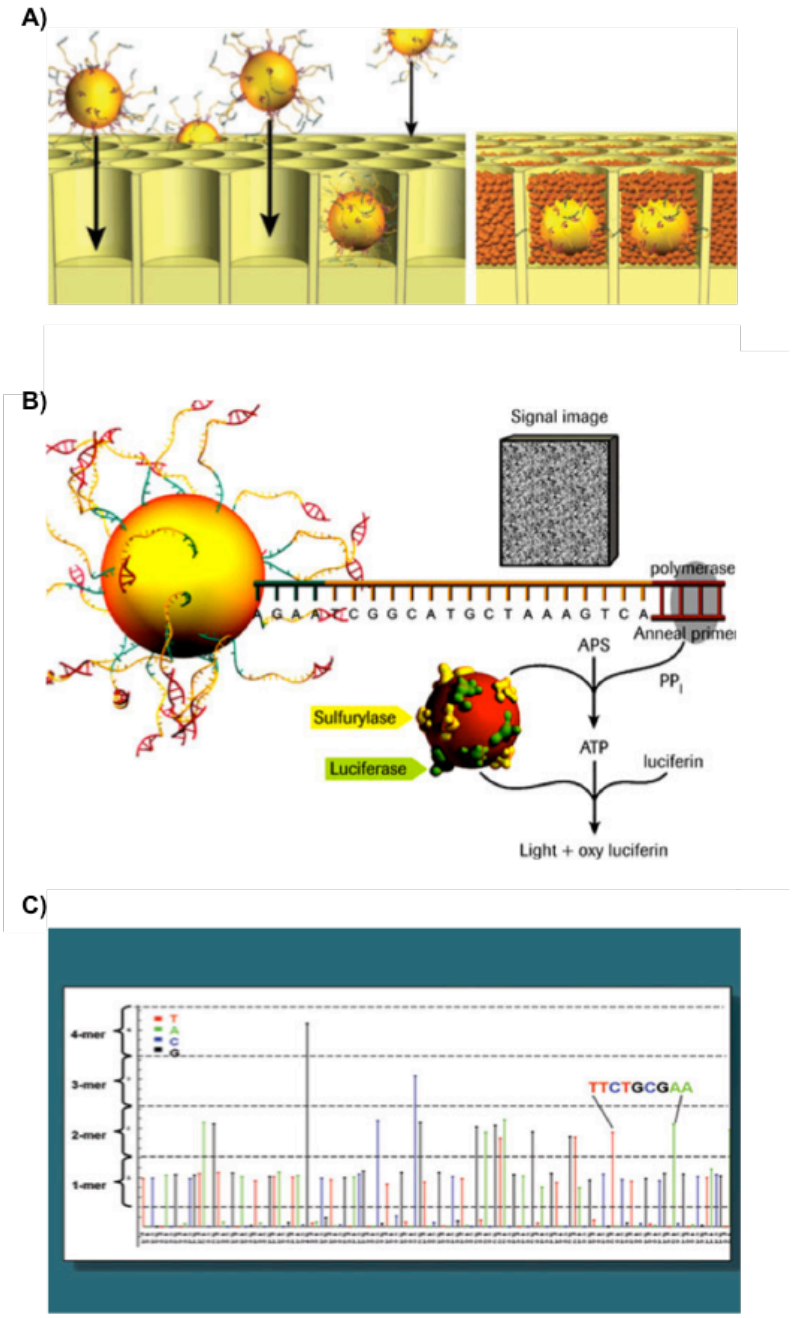


Figura 23. Pirosecuenciación. A) Deposición de esferas de agarosa con los fragmentos de DNA amplificados y deposición de las esferas que contienen las enzimas de secuenciación en la placa PicoTiter^{TM}. B) Reacciones enzimáticas generadas a partir de la incorporación de un dNTP en la cadena de DNA en síntesis. C) Fluograma. Imágenes tomadas de [130]

3.2.1 Control de calidad de los datos de secuenciación

En el caso de las plataformas 454-Roche, tras la secuenciación, se genera un único fichero en formato SFF (del inglés “Standard Flowgram Format”) [131] para todas las muestras que han de ser posteriormente transformado en ficheros FASTQ para poder ser subsecuentemente procesado. La extracción de las secuencias en formato FASTQ se realizó mediante el uso del programa sfffile, incluido en las SFF Tools como parte del paquete de programas de 454-Roche 454 Sequencing System Software Packaged. Posteriormente, se realizó un screening de las lecturas generadas para eliminar: 1) adaptadores específicos del kit de Multiplicom; 2) bases del extremo 3’ de las lecturas con un valor de calidad Phred menor de 30; 3) lecturas con una longitud superior al tamaño máximo de PCR incluyendo los adaptadores (aproximadamente 400nt). El paso 1 se realizó con el software Cutadapt [132], versión 0.4 para el pipeline inicial y versión 1.3 en el pipeline actualizado. Los pasos 2 y 3 se incluyeron en el pipeline más reciente donde se empleó Prinseq-lite-v0.19.5 [133].

El software utilizado para verificar la calidad de los datos brutos de secuenciación para este estudio fue FASTQC v0.10.1 (ver Métodos 2.3.2).

3.2.2 Alineamiento

El desarrollo del pipeline inicial incluyó el software SMALT [134], una versión mejorada del programa SSAHA2 [135] con una alta eficiencia en el mapeo de lecturas largas. Este software emplea un sistema de indexado de la secuencia de referencia mediante la generación de tablas de arreglos asociativos (en ingles ‘hash tables’) a partir de segmentos de secuencia no solapantes de una longitud menor a 21pb que se encuentran a una distancia equidistante (Figura 24). Durante la generación del índice de la referencia es posible eliminar aquellos segmentos que se repiten muchas veces en el genoma optimizando así el proceso de mapeo. Una vez generado el índice, cada lectura, en formato FASTQ, es dividida en fragmentos de la misma longitud que el índice de la secuencia de referencia. A continuación, el algoritmo trata de identificar secuencias homólogas o ‘hits’ en el índice creado para la secuencia de referencia. Posteriormente, el software elige aquellos ‘hits’ de la misma lectura que se posicionan en zonas cercanas e intenta alinearlos de forma conjunta frente a la secuencia de referencia mediante el algoritmo de Smith-Waterman bandeado. Mediante este proceso el software también es capaz de permitir inserciones y deleciones entre dos hits consecutivos asumiendo que ambos difieren en un número mínimo de bases. Este tipo de programas de mapeo son más lentos (necesitan más memoria RAM

para ser ejecutados), menos sensibles en regiones repetitivas y más tolerantes con las zonas genómicas de alta variabilidad.

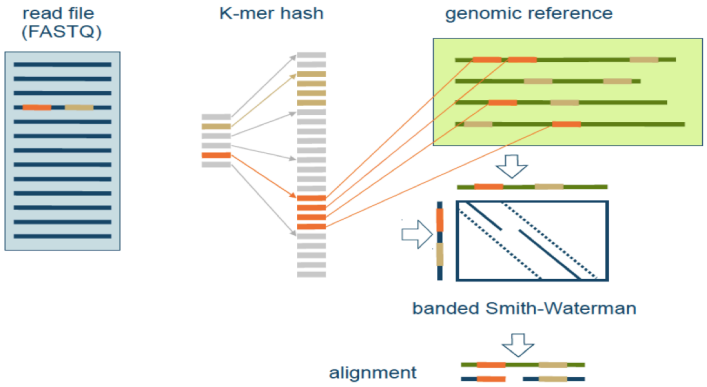


Figura 24. Alineamiento de las lecturas mediante el algoritmo SMALT.

Debido a la lentitud en el mapeo de SMALT y a la alta especificidad de un sistema de captura como el empleado, se tomó como referencia exclusivamente las secuencias génicas pertenecientes a *BRCA1* (chr17:41,196,312-41,277,500) y *BRCA2* (chr13: 32,889,611-32,973,805), según la base de datos de Ensembl. Para llevar a cabo el indexado de esta referencia se eligió una longitud de fragmento de 20. La distancia de separación entre los fragmentos de la referencia se fijó en 6 nucleótidos. El alineamiento de las lecturas de GSJunior se realizó empleando los parámetros de mapeo por defecto. Posteriormente, en el pipeline más actualizado, se modificó el programa de mapeo por Bowtie2, una pieza de software igualmente eficiente, más flexible y sofisticada basada en el algoritmo de "Burrows-Wheeler" [136] (Figura 25).

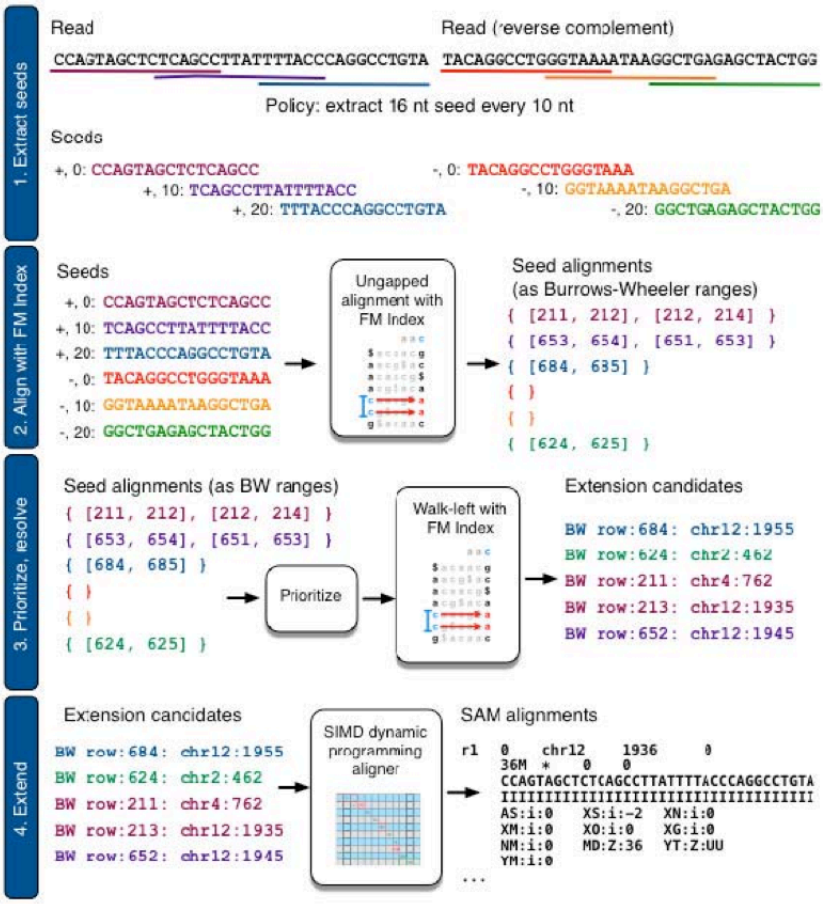


Figura 25. Sistema de alineamiento de lecturas con Bowtie 2 (Imagen tomada de [137]). El alineamiento de una lectura se divide en cuatro fases: 1) generación de sub-secuencias semilla (“seeds”) solapantes tanto para la lectura en sí como para su reverso complementario; 2) alineamiento de estas sub-secuencias frente a la referencia indexada empleando la transformación de Burrow-Wheeler (BW) comprimida (index FM); 3) priorización de los rangos BW y priorización de los mismos comenzando por los rangos más pequeños; 4) extensión de los alineamientos de las sub-secuencias para obtener alineamientos definitivos, bien de la lectura completa o bien de manera parcial, a través de un método de programación dinámica acelerada basado en SIMD (Single Instruction Multiple Data) que le permite ser más eficiente y flexible en la identificación de indels.

3.2.3 Control de calidad de la eficiencia del kit de captura y filtrado de lecturas útiles

Al igual que en los apartados 1.2.3 y 2.3.4 de esta sección, se llevaron a acabo una serie de comprobaciones sobre la eficiencia de la tecnología de captura con el objetivo de determinar posibles desviaciones durante el proceso de generación de las librerías o en su posterior secuenciación, adaptados al tipo de captura empleado y la plataforma de secuenciación.

- Sensibilidad del kit de captura: el cálculo de este parámetro se realizó a partir de los valores medios de profundidad de lectura por amplicón para cada una de las muestras pertenecientes al mismo run con el fin de confirmar una presencia mínima de lectura a lo largo de todo el panel y una distribución de lecturas 'homogénea' entre muestras. Para ello, en el pipeline inicial, se procesó el fichero BAM de los alineamientos individuales de cada muestra para obtener un fichero en formato PILEUP según Samtools v.1.9 (ver apartado 1.2.4 para más detalles sobre este formato) con el valor de profundidad de lectura por posición. Los ficheros en formato PILEUP individuales por muestra se combinaron en un único fichero a partir del cual se calculó la profundidad de lectura media para todos los amplicones en cada una de las muestras. Los resultados proporcionados por este script se visualizaron de dos formas diferentes, ordenando los valores medios de profundidad de lectura por amplicón atendiendo a la posición de la PCR en el gen o atendiendo al grupo de PCR multiplex al que pertenecía cada una de las PCRs. Para llevar a cabo este control adicional, se generaron scripts propios (scripts 'in house') escritos en Bash y Perl. En el estudio retrospectivo, donde se aplicó la última versión del pipeline, este punto de control fue actualizado sustituyendo el comando de samtools pileup por samtools depth (versión 0.1.16), que permite generar la matriz de profundidad de lectura por posición y por muestra, obtenida anteriormente mediante la concatenación de los diferentes ficheros en formato pileup empleando diferentes scripts generados para ese propósito, de una forma más eficiente. Adicionalmente, en este segundo pipeline, se generó un script en R para generar los gráficos de las figuras resultantes (Figuras 26 y 27) de forma automática.

- Especificidad del kit de captura: debido a que se trataba de un sistema de captura por PCR comercial se asumió que este parámetro sería prácticamente del 100% careciendo de importancia para la evaluación del sistema de captura.

- Porcentaje de duplicados de PCR: a diferencia de los sistemas de captura por hibridación de sondas, los protocolos de captura mediante PCR implican que todas las lecturas se generan a partir del mismo punto del amplicón y por lo tanto, a diferencia de los pipelines diseñados en los dos apartados anteriores (ver Métodos 1.2 y 2.3), no se pudieron eliminar las lecturas procedentes de duplicados de PCR.

- Reproducibilidad de la tecnología de captura: la matriz de valores de profundidad de lectura por posición generada para evaluar la sensibilidad del kit, fue utilizada para calcular el coeficiente de correlación de Pearson entre muestras.

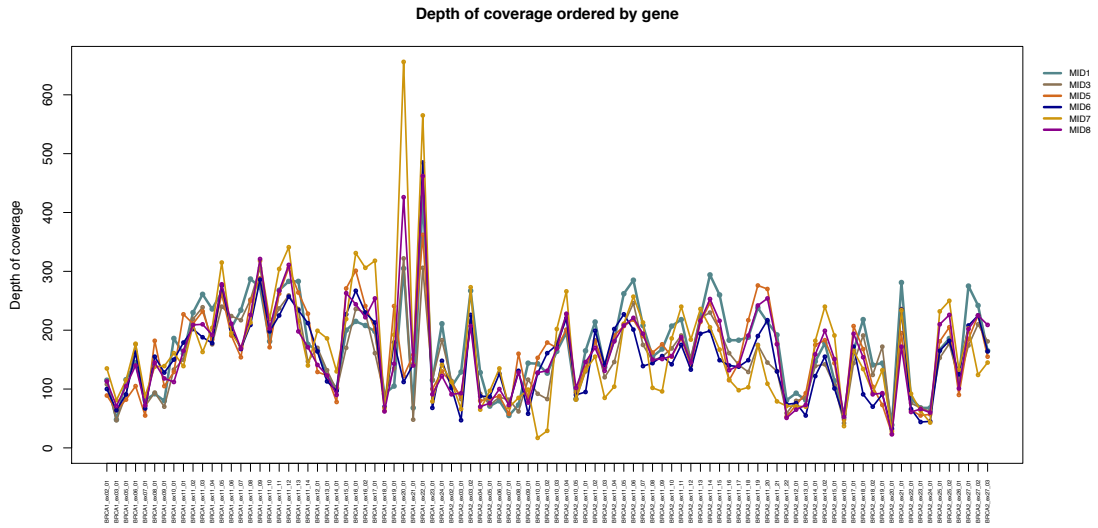


Figura 26. Profundidad de lectura media por amplicón ordenada por posición del amplicón en el gen. El nombre de cada PCR incluye la siguiente información en este orden: Gen+"_"+Exon_amplificado+"_"+Número_de_PCR_en_ese_exón. De esta forma, BRCA1_ex16_02 corresponde al gen BRCA1, exón 16 y PCR número 2 de las dos PCRs diseñadas para amplificar el exón 16.

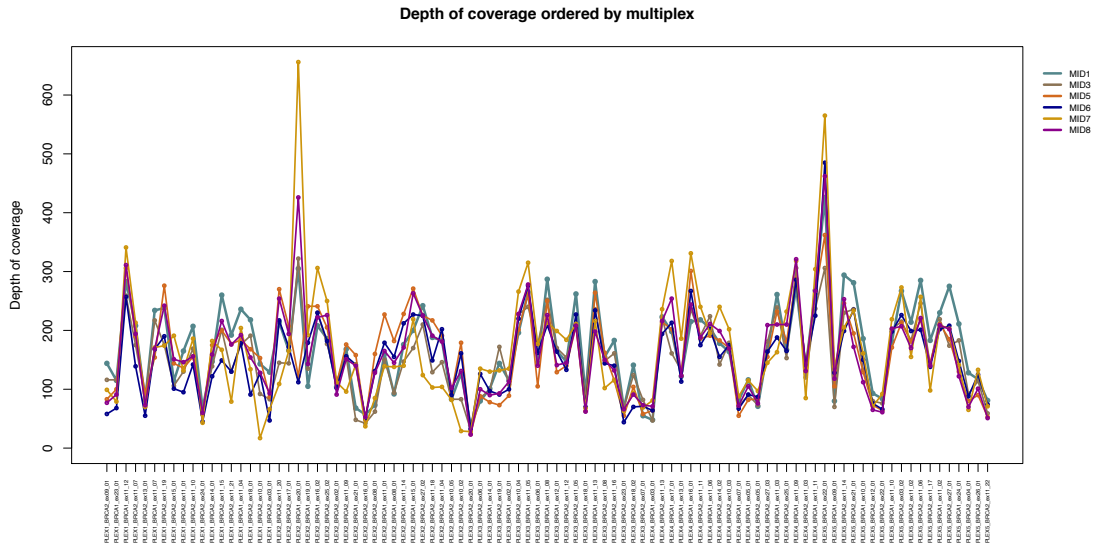


Figura 27. Profundidad de lectura media por amplicón ordenada por grupo de multiplex. El nombre de cada PCR incluye la siguiente información en este orden: Número_de_multiplex+"_"+Gen+"_"+Exon_amplificado+"_"+Número_de_PCR_en_ese_exón . Así, PLEX2_*BRCA2*_ex11_18 corresponde a la multiplex 2, gen *BRCA2*, exón 11 y PCR número 18 diseñada para amplificar el exón 11.

3.2.4 Regiones diana y variantes patogénicas o de susceptibilidad no cubiertas

Se aplicó el mismo procedimiento desarrollado en el apartado 2.3.4 de esta misma sección.

3.2.5 Identificación de variantes puntuales y pequeños indels

La detección de variantes implementada en el pipeline inicial combinó los resultados obtenidos por Samtools v.0.1.9 y GATK v.1.0.5777, herramientas y versiones previamente utilizadas en el análisis del dataset presentado el apartado 2 de esta sección. La llamada de variantes en el pipeline actualizado se llevó a cabo mediante la combinación de GATK v2.4 y un nuevo software que reemplazó a Samtools, VarScan2 v2.2.11 [138, 139]. A diferencia de GATK y Samtools, que se basan en métodos bayesianos, VarScan2 se apoya en métodos heurísticos más simples que GATK o Samtools y por lo tanto puede ser más robusto en zonas donde la profundidad de lectura es muy extrema, en el caso de analizar varias muestras en pool o cuando las muestras no son totalmente puras como en el caso de las muestras tumorales. VarScan2 toma como fichero de entrada un fichero en formato mpileup generado por Samtools similar al formato PILEUP pero preparado para presentar los datos por posición para una o más muestras a la vez en la misma línea, a partir del cual identifica tanto variantes puntuales como pequeños indels.

3.2.6 Identificación de CNVs

Se implementó una estrategia para la detección de Copy Number Variants (CNVs) a partir de los resultados obtenidos en la estrategia descrita en el apartado 3.2.3 para el estudio de la cobertura. Para ello, en cada run, se normalizaron los valores de la profundidad media de lectura por amplicón dividiendo cada uno de ellos por el número de lecturas totales y posteriormente por el valor medio para todas las muestras en ese amplicón. A continuación, para cada muestra, se dividió este valor resultante por la media de los valores para todos los amplicones del panel. Los resultados de esta normalización se presentaron inicialmente en una tabla tabulada para generar un gráfico a partir de una plantilla preconfigurada en Excel. En la última versión del pipeline, estos gráficos se obtuvieron de forma automatizada empleando un script en R (Figuras 25 y 26).

Las potenciales deleciones detectadas tras la observación de los gráficos generados fueron contrastadas con los datos obtenidos en el laboratorio mediante la visualización de los perfiles de fragmentos de las PCRs o mediante la técnica de MLPA (Multiplex ligation-dependent probe amplification).

3.2.7 Anotación de variantes

En el pipeline inicial, la anotación de las variantes se llevó a cabo empleando la estrategia descrita en la el apartado 2.3.6 de esta sección. Posteriormente, en el pipeline actualizado, se cambió la base de datos a Ensembl versión 71 [127] donde se incluían nuevas categorías para estratificar las muestras según su posición o efecto en el transcrito (Figura 27). Tras ejecutar la nueva versión del script de Ensembl "Variant Effect Predictor", vinculada a la versión 71 de la base de datos, y combinar los resultados con la información técnica de las variantes, se obtuvieron los siguientes campos de información:

HGNC_symbol: símbolo HGNC del gen.

Chr: cromosoma.

Pos: posición genómica.

Ref_Allele: alelo en la referencia.

Var_Allele: alelo en la muestra.

Sample_Genotype: genotipo identificado en la muestra en función de la frecuencia del alelo alternativo siguiendo la clasificación expuesta en la Tabla 5.

Sample_Depth: profundidad de lectura alcanzada en esta posición.

Sample_Var/Depth: frecuencia del alelo alternativo.

Gene_ID: código de identificación del gen en Ensembl.

Gene_description: descripción del gen.

HGVSc_name: nomenclatura de la variante siguiendo los estándares de la HGVS a nivel de cDNA.

Intron: número de intrón.

Exon: número de exón.

HGVSp_name: nomenclatura de la variante siguiendo los estándares de la HGVS a nivel proteico.

Variant_effect: efecto de la variante a nivel transcripcional.

ALL_MAF: frecuencia alélica global menor obtenida a partir de los datos de la fase 1 de "1000 Genomes".

AFR_MAF: frecuencia alélica menor en la población africana.

AMR_MAF: frecuencia alélica menor en la población americana.

ASN_MAF: frecuencia alélica menor en la población del este asiático.

EUR_MAF: frecuencia alélica menor en la población europea.

InterPro_IDs: código del dominio proteico afectado en InterPro [140].

InterPro_descriptions: descripción del dominio proteico en InterPro.

HGMD_info: código de identificación de la HGMD Professional, nomenclatura HGNC del cambio con el transcrito según la base de datos RefSeq [141] y nombre de la enfermedad asociada.

Related_publication: artículos relacionados con la variante.

Existing_variation: código de identificación de la variante en caso de existir.

Transcript_ID: código de identificación del transcrito en Ensembl.

RefSeq_ID: código de identificación del transcrito en RefSeq.

CCDS_ID: código de identificación del transcrito en CCDS [142].

Canonical_isoform: indica si la isoforma a la que afecta la variante es la isoforma canónica o no.

Conservation_score: valor de conservación GERP de las bases afectadas por la variante.

Grantham_distance: distancia de Grantham.

Condel_prediction: predicción del efecto de la variante a nivel proteico según el predictor Condel.

SIFT_prediction: predicción del efecto de la variante a nivel proteico según SIFT.

PolyPhen_prediction: predicción del efecto de la variante a nivel proteico según PolyPhen.

Flanking_sequence: secuencia a ambos lados de la variante, 50pb upstreams y 50pb downstreams.

Las columnas 'Sample_Genotype', 'Sample_Depth' y 'Sample_Var/Depth' se generan tantas veces como muestras se hayan analizado a la vez.

Categoría	Frecuencia en SNVs	Frecuencia en Indels
Homo_ref	0	0
P_Homo_ref	>0-0,12	NA
UNC_Hetero	>0,12-0,35	>0-0,3
P_Hetero	>0,35-0,65	>0,3-0,6
UNC_Homo_var	>0,65-0,85	NA
P_Homo_var	>0,85-1	>0,6-1

Tabla 5. Clasificación de las variantes atendiendo a su frecuencia alélica. Homo_ref=homocigota para la referencia; P_Homo_ref=probable homocigoto para la referencia; UNC_Hetero=heterocigoto incierto; P_Hetero=probable heterocigoto; UNC_Homo_var=homocigoto incierto para el alelo alternativo; P_Homo_var=probable homocigoto para el alelo alternativo.

3.2.8 Sensibilidad y Especificidad en la llamada de variantes

La sensibilidad y la especificidad en la llamada de variantes se calcularon siguiendo la metodología descrita en el apartado 1.2.6 de esta misma sección.

3.2.9 Esquema general del pipeline desarrollado

A continuación, se describen las estapas del análisis para el panel de genes BRCA (Figura 28).

Paso 1.- sg_pipe_GSJunior_BRCA.sh: script en Bash que contiene el pipeline completo para el análisis de las muestras de BRCA secuenciadas en una plataforma de pirosecuenciación 454-Roche.

$ sg_pipe_GSJunior_BRCA.sh – Toma 5 datos como input

Input1 : Fichero .sff que contiene el dato de secuenciación bruto.

Input2 : Versión del kit de captura, 2.0 ó 2.1

Input3 : Lista de MIDs separados por comas

Input4 : Fichero con el nombre de las muestras (una muestra por línea en el mismo orden que los MIDs)

Input5 : Código interno del conjunto de muestras

El script genera una serie de directorios a partir de los cuales deriva la información dependiendo del tipo de resultado que se obtenga atendiendo al siguiente esquema:

|-- analysis : contiene todos los ficheros finales resultantes del análisis
| |-- annotation : resultados de la anotación de las variantes
| |-- mapping : resultados del mapeo de las lecturas
| |-- stats : directorio que contiene las estadísticas calculadas
| | |-- coverage : estadísticas de la eficiencia del kit
| | |-- primary : estadísticas del dato de secuenciación limpio
| |-- variants : resultado de la llamada de variantes
|-- rawdata : dato bruto
`-- trash : directorio que contien los ficheros temporales
  |-- annotation : ficheros temporales de la anotación de las variantes
  |-- mapping : ficheros temporales del mapeo de las lecturas
  |-- variants : ficheros temporales de la llamada de variantes

Paso 2.- Conversión del dato de secuenciación al formato FASTQ y limpieza de las lecturas generadas.

$ sfffile -c 200 -r -s -xlr fichero.sff – Separa el fichero inicial en formato SFF [143] en ficheros individuales por muestra en el mismo formato.

$ sffinfo -s muestra.sff > lecturas.fasta – Extrae las secuencias en formato fasta para cada muestra

$ sffinfo -q muestra.sff > lecturas.qual – Extrae los valores de calidad asociados a cada lectura en formato fasta.

$ sg_fastaQual2fastq.pl – Genera un fichero en formato FASTQ a partir de las lecturas y sus valores de calidad asociados ubicados en dos ficheros independientes en formato FASTA.

Input1 -- Fichero FASTA con las lecturas.
Input2 -- Fichero FASTA con los valores de calidad asociados a las lecturas.
Output1 -- Fichero con la totalidad de lecturas generadas en formato FASTQ.

$ cutadapt -O 5 secuencia_primer muestra.fastq > muestra_sin_adaptador.fastq – Elimina los adaptadores si el número de bases homólogo entre el adaptador y la secuencia es de al menos 5pb. Este comando se ejecutó de forma recursiva para los dos adaptadores de Roche AAGACTCGGCAGCATCTCCA y TGGAGATGCTGCCGAGTCTT así como con sus reversos complementarios.

$ prinseq-lite.pl -trim_qual_right 30 -max_len 400 –fastq muestra_sin_adaptador.fastq -out_good muestra_limpia.fastq -out_bad lecturas_baja_calidad.fastq – Recorta las lecturas si la calidad en el extremo 3' es menor de 30 y permite una longitud máxima de lectura de 400pb.

Paso 3.- Control de calidad del dato de secuenciación tras la eliminación de adaptadores y lecturas de baja calidad.

Paso 4.- Alineamiento con Bowtie2 empleando la opción "--very-sensitive".

Paso 5.- Control de calidad de la eficiencia del kit de captura. Similar al 'Paso 4' de la sección 1.2.7 a excepción del cálculo del tamaño de inserto ya que se trata de muestras en las que solamente se secuenció un extremo del fragmento de DNA (secuenciación 'single-end').

- No se eliminaron lecturas duplicadas dada la naturaleza del kit de captura.
- No se descartaron lecturas con una baja calidad de mapeo debido a la alta especificidad de la tecnología de captura.
- Para obtener la matriz de comparación entre muestras con los valores de profundidad de lectura por posición para el estudio sobre la cobertura del panel se empleó el módulo 'depth' de Samtools.

Pasos 6.- Llamada de SNVs e indels con Samtools y GATK.

$ samtools mpileup -l brca_regiones_captura.bed –B -D -S -m 3 -F 0.001 -f hg19.fasta muestra.bam > muestra.mpileup && VarScan mpileup2cns muestra.mpileup --min-coverage 3 --min-freq-for-hom 0.85 --min-var-freq 0.1 --p-value 0.99 --strand-filter 0 --variants --output-vcf 1 --vcf-sample-list nombre_muestra > muestra_varscan.vcf – Llamada de variantes con VarScan a partir del fichero mpileup generado por Samtools.

$ gatk_toolkit -T UnifiedGenotyper -R hg19.fasta -I muestra.bam -o muestra_gatk.vcf -glm BOTH -L brca_regiones_captura.bed --min_base_quality_score 0 -minIndelCnt 3 -stand_call_conf 0 -stand_emit_conf 0 -minIndelFrac 0 -sample_rename_mapping_file samplelist_gatk -dcov 1000

Paso 7.- Identificación de CNVs. Implementado como función dentro del script en Bash correspondiente al pipeline general (Paso 1). A partir de la matriz de comparación obtenida en el 'Paso 4' se llevó a cabo la normalización descrita en el apartado 3.2.6 de esta sección.

Paso 8.- Anotación de las variantes. Los ficheros de resultados de variantes se parsearon y combinaron con los VCFTools para generar un único fichero de resultado para cada muestra priorizando los resultados de GATK sobre Varscan. Este fichero en formato VCF fue el input para el script VEP de Ensembl modificado para la versión 71 de esta base de datos.

Los resultados del análisis bioinformático fueron posteriormente evaluados por los expertos en Genética Médica de Sistemas Genómicos tomando como mínimo variantes con una profundidad de lectura de 20x y una frecuencia alélicas de 0,2. Inicialmente, las variantes fueron confirmadas visualmente mediante los softwares AVA e IGV debido al problema del homopolímero inherente a esta tecnología. Posteriormente, la anotación biológica proporcionada por el pipeline así como otra serie de herramientas para el análisis clínico,

descritas en el apartado 2.3.8 de esta misma sección, fueron empleadas para la toma de decisión sobre la potencial patogenicidad de una determinada variante.

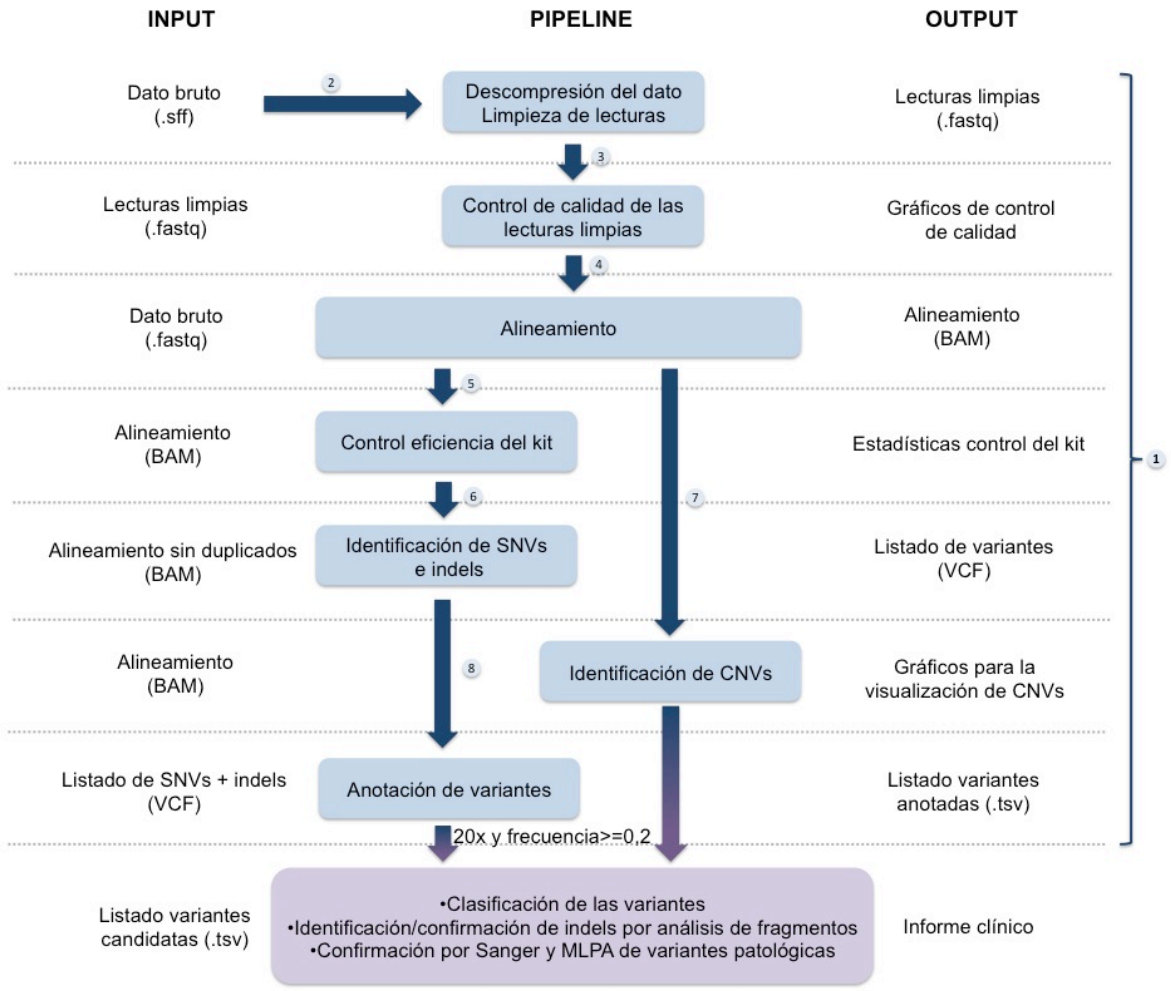


Figura 28. Pipeline aplicado al análisis de los genes BRCA para datos de las plataformas de secuenciación 454-Roche.

RESULTADOS

Los resultados de esta tesis se presentan en tres apartados diferentes. El primero de ellos está orientado al estudio del exoma completo, donde se contempla el diseño inicial del protocolo de análisis, su validación en una muestra controlada y su aplicación al estudio de enfermedades mendelianas raras. Este protocolo de trabajo evolucionó posteriormente con la introducción de nuevas herramientas de análisis más eficientes y sensibles para ser aplicado en muestras de resecuenciación dirigida con objetivos diagnósticos. Finalmente, una modificación de la estrategia de análisis desarrollada fue empleada para el estudio de los genes BRCA en mini-secuenciadores con tecnología NGS. Los distintos estudios planteados en esta tesis se analizaron siguiendo una estructura modular común tal como muestra la Figura 29. A continuación se describen los resultados obtenidos tras el análisis de los tres datasets mencionados.

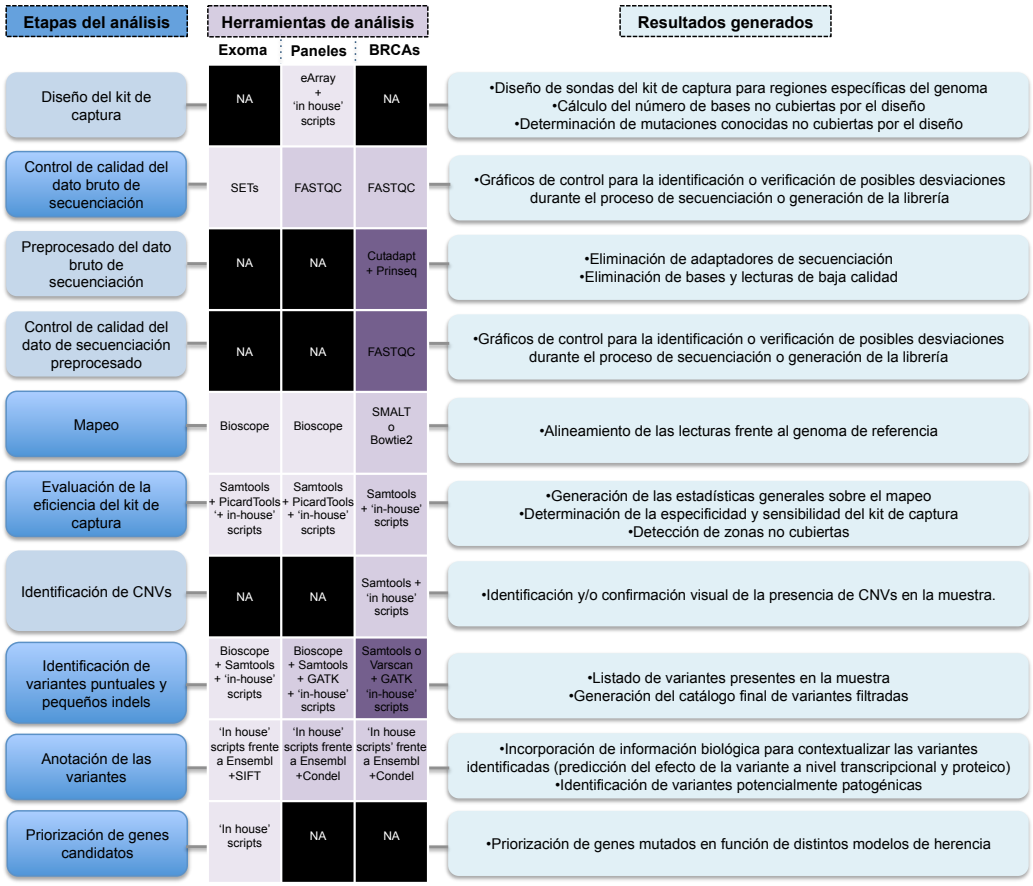


Figura 29. Visión global de los protocolos de análisis desarrollados para cada uno de los tres dataset.

1 Análisis del exoma completo y su aplicación en el estudio de enfermedades mendelianas

Las siguientes secciones detallan los resultados obtenidos tras el desarrollo de un protocolo de análisis específico para datos de secuenciación del exoma completo procedentes del estudio de una línea celular controlada. Posteriormente, este protocolo fue exitosamente aplicado en la identificación de un nuevo gen responsable de Anemia de Fanconi (FA), una grave enfermedad rara de carácter autosómico recesivo.

1.1 Desarrollo de un pipeline de análisis para el exoma completo

Para llevar a cabo la puesta a punto del pipeline de análisis para el exoma completo se partió de un material de referencia, una línea de HapMap previamente genotipada por el consorcio HapMap mediante arrays de genotipado que contaba con un total de 29.219 posiciones interrogadas dentro de las coordenadas cubiertas por las sondas del kit de captura para el exoma.

1.1.1 Evaluación de los datos de secuenciación y captura

En total, se generaron ~210 millones de lecturas emparejadas asimétricas, de 50pb y 25pb, en ¼ de placa de secuenciación (~7.88Gb). El análisis global de los valores de calidad obtenidos para cada uno de los dos ficheros generados tras la secuenciación, fichero con las lecturas en 'forward' o F3 y fichero con las lecturas en 'reverse' o F5, mostró una distribución de los valores de calidad normal según los estándares marcados en los manuales de usuario proporcionados por el fabricante para la secuenciación de muestras de DNA y la versión 4 de la plataforma SOLiD (Figura 30). Bajo estos estándares, más del 50% de las bases generadas deben tener una fiabilidad por encima de Phred 20, asumiendo por tanto 1 error por cada 100 bases. Por otro lado, la proporción de bases normal con valores de calidad más altos disminuye conforme aumentaban los ciclos de ligación de manera que cuanto más larga es la lectura mayor es la tasa de error en las últimas bases. Asimismo, y siguiendo los estándares proporcionados por el fabricante, los valores de calidad globales en las lecturas F5 resultaron más bajos en relación a los resultados obtenidos para las lecturas F3 pese a tener una menor longitud debido a las diferencias en la estrategia de secuenciación empleada en cada caso.

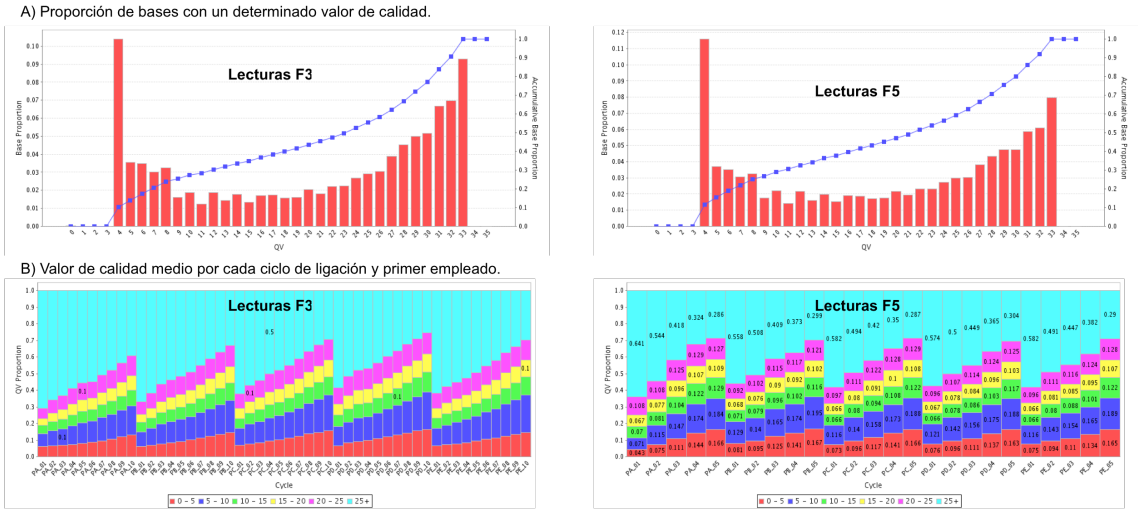


Figura 30. Resultados de la evaluación de la calidad de las bases generadas para la muestra control en ambos grupos de lecturas. A) Proporción de bases con un determinado valor de calidad. La línea azul representa la proporción acumulativa de bases respecto a los valores calidad. B) Proporción de bases correspondiente a cada uno de los seis grupos de valores de calidad, indicados al pie de la imagen, atendiendo al ciclo de ligación y al primer empleado. Así, los valores relacionados con PA_01 representan los resultados tras la secuenciación con el primer A en el ciclo de ligación 1.

Del total de lecturas generadas se obtuvo un porcentaje de mapeo del 68,75%, valor considerado normal según los estándares proporcionados por el fabricante de la plataforma de secuenciación. El porcentaje de duplicados de PCR fue del 26,9% de las lecturas mapeables. Tras la eliminación de estas lecturas, se comprobó la eficiencia del sistema de captura mediante el cálculo del porcentaje de lecturas en las zonas diana y del porcentaje de bases cubiertas a diferentes profundidades de lectura. El porcentaje total de lecturas dentro de las zonas diana fue del 51,01%, valor dentro de la normalidad según el fabricante del kit de captura [76]. El análisis de la sensibilidad del kit de captura a 1x, 10x y 20x fue de 97,8%, 85,17% y 73,78% respectivamente. La profundidad de lectura media alcanzada por base fue de ~54x mientras que la mediana resultó de 39x (Figura 31). El tamaño medio del inserto calculado para esta librería fue de 200,91pb, con una desviación estándar de 31, siendo el valor teórico para este parámetro de aproximadamente 200pb.

No se observaron por lo tanto desviaciones significativas en ninguno de los parámetros considerados.

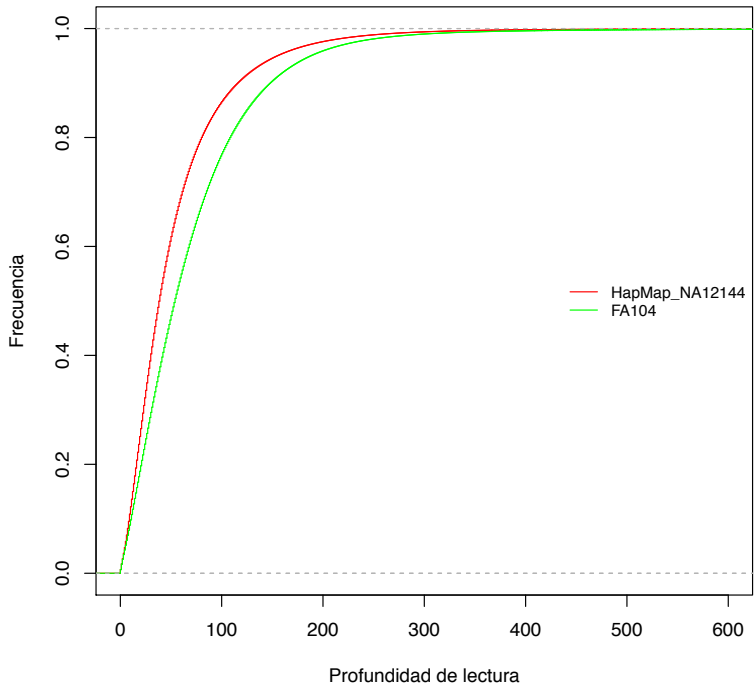


Figura 31. Proporción acumulativa de bases respecto a la profundidad de lectura para línea celular de HapMap NA12144 y la muestra de Fanconi FA104.

1.1.2 Definición de los parámetros de filtrado para la priorización de variantes

La identificación de variantes se basa en un equilibrio entre sensibilidad y especificidad y por tanto, es imprescindible determinar unos criterios mínimos de filtrado a fin de proporcionar un dataset inicial reducido que permita manipular la gran cantidad de datos generada. Debido a la gran divergencia entre programas de identificación de variantes, filtros elegidos y número de variantes reportadas (entre 14.000 y 24.000) para el mismo sistema de captura y plataforma de secuenciación entre las pocas publicaciones disponibles en el momento de iniciar este estudio [48, 144-147], se abordó el análisis de los parámetros más relevantes para la identificación de las variantes en base a la comparación frente a los datos de genotipado masivo por arrays de la línea celular usada como control. Así, inicialmente, se lanzó la llamada de variantes sin aplicar ningún filtro con el fin de obtener el mayor número de variantes posibles y poder calcular la sensibilidad y especificad 'bruta' de la tecnología. Las variantes fueron clasificadas en cuatro grupo: verdaderos positivos (VP), falsos positivos (FP), verdaderos negativos (VN) y falsos negativos (FN) (ver Métodos 1.2.6). En total, 27.924 posiciones de los arrays de genotipado se encontraban dentro de las regiones diana

y en zonas cubiertas por al menos una lectura. Para la definición de los parámetros de filtrado se estudió en cada uno de los cuatro grupos de variantes la distribución de tres parámetros reportados por el caller: la calidad del genotipo o GQ, el valor de calidad de la variante o SNVQ y la profundidad de lectura o DP (Figura 32). La distribución del valor de calidad de la variante mostró ser un parámetro de gran relevancia a la hora de discriminar entre VP y FP siendo el valor de la media de 103,70 y 10,52 respectivamente. Dado que el 75% de los FP tenían un valor de SNVQ de 12 o inferior y que el valor del primer cuartil para los VP en este valor era de 51, se tomó un valor de SNVQ de 20 (1 error por cada 100 variantes) para la priorización de las variantes. El valor medio de calidad del genotipo resultó ser del doble en el grupo de VP que en el grupo de VN y 9 veces mayor que en el grupo de FP. De la misma forma, el estudio de la distribución del valor de DP en el grupo de FP mostró que el 75% de las variantes tenían un DP de 4 o inferior. Dado que ambos valores están relacionados y con el objetivo de investigar de una forma global el comportamiento de estos dos parámetros unidos, se realizó un estudio con todas las combinaciones posibles de ambos asumiendo valores de GQ de 10, 20, 30, y valores de DP de 1x hasta 15x (Figura 33). Dada la importancia del valor de GQ para la correcta priorización de los genes en función de un modelo dominante o recesivo y teniendo en cuenta los datos observados en la Figura 32, se fijó un valor del genotipo de 30, ya que los resultados mostraban una menor dispersión en este valor. De la misma forma, se fijó un valor de 9x para filtrar las variantes en base a las pruebas realizadas donde el aumento de la sensibilidad, teniendo en cuenta las posiciones con un GQ de al menos 30, parecía estabilizarse.

Para la identificación de pequeños indels se empleó un software diferente al utilizado para la detección de SNVs a partir del cual se estableció un único parámetro de filtrado la profundidad de lectura, que por analogía a los resultados obtenidos en las variantes puntuales, y dado que no se disponía de ningún otro dataset de referencia validado con una tecnología transversal, se fijó en un 9x.

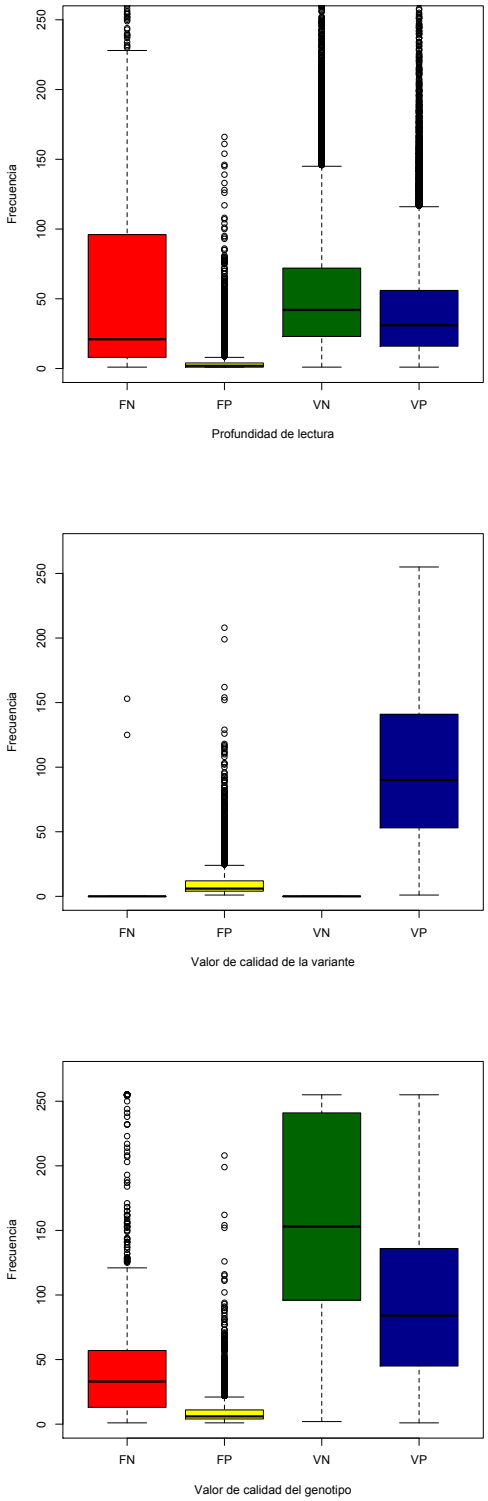


Figura 32. Distribución de los parámetros de filtrado en los cuatro grupos de variantes generados en la muestra control tras comparar con los resultados de genotipado masivo por arrays para esta misma línea celular.

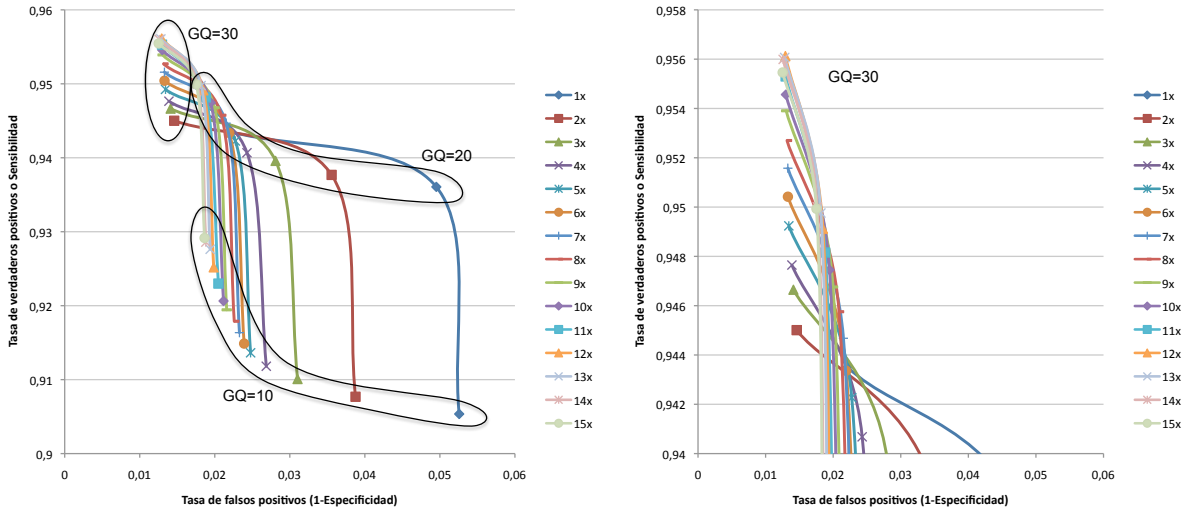


Figura 33. Selección de los parámetros de filtrado. Las líneas unen los resultados obtenidos a una misma profundidad de lectura empleando diferentes valores de GQ.

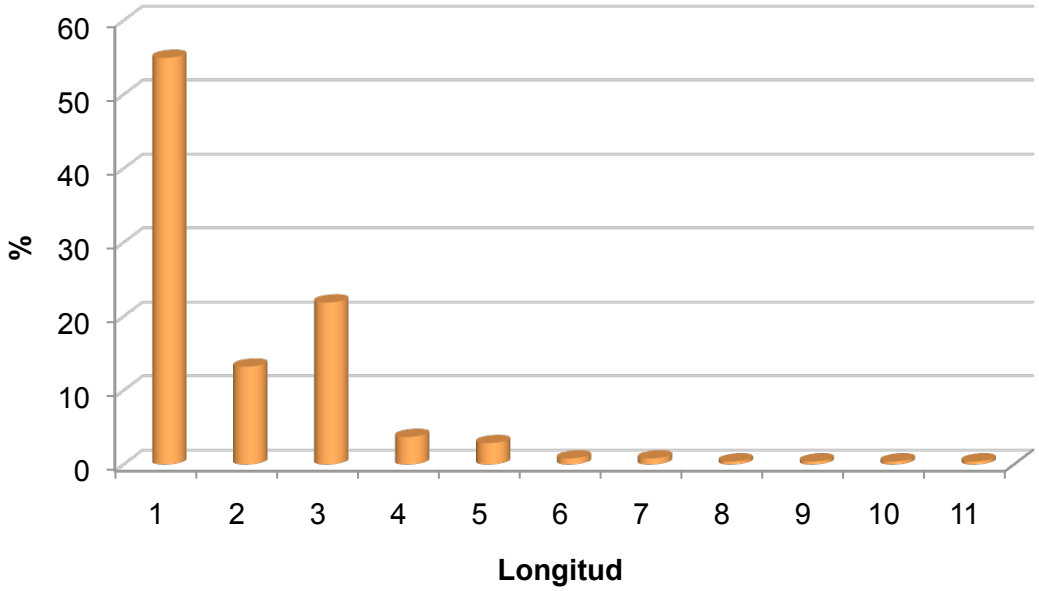


Figura 34. Distribución de la longitud de los indels identificados.

Se identificaron un total de 29.711 SNVs cuya sensibilidad y especificidad, calculados a partir de los datos de genotipado de arrays para el total de SNVs detectadas, fue del 91% y del 72,61% respectivamente. La aplicación de los criterios de priorización anteriormente expuestos (GQ=30, DP=9x, SNVQ=20), que cumplían el 89,46% de las bases diana, redujo

el número inicial de potenciales variantes a 18.804 SNVs. La sensibilidad y especificidad obtenidas en la detección de SNVs tras la aplicación de estos filtros fue de 95,39% y 98,69% respectivamente. En un 99,48% de las variantes reportadas, el genotipo fue llamado correctamente asegurando así una correcta clasificación de la mayoría de los genes candidatos en función del modelo de herencia. El 98,80% de los SNVs identificados se encontraron descritos en la base de datos de Ensembl, valores esperados al tratarse de una línea HapMap caracterizada previamente por diferentes consorcios internacionales.

Respecto a los indels, inicialmente se identificaron 993 indels de los cuales, tras aplicar el criterio de filtrado de 9x, se quedaron en 243. El 23,14% de los indels tenían un tamaño múltiplo de 3, no variando por tanto la pauta de lectura (Figura 34). De forma similar a los SNVs, el porcentaje de indels conocidos fue muy alto (95,89%) augurando así una baja proporción de falsos positivos en el dataset final.

1.2 Aplicación del exoma completo en el estudio de enfermedades raras. Identificación de un nuevo gen causante de Anemia de Fanconi (FANCQ)

El protocolo de análisis desarrollado se aplicó en una muestra de un paciente con anemia de Fanconi denominado FA104. El paciente FA104, nacido en el año 2000 e hijo de padres sin parentesco alguno, presentaba una serie de malformaciones (ausencia de pulgares, microsomía, atresia esofágica, ano anterior, hipotiroidismo y dismorfia e implantación baja de las orejas) que correspondía a un individuo con fenotipo FA. A los 2 años de edad, el paciente desarrolló una anemia aplásica que posteriormente desembocó en fallo de la médula ósea. El paciente fue sometido a un transplante medular a la edad de 4 años, tras el cual murió debido a un shock hemorrágico. El test de fragilidad cromosómica fue positivo, confirmándose así el diagnóstico de FA. Esta muestra había sido previamente estudiada mediante secuenciación tradicional Sanger para descartar los principales genes relacionados con esta enfermedad, estudio tras el cual el diagnóstico de este individuo seguía siendo una incógnita.

Se generaron un total de 7,80Gb de datos en ¼ de placa de SOLiD. El control de calidad del dato bruto no reportó desviación alguna (Anexo, Figura 1). El 74,25% de las lecturas generadas mapearon frente al genoma de referencia de las cuales el 63,16% se localizaron dentro de las zonas diana. La cantidad de datos final obtenida tras eliminar estas lecturas aseguró una profundidad de lectura media útil de aproximadamente 71x. Tanto la calidad del dato bruto como los diferentes parámetros de control calculados para la muestra se mantuvieron dentro de los valores normales según el fabricante y comparables a los datos obtenidos en la muestra de HapMap (Tabla 6).

Tabla 6. Comparación entre los resultados obtenidos para la línea de HapMap y la muestra de Fanconi FA104.

	HapMap	FA104
Gb generadas	7,88	7,80
% mapeables	68,75	74,25
% duplicados PCR	26,90	26,48
Tamaño del inserto (nt)	200,01	170,81
% lecturas 'on-target'	51,01	63,16
% bases cubiertas a 1x	97,45	98,78
% bases cubiertas a 10x	87,43	89,27
% bases cubiertas a 20x	73,52	82,46
% bases aplicando filtros (DP=9x y GQ=30)	89,46	90,01
Profundidad de lectura media	54	71
Variantes sin filtrar	30.704	33.720
Variantes filtradas (DP=9x,GQ=30,SNVQ=20 en SNVs y DP=9x en indels)	19.046	21.080
Sensibilidad	95,39	-
Especificidad	98,69	-
% SNVs conocidas	98,8	86,94
% Indels conocidos	95,89	67,38
Variantes potencialmente patogénicas noveles	5	509
Genes candidatos - modelo dominante	-	491
Genes candidatos - modelo recesivo	-	17

Se identificaron un total de 21.080 variantes de alta calidad, 20.663 SNVs y 417 indels. El 86,94% de los SNVs y el 67,38% de los indels habían sido previamente descritos, valores dentro de los estándares reportados por distintos autores para confirmar la mayor o menor presencia de falsos positivos [48, 49, 102, 145, 146]. Se consideraron variantes potencialmente patogénicas (PP) todas aquellas SNVs no descritas en la base de datos de Ensembl y cuya predicción a nivel transcripcional, en cualquiera de los transcritos conocidos para el gen, resultara en un cambio no sinónimo o se localizaran en zonas de splicing. Asimismo, se incluyeron en esta categoría aquellos indels presentes en regiones codificantes cuyo efecto a nivel transcripcional resultara en un cambio en la pauta de lectura, indels complejos e indels localizados en zonas de splicing. Se descartaron como variantes causales aquellas que habían sido previamente descritas en alguna de las grandes bases de datos recogidas en Ensembl, obteniéndose un total de 509 variantes. Dada la baja frecuencia de aparición de esta enfermedad, inicialmente se tomaron en consideración solamente los genes candidatos pertenecientes a un modelo de herencia recesivo, en total 17 (Anexo. Tabla 1). Debido a la naturaleza de esta enfermedad, uno de estos 17 genes candidatos, el gen ERCC4/XPF, dada su función en la reparación del DNA, se presentó como principal gen candidato. Las dos mutaciones encontradas en este gen, una deleción de 5pb gen en el exon 8 (c.1484_1488delCTCAA) y una mutación no sinónima en el exon 11

(c.2065C>A) fueron estudiadas en profundidad para demostrar su patogenicidad (Tabla 7). La deleción provocaba, además de un cambio en la pauta de lectura, un codón de parada prematuro (p.Thr495Asnfs*6).La otra mutación identificada en el gen ERCC4, producía un cambio no sinónimo de una arginina por una serina (p.Arg689Ser) dentro del sitio activo de la endonucleasa. Los valores de conservación GERP asociados a cada base mutada resultaron muy elevados. Las dos mutaciones fueron confirmadas mediante secuenciación Sanger a partir de sangre.

Tabla 7. Información obtenida para las dos variantes patogénicas identificadas en el individuo FA104.

Cromosoma	16	16
Inicio	14029271	14041518
Final	14029275	14041518
Tamaño	5	1
Referencia	AACTC	C
Muestra	-	M
Tipo de variante	deletion	SNV
ID del GEN	ENSG00000175595	ENSG00000175595
ID del trascrito	ENST00000311895	ENST00000311895
Efecto de la variante	FRAMESHIFT_CODING	NON_SYNONYMOUS_CODING
Cambio aminoacídico	-	R689S
ID de la variante	-	-
Predicción SIFT	N/A	DAMAGING
Score SIFT	N/A	0
Nombre del gen	ERCC4	ERCC4
Descripción del gen	DNA repair endonuclease XPF	DNA repair endonuclease XPF
Genotipos identificados	TTAACTCAAAT/TTAAAT	N/A
Distribución de las lecturas	REF,25	N/A
Lecturas no redundantes	22	121
SNVQ	N/A	228
GQ	N/A	228
Valor de conservación GERP	-1,0201;3,2201;4,3501;-4,0501;4,3501;	3,75
Gen-enfermedad (OMIM)	XFE PROGEROID SYNDROME; XERODERMA PIGMENTOSUM, COMPLEMENTATION GROUP F	XFE PROGEROID SYNDROME; XERODERMA PIGMENTOSUM, COMPLEMENTATION GROUP F

Los resultados obtenidos mediante la secuenciación del exoma completo se apoyaron en estudios funcionales adicionales y en la identificación de mutaciones causales en este mismo gen en otro individuo alemán con FA durante los siguientes dos años tras los que quedó demostrado que dependiendo de la localización de las mutaciones en el gen, además

de las dos enfermedades ya relacionadas con ERCC4 Xeroderma pigmentoso [MIM:278760] y Progeria [MIM:610965], los individuos podían desarrollar Anemia de Fanconi. Este nuevo gen recibió un nuevo alias aprobado por la HGNC (HUGO Gene Nomenclature Committee), FANCQ, siendo el 16º gen causante de esta enfermedad. Los resultados de este estudio han sido publicados en la revista American Journal of Human Genetics, bajo el título "Mutations of the DNA repair endonuclease ERCC4/XPF cause Fanconi anemia" [148].

2 Análisis de paneles de genes orientado al diagnóstico genético en enfermedades cardiovasculares

Las siguientes secciones muestran el desarrollo y validación de un panel de 72 genes dirigido hacia el diagnóstico de enfermedades cardiovasculares. El análisis de las muestras se realizó aplicando un pipeline de análisis diseñado a partir del protocolo empleado para muestras de exoma completo descrito en la sección anterior. A diferencia del pipeline para muestras de exoma, en este protocolo se incluyó nuevo software para analizar en mayor profundidad el diseño del kit de captura así como nuevos paquetes de análisis para la minimización del número de falsos positivos y negativos en la llamada de variantes. Adicionalmente, se incorporaron nuevos recursos biológicos en la anotación de las variantes para facilitar el diagnóstico.

2.1 Evaluación del diseño del kit de captura

El diseño realizado para la captura de los 72 genes de interés constó de 46.081 sondas de 120pb de longitud que cubrían un total de 770.165nt. El 94,41% del total de bases diana fue capturado por al menos 1 sonda tras realizar el diseño (Tabla 8). El 85,74% de las zonas no cubiertas por el diseño correspondían a zonas UTR donde es común encontrar elementos repetitivos del genoma como SINEs o LINEs [149] que son enmascarados automáticamente por el software de diseño y que por tanto no son incluidos en el kit de captura (Figura 35). Adicionalmente, se calcularon las mutaciones descritas en la base de datos HGMD correspondientes a las zonas diana. El total de variantes patogénicas y de susceptibilidad incluidas en las regiones diana fue de 4.233 de las cuales tan solo 78 (1,84% del total) se localizaron en zonas no cubiertas por las sondas diseñadas. Cabe destacar el alto porcentaje de mutaciones descritas en la base de datos HGMD no cubiertas en el gen MEF2A, mutaciones que se concentraron en una zona de 12pb perteneciente a una zona de mayor tamaño (144nt) marcada como elemento repetitivo en el genoma por el programa RepeatMasker [150] o TandemRepeatFinder [151].

Tabla 8. Porcentaje de bases no capturadas en los genes diana y variantes no cubiertas tras el diseño de sondas.

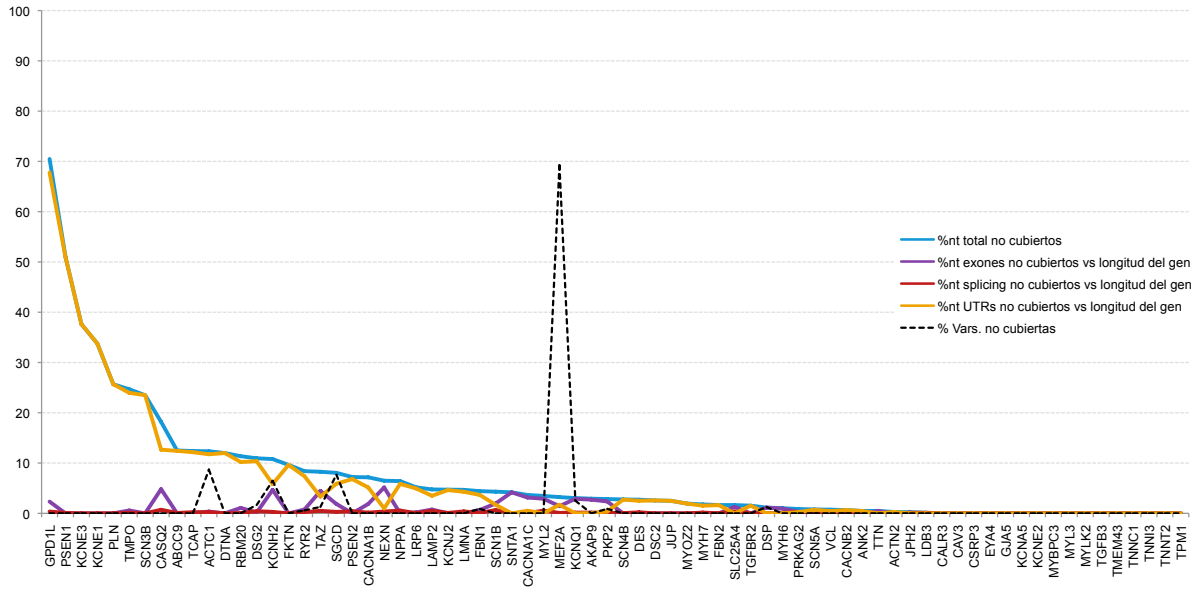
Nombre del gen	nt totales	%nt total no cubiertos	nt exones	%nt exones no cubiertos	nt splicing	%nt splicing no cubiertos	nt UTRs	%nt UTRs no cubiertos	Vars. en HGMD	% Vars.en HGMD no cubiertas
GPD1L	4160	70,48	1056	9,19	100	15,00	3004	93,87	4	0,00
PSEN1	6218	51,03	1404	0,00	140	3,57	4674	67,78	212	0,00
KCNE3	3193	37,61	312	0,00	50	0,00	2831	42,42	6	0,00
KCNE1	3204	33,77	390	0,00	50	0,00	2764	39,15	0	0,00
PLN	2041	25,67	159	0,00	40	0,00	1842	28,45	5	0,00
TMPO	3675	24,68	2085	1,01	60	8,33	1530	57,58	1	0,00
SCN3B	5755	23,49	648	0,00	90	0,00	5017	26,95	0	0,00
CASQ2	2804	18,19	1200	11,25	130	15,38	1474	24,08	16	0,00
ABCC9	6252	12,51	4650	0,00	400	1,25	1202	64,64	17	0,00
TCAP	2163	12,39	504	0,00	40	12,50	1619	16,24	0	0,00
ACTC1	4196	12,35	1134	1,32	90	11,11	2972	16,59	23	8,70
DTNA	4704	11,99	1704	0,00	186	0,00	2814	20,04	1	0,00
RBM20	7393	11,36	3684	2,17	160	3,13	3549	21,27	17	0,00
DSG2	6001	10,95	3357	0,30	170	14,71	2474	25,14	60	1,67
KCNH2	4456	10,79	3480	5,98	170	7,65	806	32,26	619	6,46
FKTN	7484	9,63	1386	0,00	120	0,00	5978	12,06	0	0,00
RYR2	17632	8,39	14904	0,90	1070	4,21	1658	78,41	168	0,60
TAZ	2020	8,27	879	10,24	130	7,69	1011	6,63	78	1,28
SGCD	1714	8,05	873	3,67	108	4,63	733	13,78	13	7,69
PSEN2	2452	7,22	1347	0,00	150	6,67	955	17,49	25	0,00
CACNA1B	10286	7,17	4602	4,17	490	3,06	5194	10,20	0	0,00
NEXN	2757	6,49	2028	7,00	150	6,67	579	4,66	5	0,00
NPPA	905	6,41	456	0,00	50	10,00	399	13,28	6	0,00
LRP6	10270	5,18	4842	0,39	250	2,00	5178	9,81	3	0,00
LAMP2	1977	4,75	1233	1,22	110	9,09	634	10,88	0	0,00
KCNJ2	5432	4,71	1284	0,00	40	12,50	4108	6,11	58	0,00
LMNA	2611	4,67	1719	0,00	150	6,67	742	15,09	0	0,00
FBN1	12436	4,41	8616	0,96	680	1,47	3140	14,52	1248	0,88
SCN1B	1531	4,31	657	4,57	80	12,50	794	3,27	9	0,00
SNTA1	2442	4,18	1518	6,72	100	0,00	824	0,00	5	0,00
CACNA1C	8915	3,66	6417	4,36	490	0,00	2008	2,29	12	0,00
MYL2	895	3,46	501	5,19	88	5,68	306	0,00	14	0,00
MEF2A	5954	3,22	1500	5,67	130	3,85	4324	2,36	13	69,23
KCNQ1	3425	3,04	2031	4,87	180	2,78	1214	0,00	398	2,51
AKAP9	12991	2,92	11724	3,01	520	3,85	747	0,80	2	0,00
PKP2	4401	2,84	2646	3,93	160	3,13	1595	1,00	112	0,89
SCN4B	4554	2,77	687	0,00	70	2,86	3797	3,27	2	0,00
DES	2358	2,71	1413	0,00	110	4,55	835	7,07	62	0,00
DSC2	5304	2,56	2706	0,00	180	0,00	2418	5,62	31	0,00
JUP	3370	2,49	2238	0,00	170	0,00	962	8,73	0	0,00
MYOZ2	2677	1,94	795	0,00	80	0,00	1802	2,89	2	0,00
MYH7	6507	1,77	5808	0,10	420	2,38	279	35,48	408	0,00

(Continúa)

Nombre del gen	nt totales	%nt total no cubiertos	nt exones	%nt exones no cubiertos	nt splicing	%nt splicing no cubiertos	nt UTRs	%nt UTRs no cubiertos	Vars. en HGMD	% Vars.en HGMD no cubiertas
FBN2	11394	1,65	8739	0,00	670	0,30	1985	9,37	32	0,00
SLC25A4	1398	1,65	897	2,01	60	8,33	441	0,00	6	0,00
TGFBR2	4711	1,53	1704	0,00	90	0,00	2917	2,47	85	0,00
DSP	10056	1,14	8616	1,33	260	0,00	1180	0,00	81	1,23
MYH6	6271	1,02	5820	1,01	400	1,25	51	0,00	0	0,00
PRKAG2	3485	0,86	1710	0,64	180	5,56	1595	0,56	13	0,00
SCN5A	8756	0,78	6051	0,00	300	1,67	2405	2,62	0	0,00
VCL	5722	0,73	3405	0,35	240	2,08	2077	1,20	3	0,00
CACNB2	3546	0,65	1821	0,00	150	0,00	1575	1,46	12	0,00
ANK2	14676	0,48	11874	0,00	480	0,83	2322	2,89	17	0,00
TTN	104658	0,48	100272	0,46	3140	1,27	1246	0,00	35	0,00
ACTN2	5093	0,27	2685	0,00	230	0,00	2178	0,64	8	0,00
JPH2	4867	0,23	2091	0,24	80	7,50	2696	0,00	4	0,00
LDB3	1698	0,12	852	0,00	100	2,00	746	0,00	6	0,00
CALR3	1402	0,00	1155	0,00	110	0,00	137	0,00	2	0,00
CAV3	1471	0,00	456	0,00	40	0,00	975	0,00	38	0,00
CSRP3	1421	0,00	585	0,00	80	0,00	756	0,00	0	0,00
EYA4	5919	0,00	1920	0,00	220	0,00	3779	0,00	5	0,00
GJA5	3217	0,00	1077	0,00	40	0,00	2100	0,00	6	0,00
KCNA5	2895	0,00	1842	0,00	30	0,00	1023	0,00	11	0,00
KCNE2	843	0,00	372	0,00	40	0,00	431	0,00	19	0,00
MYBPC3	4564	0,00	3822	0,00	350	0,00	392	0,00	0	0,00
MYL3	1014	0,00	588	0,00	90	0,00	336	0,00	12	0,00
MYLK2	2937	0,00	1791	0,00	150	0,00	996	0,00	2	0,00
TGFB3	2612	0,00	1239	0,00	90	0,00	1283	0,00	0	0,00
TMEM43	3481	0,00	1203	0,00	140	0,00	2138	0,00	4	0,00
TNNC1	794	0,00	486	0,00	80	0,00	228	0,00	14	0,00
TNNI3	940	0,00	633	0,00	100	0,00	207	0,00	63	0,00
TNNT2	1302	0,00	867	0,00	180	0,00	255	0,00	71	0,00
TPM1	1168	0,00	855	0,00	119	0,00	194	0,00	29	0,00

“Nombre del gen”: símbolo HGNC del gen. “pb totales”: nucleótidos totales incluidos en el diseño para ese gen. “%pb total no cubiertos”: porcentaje de nucleótidos totales no cubiertos. “pb exones”: nucleótidos incluidos en el diseño correspondientes a zonas exónicas excluyendo las zonas UTR. “%pb exones no cubiertos”: porcentaje de bases en zonas exónicas excluyendo las regiones UTR no cubiertas por el diseño respecto al total de nucleótidos pertenecientes a zonas de exones excepto zonas UTR. “pb splicing”: nucleótidos correspondientes a zonas de splicing (5pb hacia dentro del intron). “%pb splicing no cubiertos”: porcentaje de nucleótidos en zonas de splicing (5pb hacia dentro del intron) no cubiertos por el diseño respecto al total de nucleótidos en zonas de splicing. “pb UTRs”: nucleótidos en zonas UTR incluidos en el diseño. “%pb UTRs no cubiertos”: porcentaje de nucleótidos no cubiertos por las sondas del diseño en zonas UTR respecto al total de nucleótidos correspondientes a las zonas UTR. “Vars. en HGMD”: variantes de la base de datos HGMD localizadas en los genes del panel. “% Vars. HGMD no cubiertas”: porcentaje de las variantes incluidas en la base de datos HGMD no cubiertas tras realizar el diseño.

Figura 35. Distribución de las bases no cubiertas por las sondas diseñadas para cada uno de los genes diana y su repercusión en cuanto el porcentaje de variantes potencialmente patogénicas descritas en la base de datos HGMD no detectadas. Los genes se encuentran ordenados atendiendo al porcentaje de zonas no cubiertas.



"%nt total no cubiertos": porcentaje de nucleótidos no cubiertos del total de bases diana para ese gen. *"%nt exones no cubiertos vs %nt total no cubiertos"*: porcentaje de bases en zonas exónicas excluyendo las regiones UTR que no se cubren respecto al porcentaje total de bases no cubiertas. *"%nt splicing no cubiertos vs %nt total no cubiertos"*: porcentaje de nucleótidos en zonas de splicing (5nt hacia dentro del intron) no cubiertos respecto al porcentaje total de bases no cubiertas. *"%nt UTRs no cubiertos vs %nt total no cubiertos"*: porcentaje de nucleótidos no cubiertos en zonas UTR respecto al porcentaje total de bases no cubiertas.

2.2 Análisis de la estabilidad del panel. Comparación frente al exoma completo

El control de calidad de las muestras mostró un descenso progresivo de la calidad de secuenciación llegando a situarse en valores medios por debajo de un Phred 20 en los últimos 10 nucleótidos de las lecturas F5 (Anexo, Figuras 2-9). Debido a esta desviación en el proceso de secuenciación, se eliminaron los 10 últimos nucleótidos de todas las lecturas F5 incrementándose así el valor medio de calidad por lectura (Figura 36).

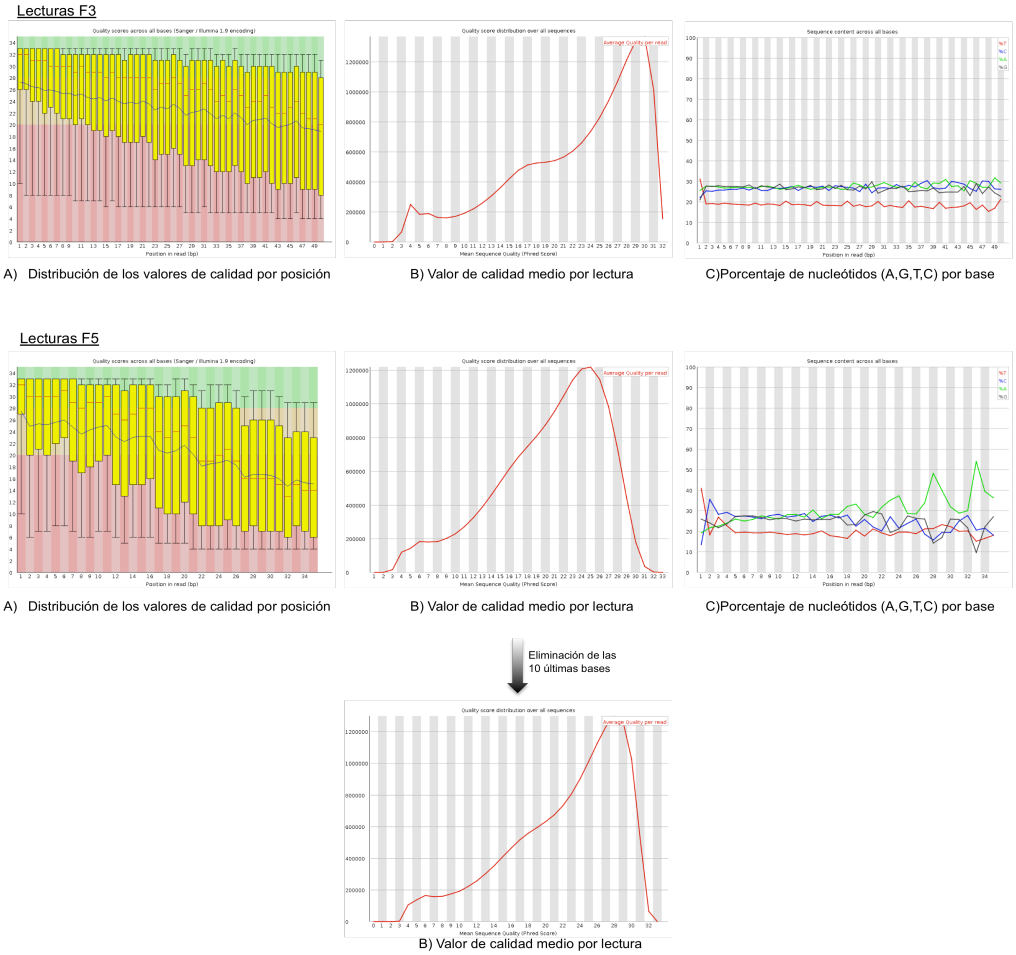


Figura 36. Resultados del control de calidad de los datos brutos para la muestra BM3062. Tras la eliminación de los 10 últimos nucleótidos en las lecturas F5 el valor de calidad medio por lectura mejoró de forma notable.

El porcentaje medio de lecturas mapeables se situó en torno al 77,91% de las cuales, un 69,90% de media, correspondían a lecturas dentro de las zonas diana, valores dentro de los parámetros estándar para la plataforma de secuenciación empleada y el kit de captura se. El porcentaje de lecturas duplicadas fue muy alto debido a la excesiva cantidad de información generada por muestra (Tabla 9). La profundidad media de lectura a lo largo del panel, una vez eliminadas las lecturas duplicadas y las lecturas con una baja calidad en el mapeo, fue de un promedio de 197x (Figura 37), valor muy superior al 50x alcanzado por la muestra de exoma de HapMap analizada en la sección anterior, valor medio ampliamente extendido en el estudio de exoma que deja sin cubrir un porcentaje de bases no despreciable y a tener en cuenta si se pretende emplear esta aproximación como método de diagnóstico de rutina [48, 144-147]. Prácticamente el 95% de las bases se encontraron cubiertas a valores de profundidad de lectura de 50x o superiores minimizándose por tanto el error relacionado con

la detección de variantes en zonas con una baja profundidad de lectura. El porcentaje medio de bases diana no cubiertas a 20x, valor mínimo recomendado por algunos autores para el filtrado inicial de variantes [152], fue del 5,56%, porcentaje muy similar al obtenido teniendo en cuenta solamente el diseño de sondas para la captura de los genes. El porcentaje de bases diana no cubiertas teniendo en cuenta estos resultados para todas las muestras fue del 6,42%. Por otro lado, se calculó la reproducibilidad entre muestras en base a los datos de profundidad de lectura por base. Los valores calculados para todas las combinaciones posibles entre muestras fueron superiores al 0,91 confirmándose así una alta tasa de reproducibilidad del sistema de captura (Tabla 10).

Tabla 9. Resumen de los datos de secuenciación y captura.

Muestra	Bases generadas (Gb)	Lecturas mapeables (%)	Profundidad de lectura media inicial	Lecturas duplicadas (%)	Lecturas con baja calidad (%)	Sensibilidad de la captura a 20x (%)	Especificidad de la captura (%)	Profundidad media final
BM3062	1,21	78,27	1225	72,82	74,82	96,04	68,69	206
BM3339	1,25	77,66	1264	76,58	78,09	96,02	72,58	196
BM3895	1,42	75,90	1398	79,03	80,45	96,27	73,09	191
BM4237	1,11	79,06	1143	70,43	72,75	95,93	64,88	197
BM5307	1,20	76,32	1186	74,56	76,17	96,13	74,37	204
BM5357	1,17	77,30	1171	73,77	75,77	96,01	67,63	186
BM6091	1,04	78,62	1066	71,15	72,65	96,15	75,70	218
BM6092	1,13	78,38	1151	69,00	71,89	95,76	61,50	193
BM6492	0,94	79,00	962	69,47	71,71	95,37	67,08	178
BM6919	1,26	78,55	1287	76,59	78,11	96,11	73,49	202

Tabla 10. Reproducibilidad del sistema de captura atendiendo a la profundidad de lectura por posición.

	BM3062	BM3339	BM3895	BM4237	BM5307	BM5357	BM6091	BM6092	BM6492	BM6919
BM3062	1									
BM3339	0,963	1								
BM3895	0,941	0,943	1							
BM4237	0,962	0,956	0,932	1						
BM5307	0,947	0,937	0,945	0,944	1					
BM5357	0,951	0,944	0,950	0,949	0,962	1				
BM6091	0,965	0,959	0,941	0,962	0,956	0,955	1			
BM6092	0,951	0,945	0,937	0,954	0,955	0,959	0,949	1		
BM6492	0,950	0,942	0,915	0,949	0,933	0,937	0,947	0,941	1	
BM6919	0,954	0,952	0,956	0,948	0,957	0,961	0,952	0,956	0,939	1

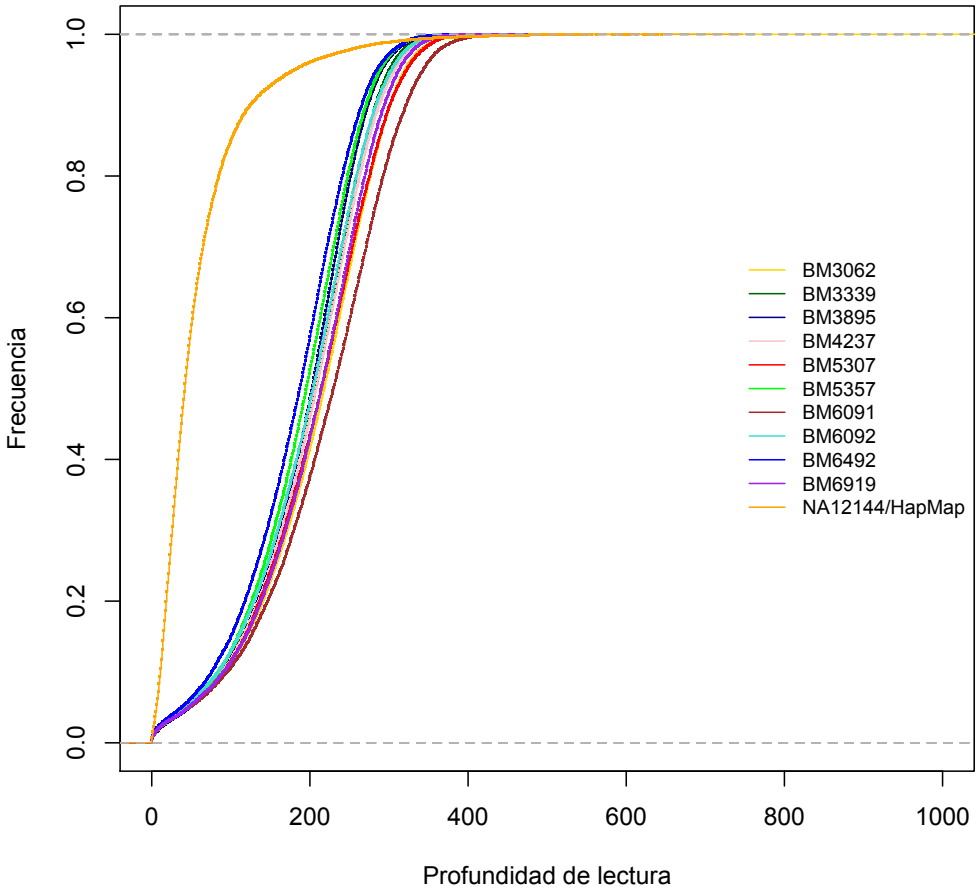


Figura 37. Distribución de la profundidad de lectura (proporción acumulativa) para el panel de enfermedades cardiacas y para el exoma de la línea HapMap en las mismas coordenadas que este panel.

2.3 Optimización de la llamada de variantes

La publicación de nuevos algoritmos para la llamada de variantes hizo necesaria la actualización del pipeline unos pocos meses después de la generación del primer protocolo de análisis desarrollado inicialmente en exoma. Así, las nuevas herramientas de llamada de variantes incluidas para el estudio de paneles de genes se testaron de nuevo en la muestra de exoma de HapMap, analizada en el apartado 1 de esta misma sección, para evaluar la sensibilidad y especificidad del nuevo pipeline de llamada de variantes y posteriormente fueron aplicadas en las muestras del panel.

Debido al descenso de calidad observado en las lecturas F5 en relación a las lecturas F3, se investigó la posibilidad de tratar la muestra teniendo en cuenta tanto las lecturas F3 como F5 (secuenciación tipo 'paired-end' o PE), tal como se había realizado en un inicio, o tratando la muestra como si se hubiera secuenciado siguiendo un protocolo de 'single-end' o SE mediante el cual solo se habrían obtenido lecturas F3. Asímismo, se evaluó la sensibilidad de un nuevo software el Genome Analyzer ToolKit o GATK tanto para llamada de SNVs como de pequeñas inserciones y deleciones.

En todas las comparaciones realizadas, se identificó un grupo de variantes que solo pudo ser detectado bien mediante SE o bien mediante PE para los tres softwares testados mostrando un cierto sesgo en los resultados debido al aumento del error de secuenciación en las lecturas reversas o F5 y a las diferencias en la funcionalidad de los programas empleados para la llamada de indels (Figuras 38). Así, en la detección de SNVs, el análisis del set de datos tratándolo como SE proporcionó un 1,8% y un 1,31% adicional de VP, con Samtools y GATK respectivamente. La comparación entre Samtools y GATK, incluyendo los resultados obtenidos con PE y SE, mostró una concordancia en la identificación de VP entre ambas herramientas del 99,31% mientras que el 0,34% y el 0,35% del total de VP fueron detectadas únicamente con Samtools o GATK, respectivamente.

La combinación de Samtools y GATK, mostró un aumento de la sensibilidad en la detección de SNVs respecto al pipeline inicial de 2,14% (Figura 39). En contra, esta nueva aproximación produjo un descenso de la especificidad de 4,16% respecto al análisis realizado inicialmente para el exoma. Las dos herramientas identificaron en común el 92,35% de las variantes totales mientras que Samtools identificó de forma exclusiva un 1,05% y GATK un 6,60%.

A diferencia de los SNVs, en el estudio de indels, se detectó un mayor porcentaje de variantes exclusivas del dataset tratado como SE que en el dataset tratado como PE, 25,68% en Bioscope y del 35,10% en GATK, debido a las diferencias en el funcionamiento de los algoritmos de llamada de indels para PE o SE tal como se especifica en sus manuales de usuario (Figura 40). En comparación con el pipeline inicial, la distribución de la longitud de los indels identificados fue muy similar (Figura 41). Para estimar la mayor o menor presencia de falsos positivos en la llamada de indels con las diferentes aproximaciones se calculó el porcentaje de indels conocidos. Los porcentajes de indels conocidos tanto tratando el dataset como SE o como PE fueron muy similares siendo de ~95% para Samtools y de ~98% para GATK, valores que confirmaban una baja presencia de potenciales falsos positivos en los resultados finales. La combinación de GATK y Bioscope

dio lugar a un incremento notable en el número de indels detectados de los cuales un 16,70% se detectaron únicamente con Bioscope y un 19,65% con GATK, siendo por tanto el porcentaje de concordancia entre ambos de 63,65%. Dado el alto porcentaje de indels conocidos identificados tanto con GATK como con Bioscope, los resultados demuestran la notable mejora que supone la inclusión de una segunda pieza de software para la llamada de indels que incluye, como paso previo a la llamada de indels, un algoritmo de realineamiento alrededor de las zonas donde existen indels.

En total se detectaron 20,887 SNVs y 407 pequeños indels de los cuales, un 95,28% de los SNVs y un 94,34% de los indels eran conocidos resultando estos porcentajes un 3,52% y un 1,21% inferiores al porcentaje de SNVs e indels descritos con el pipeline inicial respectivamente. Pese al ligero aumento en los potenciales falsos positivos, el 9,97% de los SNVs y un 34,89% de las pequeñas inserciones y deleciones identificadas con este segundo pipeline no se habían detectado con el pipeline inicial desarrollado para datos de TargetSeq suponiendo por tanto la aplicación de este segundo pipeline, una mejora importante para la identificación de SNVs y en mayor medida para la identificación de indels.

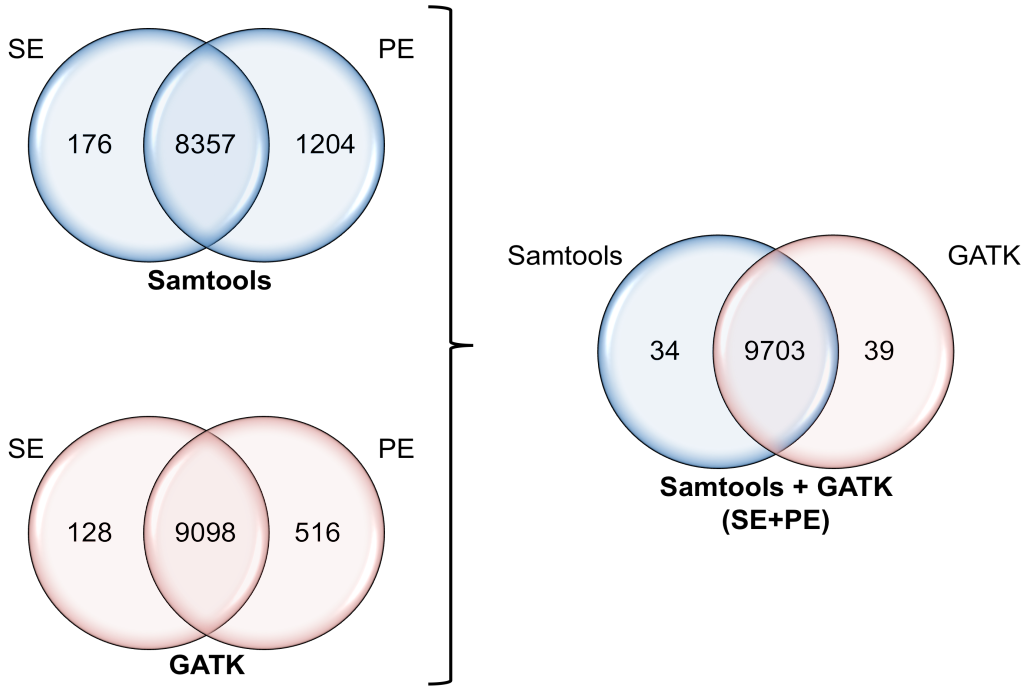


Figura 38. Comparación entre el número de verdaderos positivos llamados por cada una de las aproximaciones. SE=dataset tratado como Single-End; PE=dataset tratado como Paired-End

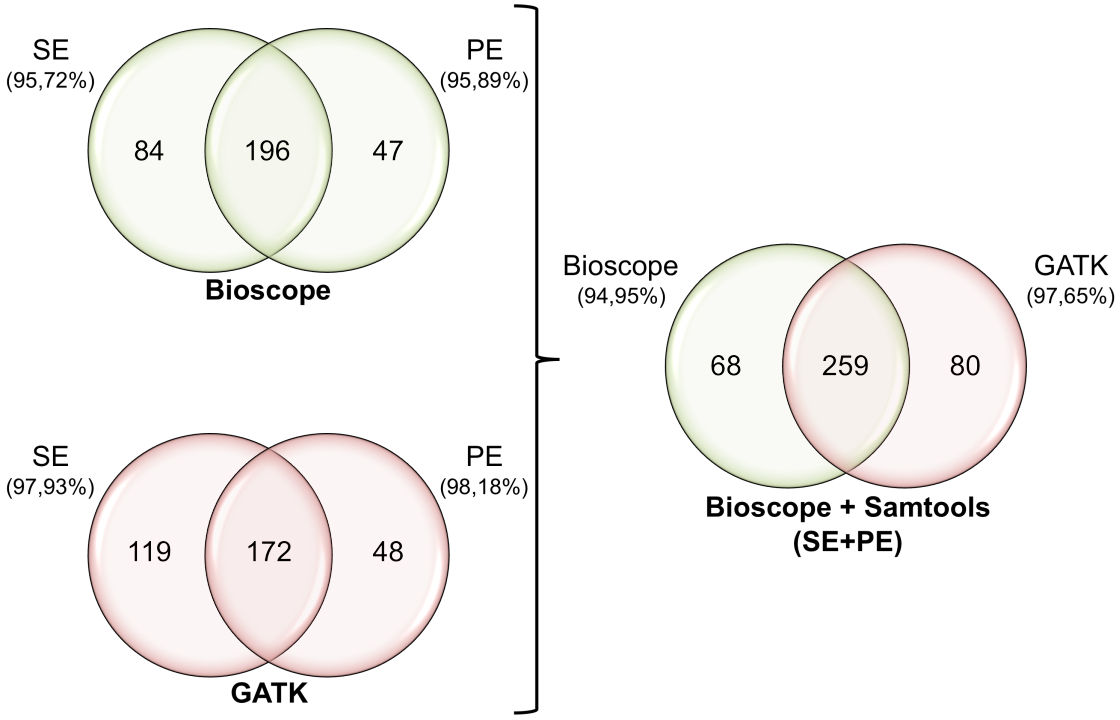


Figura 39. Indels identificados mediante las diferentes aproximaciones. El porcentaje de indels conocido se muestra entre paréntesis. SE=dataset tratado como Single-End; PE=dataset tratado como Paired-End.

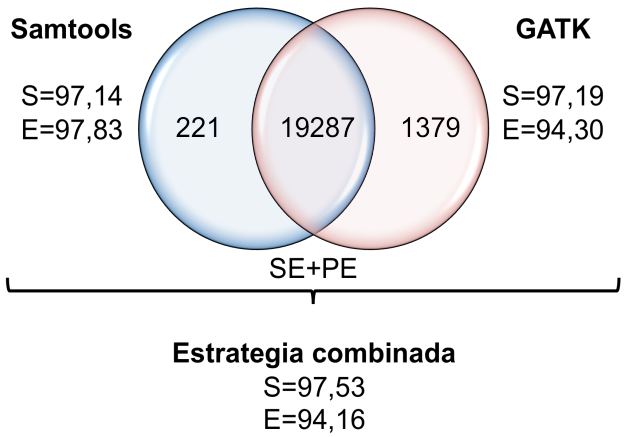


Figura 40. SNVs identificados en la muestra de exoma. S=Sensibilidad. E=Especificidad.

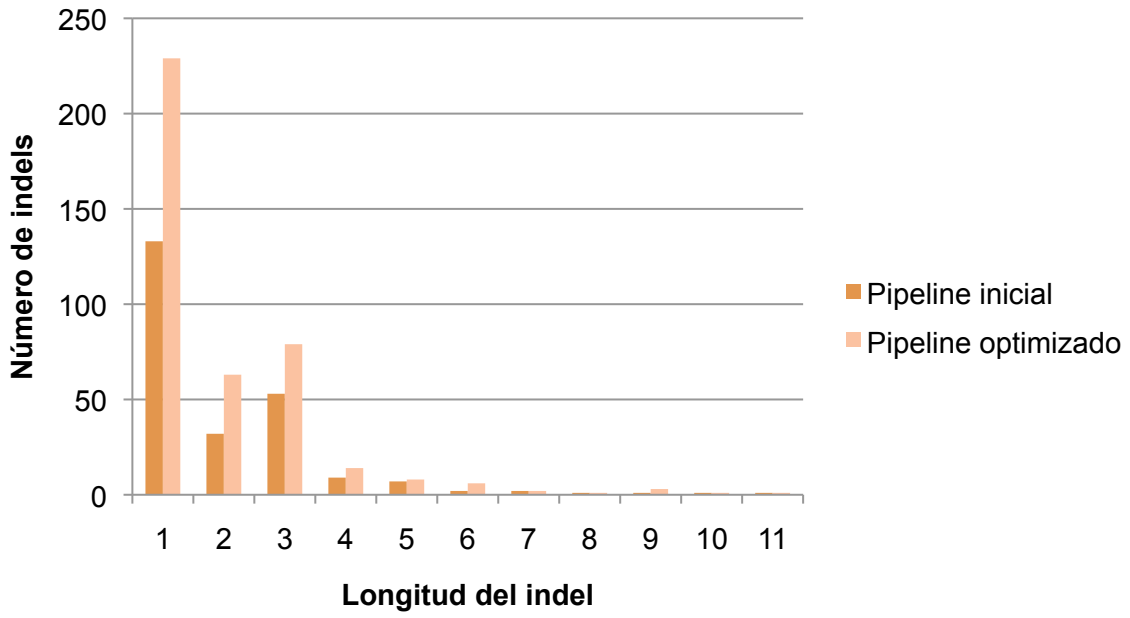


Figura 41. Distribución de la longitud de los indels identificados en el pipeline inicial donde se incluyó solamente Bioscope en comparación con el pipeline optimizado donde se empleó una combinación de Bioscope y GATK.

2.4 Identificación de variantes en las muestras control. Sensibilidad del panel

La nueva estrategia para la llamada de variantes aplicada en las 10 muestras control secuenciadas con el panel de 72 genes reportó una media de 683 SNVs y 48 indels por muestra (Tablas 11 y 12). Aproximadamente, en la detección de SNVs, Samtools reportó un 9,80% de variantes únicamente identificadas por este algoritmo mientras que GATK detectó de manera exclusiva un 13,33% del total de variantes. El porcentaje medio de variantes conocidas fue de 93,02%. En el caso de los indels, como valor promedio, el 17,57% del total fue identificado únicamente por Bioscope mientras que el 27,77% solo fue detectado por GATK. Aproximadamente, el 85,18% de los indels era conocido. Los porcentajes de variantes conocidas se mantuvieron dentro de los valores reportados por otros autores [48, 49, 102, 145, 146].

En el caso de las mutaciones patogénicas, de los 6 SNVs solamente uno de ellos fue identificado por Samtools y no por GATK. En el caso de los indels, todas las variantes fueron identificadas por GATK, sin embargo, dos de ellas fueron identificadas siguiendo la estrategia para SE pero no con PE. Por otro lado, Bioscope falló en la detección de 1 de los 4 indels. Tras la combinación de los diferentes resultados, la sensibilidad alcanzada por el nuevo pipeline desarrollado fue del 100% (Tabla 13).

Tabla 11. Identificación de SNVs en las muestras control.

Muestra	SNVs				
	Total	Solo detectado con Samtools	Solo detectado con GATK	Ambos	% conocidos
BM3062	707	68	91	548	94,06
BM3339	789	64	105	620	93,92
BM3895	672	59	92	521	90,77
BM4237	624	59	88	477	93,43
BM5307	635	58	103	474	91,97
BM5357	697	66	78	553	94,55
BM6091	690	70	101	519	92,61
BM6092	713	68	85	560	94,67
BM6492	631	63	80	488	92,23
BM6919	677	58	86	533	92,02

Tabla 12. Identificación de indels en las muestras control.

Muestra	Indels				
	Total	Solo detectado con Bioscope	Solo detectado con GATK	Ambos	% conocidos
BM3062	48	10	13	25	85,42
BM3339	49	6	19	24	89,80
BM3895	49	4	19	26	83,67
BM4237	46	10	10	26	80,43
BM5307	48	10	16	22	81,25
BM5357	49	8	14	27	83,67
BM6091	53	13	9	31	81,13
BM6092	59	8	16	35	83,05
BM6492	44	9	6	29	93,18
BM6919	41	7	13	21	90,24

Tabla 13. Detección de las mutaciones causales en las muestras control por las diferentes aproximaciones desarrolladas.

Muestra	Gen	Mutación (cDNA)	SNVS				Indels			
			Samtools		GATK		Bioscope		GATK	
			PE	SE	PE	SE	PE	SE	PE	SE
BM3062	FBN1	c.8333T>G	si	si	si	si	n/a	n/a	n/a	n/a
BM3339	FBN1	c.1148-1G>A	si	si	si	si	n/a	n/a	n/a	n/a
BM3895	FBN1	c.2248del	n/a	n/a	n/a	n/a	si	no	no	si
BM4237	KCNH2	c.2464G>A	si	si	no	no	n/a	n/a	n/a	n/a
BM5307	TGFBR2	c.1314T>A	si	si	si	si	n/a	n/a	n/a	n/a
BM5357	FBN1	c.4326dupA	n/a	n/a	n/a	n/a	no	no	si	si
BM6091	FBN1	c.5076_5078delAAG	n/a	n/a	n/a	n/a	si	no	si	si
BM6092	FBN1	c.7039_7040del	n/a	n/a	n/a	n/a	si	si	no	si
BM6492	FBN1	c.3539G>T	si	si	si	si	n/a	n/a	n/a	n/a
BM6919	MYBPC3	c.2308+1G>A	si	si	si	si	n/a	n/a	n/a	n/a

El estudio comparativo entre casos demostró que aproximadamente el 30% de las variantes identificadas eran comunes a todas las muestras siendo el porcentaje de variantes conocidas de 97,62% y 100% para SNVs e indels respectivamente (Tabla 14).

Tabla 14. Variantes comunes en las muestras control.

Número muestras	SNVs comunes	% SNVs comunes conocidos	Indels comunes	% indels comunes conocidos
1	1648	86,83	123	77,24
2	1147	93,03	84	89,29
3	890	94,72	64	89,06
4	723	95,16	54	87,04
5	616	95,45	42	88,10
6	515	95,34	34	85,29
7	428	96,03	28	85,71
8	352	96,59	24	83,33
9	306	97,39	20	85,00
10	210	97,62	13	100,00

Al igual que en el análisis del exoma, y con el fin de priorizar las variantes, se consideraron solamente aquellas potencialmente patogénicas, es decir, variantes cuyo efecto a nivel transcripcional fuera no sinónimo, produjera un cambio en la pauta de lectura o se identificara en zonas de splicing alternativo (Tabla 15). Estas variantes fueron a su vez clasificadas en diferentes grupos (Figura 42) y vinculados a niveles de prioridad en el análisis siendo 4 el nivel más prioritario. A partir de estos resultados, las variantes fueron clasificadas clínicamente por el grupo de expertos en Genética Médica tras la valoración de

los datos de la anotación generados para cada variante y la revisión bibliográfica correspondiente de la variantes conocidas.

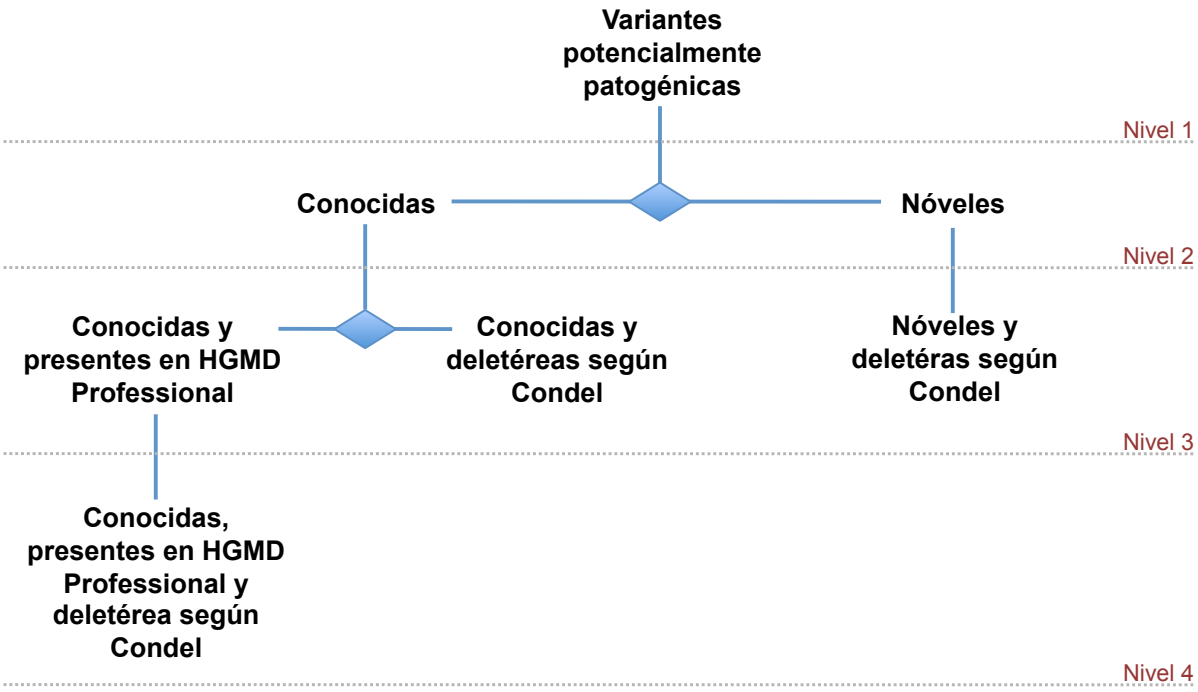


Figura 42. Clasificación de las variantes potencialmente patogénicas para priorizar las búsqueda de las variantes causales. El estudio se inicia por el grupo de variantes pertenecientes al cuarto nivel a partir del cual, en caso de no detectarse variantes de interés, se disminuye un nivel.

Asumiendo la totalidad de genes del panel, se identificó una media de 52,9 variantes por muestra, cuyo efecto a nivel transcripcional era potencialmente patológico, de las cuales 22 eran comunes al 80% de las muestras. La aplicación de los algoritmos de predicción del efecto de la variante a nivel proteico, la información sobre los posibles dominios proteicos afectados así como otra serie de datos adicionales proporcionada por el pipeline desarrollado sirvieron para discriminar las mutaciones causales en cada una de las muestra.

Tabla 15. Variantes potencialmente patogénicas en las muestras control. Las casillas coloreadas en gris marcan la clasificación más precisa para las mutaciones causales siguiendo el esquema presentado en la Figura 41. Se definió como 'variantes comunes' aquellas presentes en al menos el 80% de las muestras.

Variantes potencialmente patológicas (PP)	BM3062	BM3339	BM3895	BM4237	BM5307	BM5357	BM6091	BM6092	BM6492	BM6919	Variantes comunes
Total	67	65	49	41	50	54	42	50	46	65	22
Conocidas	55	58	38	34	38	48	36	46	35	57	20
Conocidas y deletéreas según Condel	3	6	3	5	4	8	3	3	2	6	1
Conocidas presentes en HGMD Professional	8	6	6	4	6	8	7	6	5	9	2
Conocidas, presentes en HGMD Professional y deletérea según Condel	2	3	0	1	2	2	2	0	0	3	0
Nóveles	12	7	11	7	11	6	6	4	11	8	2
Noveles y deletéras según Condel	5	0	1	1	6	0	0	1	4	4	0

2.5 Estudio retrospectivo en 163 pacientes

A finales del año 2012, se llevó a cabo un estudio retrospectivo de los resultados del panel en 163 individuos diagnosticados de patología cardiovascular con riesgo de muerte súbita. Para cada uno de ellos, se seleccionó uno o varios subpaneles de genes acorde con la historia clínica del paciente, en total se analizaron 4.795 genes (Figura 43). Del total de pacientes analizados por la Unidad de Genética Médica de Sistemas Genómicos, en 40 de ellos (24,53% del total) se detectaron un total de 45 mutaciones patogénicas de las cuales 22 habían sido descritas previamente y las otras 23, pese a no estar descritas, correspondían a proteínas truncadas. Debido a que el estudio se realizó sobre un panel con un mayor número de genes que los específicos para dicho trastorno, 8 mutaciones causales fueron identificadas en genes no relacionados con el diagnóstico cardiológico inicial. Por otro lado, se detectaron un total de 221 variantes de significado desconocido (VSD) en 93 pacientes de las cuales 187 fueron identificadas en genes correspondientes a la patología en estudio mientras que 34 de ellas se encontraron en otros genes no correspondientes a su diagnóstico cardiológico inicial. 7 de los 93 pacientes compartían además varias de estas VSD. De las 221 VSD, 14 de ellas fueron consideradas probablemente patogénicas tras un análisis más profundo de los resultados bioinformáticos obtenidos. De la misma forma 10 de ellas fueron catalogadas como variantes probablemente no patogénicas. El resto de variantes se consideraron de significado incierto (Figura 44). No se detectaron variantes

potencialmente patogénicas desde el punto de vista clínico en 30 de los 163 individuos. El listado completo de las variantes con relevancia diagnóstica se encuentran en el Anexo (Tabla 2).

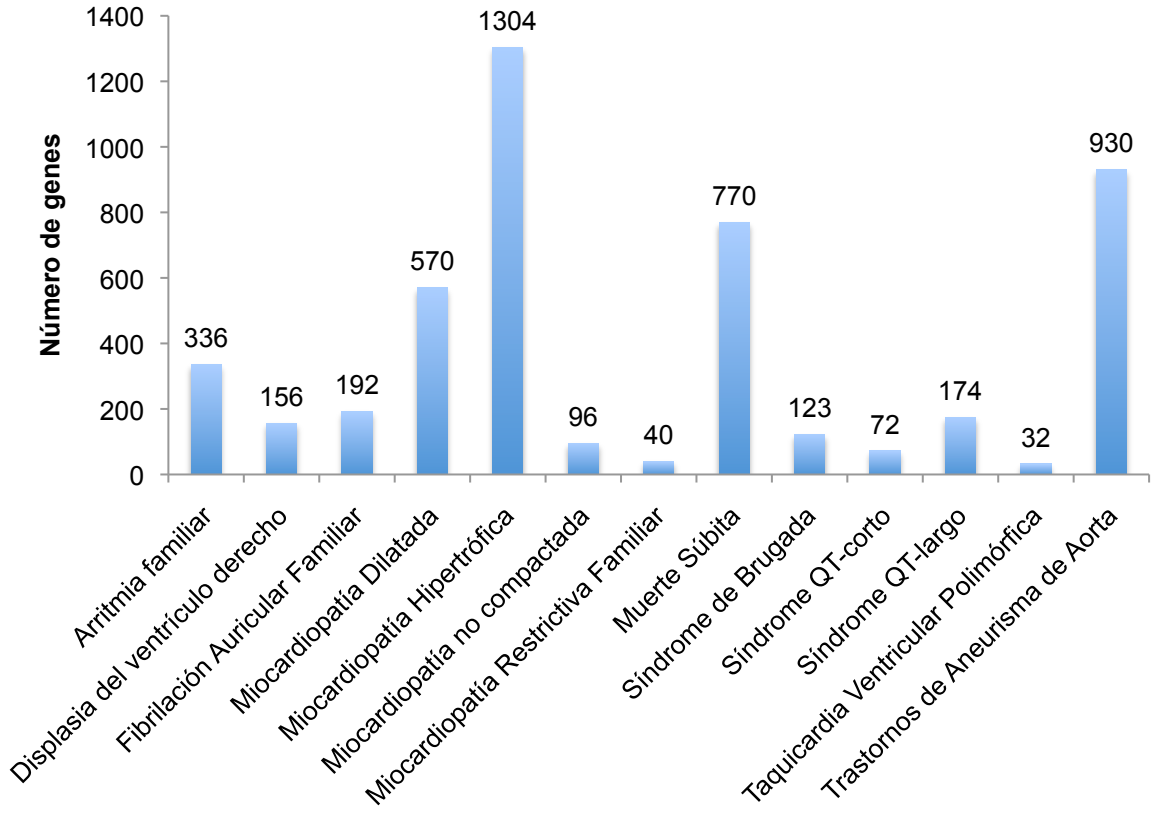


Figura 43. Número de genes analizado por enfermedad para los 163 individuos incluidos en el estudio retrospectivo.

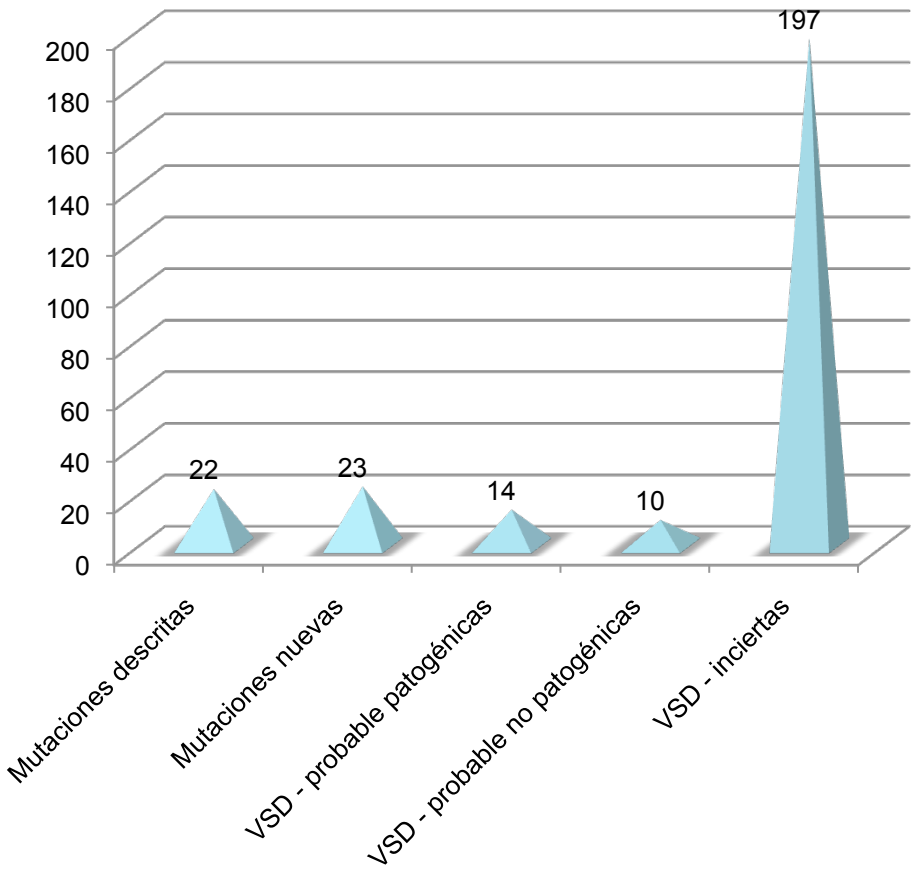


Figura 44. Resumen de las variantes identificadas en los 163 pacientes incluidos en el estudio retrospectivo.

3 Análisis de los genes *BRCA1* y *BRCA2* en mini-secuenciadores con tecnología NGS

La siguiente sección describe los resultados obtenidos tras el desarrollo y la aplicación de un sistema de diagnóstico basado en la captura de los genes *BRCA1* y *BRCA2* y su posterior secuenciación mediante mini-secuenciadores con tecnología NGS. El protocolo de análisis de las muestras se generó a partir de los protocolos desarrollados en los dos datasets anteriores adaptándolo a las necesidades de una nueva tecnología de secuenciación y un nuevo sistema de captura.
El sistema de diagnóstico fue validado en un grupo inicial de 6 muestras control previamente secuenciadas mediante tecnología Sanger incluidas dentro del mismo run (código del run BF42). En runs posteriores, para establecer las limitaciones de la tecnología, se incluyeron como controles: 4 muestras con en las que la mutación patogénica era un indel (algunas en zonas de homopolímero), 4 muestras con CNVs y una línea de HapMap, la misma empleada para la validación de los protocolos de análisis anteriores.

En el caso de este dataset, se emplearon dos pipelines diferentes que corresponden a la versión inicial y a la última actualización del mismo pipeline donde se reanalizaron y compararon los resultados en las muestras control.

3.1 Estabilidad del panel

A diferencia de la plataforma SOLiD, el secuenciador GSJunior genera lecturas largas de longitud variable que además, en este caso dada la tecnología de captura, incluyen una serie de adaptadores específicos de la plataforma que van ligados a los primers de amplificación correspondientes a cada PCR. Dado que la longitud de lectura en algunos casos supera el tamaño del amplicón, las secuencias correspondientes a estos adaptadores específicos pueden ser leídas generando así una serie de bases que no corresponden al genoma. Es por tanto, imprescindible la eliminación de estos adaptadores específicos así como de aquellas bases cuya calidad no resulte fiable.
Para cada muestra, se llevaron a cabo dos controles de calidad, en el dato inicial y una vez eliminados los adaptadores y las bases con baja calidad. Los controles de calidad correspondientes a todas las muestras, excepto la que se muestra en la Figura 45, se localizan en el Anexo (Figuras 10-23).

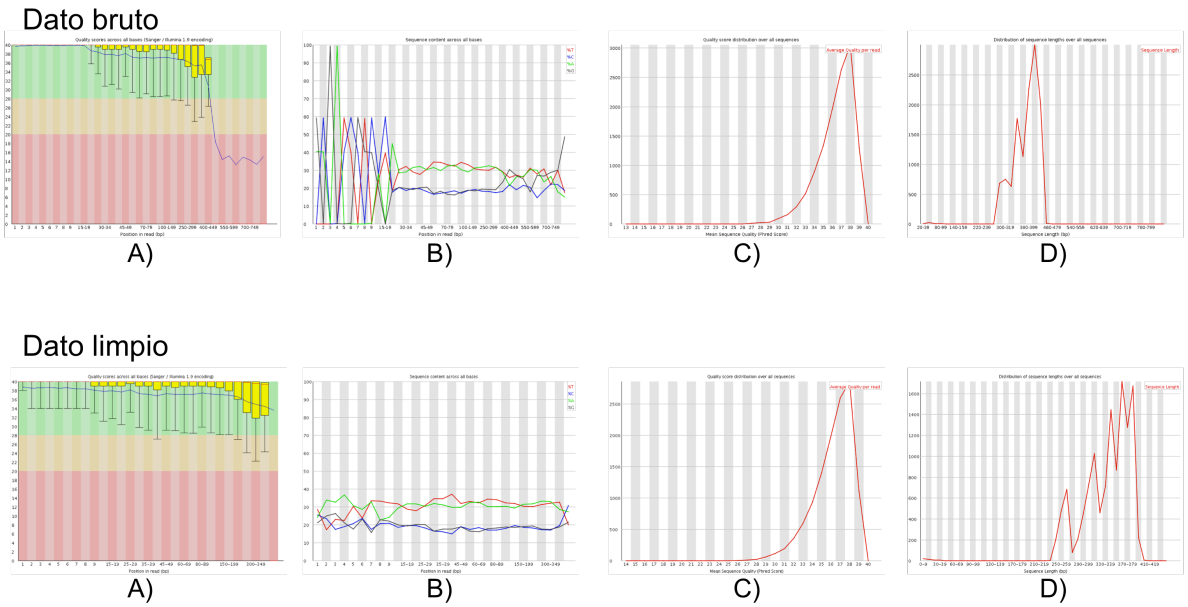


Figura 45. Control de calidad de los datos de secuenciación para la muestra 09S491.

Una vez procesados los datos iniciales, se consideraron valores óptimos para la tecnología en estudio:

1. Valor medio de calidad para cada posición de la lectura de 30 o superior.
2. Distribución del contenido de nucleótidos cercano a 41% del contenido en GC según los valores medios para el genoma humano [23].
3. Distribución de los valores medios de calidad por lectura con un pico claro por encima de un valor de Phred 30 o superior.
4. Distribución de la longitud de las lecturas mostrando un perfil en sierra con la mayoría de las lecturas incluidas en 5 picos claros.

No se observó desviación alguna respecto a los parámetros de calidad analizados en las distintas muestras.

Prácticamente el 100% de lecturas generadas pasaron los criterios de filtrado y mapearon contra la referencia asegurando así una alta sensibilidad (Tabla 16). No se detectaron diferencias significativas entre muestras.

Run	Muestra	MID	Lecturas generadas	% lecturas filtradas	Pipeline inicial		Pipeline actualizado	
					% lecturas mapeables	% lecturas mapeables en rango	% lecturas mapeables	% lecturas mapeables en rango
BF42	09S491	1	12345	0,19	99,42	99,39	99,42	99,39
	09S218	3	11034	0,14	99,48	99,46	99,48	99,47
	10S068	5	11615	0,23	99,21	99,09	99,22	99,13
	10S1106	6	11030	0,15	99,42	99,37	99,41	99,36
	09S432	7	12309	0,15	99,56	99,55	99,56	99,55
	09S523	8	11758	0,18	99,52	99,47	99,53	99,48
BF67	BRCA09880	3	14907	0,12	99,77	99,62	99,56	99,53
	BRCA09992	8	14854	0,14	99,79	99,60	99,58	99,54
BF79	BRCA10714	1	13506	0,10	99,87	99,61	99,83	99,81
	BRCA10726	2	10620	0,06	99,89	99,39	99,83	99,79
BF80	BRCA12144	6	14425	0,08	99,97	99,90	99,94	99,92
BF87	BRCA11314	2	10270	0,08	99,85	99,72	99,81	99,81
BF96	BRCA11928	6	11386	0,04	99,87	99,85	99,83	99,82
DB06	BRCA12773	6	13187	0,05	99,94	99,82	99,90	99,89
	BRCA12836	8	14177	0,06	99,88	99,68	99,83	99,75

Tabla 16. Parámetros de control del mapeo para las muestras control.

En el caso de los sistemas de amplificación mediante PCR en multiplex es imprescindible verificar que todos los amplicones se encuentran representados y que además la profundidad de lectura obtenida en cada uno de ellos alcanza unos niveles mínimos para la correcta identificación de variantes. Para ello, se generaron dos gráficas que permitieron la visualización de las muestras en estudio enfrentando los valores medios de profundidad de lectura por amplicón para las muestras dentro del mismo run, ordenando estos valores bien por la posición genómica de cada una de las PCR individuales en el gen o bien por grupos de PCRs incluidos dentro de la misma reacción en multiplex (Figuras 46-59).

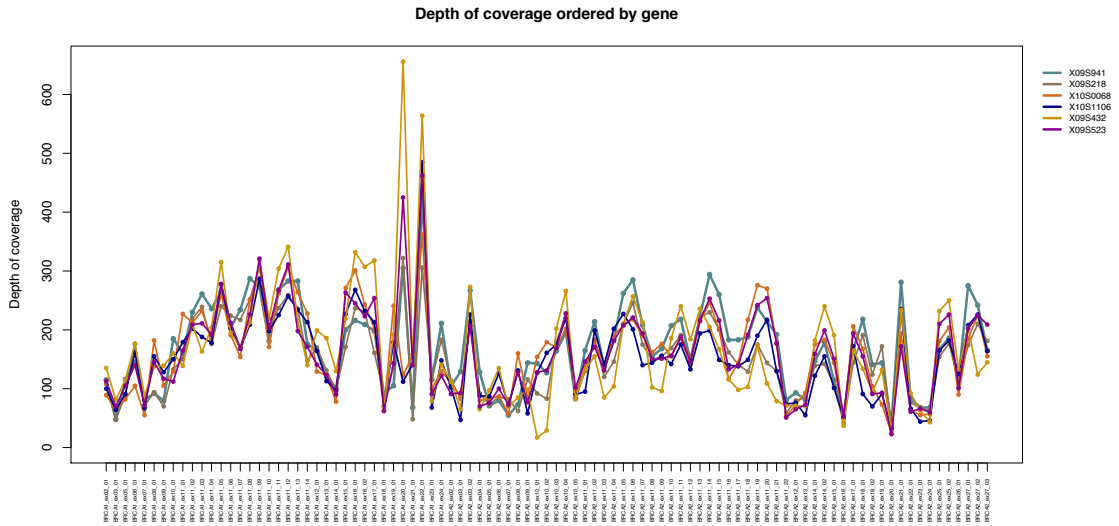


Figura 46. Evaluación de la sensibilidad del kit de captura atendiendo a la profundidad de lectura media calculada para cada PCR y ordenadas según su posición en el gen para el run BF42.

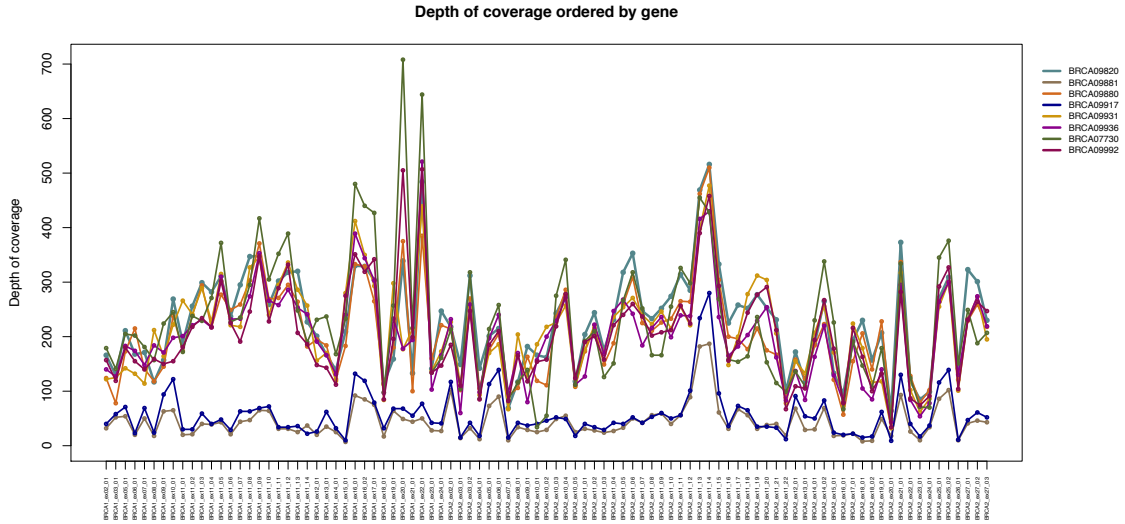


Figura 47. Evaluación de la sensibilidad del kit de captura atendiendo a la profundidad de lectura media calculada para cada PCR y ordenadas según su posición en el gen para el run BF67.

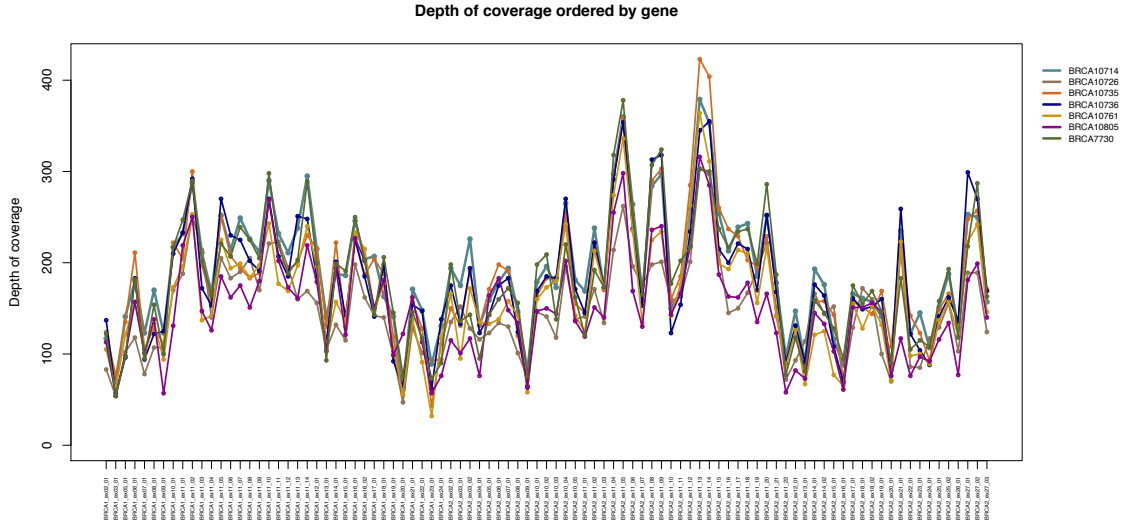


Figura 48. Evaluación de la sensibilidad del kit de captura atendiendo a la profundidad de lectura media calculada para cada PCR y ordenadas según su posición en el gen para el run BF79.

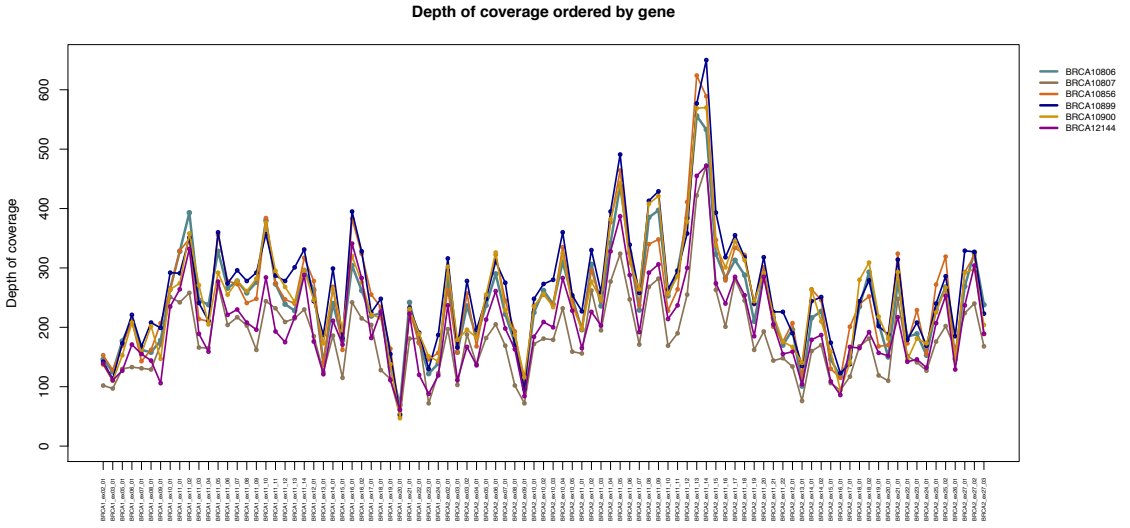


Figura 49. Evaluación de la sensibilidad del kit de captura atendiendo a la profundidad de lectura media calculada para cada PCR y ordenadas según su posición en el gen para el run BF80.

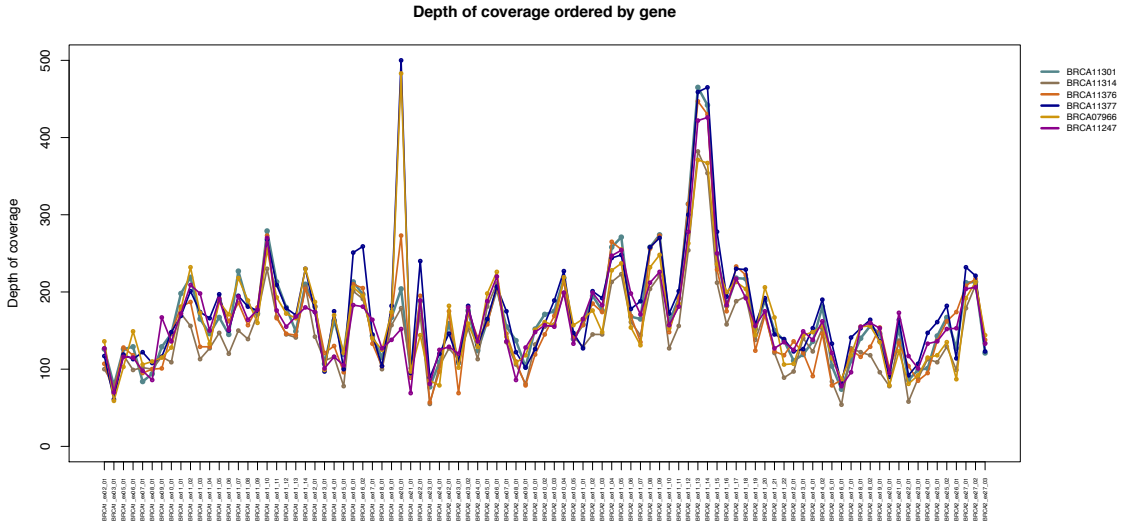


Figura 50. Evaluación de la sensibilidad del kit de captura atendiendo a la profundidad de lectura media calculada para cada PCR y ordenadas según su posición en el gen para el run BF87.

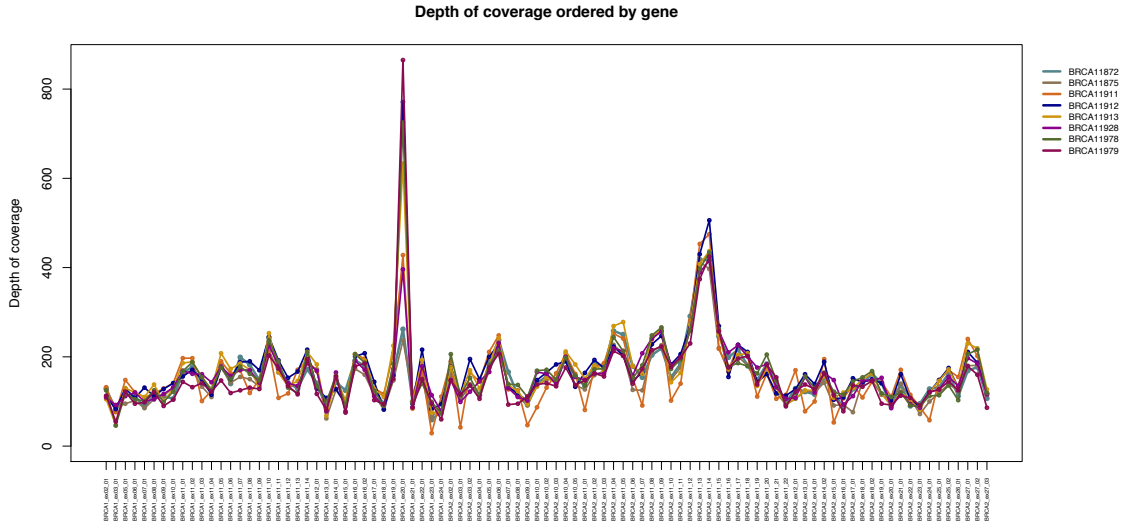


Figura 51. Evaluación de la sensibilidad del kit de captura atendiendo a la profundidad de lectura media calculada para cada PCR y ordenadas según su posición en el gen para el run BF96.

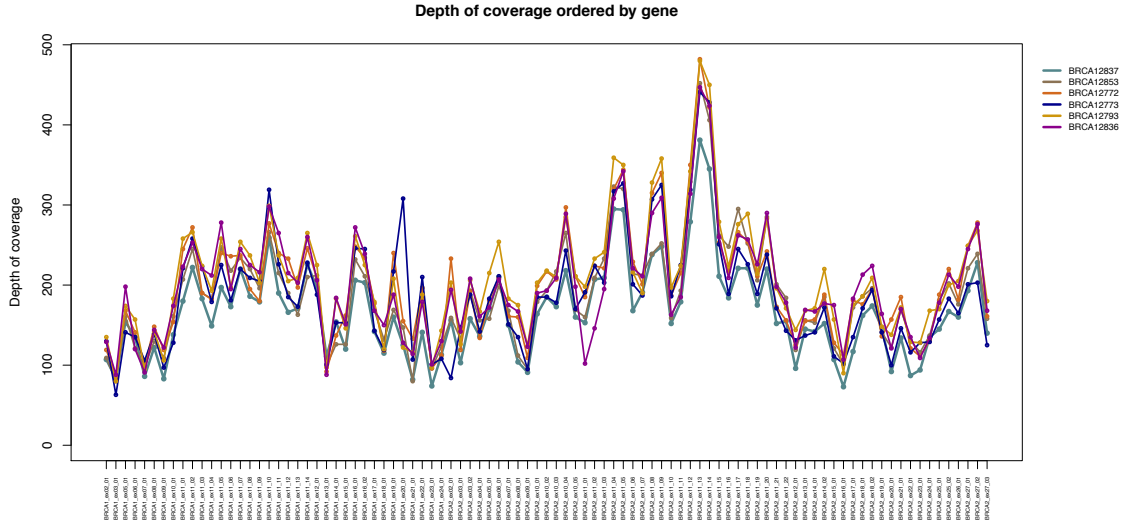


Figura 52. Evaluación de la sensibilidad del kit de captura atendiendo a la profundidad de lectura media calculada para cada PCR y ordenadas según su posición en el gen para el run DB06.

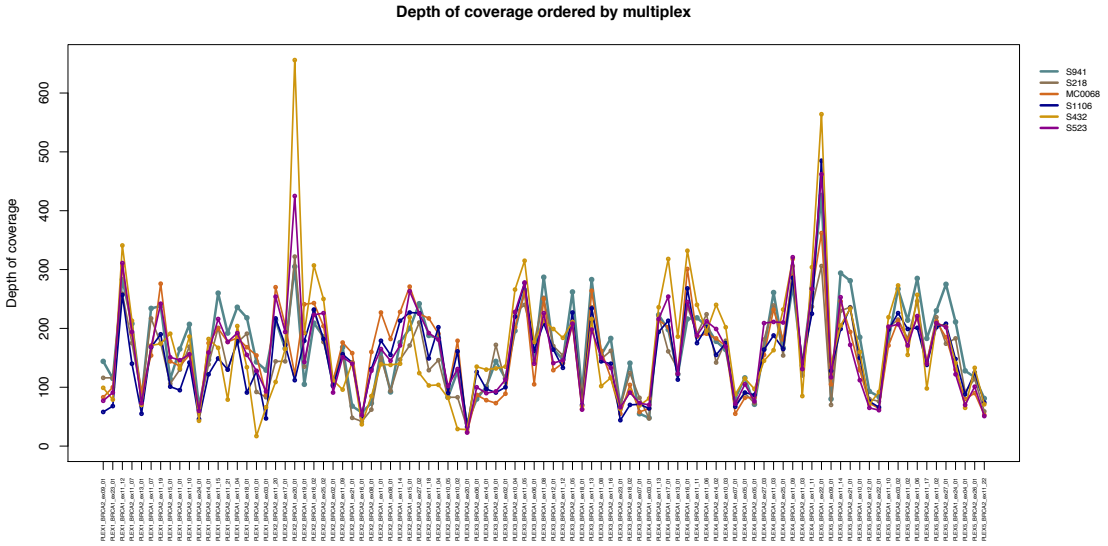


Figura 53. Evaluación de la sensibilidad del kit de captura atendiendo a la profundidad de lectura media calculada para cada PCR y ordenadas atendiendo a la multiplex para el run BF42.

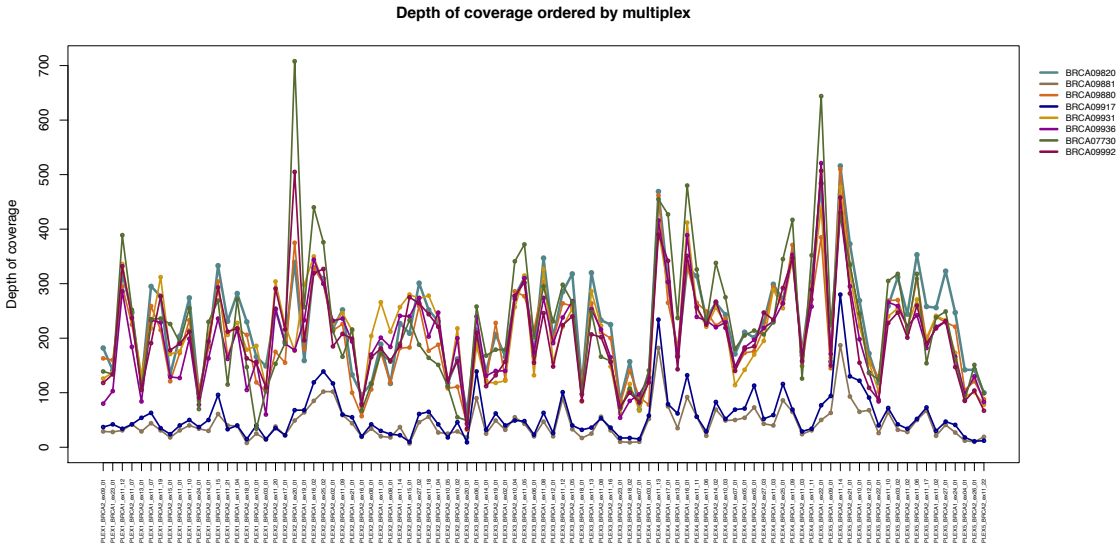


Figura 54. Evaluación de la sensibilidad del kit de captura atendiendo a la profundidad de lectura media calculada para cada PCR y ordenadas atendiendo a la multiplex para el run BF67.

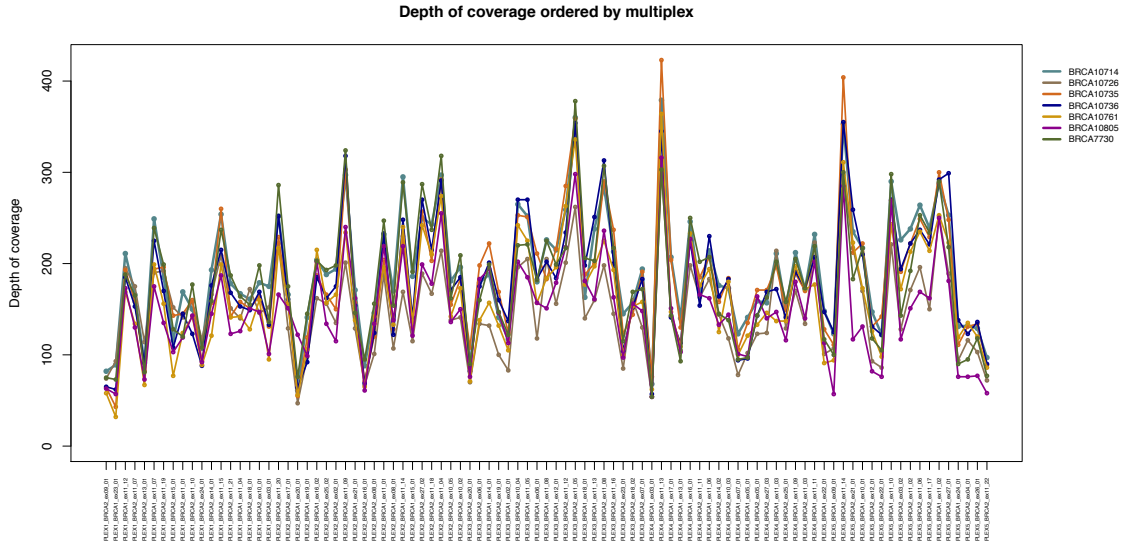


Figura 55. Evaluación de la sensibilidad del kit de captura atendiendo a la profundidad de lectura media calculada para cada PCR y ordenadas atendiendo a la multiplex para el run BF79.

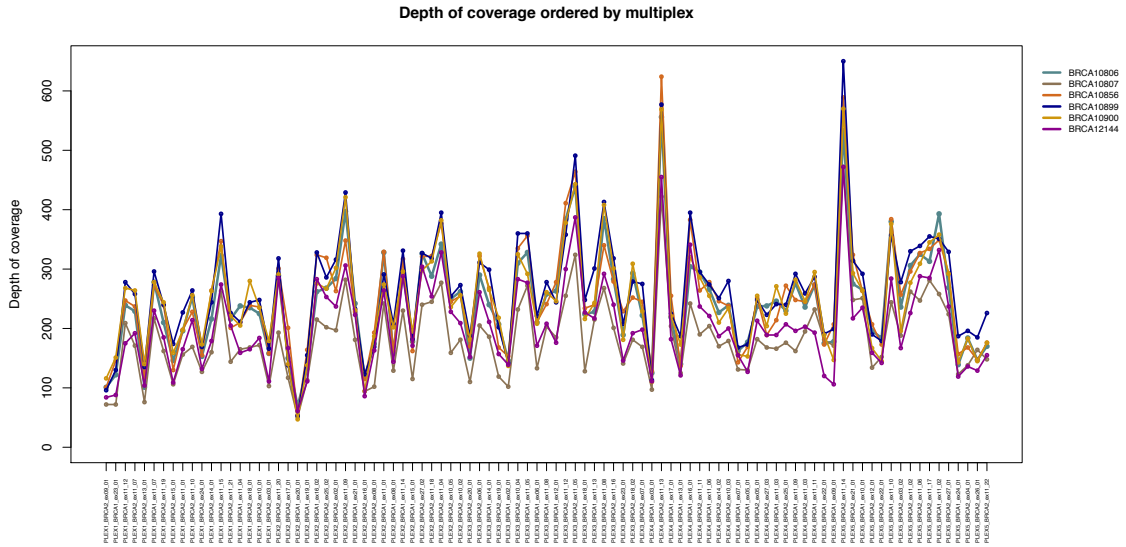


Figura 56. Evaluación de la sensibilidad del kit de captura atendiendo a la profundidad de lectura media calculada para cada PCR y ordenadas atendiendo a la multiplex para el run BF80.

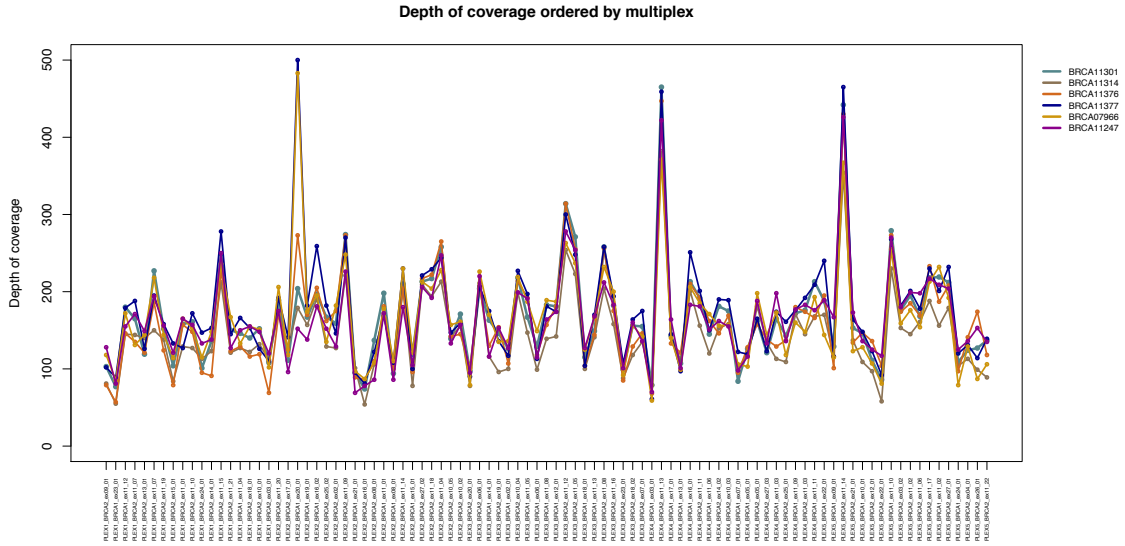


Figura 57. Evaluación de la sensibilidad del kit de captura atendiendo a la profundidad de lectura media calculada para cada PCR y ordenadas atendiendo a la multiplex para el run BF87.

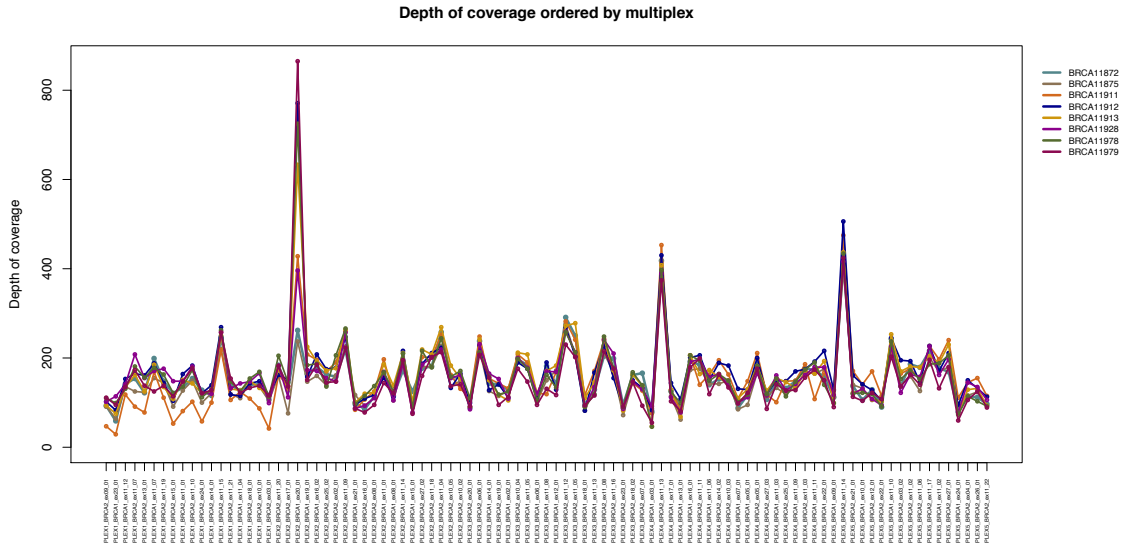


Figura 58. Evaluación de la sensibilidad del kit de captura atendiendo a la profundidad de lectura media calculada para cada PCR y ordenadas atendiendo a la multiplex para el run BF96.

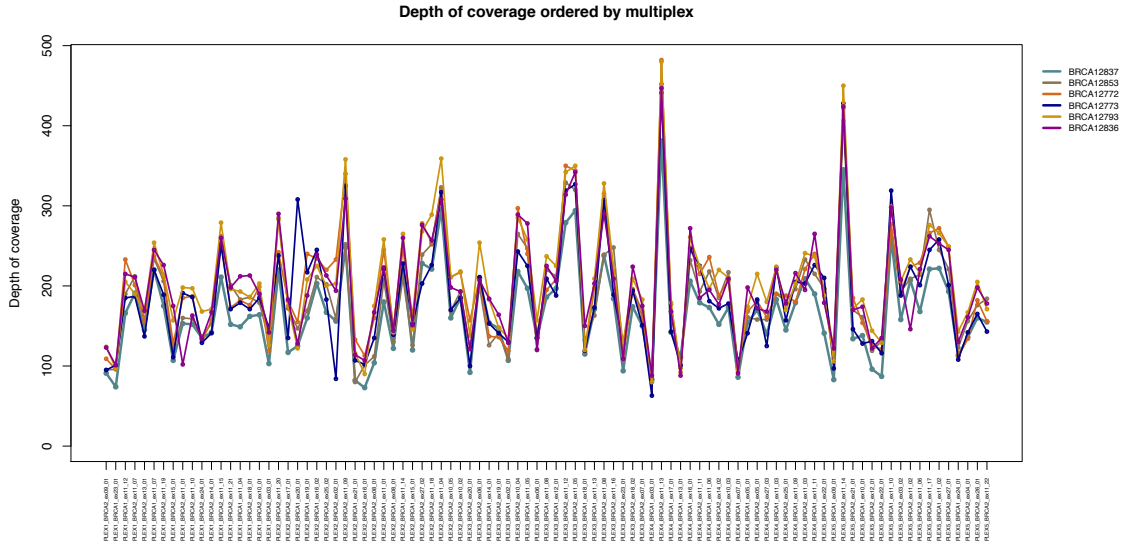


Figura 59. Evaluación de la sensibilidad del kit de captura atendiendo a la profundidad de lectura media calculada para cada PCR y ordenadas atendiendo a la multiplex para el run DB06.

A excepción de las muestras BRCA09917 y BRCA07730, cuya cobertura resultó muy deficiente debido a un problema de amplificación en el laboratorio, todas las muestras incluidas en los gráficos anteriores, controles o no secuenciadas en el mismo run, mostraron valores de profundidad de lectura muy homogéneos.

La profundidad media de lectura por amplicón en las muestras control fue de 163x (Figura 60).

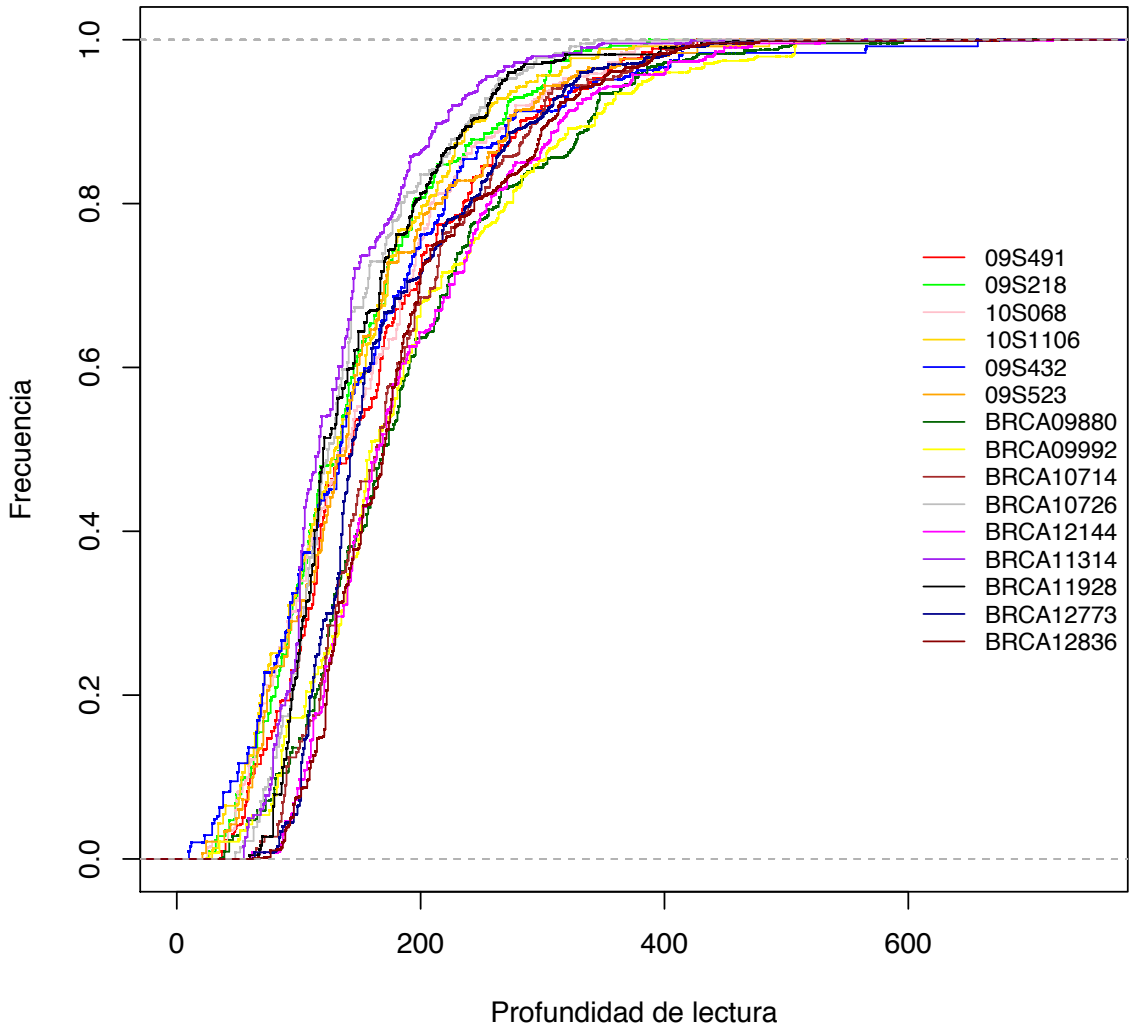


Figura 60. Distribución de la profundidad de lectura para cada una de las muestras control.

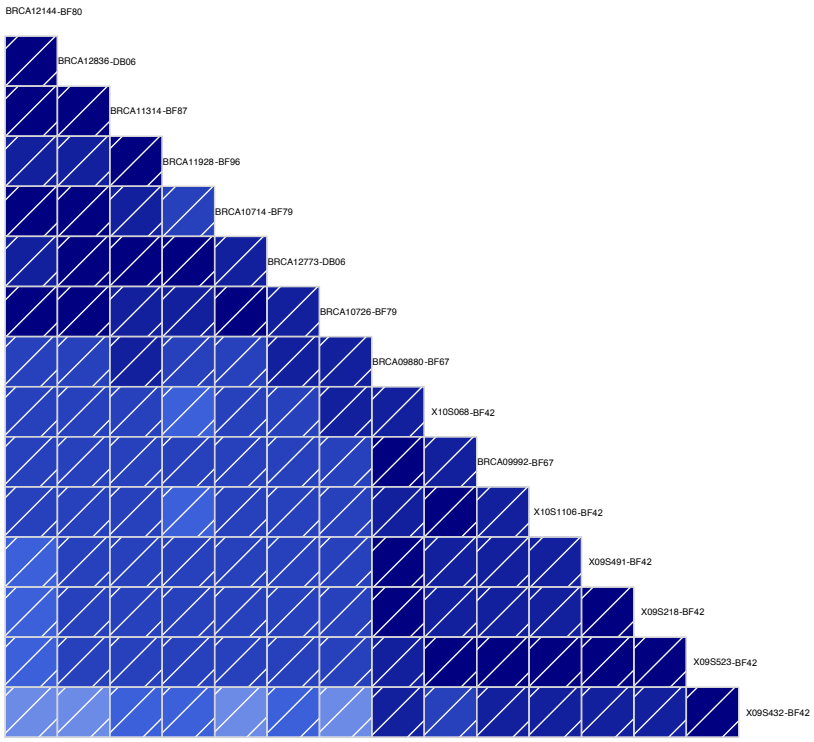


Figura 61. Correlación entre muestras atendiendo a la profundidad de lectura media alcanzada por amplicón.

El kit empleado mostró una mayor tasa de reproducibilidad entre muestras pertenecientes a la misma versión del kit de captura (versión 2.1 en todas las carreras excepto para el BF42 y BF67) y dentro del mismo run (Figura 61).

La profundidad de lectura es un factor limitante a la hora de identificar variantes en este tipo de tecnología y por lo tanto, es imprescindible contar con un umbral mínimo de profundidad de lectura por posición. Para determinar las zonas de conflicto y poder valorar la repercusión que un menor número de lecturas por base podría tener en el diagnóstico, se calculó el número de bases no cubiertas a 20x, valor mínimo establecido para la identificación de variantes [152], así como el número de variantes potencialmente patogénicas o de susceptibilidad incluidas en estas zonas y cuya detección por tanto podría verse comprometida. Tal como señalaban los datos de sensibilidad del kit de captura, solamente la muestra 09S432 se vio afectada a valores por debajo de 20x. En total 476 bases correspondientes al gen BRCA2 no fueron cubiertas a esta profundidad de lectura (regiones chr13:32906355-32906503, chr13:32906711-32906829 y chr13:32907352-32907559) de las

cuales 393 correspondían al exón 10, suponiendo un 34,90% del tamaño total de este exón. En estas regiones, se identificaron 48 variantes registradas en la base de datos HGMD (Tabla 17).

Tabla 17. Variantes patogénicas o de susceptibilidad de la base de datos HGMD no cubiertas a valores de al menos 20x en la muestra 09S432.

Cromosoma	Inicio	Final	Longitud	HGMD_ID	dbSNP ID	HGVS (cDNA)	Relevancia clínica
chr13	32906355	32906503	149	CM032200	-	NM_000059.3:c.818C>A	Breast_and/or_ovarian_cancer
chr13	32906355	32906503	149	CM118446	rs28897705	NM_000059.3:c.831T>G	Prostate_cancer
chr13	32906355	32906503	149	CM088430	rs766173	NM_000059.3:c.865A>C	Breast_cancer_in_radiographers,_decreased
chr13	32906355	32906503	149	CM1110640	-	NM_000059.3:c.880G>T	Ovarian_cancer
chr13	32906711	32906829	119	CI119504	-	NM_000059.3:c.1097dupT	Breast_and/or_ovarian_cancer
chr13	32906711	32906829	119	CM002750	rs144848	NM_000059.3:c.1114A>C	Breast_cancer,_association_with
chr13	32906711	32906829	119	CM128947	-	NM_000059.3:c.1117C>T	Breast_cancer
chr13	32906711	32906829	119	CM065030	rs80358408	NM_000059.3:c.1123C>T	Breast_cancer
chr13	32906711	32906829	119	CD086611	rs80359263	NM_000059.3:c.1128delT	Breast_cancer
chr13	32906711	32906829	119	CD063461	rs80359265	NM_000059.3:c.1147delA	Breast_cancer
chr13	32906711	32906829	119	CM065036	rs41293475	NM_000059.3:c.1151C>T	Breast_cancer
chr13	32906711	32906829	119	CM021509	rs80358411	NM_000059.3:c.1153A>T	Breast_and/or_ovarian_cancer
chr13	32906711	32906829	119	CD126350	-	NM_000059.3:c.1176_1180delCTGTG	Breast_and/or_ovarian_cancer
chr13	32906711	32906829	119	CM105154	-	NM_000059.3:c.1180G>T	Breast_and/or_ovarian_cancer
chr13	32906711	32906829	119	CM065021	rs80358412	NM_000059.3:c.1183T>G	Breast_cancer
chr13	32906711	32906829	119	CI065814	-	NM_000059.3:c.1189_1190insTTAG	Breast_cancer
chr13	32906711	32906829	119	CI126183	-	NM_000059.3:c.1189_1190insCAAC	Potential_protein_deficiency
chr13	32907352	32907559	208	CM033756	rs80358457	NM_000059.3:c.1744A>C	Breast_cancer
chr13	32907352	32907559	208	CD022531	-	NM_000059.3:c.1748delT	Breast_and/or_ovarian_cancer
chr13	32907352	32907559	208	CM119488	-	NM_000059.3:c.1748T>A	Breast_and/or_ovarian_cancer
chr13	32907352	32907559	208	CD022892	rs80359301	NM_000059.3:c.1754delA	Breast_and/or_ovarian_cancer
chr13	32907352	32907559	208	CD044112	rs80359302	NM_000059.3:c.1755_1759delGAAAA	Breast_and/or_ovarian_cancer
chr13	32907352	32907559	208	CM086612	-	NM_000059.3:c.1763A>G	Breast_cancer
chr13	32907352	32907559	208	CD032732	rs80359303	NM_000059.3:c.1763_1766delATAA	Breast_and/or_ovarian_cancer
chr13	32907352	32907559	208	CD032733	-	NM_000059.3:c.1765_1766delAA	Breast_and/or_ovarian_cancer
chr13	32907352	32907559	208	CD032734	rs80359305	NM_000059.3:c.1773_1776delTTAT	Breast_and/or_ovarian_cancer
chr13	32907352	32907559	208	CM076038	rs56328701	NM_000059.3:c.1786G>C	Breast_cancer
chr13	32907352	32907559	208	CD118428	-	NM_000059.3:c.1787_1799delATGAAACATCTTA	Prostate_cancer
chr13	32907352	32907559	208	CM035689	rs28897710	NM_000059.3:c.1792A>G	Breast_cancer
chr13	32907352	32907559	208	CD972072	rs120074213	NM_000059.3:c.1796_1800delCTTAT	Breast_cancer
chr13	32907352	32907559	208	CD129013	-	NM_000059.3:c.1797_1801delTTATA	Breast_cancer
chr13	32907352	32907559	208	CM1210124	rs80358466	NM_000059.3:c.1804G>A	Breast_cancer
chr13	32907352	32907559	208	CD113845	-	NM_000059.3:c.1805delG	Breast_and/or_ovarian_cancer
chr13	32907352	32907559	208	CD011988	rs80359309	NM_000059.3:c.1813delA	Breast_cancer
chr13	32907352	32907559	208	CI021889	-	NM_000059.3:c.1813_1814insC	Breast_and/or_ovarian_cancer
chr13	32907352	32907559	208	CI972557	rs80359308	NM_000059.3:c.1813dupA	Breast_cancer
chr13	32907352	32907559	208	CI044180	rs80359310	NM_000059.3:c.1815dupA	Breast_and/or_ovarian_cancer
chr13	32907352	32907559	208	CM043978	rs80358474	NM_000059.3:c.1832C>A	Breast_and/or_ovarian_cancer
chr13	32907352	32907559	208	CI104369	rs80359312	NM_000059.3:c.1842dupT	Breast_and/or_ovarian_cancer
chr13	32907352	32907559	208	CM065039	-	NM_000059.3:c.1847G>A	Breast_cancer
chr13	32907352	32907559	208	CI022925	-	NM_000059.3:c.1851dupA	Breast_and/or_ovarian_cancer
chr13	32907352	32907559	208	CI983042	-	NM_000059.3:c.1855dupC	Breast_cancer
chr13	32907352	32907559	208	CM066535	rs80358476	NM_000059.3:c.1855C>T	Breast_and/or_ovarian_cancer
chr13	32907352	32907559	208	CD086512	-	NM_000059.3:c.1881delA	Breast_cancer
chr13	32907352	32907559	208	CI118867	rs80359314	NM_000059.3:c.1888dupA	Ovarian_carcinoma
chr13	32907352	32907559	208	CD105672	rs80359315	NM_000059.3:c.1889delC	Poorer_survival_in_prostate_cancer
chr13	32907352	32907559	208	CI063650	-	NM_000059.3:c.1899_1900insTT	Breast_cancer
chr13	32907352	32907559	208	CD086513	-	NM_000059.3:c.1901delC	Breast_cancer

3.2 Sensibilidad y especificidad clínica del panel

3.2.1 Identificación de variantes puntuales y pequeños indels

Los datos presentados en esta sección muestran los resultados obtenidos tras el análisis de todas las muestras control con la versión inicial del pipeline y la versión más actualizada del mismo. En una primera instancia, se emplearon criterios de filtrado puramente basados en la información técnica obtenida a partir de los datos de secuenciación y, posteriormente, para tratar de identificar las variantes patogénicas, los resultados se acotaron empleando la información biológica obtenida tras la anotación de las variantes. En total, 35 SNVs, 3 inserciones y 6 deleciones pequeñas fueron utilizadas para el testeo de ambos pipelines.

La identificación de variantes en el pipeline inicial se basó en las mismas herramientas de análisis y criterios de filtrado descritos para el análisis del panel de Cardiomiopatías (ver apartado 2.3 de la sección de Métodos). Como criterio estándar para todos los runs, se tomó un valor mínimo de profundidad de lectura de 20x [152]. Adicionalmente, las variantes fueron filtradas asumiendo un valor de Phred 30, es decir, 1 error de cada 1000, tanto en el valor de calidad de la variante (SNVQ) como en el valor de calidad del genotipo (GQ). Pese al uso de estos filtros, el número de variantes identificadas, principalmente indels, resultó especialmente llamativo. La dificultad de establecer el número de nucleótidos idénticos correcto en zonas de homopolímero mediante pirosecuenciación genera una alta tasa de error en estas zonas siendo más acusada cuanto mayor sea la longitud del homopolímero [153]. No obstante, este filtro redujo ligeramente el número de indels mientras que el número de SNVs permaneció sin cambios (Tabla 18). Todas las variantes presentes en las muestras control fueron identificadas en el listado final. A partir de este punto, se tomó en cuenta la información biológica asociada a las muestras tras incorporar los datos de Ensembl relativos a cada variante. Así, las variantes restantes fueron priorizadas en función de su efecto a nivel transcripcional considerando éste como parámetro único al clasificarlas como potencialmente patogénicas (PP). Todas las mutaciones causales se encontraban dentro del listado final. De entre las 10 mutaciones patogénicas de las muestras control, 2 de ellas además estaban registradas en la base de datos HGMD como mutaciones patogénicas generándose así un diagnóstico prácticamente directo.

En la versión más evolucionada del pipeline de análisis, el software GATK fue actualizado a una versión superior, que permitía una mayor flexibilidad a la hora de ejecutar el pipeline haciendo visibles, y por lo tanto modificables, algunos de los parámetros que anteriormente se usaban por defecto. Adicionalmente, se sustituyó Samtools por VarScan2, basado en métodos heurísticos que proporcionan una detección de variantes más eficiente en zonas con una profundidad de lectura extrema [138]. Otra de las ventajas que ofrecía VarScan2 sobre la versión de Samtools inicial, es que permitía obtener la información cruda sobre la frecuencia alélicas de las variantes de forma similar a GATK permitiendo así priorizar las variantes de una forma diferente. Para cada muestra, se combinó la información resultante de ambos 'callers' prevaleciendo siempre los resultados obtenidos por GATK sobre los resultados de VarScan2 por ser una herramienta con mayor recorrido que VarScan2 en el momento del ensayo, y por lo tanto fue considerada como más robusta. La llamada de variantes se realizó en ambos casos intentando rescatar el mayor número de ellas posible, es decir, disminuyendo a valores mínimos cada uno de los límites por defecto de ambos callers para posteriormente realizar un filtrado controlado sobre los resultados (Tabla 20). Los resultados, a diferencia del pipeline inicial, se reportaron en un mismo fichero para facilitar su revisión. Para priorizar las variantes, se generó un primer filtro (Filtro 1) en el que

se tomó un valor de profundidad de lectura mínimo de 20x y un valor de frecuencia mínima para el alelo alternativo de 0,2 como parámetros por defecto teniendo en cuenta los datos de validación Sanger obtenidos tras el análisis de varios runs (Figura 62). Dado que los errores de secuenciación en las zonas de homopolímero suelen repetirse en cada muestra secuenciada, en esta última versión del pipeline, las variantes fueron combinadas en un único fichero para facilitar su comparación y posterior priorización. Así, en un paso más para reducir el número de variantes candidatas e intentar aproximarse lo más posible a la mutación causal, en cada una de las muestras control se descartaron además todas aquellas variantes presentes en más del 90% de las muestras (Filtro 2) considerándose que éstas derivarían fundamentalmente de: a) errores de secuenciación reproducidos en todos los runs, b) variantes con una frecuencia muy alta en la población y por lo tanto poco relevante para el diagnóstico, c) mutaciones propias de la misma secuencia de referencia empleada. Teniendo en cuenta las variantes potencialmente patogénicas, tras realizar el primero de los filtros, el número de indels inicial fue reducido entre un 63% y un 86%. El segundo de los filtros redujo entre un 6% y un 16% de indels adicional. El número de SNVs se redujo tras el primer filtro en un 33% y un 75% a excepción de la muestra BRCA11928 donde no produjo efecto alguno. La comparación entre muestras redujo el número de variantes candidatas entre un 6,5% y un 50% adicional. Todas las variantes se identificaron en el dataset final.

Run	Muestra	Sin filtrar				Filtro: 20x + SNVQ>30 + GQ>30 + frecuencia 0,2 en GATK			
		SNVs		Indels		SNVs		Indels	
		Total	PP	Total	PP	Total	PP	Total	PP
BF42	09S491	20	6	48	17	20	6	46	17
	09S218	22	7	44	12	22	7	42	12
	10S068	26	6	49	20	26	6	42	17
	10S1106	30	10	45	12	30	10	41	11
	09S432	11	3	37	14	11	3	35	14
	09S523	24	7	45	14	24	7	42	13
BF79	BRCA10714	22	7	52	18	22	7	50	18
	BRCA10726	26	7	44	13	26	7	43	12
BF87	BRCA11314	29	9	30	9	29	9	30	9
BF96	BRCA11928	27	6	33	16	27	6	30	14
BF80	BRCA12144	25	8	68	27	25	8	67	26

Tabla 18. Resultados de la llamada de variantes tras la ejecución del pipeline inicial. Las variantes clasificadas como potencialmente patogénicas (ver Métodos, apartado 3.2.7) se reportan bajo las columnas nombradas como 'PP'.

Pese a que los sistemas de captura mediante amplicones sean mucho más estables desde el punto de vista de la cobertura que los sistemas de hibridación por sondas, estos sistemas no están exentos de grandes variaciones en cuanto a la profundidad de lectura obtenida. Dado que la profundidad de lectura es el factor limitante más importante para la detección de variantes, se investigó la posibilidad de generar un filtro dinámico a partir del cual, dependiendo de la profundidad de lectura alcanzada se aplicaba una frecuencia alélica diferente. Así, tomando como base el modelo de error definido por De Leeneer en datos de pirosecuenciación para alcanzar un 100% de sensibilidad teórico, se generó un nuevo filtro (Filtro 3) [75]. En el dataset resultante del Filtro 3 se eliminaron todas aquellas variante identificadas en el 90% de las muestras o más (Filtro 4). Este nuevo filtro no tuvo prácticamente repercusión a nivel de SNVs, sin embargo, en comparación con el Filtro 2, el número de indels se redujo entre un 11% y un 58% del total. A pesar del descenso en el número de indels de forma significativa, esta forma de filtrado generó un falso negativo en la muestra control BRCA11928, cuya mutación patogénica era una duplicación de una A en una zona de homopolímero con una longitud de 6pb. En la misma posición de la variante (NM_000059.3:c.956dupA) se identificaron dos alelos alternativos, el primero de ellos era una deleción de una A que había sido identificada en todas las muestras del mismo run en heterocigosis a frecuencias muy bajas descartadas tras aplicar las tablas de De Leeneer. La segunda de las variantes, correspondiente a la mutación patogénica, una inserción de una A que se presentaba a una frecuencia de 0,33 y una profundidad de lectura de 209x, frecuencia alélica ligeramente inferior al límite teórico de 0,35 establecido por De Leeneer para esa profundidad de lectura. El resto de variantes de las muestras control se reportaron en el dataset final. Se obtuvo, por tanto, una sensibilidad del 98% tras aplicar las tablas teóricas de De Leeneer [153] siendo descartada por tanto esta opción de filtrado. Los resultados de estas pruebas se resumen en la Tabla 19.

Dada la dificultad que implica la detección de indels de forma automática con esta tecnología debido al problema del homopolímero, el departamento de Biomedicina de Sistemas Genómicos generó un protocolo de trabajo que combinaba los resultados de este pipeline bioinformático junto con los resultados obtenidos a partir del análisis de fragmentos de las PCRs (Figura 63) haciendo posible una identificación eficiente de los indels y solventando así las limitaciones de la tecnología de secuenciación en zonas de homopolímero.

De manera adicional, se evaluaron los porcentajes de sensibilidad y especificidad aplicando el 'Filtro 2' (profundidad de lectura de al menos 20x y frecuencia del alelo alternativo de 0,2) en la línea celular HapMap NA12144 (BRCA12144) para la que existían 8 variantes descritas para los genes BRCA y un total de 62 posiciones interrogadas. Se obtuvieron valores de sensibilidad y especificidad del 100% y del 95%, respectivamente.

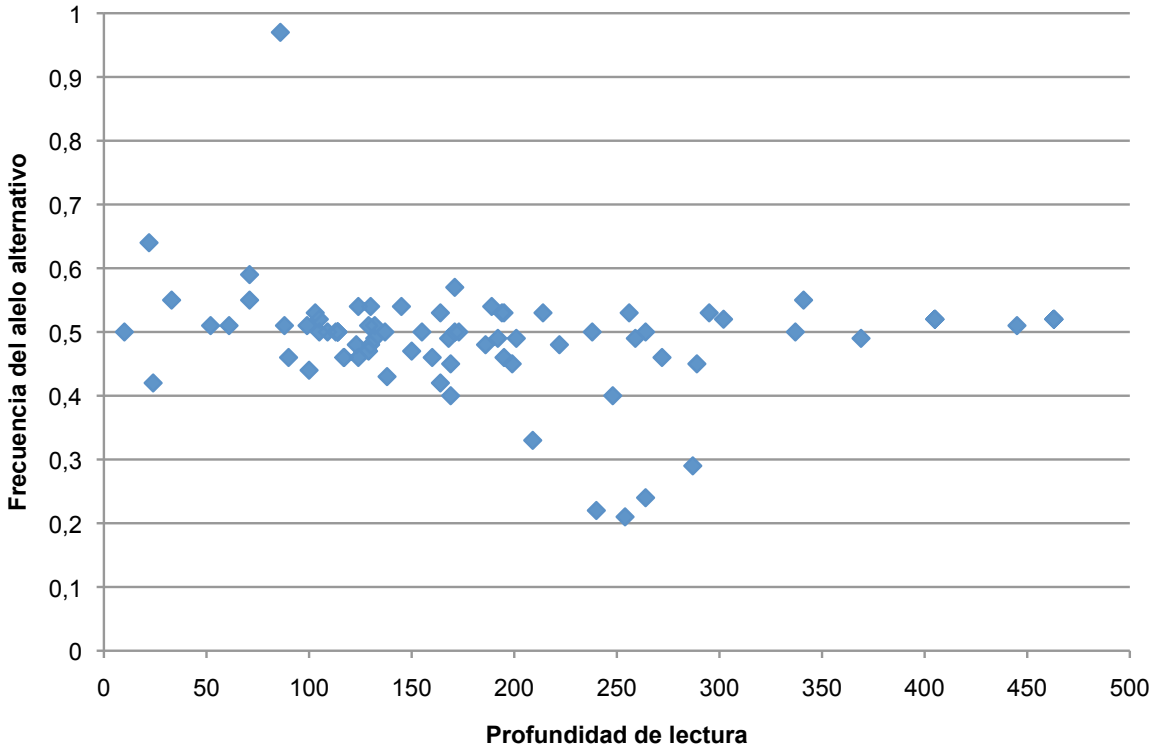


Figura 62. Frecuencia alélicas frente a la profundidad de lectura obtenidas tras la validación mediante secuenciación Sanger de 77 variantes. Las variantes identificadas a frecuencias menores a 0,4 correspondían a indels en zonas de homopolímero.

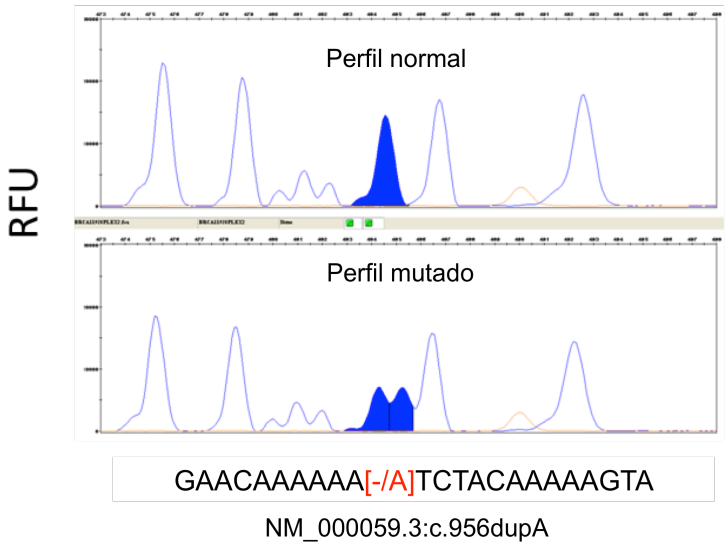


Figura 63. Identificación de pequeños indels a través del análisis de fragmentos. Si existe una deleción o una inserción el tamaño del fragmento en el perfil normal se ve reducido a la mitad en altura y aparece un nuevo pico correspondiente al alelo mutado en el perfil mutado.

Run	Muestra	Pipeline actualizado																	
		Sin filtrar				Filtro1 (>20x + >0.2)				Filtro 2 (Filtro 1 + no en 90%)				Filtro 3 (Filtro dinámico)				Filtro 4 (Filtro dinámico + no en 90%)			
		SNVs		Indels		SNVs		Indels		SNVs		Indels		SNVs		Indels		SNVs		Indels	
		Total	PP	Total	PP	Total	PP	Total	PP	Total	PP	Total	PP	Total	PP	Total	PP	Total	PP	Total	PP
BF42	09S491	48	13	276	159	28	8	68	30	22	6	43	19	33	7	46	14	26	5	29	9
	09S218	48	12	250	133	27	8	66	26	21	6	40	16	34	9	54	20	27	7	33	14
	10S068	67	19	285	154	32	6	79	35	25	4	54	25	41	10	69	28	34	8	47	21
	10S1106	64	23	264	149	32	10	77	27	25	8	52	18	41	12	64	20	34	10	47	16
	09S432	35	11	259	125	17	3	64	25	11	1	41	17	22	6	66	24	15	4	47	18
	09S523	52	13	260	141	33	7	68	34	26	5	47	25	35	7	58	20	28	5	40	14
BF79	BRCA10714	45	15	171	88	29	8	44	15	23	7	25	8	34	9	25	8	28	8	15	4
	BRCA10726	45	11	197	99	31	7	46	14	24	5	27	8	34	7	30	9	27	5	18	5
BF87	BRCA11314	49	13	208	98	28	8	89	36	21	6	56	23	37	8	59	22	30	6	33	11
BF96	BRCA11928	29	4	152	69	14	4	61	24	7	2	35	14	19	4	36	14	12	2	19	8
BF80	BRCA12144	31	8	134	63	16	2	58	22	9	0	30	12	20	3	31	10	13	1	14	5

Tabla 19. Resultados de la llamada de variantes tras la ejecución del pipeline actualizado en las muestras control aplicando los diferentes filtros. Las variantes clasificadas como potencialmente patogénicas se reportan bajo las columnas nombradas como 'PP'.

A diferencia de la información biológica reportada en el pipeline inicial, en este último pipeline se introdujo información sobre las frecuencias poblacionales de las variantes conocidas siendo posible de esta forma discriminar si una variante conocida puede ser candidata o no asumiendo como patológico frecuencias inferiores al 1% en el caso de modelos de herencia recesivos o frecuencias inferiores al 5% en el caso de modelos dominantes. La Tabla 20 muestra un ejemplo de la información obtenida para cada una de las variantes reportadas por el pipeline.

Tabla 20. Información reportada por el pipeline para cada una de las variantes. La tabla recoge los resultados de dos variantes, un polimorfismo poblacional de baja frecuencia (c.2077G>A) y una mutación patogénica (c.66_67delAG).

Anotación	**NM:007294.3:c.2077G>A**	**NM_007294.3:c.66_67delAG**
#HGNC_symbol	BRCA1	BRCA1
Chr	chr17	chr17
Pos	41245471	41276046
Ref_Allele	C	TCT
Var_Allele	T	T
09S491_S1_Depth	121	116
09S491_S1_Var/Depth	0	0
09S491_S1_Genotype	Homo_ref	Homo_ref
09S218_S1_Depth	149	109
09S218_S1_Var/Depth	0	0
09S218_S1_Genotype	Homo_ref	Homo_ref
10S068_S1_Depth	118	90
10S068_S1_Var/Depth	0	0,46
10S068_S1_Genotype	Homo_ref	P_Hetero
10S1106_S1_Depth	117	101
10S1106_S1_Var/Depth	0,46	0
10S1106_S1_Genotype	P_Hetero	Homo_ref
09S432_S1_Depth	95	136
09S432_S1_Var/Depth	0	0
09S432_S1_Genotype	Homo_ref	Homo_ref
09S523_S1_Depth	130	114
09S523_S1_Var/Depth	0	0
09S523_S1_Genotype	Homo_ref	Homo_ref
Gene_ID	ENSG00000012048	ENSG00000012048
Gene_description	breast_cancer_1&_early_onset	breast_cancer_1&_early_onset
HGVSc_name	ENST00000357654.3:c.2077G>A	ENST00000357654.3:c.66_67delAG
Intron	-	-
Exon	10/23	2/23
HGVSp_name	ENSP00000350283.3:p.Asp693Asn	ENSP00000350283.3:p.Glu23ValfsTer17
Variant_effect	missense_variant	frameshift_variant&feature_truncation
ALL_MAF	T:0,0395	-
AFR_MAF	0,01	-
AMR_MAF	0,05	-
ASN_MAF	0	-
EUR_MAF	0,08	-
InterPro_IDs	IPR011364	IPR011364
InterPro_descriptions	BRCA1	BRCA1
HGMD_info	CM960172/NM_007294.3:c.2077G>A/DP/Altered_radiation_exposure-response_relationship	CI962220/NM_007294.3:c.66dupA/DM/Breast_cancer
Related_publication	http://www.ncbi.nlm.nih.gov/pubmed/17764108	http://www.ncbi.nlm.nih.gov/pubmed/8807330
Existing_variation	rs4986850	rs80357713&rs199805151&COSM35893
Transcript_ID	ENST00000357654	ENST00000357654
RefSeq_ID	NM_007294.3	NM_007294.3
CCDS_ID	CCDS11453.1	CCDS11453.1
Canonical_isoform	YES	YES
Conservation_score	-0,431	-0,605
Grantham_distance	23	-
Condel_prediction	neutral(0,406)	-
SIFT_prediction	deleterious(0,01)	-
PolyPhen_prediction	benign(0,011)	-
Affected_prot_domains	Pfam_domain:PF04873&PIRSF_domain:PIRSF001734	Superfamily_domains:SSF57850&PIRSF_domain:PIRSF001734
Flanking_sequence	AAAAGAACCAGGTGCATTTGTTAACTTCAGCTCTGGGAAAGTATCGCTGT[C/T]ATGTCTTTTACTTGTCTGTTCATTTGGCTTGTTACTCTTCTTGGCTCCAG	AGGAATCCCAAATTAATACACTCTTGTGCTGACTTACCAGATGGGACACTC[TCT/T]TAAGATTTTCTGCATAGCATTAATGACATTTTGTACTTCTTCAACGCGAAG

3.2.2 Identificación de variaciones en el número de copias (CNVs)

Pese a que la intención de los gráficos generados en el apartado 3.1.3 de esta misma sección 'a priori' no fuera identificar CNVs, los valores de profundidad media por amplicón sin normalizar, que servían inicialmente para estudiar la cobertura global del panel, ya dejaban entrever en algunas carreras ligeros descensos en la profundidad de lectura media por amplicón que no tenían relación con una disminución de la eficiencia en alguna de las PCR multiplex. La aplicación de una serie de pasos de normalización disminuyó el ruido entre muestras destacando de forma más clara las variaciones en el número de copias presentes en las muestras control (Figuras 64 y 65). Pese a la normalización, el método implementado debe ser mejorado para eliminar una mayor cantidad de ruido en aquellos runs donde existe una menor correlación.

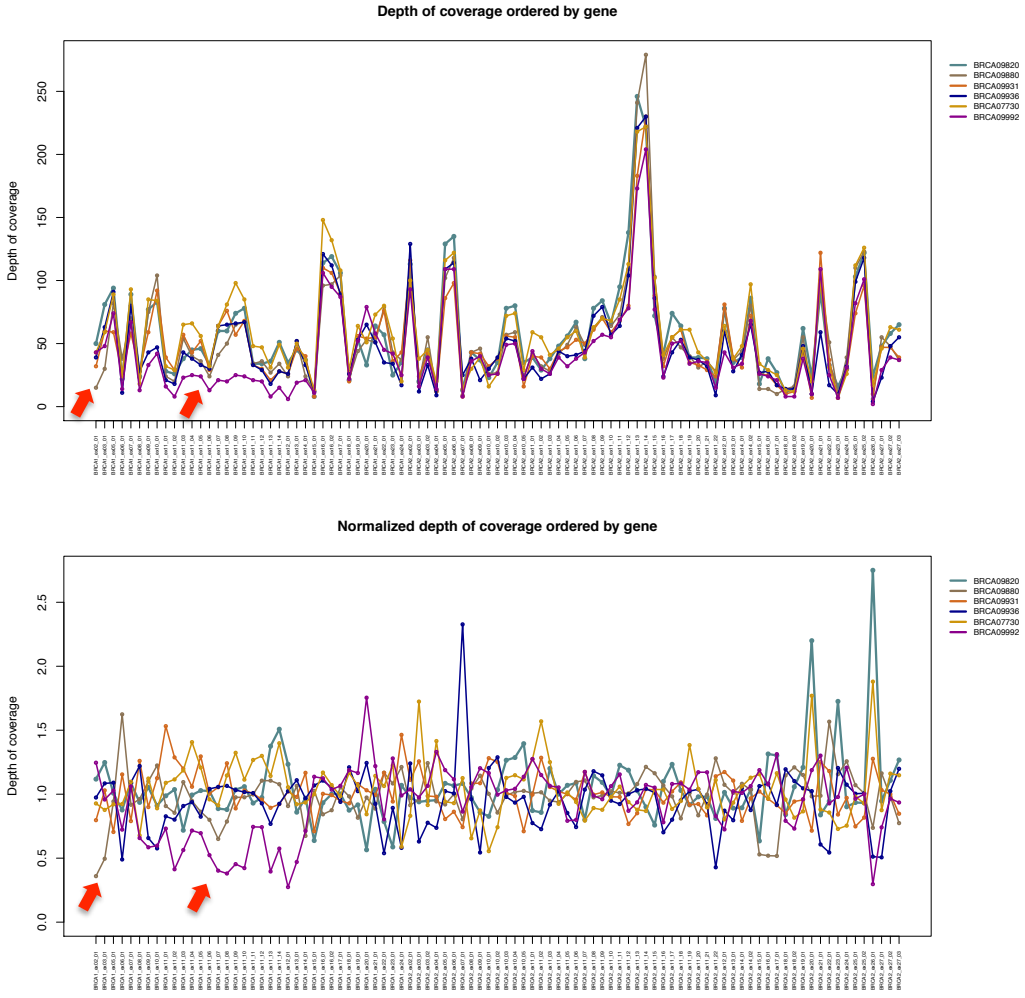


Figura 64. Detección de CNVs a partir de los datos de profundidad de lectura por amplicón normalizados para el run BF67. El gráfico superior muestra los datos sin normalizar. El gráfico inferior muestra los resultados ya normalizados.

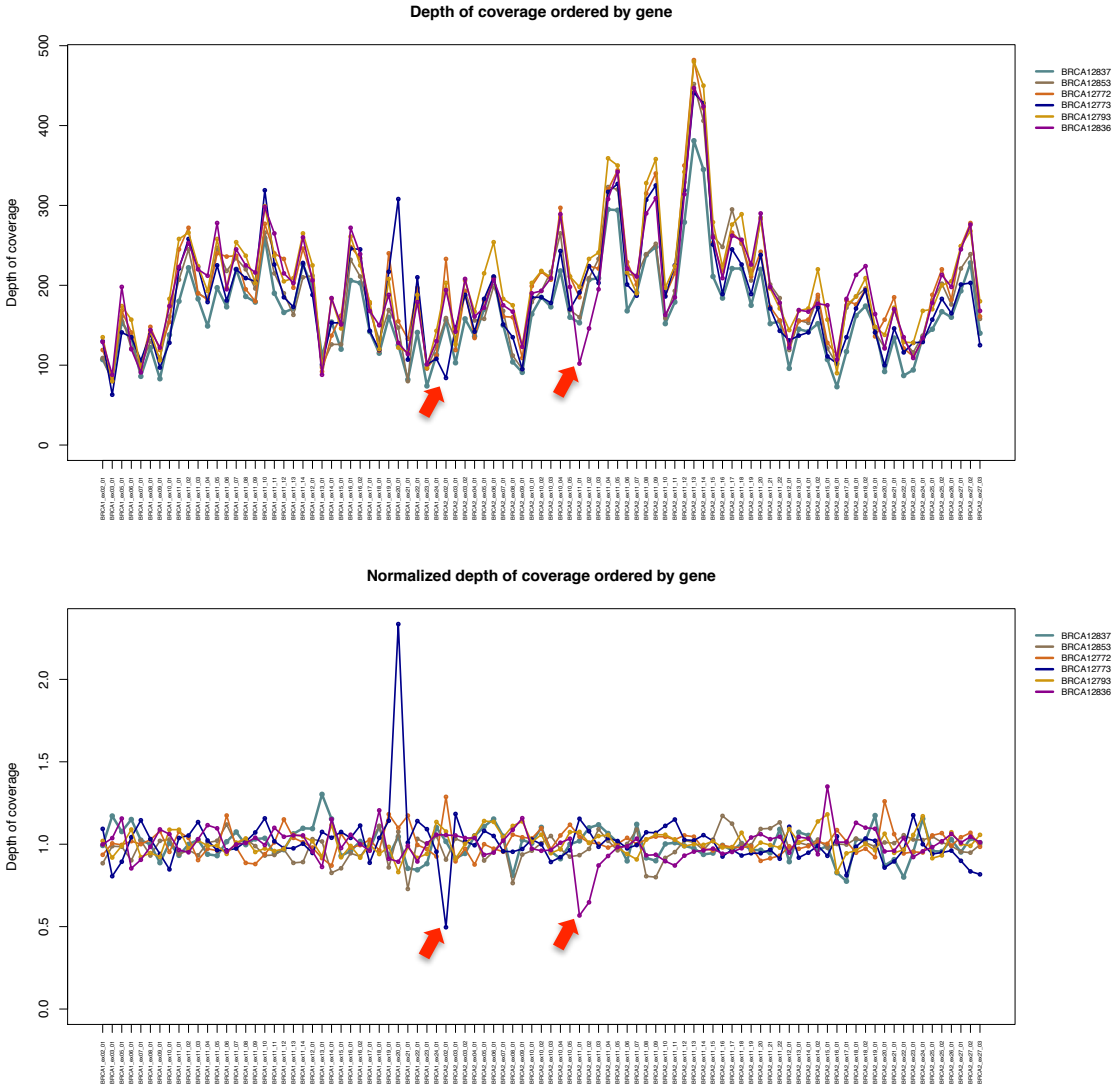


Figura 65. Detección de deleciones grandes a partir de los datos de profundidad de lectura por amplicón normalizados para el run DB06. El gráfico superior muestra los datos sin normalizar. El gráfico inferior muestra los resultados ya normalizados

DISCUSIÓN

La secuenciación masiva es uno de los avances tecnológicos más revolucionarios de nuestro tiempo donde la Bioinformática juega un papel fundamental. La gran capacidad de generación de datos de estas máquinas en periodos de tiempo muy cortos unido a su bajo precio por base en comparación con los métodos tradicionales de secuenciación han hecho que esta tecnología se extienda de forma fulminante en la comunidad científica. El continuo incremento en la cantidad de datos producidos por estas plataformas desencadenó, en su inicio, un importante desfase entre el ritmo de evolución de estos nuevos sistemas de secuenciación y el desarrollo de herramientas de análisis especializadas capaces de procesar tal magnitud de datos, situación que se ha prolongado durante años. A lo largo de este tiempo, la comunidad bioinformática ha mantenido un ritmo de desarrollo frenético para lograr traducir esta enorme cantidad de información en respuestas más simples y relacionarlas posteriormente dentro un contexto biológico. Actualmente, existen diversas herramientas de análisis de acceso libre diseñadas para llevar a cabo cada unos de los pasos que componen el procesamiento global de las muestras en estudios de resecuenciación. Sin embargo, es extremadamente importante validar cada una de las herramientas que se pretenden emplear en los protocolos de análisis de rutina y conocer a fondo la metodología de análisis subyacente y sus limitaciones a fin de determinar el grado de incertidumbre de la tecnología.

Esta tesis presenta el desarrollo de varios protocolos de análisis modulares e integrales dirigido al procesamiento de datos procedentes de estudios de resecuenciación dirigida. El protocolo inicial, aplicado en la investigación de enfermedades mendelianas raras, se desarrolló a partir de datos procedentes de la secuenciación del exoma completo tomando como base para su puesta a punto los datos disponibles del genotipado masivo por arrays para la muestra control, una línea celular HapMap. Este pipeline de análisis inicial demostró su eficiencia en la identificación de un nuevo gen responsable de Anemia de Fanconi en un paciente cuyo diagnóstico, tras el screening de los genes más frecuentes en esta enfermedad, seguía siendo una incógnita. Una versión actualizada de este protocolo se aplicó posteriormente con éxito en el análisis de los datos procedentes de paneles de genes orientados al diagnóstico en enfermedades cardiovasculares en muestras control. Este nuevo protocolo incorporó herramientas de análisis más avanzadas que aumentaron la sensibilidad y especificidad en la identificación de variantes, proporcionando una mayor eficiencia, precisión y robustez en el sistema de desarrollado con objetivos diagnósticos.

Paralelamente, el protocolo de análisis se adaptó a una nueva plataforma de secuenciación y sistema de captura distintos, en este caso dirigido a la resecuenciación de los genes *BRCA1* y *BRCA2*, mediante el uso de muestras control en las que fue capaz de identificar el 100% de los cambios descritos. Los pipelines de análisis desarrollados incluyen herramientas desde la manipulación de las secuencias brutas generadas hasta la obtención del listado final de variantes o genes de interés que sirvieron en cada uno de los estudios expuestos anteriormente para obtener un diagnóstico final. Asimismo, esta tesis proporciona una visión detallada sobre las limitaciones tecnológicas que ofrecen las aproximaciones seleccionadas en relación a la estrategia de análisis empleada en cada caso destacando los distintos puntos de control así como los valores que deben esperarse en cada uno de ellos. Los distintos protocolos de análisis obtenidos pretenden servir de guía para la correcta aplicación de la secuenciación masiva en estudios genéticos haciendo hincapié en las áreas de mayor conflicto y en la importancia de la ejecución de ciertos procesos de control que garanticen la obtención de unos mínimos criterios de calidad que confieran la seguridad que el Diagnóstico Genético exige.

1 Desarrollo y aplicación de herramientas bioinformáticas en estudios de resecuenciación dirigida

El análisis de los datos en estudios de resecuenciación dirigida puede dividirse en siete etapas principales cada una de las cuales merece una especial atención debido a las implicaciones que cada una de ellas tiene sobre el resultado final.

1.1 Diseño y evaluación de los kits de captura

Actualmente, existen diferentes sistemas de captura basados bien en hibridación mediante sondas o bien en amplificación mediante PCR de múltiples regiones en paralelo que permiten capturar zonas concretas del genoma. La presencia de repeticiones y zonas de baja complejidad en el genoma, unido a la existencia de pseudogenes y familias génicas que comparte un alto grado de homología entre sus miembros influyen en gran medida en el porcentaje de cobertura de los genes de interés. Los sistemas de captura basados en la hibridación de sondas son más sensibles a estas regiones conflictivas debido a la corta longitud de las sondas que emplean estas técnicas y por lo tanto dan lugar a menores porcentajes de captura por gen que los sistemas basados en PCR. Igualmente, estos sistemas de captura por hibridación de sondas ofrecen una heterogeneidad muy alta en la

cobertura con implicaciones importantes a la hora de identificar variantes y teniendo un mayor impacto en la identificación de CNVs. Por el contrario, los sistemas basados en PCR son capaces de cubrir zonas más amplias y por lo tanto no son tan sensibles a las zonas repetitivas o de baja complejidad proporcionando así una cobertura global del gen más óptima. En los sistemas personalizados de captura con los que se pretende llevar a cabo un diagnóstico, esta evaluación inicial del diseño, por su repercusión en el número de bases que no se pueden cubrir, es imprescindible. El estudio sobre enfermedades cardiacas presentado en esta tesis para el que se realizó un diseño a medida de un panel con objetivos diagnósticos muestra la importancia en la identificación de las zonas no cubiertas por gen reportando además el número de mutaciones descritas en estas zonas para así ofrecer una visión inicial sobre la eficiencia del sistema de captura y conocer así la posible repercusión que el uso de ese diseño pueda tener en el diagnóstico.

1.2 Control de calidad de los datos brutos de secuenciación

Al igual que ocurre en la secuenciación Sanger, en las plataformas de secuenciación masiva, cada base que se genera está asociada a un valor de calidad que refleja la probabilidad de que esa base haya sido identificada de forma incorrecta. El análisis global de los valores de calidad generados para cada base es fundamental para la identificación de desviaciones durante el proceso de secuenciación que pudieran tener consecuencias posteriores en la identificación de variantes. El software suministrado por el vendedor de la plataforma de NGS normalmente proporciona información sobre ciertos parámetros relacionados con las señales de intensidad producidas en cada ciclo, la deposición de las esferas en las placas de secuenciación y otros parámetros relacionados con el funcionamiento del sistema óptico o el sistema de fluidos. Sin embargo, este software no proporciona información suficiente para poder evaluar la calidad del dataset de forma integral e identificar el grado de magnitud de las posibles desviaciones observadas durante el proceso de secuenciación debido a que carece de un conocimiento experto del problema biológico que se trata de resolver en cada una de las aplicaciones posibles del secuenciador.

La evaluación del dataset atendiendo a sus valores de calidad permite determinar si la cantidad de datos y su calidad son suficientes para llevar a cabo el análisis o si por el contrario es necesario repetir su secuenciación o eliminar con anterioridad ciertos artefactos generados durante el proceso de generación de las librerías o durante el mismo proceso de secuenciación.

Actualmente existen diferentes herramientas para el análisis global de la calidad de los datos secuenciados y en la mayoría de los casos carecen de una completa visión sobre la problemática de la tecnología utilizada así como de unos gráficos que permita una rápida evaluación de los resultados obtenidos. El software FASTQC demostró ser la herramienta más completa, eficiente y fácilmente integrable para su ejecución de forma modular ofreciendo una serie de gráficos a través de los cuales fue posible llevar a cabo un control exhaustivo de las distintas desviaciones durante el proceso de generación de datos independientemente de la tecnología de secuenciación empleada o el tipo de librería generada.

1.3 Alineamiento de las lecturas

La comparación de cada una de las lecturas generadas frente a una determinada secuencia de referencia ha sido uno de los puntos más críticos en el paso de la secuenciación tradicional a la secuenciación masiva debido no solo al gran volumen de datos producidos sino también a la corta longitud de las lecturas [154]. Por otra parte, el genoma humano contiene secuencias repetitivas o con una alta homología entre sí de manera que es común encontrar lecturas que mapean con la misma fiabilidad en distintas zonas del genoma. La existencia de errores de secuenciación en las lecturas unidos a la presencia de variantes hacen que el número de lecturas que mapean en múltiples sitios incremente. La estrategia propuesta por Li et al en 2008 [85], plantea la asignación de probabilidades a cada posición de mapeo identificada en cada lectura de manera que tras obtener todas las posiciones genómicas posibles es posible decidir cual de ellas es más correcta en función del valor de probabilidad calculado por el algoritmo de alineamiento. De esta forma, se dice que una lectura es única si la segunda de las posiciones posibles contiene un mayor número de mismatches, una menor longitud de alineamiento o ambas a la vez, y por consiguiente un valor de mapeo más bajo que la mejor posición de mapeo encontrada. El uso de valores de calidad de mapeo ha sido adoptado por un número creciente de herramientas de análisis [155-158]. Sin embargo, cada uno de estos algoritmos procesa los datos de formas muy diferentes ofreciendo distintos parámetros de configuración que permiten realizar el alineamiento con mayor o menor eficiencia dependiendo de la tecnología de secuenciación y sus propiedades asociadas. La elección de una herramienta de mapeo eficiente, adaptada a cada plataforma de secuenciación, es de vital importancia ya que influye de forma sustancial en la precisión a la hora de detectar variantes. Conocer qué software de mapeo es el más robusto en cada caso es clave para concluir con éxito el análisis de los datos [82, 159].

El porcentaje de lecturas mapeables varía en gran medida dependiendo del software elegido, los parámetros de análisis seleccionados y el tipo de secuenciación (single-end o paired-end) llevado a cabo. La suite de herramientas de Bioscope incluye un programa de mapeo desarrollado específicamente para lecturas en colores que ofrece una alta tasa de lecturas mapeadas. Esta herramienta presenta una serie de parámetros configurables, optimizados para el análisis de las muestras estudiadas en esta tesis, a través de los cuales es posible la paralelización de la misma así como su optimización en un cluster de computación mediante el lanzamiento de los procesos a través de sistemas de colas de trabajos que permiten por lo tanto una mayor eficiencia en el proceso de mapeo. De forma similar, para el conjunto de datos analizados procedentes de la plataforma 454-Roche, se eligió el software SMALT cuyos parámetros están perfectamente optimizados para el alineamiento de lecturas largas.

Las herramientas de mapeo seleccionadas, pese a no ser las más eficientes a nivel computacional, proporcionaron unos resultados altamente sensibles tal como quedó demostrado posteriormente en la llamada de variantes. Además, estas herramientas ha evolucionado para adaptarse al formato SAM, actualmente considerado como el formato estándar para los alineamientos de datos de NGS, permitiendo así la integración de estos resultados con otra serie de utilidades más genéricas sin necesidad de depender de la plataforma de secuenciación empleada o del tipo de secuenciación realizado (‘single-end’ o ‘paired-end’) consiguiendo así una mayor flexibilidad.

1.4 Evaluación de la calidad de las librerías

Los diferentes pasos que tienen lugar durante la preparación de las muestras deben asegurar la recuperación de los fragmentos de DNA de la forma más homogénea posible minimizando las posibles desviaciones en la cobertura. Pese a todos los controles de calidad que se suceden a lo largo de la tediosa preparación de las muestras, es frecuente encontrar distintos artefactos derivados de ésta que necesitan ser eliminados o reducidos durante el análisis de los datos. Uno de los problemas más comunes en los datos procedentes de plataformas de secuenciación masiva es la generación de lecturas duplicadas, que proceden principalmente del proceso de amplificación por PCR necesario para obtener un número suficiente de copias del DNA molde para poder llevar a cabo la secuenciación [160]. Debido a los errores introducidos por la propia polimerasa durante la amplificación de las moléculas por PCR, un mayor número de lecturas duplicadas implica un mayor número de falsos positivos en la llamada de variantes y por lo tanto estas lecturas han de ser identificadas y eliminadas del dataset para la reducción de errores. En los

sistemas de hibridación por sondas, la eliminación de lecturas duplicadas provoca una reducción en la cantidad de lecturas útiles diferente en cada muestra y directamente dependiente del sistema de captura empleado. El porcentaje de lecturas útiles se ve menos afectado si la secuenciación es de tipo paired-end ya que para identificar las lecturas procedentes de duplicados de PCR es necesario que ambas lecturas compartan las coordenadas genómicas de mapeo de inicio y final mientras que si por el contrario la secuenciación es de tipo single-end solo se dispone de una única lectura y es mucho más probable que por azar su posición genómica coincida con otra secuencia pese a que procedan de fragmentos distintos de DNA. El porcentaje de lecturas duplicadas suele verse incrementado en muestras donde la cantidad de DNA de partida es baja siendo imprescindible llevar a cabo un mayor número de ciclos de amplificación que dan lugar a un mayor sesgo en la representación del genoma de la muestra. Algunos autores han demostrado la posibilidad de generar librerías sin necesidad de llevar a cabo la amplificación por PCR reduciéndose así el error derivado de la tecnología y la heterogeneidad en la cobertura [161]. No obstante, la cantidad de DNA de partida necesaria para llevar a cabo estos protocolos 'PCR free' es todavía muy alta. La adopción de estos protocolos 'PCR free' en la preparación de las muestras de resecuenciación dirigida proporcionará no solo un alivio en cuanto a la detección de falsos positivos en sistemas donde no es posible eliminar estas lecturas duplicadas, como el presentado en el estudio de los genes BRCA, sino también un aumento sustancial de la especificidad y sensibilidad del kit de captura al incrementarse el porcentaje de lecturas útiles.

Además del cálculo del porcentaje de lecturas duplicadas, tras el alineamiento de los datos es indispensable estimar la especificidad del kit de captura, parámetro mediante el cual es posible evaluar la eficiencia del diseño de los sondas. La presencia de genes con una alta homología a nivel de secuencia, como es el caso de familias génicas o de genes para los que existe un pseudogen, reduce los valores de especificidad del kit de captura debiendo por tanto aumentar la cantidad de lecturas necesarias de partida o inhabilitando el diagnóstico en el caso de que las regiones homólogas tengan un alto porcentaje de similitud y sean además zonas muy extensas. De la misma forma, para evaluar la eficiencia del kit de captura, se debe calcular su sensibilidad a diferentes profundidades de lectura con el fin de evaluar la cobertura global del panel.

Por último, la determinación del porcentaje de nucleótidos no cubiertos tras realizar la secuenciación es fundamental para determinar si existe una mayor pérdida de información en las regiones diana, además de la reportada por la ausencia de sondas en un pequeño porcentaje de estas regiones, que pueda tener repercusión en el diagnóstico.

Los pipelines diseñados toman en cuenta todos estos criterios para generar un dataset libre de elementos contaminantes que pudieran interferir en la llamada de las variantes. Del mismo modo, el protocolo de análisis desarrollado cuenta con una serie de scripts para el cálculo de las estadísticas de control necesarias para valorar la repercusión de cada uno de los puntos citados anteriormente proporcionando un análisis global e integral de la calidad de las librerías y la eficiencia del kit de captura.

1.5 Identificación de variantes puntuales y pequeños indels

El objetivo principal en estudios de resecuenciación es la identificación de variantes que puedan explicar total o parcialmente la aparición o el desarrollo de una determinada enfermedad. El creciente número de algoritmos para la llamada de variantes puede dar una falsa impresión sobre la complejidad subyacente a este paso del análisis siendo subestimada por clínicos o investigadores no expertos. Dado que no existe un "gold standard" para el análisis de los datos y que cada uno de estos protocolos que se desarrollan es único, el cálculo del porcentaje de incertidumbre que se obtiene tras el uso de la estrategia de análisis elegida es fundamental. La dificultad en la identificación de SNVs deriva fundamentalmente de tres factores: el error en la secuenciación, el alineamiento erróneo de las lecturas y la baja cobertura alcanzada en algunas zonas. Respecto al error de secuenciación y su implicación en la llamada de variantes es importante destacar que al igual que en la secuenciación tradicional, la calidad de las lecturas generadas por NGS también disminuye conforme aumenta la longitud de la lectura. Dependiendo de la calidad alcanzada tras el chequeo inicial del dato de secuenciación, es preferible por tanto bien eliminar una serie de nucleótidos concretos, como en el caso de las muestras control del panel de genes para la detección de enfermedades del corazón presentado en esta tesis, o bien eliminar aquellas lecturas con una calidad de secuenciación baja. Asimismo, dependiendo de la plataforma de secuenciación, existen diferentes desviaciones que ocurren en zonas concretas del genoma en las que existe un descenso de cobertura como las zonas con un alto contenido en GC en la plataforma SOLiD o las zonas de homopolímero en la plataforma 454-Roche donde aumenta tremendamente el error de secuenciación. Por otro lado, el alineamiento erróneo de algunas lecturas debido a su corta longitud unido a la tremenda complejidad del genoma humano pueden generar un mayor número de falsos positivos si no se es estricto en los parámetros de mapeo o en el posterior procesamiento de los ficheros de alineamiento eliminando la lecturas con baja calidad de mapeo. El último de los factores que afectan en mayor medida al proceso de identificación de las variantes es la profundidad de lectura. Las plataformas de secuenciación masiva

compensan la disminución en la longitud de lectura y su mayor frecuencia de error en comparación con la secuenciación Sanger con un incremento en el número de lecturas generadas de manera que cada posición que se pretende evaluar es leída varias veces [162, 163]. Los algoritmos diseñados para la identificación de variantes se basan en el análisis del conjunto de lecturas que cubren una determinada base a partir del cual son capaces de generar diferentes valores de probabilidad que pueden ser empleados para discernir entre errores de secuenciación y verdaderas variantes.

Actualmente, existe una gran diversidad de herramientas de análisis para la identificación de variantes que incluyen desde métodos basados en el conteo del número de lecturas discordantes entre la muestra y la secuencia de referencia junto con el uso de reglas sencillas que definen cuando determinar la presencia de un SNP y su genotipo hasta sistemas muy sofisticados que incluyen cálculos probabilísticos avanzados en los que se incorpora información sobre frecuencias alélicas poblacionales o patrones de desequilibrio de ligamiento poblacionales. Los resultados que producen los diferentes programas para la llamada de variantes pueden ser muy diversos pudiendo llegar a existir grandes diferencias entre ellos, siendo éstas más acusadas en zonas de baja profundidad de lectura [116]. Para evitar un mayor sesgo en la detección de variantes a bajas profundidades de lectura, y dada la enorme variabilidad en la cobertura que presentan los sistemas de hibridación por sondas, en muestras con fines diagnósticos es imprescindible saturar la muestra para alcanzar una profundidad de lectura media de al menos 150-200x, valor muy por encima de los valores de profundidad medios reportados en exoma, que se sitúan en torno a 50x, con el fin de asegurar una cobertura óptima en un mayor porcentaje de bases.

A pesar de que los pequeños indels sean abundantes en el genoma y generen una gran variabilidad entre individuos, han recibido una menor atención que las variantes puntuales o las grandes variaciones estructurales. Las pequeñas inserciones y deleciones han sido tradicionalmente difíciles de detectar y validar incluso mediante secuenciación Sanger [164-168]. Pese a que el número de indels depositados en las grandes bases de datos ha incrementado de forma exponencial durante los últimos años gracias a la llegada de las plataformas de NGS, estudios recientes sugieren que aún existe una proporción de falsos negativos alta [169] [170].

La detección de indels se ve afectada por los mismos factores que influyen en la detección de variantes puntuales, siendo el alineamiento de las lecturas el punto más crítico para su detección [171-173]. Desde el punto de vista computacional, el alineamiento de lecturas en zonas de indels tiene un coste mucho mayor que el mapeo en zonas donde se producen cambios puntuales debido a que para la detección de indels es necesario permitir huecos en

el alineamiento incrementando de forma sustancial las combinaciones posibles y el número de posiciones probables de mapeo. Si bien es cierto que los algoritmos para el alineamiento y detección automática de indels en datos de NGS han mejorado enormemente durante los últimos años, es importante destacar que los rangos de resolución para este tipo de variantes es limitado y raramente se han detectado indels con una longitud superior a 15pb en datos de TargetSeq [102].

En el caso de los datos de SOLiD, este trabajo demostró diferencias notables entre los resultados obtenidos dependiendo de si se dispone de lecturas 'paired-end' o 'single-end'. Estas discrepancias derivaron principalmente del funcionamiento de los softwares para la llamada de indels cuando manejan lecturas 'paired-end' o 'single-end'. Por otro lado, el realineamiento de las secuencias alrededor de los indels con GATK y la posterior aplicación de este algoritmo para la llamada de indels dio lugar a un incremento sustancial en el número total de indels identificados. La aplicación de estos cambios en el pipeline aumentó el número de indels totales un 34,85% en comparación con los resultados obtenidos en el primer pipeline.

Si bien es cierto que el aumento en la longitud de lectura facilita el mapeo y por tanto la detección de variantes, en el caso de la pirosecuenciación en la que la cuantificación de la luz emitida se emplea para determinar el número de bases repetidas, las regiones con un gran número de repeticiones de una única base no se resuelven bien, aumentando enormemente el error de secuenciación en estas zonas [153, 173, 174]. Pese a los esfuerzos realizados para reducir el número de potenciales falsos positivos en la llamada de indels, y a los prometedores resultados que se obtuvieron empleando un filtro dinámico, en el que dependiendo de la profundidad de lectura se seleccionaban las variantes en función de su frecuencia alélicas, en el estudio de los genes BRCA los resultados del análisis bioinformático tuvieron que acompañarse del análisis manual de los perfiles de los fragmentos procedentes de las reacciones de PCR en multiplex realizados en el laboratorio a fin de reducir el número de validaciones. Éste problema es tan agudo en algunas plataformas como 454 o plataformas incluso más reciente con el mismo problema en la detección de variantes en zonas de homopolímero como Ion Torrent, que se han desarrollado kits de laboratorio específicos para acompañar la secuenciación de las muestras de forma rutinaria y poder discriminar así los indels reales del alto número de indels detectados.

Actualmente, la identificación '*automatizada'* de indels sigue siendo un reto importante para la secuenciación masiva. Sin embargo, la enorme evolución de las herramientas de análisis

así como la reducción progresiva del error de secuenciación y la mejora en los algoritmos de mapeo, harán posible una detección de indels más eficiente en un futuro próximo.

La identificación de variantes es por tanto una tarea compleja en la que intervienen múltiples factores, desde su preparación y secuenciación hasta el análisis final de los datos, y donde se asume una cierta incertidumbre asociada a la combinación de una determinada plataforma de secuenciación con un sistema de captura concreto.

En el trabajo desarrollado en esta tesis, se utilizaron una serie de muestras controladas bien por tecnologías transversales como en el caso del exoma de la línea HapMap para la que se disponía de datos de genotipado masivo por arrays, o bien mediante secuenciación tradicional por Sanger. El primer pipeline desarrollado en exoma mostró una sensibilidad del 95,39% y una especificidad del 98,32% en la detección de SNVs, valores similares a los reportados por otros autores [175, 176]. Estos valores fueron mejorados en el segundo pipeline generado aplicado al panel de cardiomiopatías con la inclusión de un segundo programa para la identificación de SNVs y con el análisis en paralelo de la muestra tratada como “single-end” debido al incremento del error de secuenciación que presentaban las lecturas reversas en la plataforma SOLiD. Este nuevo pipeline para la llamada de variantes aumentó la sensibilidad en un 2,14% respecto a los resultados obtenidos en el pipeline inicial. Por el contrario, esta nueva aproximación combinada disminuyó la especificidad en un 4,16%, sin embargo, esta bajada en la especificidad se vio compensada con el aumento de variantes previamente no reportadas en el pipeline anterior, en total un 9,97% de los SNVs totales identificados demostrando así que la combinación de distintos programas para la llamada de variantes es la estrategia más adecuada para maximizar los valores de sensibilidad. Conclusiones similares fueron publicados posteriormente a la realización de este análisis en otros estudios [177] [178]. La aplicación de este nuevo pipeline en las muestras control del panel de cardio proporcionó un 100% de sensibilidad en los resultados. De la misma forma, en el análisis de los genes BRCA, el pipeline adaptado y los nuevos criterios de filtrado establecidos también reportaron un 100% de sensibilidad en las muestras control. Sin embargo, pese a los esfuerzos realizados para el abordaje de los problemas derivados del homopolímero, y para alcanzar el 100% de sensibilidad en las muestras control de una forma eficiente, fue necesaria la revisión manual del análisis de los fragmentos de las PCRs para identificar de forma unívoca los indels presentes en las muestras control y evitar así validar un alto número de variantes. Estrategias similares han sido reportadas por otro autores empleando kits comerciales diseñados para la plataforma 454-Roche con este propósito siendo ésta la única forma eficiente conocida actualmente para la identificación de indels en esta plataforma [179].

1.6 Anotación de las variantes

La anotación de las variantes implica una serie de procesos a través de los cuales se vincula la información técnica relativa a cada variante identificada como el número de lecturas, su valor de calidad, su frecuencia, etc., con información biológica que permita contextualizar las variantes. Este proceso debe integrar fuentes de información biológica adecuadas, organizadas y actualizadas que aporten la información necesaria para poder ayudar a definir clínicamente el tipo de variante ante la cual nos encontramos y su posible relación con la patología que presenta el paciente.

Dado el gran número de variantes obtenido en estudios de NGS, la automatización del proceso de anotación de las mismas juega un papel fundamental para su posterior filtrado e interpretación por un experto clínico. Ensembl es una de las mayores bases de datos de información 'ómica' que existen actualmente. Este recurso engloba una multitud de bases de datos como dbSNP, COSMIC, 1000Genomes, HapMap o ClinVar [90]. La posibilidad de generar scripts adaptados a partir de sus APIs para obtener la información biológica deseada hace de Ensembl uno de los recursos más completos y versátiles para enriquecer los resultados y abordar así el análisis clínico de las variantes identificadas.

El listado de los atributos obtenidos para cada variante como resultado final de la ejecución del pipeline presentado en esta memoria fueron establecidos y orientados por la Unidad de Genética Médica de Sistemas Genómicos ofreciendo un claro enfoque hacia la Medicina Genómica, visión de la cual carecen muchas de las herramientas comerciales o públicas generadas hasta la fecha.

2 Secuenciación del exoma completo y su aplicación en enfermedades raras

Actualmente la causa de ~3,000 enfermedades mendelianas todavía sigue siendo una incógnita. La identificación de nuevos genes candidatos responsables de su aparición en muchas de estas enfermedades es difícil de abordar mediante métodos tradicionales debido al escaso número de individuos que las padecen y al enorme coste que esto representa. Durante los últimos años, la investigación en enfermedades raras ha demostrado ser un recurso muy valioso para el estudio sobre los mecanismos moleculares que se desencadenan a partir de la mutación de un determinado gen. Dado que la mayoría las enfermedades mendelianas se desarrollan debido a la presencia de mutaciones patogénicas en regiones exónicas o en zonas de splicing, la posibilidad de secuenciar el exoma completo ha supuesto una auténtica revolución en su investigación provocando un cambio radical en el estudio de este tipo de enfermedades que se ve reflejado en el creciente número de publicaciones que hacen uso del exoma como herramienta fundamental para su estudio [180, 181] [46]

A diferencia de la secuenciación del genoma completo, la secuenciación de exomas permite caracterizar regiones concretas en un mayor número de individuos por el mismo coste. Sin embargo, existen ciertas limitaciones vinculadas al uso de sistemas de captura como la gran heterogeneidad en la captura, que difiere además entre kits de captura [38], y que representa un verdadero desafío en la identificación de variantes en zonas donde la profundidad de lectura disminuye. La mayoría de los estudios publicados de exoma han sido secuenciados a una media de profundidad de lectura de aproximadamente 30-50x, valor que deja entre un 10-20% de las regiones diana (entre 3,8 y 7,6Mb) cubiertas a menos de un 10x, profundidad de lectura muy pobre que compromete en gran medida la eficiencia en la detección de variantes [182] .

El número de variantes reportadas en el exoma por diferentes estudios publicados basados en el mismo sistema de captura y la misma plataforma de secuenciación en el momento en el que se realizó el análisis del exoma, variaban en gran medida, desde 14.000 a 24.000 [144] [145] [48] [146] [147]. Las diferencias en el uso de los distintos programas de alineamiento, de software para la llamada de variantes y el uso de estrategias de filtrado dispares son los principales factores atribuibles a estas diferencias que dejan entrever, en los casos más extremos, una alta tasa de falsos negativos y positivos. El aumento global de la profundidad de lectura media por muestra al doble o incluso triple de lo establecido en las estrategias estándar de secuenciación reduciría en gran medida el porcentaje de falsos

negativos [175], sin embargo, esta estrategia supondría un aumento significativo del coste total por muestra difícil de justificar frente al coste de secuenciación y el beneficio que representaría este tipo de secuenciación frente a la obtención de un genoma completo.

En esta tesis se presentan los resultados obtenidos tras la secuenciación del exoma completo de un único individuo que padecía Anemia de Fanconi (FA) cuyo diagnóstico seguía siendo una incógnita tras la secuenciación de varios de los genes actualmente relacionados con la enfermedad. Pese a no contar con un mayor número de individuos para el estudio, los estrictos sistemas de filtrado desarrollados en base a una línea celular de HapMap, así como la información biológica adicional suministrada para cada una de las 21,080 variantes identificadas dio como resultado un listado reducido de 17 genes candidatos bajo un modelo de herencia recesivo. De entre los genes de este listado, y dada la enfermedad en estudio, el grupo de investigación clínica seleccionó el gen ERCC4/XPF como gen candidato debido a su papel fundamental en la ruta de reparación del DNA y por su relación previa con dos enfermedades graves distintas, xeroderma pigmentoso y progeria segmental tipo XFE. Este hallazgo planteaba la hipótesis de que este mismo gen pudiera estar implicado también en el desarrollo de FA dependiendo de la zona del gen afectada. Las dos mutaciones encontradas en este gen, una deleción de 5pb en el exon 8 que provocaba un codón de parada prematuro y una mutación “missense” en el exon 11, fueron confirmadas tras los estudios de segregación correspondientes en los familiares mediante secuenciación Sanger. Adicionalmente, el grupo de investigación clínica realizó un screening en un grupo de pacientes alemanes con FA entre los que se detectó un individuo, no emparentado con el paciente previamente estudiado mediante el análisis del exoma completo, que presentaba dos mutaciones también en este gen. Tras llevarse a cabo un gran número de experimentos funcionales en el laboratorio se confirmó la hipótesis inicial planteada, dependiendo de la mutación producida en el gen XPF los pacientes pueden presentar Anemia de Fanconi, Xeroderma pigmentosa o Progeria tipo XFE [148]. Coincidiendo con la publicación de este trabajo, un nuevo síndrome fue relacionado con ERCC4, el síndrome de Cockaine [183], siendo por tanto cuatro las enfermedades relacionadas con mutaciones en este gen.

Las conclusiones alcanzadas por este trabajo permiten ahondar en el conocimiento sobre las bases genéticas del desarrollo embrionario, la hematopoyesis o la predisposición genética al cáncer a la vez que abren una nueva vía de esperanza al tratamiento de estas enfermedades.

Nuestros hallazgos soportan una vez más la idoneidad de esta tecnología en el estudio de enfermedades monogénicas incluso con un solo individuo. Sin embargo, es importante tener presente que pese a que la secuenciación del exoma completo constituya actualmente la forma más eficiente de realizar un rápido screening del conjunto de genes de un individuo, el grado de éxito de esta aplicación es todavía incierto ya que solamente salen a la luz aquellos estudios donde como en nuestro caso se identifican nuevos genes.

3 Resecuenciación dirigida y diagnóstico genético

Tradicionalmente, la combinación de sistemas de amplificación por PCR y su posterior secuenciación mediante tecnología Sanger han constituido el método principal de screening para el diagnóstico genético. Su elevado coste de secuenciación por base así como su bajo rendimiento han limitado su aplicación en el análisis de varios genes como método rutinario de diagnóstico siendo este el hándicap más importante en el diagnóstico de enfermedades monogénicas con heterogeneidad genética. La llegada de las plataformas de secuenciación masiva, unidas al desarrollo de sistemas de captura de regiones concretas del genoma mediante la generación de sistemas de enriquecimiento a medida han abierto una nueva perspectiva en el diagnóstico genético de este tipo de enfermedades [184]. La combinación de ambos protocolos, ofrece nuevos métodos de diagnóstico más resolutivos, de mayor alcance y sustancialmente más económicos en comparación con los sistemas tradicionales utilizados hasta la fecha con fines diagnósticos [185]. La generación de kits personalizados de captura para un grupo de enfermedades despierta gran interés por su descenso en el coste frente al uso del exoma completo. Debido a su menor tamaño, es posible secuenciar un mayor número de muestras empleando el mismo espacio de secuenciación a través del uso de sistemas de etiquetado por muestras (barcodes) que permiten una mayor optimización del espacio de secuenciación ya que, en el mismo espacio físico y al mismo tiempo, es posible procesar más de una muestra permitiendo así abaratar aún más los costes por muestra.

En la etiología de las enfermedades cardiovasculares se ha demostrado un claro componente hereditario y monogénico, que sigue los patrones mendelianos clásicos susceptibles de estudio por NGS debido a la alta heterogeneidad genética [67] .

En esta memoria se presentan los resultados obtenidos tras la aplicación de un pipeline integral para el análisis de un panel diseñado a medida para el diagnóstico de enfermedades cardiovasculares con riesgo de muerte súbita. En comparación con el exoma completo, la diferencia más notable y con mayor repercusión en el diagnóstico fue la profundidad de

lectura. En el caso del exoma, ~73% de las bases se encontraban cubiertas a valores de un 20x o superiores, alcanzándose una profundidad de lectura media de 50x. Por el contrario, en el panel de enfermedades cardiológicas, el 95% de las bases del panel se encontraron cubiertas a valores de 50x o superiores, siendo el valor medio de profundidad prácticamente cuatro veces superior al obtenido en exoma (197x). Los resultados muestran que el uso de paneles de genes frente al exoma con propósitos diagnósticos es más adecuado debido a la mayor profundidad de lectura alcanzada a lo largo de las regiones de interés, dotando a los paneles de una mayor fuerza estadística imprescindible para realizar una llamada de variantes más eficiente. Gracias a la introducción de nuevas herramientas de análisis en el pipeline generado para la llamada de variantes, se obtuvo un 100% de sensibilidad en el estudio llevado a cabo en un conjunto de 10 muestras control.

Siguiendo una serie de protolocos establecidos por diferentes autores y adaptados a las necesidades de nuestro centro [186], se llevó a cabo el estudio retrospectivo en un conjunto de 163 muestras pertenecientes a pacientes diagnosticados de diversas enfermedades cardiovasculares con riesgo de muerte súbita empleando este panel y la nueva estrategia de análisis desarrollada. Se identificaron mutaciones causales en el 27,6% de los casos, porcentaje muy superior al reportado por otros autores quienes establecen que solamente entre el 1% y el 2% de los casos son debidos a mutaciones monogénicas en canales iónicos cardiacos y/o proteínas asociadas [67]. En un 82% de los casos resueltos, el gen identificado se encontraba en el subpanel o grupo de genes perteneciente al diagnóstico cardiológico inicial. Sin embargo, en el 18% restante de los casos el gen causal pertenecía a un subpanel diferente dando buena cuenta de la ventaja de uso de este tipo de abordajes en el diagnóstico de enfermedades con una alta heterogeneidad genética como elemento modificador del diagnóstico cardiológico inicial.

La secuenciación masiva, sin lugar a dudas, permitirá mejorar el asesoramiento genético familiar al confirmar un mayor número de pacientes con diagnóstico cardiológico.

Actualmente, en nuestros laboratorios se emplea una versión posterior de este mismo panel, donde se aplica una estrategia de análisis similar, y al que le ha sido otorgado recientemente el marcado CE y el sello de calidad CLIA. Recientemente, el conjunto de paneles bajo el nombre comercial “GeneProfile”, diseñado y analizados siguiendo estrategias similares a las planteadas en esta tesis, han sido catalogados como el segundo producto más innovador en los premios “Life Science European Awards”.

4 Diagnóstico genético mediante el uso de mini-secuenciadores con tecnología NGS

Durante los dos últimos años, los mini-secuenciadores con tecnología NGS se han implantando rápidamente en la comunidad científica llevando la secuenciación masiva a prácticamente cualquier laboratorio independientemente de su tamaño. El desarrollo de sistemas de captura de zonas concretas del genoma ha impulsado la venta de estos equipos para la secuenciación de muestras con objetivos diagnósticos mediante el uso de paneles de genes de pequeño tamaño. Estas plataformas, presentan diferentes ventajas competitivas frente a los secuenciadores de mayor tamaño como el menor número de muestras necesarios para iniciar una carrera. Los secuenciadores de mayor tamaño producen una cantidad de datos muy superior a los mini-secuenciadores de manera que para completar un run los secuenciadores grandes necesitan o bien muestras en las que la cantidad de datos necesaria para el análisis sea grande, como es el caso de la secuenciación del exoma el genoma completo, o bien un alto volumen de muestras para que el run sea rentable. Por otro lado, los mini-secuenciadores con tecnología NGS ofrece tiempos de generación de datos inferiores a los alcanzados en plataformas de mayor tamaño, ventaja que consiguen mediante la mejora de sus sistemas ópticos y la optimización del espacio físico de secuenciación, pasando de tiempos de secuenciación de 1 a 2 semanas a tiempo de secuenciación de unas pocas horas [14] [187]. Otra de las ventajas de estos mini-secuenciadores respecto a los secuenciadores grandes es su coste de adquisición, que sin ser equipamientos económicos, supone alrededor de un tercio del coste de los secuenciadores de mayor tamaño siendo así más asequibles para un mayor número de usuarios.

En esta tesis se analizaron los genes *BRCA1* y *BRCA2* en una serie de muestras control mediante mini-secuenciadores con tecnología NGS y un sistema de captura basado en el uso de PCR en multiplex que, a diferencia de los sistemas basados en la hibridación por sondas como la tecnología SureSect utilizada en la captura del exoma completo y en el panel de cardiomiopatías, genera una cobertura más homogénea y está exento de pérdidas de cobertura debido a la inespecificidad de los primers empleados.

El análisis de las muestras control reportó un 100% de sensibilidad en la llamada de variantes. Sin embargo, el número de indels resultó especialmente alto debido al aumento en la tasa de error en zonas de homopolímero. Dado que en esta plataforma la automatización en la detección de indels mediante métodos computacionales es compleja, se abordó la identificación de los mismos con técnicas de laboratorio convencionales de

análisis de fragmentos mediante las que, tras su revisión manual y apoyándose en la información biológica obtenida mediante la anotación de las variantes, se consiguió un 100% de sensibilidad en el diagnóstico. Si bien es cierto que este tipo de alternativas es la única solución para reducir el listado de variantes, esta aproximación dista de la idoneidad ya que requiere una revisión manual de las variantes comparándolas con los resultados del análisis de fragmentos y por lo tanto es factible siempre y cuando el número de genes no sea muy alto. La gran demanda de secuenciadores de este tipo ha impulsado el desarrollo de otros mini-secuenciadores con tecnología NGS basados principios químicos distintos y que, pese a no estar exentos de otras desviaciones, carecen del problema asociado al homopolímero presentando así una menor tasa de error, secuenciadores que probablemente sustituirán en un futuro próximo a la pirosecuenciación.

La presencia de CNVs también fue investigada en esta tesis mediante un método basado en la normalización de la profundidad de lectura media por amplicón entre las muestras pertenecientes al mismo run. Pese a la presencia de ruido tras la normalización de los datos, las deleciones presentes en las muestras control pudieron ser diferenciadas mediante la estrategia de análisis propuesta.

A diferencia de paquetes de análisis ampliamente utilizados para el estudio de este tipo de datos como el software comercial AVA [www.roche.com], el protocolo de análisis obtenido no solo es capaz de proporcionar la información técnica relativa a la profundidad de lectura de la variante o la frecuencia de los alelos en cada posición sino que reporta además información adicional sobre el efecto de la variante a nivel transcripcional y proteico así como otra serie de información biológica complementaria que permite discriminar de entre el pool de variantes identificadas aquellas que pudieran ser causantes de la enfermedad en estudio.

Actualmente, se han analizado 447 casos en los laboratorios de Sistemas Genómicos siguiendo una aproximación donde se combinan los datos obtenidos por el pipeline desarrollado con las técnicas moleculares del análisis de fragmentos para definir los indels reales y la visualización de los datos mediante el software AVA e IGV que ha permitido identificar mutaciones en el 14,7% de los casos estudiados obteniendo valores ligeramente superiores al 10% de mutaciones hereditarias reportadas por otros estudios [188].

5 Perspectiva futura

La aparición de sistemas de captura de zonas específicas del genoma ha supuesto un paso significativo en la aplicación de las nuevas tecnologías de secuenciación al diagnóstico e investigación en Genética Humana. Durante los primeros años tras la aparición de las nuevas tecnologías de secuenciación hemos aprendido a manejar un volumen de datos tremendamente superior al ofrecido por las técnicas de secuenciación tradicionales. Tras superar progresivamente el desfase tecnológico entre el ritmo de desarrollo de las plataformas de NGS y el análisis bioinformático de sus datos, y una vez demostrado el alcance que esta nuevas tecnologías de secuenciación puede llegar a tener sobre la Salud Humana, se plantean nuevos desafíos científicos y analíticos para el establecimiento de unas buenas prácticas clínicas que permitan llevar la secuenciación masiva al diagnóstico genético de rutina. Actualmente, existe un vacío entre el ritmo de desarrollo de estas tecnologías y la estandarización de sus procesos, tanto a nivel de generación de los datos como en su análisis. La gran variabilidad de herramientas no validadas disponibles a esta fecha, tanto de libre acceso como paquete de análisis comerciales, ofrecen distintos grados de incertidumbre en los resultados que generalmente no son visibles o accesibles para el usuario final cuya comprensión requiere conocimientos avanzados del sistema biológico a analizar así como del significado de cada uno de los resultados parciales proporcionados de forma integrada, situación que convierte el análisis de los datos en uno de los puntos más relevantes para el éxito en la aplicación de la secuenciación masiva al diagnóstico. Gracias al exigente desarrollo bioinformático llevado a cabo durante los últimos años, actualmente se cuenta con una serie de estándares como el formato FASTQ para reportar el dato bruto, el formato SAM para describir los alineamientos o el formato VCF para listar las variantes encontradas, que permiten de forma “sencilla” comparar los resultados obtenidos tras la aplicación de diferentes pipeline. Sin embargo, la aplicación de la secuenciación masiva al diagnóstico genético precisa de manera urgente la generación de comités internacionales multidisciplinares que dictaminen unos estándares mínimos de calidad, que abarquen desde el procesamiento inicial de las muestras hasta la obtención de resultados tras el análisis de sus datos, así como el establecimiento de muestras o datasets de referencia que puedan ser empleados en las validaciones para demostrar la eficiencia del sistema de diagnóstico elegido [186].

La detección de variantes puntuales de forma automatizada en datos procedentes de estudios de resecuenciación dirigida mediante NGS presenta una baja tasa de falsos positivos y negativos. Sin embargo, la identificación automatizada de indels todavía sigue siendo un punto a mejorar en la aplicación de estas tecnologías al diagnóstico. El aumento

en la longitud de la lectura que prometen las nuevas versiones de las plataformas NGS así como la evolución de los programas de análisis para la detección automática de indels permitirán reducir de forma progresiva las limitaciones actuales. La evolución de las plataformas de secuenciación masiva de segunda y tercera generación permitirá finalmente alcanzar el hito más importante de esta nueva era genómica en la que nos encontramos inmersos, secuenciar un genoma por $1,000. La consecución de este hito en un futuro muy cercano resolverá en gran medida algunas de las limitaciones derivadas de los sistemas de captura, sin embargo, es importante no olvidar que la obtención de genomas por $1,000 es simplemente el inicio de una nueva época en la investigación genómica donde cientos de miles de individuos deberán ser secuenciados con el fin de entender concienzudamente las relaciones gen-enfermedad siendo solo posible mediante la generación de grupos multidisciplinares de profesionales altamente especializados cuyos análisis bien podrían multiplicar por 10 el precio de secuenciación [37]. Adicionalmente, la generación de un volumen tan grande de datos requerirá del desarrollo de nuevas infraestructuras computacionales adaptadas que permitan no solo realizar los cálculos necesarios para el análisis de estas muestras de forma eficiente sino también que ofrezcan nuevas alternativas para el almacenamiento de estos datos. Dado el enorme coste que este tipo de infraestructuras supone, además del gasto adicional que supone su mantenimiento y gestión, la computación de datos NGS en sistemas de tipo nube o “cloud computing”, que ofrecen un entorno flexible e ‘ilimitado’, es cada vez más popular [60] [189] [190, 191].

La secuenciación de genomas completos de forma rutinaria generará una gran cantidad de información sobre la variabilidad entre humanos provocando una auténtica explosión de información genómica que deberá ser almacenada de una forma estructurada y coherente para poder ser analizada de forma integral. Una de las iniciativas internacionales con mayor repercusión en este sentido es el proyecto del “varioma” (Human Variome Project), cuyo objetivo principal es generar un catálogo curado de las variantes descritas en el genoma humano integrando información genómica, fenotípica y farmacológica al mismo tiempo. Para lograr el éxito en la globalización genómica, el consorcio promueve el uso de sistemas de nomenclatura y definición de fenotipos estándar que aplica en la curación de datos ya depositados en las bases de datos más relevantes como OMIM o dbSNP, además de incentivar la deposición de las variantes encontradas por los diferentes laboratorios a nivel mundial en bases de datos públicas [192, 193].

Actualmente, la mayoría de las aproximaciones para la anotación de variantes se basan en métodos de asociación directos (efecto a nivel transcripcional o proteico, grado de conservación de las bases, etc), sin embargo, la complejidad del genoma humano es muy

basta y necesita de métodos de análisis más complejos capaces de estudiar la extensa información genómica generada por los usuarios de las plataformas NGS en un contexto global. La explotación de la información genómica producida por la comunidad científica necesita con urgencia el desarrollo de nuevas metodologías de análisis más globales que permitan inferir relaciones indirectas entre genes a través de la integración de diferentes fuentes de información biológica independientemente de su naturaleza. En un futuro próximo, el estudio global del individuo o del conjunto de individuos mediante estas nuevas metodologías de Biología de Sistemas permitirán identificar de forma indirecta la implicación de nuevas variantes en el desarrollo de la enfermedades comunes proporcionando importantes avances en el conocimiento de la relación gen-enfermedad favoreciendo así como una prognosis precoz, un diagnóstico más preciso o incluso un tratamiento personalizado.

CONCLUSIONES

- La correcta aplicación de la secuenciación masiva al diagnóstico genético depende en gran medida de la selección y combinación de las herramientas bioinformáticas y los parámetros de ejecución adecuados así como los criterios empleados para el filtrado, integración y presentación de la información.

- Los protocolos de análisis presentados en esta tesis nos han permitido abordar el estudio integral de los datos procedentes de la secuenciación del exoma completo, desde el control de calidad de los datos brutos hasta la obtención de un listado final de mutaciones y genes candidatos priorizados de una forma eficiente, sensible y robusta. Este pipeline inicial es actualmente empleado en Sistemas Genómicos con las actualizaciones periódicas pertinentes.

- El pipeline se optimizó para su aplicación al estudio de paneles de genes diseñados a medida para el diagnóstico de enfermedades genéticas, en concreto y en este trabajo, para el estudio de enfermedades cardiacas con riesgo de muerte súbita. Estos paneles de genes son de gran utilidad para el estudio de enfermedades con una alta heterogeneidad genética. En nuestro caso, se obtuvo una sensibilidad del 100% en la validación de muestras control. Paralelamente, el pipeline se adaptó también al análisis de los genes BRCA a través del uso de un mini-secuenciador con tecnología NGS basado en la tecnología de la pirosecuenciación que demostró ser eficiente y sensible en el 100% de las muestras tratadas. Sin embargo, para lograr esta sensibilidad los resultados 'in silicio' tuvieron que ser apoyados por técnicas de análisis de fragmentos convencionales debido al mayor error presente en las zonas de homopolímero asociado a esta plataforma de secuenciación.

- En los paneles con fines diagnósticos y con el propósito de minimizar la tasa de error y el porcentaje de bases no cubiertas, propiedades dependientes de la plataforma de secuenciación, el kit de captura empleado y el diseño específico del panel, es imprescindible secuenciar a profundidades de lectura que consigan saturar la muestra.

- Los protocolos de trabajo desarrollados en esta tesis fueron testados en muestras previamente caracterizadas por otras técnicas convencionales de manera que el porcentaje de incertidumbre asociado a la tecnología de secuenciación y captura pudo ser establecido de forma clara y transparente proporcionando una base sólida para su uso con fines diagnósticos tal como quedó demostrado en los tres datasets analizados.

BIBLIOGRAFÍA

1. Schuster, S.C., *Next-generation sequencing transforms today's biology.* Nat Methods, 2008. **5**(1): p. 16-8.
2. International Human Genome Sequencing, C., *Finishing the euchromatic sequence of the human genome.* Nature, 2004. **431**(7011): p. 931-45.
3. International HapMap, C., *A haplotype map of the human genome.* Nature, 2005. **437**(7063): p. 1299-320.
4. Mardis, E.R., *Anticipating the 1,000 dollar genome.* Genome Biol, 2006. **7**(7): p. 112.
5. Service, R.F., *Gene sequencing. The race for the $1000 genome.* Science, 2006. **311**(5767): p. 1544-6.
6. Metzker, M.L., *Sequencing technologies - the next generation.* Nat Rev Genet, 2010. **11**(1): p. 31-46.
7. Voelkerding, K.V., S.A. Dames, and J.D. Durtschi, *Next-generation sequencing: from basic research to diagnostics.* Clin Chem, 2009. **55**(4): p. 641-58.
8. Rusk, N. and V. Kiermer, *Primer: Sequencing--the next generation.* Nat Methods, 2008. **5**(1): p. 15.
9. Margulies, M., et al., *Genome sequencing in microfabricated high-density picolitre reactors.* Nature, 2005. **437**(7057): p. 376-80.
10. Ronaghi, M., *Pyrosequencing sheds light on DNA sequencing.* Genome Res, 2001. **11**(1): p. 3-11.
11. *Prepare for the deluge.* Nat Biotechnol, 2008. **26**(10): p. 1099.
12. Bentley, D.R., et al., *Accurate whole human genome sequencing using reversible terminator chemistry.* Nature, 2008. **456**(7218): p. 53-9.
13. McKernan, K.J., et al., *Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding.* Genome Res, 2009. **19**(9): p. 1527-41.
14. Loman, N.J., et al., *Performance comparison of benchtop high-throughput sequencing platforms.* Nat Biotechnol, 2012. **30**(5): p. 434-9.
15. Schadt, E.E., S. Turner, and A. Kasarskis, *A window into third-generation sequencing.* Hum Mol Genet, 2010. **19**(R2): p. R227-40.
16. Rusk, N., *Cheap third-generation sequencing.* Nat Methods, 2009. **6**(244).
17. Eid, J., et al., *Real-time DNA sequencing from single polymerase molecules.* Science, 2009. **323**(5910): p. 133-8.
18. Schneider, G.F. and C. Dekker, *DNA sequencing with nanopores.* Nat Biotechnol, 2012. **30**(4): p. 326-8.
19. Mardis, E.R., *Next-generation DNA sequencing methods.* Annu Rev Genomics Hum Genet, 2008. **9**: p. 387-402.
20. *The human genome at ten.* Nature, 2010. **464**(7289): p. 649-50.
21. Check Hayden, E., *Human genome at ten: Life is complicated.* Nature, 2010. **464**(7289): p. 664-7.
22. Mardis, E.R., *A decade's perspective on DNA sequencing technology.* Nature, 2011. **470**(7333): p. 198-203.
23. Lander, E.S., et al., *Initial sequencing and analysis of the human genome.* Nature, 2001. **409**(6822): p. 860-921.
24. International HapMap, C., et al., *Integrating common and rare genetic variation in diverse human populations.* Nature, 2010. **467**(7311): p. 52-8.
25. International HapMap, C., et al., *A second generation human haplotype map of over 3.1 million SNPs.* Nature, 2007. **449**(7164): p. 851-61.
26. *Web del Consorcio Internacional de HapMap*. Available from: www.hapmap.org.

27. International HapMap, C., *The International HapMap Project.* Nature, 2003. **426**(6968): p. 789-96.
28. *Web del Consorcio 1000Genomes*. Available from: www.1000genomes.org.
29. Genomes Project, C., et al., *A map of human genome variation from population-scale sequencing.* Nature, 2010. **467**(7319): p. 1061-73.
30. International Cancer Genome, C., et al., *International network of cancer genome projects.* Nature, 2010. **464**(7291): p. 993-8.
31. Cancer Genome Atlas Research, N., *Comprehensive genomic characterization defines human glioblastoma genes and core pathways.* Nature, 2008. **455**(7216): p. 1061-8.
32. Consortium, E.P., *The ENCODE (ENCyclopedia Of DNA Elements) Project.* Science, 2004. **306**(5696): p. 636-40.
33. *Base de datos dbSNP*. Available from: http://www.ncbi.nlm.nih.gov/SNP/.
34. Collins, F., *Has the revolution arrived?* Nature, 2010. **464**(7289): p. 674-5.
35. Venter, J.C., *Multiple personal genomes await.* Nature, 2010. **464**(7289): p. 676-7.
36. Tucker, T., M. Marra, and J.M. Friedman, *Massively parallel sequencing: the next big thing in genetic medicine.* Am J Hum Genet, 2009. **85**(2): p. 142-54.
37. Mardis, E.R., *The $1,000 genome, the $100,000 analysis?* Genome Med, 2010. **2**(11): p. 84.
38. Clark, M.J., et al., *Performance comparison of exome DNA sequencing technologies.* Nat Biotechnol, 2011. **29**(10): p. 908-14.
39. Mamanova, L., et al., *Target-enrichment strategies for next-generation sequencing.* Nat Methods, 2010. **7**(2): p. 111-8.
40. Summerer, D., et al., *Microarray-based multicycle-enrichment of genomic subsets for targeted next-generation sequencing.* Genome Res, 2009. **19**(9): p. 1616-21.
41. Gnirke, A., et al., *Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing.* Nat Biotechnol, 2009. **27**(2): p. 182-9.
42. Meuzelaar, L.S., et al., *MegaPlex PCR: a strategy for multiplex amplification.* Nat Methods, 2007. **4**(10): p. 835-7.
43. Turner, E.H., et al., *Massively parallel exon capture and library-free resequencing across 16 genomes.* Nat Methods, 2009. **6**(5): p. 315-6.
44. Antonarakis, S.E. and J.S. Beckmann, *Mendelian disorders deserve more attention.* Nat Rev Genet, 2006. **7**(4): p. 277-82.
45. Gilissen, C., et al., *Unlocking Mendelian disease using exome sequencing.* Genome Biol, 2011. **12**(9): p. 228.
46. Maxmen, A., *Exome sequencing deciphers rare diseases.* Cell, 2011. **144**(5): p. 635-7.
47. Ng, S.B., et al., *Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome.* Nat Genet, 2010. **42**(9): p. 790-3.
48. Gilissen, C., et al., *Exome sequencing identifies WDR35 variants involved in Sensenbrenner syndrome.* Am J Hum Genet, 2010. **87**(3): p. 418-23.
49. Ng, S.B., et al., *Exome sequencing identifies the cause of a mendelian disorder.* Nat Genet, 2010. **42**(1): p. 30-5.
50. Chen, W.J., et al., *Exome sequencing identifies truncating mutations in PRRT2 that cause paroxysmal kinesigenic dyskinesia.* Nat Genet, 2011. **43**(12): p. 1252-5.
51. Yan, X.J., et al., *Exome sequencing identifies somatic mutations of DNA methyltransferase gene DNMT3A in acute monocytic leukemia.* Nat Genet, 2011. **43**(4): p. 309-15.
52. Pugh, T.J., et al., *Medulloblastoma exome sequencing uncovers subtype-specific somatic mutations.* Nature, 2012. **488**(7409): p. 106-10.
53. Quesada, V., et al., *Exome sequencing identifies recurrent mutations of the splicing factor SF3B1 gene in chronic lymphocytic leukemia.* Nat Genet, 2012. **44**(1): p. 47-52.

54. Krauthammer, M., et al., *Exome sequencing identifies recurrent somatic RAC1 mutations in melanoma.* Nat Genet, 2012. **44**(9): p. 1006-14.
55. Shearer, A.E., et al., *Deafness in the genomics era.* Hear Res, 2011. **282**(1-2): p. 1-9.
56. Zaragoza, M.V., et al., *Mitochondrial DNA variant discovery and evaluation in human Cardiomyopathies through next-generation sequencing.* PLoS One, 2010. **5**(8): p. e12295.
57. Brion, M., et al., *New technologies in the genetic approach to sudden cardiac death in the young.* Forensic Sci Int, 2010. **203**(1-3): p. 15-24.
58. Haas, J., H.A. Katus, and B. Meder, *Next-generation sequencing entering the clinical arena.* Mol Cell Probes, 2011. **25**(5-6): p. 206-11.
59. McPherson, J.D., *Next-generation gap.* Nat Methods, 2009. **6**(11 Suppl): p. S2-5.
60. Baker, M., *Next-generation sequencing: adjusting to data overload.* Nat Methods, 2010. **7**: p. 495-499.
61. Koboldt, D.C., et al., *Challenges of sequencing human genomes.* Brief Bioinform, 2010. **11**(5): p. 484-98.
62. Development, C.o.A.R.D.R.a.O.P., *Rare Diseases and Orphan Products: Accelerating Research and Development*, ed. N.A. Press. 2010.
63. Ng, S.B., et al., *Massively parallel sequencing and rare disease.* Hum Mol Genet, 2010. **19**(R2): p. R119-24.
64. Adabag, A.S., et al., *Sudden cardiac death: epidemiology and risk factors.* Nat Rev Cardiol, 2010. **7**(4): p. 216-25.
65. Benito Morentina, C.A., *Estudio poblacional de la muerte súbita cardiovascular extrahospitalaria: incidencia y causas de muerte en adultos de edad mediana.* Rev Esp Cardiol, 2011. **64**(01): p. 28-34.
66. Estes, N.A., 3rd, *Predicting and preventing sudden cardiac death.* Circulation, 2011. **124**(5): p. 651-6.
67. Chopra, N. and B.C. Knollmann, *Genetics of sudden cardiac death syndromes.* Curr Opin Cardiol, 2011. **26**(3): p. 196-203.
68. Bagnall, R.D., J. Ingles, and C. Semsarian, *Molecular diagnostics of cardiomyopathies: the future is here.* Circ Cardiovasc Genet, 2011. **4**(2): p. 103-4.
69. Meder, B., et al., *Targeted next-generation sequencing for the molecular genetic diagnostics of cardiomyopathies.* Circ Cardiovasc Genet, 2011. **4**(2): p. 110-22.
70. Ramus, S.J., et al., *Ovarian cancer susceptibility alleles and risk of ovarian cancer in BRCA1 and BRCA2 mutation carriers.* Hum Mutat, 2012. **33**(4): p. 690-702.
71. Lynch, H.T., et al., *Hereditary breast cancer: part I. Diagnosing hereditary breast cancer syndromes.* Breast J, 2008. **14**(1): p. 3-13.
72. Campeau, P.M., W.D. Foulkes, and M.D. Tischkowitz, *Hereditary breast cancer: new genetic developments, new therapeutic avenues.* Hum Genet, 2008. **124**(1): p. 31-42.
73. Liu, W., et al., *Denaturing high performance liquid chromatography (DHPLC) used in the detection of germline and somatic mutations.* Nucleic Acids Res, 1998. **26**(6): p. 1396-400.
74. De Leeneer, K., et al., *Rapid and sensitive detection of BRCA1/2 mutations in a diagnostic setting: comparison of two high-resolution melting platforms.* Clin Chem, 2008. **54**(6): p. 982-9.
75. De Leeneer, K., et al., *Massive parallel amplicon sequencing of the breast cancer genes BRCA1 and BRCA2: opportunities, challenges, and limitations.* Hum Mutat, 2011. **32**(3): p. 335-44.
76. *Web de Agilent - SureSelect*. Available from: www.genomics.agilent.com.
77. Williams, R., et al., *Amplification of complex gene libraries by emulsion PCR.* Nat Methods, 2006. **3**(7): p. 545-50.
78. Janitz, M., *Applied Biosystems SOLiD™ System: Ligation-Based Sequencing*, in *Next Generation Genome Sequencing: Towards Personalized Medicine*. 2008, Wiley-VCH Verlag GmBH & Co. KGaA.

79. *Web oficial de LifeTechnologies para su secuenciador SOLiD* Available from: http://www.appliedbiosystems.com/absite/us/en/home/applications-technologies/solid-next-generation-sequencing.html.
80. Ewing, B. and P. Green, *Base-calling of automated sequencer traces using phred. II. Error probabilities.* Genome Res, 1998. **8**(3): p. 186-94.
81. Ewing, B., et al., *Base-calling of automated sequencer traces using phred. I. Accuracy assessment.* Genome Res, 1998. **8**(3): p. 175-85.
82. Li, H. and N. Homer, *A survey of sequence alignment algorithms for next-generation sequencing.* Brief Bioinform, 2010. **11**(5): p. 473-83.
83. Trapnell, C. and S.L. Salzberg, *How to map billions of short reads onto genomes.* Nat Biotechnol, 2009. **27**(5): p. 455-7.
84. Richter, B.G. and D.P. Sexton, *Managing and analyzing next-generation sequence data.* PLoS Comput Biol, 2009. **5**(6): p. e1000369.
85. Li, H., J. Ruan, and R. Durbin, *Mapping short DNA sequencing reads and calling variants using mapping quality scores.* Genome Res, 2008. **18**(11): p. 1851-8.
86. Li, H., et al., *The Sequence Alignment/Map format and SAMtools.* Bioinformatics, 2009. **25**(16): p. 2078-9.
87. *Web de Picard Tools*. Available from: http://picard.sourceforge.net.
88. Thorvaldsdottir, H., J.T. Robinson, and J.P. Mesirov, *Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration.* Brief Bioinform, 2013. **14**(2): p. 178-92.
89. Hubbard, T., et al., *The Ensembl genome database project.* Nucleic Acids Res, 2002. **30**(1): p. 38-41.
90. Flicek, P., et al., *Ensembl 2013.* Nucleic Acids Res, 2013. **41**(Database issue): p. D48-55.
91. Sherry, S.T., et al., *dbSNP: the NCBI database of genetic variation.* Nucleic Acids Res, 2001. **29**(1): p. 308-11.
92. Chen, Y., et al., *Ensembl variation resources.* BMC Genomics, 2010. **11**: p. 293.
93. Cooper, D.N. and M. Krawczak, *Human Gene Mutation Database.* Hum Genet, 1996. **98**(5): p. 629.
94. Stenson, P.D., et al., *Human Gene Mutation Database: towards a comprehensive central mutation database.* J Med Genet, 2008. **45**(2): p. 124-6.
95. Genomes Project, C., et al., *An integrated map of genetic variation from 1,092 human genomes.* Nature, 2012. **491**(7422): p. 56-65.
96. McLaren, W., et al., *Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor.* Bioinformatics, 2010. **26**(16): p. 2069-70.
97. Rios, D., et al., *A database and API for variation, dense genotyping and resequencing data.* BMC Bioinformatics, 2010. **11**: p. 238.
98. *Web de Ensembl versión 59*. Available from: http://aug2010.archive.ensembl.org/index.html.
99. Ng, P.C. and S. Henikoff, *SIFT: Predicting amino acid changes that affect protein function.* Nucleic Acids Res, 2003. **31**(13): p. 3812-4.
100. Cooper, G.M., et al., *Distribution and intensity of constraint in mammalian genomic sequence.* Genome Res, 2005. **15**(7): p. 901-13.
101. *Datos de genotipado masivo por arrays del consorcio de HapMap*. Available from: ftp://ftp.ncbi.nlm.nih.gov/hapmap/phase_3/hapmap3_reformatted.
102. Ng, S.B., et al., *Targeted capture and massively parallel sequencing of 12 human exomes.* Nature, 2009. **461**(7261): p. 272-6.
103. *Base de datos de Online Mendelian Inheritance in Man (OMIM) - NCBI*. Available from: http://omim.org/.
104. *PubMed - NCBI.*
105. *Base de datos Gene - NCBI*. Available from: http://www.ncbi.nlm.nih.gov/gene/.
106. *Base de datos HGMD*. Available from: www.hgmd.org.

107. *Base de datos de Ensembl*. Available from: www.ensembl.org.
108. *Base de datos GeneCards*. Available from: www.genecards.org.
109. *Base de datos UniProt*. Available from: http://www.uniprot.org/.
110. *Genatlas Universite Paris Descartes*. Available from: http://genatlas.medecine.univ-paris5.fr/.
111. Gilbert, D., *Shopping in the genome market with EnsMart.* Brief Bioinform, 2003. **4**(3): p. 292-6.
112. *Web del programa de diseño de sondas eArray de Agilent*. Available from: https://earray.chem.agilent.com/earray/)
113. *FASTQC*. Available from: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.
114. Cock, P.J., et al., *The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants.* Nucleic Acids Res, 2010. **38**(6): p. 1767-71.
115. McKenna, A., et al., *The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data.* Genome Res, 2010. **20**(9): p. 1297-303.
116. Nielsen, R., et al., *Genotype and SNP calling from next-generation sequencing data.* Nat Rev Genet, 2011. **12**(6): p. 443-51.
117. Albers, C.A., et al., *Dindel: accurate indel calls from short-read data.* Genome Res, 2011. **21**(6): p. 961-73.
118. DePristo, M.A., et al., *A framework for variation discovery and genotyping using next-generation DNA sequencing data.* Nat Genet, 2011. **43**(5): p. 491-8.
119. Danecek, P., et al., *The variant call format and VCFtools.* Bioinformatics, 2011. **27**(15): p. 2156-8.
120. Consortium, G. *http://www.1000genomes.org*. 2010; 1000 Genomes Consortium website].
121. *Ensembl versión 62*. Available from: http://apr2011.archive.ensembl.org/index.html.
122. *Web de la sociedad HGVS*. Available from: www.hgvs.org.
123. Gonzalez-Perez, A. and N. Lopez-Bigas, *Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel.* Am J Hum Genet, 2011. **88**(4): p. 440-9.
124. *Ensembl 64*. Available from: http://sep2011.archive.ensembl.org/index.html.
125. Adzhubei, I.A., et al., *A method and server for predicting damaging missense mutations.* Nat Methods, 2010. **7**(4): p. 248-9.
126. Grantham, R., *Amino acid difference formula to help explain protein evolution.* Science, 1974. **185**(4154): p. 862-4.
127. *Ensembl versión 71*. Available from: http://apr2013.archive.ensembl.org/index.html.
128. *Web del software Alamut*. Available from: http://www.interactive-biosoftware.com/alamut-visual/.
129. *Kit de captura BRCA MASTR de Multiplicom*. Available from: http://www.multiplicom.com/products/brca-mastr-dx.
130. *GSJunior*. Available from: www.gsjunior.com.
131. *Definición del formato SFF*. Available from: http://www.ncbi.nlm.nih.gov/Traces/trace.cgi?cmd=show&f=formats&m=doc&s=format#sff
132. Martin, M., *Cutadapt removes adapter sequences from high-throughput sequencing reads.* EMBnet.journal, 2011. **17**(1): p. 10-12.
133. Schmieder, R. and R. Edwards, *Quality control and preprocessing of metagenomic datasets.* Bioinformatics, 2011. **27**(6): p. 863-4.
134. *Sitio web del algoritmo SMALT*. Available from: http://www.sanger.ac.uk/resources/software/smalt/.
135. Ning, Z., A.J. Cox, and J.C. Mullikin, *SSAHA: a fast search method for large DNA databases.* Genome Res, 2001. **11**(10): p. 1725-9.

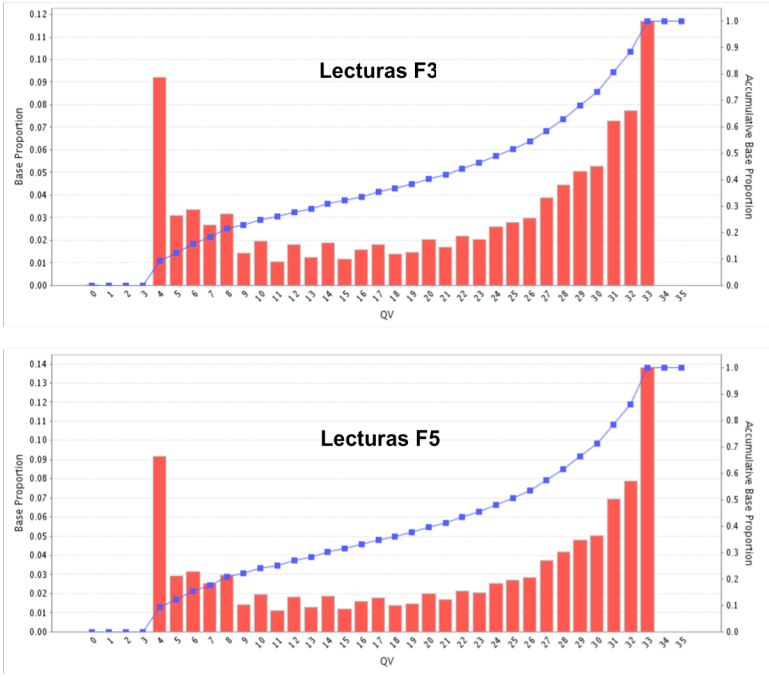
136. Langmead, B. and S.L. Salzberg, *Fast gapped-read alignment with Bowtie 2.* Nat Methods, 2012. **9**(4): p. 357-9.
137. Li, H. and R. Durbin, *Fast and accurate long-read alignment with Burrows-Wheeler transform.* Bioinformatics, 2010. **26**(5): p. 589-95.
138. Koboldt, D.C., et al., *VarScan: variant detection in massively parallel sequencing of individual and pooled samples.* Bioinformatics, 2009. **25**(17): p. 2283-5.
139. Koboldt, D.C., et al., *VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing.* Genome Res, 2012. **22**(3): p. 568-76.
140. Hunter, S., et al., *InterPro: the integrative protein signature database.* Nucleic Acids Res, 2009. **37**(Database issue): p. D211-5.
141. Kim Pruitt, G.B., Tatiana Tatusova, and Donna Maglott, *The NCBI handbook*, in *The NCBI handbook*, O.J. McEntyre J, Editor. 2002: Bethesda (MD): National Center for Biotechnology Information (US).
142. Pruitt, K.D., et al., *The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes.* Genome Res, 2009. **19**(7): p. 1316-23.
143. *Definición del formato Standard Flowgram Format (SFF)*. Available from: http://www.ncbi.nlm.nih.gov/Traces/trace.cgi?cmd=show&f=formats&m=doc&s=format#sff.
144. Vissers, L.E., et al., *A de novo paradigm for mental retardation.* Nat Genet, 2010. **42**(12): p. 1109-12.
145. Hoischen, A., et al., *De novo mutations of SETBP1 cause Schinzel-Giedion syndrome.* Nat Genet, 2010. **42**(6): p. 483-5.
146. Haack, T.B., et al., *Exome sequencing identifies ACAD9 mutations as a cause of complex I deficiency.* Nat Genet, 2010. **42**(12): p. 1131-4.
147. Krawitz, P.M., et al., *Identity-by-descent filtering of exome sequence data identifies PIGV mutations in hyperphosphatasia mental retardation syndrome.* Nat Genet, 2010. **42**(10): p. 827-9.
148. Bogliolo, M., et al., *Mutations in ERCC4, encoding the DNA-repair endonuclease XPF, cause Fanconi anemia.* Am J Hum Genet, 2013. **92**(5): p. 800-6.
149. Mignone, F., et al., *Untranslated regions of mRNAs.* Genome Biol, 2002. **3**(3): p. REVIEWS0004.
150. *Programa RepeatMasker*. Available from: http://www.repeatmasker.org.
151. Benson, G., *Tandem repeats finder: a program to analyze DNA sequences.* Nucleic Acids Res, 1999. **27**(2): p. 573-80.
152. Craig, D.W., et al., *Identification of genetic variants using bar-coded multiplexed sequencing.* Nat Methods, 2008. **5**(10): p. 887-93.
153. De Leeneer, K., et al., *Practical tools to implement massive parallel pyrosequencing of PCR products in next generation molecular diagnostics.* PLoS One, 2011. **6**(9): p. e25531.
154. Flicek, P. and E. Birney, *Sense from sequence reads: methods for alignment and assembly.* Nat Methods, 2009. **6**(11 Suppl): p. S6-S12.
155. Langmead, B., et al., *Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.* Genome Biol, 2009. **10**(3): p. R25.
156. Li, H. and R. Durbin, *Fast and accurate short read alignment with Burrows-Wheeler transform.* Bioinformatics, 2009. **25**(14): p. 1754-60.
157. Li, R., et al., *SOAP2: an improved ultrafast tool for short read alignment.* Bioinformatics, 2009. **25**(15): p. 1966-7.
158. Homer, N., B. Merriman, and S.F. Nelson, *BFAST: an alignment tool for large scale genome resequencing.* PLoS One, 2009. **4**(11): p. e7767.
159. Ruffalo, M., T. LaFramboise, and M. Koyuturk, *Comparative analysis of algorithms for next-generation sequencing read alignment.* Bioinformatics, 2011. **27**(20): p. 2790-6.

160. Aird, D., et al., *Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries.* Genome Biol, 2011. **12**(2): p. R18.
161. Kozarewa, I., et al., *Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes.* Nat Methods, 2009. **6**(4): p. 291-5.
162. Li, Y., et al., *Low-coverage sequencing: implications for design of complex trait association studies.* Genome Res, 2011. **21**(6): p. 940-51.
163. Stitziel, N.O., A. Kiezun, and S. Sunyaev, *Computational and statistical approaches to analyzing variants identified by exome sequencing.* Genome Biol, 2011. **12**(9): p. 227.
164. Cartwright, R.A., *Problems and solutions for estimating indel rates and length distributions.* Mol Biol Evol, 2009. **26**(2): p. 473-80.
165. Mullaney, J.M., et al., *Small insertions and deletions (INDELs) in human genomes.* Hum Mol Genet, 2010. **19**(R2): p. R131-6.
166. Volfovsky, N., et al., *Genome and gene alterations by insertions and deletions in the evolution of human and chimpanzee chromosome 22.* BMC Genomics, 2009. **10**: p. 51.
167. Bhangale, T.R., et al., *Comprehensive identification and characterization of diallelic insertion-deletion polymorphisms in 330 human candidate genes.* Hum Mol Genet, 2005. **14**(1): p. 59-69.
168. Mills, R.E., et al., *An initial map of insertion and deletion (INDEL) variation in the human genome.* Genome Res, 2006. **16**(9): p. 1182-90.
169. Mills, R.E., et al., *Mapping copy number variation by population-scale genome sequencing.* Nature, 2011. **470**(7332): p. 59-65.
170. Kidd, J.M., et al., *A human genome structural variation sequencing resource reveals insights into mutational mechanisms.* Cell, 2010. **143**(5): p. 837-47.
171. Krawitz, P., et al., *Microindel detection in short-read sequence data.* Bioinformatics, 2010. **26**(6): p. 722-9.
172. Vali, U., et al., *Insertion-deletion polymorphisms (indels) as genetic markers in natural populations.* BMC Genet, 2008. **9**: p. 8.
173. Neuman, J.A., O. Isakov, and N. Shomron, *Analysis of insertion-deletion from deep-sequencing data: software evaluation for optimal detection.* Brief Bioinform, 2013. **14**(1): p. 46-55.
174. Huse, S.M., et al., *Accuracy and quality of massively parallel DNA pyrosequencing.* Genome Biol, 2007. **8**(7): p. R143.
175. Bonnefond, A., et al., *Molecular diagnosis of neonatal diabetes mellitus using next-generation sequencing of the whole exome.* PLoS One, 2010. **5**(10): p. e13630.
176. Tewhey, R., et al., *Enrichment of sequencing targets from the human genome by solution hybridization.* Genome Biol, 2009. **10**(10): p. R116.
177. O'Rawe, J., et al., *Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing.* Genome Med, 2013. **5**(3): p. 28.
178. Liu, X., et al., *Variant callers for next-generation sequencing data: a comparison study.* PLoS One, 2013. **8**(9): p. e75619.
179. Feliubadalo, L., et al., *Next-generation sequencing meets genetic diagnostics: development of a comprehensive workflow for the analysis of BRCA1 and BRCA2 genes.* Eur J Hum Genet, 2013. **21**(8): p. 864-70.
180. Ku, C.S., N. Naidoo, and Y. Pawitan, *Revisiting Mendelian disorders through exome sequencing.* Hum Genet, 2011. **129**(4): p. 351-70.
181. Bamshad, M.J., et al., *Exome sequencing as a tool for Mendelian disease gene discovery.* Nat Rev Genet, 2011. **12**(11): p. 745-55.
182. Kiezun, A., et al., *Exome sequencing and the genetic basis of complex traits.* Nat Genet, 2012. **44**(6): p. 623-30.

183. Kashiyama, K., et al., *Malfunction of nuclease ERCC1-XPF results in diverse clinical manifestations and causes Cockayne syndrome, xeroderma pigmentosum, and Fanconi anemia.* Am J Hum Genet, 2013. **92**(5): p. 807-19.
184. Raju, H. and E.R. Behr, *Unexplained sudden death, focussing on genetics and family phenotyping.* Curr Opin Cardiol, 2013. **28**(1): p. 19-25.
185. Su, Z., et al., *Next-generation sequencing and its applications in molecular diagnostics.* Expert Rev Mol Diagn, 2011. **11**(3): p. 333-43.
186. Gargis, A.S., et al., *Assuring the quality of next-generation sequencing in clinical laboratory practice.* Nat Biotechnol, 2012. **30**(11): p. 1033-6.
187. Junemann, S., et al., *Updating benchtop sequencing performance comparison.* Nat Biotechnol, 2013. **31**(4): p. 294-6.
188. Nancie Petrucelli, M., Mary B Daly, MD, PhD, and Gerald L Feldman, MD, PhD, FACMG., *BRCA1 and BRCA2 Hereditary Breast and Ovarian Cancer*, in *GeneReviews*2013.
189. Pennisi, E., *Human genome 10th anniversary. Will computers crash genomics?* Science, 2011. **331**(6018): p. 666-8.
190. Schadt, E.E., et al., *Computational solutions to large-scale data management and analysis.* Nat Rev Genet, 2010. **11**(9): p. 647-57.
191. Stein, L.D., *The case for cloud computing in genome informatics.* Genome Biol, 2010. **11**(5): p. 207.
192. Editorial, N.G., *What is the human variome project?* Nat Genet, 2007. **39**(4): p. 423.
193. Kohonen-Corish, M.R., et al., *Beyond the genomics blueprint: the 4th Human Variome Project Meeting, UNESCO, Paris, 2012.* Genet Med, 2013. **15**(7): p. 507-12.

ANEXO

A) Proporción de bases con un determinado valor de calidad.



B) Valor de calidad medio por cada ciclo de ligación y primer empleado.

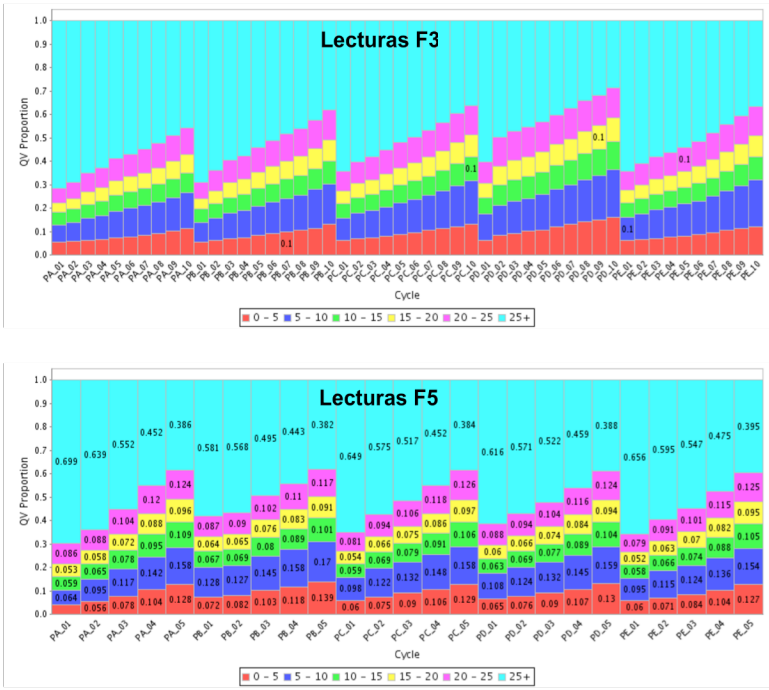


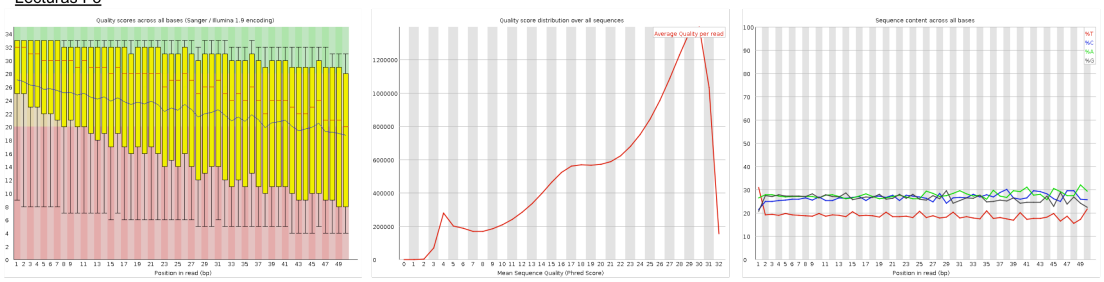
Figura 1. Control de calidad de los datos brutos generados para la muestra FA104.

Tabla 1. Listado de variantes candidatas siguiendo un modelo de herencia recesivo. Crom: cromosoma; Posición: posición genómica de la variante; Ref: Alelo en la referencia; Mues: alelo en la muestra; Tipo: efecto de la variante que puede resultar en SS=Splice Site, NFC=Non-Frameshift Coding, NSC=Non-Synonimous Coding, FC=Frameshift Coding; Cambio AA: cambio de aminoácido y posición proteica a la que afecta el cambio; Gen: nombre del gen según la nomenclatura HGNC; Lecturas: lecturas no redundantes que cubren la variante; SNVQV: valor de calidad del SNV; GQ: valor de calidad del genotipo.

Crom	Posición	Ref	Mues	Tipo	Cambio AA	Gen	Lecturas	SNV QV	GQ
1	169489751	A	W	SS	-	F5	42	171	171
1	169525877	T	Y	SS	-	F5	52	36	36
2	73675227	-	CTC	NFC	S/SP	ALMS1	16	-	-
2	73678183	G	R	NSC	G1509D	ALMS1	156	120	120
3	49094490	G	S	NSC	N381K	QRICH1	122	228	228
3	49095011	C	S	NSC	G208R	QRICH1	109	43	43
4	126238305	C	M	NSC	P247T	FAT4	52	178	178
4	126355484	C	M	NSC	A2368E	FAT4	56	190	190
5	156479444	TTG	-	NFC	TS/S	HAVCR1	61	-	-
5	156479568	-	GTT	NFC	T/TT	HAVCR1	106	-	-
6	31238942	G	W	NSC	A176V	HLA-C	23	61	39
6	31239577	A	C	NSC	S48A	HLA-C	21	90	90
6	32709309	A	R	SS	-	HLA-DQA2	29	84	84
6	32713044	C	Y	NSC	T64M	HLA-DQA2	192	228	228
6	32713188	C	Y	SS	-	HLA-DQA2	126	228	228
6	38840915	A	R	NSC	I2479V	DNAH8	72	216	216
6	38879340	A	T	NSC	E3267D	DNAH8	12	34	34
7	100686777	C	Y	NSC	T4027M	MUC17	323	228	228
7	100687107	G	R	SS	-	MUC17	66	79	79
8	30700598	T	Y	NSC	N1979S	TEX15	33	97	97
8	30701995	A	M	NSC	D1513E	TEX15	141	228	228
10	69682773	T	Y	NSC	D920G	HERC4	64	69	69
10	69785435	-	A	SS	-	HERC4	9	-	-
16	14029271	AACTC	-	FC	-	ERCC4	22	-	-
16	14041518	C	M	NSC	R689S	ERCC4	121	228	228
16	72137553	C	S	NSC	Q564E	DHX38	56	85	85
16	72142141	A	R	NSC	S994G	DHX38	52	106	106
17	74272839	C	Y	NSC	V1593M	QRICH2	54	33	33
17	74277009	T	Y	NSC	Q1264R	QRICH2	23	81	81
18	14105016	C	M	NSC	R508I	ZNF519	136	228	228
18	14105853	C	M	NSC	R229I	ZNF519	23	51	51
19	51918360	A	R	NSC	S445P	SIGLEC12	43	39	39
19	52004795	G	CT	FC	-	SIGLEC12	19	-	-
X	53561632	A	W	NSC	F4226I	HUWE1	42	53	53
X	53642759	C	M	NSC	E665D	HUWE1	16	33	33

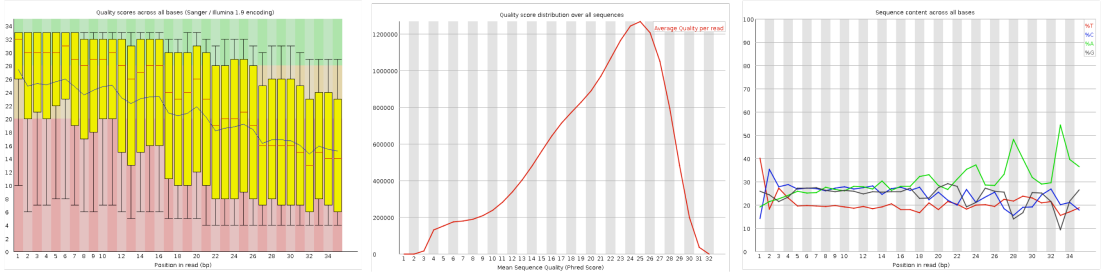
MUESTRA BM3339

Lecturas F3



A) Distribución de los valores de calidad por posición | B) Valor de calidad medio por lectura | C)Porcentaje de nucleótidos (A,G,T,C) por base

Lecturas F5

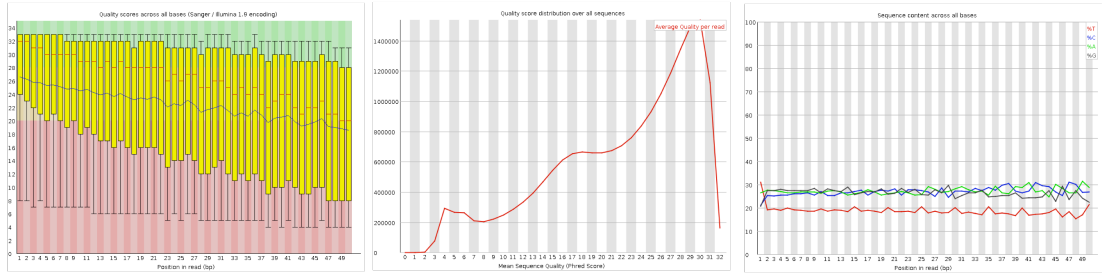


A) Distribución de los valores de calidad por posición | B) Valor de calidad medio por lectura | C)Porcentaje de nucleótidos (A,G,T,C) por base

Figura 2. Control de calidad de los datos brutos para la muestra BM3339.

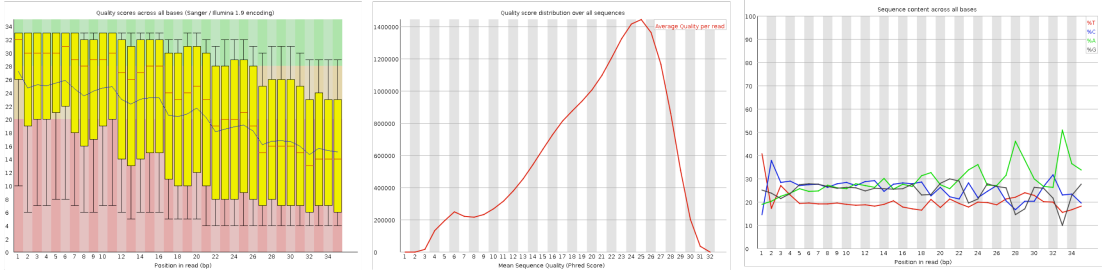
MUESTRA BM3895

Lecturas F3



A) Distribución de los valores de calidad por posición | B) Valor de calidad medio por lectura | C)Porcentaje de nucleótidos (A,G,T,C) por base

Lecturas F5

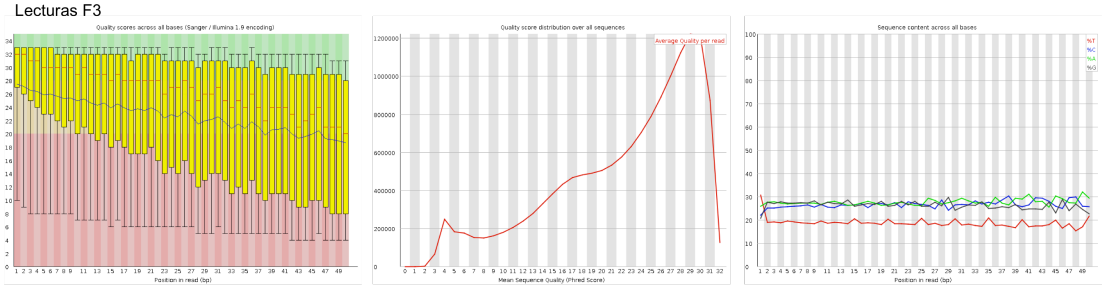


A) Distribución de los valores de calidad por posición | B) Valor de calidad medio por lectura | C)Porcentaje de nucleótidos (A,G,T,C) por base

Figura 3. Control de calidad de los datos brutos para la muestra BM3895.

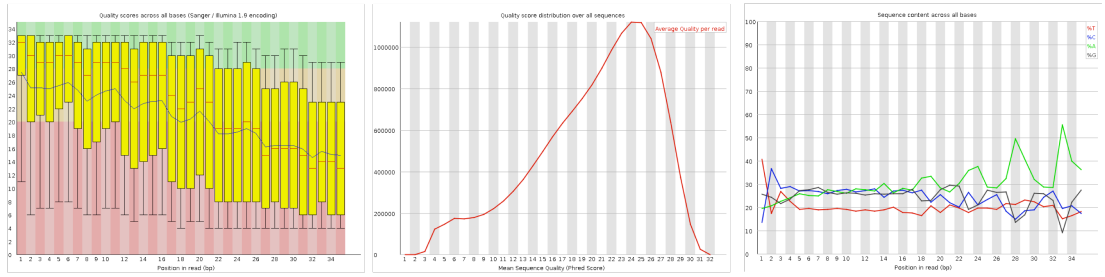
MUESTRA BM4237

Lecturas F3



A) Distribución de los valores de calidad por posición B) Valor de calidad medio por lectura C)Porcentaje de nucleótidos (A,G,T,C) por base

Lecturas F5

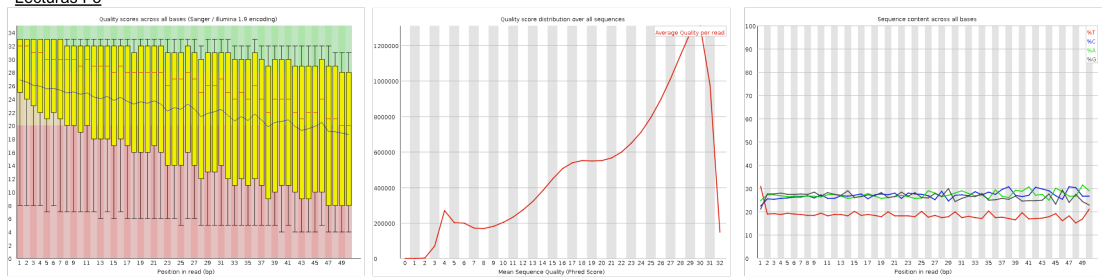


A) Distribución de los valores de calidad por posición B) Valor de calidad medio por lectura C)Porcentaje de nucleótidos (A,G,T,C) por base

Figura 4. Control de calidad de los datos brutos para la muestra BM4237.

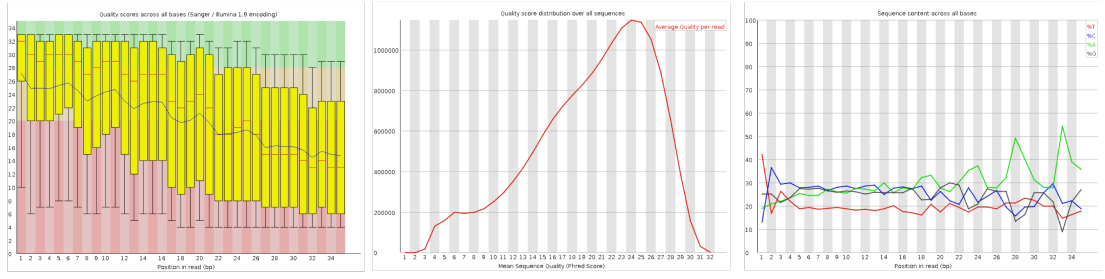
MUESTRA BM5307

Lecturas F3



A) Distribución de los valores de calidad por posición B) Valor de calidad medio por lectura C)Porcentaje de nucleótidos (A,G,T,C) por base

Lecturas F5

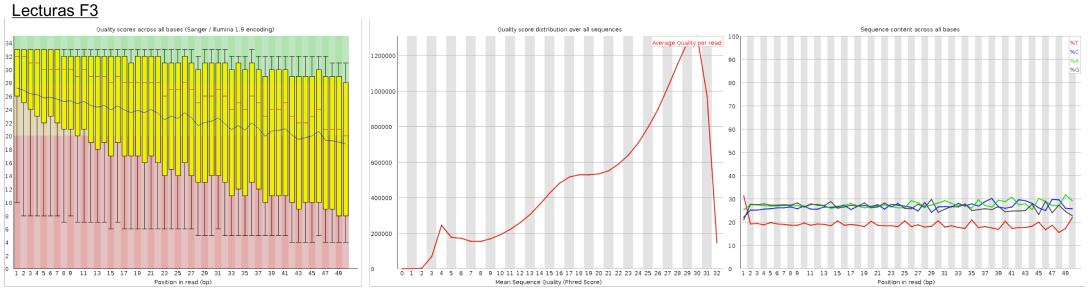


A) Distribución de los valores de calidad por posición B) Valor de calidad medio por lectura C)Porcentaje de nucleótidos (A,G,T,C) por base

Figura 5. Control de calidad de los datos brutos para la muestra BM5307.

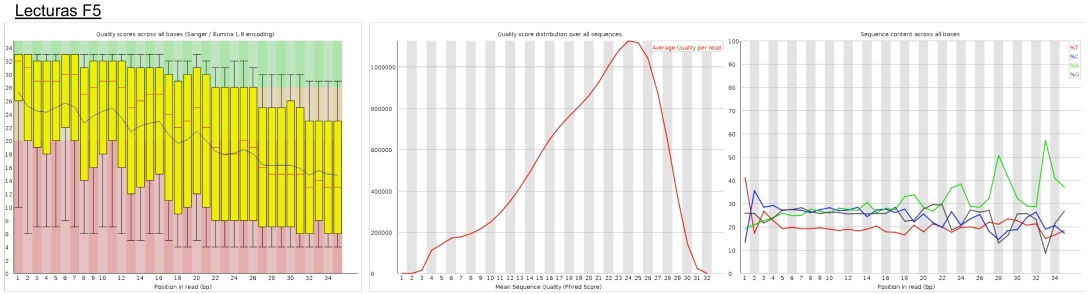
MUESTRA BM5357

Lecturas F3



A) Distribución de los valores de calidad por posición B) Valor de calidad medio por lectura C)Porcentaje de nucleótidos (A,G,T,C) por base

Lecturas F5

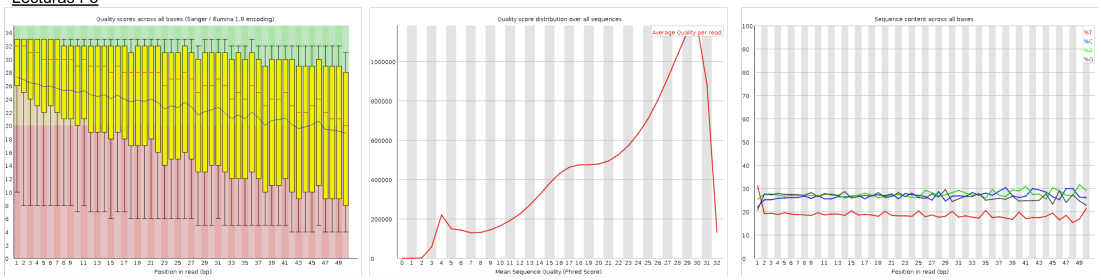


A) Distribución de los valores de calidad por posición B) Valor de calidad medio por lectura C)Porcentaje de nucleótidos (A,G,T,C) por base

Figura 6. Control de calidad de los datos brutos para la muestra BM5357.

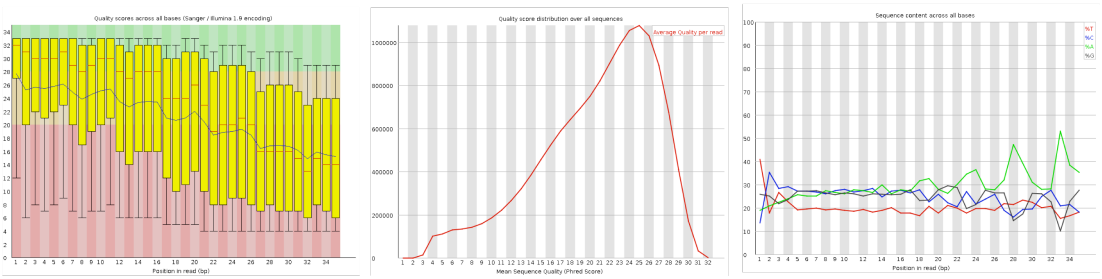
MUESTRA BM6091

Lecturas F3



A) Distribución de los valores de calidad por posición B) Valor de calidad medio por lectura C)Porcentaje de nucleótidos (A,G,T,C) por base

Lecturas F5

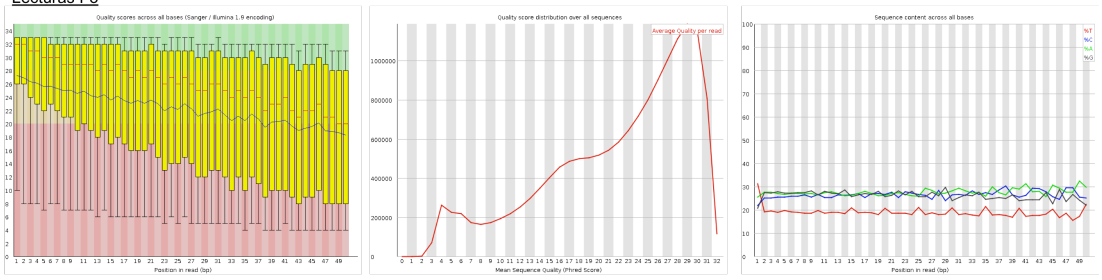


A) Distribución de los valores de calidad por posición B) Valor de calidad medio por lectura C)Porcentaje de nucleótidos (A,G,T,C) por base

Figura 7. Control de calidad de los datos brutos para la muestra BM6091.

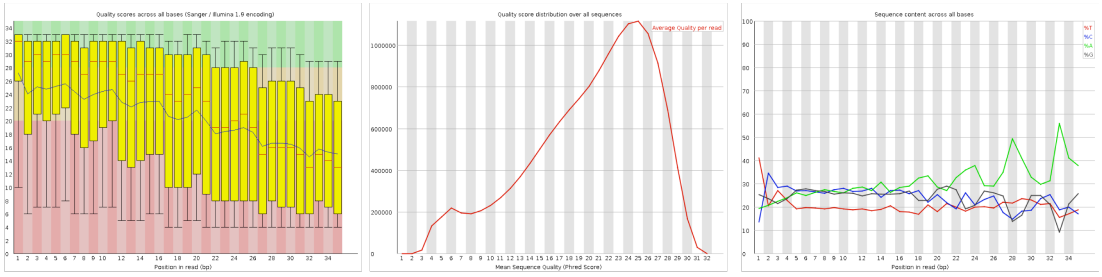
MUESTRA BM6092

Lecturas F3



A) Distribución de los valores de calidad por posición B) Valor de calidad medio por lectura C)Porcentaje de nucleótidos (A,G,T,C) por base

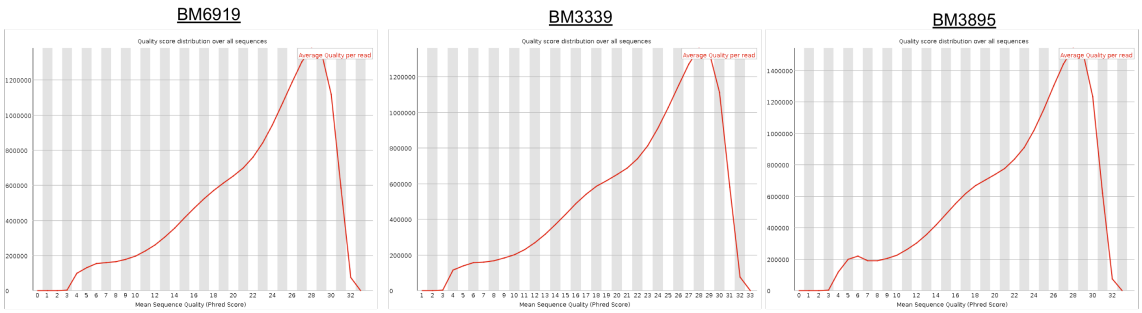
Lecturas F5



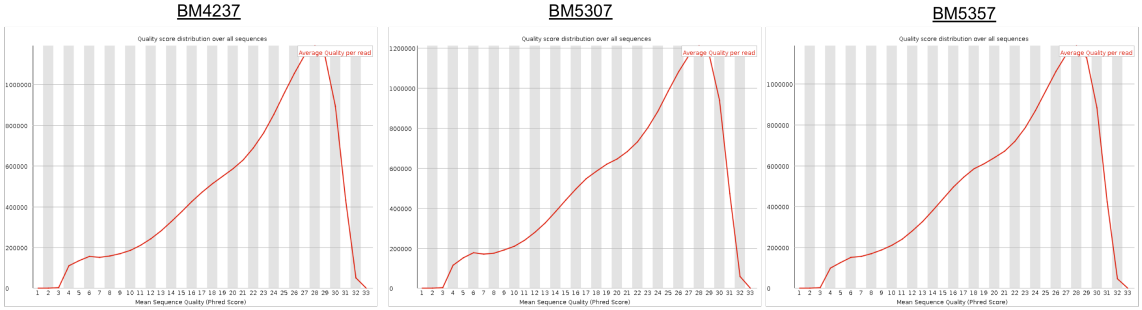
A) Distribución de los valores de calidad por posición B) Valor de calidad medio por lectura C)Porcentaje de nucleótidos (A,G,T,C) por base

Figura 8. Control de calidad de los datos brutos para la muestra BM6092.

BM6919 BM3339 BM3895



BM4237 BM5307 BM5357



BM6091 BM6092 BM6492

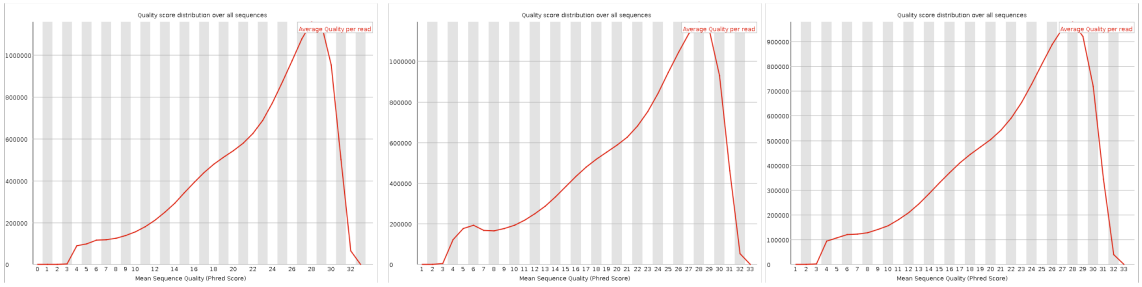
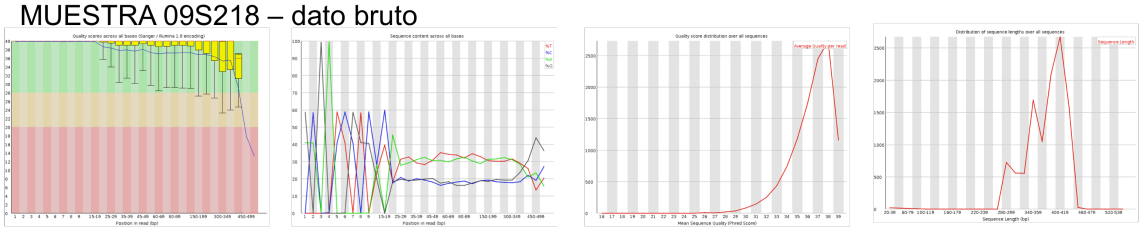


Figura 9.- Valor de calidad medio por lectura tras eliminar los últimos 10nt en las lecturas F5.

Tabla 2. Listado de variantes con relevancia diagnóstica identificadas en el estudio retrospectivo en 40 de los 163 individuos estudiados.

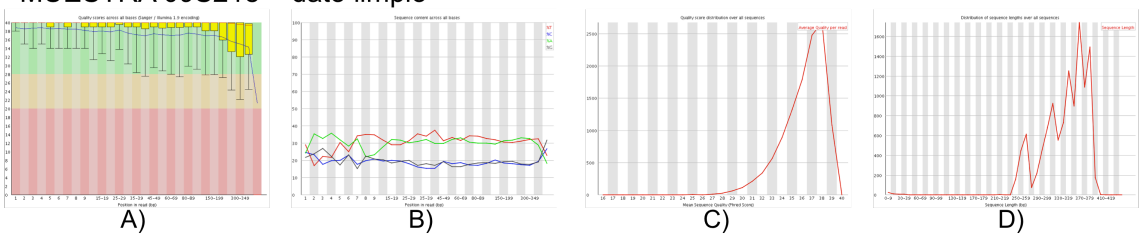
Código del paciente	Diagnóstico	Gen	Nucleótido	Aminoácido
4	Miocardiopatía Hipertrófica	MYBPC3	NM_000256.3:c.3605delC	p.Cys1202Leufs*35
7	Arritmia familiar	AKAP9	NM_005751.4:c.10129C>T	p.Arg3377*
15	Síndrome de Brugada	MYBPC3	NM_000256.3:c.2873C>T	p.Thr958Ile
18	Displasia Arritmogénica Ventrículo derecho	DSP	NM_004415.2:c.6850C>T	p.Arg2284*
19	Síndrome de Marfan	FBN1	NM_000138.4:c.478T>C	p.Cys160Arg
20	Síndrome QT largo	KCNH2	NM_000238.2:c.2587C>T	p.Arg863*
28	Síndrome de Marfan	SLC25A4	NM_001151.3:c.178A>T	p.R60*
30	Síndrome de Marfan	FBN1	NM_000138.4:c.7776C>A	p.Cys2592*
33	Miocardiopatía Hipertrófica	MYH7	NM_000257.2:c.428G>A	p.Arg143Gln
36	Síndrome QT corto	LDB3	NM_001080116.1:c.349G>A	p.Asp117Asn
37	Síndrome de Marfan	FBN1	NM_000138.4:c.2362_2370del	p.Phe788_Cys790del
42	Síndrome de Marfan	FBN1	NM_000138.4:c.1585C>T	p.Arg529*
47	Síndrome QT largo	TNNT2	NM_000364.2:c.502C>T	p.Arg168*
49	Síndrome de Marfan	ABCC9	NM_005691.2:c.1467_1468insA	p.Glu490Argfs*8
56	Miocardiopatía Hipertrófica	MYBPC3	NM_000256.3:c.2529_2530dup	p.Met844Argfs*36
59	Miocardiopatía dilatada	TTN	NM_133378.4:c.73012C>T	p.Arg24338*
59	Miocardiopatía dilatada	LDB3	NM_001080116.1:c.349G>A	p.Asp117Asn
64	Miocardiopatía dilatada	MYBPC3	NM_000256.3:c.1546G>T	p.Glu516*
65	Síndrome de Marfan	TNNT2	NM_001001430.1:c.230C>T	p.Pro77Leu
68	Síndrome de Marfan	FBN1	NM_000138.4:c.6530delG	p.Gly2177Glufs*8
70	Síndrome de Marfan	FBN1	NM_000138.4:c.1869C>A	p.Cys623*
71	Displasia Arritmogénica Ventrículo derecho	ANK2	NM_001148.4:c.4373A>G	p.Glu1458Gly
72	Miocardiopatía Hipertrófica	RBM20	NM_001134363.1:c.1364C>T	p.Ser455Leu
74	Síndrome de Marfan	FBN1	NM_000138.4:c.4096G>A	p.Glu1366Lys
76	Arritmia familiar	AKAP9	NM_005751.4:c.7438C>T	p.Gln2480*
78	Síndrome QT largo	KCNQ1	NM_000218.2:c.871_872insA	p.Ser291Tyrfs*172
82	Displasia Arritmogénica Ventrículo derecho	PKP2	NM_004572.3:c.1368del	p.Lys456Asnfs*3
84	Miocardiopatía Hipertrófica	MYBPC3	NM_000256.3:c.1090+1G>A	p.?
84	Miocardiopatía Hipertrófica	VCL	NM_014000.2:c.829C>A	p.Leu277Met
89	Miocardiopatía Hipertrófica	MYBPC3	NM_000256.3:c.772G>A	p.Glu258Lys
89	Miocardiopatía Hipertrófica	MYBPC3	NM_000256.3:c.1828G>C	p.Asp610His
91	Miocardiopatía Hipertrófica	MYBPC3	NM_000256.3:c.772G>A	p.Glu258Lys
92	Miocardiopatía Hipertrófica	MYBPC3	NM_000256.3:c.2905+1G>A	p.?
93	Miocardiopatía Hipertrófica	LMNA	NM_170707.2:c.1930C>T	p.Arg644Cys
103	Miocardiopatía dilatada	TTN	NM_003319.4:c.43015_43028delinsTTTACTCTTC	p.Glu14339Phefs*11
103	Miocardiopatía dilatada	KCNE3	NM_005472.4:c.296G>A	p.Arg99His
108	Síndrome de Brugada	MYBPC3	NM_000256.3:c.2373dup	p.Trp792Valfs*41
111	Taquicardia ventricular catecolaminérgica	KCNJ2	NM_000891.2:c.644G>A	p.Gly215Asp
114	Ventrículo izquierdo no compactado	TTN	NM_003319.4:c.30800del	p.His10267Profs*18
116	Miocardiopatía Hipertrófica	MYBPC3	NM_000256.3:c.772G>A	p.Glu258Lys
116	Miocardiopatía Hipertrófica	LDB3	NM_001080116.1:c.349G>A	p.Asp117Asn
119	Miocardiopatía dilatada	TTN	NM_003319.4:c.75039_75042del	p.Arg25014Serfs*9
150	Síndrome QT largo	KCNJ2	NM_000891.2:c.652C>T	p.Arg218Trp
161	Síndrome de Brugada	SCN5A	NM_198056.2:c.2582_2583del	p.Phe861Trpfs*90
162	Ventrículo izquierdo no compactado	TNNT2	NM_001001430.1:c.281G>A	p.Arg94His

MUESTRA 09S218 – dato bruto



A) B) C) D)

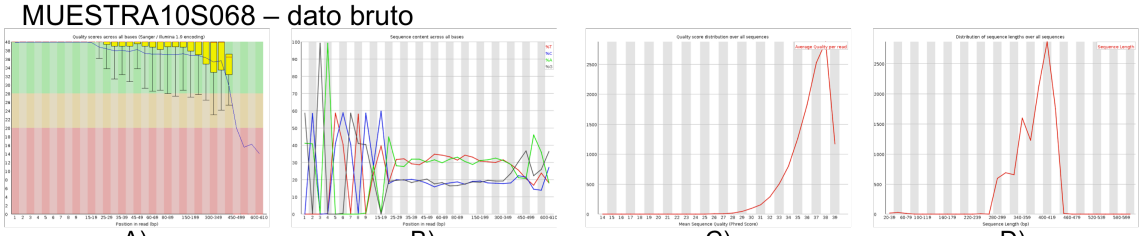
MUESTRA 09S218 – dato limpio



A) B) C) D)

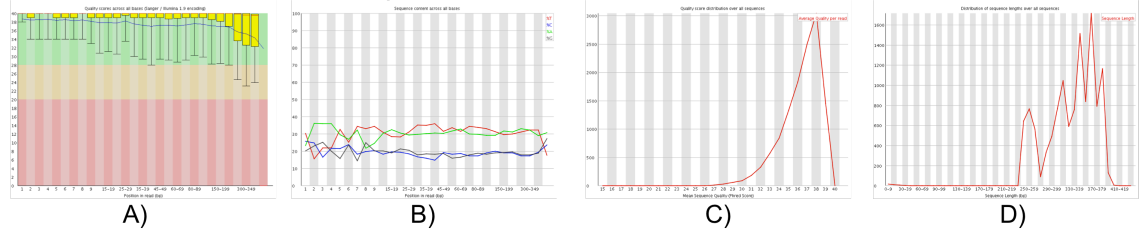
Figura 10.- Control de calidad del dato bruto y tras la eliminación de adaptadores y bases/lecturas de baja calidad en la muestra 09S218.

MUESTRA10S068 – dato bruto



A) B) C) D)

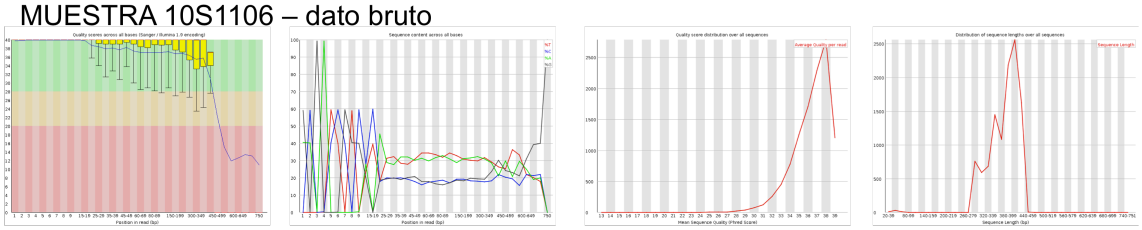
MUESTRA10S068 – dato limpio



A) B) C) D)

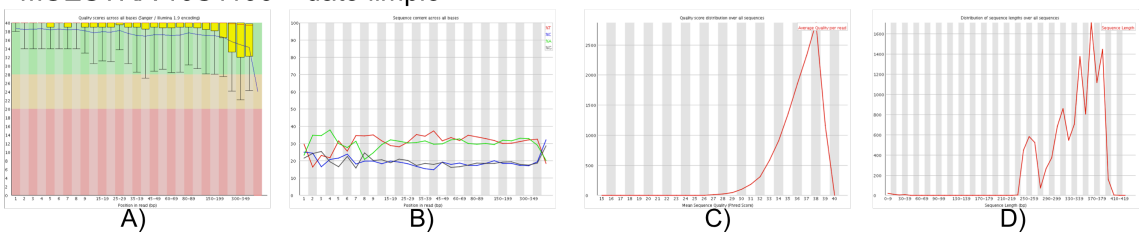
Figura 11.- Control de calidad del dato bruto y tras la eliminación de adaptadores y bases/lecturas de baja calidad en la muestra 10S068.

MUESTRA 10S1106 – dato bruto



A) B) C) D)

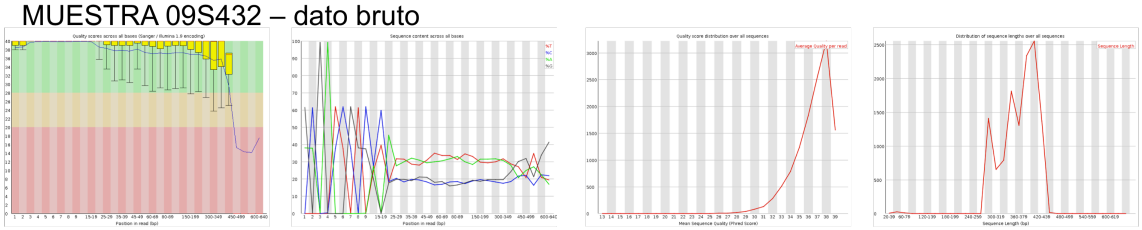
MUESTRA 10S1106 – dato limpio



A) B) C) D)

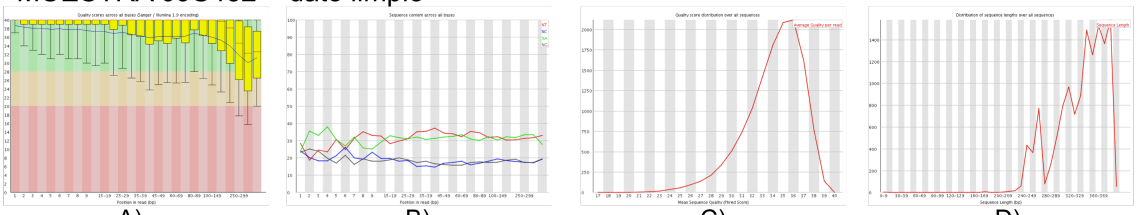
Figura 12.- Control de calidad del dato bruto y tras la eliminación de adaptadores y bases/lecturas de baja calidad en la muestra 10S1106.

MUESTRA 09S432 – dato bruto



A) B) C) D)

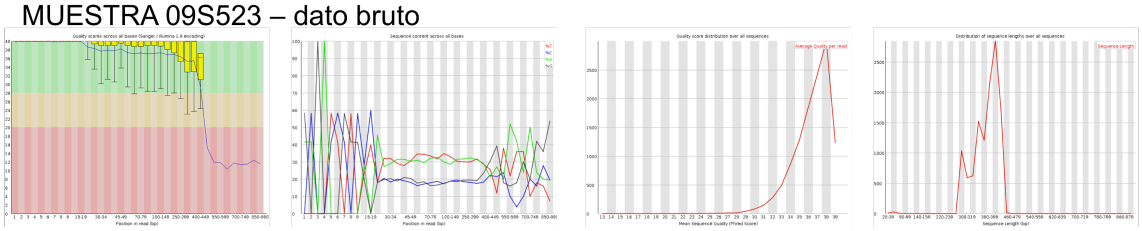
MUESTRA 09S432 – dato limpio



A) B) C) D)

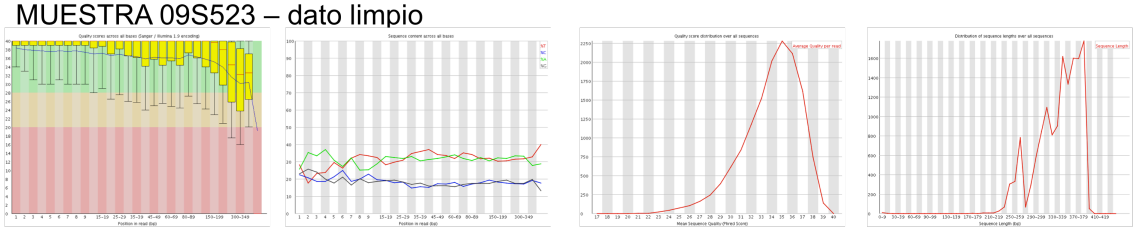
Figura 13.- Control de calidad del dato bruto y tras la eliminación de adaptadores y bases/lecturas de baja calidad en la muestra 09S432.

MUESTRA 09S523 – dato bruto



A) B) C) D)

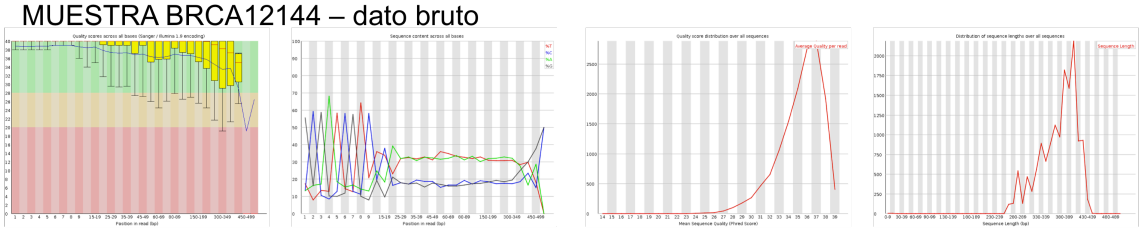
MUESTRA 09S523 – dato limpio



A) B) C) D)

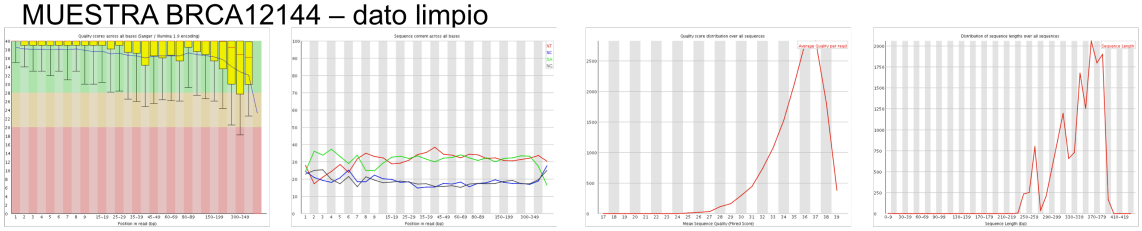
Figura 14.- Control de calidad del dato bruto y tras la eliminación de adaptadores y bases/lecturas de baja calidad en la muestra 09S523.

MUESTRA BRCA12144 – dato bruto



A) B) C) D)

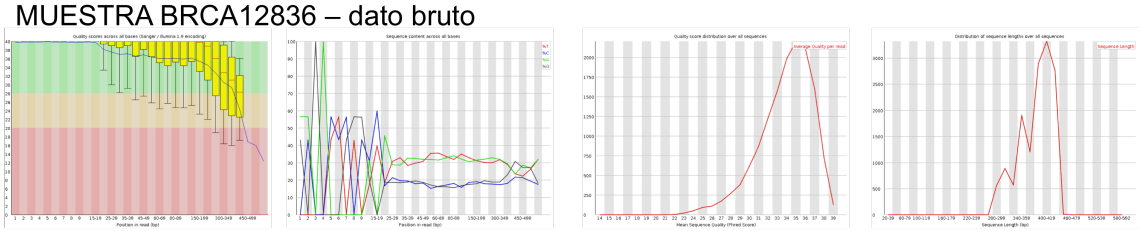
MUESTRA BRCA12144 – dato limpio



A) B) C) D)

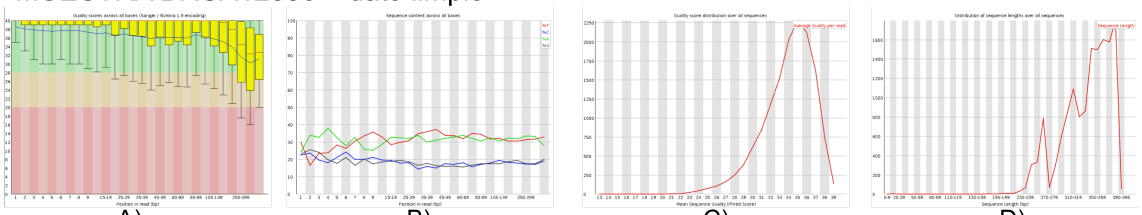
Figura 15.- Control de calidad del dato bruto y tras la eliminación de adaptadores y bases/lecturas de baja calidad en la muestra BRCA12144.

MUESTRA BRCA12836 – dato bruto



A) B) C) D)

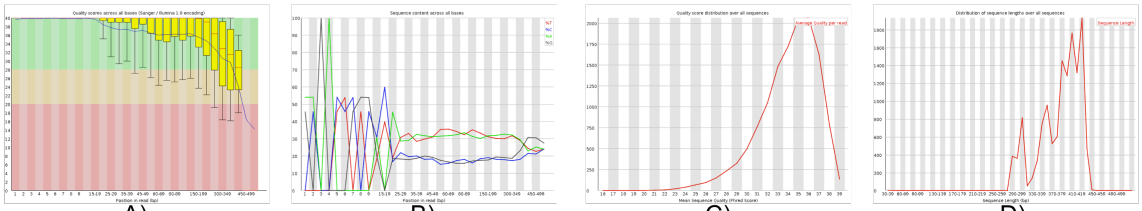
MUESTRA BRCA12836 – dato limpio



A) B) C) D)

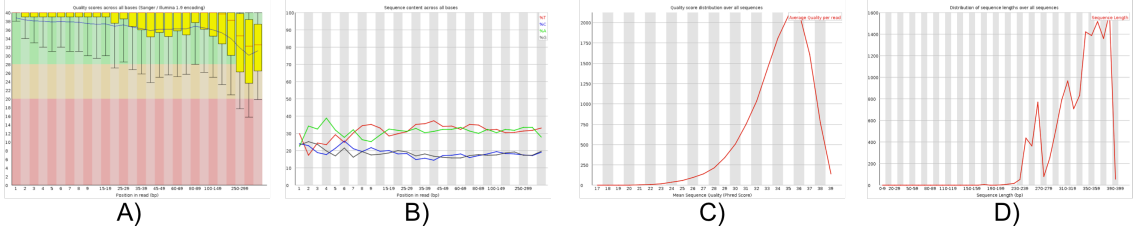
Figura 16.- Control de calidad del dato bruto y tras la eliminación de adaptadores y bases/lecturas de baja calidad en la muestra BRCA12836.

MUESTRA BRCA12733 – dato bruto



A) B) C) D)

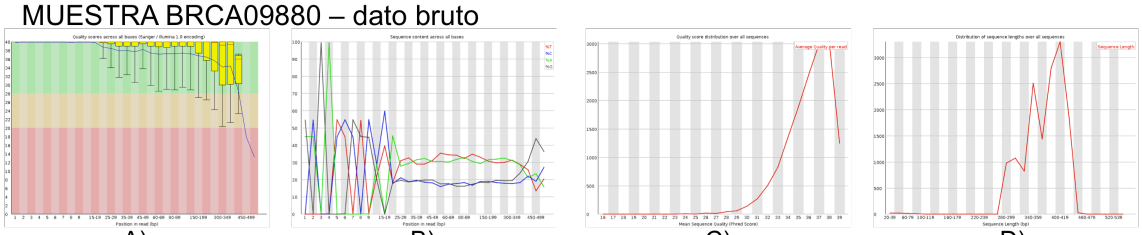
MUESTRA BRCA12733 – dato limpio



A) B) C) D)

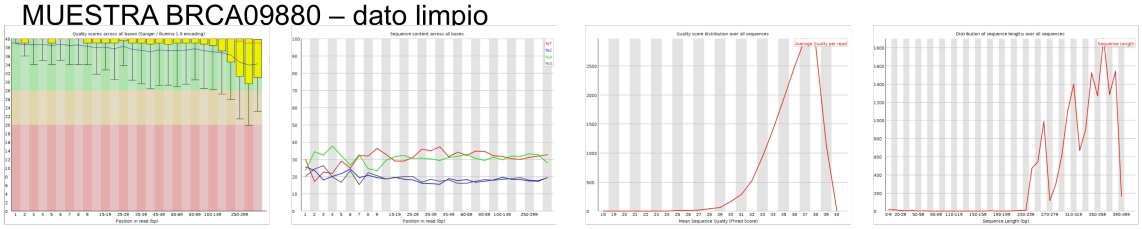
Figura 17.- Control de calidad del dato bruto y tras la eliminación de adaptadores y bases/lecturas de baja calidad en la muestra BRCA12733.

MUESTRA BRCA09880 – dato bruto



A) B) C) D)

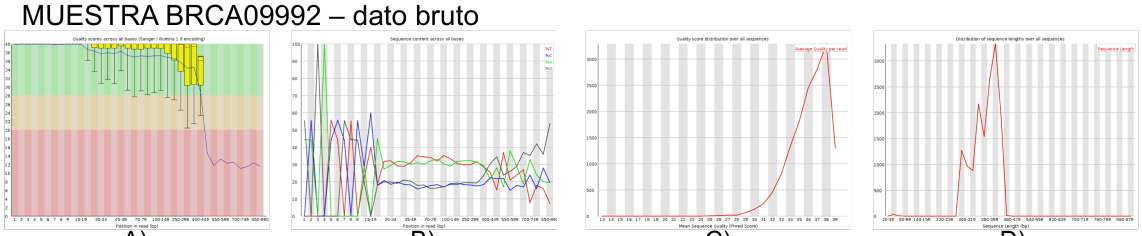
MUESTRA BRCA09880 – dato limpio



A) B) C) D)

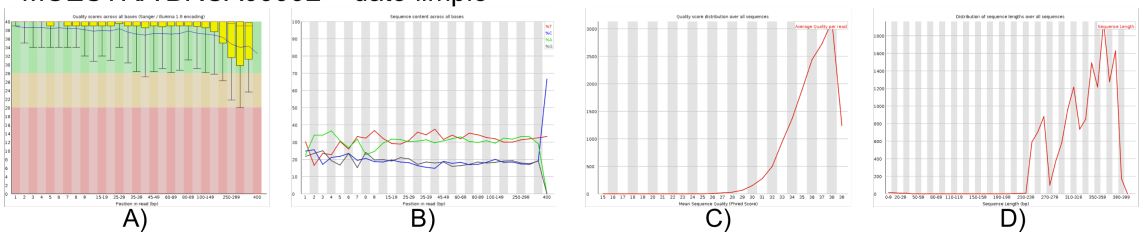
Figura 18.- Control de calidad del dato bruto y tras la eliminación de adaptadores y bases/lecturas de baja calidad en la muestra BRCA09880.

MUESTRA BRCA09992 – dato bruto



A) B) C) D)

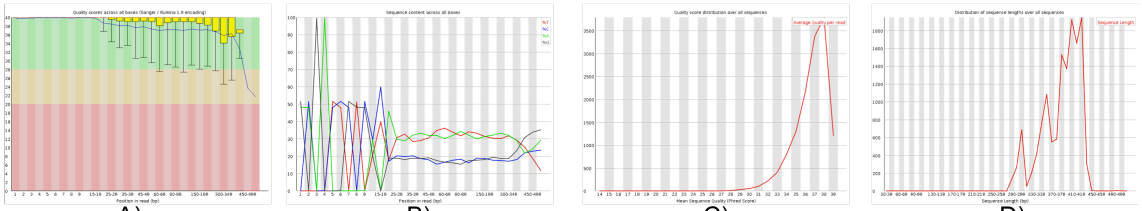
MUESTRA BRCA09992 – dato limpio



A) B) C) D)

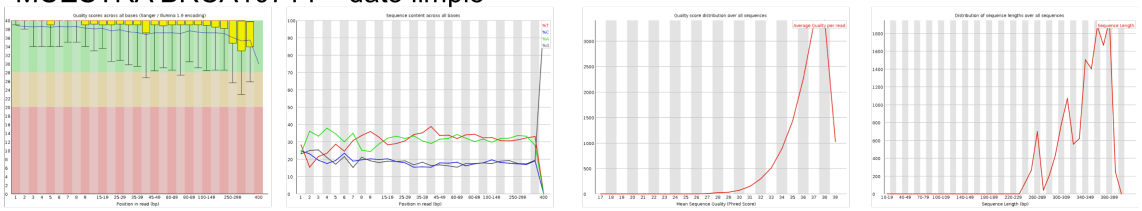
Figura 19.- Control de calidad del dato bruto y tras la eliminación de adaptadores y bases/lecturas de baja calidad en la muestra BRCA09992.

MUESTRA BRCA10714 – dato bruto



A) B) C) D)

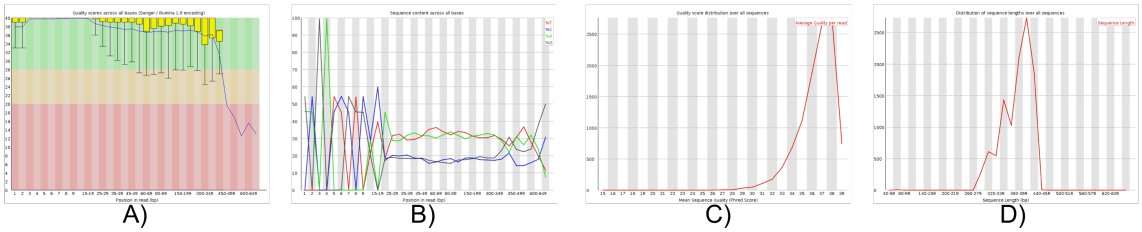
MUESTRA BRCA10714 – dato limpio



A) B) C) D)

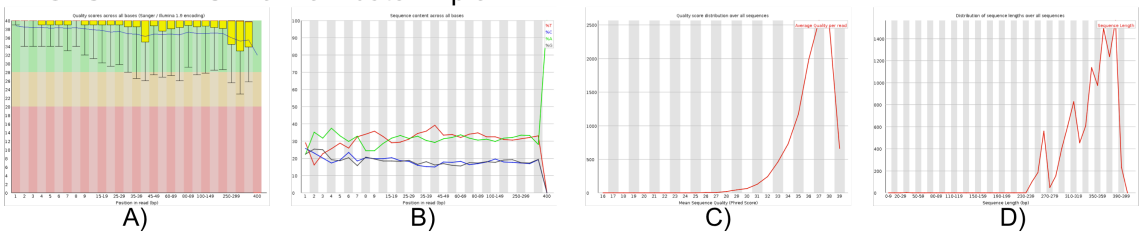
Figura 20.- Control de calidad del dato bruto y tras la eliminación de adaptadores y bases/lecturas de baja calidad en la muestra BRCA10714.

MUESTRA BRCA10726 – dato bruto



A) B) C) D)

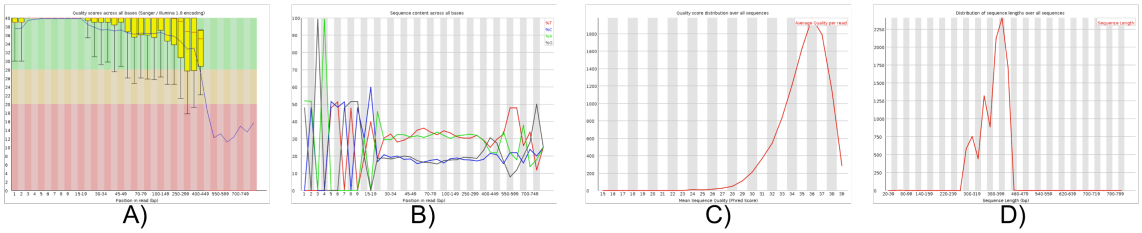
MUESTRA BRCA10726 – dato limpio



A) B) C) D)

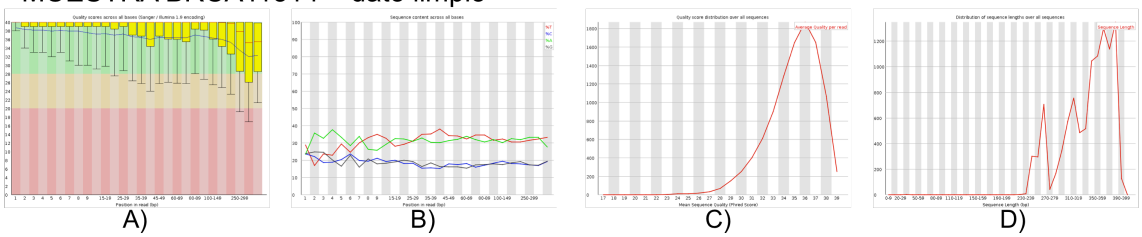
Figura 21.- Control de calidad del dato bruto y tras la eliminación de adaptadores y bases/lecturas de baja calidad en la muestra BRCA10726.

MUESTRA BRCA11314 – dato bruto



A) B) C) D)

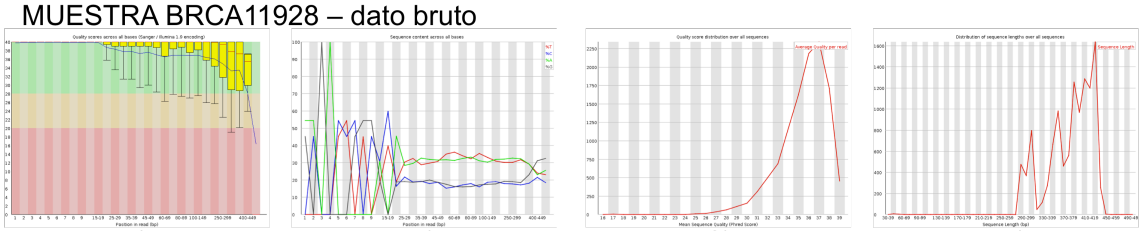
MUESTRA BRCA11314 – dato limpio



A) B) C) D)

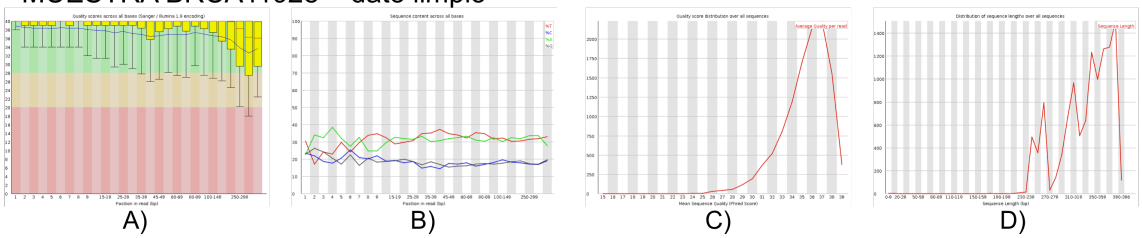
Figura 22.- Control de calidad del dato bruto y tras la eliminación de adaptadores y bases/lecturas de baja calidad en la muestra BRCA11314.

MUESTRA BRCA11928 – dato bruto



A) B) C) D)

MUESTRA BRCA11928 – dato limpio



A) B) C) D)

Figura 23.- Control de calidad del dato bruto y tras la eliminación de adaptadores y bases/lecturas de baja calidad en la muestra BRCA11928.