Desarrollo de un protocolo de detección de CNVs relacionadas con cáncer de mama y ovario hereditario a partir de datos de NGS

Marina Sánchez Soler
Máster en Bioinformática y Bioestadística
Estudios genéticos de enfermedades humanas

Consultora: Helena Brunel Montaner
Profesor Responsable de la asignatura: David Merino Arranz

Junio 2018

FICHA DEL TRABAJO FINAL

Título del trabajo:	*Desarrollo de un protocolo de detección de CNVs relacionadas con cáncer de mama y ovario hereditario a partir de datos de NGS*
Nombre del autor:	*Marina Sánchez Soler*
Nombre del consultor/a:	*Helena Brunel Montaner*
Nombre del PRA:	*David Merino Arranz*
Fecha de entrega (mm/aaaa):	*06/2018*
Titulación:	*Máster en Bioinformática y Bioestadística*
Área del Trabajo Final:	*Estudios genéticos de enfermedades humanas*
Idioma del trabajo:	*Castellano*
Palabras clave	*Copy Number Variation (CNV), Breast Cancer, Next generation sequencing (NGS)*

Resumen del Trabajo:

El cáncer de mama y/o de ovario es uno de los más frecuentes en España según las últimas estadísticas de la SEOM, siendo las alteraciones en los genes *BRCA1* y *BRCA2* las principales causas de esta patología.

Gracias a la introducción de las técnicas de secuenciación masiva (NGS) se ha conseguido aumentar la eficacia en el diagnóstico de estas enfermedades, reduciendo los tiempos de procesado y abaratando los costes. Esta tecnología presenta buenos resultados para la detección de variantes pequeñas (SNPs e *indels*), pero su utilización para la detección de variantes más grandes (CNVs) sigue siendo un reto. En este punto, surge la necesidad de protocolizar y optimizar las herramientas bioinformáticas para el análisis de estas variantes.

El presente proyecto se ha centrado en evaluar diferentes herramientas de análisis de CNVs en línea germinal a partir de datos de NGS. Para llevar a cabo dicho objetivo se ha empleado la base de datos con código *EGAS00001002428* de la plataforma EGA que contiene los resultados de 96 muestras evaluadas mediante el panel *TruSight Cancer* de Illumina. Las herramientas a evaluar han sido DECoN y panelcn.MOPS.

Las conclusiones obtenidas muestran que ambos algoritmos presentan buenos rendimientos en términos de sensibilidad y especificidad, siendo panelcn.MOPS mejor candidato para aplicar en clínica ya que posee una sensibilidad del 100%. Aun así, se necesita de más estudios que corroboren los resultados obtenidos para valorar la implementación de esta herramienta en la rutina del laboratorio.

Abstract:

The breast and/or ovarian cancer is one of the most frequent in Spain according to the latest statistics of the SEOM, being the alterations in the *BRCA1* and *BRCA2* genes the main causes of this pathology.

The introduction of Next-Generation Sequencing (NGS) has allowed to increase the efficiency in the diagnosis of these diseases, reducing processing times and costs. This technology has good results for the detection of small variants (SNPs and indels), but its use for the detection of larger variants (CNVs) remains a challenge. At this point, there is a need for protocol and optimize bioinformatics tools for the analysis of these variants.

The present project has focused on evaluating different tools to analyze germline CNVs from NGS data. To carry out this objective, the database with code *EGAS00001002428* of the EGA platform was used, which contains the results of 96 samples evaluated by the *TruSight Cancer panel* (Illumina). The tools evaluated have been DECoN and panelcn.MOPS.

The conclusions obtained show that both algorithms presented good performances in terms of sensitivity and specificity, with panelcn.MOPS being the best candidate to apply in the clinic since it has a sensitivity of 100%. Even so, more studies are needed to corroborate the results obtained to assess the implementation of this tool in the laboratory routine.

Índice

Lista de figuras

Lista de tablas

1. Introducción

1.1 Contexto y justificación del Trabajo

El cáncer es uno de los grupos de enfermedades de mayor impacto en la salud pública, diagnosticándose más de 14 millones de nuevos casos al año[1]. Según la Sociedad Española de Oncología Médica (SEOM), el cáncer de mama y/o de ovario se encuentran entre los tipos más frecuentes en España, tal y como se muestra en la Figura 1.

***Figura 1**. Incidencia estimada de los tumores más frecuentes en España en el año 2017 (ambos sexos). Datos procedentes de GLOBOCAN 2012, desglosados por edad y sexo, y extrapolados a los datos de la población española para el año 2017 proporcionada por el INE[2].*

Se estima que aproximadamente un 5-10% de este tipo de cánceres son hereditarios, con un patrón de herencia autosómico dominante, siendo *BRCA1* [MIM# 113705][3] y *BRCA2* [MIM# 600185][4] los principales genes implicados[5]. La frecuencia de portadores de mutaciones en estos genes está descrita en torno a 1/400 - 1/800 dependiendo de la etnia[6] y se calcula que el riesgo acumulado de cáncer de mama a los 70 años en pacientes portadores de mutaciones en *BRCA1* y *BRCA2* es de 65% y 45%, respectivamente, y el riesgo de cáncer de ovario de 39% y 10%, respectivamente. Además, se ha demostrado que existen otros genes involucrados en estos tipos de cánceres, pero su contribución y la penetrancia de sus mutaciones para la mayoría de ellos sigue sin ser clara[7]. En la Figura 2 se observa la distribución de las mutaciones encontradas en los diferentes genes relacionados con esta patología.

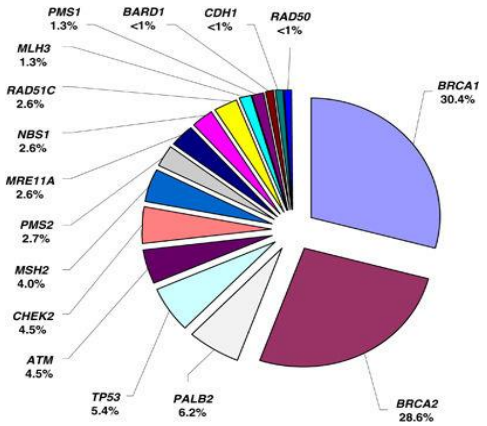


***Figura 2.** Distribución relativa de las variantes detectadas mediante NGS en un estudio de 708 pacientes diagnosticados de cáncer de mama y/o de ovario hereditario.*

Dado su alto impacto en la sociedad, el diagnóstico precoz de esta enfermedad toma un importante papel en este ámbito de la medicina, ayudando a prevenir el cáncer en el individuo y sus familiares y aumentando las tasas de éxito de los tratamientos.

La introducción de las técnicas de secuenciación masiva o *Next Generation Sequencing* (NGS) ha permitido optimizar el estudio de los genes implicados en dicha patología, reduciendo los tiempos de procesado y abaratando los costes respecto a la secuenciación *Sanger*, así como permitiendo estudiar un mayor número de genes en un mismo proceso[5]. Estos estudios permiten detectar variantes del tipo: *SNP* (*Single Nucleotide Polymorphism*) o variación de un solo nucleótido, *indels* (pequeñas deleciones e inserciones), *CNVs* (*Copy Number Variation*) o variaciones en el número de copias (deleciones, inserciones o duplicaciones de fragmentos de ADN de tamaño comprendido entre 1 Kb y 5 Mb) y *otras variantes estructurales* como grandes deleciones, inserciones o duplicaciones de un tamaño mayor a 5 Mb[8] (Figura 3).

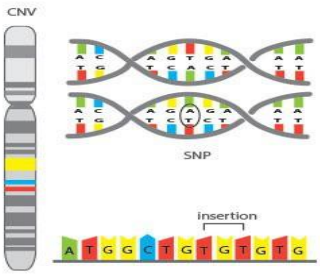


Figura 3. *Representación de los diferentes eventos que se pueden detectar: polimorfismos de un solo nucleótido (SNP), inserciones o deleciones de pequeño tamaño (indels) y variaciones en el número de copias (CNV).*[9]

Este tipo de variantes pueden encontrarse tanto en línea somática (alteraciones genéticas detectadas en el tumor) como en línea germinal. El estudio del cáncer hereditario se centra en las variantes detectadas en la línea germinal, las cuales son susceptibles de ser transmitidas a la descendencia.

En la actualidad, existen protocolos optimizados de detección de SNPs e *indels* pero hay una falta de estandarización en las herramientas y criterios a seguir para el análisis de las CNVs a partir de los datos de NGS siendo necesario emplear otras técnicas como es el MLPA[10,11](*Multiplex Ligation-dependent Probe Amplification*).

El objetivo del presente trabajo es revisar las herramientas descritas en la bibliografía para la determinación de CNVs a partir de datos de NGS, analizando su rendimiento en términos de sensibilidad y especificidad sobre un *dataset* particular. Para ello, se ha empleado los datos recogidos en el estudio con código *EGAS00001002428* de la plataforma *European Genome-Phenome Archive* (EGA) que incluye resultados de 96 muestras analizadas mediante el panel *TruSight Cancer* de Illumina[12,13]. Esta base de datos contiene también los resultados de la validación por MLPA de las CNVs detectadas.

Los resultados obtenidos de este estudio podrían ser de utilidad y aplicación clínica en el diagnóstico del cáncer hereditario.

1.2 Objetivos del Trabajo

1.2.1 Objetivo general

El objetivo principal de este trabajo consiste en definir el protocolo que presente mayor rendimiento para la detección de CNVs en línea germinal relacionadas con cáncer de mama y/o de ovario hereditario a partir de datos de NGS, y determinar su posible aplicación en clínica.

1.2.2 Objetivos específicos

Los objetivos específicos de este trabajo son:

1. Solicitar acceso al estudio *EGAS00001002428* de la plataforma *European Genome-Phenome Archive* (EGA) que contiene el resultado de 96 muestras analizadas mediante el panel *TruSight Cancer* de Illumina (NGS).
2. Seleccionar, de la bibliografía publicada, al menos 2 protocolos de detección de CNVs en línea germinal para el estudio de los genes *BRCA1* y *BRCA2* a partir de datos de NGS.
3. Aplicar los protocolos seleccionados sobre el *dataset* elegido cuyos resultados han sido validados previamente por MLPA.
4. Determinar los parámetros de sensibilidad, especificidad, VPP y VPN de los algoritmos propuestos.
5. Seleccionar el protocolo que presente un mejor rendimiento en la detección de CNVs y evaluar la posibilidad de aplicarlo en la rutina de diagnóstico clínico del cáncer de mama y/o de ovario hereditario.

1.3 Enfoque y método seguido

En primer lugar, se ha escogido la base de datos *EGAS00001002428* de la plataforma *European Genome-Phenome Archive* (EGA)[13] para aplicar los protocolos seleccionados en el análisis de las CNVs de las muestras contenidas en este estudio. De entre los datos disponibles, tanto públicamente como de acceso restringido, esta base de datos ha sido la seleccionada ya que posee las características idóneas para comprobar la eficacia y el rendimiento de los algoritmos propuestos. Este conjunto de datos contiene los datos brutos de la secuenciación masiva de 96 muestras estudiadas por un panel de genes enfocado al cáncer hereditario (*TruSight Cancer*) de la casa comercial *Illumina*. El estudio de las CNVs de este conjunto de datos ha sido validado posteriormente por la técnica MLPA, incluyéndose los resultados confirmatorios en el *dataset*. Se ha considerado emplear esta base de datos ya que contiene los resultados de las pruebas confirmatorias permitiendo determinar la eficacia de los protocolos que se van a emplear.

La estrategia propuesta para el trabajo es realizar un análisis bioinformático de los datos de partida y extraer valores de rendimiento de las tasas de detección en términos de sensibilidad, especificidad, VPP y VPN.

1.4 Planificación del Trabajo

1.4.1 Tareas del Trabajo

Las tareas que se van a llevar a cabo durante la elaboración del trabajo son:

Fase 1: Búsqueda bibliográfica

1.1 Solicitud de acceso a la base de datos *EGAS00001002428* de la plataforma *European Genome-Phenome Archive* (EGA) que incluye resultados de 96 muestras analizadas mediante el panel *TruSight Cancer* de *Illumina*.
1.2 Revisión de los datos contenidos en el *dataset* y preparación de los datos brutos para su posterior análisis.
1.3 Búsqueda bibliográfica y selección de 2 protocolos de detección de CNVs a partir de datos de NGS.

Fase 2: Análisis bioinformático

2.1 Construcción de *scripts*. Aplicación de los algoritmos seleccionados sobre las muestras de estudio.
2.2 Análisis de los resultados. Cálculo de rendimiento de los protocolos.
2.3 Comparar los rendimientos obtenidos y seleccionar el algoritmo que mejor resultados presente.

Fase 3: Redacción de la memoria del trabajo.
Fase 4: Elaboración de la presentación.
Fase 5: Defensa del TFM.

1.4.2 Hitos del Trabajo

1. Panificación temporal de trabajo.
2. Preparación del *dataset*. Familiarización con los datos de partida y preparación para su posterior análisis.
3. Selección de 2 algoritmos de los propuestos en la bibliografía para el análisis de las CNVs en el *dataset* seleccionado.
4. Resultados del rendimiento de los algoritmos propuestos.
5. Análisis comparativo de los resultados de los protocolos aplicados y selección del que mejor resultados presente, valorando su posible aplicación en clínica.
6. Memoria del trabajo.
7. Presentación del trabajo.

1.4.3 Asignación temporal de las tareas e hitos propuestos

TAREA	INCIO	FINAL	DURACIÓN (h)
FASE 0	**26/02/2018**	**19/03/2018**	**70**
0.1 Propuesta TFM	26/02/2018	05/03/2018	30
0.2 Planificación temporal TFM	06/03/2018	19/03/2018	40
FASE 1	**20/03/2018**	**23/04/2018**	**95**
1.1 Solicitud acceso *dataset*	20/03/2018	26/03/2018	15
1.2 Preparación datos de partida	27/03/2018	09/04/2018	35
1.3 Búsqueda bibliográfica y selección de 2 protocolos	10/04/2018	23/04/2018	45
FASE 2	**23/04/2018**	**21/05/2018**	**95**
2.1 Elaboración de *scripts*	23/04/2018	07/05/2018	45
2.2 Análisis resultados obtenidos	08/05/2018	15/05/2018	25
2.3 Comparación resultados	15/05/2018	21/05/2018	25
FASE 3: Redacción memoria	**22/05/2018**	**06/06/2018**	**60**
FASE 4: Elaboración presentación	**07/06/2018**	**13/06/2018**	**25**
FASE 5: Defensa TFM	**14/06/2018**	**22/06/2018**	**20**

Tabla 1. *Planificación de la asignación temporal de las tareas propuestas.*

HITOS	FECHA
1. Planificación TFM	19/03/2018
2. Preparación *dataset*	09/04/2018
3. Selección 2 algoritmos	23/04/2018
4. Rendimiento algoritmos propuestos	15/05/2018
5. Análisis comparativo rendimiento protocolos	21/05/2018
6. Memoria TFM	06/06/2018
7. Presentación TFM	14/06/2018

Tabla 2. *Asignación de la fecha de los diferentes hitos propuestos a lo largo del desarrollo del trabajo.*

Diagrama de Gantt:

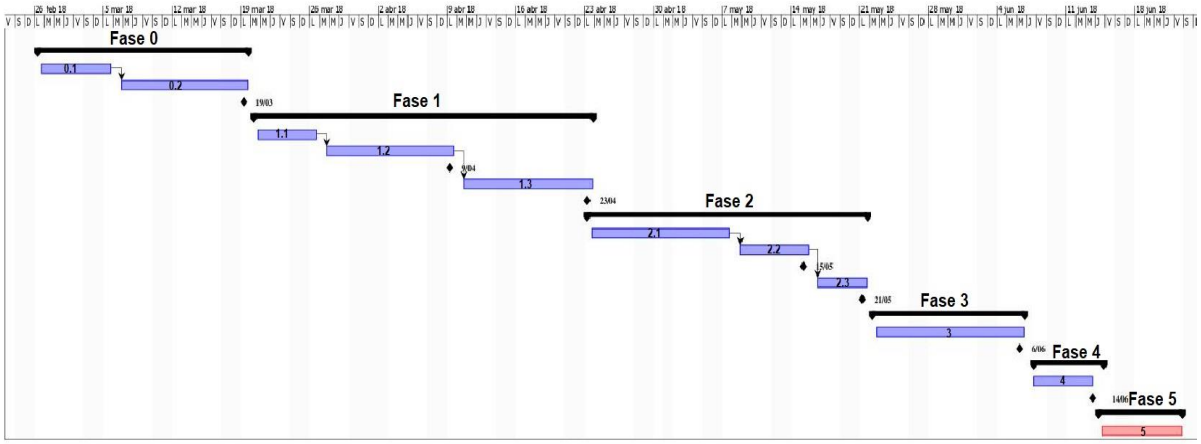


Figura 4. *Diagrama de Gantt donde se observa el tiempo previsto dedicado a cada una de las tareas. La enumeración mostrada en cada barra corresponde a la enumeración de las tareas mostradas en la Tabla 1. Los rombos corresponden a los hitos marcados a lo largo del trabajo que se encuentran reflejados en la Tabla 2. Para la distribución temporal de las tareas se han tenido en cuenta los días festivos y domingos y un tiempo dedicado de unas 4 horas diarias.*

1.4.4 Análisis de riesgos

- No conseguir acceso a la base de datos propuesta: *EGAS00001002428* de la plataforma *European Genome-Phenome Archive* (EGA). En ese caso se buscaría otro *dataset* con características similares.
- No encontrar 2 protocolos adecuados para el análisis de los resultados.
- Los *scripts* creados no funcionan con el *dataset* empleado. No detectan las CNVs o los resultados encontrados no corresponden con los resultados confirmatorios del MLPA.

1.5 Breve sumario de productos obtenidos

- Plan de trabajo
- Protocolo completo del *script* a partir del cual se va a analizar las CNVs de los datos obtenidos de procesos de secuenciación masiva.
- Listado de CNVs detectadas en las muestras procesadas del *dataset* de partida.
- Resultados del rendimiento del algoritmo de detección de CNVs.
- Memoria del trabajo.
- Presentación virtual.
- Autoevaluación del proyecto.

1.6 Breve descripción de los otros capítulos de la memoria

Capítulo 2: Materiales y métodos. Este capítulo está dividido en 3 partes:
- Descripción y preparación de las muestras obtenidas para el análisis.
- Descripción de la búsqueda realizada para la selección de los protocolos.
- Descripción del proceso de análisis del rendimiento de los algoritmos.

Capítulo 3: Resultados. Aplicación de los protocolos seleccionados de detección de CNVs a partir de datos de secuenciación masiva y cálculo del rendimiento obtenido para cada algoritmo.

Capítulo 4: Discusión. Comparativa del rendimiento obtenido de cada algoritmo y selección del que presente mejores resultados.

Capítulo 5: Conclusiones. Presentación de las conclusiones obtenidas tras la realización del trabajo.

Capítulo 6: Autoevaluación del proyecto realizado.

Capítulo 7: Glosario. Listado alfabético de definiciones de términos empleados durante la realización de la memoria del trabajo.

Capítulo 8: Bibliografía. Listado de la bibliografía empleada para la realización del trabajo.

Capítulo 9: Anexos.

2. Materiales y métodos

2.1 Descripción y preparación de las muestras obtenidas para el análisis

Documento de solicitud de acceso a la base de datos EGAD00001003335 del estudio EGAS00001002428.

Se ha solicitado el acceso a la base de datos del estudio *EGAS00001002428* de la plataforma *European Genome-Phenome Archive* (EGA) que contiene el resultado de 96 muestras analizadas mediante el panel *TruSight Cancer* de Illumina (NGS). Para ello, ha sido necesario redactar un documento argumentando los motivos por los cuales se solicitaba acceso a dicha base de datos. Tras la aprobación de la solicitud de acceso, la plataforma EGA proporciona un usuario y contraseña.

El conjunto de datos seleccionados incluye resultados de secuenciación de alta generación de un panel de cáncer hereditario (*Trusight Cancer Panel*) llevado a cabo en un secuenciador Illumina HiSeq 2500. Se trata de un panel diseñado para el estudio de 94 genes y 284 SNPs relacionados con predisposición a cáncer, proporcionando cobertura en las regiones exónicas y las regiones flanqueantes no codificantes de los exones (promedio de 50 pb), cubriendo más de 1700 exones localizados en los genes de estudio[14]. En el **Anexo 1** se adjunta el listado de los genes incluidos en el panel (no se han tenido en cuenta los 284 SNPs ya que no van a ser evaluados en el presente trabajo).

De las 96 muestras, 66 contienen al menos una CNV validada y para 30 muestras se han validado los resultados negativos en 26 genes (Tabla 3). Además, 2 de los 66 individuos tienen 1 CNV en dos genes diferentes, siendo un total de 68 CNVs detectadas[12].

	Pool 1	Pool 2
Gene	Number of exon CNVs	
ATM	1	1
BRCA1	7	8
BRCA2	5	5
CHEK2	2	3
EPCAM	0	1
EZH2	0	1
FH	1	0
MLH1	1	0
MSH2	4	4
MSH6	1	1
NF1	1	0
NSD1	3	3
PALB2	1	0
PMS2	3	2
PTEN	0	1
RAD51C	0	1
RB1	1	0
SDHB	1	1
TP53	1	2
WT1	0	1
Total	33	35
	Samples with no exon CNV	
APC, ATM, BAP1, BARD1, BMPR1A, BRCA1, BRCA2, BRIP1, CDH1, CDK4, CDKN2A, CHEK2, EPCAM, MLH1, MSH2, MSH6, MUTYH, NBN, PALB2, PMS2, PTEN, RAD51C, RAD51D, SMAD4, STK11, TP53	15	15

Tabla 3. *Tabla extraída del artículo de S. Mahamdallie et al.[12] que recoge los resultados obtenidos de la validación mediante MLPA de las CNVs detectadas en el estudio EGAD00001002428 mediante el panel Trusight Cancer de Illumina en 96 muestras.*

Este tipo de paneles dirigidos da lugar a un conjunto de datos más pequeño y manejable que si se compara con enfoques más amplios como pueden ser la secuenciación del genoma completo (WGS) o del exoma completo (WES), permitiendo obtener altas profundidades de lecturas en los genes de interés. Para ello, emplea una metodología de captura por oligonucleótidos específicos (sondas) complementarios a las regiones de estudio[14,15].

Descarga de los archivos fastq

La secuenciación llevada a cabo ha sido *paired end* o secuenciación "de segmentos emparejados" que consiste en secuenciar ambos extremos de cada uno de los fragmentos generados, mejorando la calidad de toda la secuencia y permitiendo un alineamiento óptimo de los fragmentos creados[16]. Es por ello que de este tipo de secuenciaciones se obtienen dos archivos por muestra (*forward* "R1" y *reverse* "R2").

Para la descarga de los 192 archivos *fastq* correspondientes a las 96 muestras se ha empleado el cliente *java* ofrecido por EGA. En el siguiente enlace se muestra las instrucciones que facilita EGA para la descarga de los archivos de su base de datos: https://ega-archive.org/download/using-ega-download-client[17]. El código de descarga ha sido adjuntado en el **Anexo 2**.

Desencriptación y visualización de los archivos fastq

Los archivos descargados se encuentran encriptados con la clave asignada durante la descarga de estos. Para desencriptarlos se ha empleado el *script* que se muestra en el **Anexo 3**. Este código permite la desencriptación de todos los archivos de manera simultánea.

Los archivos *fastq* son ficheros de texto planos que contienen los datos brutos extraídos del secuenciador, almacenando todas las lecturas y la calidad de estas[18]. Cada entrada o lectura de este archivo contiene 4 líneas:

- Identificador de secuencia o título, encabezado siempre por el carácter "@".
- Secuencia biológica compuesta por la combinación de los 4 nucleótidos: Adenina (A), Guanina (G), Citosina (C) y Timina (T). En caso de que el secuenciador no sea capaz de asignar un nucleótido introduce el carácter "N".
- Indicador de fin de secuencia e inicio de calidades asociadas (consiste en un carácter "+").
- Calidad de la secuencia codificado con un conjunto de caracteres ASCII (este campo tiene la misma longitud que la secuencia biológica).

El identificador de secuencia contiene la información de la carrera, tal y como se muestra en la Figura 5.

@<instrument>:<run number>:<flowcell ID>:<lane>:<tile>:<x-pos>:<y-pos> <read>:<is filtered>:<control number>:<sample number>

Element	Requirements	Description
@	@	Each sequence identifier line starts with @
<instrument>	Characters allowed: a–z, A–Z, 0–9 and underscore	Instrument ID
<run number>	Numerical	Run number on instrument
<flowcell ID>	Characters allowed: a–z, A–Z, 0–9	
<lane>	Numerical	Lane number
<tile>	Numerical	Tile number
<x_pos>	Numerical	X coordinate of cluster
<y_pos>	Numerical	Y coordinate of cluster
<read>	Numerical	Read number. 1 can be single read or Read 2 of paired-end
<is filtered>	Y or N	Y if the read is filtered (did not pass), N otherwise
<control number>	Numerical	0 when none of the control bits are on, otherwise it is an even number. On HiSeq X and NextSeq systems, control specification is not performed and this number is always 0.
<sample number>	Numerical	Sample number from sample sheet

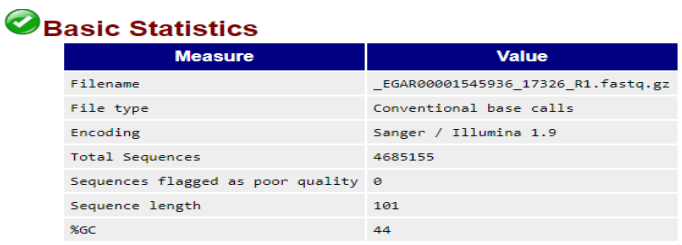
Figura 5. *Descripción de cada uno de los elementos que componen el identificador de secuencia en un archivo .fastq.*

En la Figura 6 se muestra un ejemplo de la visualización de un archivo *fastq*, donde se puede apreciar la entrada de las 4 primeras lecturas.

Figura 6. *Visualización de las 4 primeras entradas de un archivo fastq.*

Análisis de calidad de los datos de partida

Tras descargar y desencriptar los archivos *fastq*, se ha llevado a cabo un análisis para evaluar la calidad de los datos de partida. Para ello, se ha empleado la herramienta *Fastqc*[19] que muestra una serie de parámetros y valores para determinar si los *fastq* cumplen criterios mínimos de calidad (**Anexo 4**). Los resultados obtenidos muestran que los archivos contienen fragmentos de tamaño 101 pb con un contenido en CG entre el 43-45% y una media de 2-4 millones de lecturas por muestra (rango de 1.9 - 7.9 millones). En las siguientes imágenes se muestra un ejemplo del resumen de los datos (Figura 7) y el gráfico de barras (Figura 8) en el cual se observa los valores obtenidos de calidad media para cada posición de las secuencias generadas. Se observa que todas las secuencias son de buena calidad (>Q30), siendo inferior la calidad de las 5 primeras bases.



Basic Statistics

Measure	Value
Filename	_EGAR00001545936_17326_R1.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	4685155
Sequences flagged as poor quality	0
Sequence length	101
%GC	44

Figura 7. *Ejemplo del resumen estadístico de los datos de uno de los archivos fastq descargados. Se observa que el archivo contiene un total de 4685155 lecturas de 101 pb de tamaño máximo y que posee un contenido en GC del 44%.*

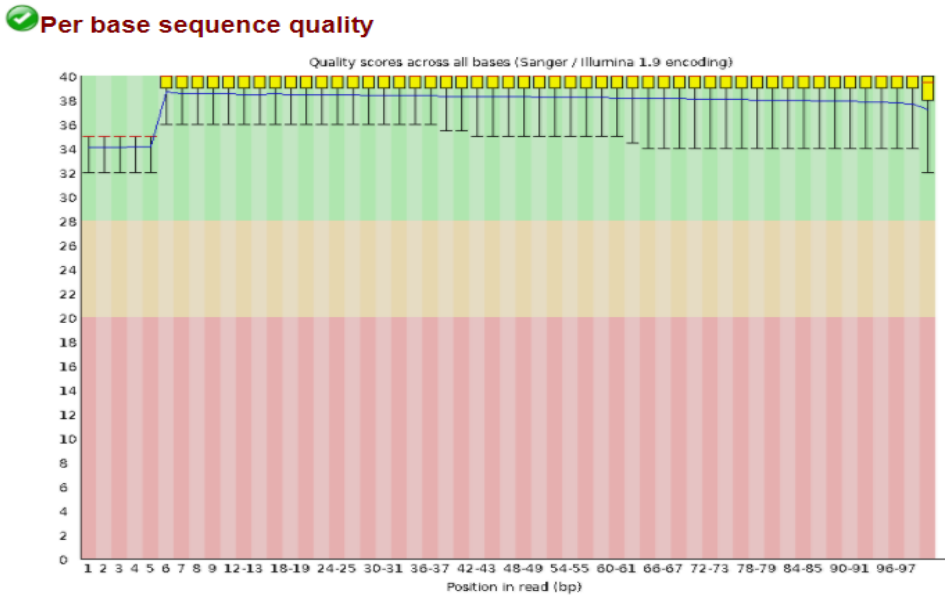


Figura 8. *Ejemplo de la visualización de la calidad media de las secuencias del archivo fastq. Se observa una calidad media de las 5 primeras bases inferior al resto.*

Preprocesado de los datos de partida

Se ha empleado la herramienta *BBmap*[20] para llevar a cabo el filtrado de los datos de partida. Esta herramienta permite eliminar las lecturas que no cumplan los criterios de calidad definidos. Para este estudio se ha establecido que los fragmentos generados tenían que tener mínimo 70 pb y una calidad superior a Q20. En el **Anexo 5** se muestra el código empleado para filtrar los datos con la herramienta *BBmap*.

Se ha observado que tras el filtrado se han eliminado aproximadamente el 5% de las lecturas generadas.

Descarga y puesta a punta de la secuencia de referencia del genoma humano

Se ha empleado la secuencia de referencia del genoma humano versión *hg19* extraída de la plataforma *Ensembl*, ya que los datos de partida incluidos en el *dataset* seleccionado fueron generados usando esta versión. Mediante el algoritmo BWA y la herramienta *SAMtools* se ha llevado a cabo el indexado para la generación de los archivos *.amb*, *.ann*, *.bwt*, *.pac*, *.sa* y *.fai* necesarios para el posterior alineamiento de las muestras frente a la secuencia de referencia. Los comandos llevados a cabo se recogen en el **Anexo 6**.

BWA[21] (*Burrows-Wheeler Aligner*) es un paquete de software que usa la transformada *Burrows-Wheeler* (BWT) para indexar el genoma de referencia y mapear secuencias parecidas contra dicho genoma.

SAMtools [22] es una herramienta que permite manipular alineamientos. Importa y exporta en formato SAM (archivo de texto que contiene los datos del

alineamiento separados por tabulaciones), ordena, une e indexa y permite recuperar las lecturas de cualquier región con rapidez, creando los archivos BAM (archivo binario del archivo SAM).

Este proceso ha generado los siguientes archivos:

- *Ensembl_GRCh37.fa.ann*
- *Ensembl_GRCh37.fa.amb*
- *Ensembl_GRCh37.fa.bwt*
- *Ensembl_GRCh37.fa.sa*
- *Ensembl_GRCh37.fa.pac*

Generación de los archivos .bam y .bam.bai

Tras preparar la secuencia de referencia, se procede al alineamiento de los *fastq* y la generación de los archivos *.bam* y *.bam.bai* necesarios para el análisis de las CNVs. El fichero *.bam* es un archivo binario que contiene la información que tiene el archivo *.sam* (producto del alineamiento de las muestras frente a la secuencia de referencia) de manera comprimida. En cuanto al archivo *.bam.bai* contiene la información indexada del archivo *.bam*, necesario para poder emplear el archivo *.bam* para su posterior análisis. El código de generación de los archivos se encuentra recogido en los **Anexos 7 y 8.**

Los archivos *.bam* contienen 2 secciones[23]:

- Sección de encabezado. Contiene información de la muestra (nombre de la muestra, tamaño, método de alineamiento, etc).
- Sección de alineamiento. Contiene varios parámetros de información para cada par de lecturas:
  AS: Calidad del alineamiento *paired end* o secuenciación “de segmentos emparejados”
  NM: Etiqueta de distancia de edición, que registra la distancia de *Levenshtein* entre la lectura y la referencia.
  XS: Calidad del alineamiento por debajo de lo óptimo.

En la Figura 9 se muestran las 5 primeras entradas de uno de los archivos *.bam* generados, empleándose la herramienta *samtools view* para la visualización.

```
usuario@usuario-X541UV:/media/usuario/Elements/BAMs$ samtools view 17296.bam | more
HWI-D00295:188:HHCL2BCXY:1:1207:10473:27741     163     1       13765   22
101M    =       13827   163     GGCTTCTCACTGGGCCTCTGCAGGAGGCTGCCATTTGTCCTGCCCACC
TTCTTAGAAGCGAGACGGAGCAGACCCATCTGCTACTGCCCTTTCTATAATAA   DDDDDIIIIIIIIIIIIIIIIIII
IIIIIHIIIIIIIIIIIIIIIIIHIIIHIIHHIIIHHHIIIIIGIHIIIIIIIIIIHIIIIIIIIIIHIHHIH
NM:i:0  AS:i:101        XS:i:101
HWI-D00295:188:HHCL2BCXY:2:1212:4102:94070      163     1       13765   22
101M    =       13827   163     GGCTTCTCACTGGGCCTCTGCAGGAGGCTGCCATTTGTCCTGCCCACC
TTCTTAGAAGCGAGACGGAGCAGACCCATCTGCTACTGCCCTTTCTATAATAA   DDDDDIIIIIHIIIIIIIIIIIH
HIIIIIIIIIIIIIIIIHIIIIIIIIIIIIHIIIIIHIIIHIIIIIIIIIIIIIGIIIIIIIIIIIIIIIIIII
NM:i:0  AS:i:101        XS:i:101
HWI-D00295:188:HHCL2BCXY:1:1207:10473:27741     83      1       13827   22
101M    =       13765   -163    ACGGAGCAGACCCATCTGCTACTGCCCTTTCTATAATAACTAAAGTTA
GCTGCCCTGGACTATTCACCCACTAGTCTCAATTTAAGAAGATCCCCATGGCC   HIIIHIHIHIIIIIIIIIIIIIII
HIIIIIHIIIIIIHIIIIIIIIIIIIIIIIIIIIIIIHIIIIIIIIIIIIIIIIIHIIIIIIIIIIHIIIDDDDD
NM:i:1  AS:i:96 XS:i:93
HWI-D00295:188:HHCL2BCXY:2:1212:4102:94070      83      1       13827   22
101M    =       13765   -163    ACGGAGCAGACCCATCTGCTACTGCCCTTTCTATAATAACTAAAGTTA
GCTGCCCTGGACTATTCACCCACTAGTCTCAATTTAAGAAGATCCCCATGGCC   IIIIIIIIIIIIIIIIIIIIIII
GIIIIIIIIIIIGHIIIIIIIIIIIIHIIIIIIIIIIIIIIIIIIIIIIIIIIHIIIIIIIIIIIIIIIIIIIIDBDDD
NM:i:1  AS:i:96 XS:i:93
HWI-D00295:188:HHCL2BCXY:2:1113:3973:99470      99      1       18912   0
34M67S  =       18912   34      CCTGCTTCCATCCTCCCTGGTGCGGGGTGGGCCCCTGTCTCTTATACA
CATCTCCGAGCCCACGAGACTAAGGCGAATCTCGTATGCCGTCTTCTGCTTGA   DDDDDIIIIIIIIIIIIIIIIIII
```

Figura 9. *Visualización de las 5 primeras entradas de uno de los archivos .bam generados.*

Existen diversas herramientas que permiten la visualización gráfica del alineamiento generado, como es el caso de IGV. Se ha empleado esta herramienta (versión 2.4.10) para observar el alineamiento de los genes de interés frente a la secuencia de referencia, tal y como se muestra en la Figura 10.

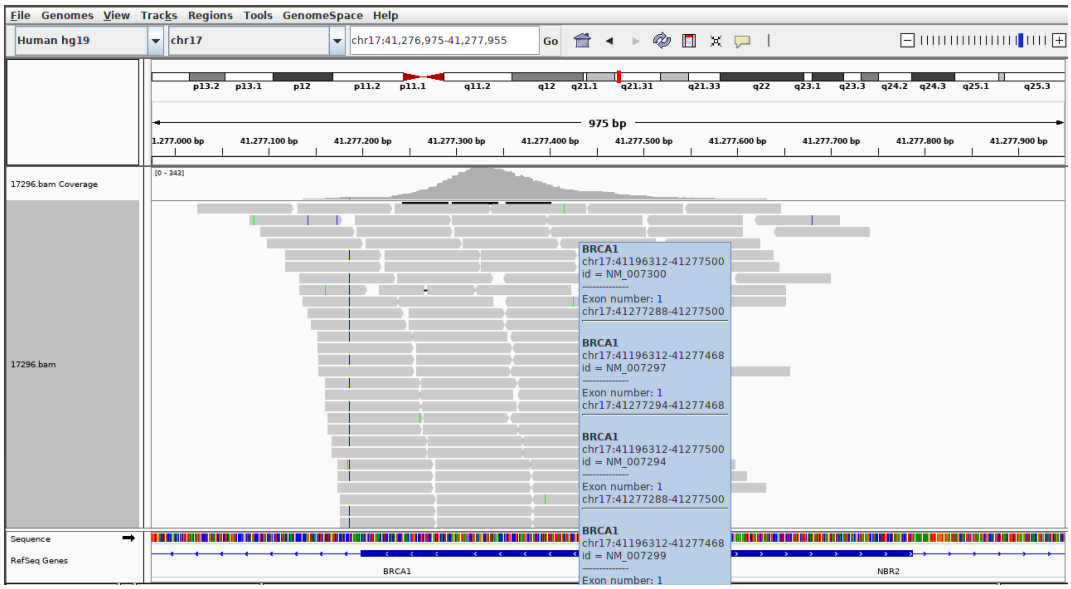
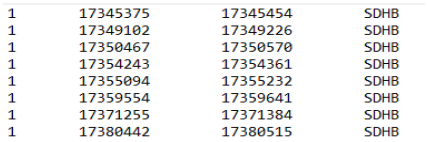


Figura 10. *Visualización gráfica de las lecturas alineadas del exón 1 del gen BRCA1 para una de las muestras estudiadas frente a la secuencia de referencia empleando el programa IGV.*

Obtención del archivo .bed

El formato *.bed*[24] (*Browser Extensible Data*) es un archivo de texto delimitado por tabulaciones que define las líneas de datos que se muestran en una anotación. Este tipo de documentos contiene 3 campos obligatorios (*chrom*, *chromStart* y *chromEnd*) y 9 adicionales (*name*, *score*, *strand*, *thickStart*, *thickEnd*, *itemRgb*, *blockCount*, *blockSizes*, *blockStarts*). Algunos de los algoritmos que se emplearán posteriormente para la detección de CNVs necesitan el archivo *.bed* con las líneas correspondientes a las regiones de estudio. Este archivo ha sido descargado del apartado *Supplementary Material* [25] del artículo de S. Mahamdallie *et al.* [12] que define el *dataset* seleccionado.

El archivo *.bed* descargado ha sido procesado para adecuarlo al formato que necesitan los algoritmos que se han empleado posteriormente. Para ello, se ha eliminado la abreviatura *chr* de la primera columna y la columna número 5 (columna resumen), así como las 284 filas correspondientes a los datos de los SNPs. De este modo, se ha quedado un archivo en formato *.bed* que incluye: número del cromosoma (columna 1), coordenada posición de inicio (columna 2), coordenada posición final (columna 3) y nombre del gen al que corresponde (columna 4). En la Figura 11 se observa las 8 primeras líneas del archivo que indica las coordenadas de cada uno de los exones del gen *SDHB*.



1	17345375	17345454	SDHB
1	17349102	17349226	SDHB
1	17350467	17350570	SDHB
1	17354243	17354361	SDHB
1	17355094	17355232	SDHB
1	17359554	17359641	SDHB
1	17371255	17371384	SDHB
1	17380442	17380515	SDHB

Figura 11. *Visualización de las 8 primeras líneas del archivo .bed correspondientes a las coordenadas de los exones del gen SDHB.*

2.2 Descripción de la búsqueda realizada para la selección de los protocolos

Se ha llevado a cabo una búsqueda bibliográfica en *Pubmed* de algoritmos de detección de CNVs para aplicarlo a datos procesados mediante NGS. Para dicha búsqueda se han empleado las palabras claves "*germline CNV NGS*", "*CNV WES*", "*CNV detection tools NGS*", "*panel trusight cancer CNV*", "*CNV detection panel NGS*", acotando la búsqueda a los 5 últimos años. Entre los algoritmos más destacados encontramos: DECoN, CoNIFER, CONTRA, XHMM, ExomeDepth, CONSERTING, CNVnator, cn.MOPS, panelcn.MOPS, CNVkit, CNVseq, entre otros (Tabla 4).

Plataforma	Herramienta	URL	Año	Lenguaje	*Input*
WES /panel	DECoN	https://www.icr.ac.uk/our-research/research-divisions/division-of-genetics-and-epidemiology/genetic-susceptibility/genetic-data-and-software-resources/decon	2016	Python	BAM
WES /panel	panelcn.MOPS	http://www.bioinf.jku.at/software/panelcnmops/	2017	R	BAM
WES	CoNIFER	http://conifer.sourceforge.net/	2011	Python	BAM
WES	CONTRA	http://contra-cnv.sourceforge.net/	2012	Python	SAM/BAM
WES	XHMM	http://atgu.mgh.harvard.edu/xhmm/index.shtml		C++	BAM
WES	ExomeDepth	http://cran.r-project.org/web/packages/	2016	R	BAM
WGS	CONSERTING	ExomeDepth/index.html		R	BAM
WGS	CNVnator	http://sv.gersteinlab.org/		C++	BAM
WGS	cn.MOPS	http://www.bioinf.jku.at/software/cnmops/	2012	R	BAM/ Matriz de recuento de lecturas
WGS	CNVkit	https://cnvkit.readthedocs.io/en/stable/	2014	Python	R
WGS	CNVseq	http://tiger.dbs.nus.edu.sg/cnv-seq/	2009	Perl, R	Posiciones de las lecturas de alineamiento

Tabla 4. *Resumen de los algoritmos evaluados* [8].

Para seleccionar el algoritmo entre los candidatos presentados, se emplearon los siguientes criterios: algoritmos válidos para WES o paneles de genes de regiones enriquecidas, estudio de CNVs en línea germinal, a ser posible que se hubieran probado en estudios *paired end* y que trabajen en R. Entre todas las herramientas se seleccionó:

- DECoN[26] (*Detection of Exon Copy Number*). Es un algoritmo con una alta sensibilidad y especificidad diseñado para la detección de CNVs a nivel de exón de secuencias específicas a partir de datos de paneles de secuenciación masiva. Se trata de una versión optimizada de ExomeDepth enfocada a paneles de genes y ha sido implementado en R. La versión más reciente es DECoN v1.0.1. Esta herramienta necesita los siguientes archivos de partida para evaluar las CNVs:

  - Secuencia de referencia del genoma humano en formato *fasta* (*Ensembl_GRCh37.fa*) con su correspondiente archivo *fasta.fai* (*Ensembl_GRCh37.fa.fai*).
  - Archivo *.txt* (*BAMlist.txt*) con el listado de muestras (archivos *.bam* y *.bam.bai*) a analizar o la ruta a la carpeta donde están localizadas los archivos *.bam* y *.bam.bai.*
  - Archivo *.bed* (*trusightcancer.bed*) que contiene las regiones de interés del panel de estudio.
  - *Output prefix*. Prefijo establecido para todas las salidas en formato *.RData* generadas.

  El *script* completo empleado para detectar CNVs mediante la herramienta DECoN se encuentra adjunto en el **Anexo 9**.

- Panelcn.MOPS [27] es una versión extendida de cn.MOPs (*Copy Number estimation by a Mixture Of PoissonS*). Ha sido diseñado para detección de CNVs en datos de paneles de secuenciación masiva. Se puede encontrar como herramienta en el paquete de *bioconductor* de R y su versión más reciente es versión 1.0.0.

  - Archivo *.bed* donde se encuentran definidas las regiones de interés a estudiar. Normalmente cada entrada corresponde a un exón de los genes estudiados.
  - Archivos *.bam* y *.bam.bai*

  El *script* completo empleado para detectar CNVs mediante la herramienta panelcn.MOPS se encuentra adjunto en el **Anexo 10**.

2.3 Descripción del proceso de análisis del rendimiento de los algoritmos

Los resultados obtenidos de cada una de las herramientas fueron evaluados en términos de sensibilidad, especificidad, valor predictivo positivo (VPP) y valor predictivo negativo (VPN).

Para evaluar su rendimiento y seleccionar el algoritmo que presente mejores características para su posible implementación en la rutina de detección de CNVs en el laboratorio, se estableció como principal criterio que fuera un algoritmo con una sensibilidad del 100%. El objetivo de establecer este criterio es evitar falsos negativos, para de este modo emplear el algoritmo seleccionado como *screening* previo a la validación de las CNVs con una técnica confirmatorio (MLPA), descartándose así los falsos positivos.

3. Resultados

Los resultados obtenidos mediante la aplicación de los diferentes algoritmos ha sido centrada al estudio de los genes *BRCA1* y *BRCA2* ya que son los que presenta una mayor correlación con el cáncer de mama y/o de ovario.

3.1 DECoN

Tras aplicar la herramienta DECoN sobre las muestras de partida observamos que 3 de ellas fueron excluidas del estudio por no pasar los filtros de calidad (mínimo una correlación con el resto de las muestras del 97% y una cobertura del 100%). El resto de las muestras fueron estudiadas detectándose 23 CNVs localizadas en los genes *BRCA1* y *BRCA2* (Tabla 6).

Las muestras excluidas del estudio se encuentran recogidas en la Tabla 5.

Sample	Details	Correlation	Median Coverage
17339	Low correlation	0.8659387	1864
17383	Low correlation	0.9615838	1511
17394	Low correlation	0.9506863	1420

Tabla 5. Listado de muestras descartadas del análisis realizado con el algoritmo DECoN por presentar una baja calidad y no superar los filtros preestablecidos.

CNV ID	Sample	First exon (BED file)	Last exon (BED file)	Number of exons	Gene	First exon (custom)	Last exon (custom)	Type	Read ratio	Correlation
45	17302	1282	1284	3	BRCA1	18	20	deletion	0.652	0.9981567
78	17305	819	819	1	BRCA2	2	2	deletion	0.565	0.9988169
81	17306	1269	1269	1	BRCA1	5	5	deletion	0.529	0.9991605
103	17317	1285	1285	1	BRCA1	21	21	deletion	0.499	0.9992543
109	17320	1265	1265	1	BRCA1	1	1	deletion	0.526	0.9992108
117	17322	824	826	3	BRCA2	7	9	deletion	0.584	0.9934506
129	17324	830	832	3	BRCA2	13	15	duplication	1.280	0.9945173
135	17326	1267	1267	1	BRCA1	3	3	deletion	0.496	0.9990329
149	17333	1276	1276	1	BRCA1	12	12	deletion	0.609	0.9989288
156	17335	830	832	3	BRCA2	13	15	deletion	0.645	0.9927816
181	17340	1281	1284	4	BRCA1	17	20	duplication	1.360	0.9960683
195	17342	818	827	10	BRCA2	1	10	deletion	0.550	0.9742160
196	17342	830	832	3	BRCA2	13	15	duplication	1.300	0.9742160
221	17357	837	840	4	BRCA2	20	23	deletion	0.575	0.9990406
252	17362	1276	1276	1	BRCA1	12	12	duplication	1.470	0.9978130
268	17364	837	837	1	BRCA2	20	20	duplication	1.330	0.9992442
284	17369	1273	1273	1	BRCA1	9	9	deletion	0.531	0.9985603
295	17371	1269	1269	1	BRCA1	5	5	deletion	0.594	0.9991092
325	17380	818	819	2	BRCA2	1	2	duplication	1.400	0.9984920
351	17384	1277	1277	1	BRCA1	13	13	deletion	0.582	0.9987479
377	17393	818	818	1	BRCA2	1	1	deletion	0.674	0.9971059
402	17398	1276	1276	1	BRCA1	12	12	duplication	1.370	0.9976965
406	17400	1265	1268	4	BRCA1	1	4	deletion	0.523	0.9972016
408	17402	818	818	1	BRCA2	1	1	deletion	0.574	0.9989050
409	17403	1272	1272	1	BRCA1	8	8	deletion	0.555	0.9990406

Tabla 6. *Resultado de las 23 CNVs detectadas con DECoN (14 en BRCA1 y 9 en BRCA2).*

3.2 Panelcn.MOPS

Tras emplear la herramienta panelcn.MOPS, se detectaron un total de 36 CNVs. De las CNVs identificadas, 20 pertenecen a *BRCA1* y 16 a *BRCA2*, tal y como se muestra en la Tabla 7.

Sample	Chr	Gene	CN
17298.bam	13	*BRCA2*	Duplication
17301.bam	17	*BRCA1*	Duplication
17302.bam	13	*BRCA2*	Deletion
17302.bam	17	*BRCA1*	Deletion
17305.bam	13	*BRCA2*	Deletion
17306.bam	17	*BRCA1*	Deletion
17317.bam	17	*BRCA1*	Deletion
17320.bam	17	*BRCA1*	Deletion
17322.bam	13	*BRCA2*	Deletion
17324.bam	13	*BRCA2*	Deletion
17326.bam	17	*BRCA1*	Deletion
17333.bam	17	*BRCA1*	Deletion
17335.bam	13	*BRCA2*	Deletion
17335.bam	17	*BRCA1*	Duplication
17335.bam	17	*BRCA1*	Deletion
17339.bam	13	*BRCA2*	Deletion
17339.bam	17	*BRCA1*	Duplication
17339.bam	17	*BRCA1*	Deletion
17340.bam	17	*BRCA1*	Duplication
17342.bam	13	*BRCA2*	Deletion
17342.bam	17	*BRCA1*	Duplication
17357.bam	13	*BRCA2*	Deletion
17362.bam	17	*BRCA1*	Duplication
17364.bam	13	*BRCA2*	Duplication
17369.bam	17	*BRCA1*	Deletion
17371.bam	17	*BRCA1*	Deletion
17380.bam	13	*BRCA2*	Duplication
17383.bam	13	*BRCA2*	Deletion
17384.bam	17	*BRCA1*	Deletion
17393.bam	13	*BRCA2*	Deletion
17394.bam	13	*BRCA2*	Deletion
17394.bam	17	*BRCA1*	Deletion
17398.bam	17	*BRCA1*	Duplication
17400.bam	13	*BRCA2*	Deletion
17402.bam	13	*BRCA2*	Deletion
17403.bam	17	*BRCA1*	Deletion

Tabla 7. *Resultado de las 36 CNVs detectadas con panelcn.MOPS (20 en BRCA1 y 16 en BRCA2).*

3.3 Rendimiento de los algoritmos propuestos

Para evaluar el rendimiento de los algoritmos presentados se compararon los resultados con las CNVs detectadas mediante una prueba confirmatoria (en el **Anexo 11** se encuentra el listado filtrado de las CNVs detectadas en los genes *BRCA1* y *BRCA2* mediante MLPA) siendo un total de 25 (15 en *BRCA1* y 10 en *BRCA2*).

En la Tabla 8 se muestra un resumen de los resultados obtenido mediante MLPA, DECoN y panelcn.MOPS.

	MLPA	DECoN	panelcn.MOPS
CNVs reales detectadas	25	23	25
Falsos positivos		0	11
Falsos negativos		2	0
Total CNVs detectadas	25	23	36

Tabla 8. *Tabla resumen con los resultados obtenidos de los listados de CNVs analizadas con los dos algoritmos y comparadas con las validaciones por MLPA.*

Los resultados obtenidos fueron comparados en términos de sensibilidad, especificidad, VPP y VPN tras comprarlos con los resultados de MLPA. A continuación, se muestra el rendimiento de los algoritmos evaluados (Tabla 9 y 10).

	DECoN	panelcn.MOPS
Sensibilidad	92%	100%
Especificidad	100%	91,8%
VPP	100%	69,44%
VPN	98,4%	100%

Tabla 9. *Rendimiento obtenido con los algoritmos empleados para la detección de CNVs relacionadas con cáncer de mama y/o de ovario hereditario.*

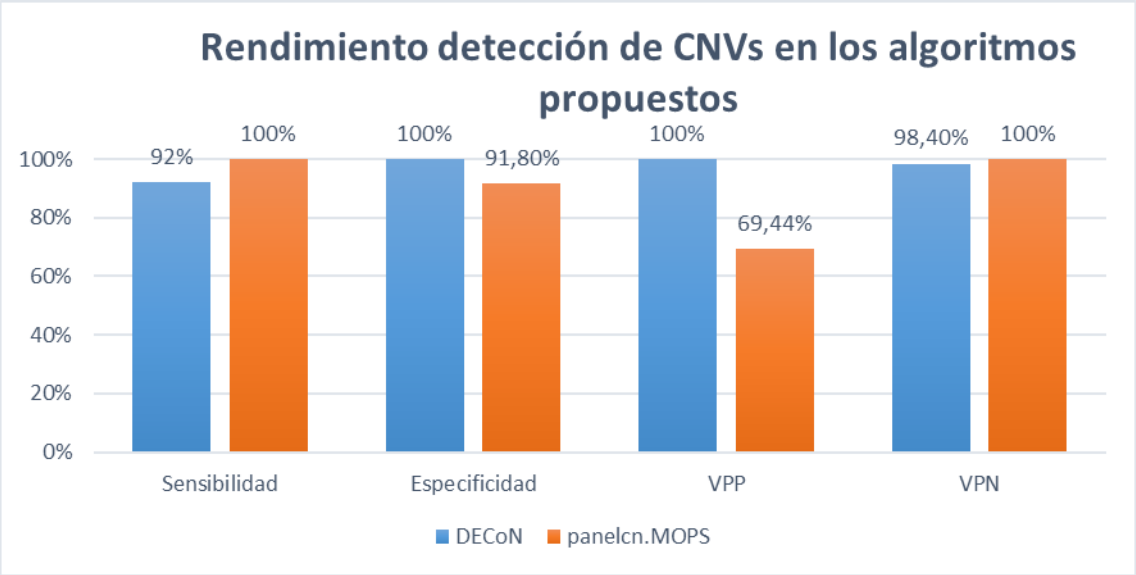


Figura 12. *Evaluación gráfica del rendimiento obtenido con los algoritmos empleados para la detección de CNVs relacionadas con cáncer de mama y/o de ovario hereditario.*

4. Discusión

Los resultados obtenidos mediante la herramienta DECoN muestran un buen rendimiento en la detección de CNVs para el *dataset* empleado. Aun así, no alcanza el 100% de sensibilidad obtenido por el grupo del profesor Nazneen Rahman[28] para la detección de CNVs exónicas en los genes *BRCA1* y *BRCA2.* Los resultados del presente estudio detectan 2 falsos negativos, disminuyendo de ese modo la sensibilidad del algoritmo al 92%. Uno de los casos de falso negativo ha sido debido a que la muestra no ha pasado el filtro de calidad, por lo que en ese caso se recomendaría evaluar las CNVs mediante otro algoritmo o recurrir a una técnica confirmatoria ya que al no pasar el filtro la muestra (17394) no ha sido evaluada. En cuanto a la otra CNV no detectada (muestra 17298), se trata de una CNV en la que está implicada únicamente el exón 21 del gen *BRCA1* por lo que es posible que el tamaño de la misma sea inferior al límite de resolución. En cuanto a los falsos positivos, el algoritmo DECoN no ha detectado ninguno presentando una especificidad del 100%. Este hecho reduciría considerablemente confirmaciones innecesarias que se tendrían que realizar en caso de emplear esta herramienta en clínica, reduciendo de ese modo el coste del estudio.

En cuanto a los resultados obtenidos con la herramienta panelcn.MOPS presenta una sensibilidad del 100%, tal y como indican los autores del algoritmo[29]. Este resultado permite valorar a panelcn.MOPS como posible candidato para su implementación en clínica ya que, tal y como se ha mencionado a lo largo de la memoria, para la introducción de un algoritmo de detección de CNVs en clínica se busca una herramienta que presente una sensibilidad del 100%, evitando errores en el diagnóstico debidos a falsos negativos. De este modo, se puede analizar todos los pacientes mediante esta herramienta y realizar únicamente una prueba confirmatoria en los pacientes cuyos resultados sean positivos para comprobar si son realmente portadores de la alteración de estudio. Por el contrario, cabe destacar que presenta una tasa de falsos positivos muy elevada (11 falsos positivos), lo que implica realizar un mayor número de MLPAs que si se compara con la herramienta DECoN.

Para poder valorar la aplicación de un algoritmo de este tipo en la rutina del laboratorio se recomienda que, a pesar de que presente falsos positivos, tenga una sensibilidad del 100% o muy cercana, para evitar que se escape ningún paciente afecto en el *screening*. Es por ello que se ha considerado que **panelcn.MOPS presenta un mejor rendimiento de detección de CNVs de cáncer de mama y/o de ovario hereditario a partir de datos de NGS.**

A pesar de haber conseguido buenos resultados de los algoritmos propuestos, con rendimientos similares a los indicados en la bibliografía, sería recomendable emplear un conjunto de datos más grande y que incluya un mayor número de genes para de ese modo, evaluar si los resultados obtenidos son consistentes. Así mismo, los algoritmos empleados han sido evaluados con

los parámetros por defecto. Sería recomendable probar otros valores para intentar aumentar el límite de detección, obteniéndose menor número de falsos positivos y/o negativos.

5. Conclusiones

- DECoN y panelcn.MOPs son herramientas útiles para el análisis de CNVs en línea germinal a partir de datos de NGS.
- La herramienta DECoN presenta una sensibilidad del 92% y una especificidad del 100% para la detección de CNVs en el panel empleado.
- La herramienta panelcn.MOPS presenta una sensibilidad del 100% y una especificidad del 91,8% para la detección de CNVs en el panel empleado.
- Ambos algoritmos presentan buenos rendimientos, siendo panelcn.MOPS el que presenta mejores cualidades para emplearlo como técnica de *screening* previo a la confirmación por MLPA de las CNVs detectadas mediante NGS, ya que no da lugar a falsos negativos (sensibilidad 100%).
- A pesar de que los resultados son prometedores, se necesita de más estudios en paneles diferentes y con un mayor número de pacientes y genes a evaluar para valorar la validez del algoritmo panelcn.MOPS y su posible aplicación en la rutina del diagnóstico de cáncer hereditario.

6. Autoevaluación

Evaluando los objetivos iniciales planteados en el apartado 1.2.1 de la presente memoria, podemos afirmar que en términos generales se ha podido cumplir los objetivos propuestos (evaluar diferentes algoritmos de detección de CNVs en línea germinal para el estudio de cáncer de mama y/o ovario).

A pesar de haber cumplido el objetivo principal, se había determinado en los objetivos específicos evaluar 3 algoritmos de CNVs, pero debido a problemas en la instalación y puesta a puesto de los *scripts* de ejecución de las herramientas seleccionadas, solo se ha podido obtener resultados de 2 algoritmos (DECoN y panelcn.MOPS).

Como perspectiva futura se propone ampliar el estudio a un tercer algoritmo que presente un rendimiento igual o superior que el obtenido con las herramientas estudiadas y reevaluar los algoritmos analizados en otros *datasets* y en un mayor número de pacientes para reafirmar la validez de los resultados obtenidos en el presente proyecto.

7. Glosario

ADN. Ácido desoxirribonucleico.

BAM. Archivo binario del archivo SAM.

BBmap. Herramienta bioinformática empleada para filtrar los datos de partida. También sirve para indexar y alinear secuencias frente a la secuencia de referencia.

BRCA1 (*breast cancer 1*). Gen supresor de tumores humano, que regula el ciclo celular y evita la proliferación descontrolada. Relacionado con cáncer de mama.

BRCA2 (*breast cancer 2*). Gen supresor de tumores humano que codificada para una proteína implicada en reparación de ADN. Relacionado con cáncer de mama.

BWA (*Burrows-Wheeler Aligner*). Software que usa la transformada *Burrows-Wheeler* (BWT) para indexar el genoma de referencia y mapear secuencias parecidas contra dicho genoma.

CNV (*Copy Number Variation)*. Variación en el número de copias de un fragmento o región de ADN respecto a la secuencia de referencia del genoma. Las CNVs más frecuentes son deleciones y duplicaciones y comprenden un tamaño de al menos 50 pb.

DECoN (*Detection of Exon Copy Number*). Algoritmo diseñado para la detección de CNVs a partir de datos de paneles de secuenciación masiva.

EGA (*European Genome-Phenome Archive*). Plataforma donde se depositan y se pueden intercambiar todo tipo de datos genéticos y fenotípicos de resultados de proyectos de investigación biomédica.

Fastq. Dato bruto extraído de los procesos de secuenciación masiva que almacena las secuencias de nucleótidos y su calidad de lectura. Se emplea como archivo de partida para el alineamiento generando el archivo SAM.

Fastqc. Herramienta de control de calidad para datos de secuenciación de alto rendimiento (NGS).

Kb. Mil pares de bases (1000 pb).

Mb. Un millón de pares de bases (1000000 pb).

MLPA *(Multiplex Ligation-dependent Probe Amplification).* Técnica diagnóstica para detección de CNVs.

NGS (*Next Generation Sequencing* o secuenciación de segunda generación). Tecnologías destinadas a llevar a cabo la secuenciación masiva a gran escala de cualquier ácido nucleico.

Panelcn.MOPS. Versión extendida del algoritmo cn.MOPs (*Copy Number estimation by a Mixture Of PoissonS*).

pb. Pares de bases.

Script. Archivo de procesamiento (de órdenes).

SAM. Archivo de texto que contiene los datos del alineamiento separados por tabulaciones.

SAMtools*.* Herramienta que permite manipular alineamientos. Importa y exporta en formato SAM, ordena, une e indexa y permite recuperar las lecturas de cualquier región con rapidez, creando los archivos BAM.

SEOM. Sociedad Española de Oncología Médica.

SNP (*Single Nucleotide Polymorphism*). Variación en la secuencia de ADN que afecta únicamente a una base.

VPP. Valor predictivo positivo.

VPN. Valor predictivo negativo.

8. Bibliografía

1. WHO. Disponible en: http://www.who.int/es/news-room/fact-sheets/detail/cancer.
2. Sociedad Española de Oncología Médica (SEOM). Full-Text. *Las cfras del cáncer en España 2018* 24 (2018). doi:M-3161-2018
3. Miki, Y. *et al.* A Strong Candidate for the Breast and Ovarian Cancer Susceptibility Gene BRCA1. *Science (80-.).* **266,** 66-71 (1994).
4. Wooster, R. *et al.* Identification of the breast cancer susceptibility gene BRCA2. *Nature* **378,** 789-92 (1995).
5. Soto, J. L. *et al.* Documento de consenso sobre la implementación de la secuenciación masiva de nueva generación en el diagnóstico genético de la predisposición hereditaria al cáncer. *Med. Clin. (Barc).* (2018). doi:10.1016/j.medcli.2017.12.010
6. Petrucelli, N., Daly, M. B. & Feldman, G. L. Hereditary breast and ovarian cancer due to mutations in BRCA1 and BRCA2. *Genet. Med.* **12,** 245-259 (2010).
7. Castéra, L. *et al.* Next-generation sequencing for the diagnosis of hereditary breast and ovarian cancer using genomic capture targeting multiple candidate genes. *Eur. J. Hum. Genet.* **22,** 1305-1313 (2014).
8. Zhao, M., Wang, Q., Wang, Q., Jia, P. & Zhao, Z. Computational tools for copy number variation (CNV) detection using next-generation sequencing data: Features and perspectives. *BMC Bioinformatics* **14,** S1 (2013).
9. CancerSeq. Disponible en: http://www.grnewsletters.com/archive/biochain/CancerSeqTM-Plus-FFPE-tissues-with-CNVSNPindel-Report-340989702.html.
10. Use, I. F. MLPA ® General Protocol Instructions For Use MLPA (Multiplex Ligation-dependent Probe Amplification) General Protocol for the detection and quantification of DNA sequences. MLPA General Protocol – Document History. 1-16 (2018).
11. Schmidt, A. Y. *et al.* Next-Generation Sequencing–Based Detection of Germline Copy Number Variations in BRCA1/BRCA2: Validation of a One-Step Diagnostic Workflow. *J. Mol. Diagnostics* **19,** 809-816 (2017).
12. Mahamdallie, S. *et al.* The ICR96 exon CNV validation series: a resource for orthogonal assessment of exon CNV calling in NGS data. *Wellcome Open Res.* **2,** 35 (2017).
13. EGAS00001002428. Disponible en: https://www.ebi.ac.uk/ega/studies/EGAS00001002428.
14. Ilumina. TruSight ™ Cancer Sequencing Panel. (2013).
15. Cancer, T. Description of TruSight Cancer. 4-7
16. Illumina. Paired end sequencing. Disponible en: https://www.illumina.com/science/technology/next-generation-sequencing/paired-end-vs-single-read-sequencing.html.
17. Ega-Download-Client Guide. Disponible en: https://ega-archive.org/download/using-ega-download-client.
18. Fastq. Disponible en: http://support.illumina.com/content/dam/illumina-support/help/BaseSpaceHelp_v2/Content/Vault/Informatics/Sequencing_Analysis/BS/swSEQ_mBS_FASTQFiles.htm.
19. FASTQc Guide. https://www.bioinformatics.babraham.ac.uk/projects
20. BBmap Guide. https://jgi.doe.gov/data-and-tools/bbtools/bb-tool

21. BWA. Disponible en: http://bio-bwa.sourceforge.net/bwa.shtml.
22. Samtools Guide. Disponible en: http://www.htslib.org/doc/samtools.html.
23. BAM Format. Disponible en: https://support.illumina.com/help/BS_App_MDProcessor_Online_1000000007932/Content/Source/Informatics/BAM-Format.htm.
24. Bed File. Disponible en: https://genome.ucsc.edu/FAQ/FAQformat.html#format1.
25. Bed Trusight Cancer. Disponible en: https://wellcomeopenresearch.s3.amazonaws.com/supplementary/11689/685e972e-7457-4dbf-8a33-3b91ae4a8f29.txt.
26. Decon, R. *et al.* DECoN v1.0.1.
27. panelcn.MOPS bioconductor. Disponible en: https://bioconductor.org/packages/release/bioc/html/panelcn.mops.html.
28. Fowler, A. *et al.* Accurate clinical detection of exon copy number variants in a targeted NGS panel using DECoN. *Wellcome Open Res.* **1,** 20 (2016).
29. Povysil, G. *et al.* panelcn.MOPS: Copy-number detection in targeted NGS panel data for clinical diagnostics. *Hum. Mutat.* **38,** 889-897 (2017).
30. Gene_List_Trusight_Cancer. Disponible en: http://www.illumina.com/documents/products/gene_lists/gene_list_trusight_cancer.xlsx.

9. Anexos

Anexo 1. *Listado de genes incluidos en el panel Trusight Cancer de Illumina.*[30]

LISTADO DE LOS 94 GENES INCLUIDOS EN EL PANEL *TRUSIGHT CANCER*						
AIP	*CDKN2A*	*EXT2*	*GPC3*	*PALB2*	*SBDS*	*WRN*
ALK	*CEBPA*	*EZH2*	*HNF1A*	*PHOX2B*	*SDHAF2*	*WT1*
APC	*CEP57*	*FANCA*	*HRAS*	*PMS1*	*SDHB*	*XPA*
ATM	*CHEK2*	*FANCB*	*KIT*	*PMS2*	*SDHC*	*XPC*
BAP1	*CYLD*	*FANCC*	*MAX*	*PRF1*	*SDHD*	
BLM	*DDB2*	*FANCD2*	*MEN1*	*PRKAR1A*	*SLX4*	
BMPR1A	*DICER1*	*FANCE*	*MET*	*PTCH1*	*SMAD4*	
BRCA1	*DIS3L2*	*FANCF*	*MLH1*	*PTEN*	*SMARCB1*	
BRCA2	*EGFR*	*FANCG*	*MSH2*	*RAD51C*	*STK11*	
BRIP1	*EPCAM*	*FANCI*	*MSH6*	*RAD51D*	*SUFU*	
BUB1B	*ERCC2*	*FANCL*	*MUTYH*	*RB1*	*TMEM127*	
CDC73	*ERCC3*	*FANCM*	*NBN*	*RECQL4*	*TP53*	
CDH1	*ERCC4*	*FH*	*NF1*	*RET*	*TSC1*	
CDK4	*ERCC5*	*FLCN*	*NF2*	*RHBDF2*	*TSC2*	
CDKN1C	*EXT1*	*GATA2*	*NSD1*	*RUNX1*	*VHL*	

Anexo 2. *Script de acceso a la plataforma EGA y descarga de los 192 fastq solicitados.*

> java -jar EgaDemoClient.jar
> login usuario
Password:
> datasets ## para visualizar los datasets a los que se ha conseguido el permiso de acceso
> files dataset EGAD00001003335 ## Muestra todos los archivos contenidos en el *dataset EGAD00001003335*
> request dataset EGAD00001003335 contraseña-descarga
request_EGAD00001003335 ## nombre descarga
> download request_EGAD00001003335

Anexo 3. *Script de desencriptado de los archivos fastq descargados del dataset EGAD00001003335.*

> java -jar EgaDemoClient.jar -p usuario contraseña-login -dc path-fastq/*.cip -dck contraseña-descarga

Anexo 4. *Análisis de la calidad mediante la herramienta Fastqc*

> fastqc *.fastq.gz –o

Anexo 5. *Preprocesado de las muestras con BBmap*

Estando en la carpeta donde se encuentran los archivos fastq, se indica el siguiente comando para filtrar los datos con BBmap.

```
> for file in *_R1.fastq.gz; do /home/usuario/Descargas/bbmap/bbduk.sh in=$file in2=${file%_R1.fastq.gz}_R2.fastq.gz
out=trimmed/${file%_R1.fastq.gz}_trimmed_R1.fastq.gz
out2=trimmed/${file%_R1.fastq.gz}_trimmed_R2.fastq.gz minlen=70 trimq=20 qtrim=rl; done
```

Anexo 6. *Indexado de la secuencia de referencia del genoma humano (versión hg19) mediante el algoritmo BWA y la herramienta Samtools.*

```
> gzip -d Ensembl_GRCh37.fa.gz
> bwa index Ensembl_GRCh37.fa
> samtools faidx Ensembl_GRCh37.fa
```

Anexo 7. *Alineamiento de las muestras frente a la secuencia de referencia del genoma humano (versión hg19) mediante el algoritmo bwa mem y conversión de los archivos .sam generados a .bam con la herramienta samtools*

Para generar varios archivos .bam simultáneamente se ha creado un archivo en lenguaje Python con el código y se ha ejecutado con el comando.

```
> python bwa_mapping.py
```

El archivo bwa_mapping.py contiene el siguiente script

```
import os, glob

files = glob.glob("*_R1.fastq.gz")

for file in files:
        id = file.split("_") [2]
        R2 = file.replace("_R1", "_R2")
        os.system("bwa mem -t 8 ~/path/Ensembl_GRCh37.fa %s %s | samtools sort -@ 8 -O BAM -o BAMs/%s.bam" % (file, R2, id))
```

Anexo 8. *Generación del archivo .bam.bai mediante la herramienta samtools*

```
> samtools index *.bam
```

Anexo 9. *Script completo para evaluar las CNVs con la herramienta DECoN.*

Abriendo el terminal en la carpeta donde se encuentra los *scripts* de DECoN
> Rscript ReadInBams.R --bams bams.file --bed bed.file --fasta fasta.file --out DECONtest

Salida *ReadInBams* genera un archivo *.RData* de resumen con el prefijo de salida especificado en la entrada que contiene valores de cobertura de muestra y nombres de muestra tomados de los archivos BAM. Este archivo se emplea como entrada en el siguiente paso.
> Rscript IdentifyFailures.R --Rdata *DECONtest.RData* --mincorr *.97* –mincov --out *DECON*

Los valores *–mincorr .97* (umbral mínimo de correlación entre la muestra de prueba y cualquier otra muestra para que se considere correlativa) y *–mincov 100* (umbral mínimo de cobertura, cobertura media mínima para cualquier muestra o exón para que se considere que está cubierta) son los filtros establecidos para la evaluación de la calidad de las muestras cuando son analizadas frente al *.bed.* Si todas las muestras y exones están por encima de los umbrales predefinidos, no se crea salida. Si se identifican muestras y / o exones no óptimos, se crea un archivo de texto separado por tabulaciones que termina en *_Failures.txt.*

makeCNVcalls realiza el análisis de todas las CNVs presenten en las muestras de estudio.
> Rscript makeCNVcalls.R –Rdata DECONtest.RData –out DECON

El siguiente paso abre un navegador web para poder visualizar de manera interactiva los resultados, donde permite filtrar por genes, muestras, etc.
> Rscript runShiny.R –Rdata DECON.RData

Filtrado de los resultados obtenidos con el algoritmo DECoN teniendo en cuenta el ratio de lectura (Read Ratio) y seleccionando las entradas de los genes BRCA1 y BRCA2. Se establecido un valor de ratio inferior a 0.7 para las deleciones y mayor de 1.3 para las duplicaciones.
> awk ‘{if ($16 > 1.3 || $16 <=0.7) print $0) /path/to/DECON_all.txt > DECON_filtrate.txt | awk ‘{if ($17 ==”BRCA1” || $17==”BRCA2”) print $0) /path/to/DECON_filtrate.txt > DECON_allBRCA.txt

Anexo 10. *Script completo para evaluar las CNVs con la herramienta panelcn.mops*

library(panelcn.mops)
data(panelcn.mops)

```
## Read bedfile
bed <- "C:/path/to/trusightcancer_exon.bed"
splitROIs(bed, "newBed.bed")
newBed <- "C: :/path/to/newBed.bed"
countWindows <- getWindows(newBed)

## Read Bamfiles
BAMFiles <- list.files("C:/path/to/panelcnmops/", pattern= ".bam$", full.names =TRUE)
print(BAMFiles)

## Test
test <- countBamListInGRanges(bam.files=BAMFiles, countWindows = countWindows,
read.width = 101)
panelcnmopstest <- test
elementMetadata(panelcnmopstest) <- cbind(elementMetadata(panelcnmopstest),
                        elementMetadata(test))

## Read sampleNames
sampleNames <- colnames(elementMetadata(test))
print(sampleNames)

## Select Genes
selectedGenes <- c("BRCA1", "BRCA2")
print(selectedGenes)

##RESULTS:
for (i in c(1:length(BAMFiles)))
{
 resultlistBRCAs <- runPanelcnMops(panelcnmopstest, testiv=c(i), countWindows =
countWindows, selectedGenes = selectedGenes, I= c(0.025, 0.57, 1, 1.46, 2),
normType = "quant",
                    sizeFactor = "quant", qu = 0.25, quSizeFactor = 0.75, norm = 1,
                    priorImpact = 1, minMedianRC = 30, maxControls = 25, sex =
"mixed")
 resulttableBRCAs <- createResultTable(resultlistBRCAs, panelcnmopstest,
countWindows = countWindows,
                        selectedGenes = selectedGenes,
                        sampleNames)
 write.table(resulttableBRCAs, file=paste(BAMFiles[i], ".CNVBRCAs.txt", sep=""))
}

## Join the individual results of each sample in a single .txt file
setwd("C:/path/to/panelcnmops")
filesBRCAsCNVs<- list.files(pattern=".txt")
allresults<- lapply(filesBRCAsCNVs, function(l) read.table(l))
out <-do.call(rbind, allresults)
write.table(out, "BRCAsCNVresults.txt")
## Filtering results
```

```
BRCAsCNVresults<-          read.table("C:/path/to/panelcnmops/BRCAsCNVresults.txt",
header = TRUE)
BRCAsCNVresults

BRCAsCNVfiltrate <- BRCAsCNVresults [BRCAsCNVresults$CN!= "CN2", ]
BRCAsCNVfiltrate <- BRCAsCNVresults [BRCAsCNVresults$lowQual!= "lowQual", ]
write.table(BRCAsCNVfiltrate, "BRCAsCNVfiltrate.txt")
```

Anexo 11. *Listado de las 25 CNVs encontradas en los genes BRCA1 y BRCA2 confirmadas por la técnica de MLPA.*

SampleID	Gene	MLPAResult	ResultType	ExonCNVType	ExonCNVSize	Chromosome
17298	*BRCA2*	Exon 21 duplication	ExonCNV	Duplication	Single	13
17302	*BRCA1*	Exon 5-7 deletion	ExonCNV	Deletion	Multi	17
17305	*BRCA2*	Exon 3 deletion	ExonCNV	Deletion	Single	13
17306	*BRCA1*	Exon 20 deletion	ExonCNV	Deletion	Single	17
17317	*BRCA1*	Exon 3 deletion	ExonCNV	Deletion	Single	17
17320	*BRCA1*	Exon 24 deletion	ExonCNV	Deletion	Single	17
17322	*BRCA2*	Exon 8-10 deletion	ExonCNV	Deletion	Multi	13
17326	*BRCA1*	Exon 22 deletion	ExonCNV	Deletion	Single	17
17333	*BRCA1*	Exon 13 deletion	ExonCNV	Deletion	Single	17
17335	*BRCA2*	Exon 14-16 deletion	ExonCNV	Deletion	Multi	13
17340	*BRCA1*	Exon 5-8 duplication	ExonCNV	Duplication	Multi	17
17342	*BRCA2*	Exon 1-11 deletion	ExonCNV	Deletion	Multi	13
17357	*BRCA2*	Exon 21-24 deletion	ExonCNV	Deletion	Multi	13
17362	*BRCA1*	Exon 13 duplication	ExonCNV	Duplication	Single	17
17364	*BRCA2*	Exon 21 duplication	ExonCNV	Duplication	Single	13
17369	*BRCA1*	Exon 16 deletion	ExonCNV	Deletion	Single	17
17371	*BRCA1*	Exon 20 deletion	ExonCNV	Deletion	Single	17
17380	*BRCA2*	Exon 1-3 duplication	ExonCNV	Duplication	Multi	13
17384	*BRCA1*	Exon 12 deletion	ExonCNV	Deletion	Single	17
17393	*BRCA2*	Exon 1-2 deletion	ExonCNV	Deletion	Multi	13
17394	*BRCA1*	Exon 1-2 deletion	ExonCNV	Deletion	Multi	17
17398	*BRCA1*	Exon 13 duplication	ExonCNV	Duplication	Single	17
17400	*BRCA1*	Exon 21-24 deletion	ExonCNV	Deletion	Multi	17
17402	*BRCA2*	Exon 2 deletion	ExonCNV	Deletion	Single	13
17403	*BRCA1*	Exon 17 deletion	ExonCNV	Deletion	Single	17