Creación de un pipeline para el análisis de datos procedentes de secuenciación masiva (MiSeq) para su incorporación en un algoritmo predictivo basado en variantes genéticas

Pablo Elso Yoldi

Máster de Bioinformática y Bioestadística

Estadística y Bioinformática

Consultor externo: Javier Campión Zabalza

Consultor UOC: Ricardo Gonzalo Sanz

Profesores responsables de la asignatura: Alexandre Sánchez Pla, Antoni Pérez Navarro, Carles Ventura Royo, Jose Antonio Morán Moreno y Maria Jesús Marco Galindo

24 de mayo de 2017

FICHA DEL TRABAJO FINAL

Título del trabajo:	*Creación de un pipeline para el análisis de datos procedentes de secuenciación masiva (MiSeq) para su incorporación en un algoritmo predictivo basado en variantes genéticas*
Nombre del autor:	*Pablo Elso Yoldi*
Nombre del consultor/a:	*Externo: Javier Campión Zabalza* *UOC: Ricardo Gonzalo Sanz*
Nombre del PRA:	*Alexandre Sánchez Pla, Antoni Pérez Navarro, Carles Ventura Royo, Jose Antonio Morán Moreno y Maria Jesús Marco Galindo*
Fecha de entrega (mm/aaaa):	05/2017
Titulación::	*Máster de Bioinformática y Bioestadística*
Área del Trabajo Final:	*Estadística y Bioinformática*
Idioma del trabajo:	*Castellano*
Palabras clave	*Pipeline, NGS, SNPs.*

Resumen del Trabajo:

El proyecto se presenta dentro de los planes de investigación y desarrollo de la empresa Making Genetics S.L. de desarrollar un pipeline que emplea herramientas bioinformáticas para trabajar con datos obtenidos de la plataforma de secuenciación masiva MiSeq de Illumina para obtener el genotipo de unas variantes genéticas previamente seleccionadas por su importancia en el riesgo a padecer una enfermedad elegida por su relevancia social. Con la elaboración de un algoritmo predictivo de riesgo que tiene en cuenta los resultados obtenidos por el pipeline se obtiene un valor de probabilidad de sufrir la enfermedad en función de las variantes y se presentan los resultados en un informe genético de fácil comprensión.

Este proyecto busca poder dar un servicio de prevención de salud a los clientes de la empresa Making Genetics S.L. al tener como resultado final un valor de riesgo a padecer una enfermedad, de forma que se pueden tomar medidas preventivas con antelación a la aparición de los síntomas de la misma, pudiendo potencialmente reducir e incluso evitar la aparición de la enfermedad.

Abstract (in English, 250 words or less):

The project is included in the research and development plans of Making Genetics S.L. of developing a bioinformatic pipeline which uses different bioinformatic tools in order to work with data obtained from the Next Generation Sequencing platform MiSeq (from Illumina). The pipeline will obtain the genotype of the genetic variants previously selected because of its importance in the risk of suffering a disease chosen for its social relevance. With the creation of a predictive algorithm that uses the results from the bioinformatic pipeline, a probability of suffering the disease value in function of the genetic variants is obtained and those results are presented in a genetic report in order to be easily understood.

This project looks to be able to give a health prevention service to Making Genetics S.L.'s clients, as the final result is a risk value to suffer a disease, so the client can take preventive measures prior to the symptoms of that disease, being able of potentially reduce or even avoid the occurrence of said disease.

Índice

Lista de figuras

1. Introducción

1.1 Contexto y justificación del trabajo

1.1.1 Breve historia de la empresa

La empresa Making Genetics (www.making-genetics.eu) se constituyó en Navarra el 13 de abril de 2013 y está orientada al campo de la medicina personalizada. Constituida por un conjunto de biólogos con amplia experiencia investigadora en el campo de la Genómica y Epigenética. El objetivo principal de Making Genetics es identificar, desarrollar, implementar en clínica y licenciar nuevos paneles de diagnóstico y pronóstico combinando biomarcadores genéticos y epigenéticos. Para ello cuenta con la experiencia necesaria para el análisis de las muestras, para los estudios bioinformáticos que provengan de análisis masivos genómicos y epigenómicos, y para la integración bioestadística de biomarcadores de diferentes orígenes biológicos. Nuestra propuesta es convertirnos en una empresa que comercializa medicina personalizada mediante el desarrollo para su licencia de un paquete de determinaciones biológicas originales (combinando diferentes biomarcadores) que puedan atender al diagnóstico de diferentes enfermedades de forma flexible.

Actualmente se ha especializado en el desarrollo de nuevos paneles fármaco (EPI)genéticos de respuesta a enfermedades autoinmunes, en el que destaca el proyecto "Desarrollo de un algoritmo predictivo de mala respuesta a fármacos para artritis reumatoide basado en el promotor del gen VIP", financiado por el programa Retos Colaboración. La empresa ha sido financiada hasta el momento por Sodena, Enisa, Gobierno de Navarra y Ministerio de Economía y Competitividad (Torres Quevedo y Retos). En noviembre de 2016 Making Genetics acaba de conseguir una ayuda correspondiente a la convocatoria para proyectos de I+D del Gobierno de Navarra con el título Desarrollo experimental de un Test Epigenético para determinar la predisposición a desarrollar enfermedad celíaca en sujetos pediátricos genéticamente predispuestos en colaboración entre Making Genetics S.L. y la Unidad de Gastroenterología Pediátrica del Complejo Hospitalario de Navarra a través de NavarraBiomed. El objetivo principal del proyecto es la identificación y validación de nuevos biomarcadores de diagnóstico de enfermedad celíaca que permitan discriminar de entre todos los pacientes pediátricos genéticamente predispuestos a aquéllos que van a desarrollar la enfermedad de los que no. Tras este primer objetivo para el que se solicita financiación los siguientes objetivos generales son la validación clínica en diferentes centros hospitalarios como técnica, la implementación clínica de estos biomarcadores y su licencia a un tercero para su comercialización. Además de estos proyectos, la empresa mantiene una amplia actividad de I+D, en la que destaca actualmente la adaptación de técnicas de secuenciación masiva para el análisis de variantes genéticas en diferentes poblaciones, y una cartera de servicios para investigadores y empresas.

1.1.2 Contexto tecnológico

La nueva generación de plataformas de secuenciación masiva o 'Next-Generation Sequencing' (NGS), inician su actividad comercial en el año 2005 generando una auténtica revolución en la investigación biológica. Durante la última década, la vertiginosa evolución de estos nuevos secuenciadores, tanto en precisión como en rendimiento, así como el abaratamiento del coste por base han permitido la rápida expansión de su uso en la comunidad científica ofreciendo nuevas alternativas para la secuenciación de genomas completos [1,2,3], resecuenciación dirigida de zonas concretas del genoma [4,5,6,7], secuenciación de transcriptomas completo (RNA-Seq) [8], identificación de microRNAs [9], estudios de interacción proteína-DNA (ChIP-seq) [10] o estudios de metilación entre otros [11].

Además, se han realizado enormes progresos en términos de velocidad, longitud de lectura y rendimiento, por lo que se ha desarrollado un gran número de nuevas aplicaciones de NGS en ciencias básicas, así como en las áreas de investigación traslacional como los diagnósticos clínicos, agrogenómica y ciencias forenses [12]. Estas plataformas de NGS permiten la obtención de miles o millones de fragmentos de ADN en un único proceso, haciendo posible el desarrollo de proyectos de secuenciación en corto plazo. A pesar de las diferencias en su química, todas estas plataformas de secuenciación masiva comparten las siguientes características:

- El ADN se fragmenta al azar y se unen adaptadores específicos a ambos lados de cada molécula directamente sin necesidad de clonar.
- La amplificación de la librería se produce mediante el anclaje del fragmento de ADN, a través de sus adaptadores, a una superficie sólida como son las microesferas o directamente a la placa de secuenciación.
- La secuenciación y detección de las bases ocurren al mismo tiempo en todas las moléculas de ADN (secuenciación masiva y paralela).
- Las lecturas generadas son cortas. El ruido producido por estas tecnologías en relación con la señal que generan limita la obtención de lecturas de mayor longitud.
- Las plataformas NGS permiten realizar secuenciación de tipo "paired-end" mediante la cual es posible leer los extremos del mismo fragmento de ADN. Esta estrategia de secuenciación facilita no solo el posicionamiento de aquellas lecturas que pueden mapear en múltiples sitios sino que también posibilita la identificación de variantes estructurales.

1.1.3 El diagnóstico genético mediante NGS: resecuenciación dirigida.

La detección de variantes genéticas a partir de datos de NGS consiste en identificar diferencias en la secuencia de ADN de un individuo al compararlo con un ADN de referencia. Los resultados dependen forzosamente de la calidad del alineamiento y ensamblaje respecto a la referencia ya que las secuencias alineadas incorrectamente pueden producir falsos positivos, mientras que las secuencias no alineadas pueden ser fuente de falsos negativos. La NGS tiene el potencial de detectar cualquier tipo de variante genómica en un único experimento, incluso puede detectar inversiones, una

clase de variación cuyo estudio resulta muy complicado para la mayoría de las otras técnicas.

Las enfermedades de interés en este Proyecto van a ser enfermedades complejas, es decir, no muestran herencia mendeliana atribuible a un único locus (una región específica en un cromosoma). Frente a las enfermedades monogénicas, las enfermedades complejas están causadas por la acción de múltiples loci, cada uno con un pequeño efecto, además estos loci pueden interaccionar entre ellos y a su vez interaccionar con factores ambientales. Los estudios de asociación genética tienen como objetivo identificar factores de susceptibilidad a una determinada enfermedad. En los estudios de asociación poblacionales los polimorfismos más habituales objeto de estudio son los polimorfismos de un solo nucleótido (en inglés, SNP - Single Nucleotide Polymorphism). Un SNP describe un cambio en una sola base, una de las bases es sustituida por otra. Para que verdaderamente pueda considerarse un polimorfismo, la variación debe aparecer al menos en el 1% de la población.

En este proyecto nos vamos a centrar en dos tipos de variantes genéticas:

- Polimorfismo de nucleótido único

La detección de variantes de nucleótido único (single nucleotide polymorphism o SNP) ha demostrado ser factible con una gran precisión cuando hay al menos una cobertura de 10-15 veces para la posición de la SNP y la tasa de error de secuenciación es razonable [9,10]. La mayoría de los algoritmos informáticos utilizados para detectar SNP emplean modelos bayesianos, calculando la probabilidad condicional de los nucleótidos en cada posición según, por ejemplo, el número de reads independientes que contienen la variante, la calidad en la asignación de la base y otros parámetros.

- Inserciones y deleciones pequeñas

La detección de pequeñas inserciones y deleciones (indels) a partir de datos de NGS ha demostrado ser más compleja de lo que inicialmente se podía prever, sobre todo por culpa de la limitada longitud de los reads que producen la mayoría de las plataformas. Las variantes de ganancia o pérdida de una única base son especialmente proclives a ser mal alineadas con el genoma de referencia, produciendo una elevada tasa de falsos positivos. Un alineamiento de novo regional, que requiere cálculos computacionales elevados, contribuye a mejorar la detección de indels aunque los niveles de sensibilidad y especificidad no logran acercarse a los de la detección de SNP [7,13,14].

1.1.4 Descripción general

El TFM que se presenta en este proyecto consiste en la creación de un pipeline que permita obtener las variantes genéticas que presenta un individuo para unos polimorfismos determinados utilizando datos de secuenciación masiva. Se va a seleccionar la enfermedad de acuerdo con los intereses de la empresa, realizando la búsqueda bibliográfica de los polimorfismos asociados a la predisposición genética a

sufrir la enfermedad seleccionada y extrayendo a partir de los datos brutos de secuenciación masiva utilizados por la plataforma MiSeq las variantes genéticas que presenta la muestra de ADN analizada. Finalmente se diseña un algoritmo que presente la predisposición a sufrir la enfermedad en cuestión para un individuo determinado.

1.1.5 Justificación del trabajo

Este TFM se desarrolla dentro de los planes de investigación y desarrollo de la empresa Making Genetics SL. La empresa necesita por motivos estratégicos implementar un servicio de análisis de polimorfismos utilizando secuenciación masiva. El estudiante se integrará en el equipo de trabajo que lo está desarrollando y se encargará de las tareas relacionadas con el apartado bioinformático.

1.2 Objetivos del trabajo

1. Identificar una enfermedad poligénica y multifactorial humana y los polimorfismos más relevantes asociados a su predisposición a desarrollarla.
    a. Seleccionar la enfermedad a estudiar.
    b. Identificar los polimorfismos más relevantes asociados a dicha enfermedad.
    c. Crear una base de datos que incluya los datos más relevantes de los polimorfismos identificados.
    d. Crear un algoritmo predictivo.
2. Desarrollar un pipeline bioinformático que implemente la identificación de las variantes genéticas en una muestra determinada.
    a. Identificar las tareas bioinformáticas y aplicaciones necesarias para la implementación del pipeline.
    b. Instalar y optimizar los paquetes necesarios para la identificación de los polimorfismos.
    c. Testar el pipeline con datos brutos obtenidos de las lecturas en MiSeq de diferentes individuos.
    d. Crear un informe genético que presente la predisposición a sufrir la enfermedad determinada.

1.3 Enfoque y método seguido

En concreto en el presente proyecto se va a utilizar una de las plataformas Illumina, secuenciador MiSeq. Esta plataforma se basa en principios químicos como la incorporación de nucleótidos marcados con terminadores reversibles de manera que en cada ciclo de ligación solamente uno de los cuatro nucleótidos posibles se une de forma complementaria al ADN molde emitiendo una señal luminosa que es captada por un sistema óptico altamente sensible. Posteriormente, el terminador se elimina para permitir la incorporación del siguiente nucleótido en ciclos sucesivos de secuenciación. La química empleada por Illumina permite generar lecturas de hasta 2x300nt llegando a producir hasta 15 Gb en datos. La figura 1 muestra el flujo de

trabajo general del procesamiento de las muestras hasta la obtención de los resultados.

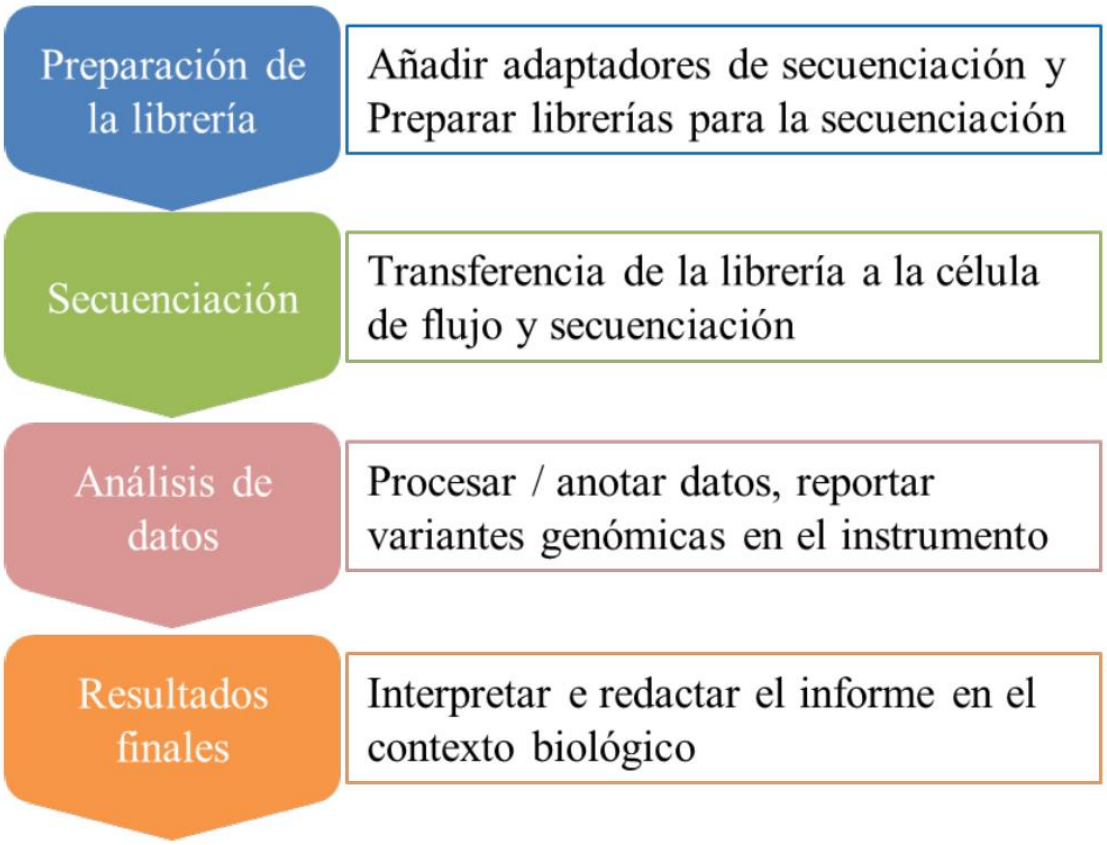

***Figura 1**. Diagrama de flujo general de MiSeq.*

Por otro lado, la dificultad de reunir un número mínimo de muestras que justifiquen económicamente la puesta en marcha de un proceso completo unido al alto coste de adquisición de grandes equipos impulsaron el desarrollo de esta nueva serie de secuenciadores más económicos, capaces de generar datos con la misma precisión que las plataformas grandes de secuenciación masiva. Estas plataformas se han extendido rápidamente dotando a pequeños laboratorios de la tecnología de secuenciación más avanzada. La aplicación de paneles de secuenciación masiva al desarrollo de test genético de enfermedades poligénicas requiere de un proceso de diseño biomédico de los genes asociados a la patología, del diseño bioinformático del panel y de la validación bioinformática y del panel genético que permita evaluar los parámetros de calidad del panel como la reproducibilidad, cobertura media, sensibilidad, especificidad, detección de deleciones e indels, confirmación de variantes [15,16]. En la figura 2 se encuentra el modelo de diseño y validación del panel para enfermedades poligénicas.

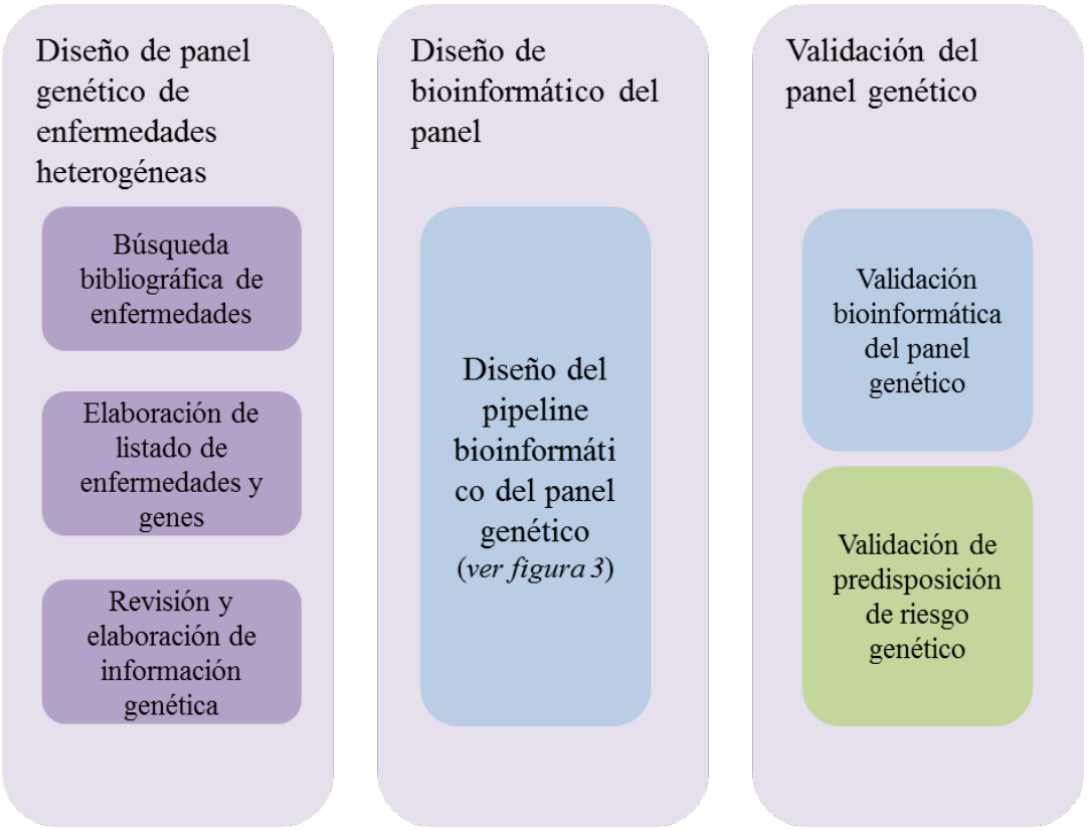

***Figura 2**. Modelo de diseño y validación del panel de enfermedades heterogéneas.*

La resecuenciación dirigida de un pequeño número de genes implica una reducción significativa en la cantidad de lecturas necesarias para la identificación de las variantes presentes con el consecuente la disminución en el coste por muestra. La llegada de los mini-secuenciadores con tecnología NGS (ej. MiSeq de Illumina) así como la disminución de los costes de secuenciación por base, forman la combinación ideal para la adopción de esta tecnología como método de preferencia en los test genéticos de rutina en el presente y futuro cercano. La descripción de enfermedades genéticas poligénicas se encuentran en todas las especialidades médicas y están asociadas a mutaciones en múltiples genes, lo que ha puesto sobre la mesa la dificultad tecnológica de estudiar de forma eficiente, todos los genes implicados en una patología. A continuación, se muestra la figura 3 con la descripción detallada sobre el procedimiento bioinformático a seguir para el análisis de la resecuenciación dirigida hasta la elaboración del informe final.

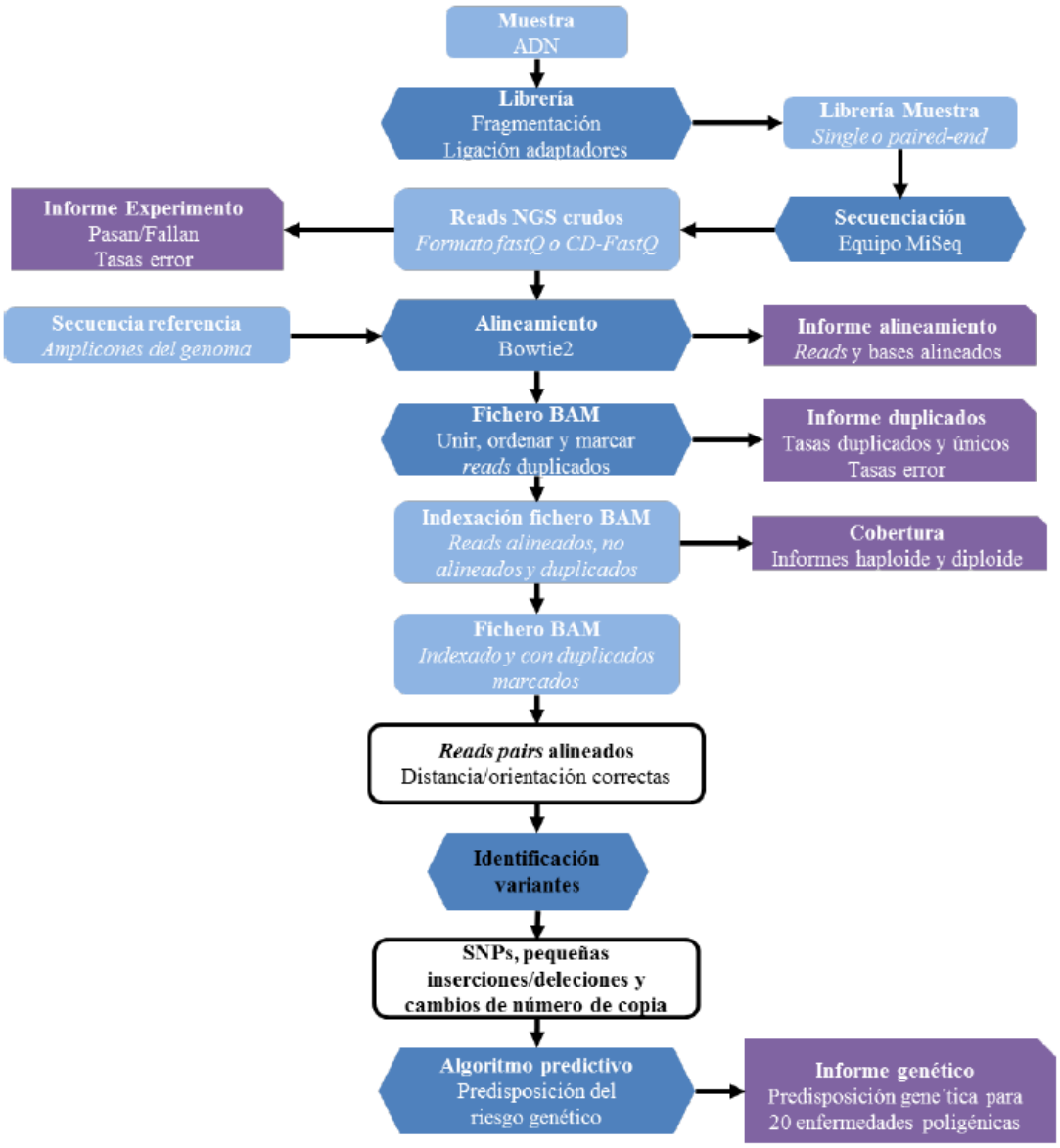

***Figura 3**. Diagrama de flujo del procesamiento bioinformático.*

1.4 Planificación del trabajo

En la figura 4 se puede ver el calendario planificado para realizar el trabajo, organizado por objetivos. Esta organización se realiza teniendo en cuenta la complejidad y necesidad de tiempo de cada uno de los apartados. Sin embargo, no es una planificación cerrada y se utiliza como un esquema orientativo, de forma que cuando se necesitó trabajar en un objetivo no planeado para esa fecha, se hizo sin afectar de forma importante al flujo de trabajo.

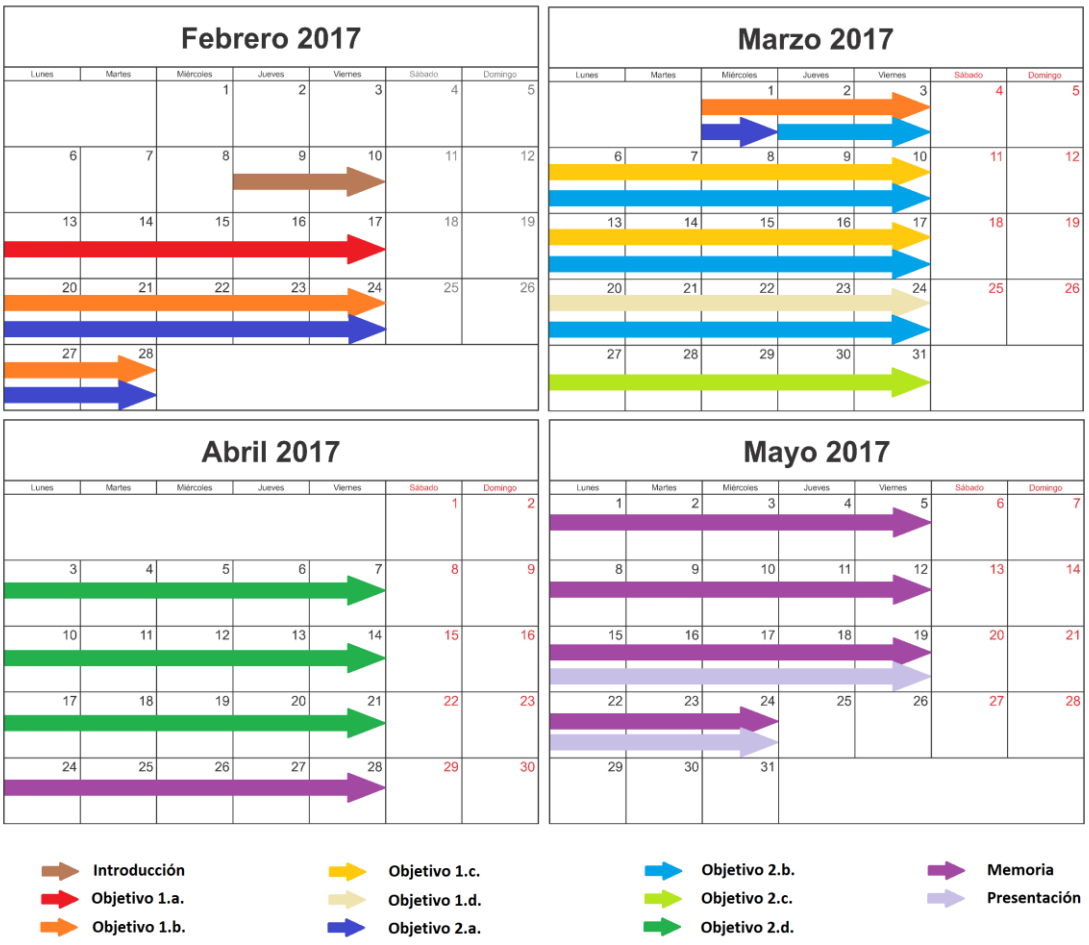

***Figura 4**. Calendario de planificación del trabajo.*

El tiempo dedicado a cada objetivo se asignó dando un margen para posibles inconvenientes para el estudiante como enfermedad, viajes, etc. De esta forma, los imprevistos no provocan retrasos ni inconsistencias en el flujo de trabajo.

1.5 Breve sumario de los productos obtenidos

1. Se obtiene un script que permite realizar el pipeline bioinformático para el análisis de detección de variantes en archivos de formato FASTQ obtenidos de secuenciación masiva. Este pipeline emplea diferentes herramientas de forma que se obtiene como producto final el genotipo de las variantes deseadas.
2. Algoritmo predictivo que utiliza información de la bibliografía y de los resultados del pipeline de análisis de detección de variantes. Este algoritmo proporciona un valor de riesgo a padecer la enfermedad estudiada.
3. Informe genético generado para entregárselo al cliente de forma que se puedan entender de forma clara y sencilla los resultados del algoritmo predictivo.

1.6 Breve descripción de los otros capítulos de la memoria

1. Identificación de la enfermedad y sus polimorfismos: En este apartado se selecciona la enfermedad que va a ser estudiada y de la cual se tratará de hacer una predicción del riesgo a padecerla. Para ello, se seleccionan los polimorfismos más adecuados y se crea una base de datos con toda la información necesaria referente a los mismos.
2. Pipeline: Se selecciona el programa más adecuado para la elaboración del pipeline para el análisis de detección de variantes y se describe el script que se elabora, mostrando que se hace en cada paso y el uso que tiene cada herramienta en el proceso.
3. Algoritmo predictivo: Se elabora el algoritmo predictivo para obtener un valor de riesgo a padecer la enfermedad seleccionada en función del genotipo obtenido en el análisis de detección de variantes.
4. Informe genético: Se crea un informe para que el cliente vea de forma clara y sencilla los resultados de su análisis.
5. Conclusiones: Se presentan las conclusiones y reflexiones obtenidas del trabajo realizado.
6. Glosario: Definición de los términos y acrónimos más relevantes utilizados dentro de la memoria.
7. Bibliografía: Citación de todos los artículos, libros y páginas web empleadas a lo largo de todo el proyecto.
8. Anexos: Inclusión del script completo del pipeline para el análisis de detección de variantes.

2. Identificación de la enfermedad y sus polimorfismos

2.1 Selección de la enfermedad

Para la selección de la enfermedad con la que se va a trabajar hace falta llevar a cabo varios filtros iniciales que hagan un barrido y nos dejen con una pequeña lista de enfermedades que cumplen los requisitos necesarios para ser consideradas apropiadas para el estudio.

Inicialmente y debido a la naturaleza del estudio, la enfermedad ha de poseer un carácter poligénico y multivariante, es decir, que se produce debido a la combinación de diferentes factores como mutaciones en diferentes genes, estilo de vida, etc. Este tipo de enfermedades tienen una mayor complejidad en su componente genético en comparación con las enfermedades monogénicas (también llamadas de herencia mendeliana) en las que las mutaciones de un único gen pueden provocarla. El número de enfermedades poligénicas es elevado y cada vez se sabe más sobre ellas gracias al avance de las tecnologías implicadas en el estudio de la genética al permitir estudiar con mayor rapidez y menor coste los polimorfismos implicados en el desarrollo de la enfermedad [17,18].

Otro elemento importante en la decisión de la enfermedad a estudiar es el impacto que tiene la misma en la sociedad. Al tener como objetivo final el desarrollo de un servicio accesible a la población de un análisis de riesgo a sufrir la enfermedad, ésta debe tener un impacto elevado en la sociedad de forma que haya un alto número de potenciales clientes que quieren saber su riesgo. Por lo tanto, quedarían descartadas las enfermedades raras o de una prevalencia e incidencia bajas, y como siguente objetivo se va a desarrollar un criterio que permita valorar el impacto social de la enfermedad.

Como primer elemento para dicho criterio se elige la lista de defunciones por causas de muerte más frecuentes. Es de gran interés poder anticipar el riesgo a padecer una de las enfermedades que más muertes causa en España presentes en la tabla 1 para actuar con tiempo y buscar soluciones que reduzcan significativamente dicho riesgo en un elevado número de personas. [19]

Causa de muerte	Número
Todas las causas	422568
Enfermedades del sistema circulatorio	124197
Tumores	111381
Enfermedades del sistema respiratorio	51848
Enfermedades cerebrovasculares	28434
Enfermedades del sistema nervioso y de los órganos de los sentidos	25835
Otras enfermedades del corazón	23043
Otras enfermedades del sistema respiratorio	22159
Tumor maligno de la tráquea, de los bronquios y del pulmón	21625
Trastornos mentales y del comportamiento	21333

Tabla 1. Defunciones según causa de muerte en España (2015).

No solo preocupan las enfermedades que causan muchas muertes, también es relevante para el estudio conocer las enfermedades que causan un gran número de altas hospitalarias, puesto que aunque no deriven en muerte tienen mucho impacto en la calidad de vida de los pacientes. Por ello, conocer la cantidad de altas hospitalarias en función del diagnóstico del paciente presentes en la tabla 2 y cuales de ellas son las más elevadas se presenta como un criterio de selección apropiado. [20]

Causa de alta hospitalaria	Número
Todas las causas	4746651
Enfermedades del sistema circulatorio	628563
Enfermedades del aparato digestivo	572778
Enfermedades del aparato respiratorio	572587
Complicaciones del embarazo, parto y puerperio	480528
Neoplasias	454891
Lesiones y envenenamientos	424014
Neoplasias malignas	357368
Enfermedades del sistema osteo-mioarticular y tejido conectivo	343501
Enfermedades del aparato genitourinario	318251

Tabla 2. Altas hospitalarias según diagnóstico en España (2015).

Relacionada con lo anterior también aparece la prevalencia de los principales problemas de salud en población de 15 años o más presentes en la tabla 3, y es un criterio más que se añade para la elección de la enfermedad [21].

Problema de salud	Ambos sexos	Hombres	Mujeres
	Tasa----- Orden	Tasa------ Orden	Tasa----- Orden
Hipertensión arterial	182,9-------- 1	171,7--------- 1	193,6-------- 1
Trastornos del metabolísmo lipídico	174,4-------- 2	166,9--------- 2	181,6-------- 2
Infección respiratoria aguda del tracto superior	159,0-------- 3	138,4--------- 3	178,5-------- 3
Enfermedades de los dientes/encías	103,3-------- 4	98,8---------- 4	107,6-------- 6
Síndromes de columna vertebral	96,1--------- 5	77,1---------- 6	114,1-------- 4
Trastornos de la ansiedad/estado de ansiedad	81,5--------- 6	52,8---------- 9	108,8-------- 5
Artrosis	76,7--------- 7	48,2--------- 14	103,9-------- 7
Bursitis/tendinitis/sinovitis no especificadas	74,6--------- 8	63,0---------- 8	85,7---------- 9
Diabetes mellitus	71,4--------- 9	77,6---------- 5	65,5-------- 13
Otras enfermedades del aparato locomotor	66,3-------- 10	51,4--------- 10	80,5-------- 10
Obesidad y sobrepeso	60,8-------- 11	49,7--------- 11	71,3-------- 11
Otras enfermedades de la piel	60,4-------- 12	50,3--------- 11	70,1-------- 12
Abuso del tabaco	57,3-------- 13	67,7---------- 7	47,3-------- 15
Otras enfermedades generales no especificadas	57,2-------- 14	49,0--------- 12	64,9-------- 14
Cistitis/otras infecciones urinarias	55,6-------- 15	18,6--------- 15	90,7---------- 8
Medicina preventiva/promoción de la salud	153,5	121,1	184,3

Tabla 3. prevalencia de los principales problemas de salud en población de 15 años o más.

Dado que se va a dar al cliente un valor de riesgo a padecer la enfermedad en función de sus variantes genéticas, es importante poder asesorarle sobre las posibles medidas que puede tomar al respecto en caso de un riesgo elevado. Por lo tanto se valora para la elección la existencia de unas pautas que colaboren a reducir el riesgo a padecer la enfermedad mediante hábitos de vida saludables o posibles cambios en el estilo de vida [22].

Otro criterio de importancia es el gasto sanitario que crean las diferentes enfermedades. Está muy relacionado con las causas de alta hospitalaria e interesa tanto a los pacientes como al sistema sanitario encontrar la manera de predecir el riesgo a padecer las enfermedades que mayor gasto general para poder tomar medidas preventivas que potencialmente mejore el estado de salud de las personas con predisposición, lo que se traduce en una reducción del gasto sanitario [23].

Directamente relacionada con el gasto sanitario se encuentra la estancia media en días y su clasificación según el diagnóstico principal. La tabla 4 reúne las principales enfermedades que mayor número de días (de media) mantienen en estancia hospitalaria a los pacientes [24]:

Causa de alta hospitalaria	Días
Demencia senil, presenil y vascular	57,41
Trastornos esquizofrénicos	37,36
Trastornos neuróticos	30,73
Trastornos mentales	26,28
Psicosis orgánicas y trastornos mentales debidos a drogas	24,45
Otras psicosis	18,71
Tuberculosis	17,77
Neoplasias malignas de tejidos hematopoyéticos	17,53
Crecimiento intrauterino retardado, desnutrición fetal, gestación acortada y bajo peso	17,1
Fiebre reumática aguda	14,93

Tabla 4. Estancia media hospitalaria según diagnóstico en España (2015).

2.2 Criterio complementario de elección

Un criterio complementario de selección de la enfermedad tiene que ver con la calidad y la cantidad de estudios conocidos sobre la enfermedad y las variantes genéticas asociadas. Se valora la calidad metodológica de los estudios incluidos según los criterios de *Scottish Intercollegiate Guidelines Network* (SIGN), una red de directrices presentes en la tabla 5 entre varios colegios profesionales que desarrolla guías de práctica clínica basada en la evidencia para el Servicio Nacional de Salud en Escocia.

Nivel de evidencia	Descripción
1++	Meta-análisis de alta calidad, RS de EC o EC de alta calidad con muy poco riesgo de sesgo
1+	Meta-análisis bien realizados, RS de EC o EC bien realizados con poco riesgo de sesgos
1-	Meta-análisis, RS de EC o EC con alto riesgo de sesgos
2++	RS de alta calidad de estudios de cohortes o de casos y controles. Estudios de cohortes o de casos y controles con riesgo muy bajo de sesgo y con alta probabilidad de establecer una relación causal
2+	Estudios de cohortes o de casos y controles bien realizados con bajo riesgo de sesgo y con una moderada probabilidad de establecer una relación causal
2-	Estudios de cohortes o de casos y controles con alto riesgo de sesgo y riesgo significativo de que la relación no sea causal
3	Estudios no analíticos, como informes de casos, series de casos o estudios descriptivos
4	Opinión de expertos

Tabla 5. Niveles de evidencia *Scottish Intercollegiate Guidelines Network* para la selección de artículos con asociación gen-enfermedad.

Conociendo los criterios de SIGN se puede establecer como objetivo óptimo en este apartado la búsqueda de meta-análisis de alta calidad de algunas de las enfermedades. Este criterio es complemetario, y se aplica en caso de duda de selección de enfermedades puesto que la presencia de un meta-análisis para una enfermedad va a ayudar de gran manera a la identificación de polimorfismos en el presente estudio.

Tras barajar los criterios mencionados y teniendo en cuenta criterios estratégicos de la empresa, se selecciona la enfermedad hipertensión por su alto impacto en la sociedad y la posibilidad de ser prevenida mediante la detección de polimorfismos implicados en el riesgo de padecerla. Además, existen pautas de recomendación para reducir las probabilidades de sufrir esta enfermedad [25].

2.3 Identificación de polimorfismos y creación de base de datos

Una vez identificada la enfermedad con la que se va a trabajar en el proyecto, es necesario conocer los polimorfismos asociados a ella y de que manera afectan al desarrollo de la misma. Diferentes polimorfismos pueden influir en la enfermedad en diferentes elementos de la misma y en diferente grado de severidad (algunos polimorfismos son más críticos que otros a pesar de no poseer la enfermedad un carácter monogénico).

Tras una extensa busqueda bibliográfica en la base de datos Pubmed, que incluye los términos "Hypertension", "GWAS", "polymorphism" y "European descendent", se

obtiene una selección de artículos que contiene una gran cantidad de información sobre las variantes implicadas en la enfermedad permite la creación de una base de datos con una gran cantidad de variantes para el presente estudio. Se han incluido en la base de datos todos los SNPs con un p value significativo y menor de 10^{-5}. Se han seleccionado 21 SNPs en 19 genes diferentes presentes en la tabla 6 [26,27,28,29].

Gen	SNP
MTHFR	rsXXXXXXXX
SH2B3	rsXXXXXXXX
CYP1A1	rsXXXXXXXX
EBF1	rsXXXXXXXX
FGF5	rsXXXXXXXX
NT5C2	rsXXXXXXXX
NPR3	rsXXXXXXXX
JAG1	rsXXXXXXXX
MECOM	rsXXXXXXXX
ZNF652	rsXXXXXXXX
CACNB2	rsXXXXXXXX
TBX5	rsXXXXXXXX
GUCY1A3	rsXXXXXXXX
SLC4A7	rsXXXXXXXX
ATP2B1	rsXXXXXXXX
ATP2B1	rsXXXXXXXX
ADM	rsXXXXXXXX
GNAS	rsXXXXXXXX
BAT2	rsXXXXXXXX
MOV10	rsXXXXXXXX
ATP2B1	rsXXXXXXXX

Tabla 6. Lista de SNPs preseleccionados y los genes a los que pertenecen. Los nombres de los SNPs no se incluyen por ser confidenciales.

A partir de esta primera base de datos inicial, se ha de reducir el número de polimorfismos que van a formar parte del análisis para reducir el coste total sin renunciar a que los resultados obtenidos sean significativos (puesto que reducir de forma excesiva el número de polimorfismos estudiados implicaría unos resultados poco descriptivos del riesgo). Para ello se van a estudiar diferentes características de las variantes que interesan para el presente estudio y se van a plasmar en una base de datos que además de ayudar a seleccionar los polimorfismos óptimos va servir de utilidad en el desarrollo y como fuente de información del algoritmo predictivo. Esta base de datos ha de contener toda la información necesaria para el algoritmo

predictivo, a partir del cual se obtendrá un valor de riesgo que se verá plasmado finalmente en el informe genético. Esta base de datos es, por lo tanto, un elemento central del proyecto y en ella ha de aparecer toda la información relevante de las variantes.

La tabla 7 reúne todos los criterios que se han tenido en cuenta para seleccionar los polimorfismos más relevantes en el estudio de la enfermedad y que van a permitir que el análisis desarrollado en el presente proyecto tenga validez a pesar de la reducción del número de polimorfismos estudiados. Al aplicar los diferentes criterios no se seleccionan aquellos SNPs que reciben penalizaciones puesto que eso significa que no son tan relevantes en el estudio como aquellos que pasan todos los criterios.

Característica	Descripción
Calidad del artículo	Valoración del artículo basado en grados de recomendación del SIGN (Scottish Intercollegiate Guidelines Network [1]).
Número de artículos	Referencia al número de apariciones del SNP en los diferentes artículos incluidos en la base de datos.
Número de fenotipos	Referencia al número de apariciones del SNP en los diferentes subfenotipos incluidos en la base de datos.
dbSNP	Variable categórica que hace referencia a la presencia o ausencia del SNP en la base de datos NCBI dbSNP Build 146 (fecha de acceso: Febrero 2016; http://www.ncbi.nlm.nih.gov/SNP/).
Tipo SNP	Clasificación del tipo de variante genética donde al tipo de SNV (variación de único nucleótido, del inglés) se le da una clasificación positiva y cualquier otro tipo de variación (Inserción/deleción, STR, CNV,..) se penaliza de forma que no pasa el criterio ya que no se puede detectar por el sistema de genotipado seleccionado para realizar la determinación. La combinación alélica ha sido obtenida a través del servicio web de SPSmart: v5.1.1 y versión dbSNP: build 132 para población CEU (http://spsmart.cesga.es/), donde el primer lugar se indica el alelo mayoritario seguido de una barra lateral y el alelo minoritario (ej. A/G).
Detección	Sistema de puntuación basado en la clasificación del tipo de combinación de variante alélicas para el polimorfismo bialélico. Penalizando las combinaciones de la misma base nitrogenadas pirimidínicas o púricas eliminándolas de la selección, en caso contrario pasa el criterio.
Frecuencia del alelo menor	Frecuencia alélica del alelo menor. Este valor es obtenido a través del servicio web de SPSmart: v5.1.1·y versión dbSNP: build 132 para población CEU (http://spsmart.cesga.es/).
Desequilibrio de ligamiento	Sistema de puntuación basado en la presencia del desequilibrio de ligamiento (LD, de las siglas del inglés) con otro polimorfismo de la base de datos para el presente fenotipo. Se penaliza la variable con menor valor en las columnas previas. El análisis de LD se realiza a través del servicio web SNAP(http://www.broadinstitute.org/mpg/snap/ldsearchpw.php) SNP Annotation and Proxy Search Versión 2.2.

Tabla 7. Criterios de selección de polimorfismos empleado.

Una vez se aplican los criterios de selección, se reúne un total de 6 SNPs localizados en 6 genes diferentes. Los genes estudiados y sus SNPs son:

- MTHFR: gen que codifica la enzima "methylenetetrahydrofolate reductase" implicada en el procesamiento de aminoácidos.

- SH2B3: gen que codifica una proteína de la familia de adaptadores SH2B, implicada en actividades de señalización por factores de crecimiento y receptores de citoquinas.
- CYP1A1: gen que codifica una enzima del citocromo P450.
- PLEKHA7: gen implicado en la estabilización y expansión de cadherinas-E (las cuales juegan un papel fundamental en la asociación entre células).
- EBF1: gen que codifica una enzima de gran importancia en la diferenciación a linfocitos B.
- FGF5: "Fibroblast Growth Factor 5" implicado en gran variedad de procesos biológicos como desarrollo embrionario, crecimiento celular, reparación de tejidos o crecimiento e invasión tumoral.

Una vez elaborada la base de datos y seleccionadas las variantes con las que se va a trabajar, el siguiente paso es la elaboración del pipeline que será capaz de obtener el genotipo de los SNPs mencionados, para a continuación desarrollar un algoritmo predictivo que a partir de la información genética obtenida del pipeline será capaz de predecir el riesgo que padece el cliente a padecer hipertensión, y con ello elaborar un informe genético en el que se detalla toda la información referente a dicho riesgo.

La secuenciación se va a realizar en la plataforma MiSeq de Illumina, y para ello se diseñan unos primers que van a permitir secuenciar regiones de 150 pares de bases. El diseño de estos primers ha de realizarse de tal manera que sean específicos para cada una de las regiones que contienen el SNP que se pretende secuenciar, dándoles una longitud suficientemente larga para evitar uniones inespecíficas a otras regiones del genoma.

Debido a complicaciones en el diseño de los primers, el cual da problemas para secuenciar la región del SNPs rsXXXXXXXX del gen PLEKHA7 por lo que finalmente no se incluye en el diseño final del algoritmo predictivo.

Esta eliminación de un SNP tiene un impacto negativo en el resultado final, pero no es un impacto lo suficientemente grande como para reducir de forma elevada la fiabilidad del estudio, por lo que se considera que no es necesario seleccionar un nuevo SNP que pueda ser incluído en el algoritmo.

3. Pipeline

3.1 Datos iniciales

Los datos con los que se va a trabajar inicialmente para llevar a cabo el estudio son datos procedentes de secuenciación masiva, en concreto de la plataforma MiSeq de Illumina.

Los datos finales son dos archivos por cada cliente en formato fastq. Los archivos en este formato son archivos de texto plano que siguen una sencilla organización. Para cada entrada que se realiza hay cuatro líneas [30]:

1. La primera línea es el identificador de la secuencia. Comienza siempre con el carácter arroba (@) y es seguido por una línea de texto que detalla y describe la información relevante de la entrada. El formato para esta primera línea en Illumina es:

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

   Donde cada uno de los elementos se detalla en la siguiente tabla:

Elemento	Requerimientos	Descripción
@	@	Cada línea de un identificador de secuencia empieza con @
<instrument>	Caracteres permitidos: a-z, A-Z, 0-9 y barra baja	Identificador del instrumento
<run number>	Numérico	Número de tirada en el instrumento
<flowcell ID>	Caracteres permitidos: a-z, A-Z, 0-9	ID de la célula de trabajo
<lane>	Numérico	Número de la columna
<tile>	Numérico	Número de la fila
<x_pos>	Numérico	Coordenada X en el clúster
<y_pos>	Numérico	Coordenada Y en el clúster
<read>	Numérico	Número de lectura (1 para "single read", 2 para paired-end")
<is filtered>	Y/N	Y si la lectura está filtrada, N si no lo está.
<control number>	Numérico	0 siempre
<index sequence>	ACTG	Secuencia índice

Tabla 8. Descripción de los elementos de un archivo en formato fastq.

2. La segunda línea contiene la secuencia de la lectura, representada con las letras A,C,G,T.

3. La tercera línea consiste únicamente en el símbolo "+".

4. La cuarta y última línea es el control de calidad. Consiste en una lista de caracteres de la misma longitud que la secuencia de la segunda línea, y cada uno de los caracteres está asociado a la base de la secuencia que ocupa su misma posición. Los caracteres representan un valor numérico que se traducirá como un score de calidad Phred.

   Cada uno de los caracteres y su score relacionado puede verse en la página de Illumina [31]

   El valor numérico asociado a cada una de las bases secuenciadas es un valor de probabilidad de que la base se haya secuenciado erróneamente, y se sigue la siguente fórmula:

   $$Q = -10log_{10}P$$

   Siendo `Q` el valor numérico y `P` la probabilidad de error de secuenciación. Así, por ejemplo, encontrar para una base concreta el símbolo de apertura de paréntesis "`(`" se asocia con una `Q=7` que tiene, siguiendo la fórmula, una probabilidad del 19.95% de haberse secuenciado erróneamente. Conociendo esta información, interesa que el valor de `Q` sea elevado puesto que se traduce en un bajo valor de P y por lo tanto la secuenciación de esa base ha sido de buena calidad.

Para cada uno de los archivos fastq puede haber cientos de lecturas (estando cada una de ellas formadas por esas cuatro líneas de información) y siempre vienen en parejas. Esto se debe a que uno de los archivos fastq contiene las lecturas de la secuenciación de la hebra inversa correspondiente a las lecturas de secuenciación del otro archivo fastq, de esta forma se analiza la calidad de ambas hebras.

Para el desarrollo del pipeline que hace el análisis de los datos obtenidos en Illumina se hace una búsqueda bibliográfica sobre posibles programas desarrollados previamente por otros investigadores para poder facilitar el trabajo y la iniciación a estudiantes a la detección de variantes genéticas. Se conocieron principalmente 2 programas que presentan diferentes aproximaciones en sus objetivos y estructura:

1. BRB-SeqTools es un programa que incluye diferentes herramientas y aplicaciones creadas para trabajar con datos de secuenciación de siguiente generación (Next Generation Sequiencing o NGS). Han de ser instalados tanto el programa como las herramientas a mano, y la función que desempeña el propio programa BRB-SeqTools es la relación entre las aplicaciones (es decir, genera el pipeline).

2. La máquina virtual TREVA (Targeted REsequencing Virtual Appliance) es, como su nombre indica, una máquina virtual del sistema operativo Ubuntu que viene con un conjunto de herramientas ya instaladas y optimizadas para el trabajo. El

pipeline de TREVA da soporte a diferentes tipos de análisis entre los que se incluye el buscado en el presente estudio, la llamada de variantes de un nucleótido único (SNPs).

TREVA y BRB-SeqTools poseen bastantes herramientas en común ya que pretenden conseguir el mismo resultado y para ello han de utilizar las aplicaciones que están mejor consideradas en la comunidad científica. Es necesario pues dar una descripción detallada de cada uno de los dos pipelines, con sus herramientas y procesos llevados a cabo, para poder encontrar las ventajas y desventajas que existen entre ambos pipelines y seleccionar el que mejor se adapte a las necesidades del presente trabajo.

3.2 BRB-SeqTools

3.2.1 Instalación y puesta a punto

Para poder instalar este programa, se recomienda trabajar en el sistema operativo Ubuntu 64-bit Desktop OS (versión 14.04), y es el sistema con el que se va a trabajar en el presente proyecto. Una vez se tiene el sistema operativo deseado, se descarga el instalador rellenando los campos de nombre, apellido, institución, país y e-mail [32].

El archivo descargado está comprimido y tras descomprimirlo con el código:

```
tar -xzvf seqtools-dl-1.0.tar.gz
```

Se obtienen los siguientes archivos:

- SeqTools: Versión de interfaz gráfica de la aplicación.
- seqtools_dge/seqtools_vc: Versión de consola de la aplicación.
- install.sh: Script para instalar la aplicación.
- INSTALL: Instrucciones de instalación.
- samples.txt: Plantilla del archivo samples.txt (necesario para el análisis).
- samples_gatk.txt: Plantilla del archivo samples.txt para GATK.
- examples.txt: Ejemplos de como usar el programa.
- icon: Carpeta que contiene los iconos de la aplicación.
- code: Carpeta que contiene scripts necesarios para ANNOVAR/SnpEff

Aunque es opcional, se corre en el terminal el archivo install.sh para que se cree un acceso directo al programa en el escritorio y sea más fácil acceder al programa.

Una vez instalado el propio BRB-SeqTools, es necesario instalar todas las herramientas que este programa utilizará para el análisis. Para ello, una vez abierto el programa hay que ir a `"Settings" -> "Tools manager"` donde aparecerá el nombre de las herramientas necesarias y una ventana de texto donde irá la localización de las herramientas. Se puede hacer de manera manual si se quiere emplear una versión concreta de una de las herramientas o si ya se tienen instaladas (de esta forma solo tenemos que decirle a BRB-SeqTools donde se encuentran localizadas las herramientas). En caso de no tener ninguna de las herramientas ya instaladas, el

programa permite al usuario una instalación automática que consiste en la ejecución de un script que descarga los programas necesarios y los organiza en unos directorios por defecto.

Como excepción, no se puede instalar de forma automática la herramienta GATK (Genome Analysis ToolKit) debido a que los desarrolladores de la misma han impuesto una restricción de licencia que lo impide. Por ello ha de ser instalado de forma manual y especificarse en BRB-SeqTools la localización de la herramienta. Para obtener GATK, hay que crear una cuenta en la página de GATK y seleccionar la versión que se quiere descargar, obteniendo un archivo (GenomeAnalysisTK-X.X.tar.bz2 siendo X.X la versión) comprimido y el cual hay que extraer en una carpeta a nuestra elección. Dicha carpeta será la que hay que especificar en la línea de GATK del "Tools Manager".

Una vez instaladas todas las herramientas, la ventana de "Tools Manager" ha de tener un aspecto similar al de la figura siguiente (aunque pueden variar las versiones de las herrramientas según la fecha en la que se realice la instalación):

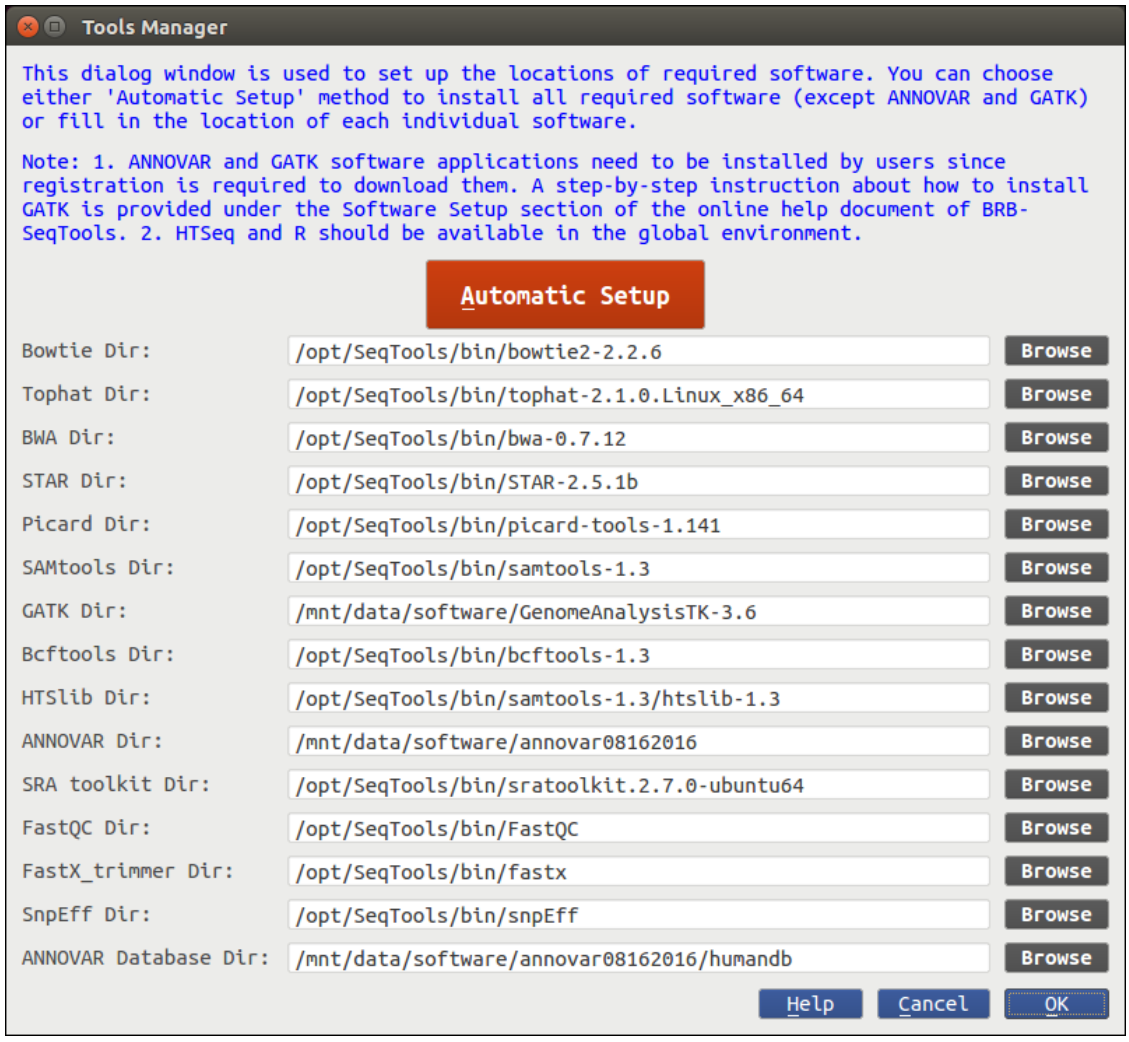

***Figura 5**. Ejemplo de lista de directorios de herramientas.*

De todas estas herramientas instaladas, no todas van a ser necesarias para el presente proyecto, puesto que algunas de ellas son necesarias para otro tipo de tareas como RNA-Seq o anotación extra de los resultados de la detección de variantes. A continuación se va a detallar el proceso que lleva a cabo el pipeline, explicando cada herramienta utilizada y su función, aunque previamente hay que preparar los archivos que se van a analizar de manera que el programa sea capaz de detectarlos correctamente.

Además de tener que instalar las herramientas, por la naturaleza de los análisis que se van a llevar a cabo es necesario tener por lo menos un genoma de referencia con el que se van a alinear las secuencias con las que se va a trabajar. Es posible crear un perfil para todos los genomas de referencia que el investigador desee y considere de forma manual, pero dado que el presente estudio va a realizarse con seres humanos, existe la opción de una instalación automática de los genomas de referencia más utilizados (entre los que se encuentran NCBI_GRCh38, Ensembl_GRCh37, UCSC_hg38 y UCSC_hg19) y es la que se sigue en este proyecto. Para ello, una vez abierto el

programa BRB-SeqTools hay que ir a "Settings" -> "Reference Genome Profile Manager" y seleccionar "Create a new profile" -> "Automatic Setup". Tras esto, se selecciona la build del genoma (en el caso de este proyecto se elige Ensembl_GRCh37) además de especificar el directorio donde se guardará el archivo de gran tamaño que es el genoma de referencia y el nombre del perfil creado. El programa, tras esto, descarga el pesado genoma y una vez concluído se crea el perfil que va a utilizarse para cada análisis.

3.2.2 Selección de análisis y samples.txt

El programa va a pedir al usuario que especifique inicialmente que tipo de análisis quiere llevar a cabo. En es caso del presente estudio, las opciones elegidas van a ser las siguentes:

- Data type: DNA-Seq
- Raw data format: fastq
- Analysis type: Variant calling

Dentro de la sección de "Analysis type" se pueden introducir opciones avanzadas, pero hacer referencia a las herramientas que se explicarán a continuación por lo que la información referente a esas opciones avanzadas se encuentra en los apartados de dichas herramientas.

Lo siguiente que necesita saber el programa es el directorio en el que se localizan los archivos fastq que se van a analizar, y hay que especificar la ruta completa del directorio donde se encuentran en la ventana de texto "Data Dir:" en la sección "Data location". También hay que especificar la ruta del directorio donde se han de volcar los archivos con los resultados del análisis, en este caso en la ventana de texto "Output Dir:"

Como ha de llevarse a cabo un alineamiento, ha de seleccionarse el perfil del genoma de referencia con el que se va a trabajar, que como se ha especificado previamente en este informe se va a utilizar la build Ensembl_GRCh37.

Por último, el programa requiere que en el mismo directorio especificado en "Data Dir:" (donde se encuentran los archivos fastq con la información de las secuencias a estudiar) se encuentre un archivo llamado samples.txt, un archivo de texto plano que consiste en una primera fila inicial con seis columnas (separadas por tabulador) con los nombres de las variables que se necesitan, y el investigador ha de añadir tantas filas como parejas de archivos fastq pretende analizar. Las seis columnas tienen los siguientes nombres y utilidades:

1. LibraryName: Especifica el nombre único que representa a la muestra.
2. LibraryLayout: PAIRED o SINGLE en función de si las muestras son "paired-end" o "single read" respectivamente. Por la naturaleza del proyecto, todas nuestras muestras pertenecen a la categoría PAIRED.
3. fastq1: Nombre de uno de los archivos fastq presentes en el directorio.
4. fastq2: Nombre del archivo fastq complementario al especificado en fastq1.

5. ReadGroup: Valor numérico para GATK, cada fila tendrá un número distinto.
6. SequencerManufacturer: Dado que los datos provienen de MiSeq (Illumina), en esta columna siempre ha de aparecer "ILLUMINA".

En la siguiente figura se puede ver un ejemplo real de un archivo sample.txt con la información necesaria para realizar el análisis de una muestra (archivo abierto desde LibreOffice Calc para una mejor visualización del contenido, pero no deja de ser un texto plano que separa las columnas con tabulador).

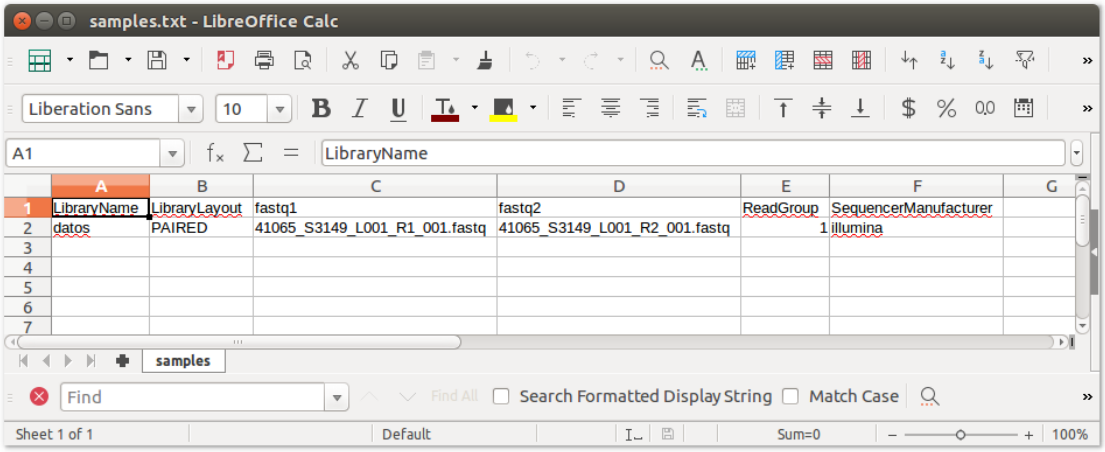

***Figura 6**. Ejemplo de un archivo samples.txt para una muestra.*

3.2.3 Script

Tras terminar los preparativos iniciales se puede comenzar el análisis haciendo click en "Run", pero haciéndolo de esta manera se ejecutan los programas con los argumentos predeterminados por el desarrollador de BRB-SeqTools, y aunque estos están correctamente codificados, son muy generales y pueden no dar los resultados exactos que se pretenden obtener. Para poder modificar los argumentos, antes de seleccionar "Run" se ha de ir a `"Settings" -> "Preferences"`, donde la única opción que se permite marcar es "Create the execution script file only". Marcando esa casilla se consigue que el programa no ejecute todo el pipeline, sino que genere el script que iba a ser corrido para poder ser modificado al antojo del investigador. Se aprecia que ya no se puede seleccionar "Run" y en su lugar se encuentra el botón "Create the execution script file", el cual al seleccionarlo crea el script denominado run_seqtools.sh en la carpeta de output previamente especificada.

Este script puede ser modificado fácilmente con el editor de texto gedit, y aunque en la mayoría de los casos de este proyecto no se necesitan modificar los argumentos, en la herramienta GATK se necesitan modificar algunos (que se encuentran especificados en el apartado correspondiente). Además, el hecho de ver el código que ejecuta el programa ayuda a su comprensión y a poder detectar posibles fallos a lo largo del pipeline e incluso añadir código extra que ayude al investigador (como comentarios explicando cada paso tras el carácter #).

Cabe destacar que el código presente en el script predeterminado no solo incluye los argumentos para la ejecución de las herramientas, sino que añade elementos que permiten al investigador llevar a cabo las tareas con mayor facilidad:

- Crea un sistema de jerarquía de directorios para almacenar ordenadamente los archivos intermedios creados a lo largo del análisis.
- Incluye comentarios con # que describen que se está realizando en cada momento.
- Tras la ejecución de cada línea de código de una herramienta, almacena la información que la consola genera en un archivo de texto mediante el código:

```
    >> "$outputDir/tmp/log.txt" 2>&1 2>&1
```

Una vez el script ha sido modificado y todo el código se ha adaptado a las necesidades del investigador y del proyecto, se debe ejecutar. Para ello se ha de abrir el terminar, navegar hasta llegar al directorio donde se encuentra el script y ejecutarlo mediante el código:

```
bash run_seqtools.sh
```

El script pone a correr diferentes herramientas en el orden necesario para que el análisis se desarrolle con normalidad, y a continuación se describen las herramientas en orden y su utilidad en la detección de variantes.

3.2.4 Burrows-Wheeler Aligner (BWA)

Lo primero que se hace en el pipeline es alinear las lecturas con el genoma de referencia, y para ello se utiliza la herramienta Burrows-Wheeler Aligner, también denominada BWA. BWA es un programa que permite mapear secuencias de baja divergencia en grandes genomas de referencia, como por ejemplo el genoma humano. Posee tres algoritmos: BWA-backtrack, BWA-SW y BWA-MEM. Para lecturas de Illumina como es el caso, los desarrolladores recomiendan utilizar el algorirmo BWA-MEM, por lo que es el que se utiliza en este pipeline [33].

Para ejecutar esta herramienta, en el script se presenta el siguiente código:

```
bwa mem -t [INT] -M ["ruta completa del genoma de referencia/genome.fa"]
[nombre del
fastq1 en samples.txt] [nombre del fastq2 en samples.txt] > "datos/bwa.sa
m"
```

- `-t [INT]` es el número de threads que se van a utilizar para llevar a cabo el trabajo, a mayor número más rápido será el proceso pero consume más recursos del ordenador.
- `-M` adapta el formato del archivo para compatibilidad con el programa Picard, es opcional.

- "datos/bwa.sam" es el archivo .sam donde se encuentran los datos obtenidos por BWA-MEM, es decir este archivo contiene la información sobre el alineamiento de las lecturas en el genoma de referencia seleccionado.

Un ejemplo real de ejecución del código de BWA-MEM es:

```
bwa mem  -t 3 -M "/home/pablo/Documents/reference/Homo_sapiens/Ensembl/GRCh37/Sequence
/BWAIndex/genome.fa" 41065_S3149_L001_R1_001.fastq 41065_S3149_L001_R2_001.fastq >
"datos/bwa.sam"
```

3.2.5 Samtools

El output que genera BWA es un archivo en formato SAM (Sequence Alignment Map), que busca ser un formato genérico para almacenar grandes cantidades de alineamientos de secuencias de nucleótidos. SAM es uno de los formatos preferidos para esta tarea debido a su flexibilidad y su amplio uso por la mayoría de los investigadores [34].

Samtools es una herramienta que permite trabajar con los alineamientos en formato SAM pudiendo clasificar, fusionar, indexar y generar alineamientos.

Samtools utiliza el archivo "bwa.sam" para refinar su contenido y optimizarlo para su uso en el resto de herramientas del pipeline. Para ello, se utiliza la función "fixmate" que elimina las lecturas que no han sido mapeadas y selecciona únicamente los resultados correctamente alineados en un nuevo archivo. Dicho nuevo archivo cambia de formato, y en este caso es BAM, cuya diferencia principal con un archivo en formato SAM es que los BAM no pueden leerse en un editor de texto al ser archivos binarios, pero se encuentran comprimidos.

El código ejecutado para obtener el nuevo archivo comprimido y optimizado es:

```
samtools fixmate -O bam "datos/bwa.sam" "datos/accepted_hits_bwa.bam"
```

- "-O bam" es la instrucción que especifica que el nuevo archivo se encuentre en formato BAM.
- "datos/bwa.sam" es el archivo con el que va a trabajar la herramienta.
- "datos/accepted_hits_bwa.bam" es el nombre del archivo que va a generar.

La siguiente instrucción consiste en clasificar los alineamientos, concretamente ordenándolos por nombre de lectura. Para ello, se utiliza la función de Samtools llamada "sort", ejecutando el siguiente código:

```
samtools sort -@ 3 accepted_hits_bwa.bam -o sorted.bam
```

- "-@ 3" es el número de threads que se van a utilizar (a mayor número, mayor rendimiento pero más recursos del ordenador que se consumen).
- "accepted_hits_bwa.bam" es el archivo con el que va a trabajar la herramienta.
- "-o sorted.bam" es el nombre del archivo que va a generar.

Tras esto, se indexa el archivo "sorted.bam" según coordenadas mediante la función de Samtools "index". Es importante saber que no se puede indexar archivos SAM con esta función. El código ejecutado es:

```
samtools index sorted.bam
```

Tras esto, se obtiene un archivo BAM por cada muestra, y se renombra cada uno de los "sorted.bam" generados por el nombre único elegido en "LibraryName" del archivo "samples.txt" para esa muestra utilizándo el código:

```
rm sorted.bam <nombre de cada muestra>.bam
```

3.2.6 Picard

Picard es una herramienta que trabaja también con datos en archivos de formato SAM o BAM y está optimizada para ejecutar líneas de comandos para manipular datos de "high-throughput sequencing" (HTS) [35].

BRB-SeqTools da al usuario la opción de aplicar la función "MarkDuplicates" de Picard para eliminar lecturas duplicadas en el archivo BAM para el futuro análisis, es un comando opcional que ha de ejecutarse o no en función del análisis realizado. En el caso del presente proyecto, no hay que utilizar este comando por las características de los datos procedentes de Illumina. Concretamente, la guía de "Best practices" [36] de GATK para la detección de SNP e indels en su apartado de "Mark Duplicates" dice que esta opción no debe utilizarse si se trabaja con datos de secuenciación por amplicones u otro tipo de datos donde las lecturas empiezan y terminan por la misma posición por diseño.

El primer comando que se realiza con Picard (al no llevarse el marcado de duplicados) es la función "AddOrReplaceReadGroups", que permite al usuario reemplazar todos los grupos de lecturas del archivo en formato BAM que se indique por un único nuevo grupo de lecturas y asignar todas las lecturas a este grupo de lecturas en un nuevo archivo en formato BAM. El código ejecutado es:

```
java -Xmx10g -jar <Directorio de instalación de Picard> AddOrReplaceReadG
roups
INPUT=<nombre de cada muestra>.bam OUTPUT=rg_added_sorted.bam RGID=1 RGLB
=dna
RGPL=illumina RGPU=UNKNOWN RGSM=<nombre de cada muestra>
```

- "java -Xmx10g -jar <Directorio de instalación de Picard> AddOrReplaceReadGroups" es el código que llama a la función "AddOrReplaceReadGroups".
- "INPUT=<nombre de cada muestra>.bam" es el archivo con el que va a trabajar Picard.
- "OUTPUT=rg_added_sorted.bam" es el archivo que se va a generar.
- "RGID=1" especifica, en este caso, el grupo 1 (que ha sido asignado en el apartado "ReadGroup" de samples.txt).

- "RGLB=dna" especifica que se está trabajando con DNA (este valor lo asigna BRB-SeqTools debido a que en la configuración inicial se selecciona "DNA-Seq").
- "RGPL=illumina" especifica la plataforma que ha llevado a cabo la secuenciación (el programa recibe esta información del apartado "SequencerManufacturer" en samples.txt)
- "RGPU=UNKNOWN" debería incluir más información sobre la unidad utilizada por plataforma de secuenciación, pero el formato de samples.txt no incluye dicha información. La función "AddOrReplaceReadGroups" exige darle un valor, por lo que BRB-SeqTools decide dar como valor por defecto "UNKNOWN".
- "RGSM=<nombre de cada muestra>" es el último argumento y es el nombre del grupo de lectura (el cual proviene del apartado "LibraryName" en samples.txt).

Tras esto, se ha de reordenar el archivo "rg_added_sorted.bam". Para ello se utiliza la función "ReorderSam", la cual reordena las lecturas del archivo para coincidir con el orden del genoma de referencia seleccionado (puesto que es necesario que se encuentre en este orden para la detección de variantes). Cabe destacar que una de las ventajas de mantener el formato BAM en vez de SAM es que esta parte del análisis es mucho más rápida con los archivos en formato BAM. El código empleado para ejecutar esta función es:

```
java -Xmx10g -jar <Directorio de instalación de Picard> ReorderSam
INPUT=rg_added_sorted.bamOUTPUT=reorder.bam REFERENCE=<Ruta del genoma de
referencia>
```

- "java -Xmx10g -jar <Directorio de instalación de Picard> ReorderSam" es el código que llama a la función "ReorderSam".
- "INPUT=rg_added_sorted.bam" es el archivo con el que va a trabajar Picard.
- "OUTPUT=reorder.bam" es el archivo que se va a generar.
- "REFERENCE=<Ruta del genoma de referencia>" es el genoma de referencia (el cual es especificado en los pasos previos a la creación del script al seleccionar el perfil del genoma elegido por el investigador).

El programa a continuación renombra el archivo "reorder.bam" como "split.bam" mediante el código:

```
mv ./reorder.bam ./split.bam
```

Y aplica al mismo la función de Samtools "index" con el código:

```
samtools index split.bam
```

3.2.7 GATK

Los pasos realizados por el pipeline hasta el momento han sido de preparación de los archivos para pasar de una pareja de archivos en formato fastq a un archivo en formato BAM que puede ser interpretado por la herramienta que va a realizar la propia detección de variantes. BRB-SeqTools da la opción de elegir entre dos herramientas para llevar a cabo esa detección: Genome Analysis Toolkit (GATK) o

Samtools dentro de las opciones avanzadas del apartado "Analysis type". Además de dar esta opción, si se elige el programa GATK se pueden modificar unas opciones en sus parámetros, como seleccionar si se quiere llevar a cabo el realineamiento de inserciones/deleciones (también conocidos como indels) o si se quiere aplicar la recalibración de la puntuación de la calidad de las bases (Base quality score recalibration).

Debido a la posibilidad de llevar a cabo un análisis más exhaustivo con GATK, es la herramienta de elección para el presente proyecto [37].

En general, una gran cantidad de regiones requieren un realineamiento debido a la presencia de una inserción o una deleción en el genoma del individuo con respecto al genoma de referencia. Dichas mutaciones generan artefactos que son confundidos con SNPs (Single Nucleotide Polymorfisms) que generan muchos problemas con la llamada de resultados. Para solucionar este problema existe la función de GATK llamada "IndelRealigner", que realiza un realineamiento local en las regiones donde hay indels de forma que genera nuevas lecturas limpias y en las posiciones del genoma de referencia correctas. El código empleado para ejecutar la función "IndelRealigner" es:

```
java -Xmx10g -jar <Directorio de instalación de GATK> -T IndelRealigner -
R
<Ruta del genoma de referencia> -I split.bam -targetIntervals <Ruta de in
tervalos de
trabajo> -known <Ruta al archivo para indels> -o realigned_reads.bam
-allowPotentiallyMisencodedQuals
```

- "java -Xmx10g -jar <Directorio de instalación de GATK> -T IndelRealigner" es para llamar a la función.
- "-R <Ruta del genoma de referencia>" es para especificar la ruta completa del genoma de referencia (el mismo que se ha utilizado en el resto de casos a lo largo del proceso).
- "-I split.bam" es el archivo con el que va a trabajar GATK.
- "-targetIntervals <Ruta de intervalos de trabajo>" es el archivo que contiene los intervalos con los que va a trabajar, los cuales se obtienen desde el archivo especificado en "-known <Ruta al archivo para indels>". En este archivo se encuentra la información para generar los intervalos de indels desde los que se va a realizar el realineamiento.
- "-o realigned_reads.bam" es el argumento que especifica el nombre del archivo que se va a generar.
- "-allowPotentiallyMisencodedQuals" hace que no salga aviso de scores de calidades poco comunes en las lecturas de las bases. Se eliminan estos avisos porque la siguiente función va estudiar dichas calidades y valorarlas.

La detección de variantes depende en gran medida de la calidad de los scores asignados a cada una de las bases secuenciadas en cada lectura. Sin embargo, aunque estos scores estiman el posible error de las máquinas de secuenciación, es posible que

por errores sistemáticos técnicos haya una sobreestimación o infraestimación de la calidad de los scores. Para tratar de solventar este problema, GATK dispone de una función llamada Base quality score recalibration (BQSR), que es un proceso en el que se aplican técnicas de machine learning para modelar los errores de forma empírica y ajustar las calidades de las bases de acuerdo a ello. Gracias a esta función, la calidad final de los resultados es de mucha mayor confianza.

Para ejecutar esta función, el código es:

```
java -Xmx10g -<Directorio de instalación de GATK> -T BaseRecalibrator -R
<Ruta del genoma
de referencia> -I realigned_reads.bam -nct 3 -knownSites <Ruta al archivo
para indels>
-o recal_data.table -allowPotentiallyMisencodedQuals
```

- "java -Xmx10g -jar <Directorio de instalación de GATK> -T BaseRecalibrator" es el código necesario para ejecutar la función.
- "-R <Ruta del genoma de referencia>" es la ruta completa del genoma de referencia.
- "-I realigned_reads.bam" es el archivo con el que va a trabajar.
- "-nct 3" especifica el número de threads que va a emplear el proceso.
- "-knownSites <Ruta al archivo para indels>" da la ruta del archivo donde se encuentra la información de SNPs (los cuales han de tenerse en cuenta a la hora de comparar con el genoma de referencia).
- "-o recal_data.table" es el archivo que va a producir el programa.
- "-allowPotentiallyMisencodedQuals" hace que no salga aviso de scores de calidades poco comunes en las lecturas de las bases.

Como se observa, el archivo obtenido no se encuentra en formato BAM, sino que tiene una extensión .table, por lo que se ha de ejecutar una función más para finalizar el BQSR en un formato BAM con el que poder seguir trabajando. Para ello existe la función "PrintReads", la cual se ejecuta mediante el siguiente código:

```
java -Xmx10g -jar <Directorio de instalación de GATK> -T PrintReads -R <R
uta del genoma
de referencia> -I realigned_reads.bam -nct 3 -BQSR recal_data.table -o re
cal.bam
```

- "java -Xmx10g -jar <Directorio de instalación de GATK> -T PrintReads" es el comando que ejecuta la función.
- "-R <Ruta del genoma de referencia>" es la ruta completa al genoma de referencia utilizado.
- "-I realigned_reads.bam" el archivo inicial con el que se había ejecutado la función "BaseRecalibrator".
- "-nct 3" es el número de threads utilizados en el proceso.
- "-BQSR recal_data.table" es el archivo obtenido tras ejecutarse "BaseRecalibrator".

- "-o recal.bam" es el nombre del archivo final que contendrá la información de "BaseRecalibrator" en formato BAM. En esencia, esta última función tiene como objetivo la conversión de formato requerida para poder continuar con el análisis.

El último código que se va a ejecutar es el más crítico de todo el proceso, puesto que va a realizar la detección de variantes en sí. No se puede decir que es el paso más importante puesto que si alguno de los pasos previos no se lleva a cabo, no se tendría la información formateada de una forma que GATK pudiera comprender y analizar. Sin embargo, al ser este un paso crucial en el análisis (puesto que va a generar una lista de las variantes genéticas detectadas) se han de especificar los argumentos concretos que GATK debe llevar a cabo y estos van a ser diferentes a los que BRB-SeqTools ofrece de forma predeterminada.

La función de GATK que va a utilizarse en este paso se denomina "HaplotypeCaller". Sus desarrolladores la definen como una función capaz de detectar SNPs e indels de forma simultánea, es decir es capaz de encontrar regiones que muestras signos de variaciones. El código empleado es:

```
java -Xmx10g -jar <Directorio de instalación de GATK> -T HaplotypeCaller
--genotyping_mode
DISCOVERY --dbsnp <Ruta a archivo dbSNP> --output_mode EMIT_ALL_SITES -L
<Ruta a archivo
.bed con localizaciones cromosómicas> -ERC GVCF -R <Ruta del genoma de re
ferencia>
-I recal.bam  -stand_call_conf 0 -o <Ruta para output deseada>/datos_raw.
g.vcf
-allowPotentiallyMisencodedQuals -nct 3
```

- "java -Xmx10g -jar <Directorio de instalación de GATK> -T HaplotypeCaller" llama a la función para realizar el análisis.
- "--genotyping_mode DISCOVERY" es un argumento que permite al investigador decirle al programa si han de descubrirse los alelos o si únicamente los dados por el investigador han de ser considerados. Como interesa descubrir la presencia de alelos alternativos, se da la opción DISCOVERY.
- "--dbsnp <Ruta a archivo dbSNP>" es un argumento que BRB-SeqTools no incluye pero es de gran interés para el trabajo realizado. Este argumento utiliza un archivo que contiene información sobre todos los SNPs conocidos por la base de datos dbSNP y si el SNP detectado por el programa se encuentra en dicha base de datos, se incluye en el archivo final su nombre, lo que ayuda a identificarlo rápidamente y comprobar que se han detectado las variantes correctas.
- "--output_mode EMIT_ALL_SITES" es otro argumento muy importante no presente en las opciones iniciales de BRB-SeqTools. Este argumento permite al investigador decidir si se van a detectar e incluir en los resultados únicamente las regiones donde hay presentes variantes o si por el contrario se han de incluir también las regiones que no contienen variantes. Por lo general, no interesa en un programa de detección de variantes incluir regiones que no contienen alelos alternativos al de referencia, pero dado que este proyecto pretende buscar el

genotipo exacto en ciertas regiones, interesa incluir en los resultados genotipos en los que no hay variantes (de hecho, serán comunes dado que las variantes son menos frecuentes que los alelos de referencia y muchos de los clientes no tendrán en varias regiones alelos alternativos). La opción predeterminada para este argumento es "EMIT_VARIANTS_ONLY" y para este proyecto se selecciona la opción "EMIT_ALL_SITES".

- `"-L <Ruta a archivo .bed con localizaciones cromosómicas>"` especifica cual es el archivo que contiene las coordenadas cromosómicas de las variantes que se pretenden estudiar. Debido a que nos interesa conocer el genotipo de unas bases muy concretas del genoma, si no establecemos las localizaciones previamente el programa hará un llamamiento de variantes y genotipado de todas las bases del genoma (imposibilitando la tarea debido a que tardaría días en llevarse a cabo y generando un archivo de decenas de gigabytes) como consecuencia de las opciones que hemos seleccionado por la naturaleza del estudio. Este archivo en formato .bed es de texto plano en el que cada una de las filas tiene este formato: , separando cada campo con tabulador.
- `"-ERC GVCF"` es un argumento opcional pero vital en este estudio. Lo que hace es generar un archivo en formato .g.vcf, que posee unas propiedades distintas a un archivo en formato .vcf y permite hacer la llamada de genotipo en las bases donde no se detectan variantes (en colaboración con EMIT_ALL_SITES, de hecho son necesarios ambos argumentos para el correcto funcionamiento). El archivo en formato .g.vcf (también llamado GVCF) será temporal y almacena la información obtenida de esta muestra.
- `"-R <Ruta del genoma de referencia>"` es la ruta completa al genoma de referencia.
- `"-I recal.bam"` es el archivo con el que se va a trabajar.
- `"-stand_call_conf 0"` indica el valor de calidad mínimo que debe tener un resultado para ser considerado como de buena calidad por el programa. En caso de que la calidad sea buena no pasa nada, pero si no se llega a la calidad mínima en los resultados finales aparece un aviso para que el investigador tenga en cuenta que esa variante no es de buena calidad.
- `"-o <Ruta para output deseada>/datos_raw.g.vcf"` es la ruta y el nombre del archivo donde se encontrarán los resultados en formato GVCF (Genomic Variant Calling Format).
- `"-allowPotentiallyMisencodedQuals"` hace que no salga aviso de scores de calidades poco comunes en las lecturas de las bases.
- `"-nct 3"` el número de threads utilizados en el proceso.

3.2.8 Nuevas muestras

El código utilizado hasta el momento permite llegar hasta obtener un archivo en formato GVCF de una muestra. No se ha terminado el análisis debido a que ese archivo no es el definitivo, pero antes de seguir analizándolo hay que llevar a cabo todo el pipeline hasta el momento con todas las muestras que se pretenden analizar. Para ello, se copia todo el código utilizado para la primera muestra, se pega debajo del

código ya descrito y se cambian los siguientes parámetros de forma que hagan referencia a la nueva muestra:

- Si por ejemplo el nombre de la primera muestra era "Muestra_1", buscar y cambiar en todo el código para la nueva muestra ese nombre al de la nueva muestra (por ejemplo, "Muestra_2"). Si se hace correctamente y se cambian todos los nombres, se van a utilizar correctamente todos los comandos para dar el nuevo análisis sin interferir con los datos de la muestra anterior.
- En BWA se pide el nombre de los archivos .fastq por lo que es importante cambiar el nombre de los antiguos archivos por el de los archivos de la nueva muestra.
- En la función `"AddOrReplaceReadGroups"` de Picard, se cambia el argumento `RGID` por un valor único. Por ejemplo, en la Muestra_1 dicho valor puede ser 1 y para la Muestra_2 puede darse el valor de `RGID=2`.

Como se puede ver, estos cambios hacen referencia a elementos del archivo samples.txt, se puede pensar que simplemente añadiendo a dicho samples.txt una fila para cada muestra se podría evitar esta modificación. Si bien eso es cierto, ese método sería útil en caso de utilizar el script sin modificar, y como en este estudio el script se ha modificado sustancialmente, se ha considerado más práctico generar el script para una muestra, hacer las modificaciones pertinentes y tras ello se copia y pega el código optimizado y cambiando estos elementos para cada muestra.

3.2.9 Unión de los archivos GVCF

Tras realizar el análisis de todas las muestras, se obtiene un archivo en formato GVCF para cada una de las mismas. El siguiente paso consiste en reunir toda la información de todas las muestras en un único archivo en formato VCF. Desarrollado específicamente para ello, GATK posee la función `GenotypeGVCFs`:

```
java -Xmx10g -jar <Directorio de instalación de GATK> -T GenotypeGVCFs
--includeNonVariantSites --dbsnp <Ruta a archivo dbSNP> -L <Ruta a archiv
o
.bed con localizaciones cromosómicas> -R <Ruta del genoma de referencia>
--variant <Ruta del archivo GVCF de la Muestra_1> --variant <Ruta del arc
hivo
GVCF de la Muestra_2> -o <Nombre del archivo con los resultados en format
o VCF>
```

- `"java -Xmx10g -jar <Directorio de instalación de GATK> -T GenotypeGVCFs"` llama a la función que va a llevar el proceso.
- `"--includeNonVariantSites"` indica a la función que se incluyan los resultados que poseen un genotipo homocigoto de referencia, es decir que incluya también los casos en los que no se han detectado variantes. Es importante este paso puesto que será habitual encontrar homocigotos de referencia en los análisis, omitiendo este argumento se perdería esa información.

- "--dbsnp <Ruta a archivo dbSNP>" se utiliza de forma idéntica a los casos anteriores, su objetivo el poder identificar y nombrar los SNPs si se encuentran presentes en la base de datos de dbSNP.
- "-L <Ruta a archivo .bed con localizaciones cromosómicas>" de forma similar a los casos anteriores, con este argumento se acotan las regiones cromosómicas donde se va a trabajar.
- "-R <Ruta del genoma de referencia>" es la ruta completa al genoma de referencia.
- "--variant" ha de incluir la ruta completa al archivo GVCF de cada muestra. Ha de incluirse un --variant por cada muestra, en este ejemplo aparece dos veces pero se pueden incluir tantos como se desee.
- "-o <Nombre del archivo con los resultados en formato VCF>" especifica la ruta y el nombre que tendrá el archivo definitivo, ahora en formato VCF.

3.2.10 Filtrado de variantes

Ya se tiene el archivo definitivo que contiene toda la información necesaria para generar el informe genético, pero antes de utilizarla se han de filtrar los resultados para comprobar si dichos resultados tienen una calidad suficiente para poder ser utilizados. Para ello se emplea la función VariantFiltration de GATK, la cual permite introducir unos valores a elección del investigador de los muchos parámetros que se pueden analizar y como resultado menciona si cada variante ha pasado el filtrado o no.

Dependiendo del tipo de estudio, muestras, resultados esperados y otros muchos elementos, el filtrado debe tener en cuenta muchos y variados parámetros. Para el presente estudio, se seleccionan los filtrados con los parámetros recomendados por los desarrolladores de GATK llamados "hard filters" [38]:

- QD < 2.0
- FS > 60.0
- MQ < 40.0
- MQRankSum < -12.5
- ReadPosRankSum < -8.0

```
java -Xmx10g -jar <Directorio de instalación de GATK> -T VariantFiltratio
n -R <Ruta al
genoma de referencia> -V <Archivo VCF a filtrar> --filterExpression "QD <
2.0 || FS > 60.0
|| MQ < 40.0 || MQRankSum < -12.5 || ReadPosRankSum < -8.0" --filterName
"FAIL" -o <Ruta
y nombre del archivo VCF con el filtrado incluido>
```

- "java -Xmx10g -jar <Directorio de instalación de GATK> -T VariantFiltration" llama a la función para hacer el filtrado.
- "-R <Ruta al genoma de referencia>" es la ruta completa al genoma de referencia.

- "-V <Archivo VCF a filtrar>" especifica cual es el archivo VCF al que se le van a aplicar los filtros.
- "--filterExpression "QD < 2.0 || FS > 60.0 || MQ < 40.0 || MQRankSum < -12.5 || ReadPosRankSum < -8.0"" especifica cuales son los filtros utilizados, en este caso son cinco y se corresponden a los "hard filters" recomendados por los desarrolladores de GATK.
- "--filterName "FAIL"" especifica que en caso de que no se pasen los filtros, aparezca el mensaje FAIL.
- "-o <Ruta y nombre del archivo VCF con el filtrado incluido>" Ruta y nombre del nuevo archivo VCF, idéntico en la mayoría de los sentidos al anterior pero que incluye los resultados del filtrado.

Con esto, se finaliza la ejecución del script y se obtiene el archivo definitivo con toda la información necesaria para aplicarla en el algoritmo predictivo.

Tras llevarse a cabo todo el proceso, se obtiene un archivo final en formato VCF el cual se divide en dos partes. La primera parte es un conjunto de líneas que comienzan con los símbolos "##" y dan información sobre el formato de la segunda parte del archivo (que contiene la información del trabajo). La información dada en esta primera parte incluye la versión del formato VCF utilizada (ya que puede ir variando y es importante para el investigador conocer la versión para evitar posibles futuras incompatibilidades), descripciones de lo que significan los diferentes datos que aparecen en varios campos de los resultados, los argumentos utilizados por GATK HaplotypeCaller (que incluye, además de los especificados por el investigador, todos aquellos optativos y que al no ser especificados llevan a cabo el argumento predeterminado por el programa), la ruta completa del genoma de referencia utilizado y los cóntigos (en inglés contigs) utilizados. Dicho de forma simple, los cóntigos son segmentos de ADN superpuestos que juntos representan una región consenso del ADN, y en este caso cada uno de los cóntigos representa un cromosoma (del 1 al 23, X, Y y mitocondrial).

La segunda parte incluye los resultados del análisis y se encuentran ordenados en una tabla que contiene 10 columnas (separadas por tabulador), cada una de ellas da información relevante para el investigador. Para cada una de las variantes detectadas hay una nueva fila en la que se muestra la información de las columnas, las cuales dan la siguiente información:

1. CHROM: especifica el cromosoma en el que se encuentra la variante, el programa toma esta información del genoma de referencia utilizado y lo especifica en el apartado de cóntigos previamente mencionado.

2. POS: posición en la que se encuentra la variante detectada en el cromosoma especificado.

3. ID: De forma predeterminada este campo se encuentra vacío ya que es un espacio reservado para añadir una identificación a la variante por parte del investigador. En el caso del presente proyecto se emplea en HaplotypeCaller el argumento que

busca la localización cromosómica de la variante en la base de datos dbSNP y si esta posición posee un SNP catalogado, su nombre será especificado en este campo.

4. REF: Especifica la base que posee el genoma de referencia en esta posición (A, G, T, C).

5. ALT: Especifica la base alternativa encontrada en esta posición, en contraste con el genoma de referencia (A, G, T, C). En caso de no encontrarse alelo alternativo, este campo contiene un valor de 0.

6. QUAL: Da un valor de la calidad de la detección de la variante. Este valor se encuentra en escala Phred (explicada previamente) por lo que interesa que el valor de este campo sea elevado. Sin embargo, este valor se ve modificado por otras variables que hacen que este valor, aunque necesario, no sea el definitivo que ha de consultarse para comprobar si la variante es de buena calidad.

7. FILTER: Especifica si la variante ha pasado o no el control de calidad. El argumento `"-stand_call_conf 0"` de HaplotypeCaller en GATK es el primer filtro que se aplica en este caso, y al emplearse el valor 0 es evidente que todas las variantes van a pasar el control, ya que 0 es el valor mínimo y en el campo de QUAL siempre habrá valores superiores. Se hace esto porque como se ha dicho previamente, el valor definitivo que hay que tener en cuenta para comprobar la calidad de la variante no es QUAL por lo que un aviso de baja calidad es irrelevante.

   Sin embargo, gracias a la función `VariantFiltration` se lleva a cabo un segundo filtrado, ya mencionado y que aplica los "hard filters" recomendados por los desarrolladores de GATK para la llamada de variantes. En caso de que se pasen estos filtros, aparecerá en este apartado el valor "PASS" mientras que si la base no pasa uno o más de los filtros aparecerá el mensaje "FAIL" de forma que el investigador sabe que estos datos no poseen la calidad necesaria para continuar trabajando con la misma.

8. INFO: Incluye información adicional sobre las variantes que pueden servir o no al investigador, dado que el abanico de estudios que se pueden realizar con estas herramientas es tan amplio que los desarrolladores incluyen todo tipo de información para satisfacer las demandas de los investigadores. Sin embargo, para este proyecto la mayoría de los diferentes datos son irrelevantes (por ejemplo, la cantidad de alelos presentes es innecesario saberla ya que al ser genoma humano diploide sabemos que el valor es 2) y se tienen en cuenta dos datos llamados QD y DP. QD es definido en el apartado de descripción como "Variant confidence/Quality by depth" que se entiende como la calidad de la variante (QUAL) dividido entre la cantidad de reads que se encuentran presentes en la variante (Depth = DP). Esto se debe a que el valor de QUAL se ve influido por la cantidad de reads de la variante, por lo que puede dar un valor aparentemente bueno pero que al verse en relación a DP se observa que la calidad es mala. Por lo

tanto, se toma QD como el valor de calidad a tener en cuenta en la detección de variantes. Es a juicio del investigador establecer un valor mínimo de calidad, y para el presente estudio se va a mantener el valor de QD>20 utilizado frecuentemente como valor estándar en diversos estudios científicos [39].

9. FORMAT: Este campo especifica el formato que tendrán las siguientes columnas, y la descripción de los elementos presentes se encuentra situada en la primera parte del archivo junto al resto de información.

10. [LibraryName]: Aparece un campo para describir la información especificada por el campo FORMAT por cada LibraryName utilizado en el archivo samples.txt inicial, es decir para cada muestra. Aparecen diferentes datos que interesan en este estudio, siendo el más relevante el genotipo (especificado en el campo FORMAT como GT). Dado que vamos a trabajar con seres humanos y para las variantes que interesan van a ser o un alelo u otro, las posibilidades en este apartado son: 0/0 para cuando el individuo es homocigoto para el alelo de referencia, 0/1 para cuando el individuo es heterocigoto y posee uno de los alelos de referencia y otro de los alelos alternativos, y 1/1 para cuando el individuo es homocigoto para el alelo alternativo. También es importante conocer el dato DP (que es distinto al valor de DP del campo INFO, ya que ese es la suma total del número de reads o DP de cada una de las muestras y en cada una de las muestras aparece el número de reads que tiene para esa variante). La DP de cada muestra es importante, ya que se puede considerar buena calidad si hay un valor DP>5 por lo que hay que tener presente también este dato a la hora de considerar la calidad de la variante. (Fuente) También se puede observar el valor PL, que es descrito como la probabilidad en escala Phred de que cada uno de los genotipos sea correcto. Ofrece 3 valores que corresponten a los genotipos 0/0, 0/1 y 1/1, y dado que se encuentran en escala Phred interesa que el valor del genotipo detectado sea muy bajo (que se traduce en un porcentaje de acierto elevado). Por ejemplo, un genotipo 0/1 tendría un buen valor PL=100,0,100 y se podría decir que el genotipo se ha detectado correctamente.

3.2.11 Testeo del pipeline

Una vez se tiene el pipeline, se ha testeado con unos primeros archivos en formato FASTQ procedentes de Illumina que contienen las regiones de los SNPs estudiados en el presente proyecto. La prueba se ha hecho con dos muestras para comprobar que no hay problemas en la creación del VCF final, y los resultados obtenidos son correctos de forma que se puede confirmar que el pipeline ha sido desarrollado correctamente y cumple su cometido.

3.3 TREVA

3.3.1 Descripción

Con el rápido desarrollo de la secuenciación masiva, además de sus mejoras en eficiencia y coste, cada vez más laboratorios buscan una manera de poder llevar a cabo análisis de dichos datos como en el presente estudio, donde se busca detectar variantes en diferentes genes humanos. Sin embargo, no es sencillo elaborar un pipeline en cada laboratorio, es un proceso que requiere mucho trabajo y conocimiento de las herramientas bioinformáticas por parte del usuario y en muchos pequeños laboratorios es difícil que haya un investigador con los conocimientos suficientes. Es por ello que Jason Li et al. [40] han desarrollado una máquina virtual llamada TREVA que contiene todas las herramientas y pipelines necesarios para poder llevar a cabo numerosos análisis bioinformáticos de forma que el investigador puede introducir sus secuencias, ejecutar un comando y recibir los resultados sin necesidad de tener que realizar la instalación de las herramientas ni de elaborar el pipeline necesario. TREVA es útil por lo tanto para el presente estudio y se exploran las posibilidades que ofrece esta máquina virtual.

3.3.2 Instalación

Instalar TREVA y todos sus componentes es muy sencillo, ya que únicamente es necesario descargar la imagen virtual disponible en su página web [41]. Dado que el presente estudio va a centrarse en análisis en genoma humano, se selecciona la imagen virtual "TREVA-1-HUMAN" de 45 GB. Ocupa tanto espacio debido a que ya contiene todo el software preinstalado, además de los diversos archivos necesarios para este tipo de análisis como el genoma completo humano y bases de datos diversas.

Una vez descargada, la imagen se importa en el programa Oracle VM Virtual Box, el cual permite trabajar en máquinas virtuales sin tener que cambiar el sistema operativo original del ordenador. Con esto ya se encuentra instalado el pipeline y está listo para ser probado. TREVA trae un gran número de herramientas preinstaladas, pero no todas son necesarias para el pipeline de detección de variantes (Puesto que TREVA permite llevar a cabo otros análisis que no son relevantes en el presente estudio). El pipeline consta de las siguientes herramientas para llevar a cabo la detección de variantes:

- BWA
- Picard
- GATK

3.3.3 Detección de variantes

Debido a la naturaleza de TREVA, lo único necesario para llevar a cabo el análisis de detección de variantes son los archivos en formato fastq que contienen la información de la secuenciación (cuyos nombres han de seguir el formato _R1.fastq y _R2.fastq) y ejecutar el siguiente comando:

```
nohup runGermLine.sh -f <nombre> -s human > nohup_<nombre>.out 2>nohup_<n
ombre>.err &
```

3.3.4 Problemas con TREVA

TREVA ha sido desarrollado para poder realizar un análisis bioinformático sin la necesidad de conocer el funcionamiento interno del pipeline, algo que va en contra del objetivo del presente estudio ya que es uno de los objetivos conseguir elaborar un pipeline que realice la detección de variantes. Además, por su naturaleza TREVA realiza un análisis muy general que no se ajusta a las necesidades específicas de cada estudio, por lo que aunque se obtendrían unos resultados válidos con esta máquina virtual, no serían óptimos como en el caso de BRB-SeqTools, el cual se puede modificar y adaptar a las necesidades del investigador. En unión a los problemas de TREVA para personalizar el pipeline se encuentra también la dificultad de encontrar una solución por parte del investigador.

En conclusión, TREVA es útil para laboratorios donde no hay conocimientos de bioinformática y se han de llevar a cabo análisis en ese área, pero para el presente estudio no es óptimo por lo que se rechaza esta opción y se selecciona BRB-SeqTools como programa de partida para la elaboración del pipeline.

4. Algoritmo predictivo

4.1 Selección del genotipo

El algoritmo predictivo va a permitir obtener un valor de probabilidad de riesgo a padecer hipertensión en función del genotipo que tiene el cliente en los cinco SNPs estudiados. Los datos necesarios para desarrollar el algoritmo predictivo se encuentran en la bibliografía de los SNPs, los cuales son descritos en la tabla 9. Utilizando dichos datos y aplicándoles distintas fórmulas se obtiene el valor de probabilidad de riesgo. Por la sencillez de las fórmulas y la también sencillez del manejo de la información, se trabaja para elaborar el algoritmo predictivo en un archivo de formato Microsoft Excel Worksheet (.xlsx).

Datos	Descripción
Nombre del SNP	Nombre de cada uno de los SNPs estudiados.
Alelo de riesgo	Alelo del SNP que influye significativamente en el riesgo a padecer hipertensión.
Alelo alternativo	Alelo del SNP alternativo al alelo de riesgo que puede formar parte del genotipo.
Genotipos	Tres columnas que contienen los tres posibles genotipos formados por el alelo de riesgo y el alelo alternativo.
Input	Columna donde el investigador introduce el genotipo obtenido en el análisis de detección de variantes.
Valores β	Valores obtenidos en la bibliografía empleada, necesarios para calcular la estimación del efecto del alelo
Estimación del efecto del alelo	Esta variable mide el efecto del alelo sobre cada fenotipo. Este dato es calculado aplicando la siguiente fórmula obteniendo como resultado valores de z-score estándar: $Z_i = (\beta_{AES} - X_{AES})/\sigma_{AES}$
Descripción cualitativa de las variantes	Se define cada uno de los genotipos de las variantes según como afecta al fenotipo.

Tabla 9. Datos empleados para la elaboración del algoritmo predictivo.

Para cada uno de los SNPs conocemos su alelo de riesgo y el alelo alternativo a través de la bibliografía empleada, de forma que existen tres genotipos posibles: homocigoto de riesgo, heterocigoto y homocigoto alternativo. En el apartado de "Input" se introducen de forma manual por el investigador los genotipos obtenidos en el análisis de detección de variantes realizado.

El objetivo del algoritmo predictivo es obtener una probabilidad de riesgo a padecer la enfermedad, y el primer paso para llegar a ella es calcular las estimaciones del efecto de los alelos y para ello han de obtenerse los denominados z-scores de los distintos genotipos de las variantes estudiadas.

4.2 Obtención de valores de z-score

Lo primero que se necesita para para conocer el z-score son los denominados valores de β para cada uno de los genotipos. Estos valores varían según el genotipo de forma que para homocigoto alternativo toman un valor de 0, para heterocigotos toman el valor obtenido en la bibliografía para cada uno de los SNPs (denominado AE) y para homocigotos de riesgo toman el valor de $2 * AE$. De esta forma se tiene para cada SNP tres valores de β, uno para cada genotipo, necesarios para z-score. En la tabla 10 se incluye un ejemplo de valores β calculados para los diferentes SNPs. β_{alt} son los valores de β para el genotipo homocigoto alternativo, β_{het} son los valores de β para el

genotipo heterocigoto y β_{rie} son los valores de β para el genotipo homocigoto de riesgo.

SNP	β_{alt}	β_{het}	β_{rie}
rs00000001	0,00	-0,12	-0,24
rs00000002	0,00	0,06	0,12
rs00000003	0,00	0,10	0,20
rs00000004	0,00	-0,05	-0,10
rs00000005	0,00	0,07	0,14

Tabla 10. Tabla con valores β de ejemplo (SNPs modificados para ejemplo).

Una vez se tienen todos los valores β se calcula X_{AES}, que es el valor promedio de todos los valores β calculados. Este valor es de vital importancia para el cálculo de los z-score. En el presente ejemplo, el valor promedio es $X_{AES} = 0.012$. También se calcula σ_{AES}, que es la desviación estándar de todos los valores β calculados. En el presente ejemplo, la desviación estándar es $\sigma_{AES} = 0.108$

Una vez se tienen los valores de β, X_{AES} y σ_{AES} se pueden calcular los valores de z-score que servirán para la estimación del efecto del alelo. La fórmula empleada (ya mencionada en la tabla 9) es la siguiente:

$$Z_i = \frac{\beta_i - X_{AES}}{\sigma_{AES}}$$

1. "β_i" es es valor de β para el SNP y el genotipo concreto del que se quiere obtener el z-score.

2. "X_{AES}" es el valor promedio de todos los valores β calculados.

3. "σ_{AES}" es un valor fijo al igual que X_{AES}, es la desviación estándar de todos los valores β calculados.

SNP	Z_{alt}	Z_{het}	Z_{rie}
rs00000001	-0.111	-1.223	-2.334
rs00000002	-0.111	0.445	1.000
rs00000003	-0.111	0.815	1.741
rs00000004	-0.111	-0.574	-1.037
rs00000005	-0.111	0.537	1.186

Tabla 11. Tabla con valores Z_i de ejemplo.

De esta forma ya se tienen los 15 valores de z-score (para cada uno de los cinco SNPs hay tres genotipos, por lo que $3 * 5 = 15$) que se utilizan como estimadores del riesgo y tienen en cuenta el efecto de cada alelo sobre el fenotipo.

4.3 Probabilidad de riesgo

En este punto se conocen los diferentes genotipos de las variantes los sus estimadores de riesgo asociados a cada uno de ellos (Z_i). El siguiente paso consiste en obtener la probabilidad de riesgo seleccionando para cada una de las variantes el valor de Z_i correspondiente al genotipo del cliente. Para ello se utiliza la columna de "Input" en la cual se han introducido los genotipos del cliente obtenidos en el análisis de detección de variantes. Con el siguiente código de Excel se consigue que para cada SNP se seleccione el valor de z-score asociado al genotipo del cliente:

```
=IF(<casilla input>=<casilla genotipo 1>;<casilla z-score genotipo 1>;
(IF(<casilla input>=<casilla genotipo 2>;<casilla z-score genotipo 2>;
(IF(<casilla input>=<casilla genotipo 3>;<casilla z-score genotipo 3>;"NA
")))))
```

Una vez se tienen los cinco valores de z-score asociados al genotipo del cliente, se aplica la siguiente fórmula para obtener la probabilidad de riesgo genético:

$$P = \frac{e^s}{1 + e^s}$$

Siendo s el valor de la suma de los cinco valores de z-score seleccionados.

En la tabla 12 se inventan unos resultados de un cliente, de forma que se incluyen sus genotipos y los z-scores asociados a dichos genotipos. Aplicando la fórmula a dichos z_scores, se calcula la probabilidad de riesgo genético para este cliente hipotético, que toma un valor de $P = 73.47\%$.

SNP	Homocigoto alternativo	Heterocigoto	Homocigoto de riesgo	Genotipo cliente	z-score cliente
rs00000001	CC	CT	TT	CT	-1.223
rs00000002	CC	CT	TT	CC	-0.111
rs00000003	GG	GA	AA	AA	1.741
rs00000004	AA	AT	TT	AT	-0.574
rs00000005	CC	CA	AA	AA	1.186

Tabla 12. Ejemplo de genotipo de un cliente inventado y selección de z-scores. Los SNPs y sus genotipos son inventados para el ejemplo y no se corresponden con los reales por mantener la confidencialidad.

Esta probabilidad de riesgo genético es el resultado final y objetivo del algoritmo predictivo, cuyo significado e importancia se verá plasmado en el informe genético entregado al cliente, que podrá conocer su riesgo a padecer hipertensión en función de sus variantes genéticas.

4.4 Descripción cualitativa de la probabilidad

Ya se posee un valor de probabilidad de riesgo a padecer hipertensión, sin embargo, es necesario saber interpretar dicho valor y presentarlo de tal forma que el cliente pueda comprender el significado del mismo aunque no tenga conocimientos previos sobre genética.

Un valor del 50% para el riesgo a padecer hipertensión significa que debido a sus variantes genéticas en los SNPs estudiados, el cliente tiene un riesgo a padecer la enfermedad idéntico al valor medio de la población. Porcentajes más bajos implican un riesgo menor que la media de la población a padecerla y porcentajes más elevados implican un mayor riesgo.

Para una mayor simplicidad y ser entendido de forma más directa, se establecen unos intervalos de riesgo:

- BAJO (<30%): Debido a sus variantes genéticas, el cliente tiene mucha menos probabilidad a padecer hipertensión que el resto de la población general. Aunque debe mantener un estilo de vida saludable para evitar padecer la enfermedad, no es necesario llevar a cabo precauciones de prevención.
- MEDIO (30%-60%): El cliente posee una selección de variantes genéticas que le otorga un riesgo a padecer la enfermedad similar al de la media de la población general.
- MODERADO (60%-80%): Debido a sus variantes genéticas, el cliente tiene un riesgo ligeramente elevado a padecer la enfermedad en relación a la población general. Ha de seguir un estilo de vida saludable y seguir unas pautas y hábitos de vida para compensar su riesgo genético y reducir las probabilidades de padecer hipertensión.
- ALTO (>80%): Debido a sus variantes genéticas, el cliente tiene un elevado riesgo a padecer hipertensión con respecto a la población general. Para tratar de evitar al máximo sufrir esta enfermedad y contrarrestar el riesgo genético, el cliente ha de seguir de forma estricta las pautas y recomendaciones dadas, llevando un estilo de vida saludable y personalizado.

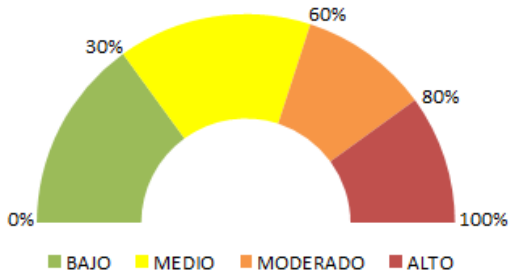

***Figura 7**. Valores de riesgo genético.*

Dichos valores se presentan al cliente para que sepa de forma rápida y sencilla su predisposición genética. Además, en base a la bibliografía empleada para la elaboración del algoritmo predictivo podemos también describir los diferentes genotipos de los cinco SNPs estudiados de tres maneras diferentes:

- Riesgo: El genotipo de esta variante es causante del aumento del riesgo a padecer hipertensión.
- Neutro: El genotipo de esta variante no tiene un efecto significativo en el riesgo a padecer hipertensión.
- Protector: El genotipo de esta variante es beneficioso para el individuo de forma que contribuye a reducir el riesgo a padecer hipertensión.

Gracias a esta descripción cualitativa de las variantes, los clientes pueden ver desglosadas sus variantes genéticas y cuales son las que tienen un mayor efecto en sus resultados.

5. Informe genético

Gracias al pipeline bioinformático desarrollado ha sido posible obtener la información deseada, en el caso del presente estudio es conocer los genotipos de cinco SNPs concretos del genoma a partir de datos de secuenciación de Illumina. Con dichos genotipos y una base de datos con extensa información obtenida a partir de la bibliografía sobre los SNPs deseados se ha aplicado un algoritmo predictivo cuyo resultado ha sido un valor de probabilidad de riesgo a padecer hipertensión.

Para presentar los resultados al cliente se elabora un informe genético que muestra y describe de forma detallada sus variantes genéticas y lo que significa tenerlas, de forma que una persona sin conocimiento previo sobre genética es capaz de entender que significan sus variantes.

Junto a esta información se presenta una lista de pautas y hábitos de vida y nutricionales personalizados que el cliente ha de tener en cuenta y seguir con rigor para mejorar su calidad de vida y poder afrontar y contrarrestar su posible riesgo a sufrir hipertensión.

6. Conclusiones

En el comienzo de este proyecto, mi conocimiento de las herramientas y el flujo de trabajo necesarios para manipular datos de NGS era prácticamente nulo. En un comienzo con mucha ayuda conseguí el suficiente impulso para comenzar a entender la idea detrás de un pipeline bioinformático y las herramientas más conocidas y utilizadas en ellos.

Hasta mi llegada al máster, no tenía apenas conocimiento de herramientas bioinformáticas y no sabía ni utilizar el terminal de Unix. Este proyecto en conjunción

con otras asignaturas me han ayudado de gran manera a desenvolverme con facilidad en estos entornos, habilidad imprescindible en este campo que es la bioinformática.

Si bien he contado con el apoyo de mis compañeros/as de trabajo para resolver mis dudas de estudiante, es evidente que es necesario aprender a resolver las dudas de forma autosuficiente y aunque es una habilidad que llevo desarrollando durante toda mi vida como estudiante, este proyecto me ha motivado a aprender a resolver mis dudas en lo referente a herramientas bioinformáticas y al ámbito de detección de variantes. He descubierto diferentes foros especializados que me han servido de gran ayuda para mis consultas, tanto en la búsqueda de soluciones a mis preguntas como en búsqueda de guías, definiciones de términos... Y es uno de los puntos de mi desarrollo como bioinformático que más valoro de este proyecto, la expansión de fuentes de conocimiento de las que voy a poder seguir aprendiendo incluso tras haber finalizado el presente proyecto.

En lo referente al cumplimiento de los objetivos, mi valoración personal es positiva ya que se ha conseguido abarcar la totalidad de los mismos, una tarea demasiado complicada a mi juicio al comienzo del proyecto, pero que gracias al apoyo de mis compañeros/as y a un horario de trabajo más amplio que el semestre pasado ha sido posible.

La planificación no ha sido seguida de forma meticulosa, dado que fue planteada inicialmente según la idea que se tenía que iban a durar las diferentes tareas y se plantearon en serie, es decir tener que terminar una para pasar al siguiente punto. En la práctica, los saltos entre diferentes objetivos se han dado constantemente, de forma que ningún objetivo realmente se comenzó al terminar el que le precedía. Este planteamiento de objetivos en serie no ha representado con fidelidad la planificación seguida, sin embargo en rasgos generales si que se ha planteado de forma y orden correcto, pero la duración de cada uno de los puntos ha sido un poco arbitraria, en gran parte por el desconocimiento de la materia en el comienzo que me impidió poder establecer unos intervalos de tiempo precisos y necesarios. Sin embargo, gracias a haber comenzado el proyecto con antelación y trabajando de forma constante, no ha habido ningún problema de tiempo y todos los objetivos se han llevado a cabo con previsión y el tiempo necesario.

Mi valoración personal del proyecto es altamente positiva. Si bien cuando me matriculé en el máster de Bioinformática y Bioestadística lo hice en base a mi contacto con la bioinformática en mis estudios de grado, este campo se me presentaba todavía muy amplio y desconocido, de forma que no había profundizado en ninguna de sus ramas. En el desarrollo de este proyecto he conocido el desarrollo de pipelines, el uso de herramientas de amplio uso en el campo y la aplicación de las mismas para dar un servicio de carácter genético, de los cuales he obtenido una gran cantidad de conocimientos que me van a ser muy útiles en mi carrera profesional.

7. Glosario

- Algoritmo predictivo: Conjunto ordenado de operaciones que permiten generar un valor de riesgo a parecer la enfermedad estudiada en función del genotipo.
- Archivo BAM: Binary Alignment Map, archivo en formato binario para almacenar grandes cantidades de alineamientos de secuencias de nucleótidos de forma binaria.
- Archivo GVCF: Genomic Variant Calling Format, archivo en formato especial usado por GATK para almacenar información sobre la llamada de variantes genómicas.
- Archivo FASTQ: Archivo en formato FASTQ para almacenar datos de secuenciación masiva.
- Archivo SAM: Sequence Alignment Map, archivo en formato genérico para almacenar grandes cantidades de alineamientos de secuencias de nucleótidos.
- Archivo VCF: Variant Calling Format, archivo en formato usado por GATK para almacenar información sobre la llamada de variantes genómicas.
- BRB-SeqTools: Herramienta capaz de crear un pipeline aplicado a datos de NGS.
- BWA: Burrows-Wheeler Aligner, un software para mapear secuencias poco divergentes frente a grandes secuencias de referencia como genomas.
- Enfermedad poligénica: Enfermedad que se produce debido a la combinación de diferentes mutaciones en diferentes genes.
- Fenotipo: Expresión del genotipo en función de un determinado ambiente
- FS: Fisher Strand, tendencia a que una de las hebras del DNA se vea favorecida sobre otra en la evaluación del análisis.
- GATK: Genome Analysis ToolKit, software que ofrece una amplia variedad de herramientas para la detección de variantes genéticas.
- Genotipo: Información genética sobre la presencia de variantes en el ADN estudiado.
- Illumina: Compañía estadounidense que desarrolla, fabrica y comercializa sistemas integrados para el análisis de variación genética y función biológica.
- Indel: INserción/DELeción, tipo de mutación genética que abarca las inserciones y deleciones de bases en el ADN.
- Informe genético: Documento final que llega al cliente con toda la información genética obtenida de la secuenciación de su ADN y el riesgo a padecer la enfermedad estudiada.
- Máquina virtual: Software que simula a una computadora y puede ejecutar programas como si fuese una computadora real.
- MiSeq: Secuenciador de la plataforma Illumina.
- MQ: Mapping Quality, estimación de la calidad del mapeado de las lecturas de la variante detectada.
- MQRankSum: Resultado de la aplicación de "Rank Sum Test" a los datos de MQ.

- NGS: Next Generation Sequencing, tecnología de secuenciación del ADN de gran potencia y que está superando a tecnologías anteriores de secuenciación tanto en eficiencia como en precio.
- Picard: Herramienta que permite trabajar con datos de NGS en formato SAM, BAM y VCF.
- Pipeline: Cadena de procesos elementales de forma que el resultado del primer proceso es el material con el que comienza a trabajar el siguiente.
- Primers: Hebra corta de ADN que sirve como punto de inicio de síntesis de ADN.
- QD: Quality by Depth, valor de calidad medio entre todas las lecturas de la variante estudiada.
- ReadPosRankSum: Resultado de la aplicación de "Rank Sum Test" a los datos de tendencia de desplazamiento de la posición de las variantes en las lecturas de la variante estudiada.
- Samtools: Software que ofrece gran cantidad de herramientas para trabajar con datos de NGS y detección de variantes genéticas.
- SNP: Single Nucleotide Polymorphism, variación en la secuencia de ADN que afecta a una sola base de una secuencia del genoma.
- TREVA: Targeted REsequencing Virtual Appliance, máquina virtual que contiene de forma base un conjunto de herramientas necesarias para el estudio de detección de variantes genéticas.
- Variante genética: Variación las bases de una secuencia en el ADN con respecto a una secuencia de referencia, por ejemplo la variación entre la secuencia del ADN de un individuo con respecto a la secuencia del genoma de referencia del ser humano.

8. Bibliografía

1. Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., ... & Dewell, S. B. (2005). Genome sequencing in microfabricated high-density picolitre reactors. Nature, 437(7057), 376-380.

2. Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., ... & Boutell, J. M. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. nature, 456(7218), 53-59.

3. Fedurco, M., Romieu, A., Williams, S., Lawrence, I., & Turcatti, G. (2006). BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. Nucleic acids research, 34(3), e22-e22.

4. Shendure, J., Porreca, G. J., Reppas, N. B., Lin, X., McCutcheon, J. P., Rosenbaum, A. M., ... & Church, G. M. (2005). Accurate multiplex polony sequencing of an evolved bacterial genome. Science, 309(5741), 1728-1732.

5. Horner, D. S., Pavesi, G., Castrignanò, T., De Meo, P. D. O., Liuni, S., Sammeth, M., ... & Pesole, G. (2010). Bioinformatics approaches for genomics and post genomics

applications of next-generation sequencing. Briefings in bioinformatics, 11(2), 181-197.

6. Richter, B. G., & Sexton, D. P. (2009). Managing and analyzing next-generation sequence data. PLoS Comput Biol, 5(6), e1000369.

7. Koboldt, D. C., Ding, L., Mardis, E. R., & Wilson, R. K. (2010). Challenges of sequencing human genomes. Briefings in bioinformatics, 11(5), 484-498.

8. Talkowski, M. E., Ernst, C., Heilbut, A., Chiang, C., Hanscom, C., Lindgren, A., ... & Shen, Y. (2011). Next-generation sequencing strategies enable routine detection of balanced chromosome rearrangements for clinical diagnostics and genetic research. The American Journal of Human Genetics, 88(4), 469-481.

9. Durbin, R. M., Abecasis, G. R., Altshuler, D. L., Auton, A., Brooks, L. D., Gibbs, R. A., ... & 1000 Genomes Project Consortium. (2010). A map of human genome variation from population-scale sequencing. Nature, 467(7319), 1061-1073.

10. Smith, D. R., Quinlan, A. R., Peckham, H. E., Makowsky, K., Tao, W., Woolf, B., ... & Stewart, D. A. (2008). Rapid whole-genome mutational profiling using next-generation sequencing technologies. Genome research, 18(10), 1638-1642.

11. Shen, Y., Wan, Z., Coarfa, C., Drabek, R., Chen, L., Ostrowski, E. A., ... & Yu, F. (2010). A SNP discovery method to assess variant allele probability from next-generation resequencing data. Genome research, 20(2), 273-280.

12. Dressman, D., Yan, H., Traverso, G., Kinzler, K. W., & Vogelstein, B. (2003). Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. Proceedings of the National Academy of Sciences, 100(15), 8817-8822.

13. Dalca, A. V., & Brudno, M. (2010). Genome variation discovery with high-throughput sequencing data. Briefings in bioinformatics, 11(1), 3-14.

14. Harismendy, O., Ng, P. C., Strausberg, R. L., Wang, X., Stockwell, T. B., Beeson, K. Y., ... & Frazer, K. A. (2009). Evaluation of next generation sequencing platforms for population targeted sequencing studies. Genome biology, 10(3), R32.

15. Yohe, S., Hauge, A., Bunjer, K., Kemmer, T., Bower, M., Schomaker, M., ... & Deshpande, A. (2015). Clinical validation of targeted next-generation sequencing for inherited disorders. Archives of Pathology and Laboratory Medicine, 139(2), 204-210.

16. Rehm, H. L., Bale, S. J., Bayrak-Toydemir, P., Berg, J. S., Brown, K. K., Deignan, J. L., ... & Working Group of the American College of Medical Genetics. (2013). ACMG clinical laboratory standards for next-generation sequencing. Genetics in Medicine, 15(9), 733-747.

17. Mitchell, K. J. (2012). What is complex about complex disorders?. Genome biology, 13(1), 237.

18. http://www.eupedia.com/genetics/heart_disease_snp.shtml Febrero 2017.

19. http://www.ine.es/jaxiT3/Tabla.htm?t=7947 Febrero 2017.

20. http://www.ine.es/jaxi/Tabla.htm?path=/t15/p414/a2015/l0/&file=01001.px&L=0 Febrero 2017.

21. https://www.msssi.gob.es/estadEstudios/estadisticas/sisInfSanSNS/tablasEstadisticas/SaludSistemaSanitario_100_Tablas1.pdf Febrero 2017.

22. http://www.who.int/dietphysicalactivity/publications/trs916/summary/en/ Febrero 2017.

23. https://www.hcup-us.ahrq.gov/reports/statbriefs/sb204-Most-Expensive-Hospital-Conditions.pdf Febrero 2017.

24. http://www.ine.es/jaxi/Tabla.htm?path=/t15/p414/a2015/l0/&file=01013.px&L=0 Febrero 2017.

25. Reddy, K. S., & Katan, M. B. (2004). Diet, nutrition and the prevention of hypertension and cardiovascular diseases. Public health nutrition, 7(1A; SPI), 167-186.

26. Newton-Cheh, C., Johnson, T., Gateva, V., Tobin, M. D., Bochud, M., Coin, L., ... & Papadakis, K. (2009). Eight blood pressure loci identified by genome-wide association study of 34,433 people of European ancestry. Nature genetics, 41(6), 666.

27. International Consortium for Blood Pressure Genome-Wide Association Studies. (2011). Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. Nature, 478(7367), 103-109.

28. Johnson, T., Gaunt, T. R., Newhouse, S. J., Padmanabhan, S., Tomaszewski, M., Kumari, M., ... & Sever, P. (2011). Blood pressure loci identified with a gene-centric array. The American Journal of Human Genetics, 89(6), 688-700.

29. Levy, D., Ehret, G. B., Rice, K., Verwoert, G. C., Launer, L. J., Dehghan, A., ... & Aulchenko, Y. (2009). Genome-wide association study of blood pressure and hypertension. Nature genetics, 41(6), 677-687.

30. https://support.illumina.com/help/SequencingAnalysisWorkflow/Content/Vault/Informatics/Sequencing_Analysis/CASAVA/swSEQ_mCA_FASTQFiles.htm Marzo 2017.

31. https://support.illumina.com/help/BaseSpace_OLH_009008/Content/Source/Informatics/BS/QualityScoreEncoding_swBS.htm Marzo 2017.

32. https://brb.nci.nih.gov/seqtools/download_seqtools.html Marzo 2017.

33. http://bio-bwa.sourceforge.net/ Marzo 2017.

34. http://samtools.sourceforge.net/ Marzo 2017.

35. https://broadinstitute.github.io/picard/ Marzo 2017.

36. https://software.broadinstitute.org/gatk/best-practices/ Marzo 2017.

37. https://software.broadinstitute.org/gatk/ Marzo 2017.

38. http://gatkforums.broadinstitute.org/gatk/discussion/2806/howto-apply-hard-filters-to-a-call-set Marzo 2017.

39. Bodi, K., Perera, A. G., Adams, P. S., Bintzler, D., Dewar, K., Grove, D. S., ... & Singh, S. (2013). Comparison of commercially available target enrichment methods for next-generation sequencing. J Biomol Tech, 24(2), 73-86.

40. Li, J., Doyle, M. A., Saeed, I., Wong, S. Q., Mar, V., Goode, D. L., ... & Hunter, S. M. (2014). Bioinformatics pipelines for targeted resequencing and whole-exome sequencing of human and mouse genomes: a virtual appliance approach for instant deployment. PloS one, 9(4), e95217.

41. http://bioinformatics.petermac.org/treva/ Marzo 2017.