Análisis de herramientas bioinformáticas para la detección de CNVs en muestras de pacientes mediante secuenciación de exoma completo (WES).

Enrique Sevilla Romero
Máster Bioinformática y Bioestadística (UOC-UB)
Área 1; Subárea 6: Bioinformática clínica
Joan Maynou Fernández
Carles Ventura Moreno

En memoria a mi Madre,
por su apoyo incondicional.

Ficha del trabajo final

Título del Autor	Análisis de herramientas bioinformáticas para la detección de CNVs en muestras de pacientes mediante secuenciación de exoma completo (WES).
Nombre del autor:	Enrique Sevilla Romero
Nombre del consultor/a:	*Joan Maynou Fernández*
Nombre del PARA	Carles Ventura Moreno
Fecha de entrega(mm/aa)	
Titulación	Máster universitario en Bioinformática y Bioestadística UOC-UB
Área del trabajo final	Área 1; Subárea 6: Bioinformática clínica Subárea 6: Bioinformática clínica
Idioma del trabajo	Castellano
Palabras Claves	CNVs, NGS, Algoritmo, pipeline, WES, línea somática

Resumen del Trabajo (máximo 250 palabras): Con la finalidad, contexto de aplicación, metodología, resultados y conclusiones del trabajo

En los últimos años la secuenciación masiva (NGS) se ha convertido en la herramienta estándar para el estudio de la variabilidad genética. Se ha demostrado tener un enorme potencial a la hora de detectar variaciones tanto a nivel de un nucleótido (SNV), como de inserciones o delección de 3 -4 nucleótidos, así como variaciones estructurales tanto equilibradas (inversiones y translocaciones) como desequilibradas (delección y duplicaciones)

La secuenciación de los exomas de todo el genoma (WES) se ha convertido en la secuenciación más adecuada para el análisis de CNVs por que obtenemos una muy buena cobertura de las regiones de análisis a un coste menor. El presente Trabajo de fin de Máster (TFM) se ha centrado en la detección y en el análisis de estas variaciones estructurales a nivel de número de copias (CNVs). De muestras secuenciadas por WES. Actualmente, existen dos aproximaciones para la detección de las CNVs a partir de los datos de secuenciación de WES de línea somática; una de ellas es la profundidad de cobertura COD) y la otra es el análisis de mapeo de extremos emparejados PEM.

Las herramientas bioinformáticas que hemos seleccionado para este TFM han sido aquellas que utilizan el método de profundidad de cobertura: ExomeDepth, CNVkit y VarScan. Estas herramientas Bioinformáticas tienen en común que comparan las diferentes profundidades de coberturas obtenidas en la secuenciación de muestras control vs muestras tumorales. Esta comparación se realiza mediante una

segmentación de los exomas a analizar en pequeñas bins/ventanas donde se comparan las coberturas entre muestras tumorales vs muestras control. Las diferencias en las coberturas obtenidas frente a las esperadas se interpretan como ganancias o pérdidas en el número de copias CNVs.
Los datos fueron obtenidos de la base de datos públicas European-Genome phemome Archive (EGA). Se trata de 96 muestras en formato Fastq, muestras control y muestras validadas mediante MLPAs para unas determinadas CNVs. Los datos crudos, fueron analizados y procesados mediante un pipeline in-house para obtener los datos alineados. A partir de los datos alineados se realizo los pipelines correspondientes para la detección de las CNVs para cada una de las tres herramientas.
Con estos datos se realizaron los análisis estadísticos de Sensibilidad y Especificidad para cada herramienta bioinformática y para cada pool. Los resultados para las tres herramientas mostraron valores bajos para la detección de Sensibilidad y especificidad. Sobre todo, para el e pool1 donde CNVkit fue la que mayor Sensibilidad obtuvo con un valor de 0,4. Siendo un valor relativamente bajo para la detección de CNVs. En cambio, para el pool2 tanto CNVkit como VarScan obtuvieron valores algo superiores. En concreto VarScan obtuvo un valor de 0,7 para este pool.
La baja Sensibilidad y especificidad obtenida por las tres herramientas bioinformáticas como la baja tasa de detección de CNVs comunes en las tres herramientas nos hace pensar que la detección y análisis de CNVs mediante estas herramientas tiene que mejorarse y optimizarse para poder obtener unos valores de detección muchos mayores y poder tener una confianza mayor a la hora de confirmar CNVs.

Abstract (in English, 250 Words or less):

In recent years, the next generation sequencing (NGS) has become the standard tool for the study of genetic variability. It has shown to have a huge potential to detect variations at the level of a single nucleotide (SNPs), insertions or deletions of 3 -4 nucleotides, as well as structural variations; balanced (inversions and translocations) and unbalanced (deletions and duplications)
The sequencing of whole-exome sequencing (WES) has become the most suitable sequencing for the analysis of CNVs because we obtain a very good coverage of the regions of analysis at a lower cost. In this work, we have focused on the detection and analysis of these structural variations at the number of copies (CNVs) in samples sequenced by WES. Currently, there are two approaches for the detection of CNVs from WES somatic line data; one of them is the depth of coverage (COD) and the other is the Paired-End Mapper (PEM).

The bioinformatic tools that we have selected for this work have been those that use the depth of coverage method: ExomeDepth, CNVkit and VarScan. These Bioinformatics tools have in common that they compare the different depths of coverage obtained in the sequencing of control samples vs tumour samples. This comparison is made by segmentation of the exomes to be analysed in small bins / windows where the coverage between tumour samples vs control samples is compared. The differences in this coverage are shown as gains or losses in the CNVs.
The samples were obtained from the European-Genome phemome Archive (EGA) public database. It consists of 96 samples in Fastq format, control samples and samples validated by MLPAs for certain CNVs. The raw data was analyzed and processed using an in-house pipeline to obtain the aligned data. This data is processed with the equivalent pipelines for detect the CNVs for each of the three tools.
Statistical analyzes of Sensitivity and Specificity were made for each bioinformatics tool and for each pool. The results for the three tools showed low values for detection of sensitivity and specificity. The pool1 CNVkit was the one with the highest sensitivity with a value of 0.4. On the other hand, somewhat higher values were obtained in the pool2, for both CNVkit and VarScan. Specifically, VarScan obtained a value of 0.7 for this pool.
It is thought that the low sensitivity and specificity obtained by the three bioinformatics tools, such as the low detection rate of common CNVs in the three tools, have to be improved and optimized in order to obtain better values to be able to have greater confidence when it comes to confirming

Índice

1. Introducción.

1.1 Contexto y Justificación del Trabajo.

La tecnología de next generation sequencing (NGS) ha reemplazado rápidamente a la Secuenciación Sanger en la detección de variantes genéticas en el diagnóstico genético. Siendo una herramienta fundamental en la detección de pequeñas mutaciones (SNVs), inserciones / deleciones (Indels) así como variantes genómicas de más de una 1kbs (CNVs).(1)

La técnica más empleada hasta ahora para la detección de CNVs ha sido los CGH arrays(2) y MLPAs, (3) pero estas técnicas tienen limitaciones en cuanto al número de bases que son capaces de detectar, no pueden bajar de 10 kbs; como en sus costes tanto económicos como de tiempo.

En los últimos años, se ha visto que las CNVs están teniendo una relación importante en la variación genómica/genética. Se ha relacionado ciertas variaciones genómicas con enfermedades neurológicas como Parkison, Hirschspring, Alzheimer o Esquizofrenia y se ha comprobado una relación estrecha con el Cáncer(4). La fuerte demanda de análisis de CNVs basado en NGS ha impulsado el desarrollo de numerosos métodos computacionales y herramientas para la detección de CNVs(4).

Recientemente, la secuenciación del exoma completo(WES) se ha convertido en la estrategia principal para secuenciar muestras de pacientes de línea somática para analizar sus alteraciones genómicas(4).Sin embargo, tanto la secuenciación de WES como la complejidad de las muestras presentan retos importantes que deben ser solucionados para poder dar resultados fiables y precisos. En la actualidad existen diferentes tipos de herramientas bioinformáticas que van a tratar de solucionar estos restos pero cada una de ellas aborda el problema de forma diferente(5).

En este TFM se realizará un análisis de varias herramientas bioinformáticas que utilizan la metodología de profundidad de cobertura (DOC) para la detección de CNVs. Los datos analizados son datos obtenidos por secuenciación NGS de pacientes secuenciados por WES de línea somática. Los algoritmos que vamos a analizar son VarScan(6), ExomeDepth(7) , CNVkit(8)**.** Se realizará un análisis estadístico de la Sensibilidad y especificidad de detección de CNVs. Con estos datos determinaremos qué herramienta bioinformática es la más adecuada para el análisis de muestras clínicas.

1.2. Justificación del TFG.

Se ha elegido este tema por su utilidad clínica y traslacional. Se trata de encontrar la herramienta bioinformática que mejor se adapte a este tipo de muestras y que esta pueda detectar CNVs con una fiabilidad y precisión alta. Ya que estos

resultados obtenidos en del proceso bioinformático van a tener una repercusión clínica en el diagnóstico genético de los pacientes.

2. Objetivos.

2.1. Objetivos generales.

El objetivo del presente TFM es comparar varias herramientas bioinformáticas de detección de CNVs sobre muestras reales obtenidas de la base de datos publica European-Genome phemome Archive (EGA). De este modo podremos evaluar el rendimiento de cada herramienta bioinformática en cuanto a la Sensibilidad y especificidad con el objetivo de determinar que herramienta bioinformática es la más adecuada.

2.2. Objetivos específicos.

Los objetivos específicos son los que se detallan a continuación:

a) Seleccionar y obtener acceso a un conjunto de datos de WES de línea somática secuenciados en NGS, accesible públicamente, y obtener los mismos datos validados por MLPAs. Estos datos se utilizarán como valores de referencia. La base de datos a partir de donde se obtienen los datos fue European-Genome phemome Archive (EGA).
b) Determinar qué herramientas bioinformáticas, de toda la bibliografía publicada, se ajusta a la temática del TFM: detección de CNVs mediante análisis de profundidad de cobertura (DOC) de muestras de línea somática y secuenciadas a nivel de exón (WES) a partir de datos secuenciados por (NGS).
c) Los datos crudos en formato fastq se procesaran mediante un a pipeline in-house hasta obtener datos alineados y optimizados para ser implementados en las diferentes herramientas bioinformáticas seleccionadas.
d) Estos datos obtenidos por las diferentes herramientas bioinformáticas se compararán con los datos validados por MLPAS para comprobar la especificidad y Sensibilidad de cada herramienta bioinformática.
e) Realizaremos dos comparaciones; una a nivel de genes mediante diagramas de Venn , buscando los genes que se detectan en común y otra a nivel de CNVs detectado por la herramienta y que esta CNV tiene que ser detectada en la misma muestra, tiene que ser el mismo gen y el mismo tipo de delección /amplificación que la detectada por el trabajo mediante MLPAs. Muestra/tipo.

3.Enfoque y método.

3.1 Enfoque.

Aunque la identificación de CNVs a partir de secuencia de exoma sigue siendo un desafío, una estrategia que está dando muy buenos resultados a la hora de detectar CNVs son las herramientas que utilizan los enfoques de profundidad de lectura (DOC) ya son capaces de correlacionar número de lecturas con CNVs. Además, para disminuir al máximo las posibles variabilidades producidas por los pasos de captura y de secuenciación, se analizarán y se comparará en el mismo experimento una combinación de muestras (control/mutado). Las herramientas bioinformáticas elegidas son: VarScan(6), ExomeDepth(7) y CNVkit(8)**.**

Cada herramienta informática trata los datos de forma específica, ya sea aplicando diferentes modelos estadísticos, normalización de las muestras frente a muestras control, segmentación de los datos, etc… Y, por consiguiente, tendremos variabilidad en la identificación de CNVs. Por ello, la importancia de comparar y analizar diferentes herramientas bioinformática para poder determinar cuál de ellas, es la más precisa para poder dar el mejor diagnóstico genético.

3.2 Método a seguir.

3.2.1 Obtener datos de secuencian por WES de línea somática secuenciados por NGS, los datos se van a obtener de la página web publica European-Genome phemome Archive (EGA). Los datos se descargarán en formato fastq. Estos datos crudos se analizan y se comprueba su calidad de secuencia. Si fuera necesario se procesarán para obtener la calidad suficiente para ser procesados por las herramientas bioinformáticas.

3.2.2 Seleccionar varias herramientas bioinformáticas para trabajar en con estos datos.

3.2.3 Procesar los datos con las diferentes herramientas bioinformáticas y obtener las CNVs que identifican los algoritmos. Estos datos obtenidos por las diferentes herramientas bioinformáticas se analizarán estadísticamente para obtener la Sensibilidad y especificidad de las herramientas bioinformáticas.

4. Planificación con hitos y temporización.

4.1 Tareas

Los 4 objetivos detallados en el apartado de objetivos 2.2 se han llevado acabo siguiendo la siguiente estimación:

1. Búsqueda bibliográfica de trabajos científicos, mediante el buscador pubmed, relacionados con el tema del trabajo "Análisis de Herramientas bioinformáticas para la detección de CNVs en secuenciación de exoma completo (WES)". (60 horas)
2. Determinar qué herramientas bioinformáticas, de toda la bibliografía publicada, se ajustan a la temática del TFM: detección de CNVs mediante análisis de profundidad de cobertura, DOC, de muestras de línea somática y secuenciadas a nivel de exón, WES a partir de datos de panel (NGS). (60 horas)
    i. Realizar un análisis de todas las herramientas bioinformáticas analizando los diferentes tipos de análisis estadísticos que realizan y como solucionan los problemas de variabilidad de la muestra. (40 horas.)
    ii. Elegir 3 algoritmos, de toda la bibliografía analizada, que puedan ser buenos candidatos para la comparativa estadística que se va a realizar. (20 horas.)
3. Seleccionar y obtener acceso a un conjunto de datos de panel NGS, accesible públicamente, a partir del cual se puedan obtener los resultados CNVs validados. (130 horas)
    i. Búsqueda de datos de secuenciación de WES de línea somática secuenciados por NGS en las bases de datos públicas European-Genome phemome Archive (EGA), datos crudos en formato FastQ. (100 horas.)
    ii. Procesar los datos mediante herramientas bioinformáticas hasta obtener los archivos bam; de muestras normales/tumorales. (40 horas.)
4. Procesar las muestras, mediante estas herramientas bioinformáticas y obtener los datos de los análisis (90 horas)
    i. Descarga y puesta a punto de las 3 herramientas bioinformáticas. Comprobación de su funcionamiento. (20 horas.)
    ii. Procesamiento las muestras de estudio y obtención de los resultados. (70 horas.)

5. Analizar estadísticamente los resultados y determinar cuál de ellos es el más preciso y sensible al comparar los datos obtenidos por las diferentes herramientas bioinformáticas contra las CNVs del control de referencia. (100 Horas)
    i. Realizar los test siguientes;
        a) Anotación de los CNVs detectados por las herramientas bioinformáticas calcularemos la Sensibilidad, especificidad y ratios de falsos positivos.
        b) Comparación del número de CNVs. Representaremos las delecciones de genes y amplificaciones de genes, que se detectaran mediante un diagrama de VENN.
6. Redacción de la memoria del TFM. (75 horas.)

7. Elaboración de la presentación. (35 horas.)

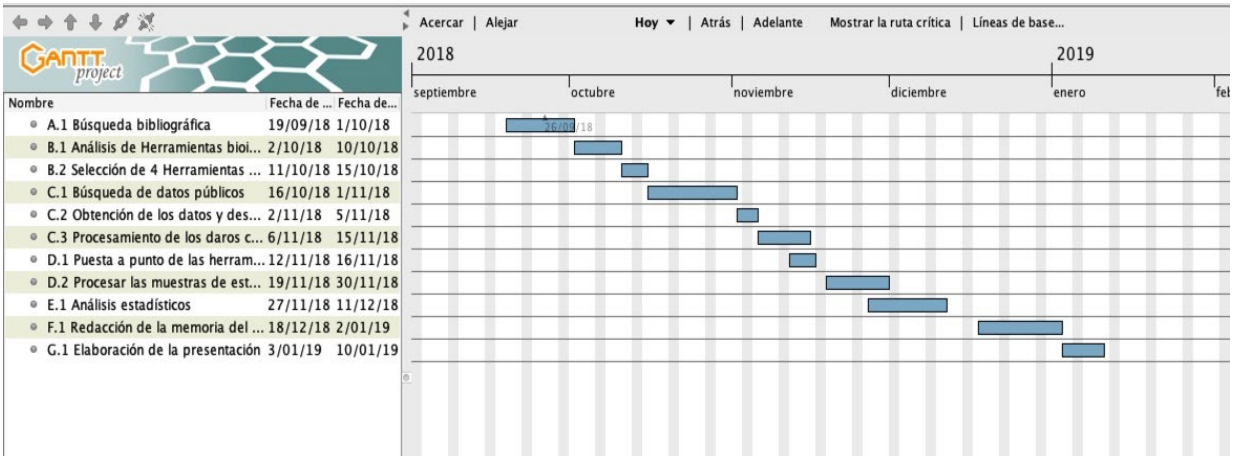
4 Calendario.

Las dos siguientes tablas 1 y gráfica 2 , muestran el calendario con el desglose de las tareas programadas y las fechas que comprenden, también la correspondiente fase de desarrollo. En la tabla 1 se muestra las tareas programadas.

Descripción de la tarea	Tarea	F. inicio	F. Finalización	Fase Desarrollo	Días	Horas	Total
Búsqueda bibliográfica	A.1	19-sept	1-oct	1	12	5	60
Análisis de Herramientas bioinformáticas	B.1	2-oct	10-oct	1	8	5	40
Selección de 3 Herramientas bioinformáticas	B.2	11-oct	15-oct	1	4	5	20
Búsqueda de datos públicos	C.1	16-oct	1-nov	1	16	5	80
Obtención de los datos y descarga	C.2	2-nov	5-nov	1	4	5	20
Procesamiento de los datos crudos.	C.3	6-nov	15-nov	1	6	5	30
Puesta a punto de las herramientas bioinformáticas	D.1	12-nov	16-nov	2	5	5	25
Procesar las muestras de estudio	D.2	17-nov	30-nov	2	14	5	70
Análisis estadísticos	E.1	27-nov	17-dic	2	20	5	100
Redacción de la memoria del TMF	F.1	18-dic	2-ene	3	15	5	75
Elaboración de la presentación	G.1	3-ene	10-ene	4	7	5	35
							555

Tabla 1. Tareas programadas del TFM

A continuación, se muestra el diagrama de Gantt , gráfica 1, con la relación de las tareas programadas y el tiempo dedicado a cada una de ellas.



Gráfica 2. Gráfica de Gantt. Descripción de las tareas del TFM

4.3 Hitos

A.1 Búsqueda bibliográfica. Elección del tema a desarrollar en el TFM.
B.1 Análisis de herramientas bioinformáticas. Analizar que herramientas bioinformáticas, de todas las existentes públicamente, se ajustan a nuestro TFM. Elegir las herramientas que procesen los datos WES de línea somática secuenciados por NGS
B.2 Selección de 3 Herramientas. Seleccionar 3 herramientas que trabajen con las condiciones descritas anteriormente y que utilicen el método de análisis de profundidad de lectura (DOC) para la obtención de las CNVs.
C.1 Búsqueda de datos públicos. Obtener de los datos de la página web publica European-Genome phemome Archive (EGA).
C.2 Obtención de los datos y descarga. Obtener los datos de 96 pacientes controles y con CNVs, procesados por WES de línea somática secuenciados por NGS.
C.3 Procesamiento de los datos crudos mediante herramientas bioinformáticas hasta obtener unos datos de calidad suficiente y formato adecuado para su utilización en las herramientas seleccionadas.
D.1 Puesta a punto de las herramientas bioinformáticas. Comprobar que las herramientas bioinformáticas funcionan correctamente.
D.2 Procesar las muestras de estudio. Procesar los datos obtenidos en cada una de las herramientas bioinformáticas y obtención de las CNVs.
E.1 Análisis estadísticos. Con los datos obtenidos en el apartado anterior realizar la estadística mediante el paquete R.
F.1 Redacción de la memoria del TMF. Presentar las conclusiones.
G.1 Elaboración de la presentación.

4.4. Análisis de riesgos.

Entre los riesgos que podemos encontrar en el desarrollo del TFM están:

Problemas en la búsqueda y/o obtención de los datos públicos para el estudio.

Posibilidad de poder cambiar de base de daos públicos y formato de los datos para el análisis.

Retraso en la puesta a punto y depuración de los algoritmos utilizados.

Retraso en la realización de la estadística.

5. Resultados esperados.

Los resultados esperados serán presentados mediante una tabla de rendimiento general de las herramientas bioinformáticas analizadas. En dicha tabla se expresará la Sensibilidad, especificidad y ratio de falsos positivos en porcentajes para cada una de las herramientas bioinformáticas analizadas. También, se presentará un diagrama de Venn, donde se representará las intersecciones entre las diferentes herramientas bioinformáticas y los datos del trabajo.

2.Material y Métodos.

2.1. Obtención de las muestras.

Los datos han sido obtenidos del European-Genome phemome Archive (EGA) con el nombre del repositorio EGAS00001002428. (9) Este organismo promueve la distribución y el intercambio de datos genéticos y fenotípicos consentidos para usos aprobados específicos, pero no totalmente abiertos, de distribución pública. El EGA sigue protocolos estrictos para la gestión de la información, el almacenamiento de datos, la seguridad y la difusión. El acceso autorizado a los datos se gestiona en asociación con las organizaciones que proporcionan datos.

EGA fue lanzado en 2008 por el Instituto Europeo de Bioinformática del Laboratorio de Biología Molecular (EMBL-EBI) para almacenamiento y distribución solo a usuarios autorizados. Recientemente, EGA se ha expandido de un proyecto exclusivamente EMBL-EBI a una colaboración con el Centro para la Regulación del Genoma (CRG) en Barcelona, España, en lo que puede ser un primer paso hacia una red más amplia de servicios de archivo y difusión de datos. Tanto EMBL-EBI como el CRG son organizaciones financiadas con fondos públicos.

Todos los pacientes dieron su consentimiento informado para el uso de su ADN en la investigación genética. Los estudios han sido aprobados por el Comité de Ética de Investigación Multicéntrico Londres (MRec / 01/2/18, MRec / 01/2/044, 05 / MRE02 / 17, respectivamente).

Para el estudio del ICR96 exón CNV se ha usado el Panel cáncer TruSight v2 (TSCP) que se dirige a los exones de 100 genes de predisposición al cáncer(archivo). Se prepararon las bibliotecas de ADN específicas a partir de 50 ng de ADN genómico utilizando el kit de captura TSCP y TruSight Rapida (Illumina). Siguiendo el protocolo del fabricante, con la excepción de la preparación de la biblioteca. Se ha preparado en dos tantas de 48-Plex. Hemos secuenciado una biblioteca en un equipo de secuenciación HiSeq 2500. Se utiliza el Kit HiSeq® rápido SBS v2, con lecturas de 101 bps de extremos emparejado. La incorporación de los adaptadores a la libraría genera se hace mediante el kit HiSeq PE Cluster Kit v4 cBot y el demultiplexado se realiza con el software Cluster HiSeq® rápido PE v2. V1.8.1 Casava (Illumina) para crear archivos FASTQ.

2.2 Descarga y tratamiento de los Datos.

2.2.1. Dataset Validado.

Para poder acceder a los datos hay que redactar un documento justificando la motivación de su uso para que los administradores del repositorio nos den un acceso a los fastq. Pagina de descarga

El benchmarck contiene 96 muestras de pacientes analizadas mediante el kit Trusight Cancer Panel, ver tabla pero las muestras también fueron analizadas mediante la técnica Multiplex Ligation-dependent Probe Amplification (MLPA). De este modo podemos confirmar los datos obtenidos mediante el panel.
De las 96 muestras analizadas, 66 contiene al menos un CNV validado. El resto de las muestras, 30 no presentaban ninguna CNV validada en los 32 genes analizados. Los genes analizados son; APC, ATM, BAP1, BARD1, BMPR1A, BRCA1, BRCA2, BRIP1, CDH1, CDK4, CDKN2A, CHEK2, EPCAM (exón 9 only) , EZH2, FH, MLH1,MSH2, MSH6, MUTYH, NBN, NF1, NSD1, PALB2, PMS2 (excluding exons 12-15), PTEN, RAD51C, RAD51D, RB1, SDHB, SMAD4, STK11, TP53 and WT1, lo que proporciona una excelente representación de los genes de predisposición al cáncer probados con mayor frecuencia en la práctica clínica. Ver tabla 2.

Nº de muestras	96
Positivos	66 positivos: 1. 24 single-exon 2. 42 multi-exons
Negativos	30 negativos
Genes 32	*APC, ATM, BAP1, BARD1, BMPR1A, BRCA1, BRCA2, BRIP1, CDH1, CDK4, CDKN2A, CHEK2*, *EPCAM* (exón 9 only), *FH, MLH1*, *MSH2*, *MSH6*, *MUTYH, NBN, NF1, NSD1, PALB2, PMS2* (excluding exons 12-15), *PTEN, RAD51C, RAD51D, RB1, SDHB, SMAD4, STK11, TP53* and *WT1*.
Panel y Secuenciador	TruSight Cancer Panel v2 (100 genes) en HiSeq 2500

Tabla 2. Dataset de European-Genome phemome Archive (EGA) con el nombre del repositorio EGAS00001002428

El conjunto de datos completo incluye los archivos FASTQ, 2 archivos por muestra el forward y el reverse, el archivo Bedfile y los resultados de las MLPAs. Ver tabla 3

MLPA resultados para cada pool en el ICR96 exon CNV Validada		
Gen	pool1	pool2
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

Tabla 3. Genes analizados en el benchmarck es el ICR 96 Exón CNV validación

La validación de las muestras mediante la técnica de MLPAs nos permite tener controles positivos y controles negativos para poder posteriormente analizar estadísticamente si las herramientas bioinformáticas que vamos a analiza tienen un rendimiento mayor o menor a la hora de detectar las CNVS y cuantos falsos positivos o verdaderos negativos obtiene.

2.2.2. Descarga y puesta a punto del dataset

Una vez obtenido el permiso de acceso al dataset desde el repositorio EGAS00001002428, mediante la cumplimentación correspondiente, nos dieron un *username* y un *login* para poder tener acceso a los datos y poder descargarlos.

2.2.3 Descarga del dataset

En el repositorio, el dataset se compone de 192 ficheros Fastq encriptados con la extensión *.cip. dos por cada muestra. Para la descarga de todos los fastq ha sido necesario descargar el programa, cliente java, ofrecido por EGA y seguir los pasos descritos:

1. Lanzamos el programa con:

   $ java -jar EgaDemoClient.jar

2. Entramos dentro del programa EGA.

   EGA > login namexxx@uoc.edu

   Password: xxxxx

   Login Success!

3. Comprobamos el dataset que vamos a descargar es el correcto. En nuestro caso es el EGAD00001003335.

   EGA > datasets

   EGAD00001003335

4. Llamada a la dataset EGAD00001003335

   Introducimos la orden para la solicitud de la descarga.

   EGA > request dataset EGAD00001003335 abc
   request_EGAD00001003335

   Requesting....

Resulting Request:

request_EGAD00001003335 (1 new request(s)).

La clave para desemcriptar los archivos es “abc”.

5. Descarga de los archivos.

Mediante la orden download, descrita abajo, los archivos se descargan en la carpeta correspondiente donde se ha abierto el programa EGA. En nuestro caso es;

~Archivos desencriptados.cip

EGA > download request_EGAD00001003335

6. Comprobación de la descarga de los archivos.

Se comprueba que se han descargado los archivos en la carpeta determinada. Los archivos de descargan con en formato; “EGAR00001546001_17403_R1.fastq.gz.cip”

Hay que comprobar si se han descargado todos los archivos, son 192.

En este punto se ha tenido muchos problemas con la descarga ya que la descarga daba errores de conexión. Se ha ejecutado el código más de 10 veces y no se ha podido descargar todos los archivos.

Solo se han podido descargado 178 muestras. 89 muestras pareadas. Entre estas muestras tenemos 55 muestras validadas por MLPAS que nos servirán de control positivo y 34 muestras normales. Estos datos han sido revisados y confirmados con los archivos facilitados por el trabajo.

2.2.4.Desencriptación de los archivos.

Ahora realizamos la desencriptación de los archivos para pasarlos de *.cip a fastq.gz. El script que lanzamos es ;

Para la primera muestra.

$ java -jar EgaDemoClient.jar -p esevillar@uoc.edu EnricyRaul2 -dc EGAR00001546001_17403_R1.fastq.gz.cip -dck abc

Ega Demo Download Client Version: 2.2.3

Verbose set: true

Decrypting 1 file(s).

Logout!

Para la segunda muestra.

$ java -jar EgaDemoClient.jar -p esevillar@uoricyRaul2 -dc EGAR00001546001_17403_R2.fastq.gz.cip -dck abc

Ega Demo Download Client Version: 2.2.3

Verbose set: true

Decrypting 1 file(s).

Logout!

Estos dos archivos son de un mismo sujeto, pero con dos reads de sentido de lectura diferente, una R1 en forward y otra R2 en reverse. Siendo muestras emparejadas/pareadas.

2.2.5.Descomprimir los archivos descargados.

Todos los archivos están comprimidos en formato .gz. El script que lanzamos para cada muestra es;

tar -xzvf EGAR00001546001_17403_R1.fastq.gz

2.3. Procesamiento de los datos

Los datos crudos han sido analizados y tratados bioinformáticamente hasta obtener los archivos en formato bam, ver figura 1, se muestra una representación gráfica de como se va a realizar.

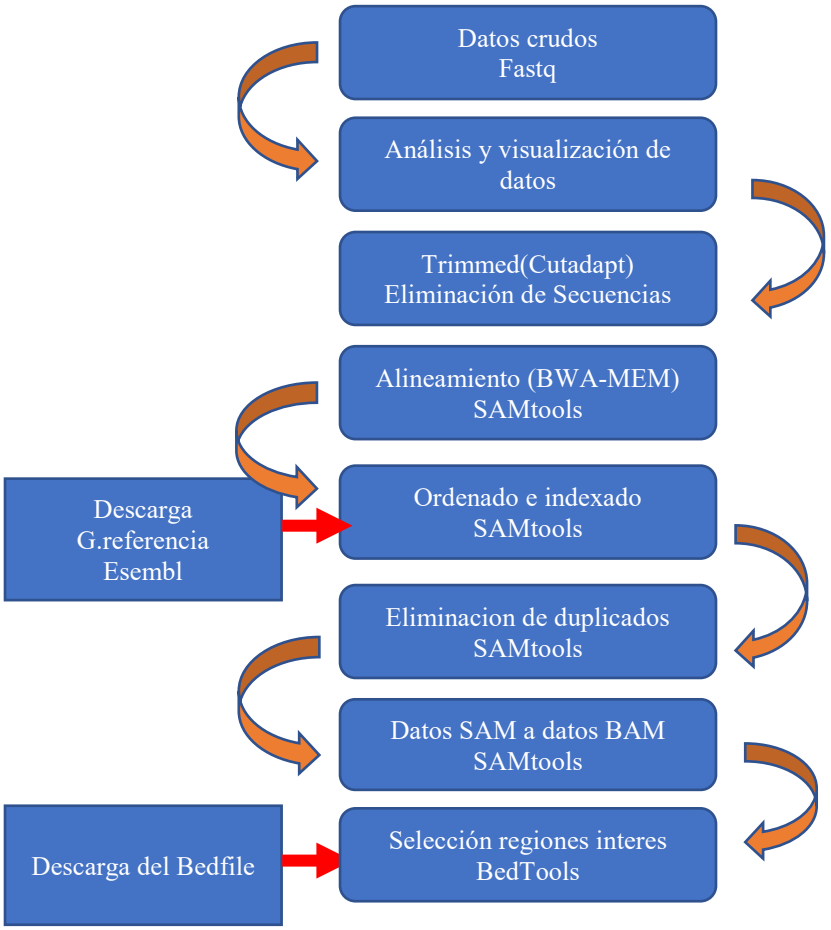


Figura 1. Representación del procesamiento de los datos crudas

2.3.1. Control de calidad

Los archivos con los que vamos a trabajar son archivos en bruto enformato fastq y obtenidos directamente desde el repositorio de EGA. El primer paso que vamos a realizar es un control de la calidad de los archivos fastq. Esto lo realizaremos mediante el programa, FastQC(10). El programa nos va a mostrar una serie de parámetros y valores, a través de los cuales vamos a poder determinar si los archivos cumplen los parámetros de calidad mininos para poder continuar con los análisis posteriores.

Cuando los archivos cumplen la calidad deseada, todos los parámetros que muestra el FastQC deben estar en verde o como mínimo en naranja. Los parámetros que vamos a revisar son Basic Statistics, para comprobar la información básica de secuencia, la calidad de la secuencia de cada lectura, el contenido de bases desapareadas por muestra y por último, los duplicados de PCR. Estos dos ultimas puntos los tenemos en rojo por lo que tendremos que solucionarlo mediante los programas bioinformáticas correspondientes.

1. **Basic Statistics**

Se aporta información básica de la secuencia, como el nombre del archivo y la versión actual del programa. En esta página se aporta información del número total de secuencias y aquellas etiquetadas como que poseen baja calidad. También se indica la longitud de la secuencia y el porcentaje de G/C.

Filename EGAR00001545906_17296_R1.fastq

File type Conventional base calls

Encoding Sanger / Illumina 1.9

Total Sequences 2773641

Sequences flagged as poor quality 0

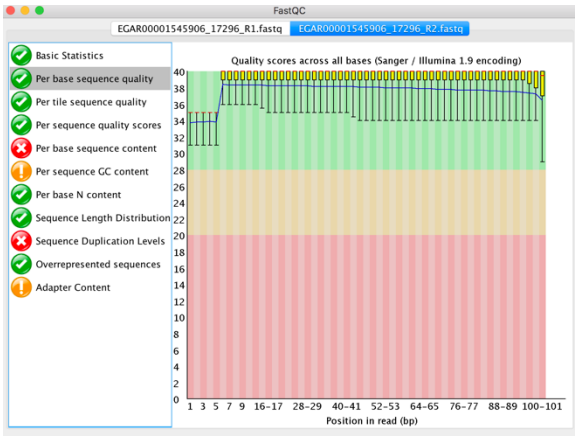
Sequence length 101

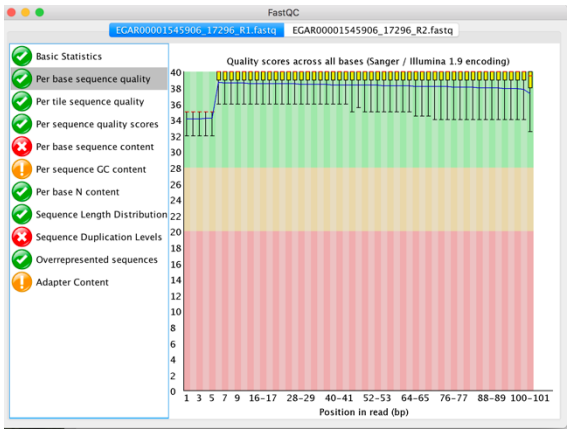
%GC 4

2. **Per base sequence quality**

En este apartado se aporta información de la calidad de cada uno de los nucleótidos que forman la secuencia, es decir, la calidad con la que se puede precisar que cada base esté en cada una de esas posiciones. En la mayoría de los casos, conforme se va alejando la secuencia se va perdiendo calidad debido a la propia naturaleza de los cebadores, los cuales se van despegando y la secuenciación va perdiendo calidad. Por lo tanto, se busca que, en las bases de posiciones superiores, la calidad Q del sistema sea mejor. Lo que se hace en la mayoría de los casos es descartar las bases con una calidad de Q que se desee.

En cada posición, se dibuja un diagrama de caja, en el que la línea central roja indica el valor de la mediana, la caja amarilla representa el rango de inter-cuartiles (los valores comprendidos entre el 25 y el 75%), las líneas inferiores y superiores se corresponden con los puntos del 10 y 90% y la línea azul representa la calidad de la media. En el eje Y se representa la calidad de la secuencia, y cuanto mayor sea, mejor calidad de la base. Este eje se divide en varios colores, en función de la calidad, representando el verde la mejor y el rojo la peor. En el eje X abscisas, se representan las bases leídas. Como se puede ver figuras 2 y 3, nuestros datos tienen una muy buena calidad pues están todas las reads en la zona verde.





Figuras 2 y 3. FastQC. gráfica de calidad de secuencia por base. Secuencias emparejadas, read 1 y read2

3.Per base sequence content

Indica la proporción de la posición de cada base para cada una de las cuatro bases que forman el ADN. En una librería al azar, se espera que no haya una gran diferencia entre las diferentes bases de la secuencia, por lo que las líneas deberían correr paralelas las unas a las otras, ya que la cantidad relativa de cada base refleja la cantidad total de estas bases en el genoma de estudio, pero en ninguna circunstancia, deben estar desproporcionadas las unas con las otras. En muchos casos, las desviaciones se producen al principio de la secuencia debido al uso de primers, pero en ningún caso la variación debe de ser del 20%, ya que si no se considera que la secuenciación ha sido un fallo.

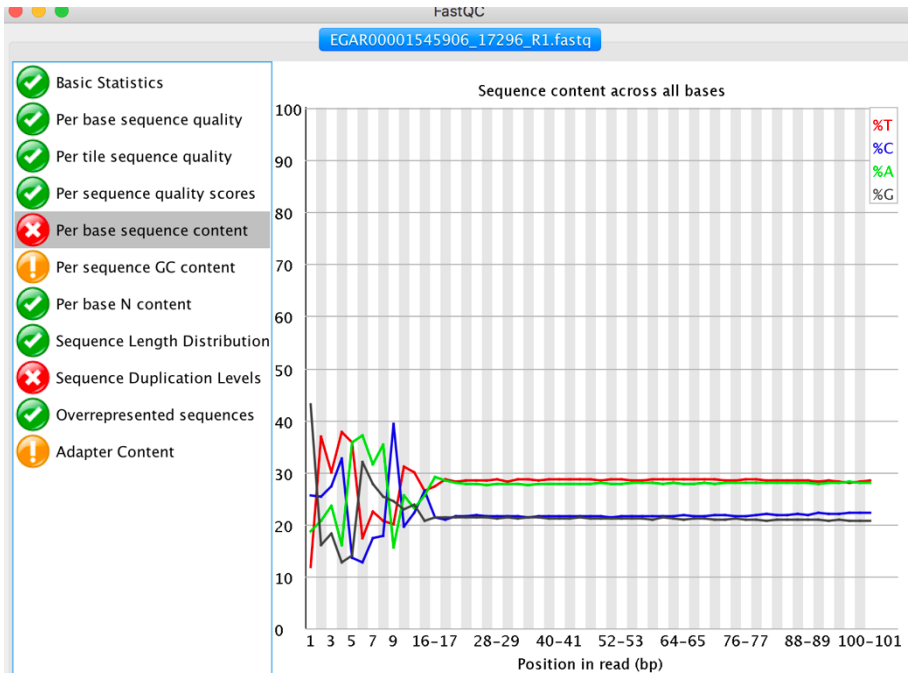


Figura 4.FastQC. Gráfica del contenido de bases desapareadas por secuencia

Según se ver en la figura 4, nuestros datos tienen una dispersión en las primeras pares de bases, con lo que se recomienda eliminarlas.

3. **Sequence Duplication Levels**

La parte de secuencias duplicadas en FastQC es sin duda una de las más difíciles de interpretar. Permite detectar la presencia de secuencias que se encuentran presentes varias veces en todo el análisis del programa.

Según se puede observar en la figura 5, nuestros datos tienen unas cuantas secuencias duplicadas, debido fundamental al método de amplificación por PCR en el proceso de creación de la librería. Estas secuencias serán eliminadas posteriormente con los programas adecuados, SAMtools o Picard

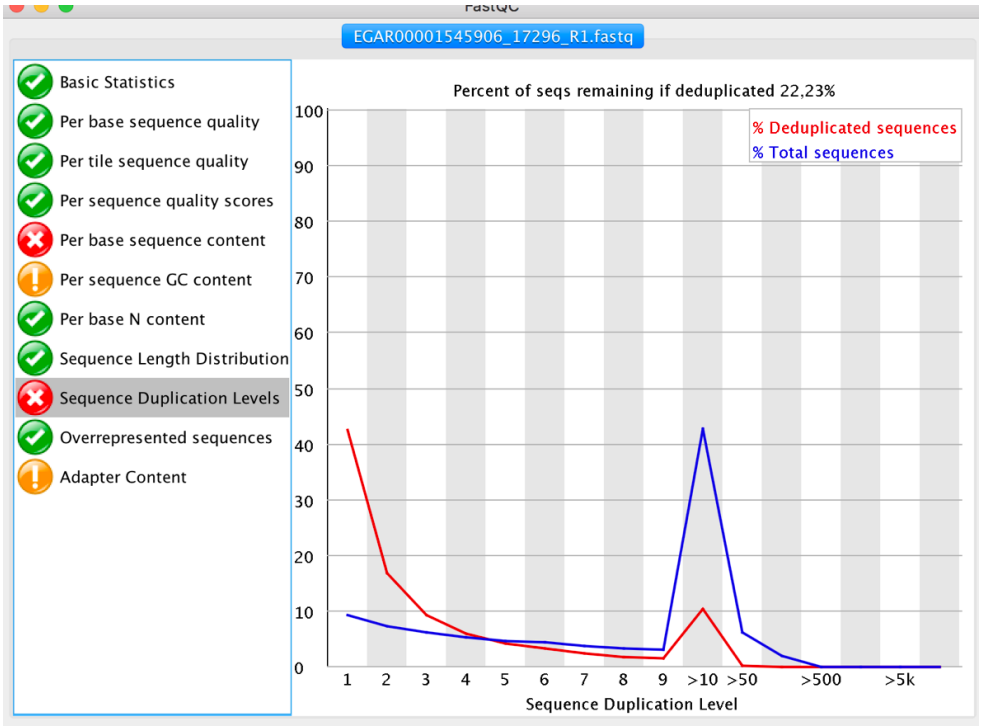


Figura 5. FastQC. Gráfica muestra los niveles de duplicados por secuencia

2.3.2. Pre-procesado

El pre-procesado consiste en obtener los archivos fastq con la mejor calidad posible para utilizarlo posteriormente en los diferentes pasos.

Uno de los grandes problemas que tiene la secuenciación masiva es la tasa de error que comete, por ello debemos realizar un pre-procesamiento de los datos (3,4).

Existen varias herramientas bioinformáticas para realizar ese pre-procesamiento:

- Trimmomatic recorte adaptador y filtrado (6,7).
- Cutadapt adaptador de recorte y filtrado con una amplia variedad de opciones de adaptador adecuado (8).
- SAMTools. Una suite para la manipulación de archivos SAM (9,10).

Nosotros vamos a trabajar con el programa Cutadapt.

2.3.2.1 Trimmed o eliminación de fragmentos de los datos en crudo.

Como hemos visto en el apartado anterior, tenemos un error de calidad en el apartado de "Per base sequence content", podemos observar que la proporción de nucleótidos al principio del read no es correcta, concretamente las 15 primeras bases.

Durante la fabricación de las librerías para la secuenciación, y concretamente para las librerías de Illumina, se utilizan primers para el proceso de amplificación de las secuencias de interés. Durante este proceso se pueden producir errores de amplificación que pueden intereferir en el proceso de análisis.

Para ello vamos a utilizar la herramienta bioinformática cutadaptad.

El código utilizado es el siguiente;

cutadapt -u 15 -U 15 -q 20 -m -50 -o ~/output2/17297_R1trim_cut.fastq -p ~/output2/17297_R2trim_cut.fastq ~/DATOS/Archivos_fastq.cip/EGAR00001545907_17297_R1.fastq ~/DATOS/Archivos_fastq.cip/EGAR00001545907_17297_R2.fastq --too-short-output ~/DATOS/output2/17297_shortR1_cuti.fastq

--too-short-paired-output ~/DATOS/output2/17297_shortR2_cuti.fastq

- -u 15 : la abreviatura -u , es la orden -cut, que elimina un determinado número de pares de bases. Si la letra "u" esta en minúscula solo elimina las 15 pbs de la lectura R1. Como he puesto las dos secuencias juntas en el mismo script, por ser secuencias pareadas, tengo que añadir -U para que también elimine la lectura R2.
- -q 20: la abreviatura -q, es la orden --quality-cutoff, Este parámetro se usa para eliminar la secuenciadas con una calidad inferior a la que marcamos en la orden. Con la orden
- -q 20 se realiza en las dos lecturas pareadas.

- -m -50: con esta orden se elimina todas aquellas lecturas que tengan una longitud inferior al número que hemos marcado. Longitud minina de 50 pbs.
- -o : con esta orden , --output introducimos la ruta y el nombre de los archivos que se obtienen.
- -p : esta orden indica que las dos lecturas son pareadas
- -Input : Se escribe la ruta y los archivos que se van a procesar *"EGAR00001545907_17297_R1.fastq" y "EGAR00001545907_17297_R2.fastq" por ser archivos pareados.*

Por último, se generan dos archivos, uno que almacena las lecturas cortas no apareadas y otro las lecturas cortas apareadas. Esto nos servirá de control del proceso.

--too-short-output. Se escribe la ruta de los archivos de salida sin aparear.

--too-short-paired-output. Se escribe la ruta de los archivos de salida apareados .

A continuación, volvemos a revisar la calidad de las lecturas después de haber realizado la orden del cutadapt., Como se ve en la figura 6, la calidad de las lecturas esta toda en la región verde con una Q40 y las secuencia " Per base Sequence content" se han corregido y ya no tenemos dispersión.

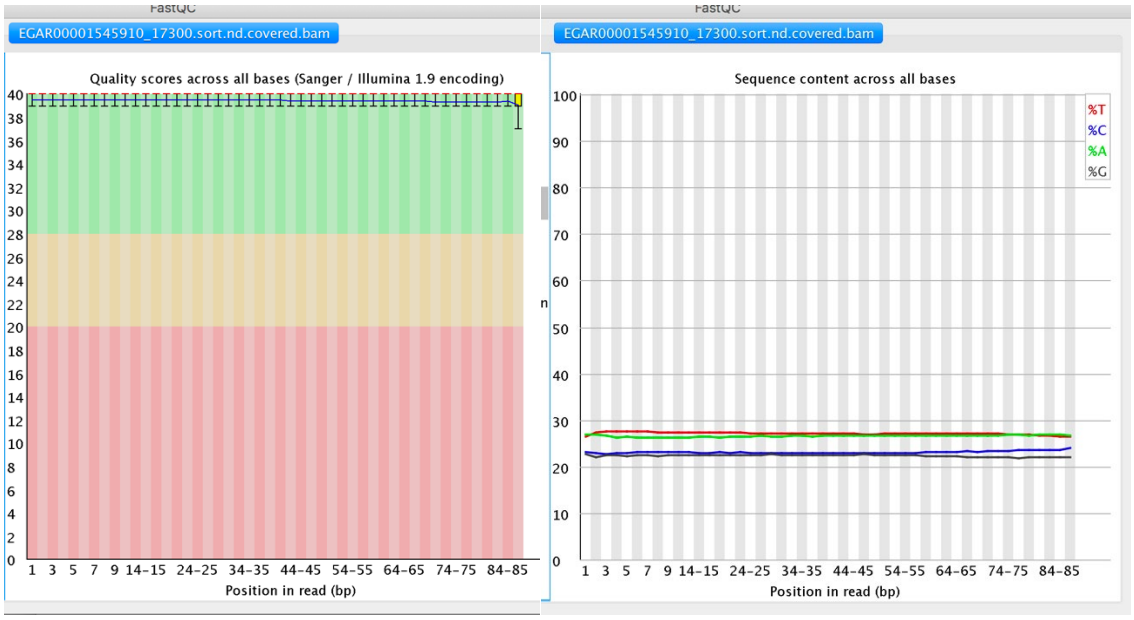


Figura 6. FastQC. Gráfica por base sequence quality y por base sequence content.

2.3.4 Alineamiento.

Una vez obtenidos los archivos de las muestras en formato fastq con unas calidades de lecturas adecuadas, se lleva a cabo el proceso de alineamiento con estos nuevos archivos.

Este proceso consiste en mapear las lecturas obtenidas frente al genoma de referencia, descargado desde pagina web del Esembl.se utilizara la versión Homo_sapiens, GRCh37.dna_sm.primary_assembly.fa.gz (9,10).

Existen varias herramientas bioinformáticas para llevar a cabo el alineamiento de las lecturas. Pero la que hemos elegido ha sido BWA que utiliza el algoritmo de Burrows - Wheeler Trasformada, BWT. Esta herramienta bioinformática puede realizar el alineamiento con: BWA-backtrack, BWA-SW y BWA-MEM.

Se ha elegido el BWA-MEM, ya que su algoritmo maximal exact matches (MEM), es el recomendado por su alta calidad y rapidez en el alineamiento secuencias muy grandes, como el genoma humano y es el que mejor mapea con lecturas de secuencia cortas. Funciona con datos emparejados. Por lo que es el algoritmo que mejor puede funcionar para nuestras muestras. Mientras que BWA-SW es muy similar a BWA-MEN pero no funciona tan bien con secuencias de alta calidad y tan grandes. Por último, BWA-Backtrach solo funciona correctamente para lecturas de secuencias de Illumina hasta 100pb.

Pero antes de la indexación de las muestras contra el genoma de referencia, tenemos que preparara el genoma de referencia contra el que se va a indexar, para que se realice este proceso satisfactoriamente.

2.3.5 Indexado genoma de referencia.

Se lanza el script desde terminal es desde SAMtools:

$ bwa-0.7.17/bwa index -a bwtsw Homo_sapiens.GRCh37.dna.primary_assembly.fa >

Una vez indexado el genoma de referencia vamos a alinear nuestras muestras contra este genoma de referencia, añadiendo una cabecera al archivo indicador de grupo.

El script que lanzamos desde terminal es:

$ ~bwa-17 enriquesevillaromero$ bwa mem -t 3 -R '@RG\tID:'"1545907_17297"'\tLB:'"1545907_17297"'\tSM:'"1545907_17297"'\tPL:ILLUMINA' ~/genoma_indexado/Homo_sapiens.GRCh37.dna.primary_assembly.fa ~17297_R1trim_cut.fastq ~/17297_R2trim_cut.fastq > ~/bwa_all/bwa1545907_17297.sam

- Bwa mem: Esta es la orden para alinear el genoma de referencia ya indexado contra nuestras muestras, hemos utilizado el alineamiento de maximal exact matches (MEM) ya que este es el que mejor funciona para alineamientos entre 70pbs y 1Mbp.
- *-t 3 :* es el numero de hilos o hebras que utiliza el sistema para realizar esta función, cuanto más alto más consumo de recursos.

- *-R* : Es el parámetro que ponemos para que el software reconozca que hay un identificador de grupo y que tiene que incluirlo en el archivo de salida.

  Este parámetro lo necesitaremos para posteriores análisis de las muestras.

  '@RG\tID:'"$i"'\tLB:'"$i"'\tSM:'"$i"'\tPL:ILLUMINA':

Identificador de grupo. Hay que cambiar el símbolo $i por el nombre de cada una de las muestras.

- Genoma de Referencia *Homo_sapiens.GRCh37.dna.primary_assembly.fa*.
- Datos trimeados para las lecturas pareadas, R1 y R2. ~/17297_R1trim_cut.fastq y ~/17297_R2trim_cut.fastq
- Output de salida en formato .sam

  bwa1545907_17297.sam

2.3.6. Sort e indexado.

Los archivos sam que se ha obtenido en el apartado anterior, son archivos que pesan muchos megas de información, por lo que se transforman/comprimen a un archivo más manejable, bam. Este archivo es idéntico al anterior, pero en formato binario.

Este tratamiento bioinformático lo realizaremos con la herramienta bioinformática SAMtools (10).

SAMtools es un conjunto de utilidades para manipular alineamientos. SAMtools importa y exporta en formato SAM, ordena, une e indexa, y permite recuperar las lecturas de cualquier región con rapidez. SAMtools está diseñado para trabajar dentro de un flujo de trabajo.

Cabe destacar los siguientes comandos:

- View: extrae todos o una porción de los alineamientos en formato sam o bam según las especificaciones indicadas.
- Sort: ordena alineamiento desde el extremo izquierdo, creando un fichero nuevo en formato BAM con el contenido ordenado.
- Index: indexa los alineamientos ordenados para un acceso más rápido, creando un fichero de índice con extensión bai.

Transformación de sam a bam

El script que lanzamos en terminal;

$~samtools view -S ~/bwa1545906_17296.sam -b -o ~/bwa1545906_17296.bam

A continuación realizamos una ordenación de las lecturas (sort). Esto es un proceso que permite una agilización en el tratamiento de los datos por las herramientas bioinformáticas.

El script que lanzamos en terminal;

```
$~samtools sort ~/bwa_all/bwa1545906_17296.bam >
~Datos_alineados_sorter.bam/bwa1545906_17296_sort.bam
```

Una vez obtenidos los datos ordenados, vamos a eliminar los duplicados de PCR que habíamos observado en el programa FASTQC. Estos duplicados de PCR son muy comunes cuando la técnica que se ha utilizado para la secuenciación de la muestra ha sido mediante amplificación de clones por PCR, como ocurre en el caso de los secuenciadores de Illumina,

La eliminación de duplicados de PCR se eliminan mediante la herramienta bioinformática SAMtools, descrita anteriormente. Para ello utilizamos el comando rmdup.

El script que lanzamos en terminal;

```
$~samtools rmdup -S  ~/Datos_alineados_sorter.bam/bwa1545906_17296_sort.bam
~/Datos_a_s_ND/bwa1545906_17296_sort_ND.bam
```

A continuación, indexamos el archivo bam para obtener también el archivo .bai que necesitaremos para la visualización en el IGV

```
$~Samtools index ~ /Datos_alineados_sorter.bam/bwa1545906_17296_sort.bam
```

2.3.7.Bedfile

Además de los ficheros .bam, las herramientas bioinformáticas que vamos a evaluar requieren un fichero denominado Bedfile. En este archivo en formato .txt o .bed describe las regiones del genoma que se han secuenciado. En concreto, describe todos los exones secuenciados del panel Trusight cáncer Panel v2 (100 genes)(11) El archivo BED se ha descargado siguiendo las indicaciones del trabajo publicado(11). Una vez descargado el archivo tiene que ser tratado para poder ser utilizado por las herramientas bioinformáticas.

2.3.8 Bedtools

Por último, vamos a realizar una agrupación de las regiones secuenciadas para cada paciente con las regiones descritas en nuestro archivo Bedfile descargado del trabajo.

Esto nos va a generar un archivo donde solo tendremos las regiones de interés que hemos secuenciado. Ver figura 8.

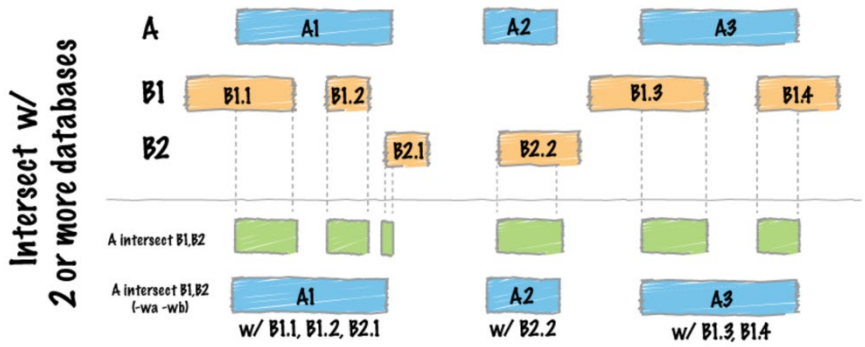


Figura 8. Bedtools. Esquema de como seleccionamos las regiones de interés(12)

Este proceso lo vamos a realizar mediante la herramienta bioinformática Bedtools.(12)

El script que lanzamos en terminal;

~ $ bedtools intersect -abam ~/Muestra_17296.sort.nd.bam -b ~/Bedfile_n_chr.bed > ~ /Muestra_17296_covered.bam

El archivo que obtenemos es un archivo .bam con solo las regiones definidas por el Bedfile.

Se ha cargado todo el código del pipeline in-house para la obtención de datos alineados optimizados en el repositorio Github;

https://github.com/kike1790/Procesamimiento-Raw-Data-TFM-E.Sevilla/blob/master/Procesamiento%20Raw%20data

2.3.9 IGV

El Integrative Genomics Viewer (IGV) (13) es una herramienta de visualización de alto rendimiento para la exploración interactiva de grandes conjuntos de datos genómicos integrados.

Vamos a comprobar que los archivos generados con el programa bioinformático Bedtools han funcionado correctamente. Para ello vamos a confrontamos el archivo bam obtenido por la función Bedtools intersect vs el archivo de Referencia CGRh37.

Figura 9. IGV. Imagen donde muestras las regiones de interés con sus coberturas y relacionado con el genoma de referencia

Como podemos ver en la figura 9, se corresponde perfectamente las read y regiones secuenciadas de nuestro archivo bam con las mismas regiones del genoma de referencia

3. Análisis de las herramientas Bioinformáticas y resultados

3.1 Análisis de las herramientas bioinformáticas.

3.1.1.ExomeDepth

La herramienta bioinformática ExomeDepth, es un algoritmo desarrollado en R (14).

ExomeDepth es una herramienta bioinformática que trabaja con profundidades de lecturas para hacer las llamadas de para cada exón de muestras secuenciadas mediante NGS. La idea principal es que a partir de un conjunto de muestras control se crea una matriz de referencia (archivo normalizado de todas las lecturas para las muestras normales), donde se genera una optimación de las profundidades de lectura para cada exón. Esta optimación o normalización de las profundidades de lectura nos servirá para comparar cada muestra tumor frente a este archivo normalizado de lecturas.

La profundidad de lectura depende de muchos factores además del número de copias; el contenido de GC, la eficacia de la captura, problemas en la secuenciación, el tipo de alineación, etc. Para minimizar estos errores, la profundidad de lectura se debe considerar no en términos absolutos sino en relación con una muestra de referencia (promedio de muchas muestras control con exomas sin CNVs). Casi todas las herramientas bioinformáticas utilizan una referencia promedio, donde la profundidad de lectura relativa se distribuirse normalmente entorno a una profundidad de lectura esperada. La herramienta bioinformática ExomeDepth defiende un modelo más complejo basándose en un modelo estadístico beta binomial para calcular las profundidades de lecturas relativas con las que comparamos nuestras muestras(15).

La distribución binomial es una distribución de probabilidad discreta que cuenta el número de éxitos en una secuencia de n ensayos independientes entre sí, con una probabilidad fija p de ocurrencia del éxito entre los ensayos. Se caracteriza por ser dicotómico, esto es, solo dos resultados son posibles. A uno de estos se denomina «éxito» y tiene una probabilidad de ocurrencia p y al otro, «fracaso», con una probabilidad2 q = 1 - p. La distribución beta-binomial es la distribución binomial en la que la probabilidad de éxito en cada ensayo no es fija sino aleatoria y sigue la distribución beta(16).

ExomeDepth utiliza el modelo probabilístico beta binomial como si el recuento de lecturas exónicas de las muestras fuera X para la muestra de prueba e Y para la referencia agregada(15). El modelo asigna una probabilidad a la lectura aleatoria de la muestra de prueba: dónde la probabilidad de lectura aleatoria pertenezca a la muestra de prueba (en lugar de la referencia). El índice *i* denota el exón y la covariable

se relaciona con el estado del número de copias para el exón *i*. La proporción de lecturas se correlacionan con la muestra prueba para las eliminaciones / duplicaciones y se calcula en función de la proporción esperada para el número de copias normal y suponiendo una relación de lectura de 0.5 (para una eliminación) o 1.5 (para una duplicación). Pero este modelo binomial no captura completamente los sesgos específicos de la muestra. Por lo que se propuso un cambio a un modelo beta-binomial (17) donde el parámetro de dispersión excesiva se estima numéricamente a partir de los datos de recuento de lecturas y su varianza se convierte en φ , agregando a la varianza binomial un término de dispersión adicional. Con este cambio conseguimos un ajuste mejor a la hora de llamar a las CNVs

Otro factor interesante en ExomeDepth es como se comparan las muestras de referencia contra el resto de las muestras a analizar. El sesgo que se produce, debido a la eficiencia en la captura, secuenciación o en cuanto al contenido GC, varia de unas librerías a otras. Para disminuir este sesgo, ExomeDepth compara la muestra de análisis contra la muestra de referencia, pero selecciona progresivamente aquellas que tienen una mayor correlación (en la profundidad en todos los exones).

La compensación óptima para compensar estos errores se alcanza cuando su usa un promedio de 10 muestras más o menos correlacionadas como referencia(8).

Por ello nosotros realizamos dos pruebas, una para todas aquellas muestras control y tumorales que se han secuenciado en el pool1, y las que se han secuenciado en el pool2. Además, utilizamos más de 10 muestras control para generar la muestra referencia.

ExomeDepth genera una muestra de referencia normalizada con todas las muestras control(15). Por lo tanto, los individuos a estudiar/tumorales deben ser excluidos de la referencia agregada. También significa que ExomeDepth puede fallar CNV comunes, si la llamada también está presente en la referencia agregada. ExomeDepth es realmente adecuado para detectar llamadas CNV raras (por lo general para análisis de trastornos mendelianos raros).

3.1.1.2 Flujo de trabajo para ExomeDepth.

Lo primero que hacemos es generar un archivo con todas las lecturas de conteos, para cada muestra y para cada región de interés, exones.

Para ello, preparamos los archivos de entrada, bam de todos los controles.

Introducimos nuestro archivo Bedfile que nos indicara las regiones de análisis/estudio. Y por último el genoma de referencia en formato fasta.

Vamos a generar dos matrices de conteos, una para generar la matriz de referencia solo con las muestras control. Y otra con las muestras tumorales a analizar.

```
### Cargamos las librerías
library(ExomeDepth)
###Cargamos el Bedfile y el gemona de referencia.
Bedfile <- file.path("~/"Bedfile_n_rs.txt")
targets <-read.table(Bedfile, header = T)
b37 <- file.path ”~/genoma_referencia/Homo_sapiens.GRCh37.dna.primary_assembly.fa")
########################################
####Cargamos las muestras control y generamos la matriz de contajes
###################################
my.bam_control <- list.files("~/Muestras_Control/pool1", pattern = '*.bam$')
bamFile_control <- file.path("~/Muestras_Control /pool1",my.bam_control)
```

Generamos la matriz de contajes para las muestras control, generando nuestra Control de referencia.

```
data_Control <- getBamCounts(bed.frame = targets,
                             bam.files = bamFile_control,
                             include.chr = FALSE,
                                          referenceFasta = b37)
Control_ExomeCount.dafr <- as(data_Control[, colnames(data_Control)], 'data.frame')###
Control_ExomeCount.dafr$chromosome <- gsub(as.character(Control_ExomeCount.dafr$space), pattern
= 'chr',replacement = '')
print(head(Control_ExomeCount.dafr))
Control_ExomeCount.mat <- as.matrix(Control_ExomeCount.dafr[,
grep(names(Control_ExomeCount.dafr), pattern = '.*bam')])
Control_nsamples <-ncol(Control_ExomeCount.mat)
```

Hacemos lo mismo para las muestras a analizar, muestras tumor.

```
########################################
```

```
#### Cargamos las muestras tumor y generamos la matriz de contajes

###################################

my.bam.tumor <- list.files("~/Muestras_Tumor/pool1",  pattern = '*.bam$')

bamFile.tumor <- file.path("~/Tumor/pool1",my.bam.tumor)


data_Tumor <- getBamCounts(bed.frame = targets,

                    bam.files = bamFile.tumor,

                    include.chr = FALSE,

                   referenceFasta = b37)
```

Generamos la matriz de conteos para las muestras tumorales

```
Tumor_ExomeCount.dafr <- as(data_Tumor[, colnames(data_Tumor)], 'data.frame')

Tumor_ExomeCount.dafr$chromosome <- gsub(as.character(Tumor_ExomeCount.dafr$space), pattern =
'chr',replacement = '')


Tumor_ExomeCount.mat <- as.matrix(Tumor_ExomeCount.dafr[, grep(names(Tumor_ExomeCount.dafr),
pattern = '.*bam')])

Tumor_nsamples <-ncol(Tumor_ExomeCount.mat)

Tumor_samples <- colnames(Tumor_ExomeCount.dafr)
```

Preparamos la muestra de referencia, es decir seleccionamos todas las muestras control

Y preparamos las muestras tumorales para que sean analizadas de una en una contra la muestra de referencia. Esta comparación se realizará comparando sus correlaciones.

Hacemos una llamada de las CNVs y generamos un archivo de salida para poder ver y analizar las CNVs por muestra y por exón.

```
my.ref.samples <- c ("Control_1.bam"," Control_2.bam"," Control_3.bam"," Control_4.bam","
Control_5.bam"," Control_6.bam"," Control_7.bam"," Control_8.bam"," Control_9.bam","
Control_10.bam"," Control_11.bam")  ## selección de los controles

my.ref.sample.set <- as.matrix(Control_ExomeCount.dafr[, my.ref.samples])
```

```
###Generamos el bucle para procesar todas las muestras tumorales
for (i in 1:Tumor_nsamples) {
 samplename <- Tumor_samples[i + 6]
my.choice_1t_Ac <- select.reference.set
 (test.counts = Tumor_ExomeCount.mat[,i],
 reference.counts = my.ref.sample.set,
 bin.length = (Control_ExomeCount.dafr$end -    Control_ExomeCount.dafr$start)/1000,
 n.bins.reduced = 10000)
###Realizamos la comparación de matriz de contaje control vs matriz de contaje tumoral#####
 my.matrix <- as.matrix(   Control_ExomeCount.dafr[,my.choice_1t_Ac$reference.choice])
 my.reference.selected <- apply(X = my.matrix,MAR = 1, FUN = sum)
 all.exons_1t_ac <- new('ExomeDepth',
 test = Tumor_ExomeCount.mat[,i],
 reference = my.reference.selected,
 formula = 'cbind(test, reference) ~ 1')
```

```
##Llamanos a las CNVs
show(all.exons_1t_ac)
all_exons_CNVs <- CallCNVs(x = all.exons_1t_ac,
 transition.probability = 10^-3,
 chromosome = Control_ExomeCount.dafr$space,
 start = Control_ExomeCount.dafr$start,
 end = Control_ExomeCount.dafr$end,
 name = Control_ExomeCount.dafr$names)
head(all_exons_CNVs@CNV.calls)
######Imprimimos en un archivo de salida las CNVs detectadas######
if(nrow(all.exons_CNVs@CNV.calls)>0){
      output.file <- paste0(Tumor_nsamples ,".cnv.txt")
      write.table(file = file.path("~/DATOS/ExomeDepth/All",output.file),x = cbind(Tumor_nsamples
,all.exons_CNVs@CNV.calls),row.names = FALSE,quote=F,sep="\t")
}
}
```

Se ha cargado todo el código en el repositorio Github; https://github.com/kike1790/ExomeDepth-TFM-CNVs/blob/master/Exomedepth%20E.Sevilla

Generamos unos archivos de salida donde nos indica que muestra es la que tiene CNVs, ver tabla 4. Además, cada archivo en formato .txt podemos ver que secuencia.

i	start.p	end.p	type	nexons	start	end	chromosome	id	BF	reads.expected	reads.observed	reads.ratio
6	238	238	deletion	1	10183532	10183872	3	chr3:10183532-10183872	4.03	207	134	0,647
6	298	298	deletion	1	128205646	128205875	3	chr3:128205646-128205875	12.1	87	20	0,230
6	537	540	deletion	4	145742434	145743169	8	chr8:145742434-145743169	19.4	242	108	0,446
6	601	601	deletion	1	100459403	100459575	9	chr9:100459403-100459575	7.03	109	52	0,477
6	625	625	deletion	1	43572707	43572780	10	chr10:43572707-43572780	4.23	75	39	0,520
6	647	647	deletion	1	88516385	88516688	10	chr10:88516385-88516688	7.61	100	44	0,440
6	685	686	deletion	2	2905234	2906720	11	chr11:2905234-2906720	5.17	323	217	0,672
6	870	870	deletion	1	48878049	48878186	13	chr13:48878049-48878186	5.08	63	28	0,444
6	1202	1202	deletion	1	89882945	89883024	16	chr16:89882945-89883024	4.95	64	29	0,453
6	1393	1393	deletion	1	1220580	1220717	19	chr19:1220580-1220717	3.85	130	80	0,615
6	1413	1413	deletion	1	45867244	45867378	19	chr19:45867244-45867378	8.65	112	48	0,429
6	1425	1425	deletion	1	36164432	36164908	21	chr21:36164432-36164908	4.43	90	49	0,544

Tabla 4. Archivo de salida de ExomaDepth .

El archivo de salida nos nuestra los siguientes campos;

- I, la muestra analizada.
- Inicio de la secuencia
- El final de la secuencia
- El tipo de mutación

- En que exón se encuentra la CNV
- El inicio de la secuencia
- El final de la secuencia
- El cromosoma
- Id del cromosoma y el inicio y final
- Factor de Bayes, BF
- Número de lecturas esperadas
- Número de lecturas observadas
- Ratio de las lecturas.

La ratio de lecturas va a mostrar unos valores entre 0 y 2. Este valor nos indica si la delección está en homozigosis, valores de 0 o en heterozigosis a partir de 0,5.

Una información importante es la columna BF, que representa el factor de Bayes. Cuantifica el valor estadístico para cada CNV. De hecho, es el log10 de la relación de probabilidad.

Cuanto más alto sea ese número, más confianza tendrá una vez sobre la presencia de una CNV. Si bien es difícil dar un umbral ideal, y para los exones cortos el Factor de Bayes no es convincente, las llamadas grandes más obvias deben marcarse fácilmente clasificándolas de acuerdo con esta cantidad.

Por último, hemos añadido el nombre del gen, tal y como se puede ver en la tabla 5, para poder comparar las herramientas bioinformáticas.

Muestra	i	start.p	end.p	type	nexons	Gen	start	end	chromosome	id	BF	reads.expected	reads.observed	reads.ratio
17364	6	238	238	deletion	1	RET	10183532	10183872	3	chr3:10183532-10183872	4.03	207	134	0,647
17364	6	298	298	deletion	1	BMPR1A	128205646	128205875	3	chr3:128205646-128205875	12.1	87	20	0,230
17364	6	537	540	deletion	4	CDKN1C	145742434	145743169	8	chr8:145742434-145743169	19.4	242	108	0,446
17364	6	601	601	deletion	1	RB1	100459403	100459575	9	chr9:100459403-100459575	7.03	109	52	0,477
17364	6	625	625	deletion	1	FANCA	43572707	43572780	10	chr10:43572707-43572780	4.23	75	39	0,520
17364	6	647	647	deletion	1	STK11	88516385	88516688	10	chr10:88516385-88516688	7.61	100	44	0,440
17364	6	685	686	deletion	2	ERCC2	2905234	2906720	11	chr11:2905234-2906720	5.17	323	217	0,672
17364	6	870	870	deletion	1	RUNX1	48878049	48878186	13	chr13:48878049-48878186	5.08	63	28	0,444
17364	6	1202	1202	deletion	1	VHL	89882945	89883024	16	chr16:89882945-89883024	4.95	64	29	0,453
17364	6	1393	1393	deletion	1	GATA2	1220580	1220717	19	chr19:1220580-1220717	3.85	130	80	0,615
17364	6	1413	1413	deletion	1	RECQL4	45867244	45867378	19	chr19:45867244-45867378	8.65	112	48	0,429

17364	6	1425	1425	deletion	1	XPA	36164432	36164908	21	chr21:36164432-36164908	4.43	90	49	0,544

Tabla 5. Archivo de salida de ExomeDepth con el nombre del gen

3.1.1.3 Comprobación con IGV

Cargamos el IGV, y buscamos en el cromosoma 3 en la región 128205646-128205875, GATA2.chr3.128199861.128200162 para comprobar que hay una CNV.

Como podemos ver en la figura 10, existen una bajada importante en la cobertura de la región de análisis frente a la cobertura de la muestra control. La imagen superior corresponde a la muestra tumor y la inferior a un control.

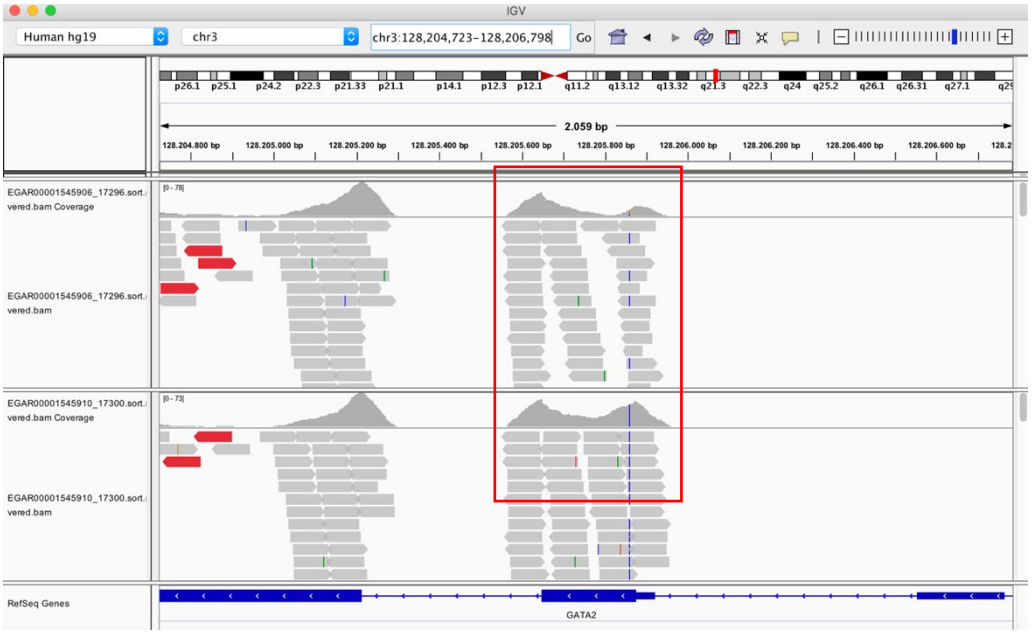


Figura 10. I Representación en GV de la delección en el cromosoma 3 en la región 128205646-128205875, GATA2.chr3.128199861.128200162 vs a muestra control

3.2.1 CNVkit

CNVkit es una Herramienta bioinformática que utiliza un método estadístico robusto para extraer las CNVs.(8) Los métodos clásicos se basan en el Teorema de Límite Central para producir estimaciones normalmente distribuidas.

Desafortunadamente, cuando hay valores atípicos en los datos, los resultados producidos por los métodos clásicos son a menudo de baja calidad. La estadística robusta es una aproximación alternativa a los métodos estadísticos clásicos pero el objetivo es producir estimadores que no se vean afectados por variaciones pequeñas en los valores atípicos o outlander. Como podemos observar en la figura 11

Las muestras para analizar se fragmentan en pequeños contenedores/bins para poder ser analizados.

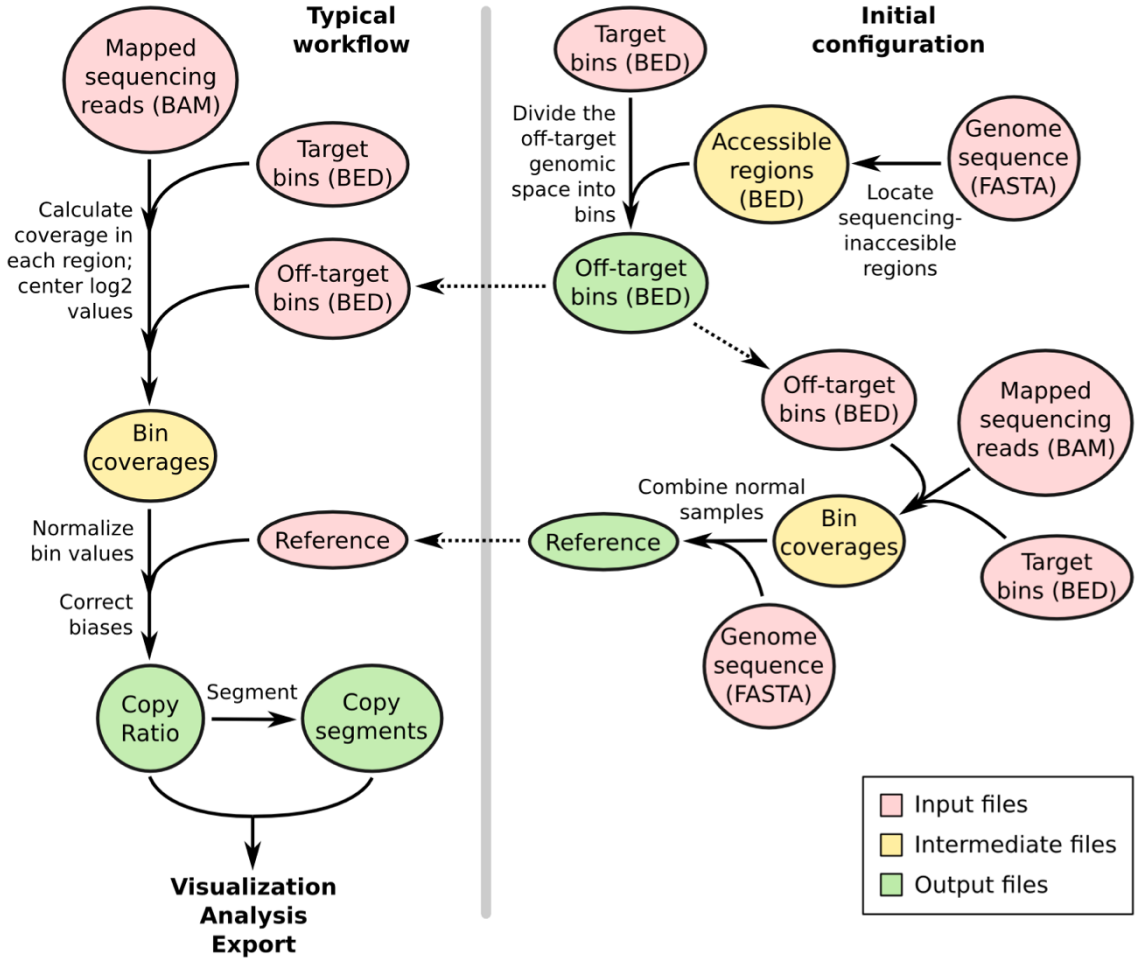


Figura 11 CNVkit. Flujo de trabajo de la herramienta bioinformática.(18)
https://doi.org/10.1371/journal.pcbi.1004873.g001

El cálculo de las profundidades de lectura se realiza tanto para las lecturas dentro de la región de análisis, las regiones definidas por nuestro Bedfile, como las que están fuera de estas regiones. Tal y como podemos ver en la figura 11. De este modo calculamos las profundidades de lectura de las regiones de captura. Posteriormente se calcula la media centrada y a continuación el sesgo se normaliza mediante el log2 para calcular el promedio de las profundidades de lectura ponderada. Estos métodos se basan en una estadística robusta(19) (20,21).

Estos valores se guardan en un archivo de salida donde tenemos los valores log2 y los valores “spread” que indica la profundidad de lectura esperada y la fiabilidad de la estimación.

CNVkit filtra los valores obtenidos para cada bin/ventana y solo deja pasar aquellos que cumplen los criterios predefinidos:

1.Cuando los datos de la muestra en la profundidad de lectura en Log2 del valor de referencia es inferior a un umbral (-5) .

2. Cuando las profundidades de lectura entre las muestras normales en la referencia son superiores a un umbral por defecto 1.0.

Métodos de segmentación.

La herramienta bioinformática CNVkit utiliza el método de segmentación binaria circular (CBS), este método es el que mejor se adapta a panales de exones medianos, como el muestro.

3.2.1.1 Flujo de trabajo para CNVKit.

Los archivos necesarios son;

- El genoma de referencia en formato fasta.
- El archivo Bedfile de las regiones de interés en formato .bed
- Los archivos control y tumoral en formato .bam y alienados con el algoritmo BWA

Lo primero que vamos a hacer es generar un archivo de referencia con todas las muestras control, para el pool 1, se llamara "my_reference.cnn". De este modo obtengo una línea basal de profundidad de lecturas por exón.

Con este archivo de referencia será contra el que procesemos todas las muestras tumorales.

cnvkit.py batch

Seleccionamos los diferentes archivos control

~/Normal_1.bam **--normal** ~/ Normal_2.bam ~/ Normal_3.bam ~/ Normal_4.bam / Normal_5.bam

##Seleccionamos el Bedfile

--targets ~ /Bedfile_nrsncbxy.bed

##Seleccionamos el genoma de referencia

--fasta ~ genoma_indexado/Homo_sapiens.GRCh37.dna.primary_assembly.fa

##Generamos los archivos de salida output con el archivo de referencia normalizado.

--output-reference ~ OUTDIR_2/my_reference.cnn **--output-dir ~**OUTDIR_2/out_17432

Obtenemos un archivo de salida con my_reference.cnn y otro archivo de salida out _17296, donde se guardan los archivos generados;

Ahora, vamos a confrontar nuestro archivo de referencia, nuestras normales contra cada muestra tumor que vamos a analiza.

Utilizamos el siguiente script.

cnvkit.py batch

##Selecionamos la muestra tumora analizar

~/Tumor/Tumor_17342.sort.nd.covered.bam

##Cargamos la muestra referencia normalizada control

-r ~/DATOS/CNVkit/OUTDIR_2/my_reference.cnn

##Generamos los archivos de salida y gráficas de deleciones y duplicaciones

-p 0 --scatter

##Archivo de salida

~/Archivo de salida_example 4

Generamos un archivo de salida 'example 4' donde se almacenará el análisis para esta muestra, 17342.'.

Mediante las opciones **--scatter,** generamos un archivo de salida donde podemos visualizar donde se ha producido la ganancia o perdida. Ver figura 13

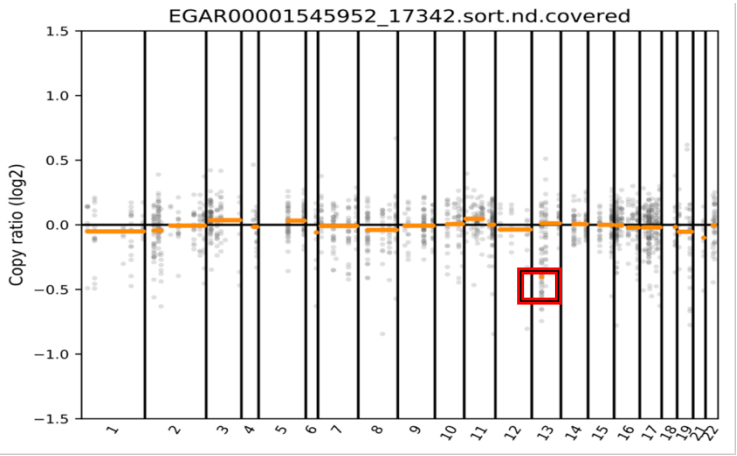


Figura 12 Representa el ratio del log2 para la muestra tumoral 17342. Como podemos ver en el cromosoma 13 hay una delección, pues su valor esta por debajo de 0,0 alrededor de -0,5 de ratio del log2

CNVkit genera varios tipos de gráficos utilizando las bibliotecas de software Biopython(22), Reportlab (http://www.reportlab.com/opensource/) y matplotlib (http://matplotlib.org):

A continuación, vamos a llamar a las CNVs y vamos a generar un archivo de salida. Los parámetros que utilizamos son la opción de genemetrics que generara un archivo con los posibles genes y posiciones cromosómicas que tenemos CNVs.

cnvkit.py **genemetrics**

##Cargamos el archivo de las muestra tumora a analizar con las coberturas

~ /Tumor_17342.sort.nd.covered.cnr

##Generamos un archivo con las coberturas donde se muestra el log2 que nos indicara ganancias o perdidas, ver tabla 6.

Found 1 gene-level gains and losses

gene	chromosome	start	end	log2	depth	weight	n_bins
BRCA2	13	32889599	32921034	-0.403	43.159	25.439	33

Tabla 6 de salida de CNVkit

La tabla de salida nos muestra;

- Nombre del gen
- Chormosoma
- Inicio de la CNV
- Final de la CNV
- *log2* : Media ponderada de las relaciones log2 de todos los bins del gen, incluidos los bins intrónicos fuera del objetivo.
- *depth* : media ponderada de profundidades de lectura no normalizadas en todos los contenedores de este gen.
- weight : Suma de los pesos de los contenedores de este gen.
- *nbins* : el número de contenedores asignados a este gen.

Los valores por defecto de la herramienta son los siguiente;

-t / --thresholdy de 0.2 , este parámetro nos informará ganancias y pérdidas de copia única en una muestra de tumor completamente pura (o CNV de línea germinal), pero sería necesario un umbral más bajo para llamar CNA somáticas si hay una contaminación significativa de células normales.

-m / --min-probes este parámetro predeterminado en 3, nos informara del vambio en el número de copias si las CNVs están en más de 2 bins

Se ha cambiado el parámetro del thresholdy a 0,05 ya que son muestras con somáticas. En cambio el min-probes se ha dejado el valor por defecto.

cnvkit.py **genemetrics**

##Cargamos los dos archivos .cnr que el peso proporcional de cada bin
~/ 17342.sort.nd.covered.cnr

##Cargamos los dos archivos .cns donde indica el número de contenedores cubiertos por el segmento.

-s ~/ 17342.sort.nd.covered.cns **-t 0.05 -m 3** >

##Generamos el archivo de salida

~/Archivo de salida_17342.txt

Muestra	gene	chro	start	end	log2	depth	weight	n_bins	segment_weight	segment_probes
17342	**BRCA2**	13	32889599	32921034	-0.4031	43.159	25.439	33	25.439	33

Tabla 7 Resultados de CNVkit para la muestra 17342.

Las columnas que tenemos son:

- Nombre del gen
- Chromosoma
- Inicio de la CNV
- Final de la CNV
- *log2* : el valor de la relación log2 del segmento que cubre el gen, es decir, la media ponderada de todos los contenedores cubiertos por el segmento completo, no solo este gen.
- *Profundidad* , *peso* , *sondas* : como arriba.
- *depth* : media ponderada de profundidades de lectura no normalizadas en todos los contenedores de este gen.
- weight : Suma de los pesos de los contenedores de este gen.
- *nbins* : el número de contenedores asignados a este gen.
- *seg_weight* : La suma de los pesos de los contenedores que soportan el segmento.
- *seg_probes* : el número de sondas que soportan el segmento.

Todo el código esta subido al repositorio de Github
https://github.com/kike1790/CNVkit-TFM-E.-Sevilla

3.2.1.3 Comprobación con IGV

Comprobamos la delección en la muestra 17342 en el cromosoma 13: BRCA2, 32889599-32921034 frente a la muestra control. Podemos ver una pérdida de cobertura, ver figura 11, relacionando esta pérdida de cobertura con la delección que encontramos en la herramienta.

Figura 13. Representación en el IGV de la delección en el cromosoma 19: BRCA2 33792243-33793321, vs a muestra control

3.3.1 VarScan

VarScan es una herramienta bioinformática independiente desarrollada en el Instituto del Genoma de la Universidad de Washington para detectar variantes en los datos de NGS. (23)

Las herramientas bioinformáticas que se utilizan para hacer llamadas de variantes emplean un marco probabilístico, como estadística bayesiana, para detectar variantes y evaluar confianza en ellos. Estos enfoques generalmente funcionan bastante bien, pero pueden ser confundidos por numerosos factores tales como la profundidad extrema de lectura, muestras combinadas, y las muestras contaminadas o impuras. En contraste, VarScan emplea un test estadístico heurístico robusto para llamar a variantes que cumplan los umbrales deseados para la profundidad de lectura, la calidad de base, frecuencia del alelo variante, y la significación estadística.

Se puede utilizar para detectar diferentes tipos de variación:

- **Variantes de la línea germinal** (SNPs and indels) en muestras individuales o pools de muestras.
- **Variantes multi-muestra** (compartido o privado) de los conjuntos de datos multi-muestra (con mpileup).
- **Las mutaciones somáticas** , eventos LOH, y variantes de la línea germinal en pares-tumorales normales.
- **Somáticas alteraciones del número de copias** (CNA) en los datos exoma-tumorales normales.

Nosotros vamos a utilizar la variante de somáticas CNA.

Este está diseñado específicamente para la secuenciación del exoma, en el que una muestra de tumor y su normal correspondiente se capturaron y se secuenciaron en condiciones idénticas. Mediante la realización de una comparación del par de profundidad de lectura entre las muestras en cada posición en el exoma, se hace posible inferir los cambios relativos en el número de copias en la muestra de tumor.

Tras la imposibilidad de tener muestras pareadas, se optó por seleccionar una muestra control en cada pool como línea basal. El criterio en la elección de la muestra control se ha realizado analizando la cobertura media de la muestra y eligiendo la muestra con una cobertura media parecida a todas las muestras.

Este valor se ha obtenido mediante la herramienta SAMtools, utilizando el script;

```
samtools depth  /~ muestra NORMAL.bam  |  awk '{sum+=$3} END { print "Average = ",sum/NR}'
```

Mediante la realización de una comparación del par de profundidad de lectura entre las muestras en cada posición en el exoma, se hace posible inferir los cambios relativos en el número de copias en la muestra de tumor.

La salida de los archivos es; regiones del cromosoma, el incido y el final de la CNV y log-base-2 del cambio de número de copias, que es similar a los datos de número de copia basados en matrices y por lo tanto susceptibles de los mismos algoritmos de segmentación.

VarScan va a comparar nuestra muestra control frente a nuestra muestra tumoral. Está diseñada para ser utilizado en exomas donde la muestra control y tumoral se han capturado, secuenciado y procesan al mismo tiempo. De este modo disminuimos el error de sesgo debido a los errores de captura, secuenciación y procesamiento

Al realizar una comparación de las profundidades de lecturas entre los pares de muestras para cada posición en el exoma, es posible inferir cambios relativos en el número de copias en la muestra del tumor. El resultado de esta herramienta es un conjunto de regiones definidas por el cromosoma, inicio, parada de la región del cromosoma analizada y log-base-2 del cambio de número de copia.

3.3.1.2 Flujo de trabajo para VarScan

Vamos a seguir las recomendaciones de Koboldt, Daniel C. et al. (6) .

Lo primero que realizamos es detectar cambios en el número de copias somáticas

Este paso lo vamos a hacer mediante el programa SAMtools , para generar el archivo .mpileupo mediante el código :

samtools mpileup –B –q 1 –f reference.fasta normal.bam tumor.bam >normal-tumor.mpileup

Los datos que necesitamos introducir son;

- -q Calidad mínima de mapeo al menos 1.
- -B Deshabilitar el ajuste de BAQ. Se recomienda usarla para hacer la llamada de variantes con VarScan.
- -f Nuestro genoma de referencia en formato fasta
- Muestra control en formato. bam
- Muestra tumor a analizar en formato .bam
- Archivo de salida normal-tumor en formato .mpileup.

El script es;

samtools mpileup -B -q 1 -f

##Cargamos el genoma de referencia

~/Homo_sapiens.GRCh37.dna.primary_assembly.fa

##Cargamos la muestra control y la muestra tumoral como dos muestras pareadas

~/control/control.bam

~/ Tumor_17316.sort.nd.covered.bam

##Generamos un archivo de salida en formato mpileup

>/Tumor_17316.sort.nd.covered.mpileup

El segundo paso es realizar el copynumber mediante el programa VarScan

Las opciones son;

1. Archivo normal-tumor mpileup generado por SAMtools
2. Archivo de salida en formato .copynumber
3. --min-base-qual - Minimum base quality to count for coverage [20]
4. --min-map-qual - Minimum read mapping quality to count for coverage [20]
5. --min-coverage - Minimum coverage threshold for copynumber segments [20]
6. --min-segment-size - Minimum number of consecutive bases to report a segment [10]
7. --max-segment-size - Max size before a new segment is made [100]
8. --p-value - P-value threshold for significant copynumber change-point [0.01]
9. --data-ratio - The normal/tumor input data ratio for copynumber adjustment [1.0]

VarScan copynumber

Cargamos el archivo de mpileup generado anteriormente

~Tumor_17298.sort.nd.covered.mpileup

##Generamos un archivo de salida en formato copynumber

~ /Tumor_17298.copynumber

Introducimos los parámetros para el análisis.

--min-coverage 10 --data-ratio 1.0 --min-segment-size 20 --max-segment-size 100

El comando anterior informará todas las regiones contiguas que cumplan con el requisito de cobertura (10) tanto en normal como en tumor. Solo se informarán las

regiones de al menos 20 pb, y después de alcanzar las 100 bases, se iniciará una nueva región. Para cada región, VarScan calcula la cobertura media en tumor y normal, el valor log2 de la relación del tumor a las coberturas medias normales y el contenido de GC.

Los archivos de salida para VarScan son:

Campo	Descripción
• cromo	Nombre de cromosoma o referencia
• chr_start	Posición inicial del cromosoma
• chr_stop	Posición final del cromosoma
• normal_depth	Profundidad media de la secuencia en la normal.
• tumor_depth	Profundidad media de la secuencia en el tumor.
• log_ratio	Relación log-base-2 de la relación tumor / profundidad normal
• gc_content	Proporción de bases de GC en la región, entre 0 y 100 (v2.2.7 y posteriores)

El archivo que hemos obtenido tiene el siguiente formato.

chrom	chr_start	chr_stop	num_positions	normal_depth	tumor_depth	log2_ratio	gc_content
1	10385395	10385494	100	46,4	45,2	-0,037	36,0
1	10385495	10385549	55	34,7	30,0	-0,210	27,3
1	11046778	11046877	100	59,8	56,7	-0,077	39,0
1	11046878	11046931	54	38,6	37,7	-0,034	35,2
1	14096748	14096847	100	47,8	38,7	-0,304	30,0
1	14096848	14096893	46	33,1	27,1	-0,290	30,4
1	14804796	14804895	100	64,2	52,6	-0,288	39,0

Tabla 5. Datos obtenidos por VarScan

Ahora, cargamos los datos en R para realizar una segmentación binaria circular(24).

El algoritmo de segmentación binaria circular (CBS) divide el genoma en regiones de igual número de copias y hacer una estimación del número de copias de las regiones de análisis mediante un T estadístico para obtener el valor P correspondiente y decidir si hay cambio o no en el número de copias.

Mediante la función read.table y transfórmanos los datos en una variable CNAobject, seleccionando los datos genomdat como el log2_ratio, Chrom como las filas de los cromosomas y maploc como el inicio de la secuencia.

```
##Cargamos la librería

library(DNAcopy)

##Cargamos el archivo para analizar

P2 <- read.table( "Tumor_1.mpileup.copynumber", header = T, dec = ",")

##Realizamos la segmentación

CNA.object_P2 <- CNA( genomdat = P2[,7], chrom = P2[,1], maploc = P2[,2], data.type = 'logratio')

CNA.object.smoothed_P2 <- smooth.CNA(CNA.object_P2)

CNA.object.smoothed.seg_P2 <- segment(CNA.object.smoothed_P2, alpha=0.1, verbose = 0, min.width = 2)

seg.pvalue_P2 <- segments.p(CNA.object.smoothed.seg_P2, ngrid=100, tol=1e-6, alpha=0.05, search.range=100, nperm=1000)

##Obtenemos el archivo de salida con los valores en el número de copias

write.table (seg.pvalue_P2, file= "~out.file_N_T_1_P2.txt", sep="\t")
```

El archivo con el código esta en Github
https://github.com/kike1790/DNAcopy-Segmentation/blob/master/DNAcopy%20TFM%20E.Sevilla

Si abrimos el archivo out.file_N_T_1_P2.txt, ver tabla 8, podemos ver los siguientes campos, los más importantes son :

- Id de la muestra
- Cromosoma; nombre de cromosoma
- Inicio de la CNV
- Final de la CNV
- Num.mark; Número de marcadores en cada segmento
- Seg.mean; la media de los segmentos
- Pval; P valor

Los parámetros de valor mun.marck y segmentación media, nos indica si hay perdida o ganancia de CNVs.

ID"	chrom	loc.start	loc.end	num.mark	seg.mean	bstat	pval	lcl	ucl
Sample.1	1	10385395	241683015	170	-179	NA	NA	NA	NA
Sample.1	2	8178659	242371126	524	-0.1742	NA	NA	NA	NA
Sample.1	3	10070277	194858404	339	-0.1723	NA	NA	NA	NA
Sample.1	4	1734166	106061558	95	-0.1999	NA	NA	NA	NA
Sample.1	5	1279714	176722455	270	-0.1489	NA	NA	NA	

Tabla 8 Datos obtenidos en R, despúes de palicar CBS

Una vez realizado el CBS obtenemos un archive out.file_N_T_1_P2.txt que vamos a transformar en un formato de tabla compatible para poder utilizar el programa MergeSegment.pl. (Copyright 2009−2012 Daniel C. Koboldt and Washington University All rights reserved)

Ahora desde Shell vamos a realizar el MergeSegmentacion,. Este programa se ejecuta desde perl.

$ perl mergeSegments.pl

Archivo de la CBS/ out.file_N_T_1_P2.txt

--ref-arm-sizes $refArmSize

##Archivo de salida

--output-basename out.file_N_T_1_P2_VarScanOutput

Obtenemos dos archivos un archivo con extensión .tsv y otro con extensión txt.

Archivo de salida .txt

segments	merged_events	amps	large-scale	focal	dels	large-scale	focal	amp_regions	del_regions
22	20	0	0	0	13	11	2	NA	chr19p,chr21q

Tabla 8. Datos obtenidos despues de aplicar MergeSegmen. Archivo de salida .txt

Archivo de salida .TSV

chrom	chr_start	chr_stop	seg_mean	num_segments	num_markers	p_value	event_type	event_size	size_class	chrom_arm	arm_fraction	chrom_fraction
chr1	10385395	241683015	-179	1	170	0	neutral	231297621	large-scale	chr1q	94 %	93 %
chr3	10070277	194858404	-0.1723	1	339	0	neutral	184788128	large-scale	chr3q	97 %	93 %

chr5	127971455	176722455	-0.1489	1	270	0	neutral	175442742	large-scale	chr5p	97 %	97 %
chr7	601297025	148544425	-0.1406	1	284	0	neutral	142531456	large-scale	chr7p	90 %	90 %
chr8	145738541	145738575	-973	1	2	1.90467173890091e-45	deletion	35	focal	chr8q	0 %	0 %
chr8	145741288	145742188	-0.0443	1	10	3.10779538898308e-07	neutral	901	focal	chr8q	0 %	0 %
chr8	145742358	145743124	-1.1099	2	8	0.0067104615670533	deletion	767	focal	chr8q	0 %	0 %
chr9	100459333	100556134	-1.03725	2	6	0.00365590147313235	deletion	96802	focal	chr9q	0 %	0 %
chr9	135819873	136149256	-0.4565	1	4	0	deletion	329384	focal	chr9q	0 %	0 %
chr10	88516318	88516627	-1.422	1	5	6.5592292159473 2e-33	deletion	310	focal	chr10q	0 %	0 %
chr11	532559	2906419	-1.80065	2	27	1.16096234389563e-18	deletion	2373861	focal	chr11p	4 %	2 %
chr11	32456471	32456847	-1.559	1	6	6.5703682047846 1e-71	deletion	377	focal	chr11p	0 %	0 %
chr12	12870695	13325025	-0.1886	1	115	0	neutral	120379571	large-scale	chr12q	99 %	90 %
chr13	48877972	48878147	-1.9723	1	3	1.94683124378057e-86	deletion	176	focal	chr13q	0 %	0 %
chr14	36649167	96103121	-0.1266	1	217	0	neutral	594539 55	large-scale	chr14q	66 %	55 %
chr16	2097390	2098003	-1.2897	1	8	5.0637492118378 5e-76	deletion	614	focal	chr16p	0 %	0 %
chr16	89882970	90066963	-621	1	6	0	deletion	183994	focal	chr16q	0 %	0 %
chr18	43309834	48556969	-1.48505	2	12	7.7799123950080 5e-27	deletion	5247136	focal	chr18q	9 %	7 %
chr19	12068435	51364649	-1.19485	6	121	0.0111081255324071	deletion	50157807	large-scale	chr19p	95 %	85 %
chr21	30717664	43778920	-0.4753	1	33	0	deletion	13061257	large-scale	chr21q	37 %	27 %

Tabla 9. Datos obtenidos despues de aplicar MergeSegmen. Archivo de salida .tsv

Una vez obtenida la tabla de variantes, tenemos que introducir los genes que coinciden con las posiciones cromosómicas. Para ello, vamos a realizar un bedtools (12) con la función intersect con los archivos de salida de VarScan y el Bedfile del trabajo. De esta manera seleccionaremos aquellos genes que coinciden con el Bedfile y los que no coinciden serán regiones/CNVs de novo.

Lo primero que hacemos es tener los archivos en el formato adecuado para poder lanzar Bedtools, los archivos que vamos a analizar tienen que estar en extensión. bed.

A continuación transformamos el archivo de salida de VarScan a un archivo .bed

cat ~/VarScanchrstarend.txt > ~/VarScanchrstarend.bed

A continuación, lanzamos Bedtools.

Bedtools intersect -a ~/PEC4/VarScanchrstarend.bed -b ~/Bedfile_1/Bedfile_copia.bed -f 0.50 -wa -wb > VarScan.txt

Generamos una salida en .txt para poder ser visualizada. De este modo, se selecciona aquella CNVs que se encuentran en las regiones de captura.

El archivo con el código esta en Github
https://github.com/kike1790/VArScan-TFM-E.Sevilla/blob/master/VarScan%20TFM%20E.Sevilla

3.3.1.3 Comprobación con IGV

Comprobamos la delección en la muestra 17343 en el cromosoma chr1348877972-48878172, frente a la muestra control. Podemos ver, figura 13, hay una perdida de cobertura, relacionada con una delección.

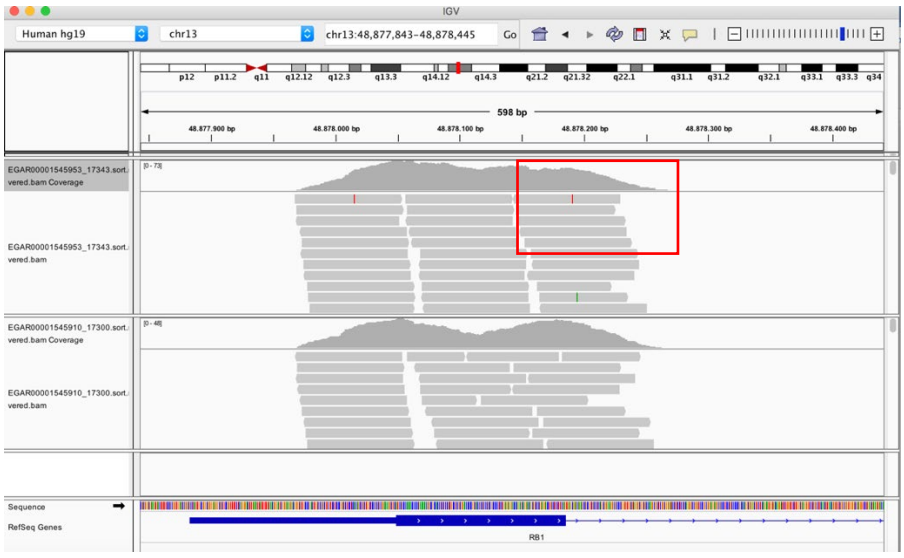


Figura 13. Representación en IGV de la delección en el cromosoma 13 vs a muestra control para los datos obtenidos de VarScan.

4. Análisis Estadístico.

4.1. Sensibilidad y Especificidad

El Trabajo ICR96, contenía 96 muestras independientes, de las cuales 66 son muestras que al menos tienen una CNVs validada por MLPAs.

Por otro lado, 30 muestras que ha dado resultados negativos para CNVs mediante la técnica de MLPAs(25).

Estos datos están resumidos en la tabla 2, descrita en material y métodos(26).Donde se muestra todas las muestras validadas por MLPAs para tumor y para control.

Se ha trabajo con menos datos ya que tuvimos problemas con la descarga del servidor y no pudimos obtener todos los datos.

Después de filtrar y quitar aquellos datos que no pudimos descargar obtenemos la siguiente tabla 10:

Nº de muestras	89
Positivos	55 positivos: • Tumor_pool1: 28 • Tumor_pool2: 27
Negativos	34 negativos • Control_pool1:15 • Control_pool2: 19
Genes 32 46 exones CNVs	*APC, ATM, BAP1, BARD1, BMPR1A, BRCA1, BRCA2, BRIP1, CDH1, CDK4, CDKN2A, CHEK2*, *EPCAM* (exon 9 only) , *FH, MLH1*, *MSH2*,*MSH6*, *MUTYH, NBN, NF1, NSD1, PALB2, PMS2* (excluding exons 12-15), *PTEN, RAD51C, RAD51D, RB1, SDHB, SMAD4, STK11, TP53* and *WT1* .

Tabla 10. Una vez seleccionado las muestras descargadas ajustamos el nº de CNVs para cada Tumor/Control y Pool1 y pool2.

Las muestras por analizar se han separado en dos pools ya que fueran secuenciadas en dos pools diferentes en la cellflow, de esta manera disminuimos el sesgo estadístico respecto al procesamiento.

Vamos a calcular la Sensibilidad y Especificidad de cada una de las herramientas bioinformáticas. Para ello utilizaremos el test de diagnóstico, el cual nos va a servir para calcular una serie de valores estadísticos para determinar si el test que estamos realizando, herramienta bioinformática, es un buen predictor a la hora de estimar si la muestra de análisis tiene o no CNVs. (27)

Para ello vamos a calcular dos valores; la Sensibilidad y la Especificidad del test.

La Sensibilidad se define como la capacidad de clasificar correctamente a una muestra con una CNV, es decir, la capacidad de que en una muestra se obtenga una CNV como resultado positivo. En otras palabras, la Sensibilidad viene a ser la capacidad de que la herramienta bioinformática que estamos analizando pueda detectar si la muestra tiene o no CNV. La ecuación para el cálculo es:

Sensibilidad=(VP)/(VP+FN)

Por otra parte, la Especificidad es la capacidad de clasificar correctamente a una CNV, es decir, la capacidad de que para una muestra sin CNV se obtenga un resultado negativo. La capacidad para detectar a los individuos sanos. La fórmula en este caso es:

Especificidad=(VN)/(VN+FP)

Ver tabla 11 como representación de los parámetros descritos

Resultado de la prueba	**Verdadero diagnóstico**	
	Validado MLPA SI	**Validado MLPA NO**
Positivo	Verdaderos Positivos	Falsos Positivos
	(VP)	(FP)
Negativo	Falsos Negativos	Verdaderos Negativos
	(FN)	(VN)

Tabla 11 Test Diagnostico para los VP;FP;FN y VN

Después de aplicar las Herramientas bioinformáticas sobre las muestras a analizar hemos obtenido los siguientes datos;

Se muestra la tabla 12, con la descripción de los genes, coordenadas y muestra de las CNVs validadas por MLPAs del pool1.:

SampleID	ICR96 Pool	Gene	MLPAResult	Result Type	ExonCNVType	ExonCNVSize	Chromosome	5PrimeExon37	3PrimeExon37
17296	1	SDHB	Exon 1 deletion	ExonCNV	Deletion	Single	1	17380443	17380514
17297	1	FH	Exon 1-10 deletion	ExonCNV	Deletion	Multi	1	241661128	241683022
17298	1	BRCA2	Exon 21 duplication	ExonCNV	Duplication	Single	13	32950807	32950928
17301	1	MSH2	Exon 8 duplication	ExonCNV	Duplication	Single	2	47672687	47672796
17302	1	BRCA1	Exon 5-7 deletion	ExonCNV	Deletion	Multi	17	41256139	41258550
17303	1	MSH2	Exon 9-16 deletion	ExonCNV	Deletion	Multi	2	47690170	47710088
17304	1	PALB2	Exon 13 deletion	ExonCNV	Deletion	Single	16	23614780	23614990
17305	1	BRCA2	Exon 3 deletion	ExonCNV	Deletion	Single	13	32893214	32893462
17306	1	BRCA1	Exon 20 deletion	ExonCNV	Deletion	Single	17	41209069	41209152
17307	1	CHEK2	Exon 3-4 deletion	ExonCNV	Deletion	Multi	22	29120965	29121355

17316	1	MSH2	Exon 8 duplication	ExonCNV	Duplication	Single	2	47672687	47672796
17317	1	BRCA1	Exon 3 deletion	ExonCNV	Deletion	Single	17	41267743	41267796
17318	1	MSH6	Exon 1-8 deletion	ExonCNV	Deletion	Multi	2	48010373	48033497
17319	1	PMS2	Exon 11 duplication	ExonCNV	Duplication	Single	7	6026390	6027251
17320	1	BRCA1	Exon 24 deletion	ExonCNV	Deletion	Single	17	41197695	41197819
17325	1	NSD1	Exon 9-13 deletion	ExonCNV	Deletion	Multi	5	176671196	176684152
17326	1	BRCA1	Exon 22 deletion	ExonCNV	Deletion	Single	17	41201138	41201211
17331	1	PMS2	Exon 6-8 deletion	ExonCNV	Deletion	Multi	7	6035165	6038906
17332	1	CHEK2	Exon 3-15 duplication	ExonCNV	Duplication	Multi	22	29083885	29121355
17333	1	BRCA1	Exon 13 deletion	ExonCNV	Deletion	Single	17	41234421	41234592
17335	1	BRCA2	Exon 14-16	ExonCNV	Deletion	Multi	13	32928998	32932066

			deletion						
17336	1	RB1	Exon 24-27 deletion	ExonCNV	Deletion	Multi	13	49047496	49054207
17337	1	MLH1	Exon 1-19 duplication	ExonCNV	Duplication	Multi	3	37035039	37092144
17338	1	NF1	Exon 37-57 deletion	ExonCNV	Deletion	Multi	17	29652838	29687721
17340	1	BRCA1	Exon 5-8 duplication	ExonCNV	Duplication	Multi	17	41251792	41258550
17341	1	PMS2	Exon 9-10 deletion	ExonCNV	Deletion	Multi	7	6029431	6031688
17342	1	BRCA2	Exon 1-11 deletion	ExonCNV	Deletion	Multi	13	32889611	32915333
17343	1	MSH2	Exon 9-10 deletion	ExonCNV	Deletion	Multi	2	47690170	47693947

Tabla 12. Tabla de los datos validados por gen, coordenadas y muestra, para el pool1

Se muestra la tabla 13, con la descripción de los genes, coordenadas y muestra de las CNVs validadas por MLPAs del pool2.:

SampleID	ICR96 Pool	Gene	MLPA Result	Result Type	ExonCNVType	ExonCNVSize	Chromosome	5PrimeExon37	3PrimeExon37
17357	2	BRCA2	Exon 21-24	ExonCNV	Deletion	Multi	13	32950807	32954282

			deletion						
17358	2	SDHB	Exon 2-7 deletion	ExonCNV	Deletion	Multi	1	17349103	17371383
17359	2	MSH2	Exon 1-2 deletion	ExonCNV	Deletion	Multi	2	47630331	47635694
17362	2	BRCA1	Exon 13 duplication	ExonCNV	Duplication	Single	17	41234421	41234592
17363	2	CHEK2	Exon 9-10 deletion	ExonCNV	Deletion	Multi	22	29092889	29095925
17364	2	BRCA2	Exon 21 duplication	ExonCNV	Duplication	Single	13	32950807	32950928
17365	2	TP53	Exon 1 deletion	ExonCNV	Deletion	Single	17	7590695	7590856
17366	2	EPCAM	Exon 9 duplication	ExonCNV	Duplication	Single	2	47613711	47613752
17366	2	MSH2	Exon 1-16 duplication	ExonCNV	Duplication	Multi	2	47630331	47710088
17369	2	BRCA1	Exon 16 deletion	ExonCNV	Deletion	Single	17	41222945	41223255
17372	2	EZH2	Exon 1-20 deletion	ExonCNV	Deletion	Multi	7	148504738	148504798

17375	2	CHEK2	Exon 3-15 duplication	ExonCNV	Duplication	Multi	22	29083885	29121355
17378	2	CHEK2	Exon 8-12 deletion	ExonCNV	Deletion	Multi	22	29091115	29099554
17379	2	PMS2	Exon 9-10 deletion	ExonCNV	Deletion	Multi	7	6029431	6031688
17380	2	BRCA2	Exon 1-3 duplication	ExonCNV	Duplication	Multi	13	32889611	32893462
17381	2	NSD1	Exon 1-2 deletion	ExonCNV	Deletion	Multi	5	176560926	176563031
17383	2	NSD1	Exon 1-5 deletion	ExonCNV	Deletion	Multi	5	176560926	176639196
17384	2	BRCA1	Exon 12 deletion	ExonCNV	Deletion	Single	17	41242961	41243049
17385	2	MSH6	Exon 5-6 deletion	ExonCNV	Deletion	Multi	2	48030559	48032166
17386	2	MSH2	Exon 7 deletion	ExonCNV	Deletion	Single	2	47656881	47657080
17388	2	RAD51C	Exon 4-9 deletion	ExonCNV	Deletion	Multi	17	56780557	56811583
17390	2	MSH2	Exon 4-5	ExonCNV	Deletion	Multi	2	47639553	47641557

			deletion						
17390	2	PMS2	Exon 1-15 duplication	ExonCNV	Duplication	Multi	7	6013030	6048650
17392	2	PTEN	Exon 1 deletion	ExonCNV	Deletion	Single	10	89624227	89624305
17393	2	BRCA2	Exon 1-2 deletion	ExonCNV	Deletion	Multi	13	32889611	32890664
17394	2	BRCA1	Exon 1-2 deletion	ExonCNV	Deletion	Multi	17	41276034	41277387
17395	2	WT1	Exon 1-10 deletion	ExonCNV	Deletion	Multi	11	32410604	32456891
17398	2	BRCA1	Exon 13 duplication	ExonCNV	Duplication	Single	17	41234421	41234592
17400	2	BRCA1	Exon 21-24 deletion	ExonCNV	Deletion	Multi	17	41197695	41203134
17403	2	BRCA1	Exon 17 deletion	ExonCNV	Deletion	Single	17	41219625	41219712

Tabla 13. Tabla de los datos validados por gen, coordenadas y muestra para pool2

La evaluación de la Sensibilidad y estabilidad para las herramientas bioinformáticas han sido por pool.

4.1.1.1 ExomeDepth pool1

Los resultados obtenidos para la herramienta bioinformática ExomeDepth fueron, ver tabal 14:

Herramientas Bioinf.	Total	CNVs
ExomeDepth_P1 ICR96_P1	2	RB1 BRCA2
ICR96_P1	13	SDHB MLH1 BRCA1 FH TP53 PMS2 NSD1 MSH6 MSH2 ATM NF1 PALB2 CHEK2
ExomeDepth_P1	13	XPA FANCA VHL GATA2 FANCE RECQL4 EPCAM TSC2 CDKN1C BMPR1A RET CEBPA ERCC2

Tabla 14 Número de CNVs y genes para ExomeDepth pool1

CNVs obtenidas por ExomeDepth para el pool1: 15 CNVS en total.

Las CNVs no coincidentes para ExomeDepth para el pool1 han sido 13 CNVS.

Cromosoma	inicio	final	gen	Tipo
chr2	47596644	47596721	EPCAM	deletion
chr3	128205645	128205875	GATA2	deletion
chr3	10183531	10183872	VHL	deletion
chr6	35420322	35420571	FANCE	duplication
chr8	145742797	145742893	RECQL4	duplication
chr9	100459402	100459575	XPA	deletion
chr10	43572706	43572780	RET	deletion
chr10	88516384	88516688	BMPR1A	duplication
chr11	2905233	2905365	CDKN1C	duplication
chr16	2097454	2098078	TSC2	deletion
chr16	89882944	89883024	FANCA	deletion
chr19	33792243	33793321	CEBPA	duplication
chr19	45867243	45867378	ERCC2	deletion

Aunque ExomeDepth detecto 2 CNV coincidentes en el pool, solo 1 de ellas coincidía con la misma muestra con el mismo cromosoma y con mismo tipo (delección o amplificación) que los datos validados por MLPAs por el trabajo. A este concepto de coincidencia en muestra, posición cromosómica, gen, y tipo de delección o amplificación se le va a llamar muestra/tipo.

Tabla de Coincidencias					
Herramienta Bioinformática	Cromosoma	Incio	Final	Gen	Tipo de deleccion
EDepth pool1	chr13	32900378	32900420	BRCA2	deleccion

Verdaderos positivos serán las CNVs que coinciden entre las CNVs Validadas por el trabajo por MLPAs y la encontradas por ExomeDepth

Falsos positivos serán el número de CNVs detectadas por ExomeDepth menos el número de Verdaderos positivos.

Falsos negativos serán el número de CNVs validadas por el trabajo por MLPAs y que no han sido detectadas por ExomeDepth.

Verdaderos negativos serán aquellas CNVs que no han sido validadas por el trabajo mediante MLPAs ni detectadas por ExomeDepth. 46 son las posibles CNVs que hay descritas para los genes analizados y están descritas por el trabajo. Restaremos 46-15-13 “ICR-MLPAs-FP”

Test diagnóstico

Resultado de la prueba	**ExomeDepth pool1**	
	CNV	**No CNV**
test Positivo	VP	FP
	1	14
test Negativo	FN	VN
	14	7

Tabla 15 Test Diagnostico para los VP;FP;FN y VN de ExomeDepth pool1

Sensibilidad	VP/(VP+FN)	0,07
Especificidad	VN/(VN+FP)	0,33
Valor predictivo positivo	VP/(VP+FP)	0,07
Valor predictivo negativo	VN/(VN+FN)	0,33
Razón de verosimilitud positiva	Sensibilidad/1-Especificidad	0,10
Razón de verosimilitud negativa	1-Sensibilidad/Especificidad	2,80

Tabla 16 de Resultados para la Sensibilidad, y Especificidad de ExomeDepth pool1

4.1.1.2 ExomeDepth pool 2

Los resultados obtenidos para la herramienta bioinformática ExomeDepth fueron, ver tabla 17:

Herramientas Bioinf.	Total	CNVs
ExomeDepth_P2 ICR96_P2	4	BRCA2 PMS2 EPCAM CHEK2
ICR96_P2	11	RAD51C BRCA1 NSD1 MSH2 WT1 SDHB TP53 EZH2 PTEN MSH6 ATM
ExomeDepth_P2	17	RB1 KIT MUTYH XPA FANCA VHL GATA2 RECQL4 TSC2 CDKN1C BMPR1A RET RUNX1 MET CEBPA ERCC2 STK11

Tabla 17 Número de CNVs y genes para ExomeDepth pool2

CNVs obtenidas por ExomeDepth para el pool2: 21 CNVS en total.

No coincidentes 17

Cromosoma	inicio	final	gen	Tipo
chr1	45797091	45797229	MUTYH	deletion
chr3	128205645	128205875	GATA2	deletion
chr3	10183531	10183872	VHL	deletion
chr4	55593383	55593491	KIT	deletion
chr7	116339138	116340339	MET	deletion
chr8	145742797	145742893	RECQL4	duplication
chr9	100459402	100459575	XPA	deletion
chr10	43572706	43572780	RET	deletion
chr10	88516384	88516688	BMPR1A	deletion
chr11	2905233	2905365	CDKN1C	deletion
chr13	48878048	48878186	RB1	deletion
chr16	2097454	2098078	TSC2	deletion
chr16	89882944	89883024	FANCA	deletion
chr19	45860527	45860630	ERCC2	deletion
chr19	1220579	1220717	STK11	deletion
chr19	33792243	33793321	CEBPA	deletion
chr21	36164431	36164908	RUNX1	deletion

Aunque ExomeDepth detecto 4 CNV coincidentes en el pool, solo 1 de ellas coincidía muestra/tipo

Tabla de Coincidencias					
Herramienta Bioinformática	Cromosoma	Incio	Final	Gen	Tipo de deleccion
EDepth pool2	chr22	29120964	29121113	CHEK2	deleccion

Test Diagnostico

Resultado de la prueba	ExomeDepth pool2	
	CNV	No CNV
test Positivo	VP	FP
	1	20
test Negativo	FN	VN
	14	11

Tabla 18 Test Diagnostico para los VP;FP;FN y VN de ExomeDepth pool2

Sensibilidad	VP/(VP+FN)	0,07
Especificidad	VN/(VN+FP)	0,35
Valor predictivo positivo	VP/(VP+FP)	0,05
Valor predictivo negativo	VN/(VN+FN)	0,44
Razón de verosimilitud positiva	Sensibilidad/1-Especificidad	0,10
Razón de verosimilitud negativa	1-Sensibilidad/Especificidad	2,63

Tabla 19 Resultados para la Sensibilidad, y Especificidad de ExomeDepth pool2

4.1.2.1 CNVkit pool 1

Los resultados obtenidos para la herramienta bioinformática CNVkit fueron, ver tabla 20. Para pool 1.

Herramienta Bioinf.	Total	CNVs
CNVkit_P1 ICR96_P1	6	MLH1 FH BRCA2 NSD1 MSH6 NF1
ICR96_P1	9	RB1 SDHB BRCA1 TP53 PMS2 MSH2 ATM PALB2 CHEK2
CNVkit_P1	9	POLD1_S478N BAP1 POLD1_L474P RUNX1 FANCL CEBPA SMAD4 ERCC2 STK11

Tabla 20 Número de CNVs y genes para CNVkit para pool1

CNVs obtenidas por CNVkit para el pool1: 15 CNVS en total.

No coincidentes 9

Cromosoma	inicio	final	gen	tipo
19	1206912	1228446	STK11	deletion
19	33792243	33793321	CEBPA	deletion
19	45854886	45873799	ERCC2	deletion
19	50909689	50909725	POLD1_L474P	deletion

19	50909689	50909725	POLD1_S478N	deletion
21	36164431	36421197	RUNX1	amplificación
18	48556571	48604838	SMAD4	deletion
2	58386899	58468449	FANCL	deletion
3	52436303	52436438	BAP1	amplificación

Y 6 coincidentes entre muestra/tipo y validadas por MLPAs

Tabla de Coincidencias					
Herramienta Bioinformática	Cromosoma	Incio	Final	Gen	Tipo de deleccion
CNVkit pool1	chr1	241661127	241683023	FH	deleccion
CNVkit pool1	chr2	48010372	48033498	MSH6	deleccion
CNVkit pool1	chr5	176671195	176684153	NSD1	deleccion
CNVkit pool1	chr3	37035038	37092145	MLH1	amplicifacion
CNVkit pool1	chr17	29652837	29687722	NF1	deleccion
CNVkit pool1	chr13	32889599	32921034	BRCA2	deleccion

Test diagnostico

Resultado de la prueba	**CNVkit pool1**	
	CNV	**No CNV**
test Positivo	VP	FP
	6	9
test Negativo	FN	VN
	9	22

Tabla 21 Test Diagnostico para los VP;FP;FN y VN de CNVkit para pool1

Sensibilidad	VP/(VP+FN)	0,40
Especificidad	VN/(VN+FP)	0,71
Valor predictivo positivo	VP/(VP+FP)	0,40
Valor predictivo negativo	VN/(VN+FN)	0,71
Razón de verosimilitud positiva	Sensibilidad/1-Especificidad	1,38
Razón de verosimilitud negativa	1-Sensibilidad/Especificidad	0,85

Tabla 22 Resultados para la Sensibilidad, y Especificidad de CNVkit para pool1

4.1.2.2 CNVkit pool2

Los resultados obtenidos para la herramienta bioinformática CNVkit fueron, ver tabla 23. Para el pool2

Herramienta Bioinf.	Total	CNVs
CNVkit_P2 ICR96_P2	9	RAD51C NSD1 MSH2 EPCAM WT1 CHEK2 TP53 PMS2 MSH6
ICR96_P2	6	BRCA1 SDHB BRCA2 EZH2 PTEN ATM
CNVkit_P2	28	PPM1D CDK4 CDKN1B WRN ERCC3 RECQL4 CYLD CEBPA MET NBN ERCC2 BLM FANCI BARD1 HRAS BRIP1 FANCF PMS1 CDKN1C RUNX1 HNF1A FANCL FLCN APC SMAD4 STK11 NF1 FANCM

Tabla 23 Número de CNVs y genes para CNVkit para pool2

CNVs obtenidas por CNVkit para el pool2: 52 CNVS en total.

No coincidentes 28

Cromosoma	inicio	final	gen	**tipo**
2	58386899	58468449	FANCL	deleccion
2	128015171	128051658	ERCC3	deleccion
2	190656535	190742163	PMS1	deleccion
2	215593388	215674305	BARD1	deleccion
5	112073573	112163704	APC	deleccion
7	116339138	116419012	MET	deleccion
8	30915963	31030619	WRN	deleccion
8	90947809	90983519	NBN	deleccion
8	145736813	145743169	RECQL4	deleccion
11	532635	534323	HRAS	deleccion
11	2905233	2906720	CDKN1C	deleccion
11	22646231	22647357	FANCF	deleccion
12	12870762	12875317	CDKN1B	amplificación
12	58142307	58146316	CDK4	deleccion
12	121416571	121438996	HNF1A	amplificación
14	45605234	45658566	FANCM	deleccion
15	89828326	89859691	FANCI	amplificación
15	91290622	91358510	BLM	amplificación
16	50783609	50815323	CYLD	deleccion
17	17116968	17119818	FLCN	amplificación
17	29422327	29665158	NF1	amplificación

17	58740355	58740914	PPM1D	amplificación
17	59760656	59940767	BRIP1	amplificación
18	48556571	48604838	SMAD4	amplificación
19	1206912	1228446	STK11	amplificación
19	33792243	33793321	CEBPA	amplificación
19	45854886	45873799	ERCC2	amplificación
21	36164431	36421197	RUNX1	amplificación

Aunque ExomeDepth detecto 9 CNV coincidentes en el pool, solo 7 de ellas coincidía muestra/tipo

Tabla de Coincidencias					
Herramienta Bioinformática	Cromosoma	Incio	Final	Gen	Tipo de deleccion
CNVkit pool2	chr11	32410603	32456892	WT1	deleccion
CNVkit pool2	chr7	6013029	6048651	PMS2	amplificacion
CNVkit pool2	chr5	176618884	176639197	NSD1	deleccion
CNVkit pool2	chr22	29083884	29121356	CHEK2	amplificacion
CNVkit pool2	chr2	47596644	47613753	EPCAM	amplificacion
CNVkit pool2	chr2	47630330	47710089	MSH2	amplificacion
CNVkit pool2	chr7	148504737	148544391	EZH2	deleccion

Test diagnostico

Resultado de la prueba	**CNVkit pool2**	
	CNV	**No CNV**
test Positivo	VP	FP
	7	30
test Negativo	FN	VN
	8	1

Tabla 24 Test Diagnostico para los VP;FP;FN y VN de CNVkit para pool2

Sensibilidad	VP/(VP+FN)	0,47
Especificidad	VN/(VN+FP)	0,10
Valor predictivo positivo	VP/(VP+FP)	0,20
Valor predictivo negativo	VN/(VN+FN)	0,27
Razón de verosimilitud positiva	Sensibilidad/1-Especificidad	0,52
Razón de verosimilitud negativa	1-Sensibilidad/Especificidad	5,51

Tabla 25 Número de CNVs y genes para CNVkit para pool2

4.1.3.1 VarScan pool1

Los resultados obtenidos para la herramienta bioinformática VarScan fueron, ver tabla 26. Para el pool1

Herramienta Bioinf.	Total	CNVs
ICR96_pool1 VarScan_pool1	1	RB1
ICR96_pool1	14	SDHB MLH1 BRCA1 FH BRCA2 TP53 PMS2 NSD1 MSH6 MSH2 ATM NF1 PALB2 CHEK2
VarScan_pool1	12	ERCC4 MEN1 XPA GATA2 WRN ALK TSC2 CDKN1C CEBPA SMAD4 STK11 WT1

Tabla 26 Número de CNVs y genes para VarScan pool1

CNVs obtenidas por VarScan para el pool1: 13 CNVS en total.

No coincidentes 14

chromosome	start	end	Gen
chr11	2905899	2906720	CDKN1C
chr11	32456245	32456892	WT1
chr11	64571805	64572289	MEN1
chr16	2097454	2098078	TSC2
chr16	14014022	14014230	ERCC4
chr18	48556571	48557005	SMAD4
chr19	33792243	33793321	CEBPA
chr19	1227580	1228446	STK11
chr2	30142858	30143526	ALK
chr3	128204569	128205212	GATA2
chr9	100459402	100459575	XPA
chr8	31024468	31024768	WRN

Aunque ExomeDepth detecto 1 CNV coincidentes en el pool, hubo nulos resultados en la detección con muestra/tipo. Por lo que no se pudo calcular los valores de Sensibilidad y especificidad para el pool1

Resultado de la prueba	**VarScan pool1**	
	CNV	**No CNV**
test Positivo	VP	FP
	0	13
test Negativo	FN	VN
	15	18

Sensibilidad	VP/(VP+FN)	0,00
Especificidad	VN/(VN+FP)	0,58
Valor predictivo positivo	VP/(VP+FP)	0,00
Valor predictivo negativo	VN/(VN+FN)	0,55
Razón de verosimilitud positiva	Sensibilidad/1-Especificidad	0,00
Razón de verosimilitud negativa	1-Sensibilidad/Especificidad	1,72

4.1.3.2 VarScan pool2

Los resultados obtenidos para la herramienta bioinformática VarScan fueron, ver tabla 29. Para el pool2

	Total	CNVs
ICR96_pool1 VarScan_pool1	11	BRCA1 NSD1 MSH2 EPCAM WT1 CHEK2 BRCA2 PMS2 PTEN MSH6 ATM
ICR96_pool1	4	RAD51C EZH2 SDHB TP53
VarScan_pool1	27	ERCC4 RB1 PRF1 MEN1 FANCA GATA2 FANCE CYLD SMARCB1 DIS3L2 RHBDF2 MET NBN ERCC2 MUTYH MLH1 TMEM127 FANCD2 HRAS ALK TSC2 RET CDKN1C SLX4 APC NF1 FANCM

Tabla 29 Número de CNVs y genes para VarScan pool2

CNVs obtenidas por VarScan para el pool2: 38 CNVS en total.

No coincidentes 27

chr1	45800062	45800184	MUTYH
chr10	43623559	43623718	RET
chr10	72357808	72358938	PRF1
chr11	64571805	64572289	MEN1
chr11	532635	532756	HRAS
chr11	2905899	2906720	CDKN1C

chr13	49051490	49051541	RB1
chr14	45652976	45653106	FANCM
chr16	50830234	50830420	CYLD
chr16	2098616	2098755	TSC2
chr16	3632342	3632695	SLX4
chr16	14014022	14014230	ERCC4
chr16	89849414	89849511	FANCA
chr17	74471111	74471224	RHBDF2
chr17	29556042	29556484	NF1
chr19	33792243	33793321	ERCC2
chr19	1206912	1207203	ALK
chr2	29445382	29445474	TMEM127
chr2	47639552	47639700	DIS3L2
chr2	218310339	218310341	SMARCB1
chr22	24176327	24176368	MLH1
chr3	10183531	10183872	GATA2
chr3	37090007	37090101	FANCD2
chr3	128038372	128038374	APC
chr4	1734238	1734240	FANCE
chr5	176562104	176563032	MET
chr5	1279789	1279791	NBN

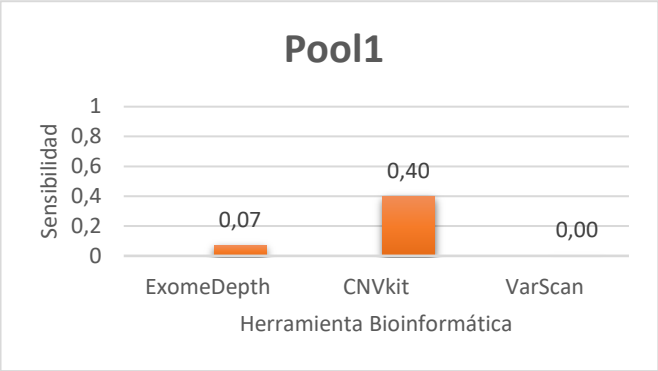
Aunque ExomeDepth detecto 11 CNV coincidentes en el pool, hubo nulos resultados en la detección con muestra/tipo.

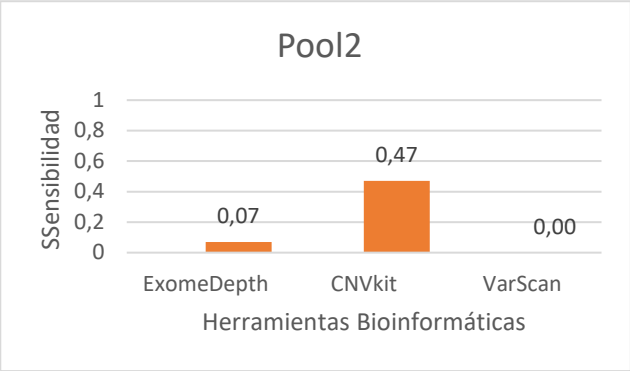
Resultado de la prueba	**VarScan pool2**	
	CNV	**No CNV**
test Positivo	VP	FP
	0	38
test Negativo	FN	VN
	15	4

Sensibilidad	VP/(VP+FN)	0,00
Especificidad	VN/(VN+FP)	0,10
Valor predictivo positivo	VP/(VP+FP)	0,00
Valor predictivo negativo	VN/(VN+FN)	0,21
Razón de verosimilitud positiva	Sensibilidad/1-Especificidad	0,00
Razón de verosimilitud negativa	1-Sensibilidad/Especificidad	10,50

4.2 Gráficas de Sensibilidad y Especificidad

Las representaciones gráficas de los datos de Sensibilidad para las tres herramientas bioinformáticas se muestran en las gráficas 2 y 3



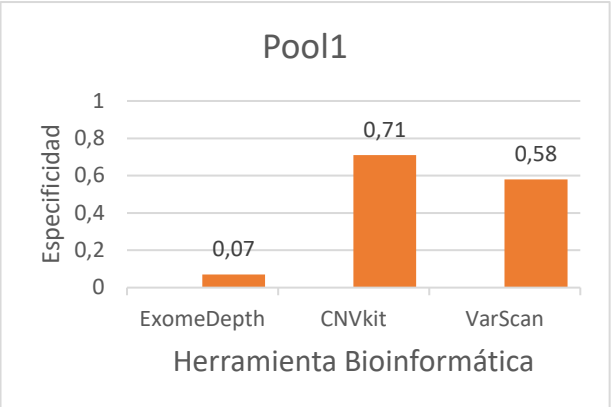


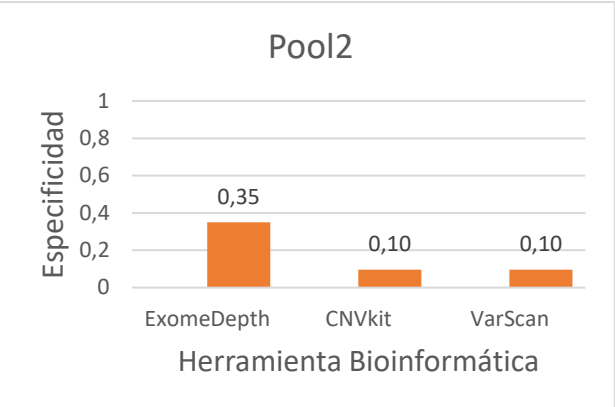
Gráfica 2 y 3 Sensibilidad de las herramientas bioinformáticas para las tres herramientas bioinformáticas para los pools 1 y 2.

Como podemos analizar en las gráficas 2, para el pool 1, la herramienta bioinformática CNVkit fue la que más alta Sensibilidad obtuvo, alrededor de 0,4. Aunque esta Sensibilidad es muy baja para poder determinar que sea una herramienta adecuada. Con las otras dos herramientas se obtuvieron valores más bajos y para VarScan ninguna coincidencia. Estos malos resultados pueden ser debidos a muchos factores, pero podría ser que los datos no han sido correctamente procesados o pueda existir algún error en la implantación del código.

En el análisis de los datos para el pool2 nos revelo unos valores de Sensibilidad más altos fueron los de la herramienta bioinformática CNVkit con un valor de 0,47. Aunque sigue siendo valores muy bajos para un detector de CNVs. Para la herramienta ExomeDepth el valor fue casi nulo y para VarScan nulo.

Las gráficas obtenidas para la Especificidad nos dieron como resultado





Gráfica 8 y 9 : Valores de especificidad de las herramientas bioinformáticas para los pool1 y pool2

Las herramientas bioinformáticas CNVkit y VArScan dieron valores relativamente altos 0,71 y 0,58 respectivamente para el pool1. Esto es fundamentalmente debido a los pocos FP que se encontramos en las herramientas. En Cambio, para el pool2 la herramienta bioinformática con valores más altos fue ExomeDepth con 0,35

4.2 Representación de los diagramas de Venn

Los diagramas de Venn se usan para mostrar gráficamente la agrupación de elementos en conjuntos, representando cada conjunto mediante un círculo. En nuestro caso vamos a representar 3 conjuntos de datos y sus intersecciones.(28)

Hemo utilizado la herramienta bioinformática de la pagina web;
(29) http://bioinformatics.psb.ugent.be/webtools/Venn/

Vamos a representar mediante diagrama de Venn el cruce de las 3 herramientas bioinformáticas analizadas respecto a los datos validados por el ICR96.

4.2.1 Datos obtenidos para el cruce entre las tres herramientas bioinformáticas para el pool1

Herramienta Bioinf.	Total	CNVs
CNVkit p1 ExomeD p1 VarS p1	1	CEBPA
CNVkit p1 ExomeD p1	2	ERCC2 BRCA2
ExomeD p1 VarS p1	6	RB1 GATA2 BMPR1A XPA TSC2 CDKN1C
CNVkit p1 VarS p1	2	SMAD4 STK11
ExomeD p1	6	FANCA VHL FANCE RECQL4 EPCAM RET
CNVkit p1	10	POLD1_S478N FH BAP1 NSD1 MLH1 MSH6 POLD1_L474P RUNX1 FANCL NF1
VarS p1	5	ERCC4 MEN1 WRN WT1 ALK

Tabla 32 Número de CNVs y genes de las tres herramientas bioinformática para el pool1

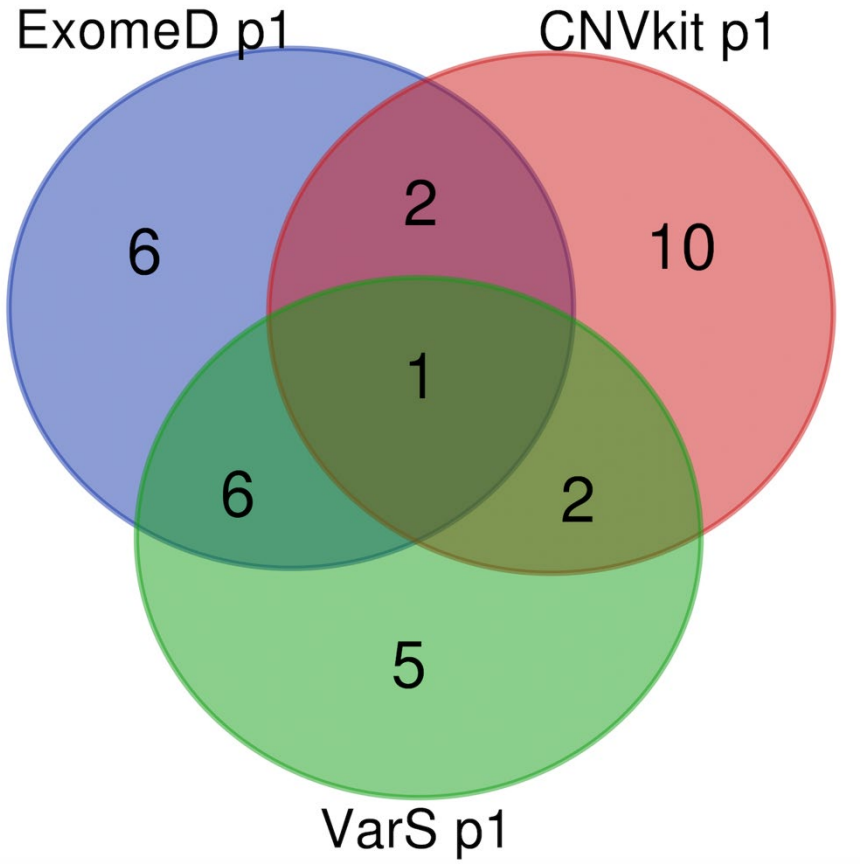


Figura 14 Diagrama de Venn para las tres herramientas bioinformáticas vs ICR96 datos validados por MLPAS para el pool1

Los acrónimos del diagrama de Ven son ;

ExomeD p1, representa los datos de las CNVs encontradas con la herramienta ExomeDepth para el pool1

CNVkit p1, representa los datos de las CNVs encontradas con la herramienta CNVkit para el pool1

VarS p1, representa los datos de las CNVs encontradas con la herramienta VarScan para el pool1

Como podemos ver en la figura 14, del diagrama de Venn, las tres herramienta bioinformática detectaron la misma CNV, CEBPA. Esta CNV se encuentra entre las CNVs validadas por MLPAs, aunque no se encontró validad en la muestra/ tipo de ninguna de ellas.

Cuando se observa las relaciones dos a dos podemos comprobar que CNVkit y Exomedepth detectaron dos CNVs; ERCC2 y BRCA2. Aunque solamente la herramienta CNVkit detecto esta CNVs en la posición muestra/ tipo validada MLPAS.

4.3.2 Datos obtenidos para el cruce entre las tres herramientas bioinformáticas para el pool2

Herramienta Bioinf.	Total	CNVs
CNVkit p2 ExomeD p2 VarS p2	8	CHEK2 STK11 EPCAM CEBPA MET ERCC2 PMS2 CDKN1C
CNVkit p2 ExomeD p2	3	KIT RUNX1 RECQL4
ExomeD p2 VarS p2	7	GATA2 MUTYH RET RB1 FANCA BRCA2 TSC2
CNVkit p2 VarS p2	17	FANCE SMARCB1 DIS3L2 NBN WT1 TMEM127 HRAS TP53 APC NF1 NSD1 MSH2 CYLD RHBDF2 SDHB MSH6 FANCM
ExomeD p2	3	BMPR1A XPA VHL
CNVkit p2	24	PRKAR1A EXT2 CDK4 CDKN1B POLD1_S478N BARD1 POLE_L424V BRIP1 POLD1_L474P HNF1A FANCL SMAD4 PPM1D RAD51C WRN EXT1 ERCC3 BLM PHOX2B FANCI EZH2 FANCF PMS1 FLCN
VarS p2	11	ERCC4 PRF1 BRCA1 MLH1 PTEN SLX4 MEN1 PTCH1 FANCD2 ALK ATM

Tabla 33 Número de CNVs y genes de las tres herramientas bioinformática para el pool1

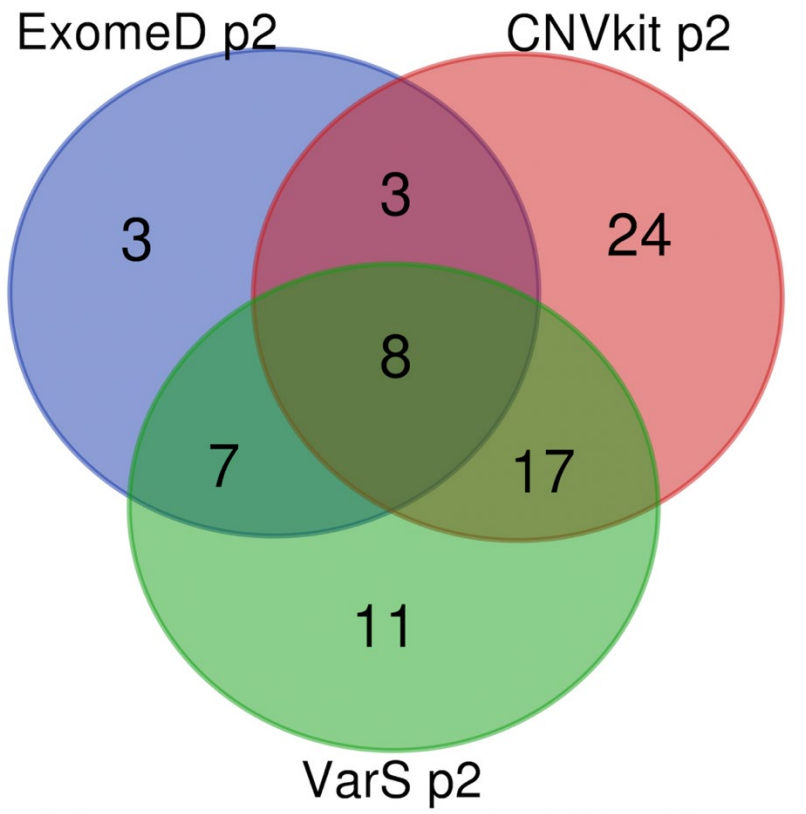


Figura 15 Diagrama de Venn para las tres herramientas bioinformáticas vs ICR96 datos validados por MLPAS para el pool1

Los acrónimos del diagrama de ven son ;

ExomeD p2, representa los datos de las CNVs encontradas con la herramienta ExomeDepth para el pool2

CNVkit p2, representa los datos de las CNVs encontradas con la herramienta CNVkit para el pool2

VarS p2, representa los datos de las CNVs encontradas con la herramienta VarScan para el pool2

Como podemos ver en la figura 15, del diagrama de Venn, las 3 herramientas bioinformáticas detectaron 8 CNVs comunes CHEK2 STK11 EPCAM CEBPA MET ERCC2 PMS2 CDKN1C. las CNVs CHEK2, PMS2 y EPCAM, son CNVs detectadas por CNVkit en Muestra/tipo validada por MLPAs, y CHEK2 es también detectada por ExomeDepth para muestra/tipo y validad por MLPAs. Cuando analizamos dos a dos las intersecciones entre herramientas las que más intersecciones detecto fue el cruce de CNVkit vs VarScan fundamentalmente porque detecto muchos falsos negativos.

5.Conclusiones generales.

En este TFM se ha implementado un pipeline in-house para la obtención y procesamiento de datos crudos desde una base de datos publica hasta la obtención de datos alineados en formato bam con una calidad óptima.

Estos datos alineados en formato bam han sido implementados en tres diferentes herramientas bioinformáticas basadas en profundidades de cobertura para la detección de CNVs .

Los rendimientos de las herramientas bioinformáticas para la detección de CNVs no han sido los esperados. Dando valores muy bajos de detección.

El análisis estadístico nos muestra que la herramienta bioinformática que mejores resultados obtuvimos fue CNVkit con 6 CNVs detectadas en el pool 1 de las 15 detectadas por el trabajo. Se obtuvieron unos valores de Sensibilidad de 0,4. Aunque estos valores siguen siendo valores muy bajos para una herramienta bioinformática. En cambio, para el pool2 se obtuvieron 9 CNVs de las 15 detectadas por el trabajo. Con una Sensibilidad 0,47. Siendo este valor mucho más bajo del esperado.

Otro parámetro que se puede analizar es que CNVkit detecto bastantes duplicaciones, sobre todo en el pool2. Con lo que es una herramienta que detecta bien delecciones y duplicaciones.

Los valores de especificidad fueron del orden de 0,71 para el pool 1 esto fue debido a que no se detectaron muchos FP . En cambio para el pool2 se obtuvo un valor de 0,10 esto puede ser debido a que hubo muchos FP, se detectaron muchas CNVs que no estaban validadas.

Para mejorar los resultados obtenidos, a nivel de CNVkit, se podría ajustar algunos parámetros como es el *-t (threshold*). Este parámetro está fijado en 0,2 e informa de las ganancias y pérdidas de copia única en una muestra de tumor completamente pura o CNV de línea germinal, este parámetro fue ajustado a 0,05 para detectar CNVs somáticas, pero sería interesante modificarlo y comprobar si mejoramos o no la Sensibilidad.

La literatura científica describe la herramienta bioinformática ExomeDepth como una herramienta con la que se obtiene buenos resultados a la hora de la detección de CNVs a nivel de Exón. En cambio, en el trabajo realizado, los resultados no son los esperados. Los valores que se obtuvieron una CNVs que coincidía con su muestra/tipo, es decir una CNV detectada por la herramienta de análisis y que coincide con la muestra, gen, posición y tipo de delección o amplificación con el trabajo validado por MLPAs, tanto para el pool1 y otra para el pool2. Los valores de especificidad están sobre 0,07 tanto para el pool 1 y como para el pool2.

Con el propósito de mejorar la detección de CNVs fueron realizadas varias aproximaciones. Una de ellas consistió en bajar el parámetro de “transition.probability = 10^4” a 10^3 para ser menos restrictivos y poder obtener más llamadas de CNVs, a pesar de que supone el riesgo de aumentar los falsos positivos. Sin embargo, se obtuvieron resultados muy semejantes. Otro parámetro que se tuvo en cuenta fue el hecho de incluir más muestras control para la normalización. En el trabajo de Plagnol, V. et al (15) se describe que con 10 muestras control se obtiene el número máximo de muestras para generar el control normalizado. Con el propósito de obtener una mejoría a la hora de generar la muestra control normalizada se incluyeron más muestras control, sin que los resultados fueran positivos. También se podría realizar el estudio con una sola muestra tumoral frente a la muestra control normalizada y comprobar los resultados. Ya que el análisis fue realizado automatizando la herramienta para la detección de todas las muestras tumorales en un solo paso.

En cuanto a los valores de especificidad estos fueron del orden de 0,33 para el pool1 y de 0,15 para el pool2, esto fue debido a que se obtuvo muchos FP, se detectaron muchas CNVs que no estaban validadas.

Por último, la Herramienta bioinformática VarScan mostro valores muy inferiores tanto para el pool1 como para el pool2. Se Obtuvo un valor de cero en cuanto a parametro de Sensibilidad . No se detectó ninguna CNV validad por la muestra/ tipo.

Los valores de especificad fueron para el pool1 del orden de 0,58. Para el pool 2 de 0,10

Estos valores tan malos pueden ser debidos a que VarScan trabaja con muestras pareadas, es decir que necesita generar un archivo mpileup entre muestra normal y muestra tumoral del mismo paciente, secuenciados juntos y en el mismo pool. Debido a la imposibilidad de tener muestras pareadas, se ha seleccionado una muestra control en cada pool como línea basal. El criterio en la elección de la muestra control se ha realizado analizando la cobertura media de la muestra y eligiendo la muestra con una cobertura media parecida a todas las muestras.

Este valor se ha obtenido mediante la herramienta SAMtools, utilizando el script;

samtools depth /~ muestra NORMAL.bam | awk '{sum+=$3} END { print "Average = ",sum/NR}'

Se ha intentado mejorar la Sensibilidad de la herramienta trabajando los parámetros de segmentación mínimo y máximo ya que han sido obtenidas muchas CNVs con fragmentos de delección muy grandes.

También, podríamos mejorar algunos parámetros a la hora de la segmentación. En el programa DNAcopy se podría cambiar el valor alpha=0.05 a un valor mayor para ser más estrictos y obtener menos regiones de análisis.

Por último, la utilización de Bedtools para seleccionar las regiones de captura solamente puede haber afectado a las reads eliminando o disminuyendo su número e interfiriendo en los resultados. También, se puede haber eliminado reads en la eliminación de los duplicados de PCR.

En cuanto a los datos obtenidos por los cruces de genes para las tres herramientas, nos revelo que en el pool 2 las tres herramientas detectaron 8 genes comunes, aunque solo 3 de ellos estaban validados muestra/tipo por el trabajo. Estos datos son muy bajos, con lo que se debería revisar las profundidades de lecturas de las muestras antes y después de haber realizado el Bedtools para analizar si hemos perdido cobertura en el número de lecturas y puede influir en los resultados tan bajos.

También debería revisarse el código y realizar varios puntos de control después de cada paso que hayamos hecho. Mediante IGV se puede comprobar que las CNVs validadas por el trabajo pueden detectadas antes de procesar las herramientas. Para intentar descubrir donde puede estar el error.

Los valores detectados por defecto de los detectores no han proporcionado unos valores satisfactorios en la detección de CNVs. Para futuros trabajos se debería probar otras herramientas bioinformáticas y realizar una validación de los parámetros de las diferentes herramientas para obtener mejores resultados. Además de evaluar cómo afecta en la detección de CNVs seleccionar solo las regiones de captura y no todo los reads alineados y por último, como afecta la eliminación de duplicados en las coberturas e eliminación de reads.

5.1. Objetivos planteados y planificados.

Respecto a los objetivos planteado en la PEC1, podemos afirmar que en términos generales se han cumplido todos los objetivos. Además, se han realizado otros que no estaban descritos en los objetivos ni en las tareas, pero por necesidad se han tenido que realizar. Hemos tenido que empezar el trabajo con la obtención de datos crudos en formato fastq y procesarlos hasta obtener los archivos bam. Esta tarea nos ha ocupado una parte importante de la planificación general

Otro punto que se ha cumplido, puesta a punto de las herramientas bioinformáticas, ha sido muy tedioso en algunos puntos, ya que al tener que procesar tal cantidad de muestras y tener que automatizarlo ha sido una tarea complicada y nos ha ocupado muchas horas y días en su puesta a punto.

El análisis estadístico se ha desarrollado en la forma y tiempo previsto. Aunque los resultados no han sido los esperados. Creo que deberíamos corregir algunos filtros y opciones de los programas para obtener unos resultados más precisos.

Por último, quería comentar que de las herramientas bioinformáticas que propusimos en el PEC1 dos de ellas no se han utilizado.

La primera Exomecopy, herramienta bioinformática desarrollado en R, nos dio un error en el punto de cálculo de la regresión y no fuimos capaces de encontrar el error y de obtener resultados.

La segunda herramienta bioinformática CONTRA que seleccionamos en el PEC1 como herramienta para su análisis, fue descartada pues después de dar los parámetros de entrada nos daba un error en las lecturas que no hemos sido capaces de subsanar.

Por ello cambiamos a la herramienta bioinformática CNVkit. Esta herramienta resulto cómoda y fácil en su instalación y los resultados han sido muy buenos.

6.Glosario

Flujo de trabajo: sistema informático que permite la automatización de procesos y agilizar los análisis a realizar.

Pipeline: procedimiento que transforma un flujo de datos en un proceso comprendido por varias fases secuenciales, siendo la entrada de cada una la salida de la anterior

Input: información (archivos, texto, datos....) de entrada que se incorpora en proceso determinado.

Output: información (archivos, texto, datos...) de salida que genera un proceso determinado.

Fasta: formato de fichero informático basado en texto, utilizado para representar secuencias biológicas como ácidos nucleicos o proteínas en la que cada nucleótido o aminoácido está representado por una letra. Cada secuencia está representado por un ID precedido por el símbolo “>”.

Bedfile: formato de fichero informático basado en texto utilizado en anotación genómica. Debe contener 3 campos obligatorios: nombre de cromosoma, posición de inicio del elemento y posición final del elemento.

SAM. Formato de texto tabulado que almacena secuencias de bases alineadas respecto al genoma.

BAM. Versión binaria de un fichero SAM.

CNV. Variación en el número de copias (copy number variation) consistente en la delección o inserción de una o varias regiones de al menos 50 pares de bases.

Coverage. Número de reads existentes para una nucleótido en la secuencia obtenida tras el alineamiento.

Heurística. Criterio que modifica la lógica de un algoritmo para dotarlo de mayor velocidad o para alcanzar una solución no óptima.

MLPA. Multiplex ligation-dependent probe amplification es una técnica basada en la reacción en cadena de la polimerasa (RCP), de cuantificación relativa del número de copias normales y anormales de ADN de hasta 40 secuencias genómicas diferentes [7].

NGS. Tecnología de secuenciación en paralelo que ha cambiado el paradigma de la secuenciación genómica, también conocida como high-throughput sequencing, massively parallel o deep sequencing.

7.Bibliografía

1. van Dijk EL, Auger H, Jaszczyszyn Y, Thermes C. Ten years of next-generation sequencing technology. Trends Genet [Internet]. 2014 Sep [cited 2018 Aug 28];30(9):418–26. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25108476

2. Carter NP. Methods and strategies for analyzing copy number variation using DNA microarrays. Nat Genet [Internet]. 2007 Jul [cited 2018 Sep 3];39(7s):S16–21. Available from: http://www.ncbi.nlm.nih.gov/pubmed/17597776

3. Shen Y, Wu B-L. Designing a simple multiplex ligation-dependent probe amplification (MLPA) assay for rapid detection of copy number variants in the genome. J Genet Genomics [Internet]. 2009 Apr [cited 2018 Sep 8];36(4):257–65. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19376486

4. Shyr D, Liu Q. Next generation sequencing in cancer research and clinical application [Internet]. 2013 [cited 2018 Sep 18]. Available from: http://www.biologicalproceduresonline.com/content/15/1/4

5. Zhao M, Wang QQ, Wang QQ, Jia P, Zhao Z. Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives - Springer. BMC Bioinformatics [Internet]. 2013 [cited 2018 Sep 3];14 Suppl 1(Suppl 11):S1. Available from: http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-14-S11-S1

6. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res [Internet]. Cold Spring Harbor Laboratory Press; 2012 Mar [cited 2018 Sep 28];22(3):568–76. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22300766

7. Plagnol V, Curtis J, Epstein M, Mok KY, Stebbings E, Grigoriadou S, et al. A robust model for read count data in exome sequencing experiments and implications for copy number variant calling. Bioinformatics [Internet]. Oxford University Press; 2012 Nov 1 [cited 2018 Sep 28];28(21):2747–54. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22942019

8. Talevich E, Shain AH, Botton T, Bastian BC. CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. PLOS Comput Biol [Internet]. Public Library of Science; 2016 Apr 21 [cited 2018 Nov

29];12(4):e1004873. Available from: https://dx.plos.org/10.1371/journal.pcbi.1004873

9. Lappalainen I, Almeida-King J, Kumanduri V, Senf A, Spalding JD, ur-Rehman S, et al. The European Genome-phenome Archive of human data consented for biomedical research. Nat Genet [Internet]. 2015 Jul 1 [cited 2018 Nov 19];47(7):692–5. Available from: http://www.nature.com/articles/ng.3312

10. Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data [Internet]. [cited 2018 Nov 10]. Available from: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

11. Mahamdallie S, Ruark E, Yost S, Ramsay E, Uddin I, Wylie H, et al. The ICR96 exon CNV validation series: a resource for orthogonal assessment of exon CNV calling in NGS data. Wellcome open Res [Internet]. The Wellcome Trust; 2017 [cited 2018 Oct 29];2:35. Available from: http://www.ncbi.nlm.nih.gov/pubmed/28630945

12. bedtools: a powerful toolset for genome arithmetic — bedtools 2.27.0 documentation [Internet]. [cited 2018 Dec 13]. Available from: https://bedtools.readthedocs.io/en/latest/index.html

13. Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform [Internet]. 2013 Mar 1 [cited 2018 Nov 29];14(2):178–92. Available from: https://academic.oup.com/bib/article-lookup/doi/10.1093/bib/bbs017

14. Package “ExomeDepth” Type Package Title Calls Copy Number Variants from Targeted Sequence Data [Internet]. 2016 [cited 2018 Dec 11]. Available from: https://cran.r-project.org/web/packages/ExomeDepth/ExomeDepth.pdf

15. Plagnol V, Curtis J, Epstein M, Mok KY, Stebbings E, Grigoriadou S, et al. A robust model for read count data in exome sequencing experiments and implications for copy number variant calling. Bioinformatics [Internet]. 2012 Nov 1 [cited 2018 Nov 29];28(21):2747–54. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/bts526

16. Mode EB, Mode EB. Elementos de probabilidad y estadística [Internet]. Reverté; 1990 [cited 2018 Dec 12]. 171 p. Available from: https://books.google.cl/books?id=5kPe6AkpOmIC&lpg=PA116&dq=Mode distribución binomial&pg=PA171#v=onepage&q=Mode distribución binomial&f=false

17. Agresti A. An Introduction to Categorical Data Analysis Second Edition [Internet].

[cited 2018 Dec 11]. Available from: https://mregresion.files.wordpress.com/2012/08/agresti-introduction-to-categorical-data.pdf

18. Talevich E, Shain AH, Botton T, Bastian BC. CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. PLOS Comput Biol [Internet]. 2016 Apr 21 [cited 2018 Dec 11];12(4):e1004873. Available from: https://dx.plos.org/10.1371/journal.pcbi.1004873

19. Maronna RA, Martin RD, Yohai VJ. Robust Statistics [Internet]. Chichester, UK: John Wiley & Sons, Ltd; 2006 [cited 2018 Dec 11]. (Wiley Series in Probability and Statistics). Available from: http://doi.wiley.com/10.1002/0470010940

20. Randal JA. A reinvestigation of robust scale estimation in finite samples. Comput Stat Data Anal [Internet]. North-Holland; 2008 Jul 15 [cited 2018 Dec 11];52(11):5014–21. Available from: https://www.sciencedirect.com/science/article/pii/S0167947308002272?via%3Dihub

21. Lax DA. Robust Estimators of Scale: Finite-Sample Performance in Long-Tailed Symmetric Distributions. J Am Stat Assoc [Internet]. 1985 Sep [cited 2018 Dec 11];80(391):736–41. Available from: http://www.tandfonline.com/doi/abs/10.1080/01621459.1985.10478177

22. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics [Internet]. 2009 Jun 1 [cited 2018 Dec 11];25(11):1422–3. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19304878

23. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, et al. VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res [Internet]. 2012 Mar 1 [cited 2018 Nov 11];22(3):568–76. Available from: http://genome.cshlp.org/cgi/doi/10.1101/gr.129684.111

24. Venkatraman ES, Olshen AB. A faster circular binary segmentation algorithm for the analysis of array CGH data. Bioinformatics [Internet]. 2007 Mar 15 [cited 2019 Jan 1];23(6):657–63. Available from: http://www.ncbi.nlm.nih.gov/pubmed/17234643

25. Talevich E, Shain AH, Botton T, Bastian BC. CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. PLOS Comput Biol [Internet]. Public Library of Science; 2016 Apr 21 [cited 2018 Dec 4];12(4):e1004873. Available from:

https://dx.plos.org/10.1371/journal.pcbi.1004873

26. Mahamdallie S, Ruark E, Yost S, Ramsay E, Uddin I, Wylie H, et al. The ICR96 exon CNV validation series: a resource for orthogonal assessment of exon CNV calling in NGS data. Wellcome open Res [Internet]. The Wellcome Trust; 2017 [cited 2018 Oct 27];2:35. Available from: http://www.ncbi.nlm.nih.gov/pubmed/28630945

27. Sensibilidad y especificidad (estadística) - Wikipedia, la enciclopedia libre [Internet]. [cited 2018 Dec 12]. Available from: https://es.wikipedia.org/wiki/Sensibilidad_y_especificidad_(estadística)

28. Baron ME. A Note on the Historical Development of Logic Diagrams: Leibniz, Euler and Venn [Internet]. Vol. 53, The Mathematical Gazette. 1969 [cited 2018 Dec 8]. p. 113. Available from: https://www.jstor.org/stable/3614533?origin=crossref

29. Guo Y, Sheng Q, Samuels DC, Lehmann B, Bauer JA, Pietenpol J, et al. Comparative study of exome copy number variation estimation tools using array comparative genomic hybridization as control. Biomed Res Int [Internet]. Hindawi Limited; 2013 [cited 2018 Oct 6];2013:915636. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24303503