Incorporación de un *pipeline* informático para la identificación de mutaciones en secuenciación de DNA tumoral

Autora: Valentina | Lorenzo Camacho

Tutor: Eduardo | Salido

Cotutora: Ayelen | Ramos Navas

Departamento de ciencias médicas básicas, Servicio de Anatomía Patológica del Hospital Universitario de Canarias

Universidad de La Laguna

Agradecimientos

Este TFG se lo dedico a todas las personas fracasadas (o más bien, a todas aquellas que fracasan). Se lo dedico a todas las personas que no aprobaron a la primera, que no supieron contestar las preguntas de los doctores en las prácticas del hospital, a todas aquellas que sudaron por no cumplir los requisitos de la beca y a las que se tuvieron que presentar a Tribunal con el corazón en un puño.

Este TFG me gustaría dedicárselo también a Eduardo Salido y Ayelen Ramos, dos personas que entienden y empatizan con el fracaso, viéndolo como una oportunidad para la reconstrucción y la exploración de la Medicina desde vías alternas.

Finalmente, me gustaría dedicarle este TFG a todas las activistas que, acampando, están luchando contra la construcción del Macroproyecto Hotelero de Cuna del Alma. Este TFG ha sido realizado en el lapso de tiempo que ha durado esta reivindicación.

"Todas somos quimeras, híbridas teorizadas y fabricadas de máquina y organismo; en definitiva, somos ciborgs."
Donna Haraway

Índice

Resumen:

En el estudio de mutaciones del genoma, las herramientas desarrolladas en el ámbito de la bioinformática han marcado un antes y después, facilitando la rápida detección de variaciones en el ADN y favoreciendo el ejercicio de la medicina personalizada dentro del ámbito de la oncología.

En este trabajo se ha hecho un estudio experimental con la finalidad de aprender sobre las diferentes herramientas bioinformáticas para el análisis de genoma humano y la detección de variantes o mutaciones en el mismo. Para ello, se utilizaron muestras de ADN y ARN de pacientes anonimizados (identificados por códigos). Estas muestras fueron aportadas y secuenciadas por el servicio de Anatomía Patológica del Hospital Universitario de Canarias. En ellas, se detectaron mutaciones o variaciones concordantes con diferentes patologías.

En la actualidad, la bioinformática a pesar de aportar una mejoría exponencial en la detección y entendimiento de las patologías, sigue siendo una subespecialidad de “nicho”, principalmente debido a su larga curva de aprendizaje. Sin embargo, es necesario dedicar esfuerzos a la formación y la divulgación de la bioinformática entre los profesionales de la salud, ya que esta subespecialidad tiene una previsión de desarrollo muy grande y podría ser indispensable a la hora de hacer más eficiente el diagnóstico y la terapéutica médica.

Abstract

In the study of genome mutations, the tools developed in the field of bioinformatics have marked a turning point, facilitating the rapid detection of DNA alterations and favoring the exercise of personalized medicine in the field of oncology.

In this work, an experimental study has been carried out with the aim of learning about the different bioinformatic tools for the analysis of the human genome and the detection of variants or mutations in tumors. For this purpose, anonymized (code-identified) DNA and RNA samples from breast cancer patients were used. These samples were provided and sequenced by the Pathology Department of the Hospital Universitario de Canarias. In them, mutations or pathogenic variations that conditioned the different pathologies were detected.

At present, bioinformatics, despite providing an exponential improvement in the detection and understanding of disease, continues to be a "niche" subspecialty, mainly due to its long learning curve. However, it is necessary to devote efforts to the training and dissemination of bioinformatics among health professionals, in order to create an even stronger and more interconnected international network.

Palabras clave

Bioinformática, secuenciación, alineación, Galaxy, Ensembl, BWA, Samtools, Trimomatic, Cáncer, mutaciones, variaciones, Linux

Introducción

¿Qué es la bioinformática?

Es una subdisciplina científica que involucra el uso de tecnología informática para recoger, organizar, acumular y analizar información biológica. En este TFG se analizará específicamente genoma humano para la detección de variaciones significativas y consecuentemente ayudar al diagnóstico del cáncer, particularmente en la orientación de una terapia dirigida a dianas concreta. [7]

¿Qué es un pipeline?

En informática, un pipeline (“*tubería*” en inglés) es la combinación de procesos informáticos con un orden definido que están involucrados en la ejecución de los *flujos de trabajo* o *workflow* (aunque se puede utilizar también como sinónimo de *flujo de trabajo*) de forma rutinaria y altamente automatizada. Estos flujos de trabajo no son más que un conjunto de tareas individuales. En este gráfico adjunto, podemos ver un ejemplo de lo mencionado:

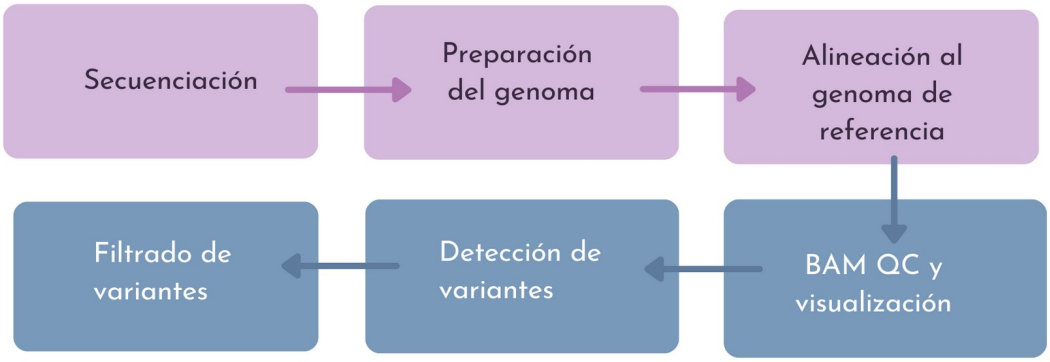


Figura 1*: Creación propia*

Como podemos ver, en esta Figura se enumeran los procedimientos necesarios para la consecución del objetivo de este TFG (la identificación de variantes para el diagnóstico patólogico). En este caso, cada "flecha" simbolizaría al pipeline, siendo este el conjunto de tareas (p.ej: indexar genoma) que nos permitirían ejecutar el proceso nombrado y nos darían los datos para la ejecución del siguiente, siendo el resultado de un proceso *(output)* el elemento de entrada *(input)* del siguiente proceso. [2] [3]

¿Cómo se incorpora el pipeline ?

A continuación, se describirán más detalladamente los procesos del pipeline que se ejecutarán siguiendo el esquema de la *Figura 1*. En este cuadro resumen se nombran algunas de las herramientas que se pueden usar y en negrita están los usados en este trabajo:

Proceso	Herramienta (s)
Secuenciación	**Illumina**
Preparación del Genoma	**FASTQ**, **Trimmomatic**, FastX Toolkit, Cutadapt
Alineación al genoma de referencia	**BWA**, **Galaxy**
BAM QC y visualización	**Samtools**, **Qualimap**, Freebayes, IGV
Detección de variantes	**VarScan**, **GATK**, Freebayes
Filtrado de variantes	**Ensembl VEP**, **GATK**, VCFtools

Secuenciación

La secuenciación se define como la determinación exacta de nucleótidos y bases en una molécula de ADN. Este proceso puede hacerse de varias maneras. En este TFG, la secuenciación se hace mediante síntesis por polimerasas de lecturas cortas del genoma mediante Illumina, proceso denominado *next-generation sequencing* (NGS) en contraposición con la forma original de secuenciar *Sanger sequencing.*

El resultado final de esta secuenciación será el archivo digital que será utilizado en los procesos posteriores [3][4][5]

Preparación del genoma

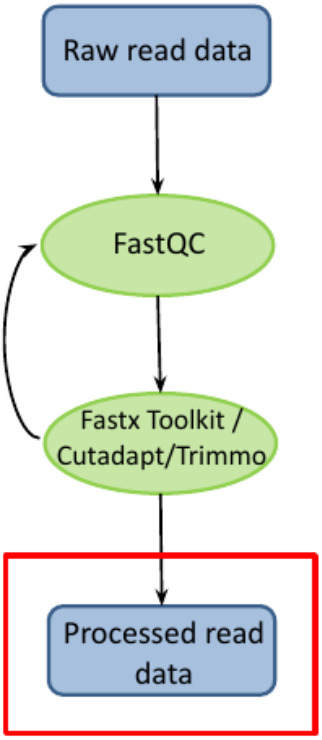


Figura 2: *Imagen resumen del proceso de preparación del genoma[1]*

Para poder trabajar con el archivo, hace falta primero pasar por el proceso que en este TFG denominaremos "de preparación". Este consiste en varios pasos:

1. Estudio de la calidad de la muestra: para esto se utilizará FASTQC, una herramienta de estudio de control de calidad frecuentemente utilizada en la bioinformática.
2. "Recorte" o *trimming* de las partes del genoma que no tengan buena calidad, manteniendo aquellas que sean buenas para su estudio. El "recorte" se ejecutará con Trimmomatic, una herramienta especializada que se puede utilizar tanto en la terminal como en la web.

A la hora de representar la calidad de nucleótidos concretos se utiliza el código ASCII, que permite asociar un valor específico a cada nucleótido sin comprometer la correspondencia de los caracteres, como se muestra a continuación:

```
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHI
Q0        Q10        Q20        Q30        Q40
```

Si este código no se usara y tuviéramos una secuencia tipo ACGT con una calidad determinada Q de 25(A), 34(C), 40(G) y 35(T) a la hora de incorporarla a un archivo de dos líneas relacionando nucleótidos y calidad quedaría de la siguiente manera:

```
ACGT
25344035
```

Como se percibe en el ejemplo anterior, no se mantendría la correspondencia de los caracteres. Sin embargo, con el código ASCII (! = Q0 y hasta I = Q40) la secuencia se leería de la siguiente manera:

```
ACGT
:CID
```

De esta manera se puede asignar la calidad debajo de la secuencia consumiendo un solo carácter sin que se pierda la correspondencia y haciendo más fácil su lectura.

Una vez identificada la calidad de las bases, hay que proceder al filtrado mencionado. Se considerará de "mala calidad" cualquier base que tenga un valor de calidad < Q30. Estos valores de calidad se definen como la probabilidad de error: $Q(A) = -10 \log_{10}(P(\sim A))$. Este criterio de calidad (originalmente propuesto para el uso del software Phred como parte de la bioinformática necesaria para ensamblar el proyecto Genoma Humano) corresponde a aceptar un nucleótido concreto en un punto de la secuenciación sólo si la probabilidad de que sea erróneo es inferior a 1/1000. [1][6]

Quality score, Q(A)	Error probability, P(~A)
10	0.1
20	0.01
30	0.001

***Figura 3**: Relación entre Q y probabilidad de error[6]*

Alineamiento al genoma de referencia

Para poder detectar variaciones, es importante comparar el genoma obtenido con otro de "referencia". A esto se llamará *alineamiento* o *mapeado* y consiste en definir (o hacer un *mapa* de) la ubicación más probable dentro del genoma para la lectura de ADN observada, además de identificar la secuencia genómica que mejor se ajusta a cada secuencia de referencia *(Figura)* . Una vez alineado, será transformado en un archivo de *Sequence Alignment Map* (SAM) o mapa de secuencias alineadas. Este alineamiento se hará en base al tipo de genoma que empleen los paneles de secuenciación del secuenciador. En este caso el genoma de referencia es el GRCh38. [1]

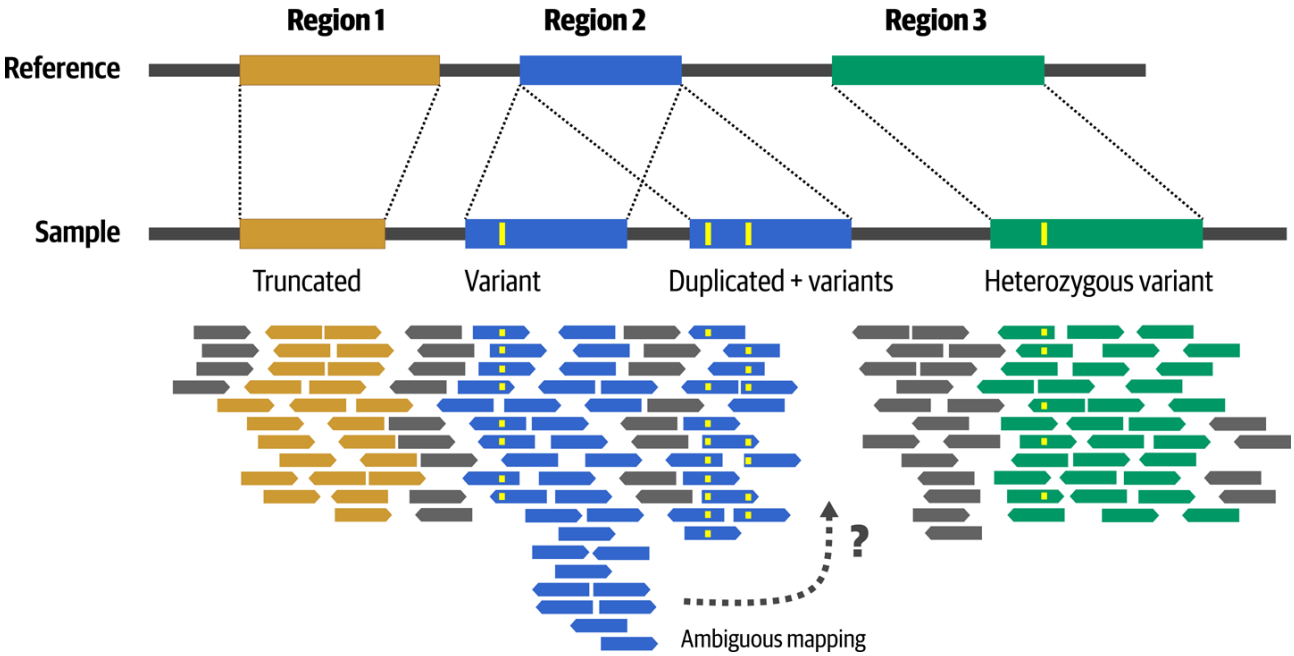


***Figura 4**: En esta imagen se describe el proceso de mapeado. Como podemos ver, el genoma es "organizado" en base al genoma de referencia, detectando variantes. A veces, si el genoma es antiguo, se pueden detectar variantes duplicadas ancestrales. [3]*

Detección y filtrado de variaciones

En este paso, se identifican potenciales variantes a través de la búsqueda de evidencia de un tipo particular de variación en los datos de secuenciación disponibles. El proceso es bastante simple: se produce una pila de lecturas, se comparan con la secuencia de referencia y se identifica dónde hay desajustes. El formato estándar que se usa es VCF, que es el *Variant Call Format* o formato de llamada de variaciones.

Habrá diferentes tipos de variaciones. Utilizaremos de ejemplo la frase "Esta variación me da magua" para facilitar el entendimiento de sus diferencias. Todas ellas se describen en base a su relación con el genoma de referencia. En este trabajo nos centraremos en la identificación de variaciones de un solo nucleótido.

Tipo de variación	Descripción	Ejemplo
Variaciones de un solo nucleótido	Una sola base se intercambia por otra	Esta variación me da **P**agua
Inserción y deleciones	Una o varias bases se añaden o eliminan	Esta _____ me da **mucha** magua
Variaciones del número de copias	Una o varias bases se repiten	Esta **variación** me da mucha magua **variación**

Una vez las variaciones son detectadas, se procederá a un proceso de “filtrado” basado en parámetros previamente delimitados que permitirán seleccionar las variantes a considerar. En este trabajo, estas serán las variantes que tengan >300 lecturas y >10% de frecuencia de presentación. Estos valores son elegidos de manera aleatoria y principalmente se establecen con la intención de evitar el marco de error que puede producir la lectura del genoma, ya que a partir en el el proceso de fijación del formaldehído con el ADN introduce cambios químicos en el ADN interpretables como mutaciones aleatorias (con más frecuencia C>T) que pueden ser amplificadas por PCR . Otro dato que se tiene en cuenta a la hora de estipular estos valores, es el Q. Teniendo en cuenta que Q30 = probabilidad de error de 1 de cada 1000 lecturas, identificar las variantes con >300 lecturas supondría permitir, según la ecuación $Q(A) = -10 \log_{10}(P(\sim A))$ un Q de aproximadamente 25, posicionándose en la mitad entre Q30 y Q20 y permitiendo así un margen a la hora de identificar variantes.
[1]

¿Por qué usar Software libre y GNU/Linux?

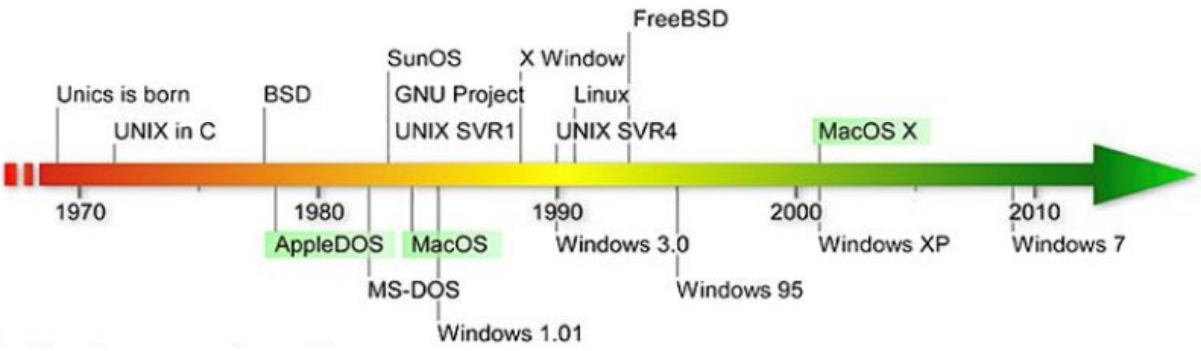


***Figura 5**: Evolución de los sistemas operativos*

El software libre es un software computacional presentado en términos que permiten a usuarios ejecutar software para cualquier objetivo, de la misma manera que permite estudiarlo, cambiarlo y distribuirlo. Estas características son importantes para el desarrollo científico ya que al ser de uso, modificación y contribución libre, no atan el desarrollo científico a empresas o instituciones particulares. Además, permite escribir capas de compatibilidad abiertas (como *drivers*) para hardware específico, como pueden ser secuenciadores genéticos.

El principal sistema operativo que trabaja con software libre es GNU/Linux. Este sistema operativo es utilizado principalmente en líneas de estudios donde es necesario el manejo de conjuntos de datos muy grandes y la ejecución de procesos que consumen muchos recursos, ya que permite paralelizar estos procesos y hacerlos más eficientes y rápidos a través de clusters y nubes de alto rendimiento que muchas veces están interconectados con múltiples sistemas GNU/Linux. Este sistema operativo aporta, en comparación con otros:

1. La posibilidad de trabajar con grandes cantidades de datos de manera muy eficiente a través del uso de clusters y nubes específicos a GNU/Linux.
2. Estabilidad, seguridad, ajustabilidad y precios asequibles.
3. Una mayor facilidad para usar aplicaciones que no son de escritorio. Este tipo de aplicaciones son importantes porque permiten flexibilidad y libertad a la hora de ejecutar tareas , posicionándolo como la única verdadera opción para uso científico a gran escala.

Además, el uso de software libre permite:

4. Personalizar las tareas ejecutadas de los programas y los programas mismos.
5. Automatizar gran parte del análisis y mantener registros precisos y detallados de cada paso, lo que proporciona beneficios en:
    a. La reproducibilidad y el reanálisis
    b. El entendimiento de problemas y los fallos de los mismos.

Todo esto ha propiciado el posicionamiento de GNU/Linux como el sistema operativo por excelencia en la bioinformática, haciendo que el software desarrollado para el análisis del genoma se enfoque a este sistema operativo. Por lo tanto, es difícil y limitante el trabajar con genoma sin tener que usar en algún punto del proceso GNU/Linux. [8][9][10][11]

Hipótesis del trabajo

El aprendizaje de herramientas bioinformáticas necesarias para la identificación de mutaciones tras la secuenciación masiva de ADN de tumores se puede incorporar en un programa de formación de médicos / residentes de Anatomía Patológica y contribuye a la implantación de terapias dirigidas y medicina personalizada.

Justificación del trabajo

Con el desarrollo de la secuenciación masiva y su implementación en las técnicas diagnósticas aplicadas a tumores, la bioinformática es una subespecialidad que no solo se ha desarrollado exponencialmente en los últimos años, sino que tiene una previsión de desarrollo muy prometedora. Sin embargo, todavía es muy ignorada, lo que conlleva a un desconocimiento de esta subespecialidad, que se presenta como elemento intermediario de conexión entre el sofware, diagnóstico final y su posterior interpretación. Esta necesidad de profesional capacitado ha requerido la formación de una estudiante de doctorado (Ayelen Ramos) para no depender de empresas externas (algo que es además difícil de financiar en nuestro centro) en la ejecución de dichas tareas.

Este estudio pretende desarrollar una interfaz de conexión entre estos dos puntos: el sofware y el diagnóstico final en nuestro entorno, con un acercamiento simplificado y asequible de las principales herramientas de análisis y estudio del genoma para facilitarle el entendimiento a las personas que puedan estar interesadas en el futuro. Hasta este momento, el plan de formación en Anatomía Patológica no ha contemplado la formación en bioinformática suficiente para el análisis genómico de mutaciones oncogénicas. Con la realización de este TFG hemos puesto a punto un plan de formación básico complementario que facilite la aproximación de este tipo de herramientas a futuros residentes de Anatomía Patológica y de otras especialidades que rotan por el servicio, si bien en su inicio formativo se recomienda el respaldo de una empresa externa especializada. En última instancia, la compartición de conocimientos transversales debe mejorar el posicionamiento del Hospital Universitario de Canarias dentro de la esfera del manejo de pacientes oncológicos.

Objetivo

Objetivo principal

Aprendizaje de las técnicas para el análisis del mutaciones a partir de la muestras secuenciadas del genoma de los pacientes a estudio y relación de los resultados con un perfil oncológico. Dada la diversidad de herramientas informáticas disponibles para cada problema, nuestro objetivo principal ha sido experimentar con varias de ellas y elegir las que mejor se adecuan a nuestras necesidades diagnósticas, configurando un "pipeline" de análisis bioinformático de mutaciones en ADN de muestras de cáncer.

Objetivos secundarios

1. Estudio de los principales genes en el cáncer de mama y su expresión genética
2. Diseño y desarollo de pipelines para obtener, seleccionar y detectar variaciones genómicas relevantes.
3. Uso de bases de datos de bases científicas y bioinformáticas: Ensembl, NCBI, OncoKB, InTOGEN, Galaxy

Material y métodos:

Diseño del estudio

Se trata de un estudio experimental de muestras de genoma secuenciadas de pacientes diagnosticados de patologías oncológicas en el Hospital Universitario de Canarias. Se pretende estudiar, modificar en caso de necesidad, mapear y filtrar el genoma para analizar las mutaciones relevantes del mismo mediante técnicas de bioinformática. Este proceso lo realizaremos con las muestras de pacientes previamente diagnosticados de diferentes patologías oncológicas. Finalmente compararemos los resultados obtenidos con bibliografía relevante.

Sujetos del estudio

Se utilizaron archivos informáticos de secuenciación masiva de muestras de ADN obtenidas de tumores que precisaban diagnóstico de mutaciones en genes dianas de posible tratamiento dirigido.

Secuenciación

Para la secuenciación, es importante "preparar el ADN": dividirlo en fragmentos más pequeños a los que hay que introducirle *adapters* para poder "marcar" el ADN que se quiere secuenciar. La máquina de secuenciación usa láseres y cámaras para detectar la luz emitida por la reacción de un nucleótido pareado. Cada nucleótido se corresponde con una específica longitud de onda, por lo que capturando las imágenes producidas por cada ciclo, la máquina puede leer líneas de bases llamadas "reads" o lecturas que se presentarán en formato digital a través de un archivo *.fq* o *fastq*, como se ve en la imagen mostrada:

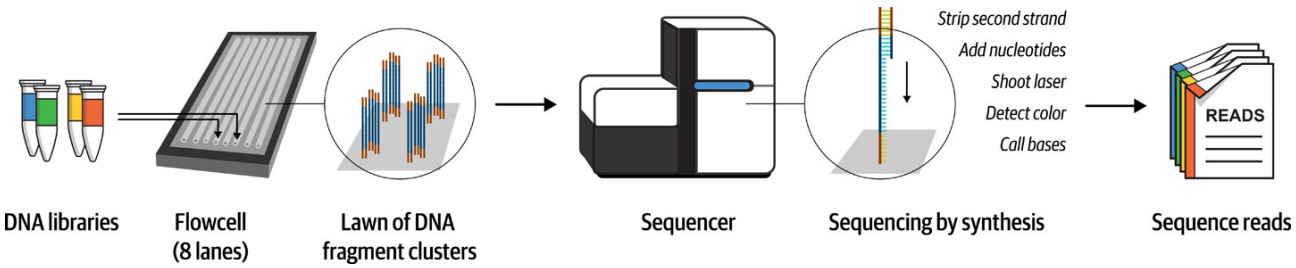


Figura 6: *Proceso de secuenciación [3]*

Para que el entendimiento de los siguientes pasos sea mejor, vamos a utilizar como ejemplo de lecturas las siguientes: Paciente1_R1.fastq y Paciente1_R2.fastq

[1][3][4][12]

Fases del estudio bioinformático

Una vez obtenido el archivo de la secuencia que proporciona el secuenciador, empieza el proceso de análisis bioinformático que con frecuencia se considera en tres fases: primario (control de calidad y pulido de las secuencias), secundario (alineamiento de las secuencias en el genoma humano y detección de variantes) y terciario (priorización de las mutaciones encontradas, interpretación de las mismas y su significado clínico). En cada una de las fases hay una diversidad de opciones y en el

presente TFG hemos explorado diversas opciones y optado por una solución adecuada a nuestras necesidades, construyendo de este modo un "pipeline" específico. A continuación se resumen los puntos principales y aprendizajes obtenidos durante este proceso.

0. GNU/Linux vs Galaxy

Como fase previa, es fundamental familiarizarse con GNU/Linux para así poder trabajar directamente desde la "terminal", con la línea de comandos. Esto es un poco intimidante si solo estamos acostumbrados a utilizar ordenadores con sistemas operativos de entorno gráfico (Windows y MacOS) pero es altamente recomendado a la hora de analizar genoma, ya que nos dará información más exacta de cada paso concreto que ejecutemos y sus respectivos errores. Para usar la línea de comandos no hace falta instalarse GNU/Linux: cualquier ordenador con Windows o MacOS puede incorporar la terminal de GNU/Linux como sistema operativo subyacente. Aunque la transición es más fácil si se trabaja directamente con GNU/Linux como sistema operativo principal. Los comandos Unix/Linux más importantes que necesitaremos usar son los que se listan en la siguiente tabla. Utilizaremos los símbolos *menor que* y *mayor que* [<>] para señalar la información que hay que introducir en el espacio concreto de la línea de comando. Es importante mencionar que dicha línea de comandos es sensible a mayúsculas y a espacios y el orden de aparición de los comandos que se especifique en los manuales tiene que seguirse rigurosamente [8]

`man <comando>`	Nos presenta el manual del programa que vamos a ejecutar, muchas veces aporta ejemplos que pueden ayudar a entender mejor su aplicación.
`index <archivo>`	Nos permite indexar el archivo con el que vamos a trabajar. Esto se hará con archivos grandes como el del Genoma Humano y el Archivo secuenciado una vez haya sido ordenado por coordenadas
`bwa mem`	Necesario para alinear las dos lecturas del genoma con el genoma de referencia (GRCh38)
`Samtools`	Permite limpiar y ordenar el archivo alineado
`java -jar <programa o directorio hasta programa>`	Necesario para ejecutar programas usando java

Si, por el contrario, no se quisiera usar la línea de comandos, o se quisiera usar en momentos puntuales, se puede introducir el uso de Galaxy. Esta es una herramienta de código abierto que

permite la ejecución de las tareas que mencionaremos en este Trabajo a través de la comodidad de un entorno web. Se puede elegir entre tres servidores: Galaxy EEUU (https://usegalaxy.org/), Galaxy Europa (https://usegalaxy.eu/) y Galaxy Australia (https://usegalaxy.org.au/). En este trabajo se utilizó el servidor europeo, ya que en el momento de uso, el estadounidense ocasionaba problemas en el alineamiento relacionados con fallos por mantenimiento. Sin embargo, se recomienda usar servidores no-europeos una vez estos conflictos hayan sido solventados, ya que por la diferencia horaria hay menos saturación por uso y las tareas se pueden ejecutar con mayor rapidez. [13]

1. Estudio de calidad

En las muestras seleccionadas, se utilizó FastQC para el análisis de calidad de ambas lecturas del paciente. Este es el software más utilizado, ya que aporta múltiples gráficas intuitivas sobre la calidad de la secuencia. Su estudio nos permitirá valorar si merece la pena seguir con el estudio o hay que repetir la secuenciación. También nos indica si es necesario hacer trimming de nuestros resultados de secuenciación. Este análisis se puede hacer a través de la línea de comandos y de Galaxy

Galaxy:

Después de subir los archivos que se quieren usar, se va a Tools> Genomic File Manipulation> Quality Control > FASTQ info.

Línea de comandos para la terminal:

```
fastqc <fastq_input_file>
```

Ejemplo:

```
fastqc Paciente1_R1.fastq
fastqc Paciente1_R2.fastq
```

Tras hacer el análisis, aparecerán dos archivos que se verán de la siguiente manera:

```
Paciente1_R1_fastqc.html
Paciente1_R2_fastqc.html
```

Estos hay que abrirlos para poder ser visualizados en la web y poder ver los informes pertinentes.

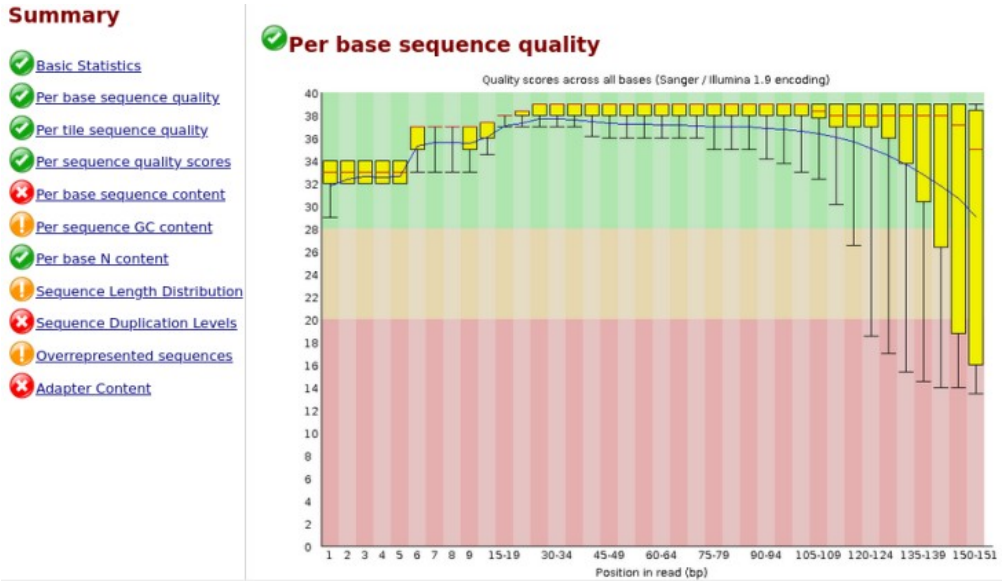
```
xdg-open <fastq_input_file_html>
```

Ejemplo:

```
xdg-open Paciente1_R1_fastqc.html
```

Es importante tener en cuenta el tipo de datos secuenciado: si son capturas de fragmentos de genoma y secuenciación directa de los mismos o si bien son producto de reacciones de amplificación por PCR. Esto último es lo habitual en nuestro caso, ya que partimos de cantidades muy pequeñas de ADN de baja calidad, obtenido a partir de cortes de tejido en parafina.

***Figura 7**: Imagen de selección propia. En esta imagen podemos ver el resultado del estudio de calidad de la muestra de una paciente. En las bases finales de la lectura la calidad disminuye por debajo de 30 (Q<30)*



Aquellas que no daban resultados óptimos se recortan con Trimmomatic usando los siguientes recursos:

Galaxy:

Se utilizan los archivos previamente analizados por FASTQC y se procede de la siguiente manera:

Genomic File Manipulation > FASTA/FASTQ> Trimmomatic.

Antes de ejecutar esta tarea, se incorporan los datos de calidad que se especifican en los pasos siguientes.

Línea de comandos:

Cuando ambas lecturas tenían problemas de calidad, se invoca el programa TrimmomaticPE, con los atributos que definen los puntos de corte de calidad (en este caso con un valor de Phred Q33):

```
      TrimmomaticPE    s_1_1_sequence.txt.gz    s_1_2_sequence.txt.gz
 lane1_forward_paired.fq.gz    lane1_forward_unpaired.fq.gz     lane1_re-
      verse_paired.fq.gz  lane1_reverse_unpaired.fq.gz ILLUMINACLIP:/usr
/share/trimmomatic/illuminaClipping.fa:2:40:15 LEADING:3 TRAILING:3 SLID-
      INGWINDOW:4:15 MINLEN:36
```

Ejemplo:

```
valentina@comrade:~$ Paciente1_R1.fastq Paciente1_R2.fastq Paciente1_R1_paired.fastq
Paciente_R1_unpaired.fastq Paciente_R2_paired.fastq Paciente_R2_unpaired.fastq ILLUMI
NACLIP:/usr/share/trimmomatic/illuminaClipping.fa:2:40:15 LEADING:3 TRAILING:3 SLIDIN
GWINDOW:4:15 MINLEN:36
```

Cuando solo una de las lecturas tenía problemas de calidad, se invoca el programa TrimmomaticSE:

```
TrimmomaticSE   s_1_1_sequence.txt.gz   lane1_forward.fq.gz   ILLUMINA-
CLIP:/usr/share/trimmomatic/illuminaClipping.fa:2:40:15       LEADING:3
TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36
```

Ejemplo:

```
valentina@comrade:~$ Paciente1_R1.fastq Paciente1_forward.fastq ILLUMINACLIP:
IlluminaClipping.fa:2:40:15 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36
```

Consideraciones a tener en cuenta:

1. Si no se especifica nada en la Phred score, se utilizará automáticamente una phred score de 33. Phred-64 no es recomendable, ya que dejaría poco genoma para analizar.
2. En los manuales se incorpora con frecuencia el uso de ILLUMINACLIP, una herramienta que se encarga de cortar adaptadores y secuencias específicas de Illumina, en este trabajo no ha sido necesario para ejecutar las tareas ya que el propio secuenciador usado Illumina MiSeq realiza esta tarea antes de proporcionar el archivo de salida.
3. Los valores más importantes son LEADING y TRAILING, ya que determinan el valor fronterizo a través del cual se “corta” la lectura al inicio y al final de la misma (que son las zonas donde suele haber mayores problemas de calidad). En este trabajo se le aplicó un valor de 30 (QB=30)
4. SLIDINGWINDOW nos determina el número de bases que se analizan de una vez (en este ejemplo 4) y el valor de calidad medio que tienen que tener para poder ser leídas (en este ejemplo 15). Su uso dependerá de la calidad general de la lectura analizada. Muy recomendable si hay varias zonas heterogéneas de la lectura que tienen mala calidad.

5. MINLEN determina el mínimo de longitud que tendrán las lecturas que se van a analizar (en este ejemplo, cualquiera <36 pares de bases). Este herramienta sería necesaria cuando el genoma analizado fuera muy heterogéneo encuanto a la longitud de las bases
6. Aunque en los ejemplos los archivos estén en formato .gz, (formato comprimido, con la esta abreviación proviene de “GNU Zip” y es un tipo de compresión para los sistemas UNIX). es posible usar archivos directamente en formato .fastq

Una vez se hace este “corte” de las zonas de peor calidad. Volvemos a hacer la evaluación de calidad de la muestra. [14][15] [16][17]

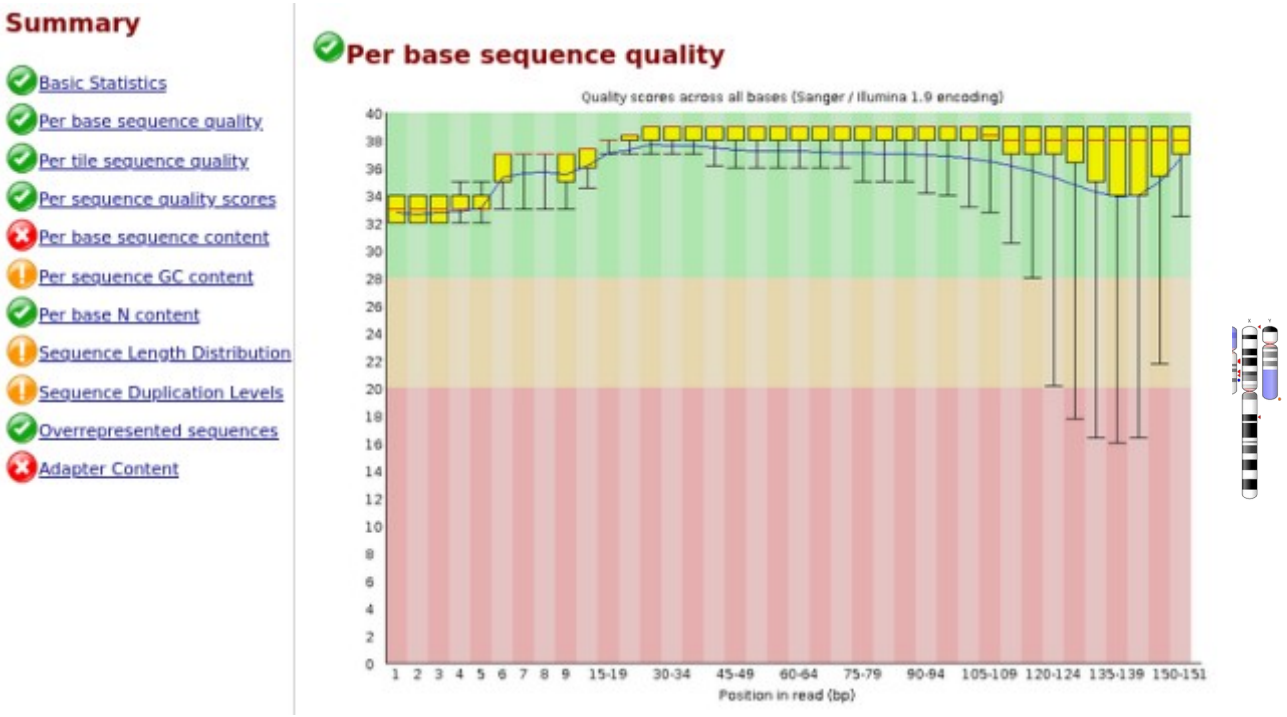


Figura 8*: Imagen de obtención propia. Aquí se presenta el resultado del corte de las bases de la lectura previa. Como podemos ver, todas se encuentran dentro del rango “verde”, determinando una buena calidad de la muestra*

2. Alineación

El siguiente paso consistirá en la alineación de las dos lecturas al genoma de referencia, que en este caso es el genoma humano *Homo Sapiens GRCh38.* Es importante tener en cuenta que a lo largo del proyecto Genoma Humano se han generado múltiples versiones del genoma de referencia y que algunas herramientas y comandos que utilicemos pueden estar optimizadas para una versión u otra. Tendemos a utilizar la última versión completa del genoma, GRCh38, pero hemos de tener presente que algunos datos podrían estar referidos a la versión previa, GRCh37.

Descargamos el genoma de referencia de Ensembl: Ensmbl →Downloads → FTP Downloads→ archivo FASTA de Homo.Sapiens → Homo_sapiens.GRCh38.dna.primary_assembly.fa.gz.

Para descomprimir:

```
gzip [ -acdfhklLnNrtvV19 ] [-S suffix] [ name ... ]
```

Ejemplo:

```
gzip -d Homo_sapiens.GRCh38.dna.primary_assembly.fa.gz
```

Este genoma de referencia debe ser "indexado" para facilitar que el ordenador encuentre las secuencias con las que hay buena identidad de una manera más rápida, fragmentando el proceso en unidades que son más asequibles a un computador convencional. Si nos imagináramos que tenemos que cocinar arroz para 60 personas, podríamos hacerlo con una cacerola, o podríamos hacerlo utilizando 10 cacerolas e ir las cocinando a la vez. Una vez planteado, se puede intuir que la segunda opción agilizaría el proceso más eficientemente, puesto que usar una cacerola para un trabajo tan grande requeriría cocinar, esperar a que se termine el arroz y volver a empezar de cero con la ración de la siguiente persona. Esto es lo que se pretende evitar indexando los archivos más pesados, dividiéndolos en “pequeños paquetes” de información para que puedan ser analizados paralelamente. Por ello, los archivos producto del proceso de indexado no deben borrarse en ningún momento.

En Galaxy no es necesario indexar el Genoma de referencia, ya que la propia web lo gestiona. Esta es una labor exclusiva de la línea de comandos.

Para indexarlo hay varios programas que lo ejecutan: bwa, samtools y picard. Todas ellas son herramientas enfocadas al trabajo del genoma y no hay gran diferencia entre ellas a la hora de ejecutar la tarea.

```
bwa index ref.fa
```

```
samtools faidx ref.fasta [region1 [...]]
```

```
CreateSequenceDictionary [arguments]
```

Ejemplo:

```
bwa index Homo_sapiens.GRCh38.dna.primary_assembly.fa
```

```
samtools faidx Homo_sapiens.GRCh38.dna.primary_assembly.fa
```

```
java -jar picard.jar CreateSequenceDictionary -R Homo_sapiens.GRCh38.dna.primary_assembly.fa
```

El indexado del genoma de referencia generará los siguientes archivos:

Homo_sapiens.GRCh38.dna.primary_assembly.fa.amb	Producto de usar BWA
Homo_sapiens.GRCh38.dna.primary_assembly.fa.ann	
Homo_sapiens.GRCh38.dna.primary_assembly.fa.bwt	
Homo_sapiens.GRCh38.dna.primary_assembly.fa.pac	
Homo_sapiens.GRCh38.dna.primary_assembly.fa.sa	
Homo_sapiens.GRCh38.dna.primary_assembly.fa.dict	Producto de PICARD
Homo_sapiens.GRCh38.dna.primary_assembly.fa.fai	Producto de SAMTOOLS

El archivo que vamos a usar para el alineamiento es el .fa, es decir , el previo al indexado.

Una vez tenemos el genoma de referencia indexado. Procedemos al alienamiento. Para ello utilizaremos la herramienta BWA y concretamente su algoritmo BWA MEM, que es la preferida para alinear de manera eficiente la elevada cantidad de secuencias cortas generadas mediantes NGS con un genoma de referencia del tamaño del genoma humano. Utiliza la teoría de alineamiendo de cadenas ("string matching") mediante la transformación de Burrows–Wheeler (BWT)). Este alineado permite "mismatches" y huecos ("gaps") facilitando que secuencias con mutaciones se alineen en el sitio correcto. Este proceso se puede ejecutar tanto en Galaxy como con la terminal.

Galaxy:

Se utilizan los archivos previos de FASTQC o de Trimmomatic

Genomics Analysis > Mapping > BWA-MEM2

Línea de Comandos:

```
valentina@comrade:~$ bwa mem -MY -R '@RG\tID:IDdelapersona\tSM:paciente1'
Homo.sapiens.GRCh38.fa -o archivodesalida.sam Paciente1_R1.fastq Paciente1
_R2.fastq
```

Consideraciones a tener en cuenta con `bwa mem`:

1. **-M** es necesario para el marcado de los alineamientos secundarios.
2. **-Y** es necesario para hacer softclipping. Esto conseguirá que partes de las lecturas que no encajen bien con el genoma de referencia en ambas lecturas serán ignoradas. Este proceso facilita el alineamiento y la detección de bases mutadas.
3. **-R** es necesario para añadir un encabezado y evitar problemas en el llamado de variantes.
4. **-o** indica el *output* o archivo de salida.

Una vez alineado, y antes de pasar al siguiente punto, es importante ordenar y limpiar el archivo alineado, además de convertirlo en un formato más asequible y compacto para trabajar: el BAM (*Binary Aligment Map*)

```
valentina@comrade:~$ samtools view -b archivodesalida.sam - | samtools fixmate - |
samtools sort - -o Paciente1_sorted.bam && samtools index Paciente1_sorted.bam
```

Consideraciones a tener en cuenta en esta línea de comandos en general:

1. El símbolo `'|'` es el pipeline. Indica que el resultado del comando antes de este símbolo (*output*) se usará como los datos de entrada (*input*) del siguiente.
2. El símbolo `'-'` es el dash. Indica que el *output* del comando anterior se leerá como los datos de entrada estandarizados del siguiente *input*.
3. El símbolo `'&&'.` Indica que el siguiente comando se ejecutará solamente si los anteriores se han ejecutado correctamente.

Consideraciones a tener en cuenta con `samtools`:

1. `view` convierte el archivo sam en bam (*Binary Alignment Map)* , permitiendo de esta manera trabajar con el archivo de una manera más rápida y efectiva, ya que el BAM es una compresión del SAM.

   Comando suelto: `samtools view -b Paciente1.sam -o Paciente1.bam`

   En este comando, la opción `-b` señala el archivo sobre el que se quiere trabajar. En este caso "`Paciente1.sam`"

2. `fixmate` rellena los espacios de las coordenadas.

   Galaxy: Genomic File Manipulator → SAM/BAM → Samtools fixmate

   Comando suelto: `samtools fixmate Paciente1.bam -o Paciente1_fixed.bam`

3. `Sort` ordena el archivo bam por coordenadas. Esto es necesario para en los siquientes pasos poder hacer estudios más eficientemente.

   Comando suelto:

   ```
   samtools sort Paciente1_fixed.bam -o Paciente1_fixedsorted.bam
   ```

4. `index` para poder trabajar con el archivo final más eficientemente. Este paso no es necesario con **Galaxy**

[18][19][20]

3. BAM QC y visualización

Una vez finalizada la alineación, es necesario comprobar la calidad de la misma. Para ello, se procede primero a averiguar cómo de exitosos han sido los emparejamientos. A continuación se presentan diferentes herramientas que pueden utilizarse:

Galaxy: Genomic File Manipulator → SAM/BAM → Samtools flagstat

Con Samtools:

```
samtools flagstat Paciente1_fixedsorted.bam
```

Con GATK:

```
java -jar gatk-package-4.2.6.1-local.jar FlagStat -I Paciente1_fixedsorted.bam
```

Con Qualimap:

```
qualimap Paciente1_fixedsorted.bam
```

Una vez se elige la herramienta y se ejecuta, nos aparecerá una imagen similar a la incorporada (*Figura 9*) que nos da información más específica sobre el tipo de emparejamientos que se han producido y si estos se han producido de manera correcta. En la primera columna se observan los emparejamientos exitosos y en la segunda los fallidos. Esta imagen es un ejemplo de una muestra que ha sido correctamente alineada, ya que el porcentaje de emparejamiento es muy alto (98.90%).

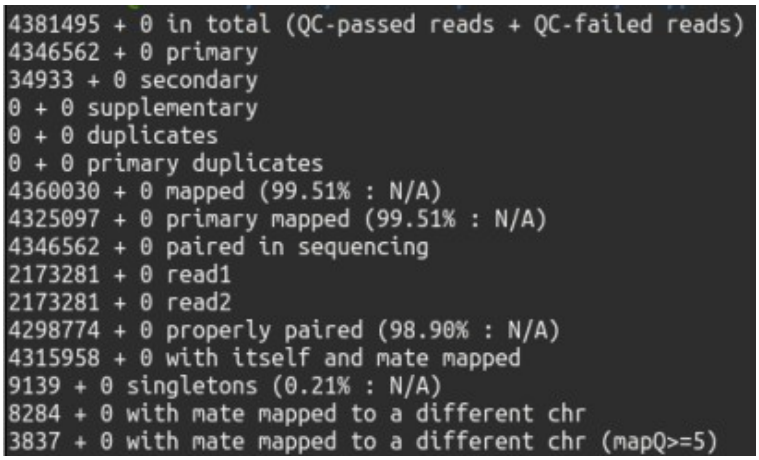


Figura 9: Imagen de obtención propia

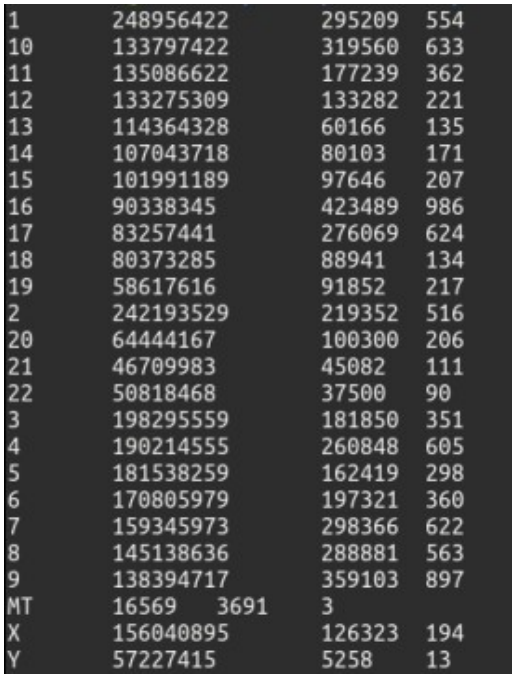
Cuando la calidad de la alineación se ha comprobado, se procede a observar el número de alineaciones por cromosoma:

Galaxy:

Genomic File Manipulation → Quality Control → QualiMap Counts QC

Línea de comandos:

```
samtools idxstats Paciente1_fixedsorted.bam
```



1	248956422	295209	554
10	133797422	319560	633
11	135086622	177239	362
12	133275309	133282	221
13	114364328	60166	135
14	107043718	80103	171
15	101991189	97646	207
16	90338345	423489	986
17	83257441	276069	624
18	80373285	88941	134
19	58617616	91852	217
2	242193529	219352	516
20	64444167	100300	206
21	46709983	45082	111
22	50818468	37500	90
3	198295559	181850	351
4	190214555	260848	605
5	181538259	162419	298
6	170805979	197321	360
7	159345973	298366	622
8	145138636	288881	563
9	138394717	359103	897
MT	16569	3691	3
X	156040895	126323	194
Y	57227415	5258	13

Figura 10*: Imagen de obtención propia*

La información se presentará en 4 columnas (*Imagen 10*) con la siquiente información:

1º columna: nombre del cromosoma

2º columna: longitud de la secuencia

3º columna: nº de secuencias alineadas

4º columna: nº de secuencias no alineadas

Una vez se corrobora el correcto mapeado de la secuencia, "apilamos" todas las secuencias que se han mapeado al mismo punto del genoma de referencia y generamos un archivo de texto .txt, que ya es de nuevo inteligible (a diferencia de los archivos en código binario).

Galaxy: Genomic File Manipulation → SAM/BAM → Samtools mpileup

Con la línea de comandos:

Samtools:

```
Samtools mpileup -f Homo.Sapiens.GRCh38.fa Paciente1_fixedsorted.bam -o Paciente1.pileup
```

En esta línea de comandos, la opción `-f` indica el genoma de referencia que hay que posicionar después

GATK:

```
java -jar gatk Pileup -R <genoma.referencia> -I <archivo.entrada> -O <archivo.salida>
```

Ejemplo:

```
valentina@comrade:~$ java -jar gatk-4.2.6.1/gatk-package-4.2.6.1-local.jar Pileup -R
 Homo_sapiens.GRCh38.fa -I Paciente1_fixsort.bam -O Paciente1_pileup.bam
```

El archivo una vez apilado presentará su información a modo de tabla, donde cada columna tiene un valor:

chr1	10000	N	1	^!a	3
chr1	10001	t	1	,	1
chr1	10002	a	1	,	1
chr1	10003	a	1	,	F
chr1	10004	c	5	,^!.^!.^!.	GEFFF
chr1	10005	c	6	,....^!.	BABFF2
chr1	10006	c	6	,.....	?1FBB6
chr1	10007	t	6	,.....	11FFFF
chr1	10008	a	6	,.....	BDFFFF
chr1	10009	a	6	,.....	31FFFF

1º columna: nombre del cromosoma

2º columna: posición de la base en el cromosoma

3º columna: base de referencia

Figura 11: Imagen de obtención propia

4º columna: nº de lecturas en esa posición

5º columna: determinación de la calidad del alineamiento en base a ASCII

6º columna: calidad de la base según el sistema ASCII[3][21][24]

4. Estudio y filtrado de variaciones

Para el estudio de variaciones se pueden utilizar varias herramientas. En este trabajo se utilizarán Varscan y GATK. Ambos son programas de código abierto muy utilizados dentro de la comunidad a la hora de hacer llamado de variaciones y su principal diferencia radica en que Varscan está más centrado en el llamado de variaciones, mientras que GATK cuentas con numerosas herramientas como Picard, dando cabida al uso de otras tareas como dan el cambio de formato y preparación del genoma a filtrar entre otras opciones. Dentro de GATK, utilizaremos `HaplotypeCaller.` De estos dos programas, solo Varscan puede usarse en Galaxy.

Galaxy:

Genomic File Manipulation→ VCF/BCF → Varscan

Para ejecutar la línea de comandos, es necesario usar java:

Varscan

```
java -jar VarScan.jar mpileup2snp Paciente1.pileup > archivosalida.txt
```

GATK:

```
valentina@comrade:~$ java -jar /gatk-4.2.6.1/gatk-package-4.2.6.1-local.jar
HaplotypeCaller -R Homo_sapiens.GRCh38.fa -I Paciente1.pileup.bam -O gatk_va
riants_call.vcf
```

Chrom	Position	Ref	Var	Cons:Cov:Reads1:Reads2:Freq:P-value
1	629750	G	A	A:9:1:8:88.89%:2.0568E-4
1	2656806	C	G	G:11:0:11:100%:1.4176E-6
1	2656819	C	G	G:12:0:12:100%:3.698E-7
1	2661984	C	G	G:14:0:14:100%:2.4927E-8
1	2661997	C	G	G:14:0:14:100%:2.4927E-8
1	2666600	C	G	G:16:1:15:93.75%:2.8282E-8
1	2666613	C	G	G:16:1:15:93.75%:2.8282E-8
1	2671216	C	G	G:16:0:16:100%:1.6637E-9
1	2671229	C	G	G:17:0:17:100%:4.2852E-10
1	2675872	C	G	G:13:0:13:100%:9.6148E-8
1	2675885	C	G	G:15:0:15:100%:6.4467E-9

Figura 12: Imagen de obtención propia

Con el archivo de texto resultante (*Figura 12*), se procederá al filtrado de variantes.

Para ello, creamos un archivo con las variantes seleccionadas en un formato que le permita a V programa trabajar con él. Este formato consiste posicionar en 4 columnas la siguiente información:

1. Cromosoma
2. Coordenada de inicio

```
14      104776629       104776629
16      2060607 2060607 G/A
16      2064406 2064406 C/T
16      2070401 2070401 T/C
16      2088217 2088217 A/C
17      7674797 7674797 T/C
```

Figura 13: Imagen de obtención propia

3. Coordenada final
4. Base supuesta /base mutada

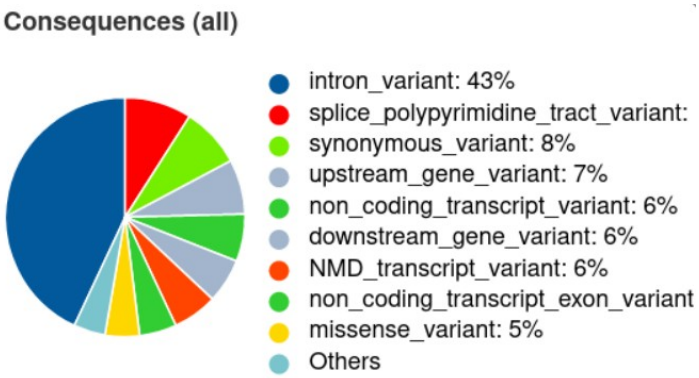
Una vez filtradas las secuencias encontradas, interesa confeccionar un archivo en un formato compacto que pueda ser interpretado por programas disponibles online para predecir las consecuencias de los cambios. El formato más útil es el VCF (variant call format).
Para convertir el archivo obtenido con pileup en VCF se usa el siguiente comando:

```
java -jar /Ruta-hasta-VarScan.v2.3.9.jar mpileup2snp Paciente1.pileup --output-vcf > Paciente1.vcf
```

[22][23]

5. Interpretación de variaciones

Una vez tenemos el archivo VCF, podemos trabajar con varias plataformas que ayudan a la interpretación de cambios de nucleótido y permiten priorizar los cientos de cambios encontrados en función de sus repercusiones biológicas predecibles (patogenicidad) y su significación clínica.
Hemos seleccionado dos plataformas muy útiles en este aspecto:
- VEP (Variant Effect Predictor: https://www.ensembl.org/info/docs/tools/vep/index.html) analiza los resultados de nuestro filtrado de mutaciones, nos hace un informe como el que mostramos en la imagen de la izquierda. Dentro de las mutaciones que se informan, seleccionamos solamente las missense, que nos dará la información de la variación en concreto y a qué gen se relaciona.



- CGI (Cancer Genome Interpreter: https://www.cancergenomeinterpreter.org) también utiliza archivos VCF para computar las consecuencias de los cambios tabulados y ofrece un listado de mutaciones de consecuencias patogénicas. Además, CGI incorpora un análisis terciario de "accionabilidad" de las mutaciones encontradas, presentando opciones terapéuticas dirigidas a las mutaciones encontradas. [22][23][26]

6. Accionabilidad de las mutaciones

En un último paso, nos interesa integrar las mutaciones encontradas en genes implicados en cáncer con los conocimientos actualizados sobre la utilidad terapéutica de las mutaciones oncogénicas,

conocido como "accionabilidad". Para este paso, la plataforma que nos ha parecido más útil es OncoKB .

Resultados

Las mutaciones que encontramos en las muestras analizadas se dividen en mutaciones conductoras (driver) y pasajeras (passenger). Las primeras son mutaciones que favorecen la proliferación tumoral, mientras que las segundas pueden tener un efecto en la inactivación de genes supresores de tumores, acelerando la supresión . El compendio de mutaciones identificadas se puede resumir en el siguiente cuadro:

Paciente	Cromosoma	Variación	Gen
Paciente 1	17	7676154G>C	TP53 (passenger)
	17	39723335A>G	ERBB2 (driver)
	9	132910596T>C	TSC1 (passenger)
Paciente 2	17	7675145C>G	TP53 (driver)
	17	39723335A>G	ERBB2 (driver)
Paciente 3	8	38427976C>T	FGFR1 (passenger)
	7	140749365A>T	BRAF (driver)
	7	55191788A>T	EGFR (driver)
Paciente 4	8	31640694A>T	NRGI (passenger)
	2	29225544T>G	ALK (passenger)

En el paciente 1 se detectaron tres mutaciones: una mutación conductora y dos pasajeras:

- La primera mutación (conductora) es una sustitución en el codón 655 del gen ERBB2 con nomenclatura de proteína p.I655V (p.Ile1655PVal) y con nomenclatura de ADN

c.39723335A>G (cambio de adenina por guanina en la base 39723335A, que provoca el cambio de aminoácido Isoleucina por una Valina). Esto codifica para la modificación conductora ERBB2.

- La segunda mutación (pasajera) es una sustitución en el codón 72 del gen TP53 con nomenclatura de proteína p.P72R (p.Pro72Arg) y con nomenclatura de ADN c.7676154G>C (cambio de guanina por adenina en la base 7676154, que provoca el cambio de aminoácido Prolina por una Arginina). Esto codifica para la modificación pasajera TP53.
- La tercera mutación es una sustitución en el codón 413 del gen TP5 con nomenclatura de proteína p.Q413R (p.Gln413Arg) y con nomenclatura de ADN c.132910596T>C (cambio de Timina por Citosina en la base 132910596, que provoca el cambio de aminoácido Glutamina por una Arginina). Esto codifica para la modificación conductora TSC1.

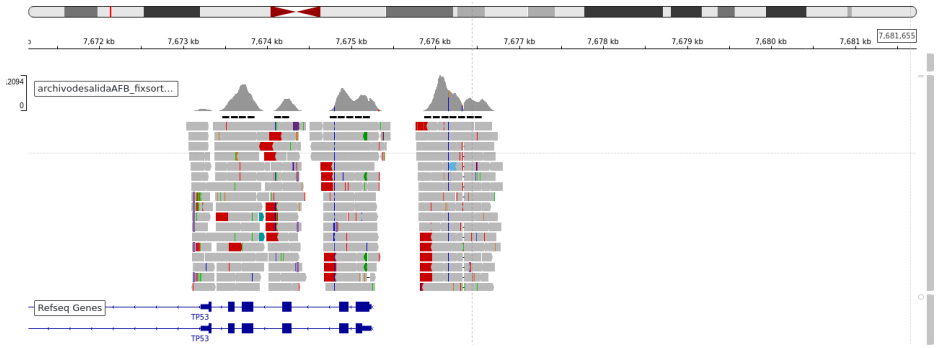


Figura 14: En esta Imagen solo se presenta la mutación conductora TP53, puesto que es la representativa. Imagen de obtención propia

En el Paciente 2 se detectaron 2 mutaciones conductoras:

- La primera es una sustitución en el codón 156 del gen TP53 con nomenclatura de proteína p.R156P (p.Arg156Pro) y con nomenclatura de ADN c.7675145C>G (cambio de citosina por guanina en la base 7675145, que provoca el cambio de aminoácido Arginina por una

Prolina). Esto codifica para la modificación conductora TP53(*Figura 14*). .

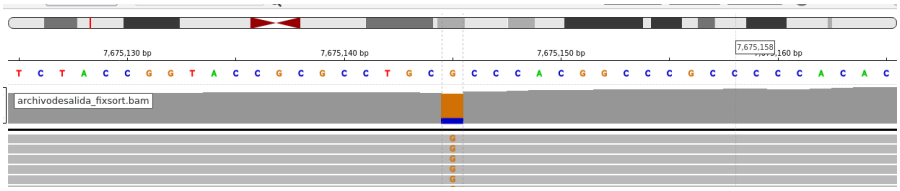


Figura 14: En esta imagen obtenida de IGV se puede percibir la presencia de Guanina (G) en la exacta localización mencionada en el trabajo. Imagen de obtención propia

- La segunda es una sustitución en el codón 655 del gen ERBB2 con nomenclatura de proteína p.I655V (p.Ile156PVal) y con nomenclatura de ADN c.39723335A>G (cambio de adenina por guanina en la base 39723335A, que provoca el cambio de aminoácido Isoleucina por una Valina. Esto codifica para la modificación conductora ERBB2.

Figura 15: En esta imagen podemos ver una representación de las dos mutaciones encontradas. Imagen de obtención propia

En el Paciente 3 se detectaron tres mutaciones: dos mutaciones conductoras en el cromosoma y una mutación pasajera.

- La primera es una sustitución en el codón 638 del gen BRAF con nomenclatura de proteína p.D638E (p.Asp638Glu) y con nomenclatura de ADN c.140749365A>T (cambio de Adenina por Timina en la base 140749365, que provoca el cambio de aminoácido Aspártico por Glutámico). Esto codifica para la mutación conductora BRAF.
- La segunda es una sustitución en el codón 847 del gen EGFRcon nomenclatura de proteína p.T847S (p.Thr156Ser) y con nomenclatura de ADN c.55191788A>T (cambio de Adenina por Timina en la base 55191788, que provoca el cambio de aminoácido Treonina por una Serina). Esto codifica para la mutación conductora EGFR.

- La tercera es una sustitución en el codón 220 del gen FGFR1con nomenclatura de proteína p.R220H (p.Arg156His) y con nomenclatura de ADN c.38427976C>T (cambio de Citonina por Timina en la base 38427976, que provoca el cambio de aminoácido Arginina por una Histidina). Esto codifica para la mutación pasajera FGFR1.

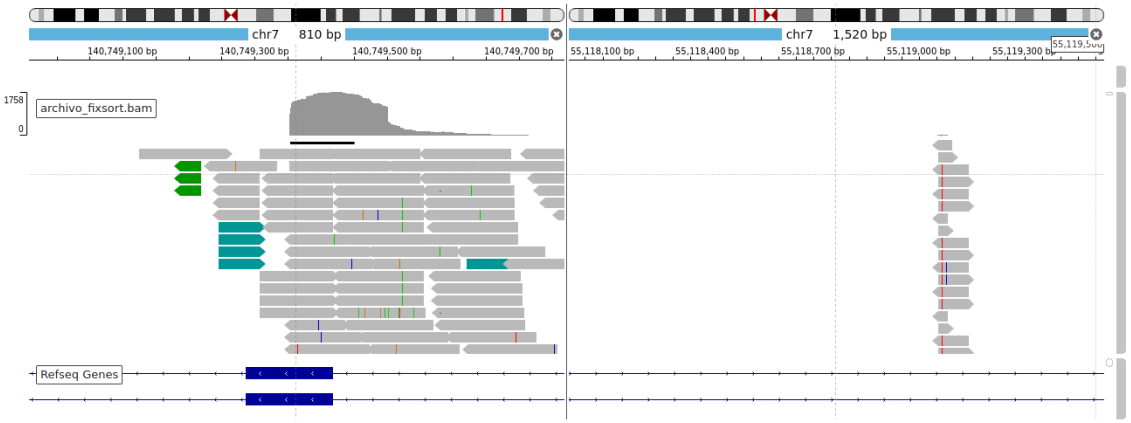


Figura 16: En esta imagen se pueden observar las dos mutaciones conductoras. Imagen de obtención propia

En el paciente 4 se detectaron dos mutaciones pasajeras:

- La primera es una sustitución en el codón 90 del gen NRG1 con nomenclatura de proteína p.K90M (p.Lys638Met) y con nomenclatura de ADN c.31640694A>T (cambio de Adenina por Timina en la base 31640694, que provoca el cambio de aminoácido Lisina por Metionina). Esto codifica para la mutación pasajera NRG1.
- La segunda es una sustitución en el codón 1030 del gen ALK con nomenclatura de proteína p.H1030P (p.His1030Pro) y con nomenclatura de ADN c.29225544T>G (cambio de Adenina por Timina en la base 29225544, que provoca el cambio de aminoácido Histidina por una Prolina). Esto codifica para la mutación pasajera ALK.

:

Discusión

En todos los pacientes estudiados, las mutaciones encontradas se contrarrestaron con las bases de datos actuales: COSMIC, IntoGen, CGI y Ensembl. En este apartado solo se tratarán las mutaciones con interés clínico y consideradas patogénicas: las mutaciones conductoras.

En el paciente número 1, la mutación en el cromosoma 17 (p.I655V) identificada como ERBB2 se determina como patogénica para Adenocarcinoma de Mama (score 0.55 en Intogen; score 0.72 en COSMIC).

En el paciente número 2, la mutación del gen ERBB2 se comparte con el paciente 1, por lo que se identifica como patogénica para Adenocarcinoma de Mama. La segunda mutación puntual y conductora que presenta en TP53 (p.R156P) presenta numerosas posibilidades diagnósticas, ya presenta una score =1 para varias patologías: cáncer de vejiga biliar, adenocarcinoma de mama, cáncer hepático, adenocarcinoma de pulmón, sarcoma y glioma de bajo grado entre otros.

En el paciente número 3 se identifican dos mutaciones conductoras La primera es (p.D638E) que codifica para el gen BRAF. En las bases de datos consultadas no hay asociaciones directas claras con vías patogénicas concretas, pero en la base de datos de IntoGen se identifica una alta presencia de este tipo de mutación BRAF en el adenocarcinoma de tiroides y melanoma cutáneo. En Pubmed se encontró una sola mención a esta mutación concreta asociándola al síndrome cardio-fascio-cutáneo. La segunda mutación (p.T847S) que codifica para el gen EGFR tampoco presenta una clara relación a una patología concreta, pero IntoGen la asocia más cercanamente con el cáncer de pulmón de células pequeñas, el glioblastoma y el adenocarcinoma de pulmón. No se ha encontrado literatura al respecto en Pubmed

En el paciente 4, las 2 mutaciones pasajeras se desestiman al no considerarse de características patogénicas.

Dentro del pipeline utilizado en este trabajo, no se han detectado diferencias significativas con respecto al uso de la línea de comandos y Galaxy a la hora de analizar el genoma. Si bien la solución de errores ha sido más eficiente a través de la línea de comandos, Galaxy aporta una

interfaz gráfica mucho más amable para el consumidor medio, además de que elimina la necesidad de indexar los datos que se vayan a trabajar, puesto que lo gestiona el mismo programa.

Con respecto a la ejecución del “trimmeado”, se detectó una dificultad de detección de mutaciones en genoma trimmeado:

- En el paciente 1 y 2 la detección de mutaciones no presentó ninguna variación tras el proceso de trimmeado con los siguientes valores: SLIDINGWINDOW:4:15 LEADING:3 ENDING:3
- En el paciente 3, la detección de mutaciones tras el trimmeado fue defectuosa: se detectó la mutación pasajera que codifica para el gen FGFR1, pero se obvió la mutación conductora EGFR, la cual solo fue identificada cuando se analizaron ambas lecturas del genoma sin modificar.
- En el paciente 4, la detección de mutaciones tras el trimmeado también fue defectuosa: se detectó la mutación pasajera que codifica para el gen NRG1, pero se obvió la mutación pasajera ALK, que solo fue identificada cuando se analizaron ambas lecturas del genoma sin modificar.

Esto hace pensar que la herramienta de trimmeado, si bien es muy útil para poder filtrar las lecturas de mala calidad y hacer más eficiente y rápido el proceso de selección, solo debería recomendarse en genoma con una calidad general por base y por ranura de secuenciado mala, ya que la diferencia de resultado no varía cuando el genoma analizado solo presenta una disminución de la calidad en las lecturas finales. También es necesario usarla solo cuando las lecturas tienen un mínimo de bases >1Mbp, porque trimmear lecturas más pequeñas dará error en el proceso de mapeado.

Cuando los resultados son comparados con una empresa especializada (ArcherDX), se detecta una variación en pacientes analizado:

- En el paciente 4, que según nuestros análisis había dado negativo para mutaciones patogénicas se detecta una mutación patogénica con nomenclatura de la proteína p.K591E p.Lys591Glu y con nomenclatura de ADN c.1771A>G que codifica para BRAF
- En el paciente 3, si bien también se detectan mutaciones que codifican para el gen BRAF, como nuestro informe determina, se detectan mutaciones en 3 localizaciones diferentes que

codifican para el mismo gen BRAF: p.Lys591Glu con DNA c.1771A>G; p.Arg444Gln con DNA c.1331G>A; p.Phe583Val con DNA c.1747T>G.

Dentro de las razones de discrepancia en la localización de mutaciones, no se encontraron causas específicas. Como la imagen adjunta nos muestra (*Figura 17*), la variación de identificación de mutaciones presentada en comparación con ArchDX podría estar relacionada tanto con las herramientas utilizadas para el análisis de genoma, el tipo de bases de datos que se usen para comparar y diferencias en el error humano a la hora de programar, si bien cada paciente presentado en este trabajo fue analizado varias veces, con genoma trimmeado y sin trimmear y utilizando tanto la línea de comandos como Galaxy. Tampoco contamos con los datos de filtrado de las variaciones que estas empresan utilizan y si esa diferencia de límite podría suponer una causa de peso. En este trabajo se ha identificado también la discrepancia a la hora de identificar mutaciones, la cual es dependiente de la base de datos que se use y la cantidad de datos que tenga cada base de datos sobre mutaciones concretas, por lo que se recomienda siempre utilizar varias fuentes de información para detectar mutaciones y catalogarlas lo más correctamente posible.

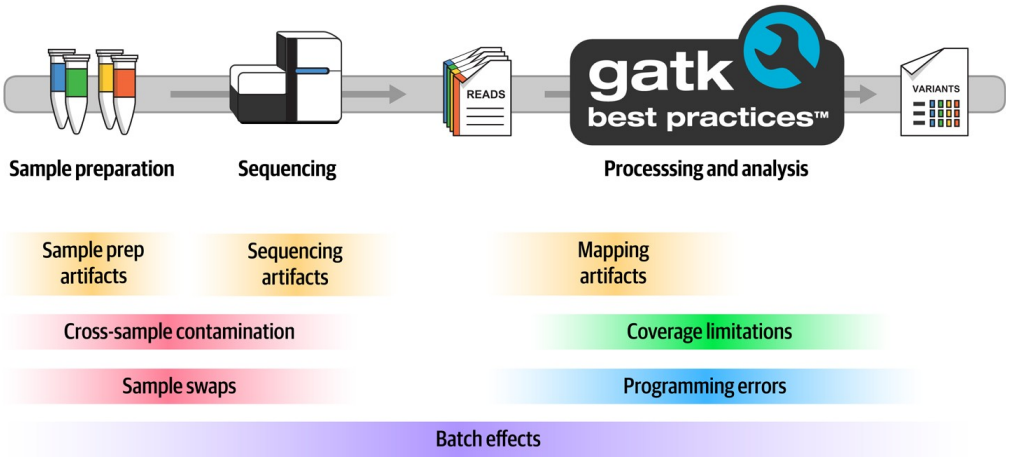


Figura 17: En esta imagen se pueden identificar todas las causas de error en el proceso bioinformático[3]

Conclusiones

El objetivo principal de este trabajo se ha cumplido, ya que se han estudiado y podido utilizar diferentes herramientas informáticas para correctamente identificar y catalogar mutaciones en 4

pacientes, asociándolos a perfiles oncológicos concretos y, en caso de no ser posible, a estimaciones plausibles. El uso de pipelines y su diseño y modificación para cada muestra genómica fue fructífera, ya que se pudo analizar y arreglar errores concretos a base de modificaciones puntuales en la línea de comandos. En cuanto a los objetivos secundarios, se consiguió hacer un uso eficiente de las herramientas de predicción de variaciones y su consecuente comparación con diferentes bases de datos (COSMIC, OncoKB, ClinVar y CGI) seleccionando entre las resultantes aquellas consideradas relevantes.

¿Qué he aprendido en este TFG?

Este TFG ha supuesto un gran aprendizaje, ya que ha supuesto un cambio del enfoque clínico para volver a retomar el estudio y el recuerdo de las bases moleculares y fisiológicas de la enfermedad de los primeros años de carrera. Desde el punto de vista de la informática, ha supuesto una oportunidad para conocer mejor el potencial del sistema GNU/Linux en su conjunto. Ha sido especialmente significativo el poder encontrar un nicho que permitiera la compaginación de dos campos de interés personal en uno solo: la informática y la medicina.

De este TFG he aprendido sobre todo de la capacidad exponencial de crecimiento que tiene la bioinformática, y lo importante que es desarrollarla dentro del marco de la salud pública,para poder democratizar el acceso a diagnósticos de poblaciones y heterogeneizar los genoma de referencia y las bases de datos presentes con datos de variaciones de todas las poblaciones posibles. La poca necesidad de recursos que requiere la identificación y filtración de variaciones (un secuenciador y un ordenador con conexión a internet) brinda una oportunidad de oro para que los países de Sur Global puedan invertir en un desarrollo de la vía diagnóstica oncológica de sus poblaciones mucho más horizontal y versátil. También, este acceso tan abierto a la posibilidad diagnóstica proporciona una autonomía clínica a los profesionales que difícilmente sería cohartada por conflictos políticos o bélicos que podrían condicionar el acceso a infraestructura solamente accesible en el Norte global.

Este TFG pretende ser una llamada a la curiosidad y el interés por esta subespecialidad, demostrando que con poco recursos y un poco de voluntad, también se puede acceder a ramas de la Medicina desconocidas que puedan interesarnos.

Bibliografía:

Figura 2:
1. Batini C. EMBL-EBI NGS bioinformatics. File Formats & Data QC [Internet]. 2021 ene. Disponible en:https://docs.google.com/presentation/d/1bleyP3NTR4e7lRb68EITwykG91S5TFyIU-y7gReCf7w/edit?usp=sharing

2. Wikipedia contributors. Pipeline [Internet]. Wikipedia, The Free Encyclopedia. 2022. Disponible en: https://en.wikipedia.org/w/index.php?title=Pipeline&oldid=1105701087

3. van der Auwera G, O'Connor BD. Genomics in the cloud: Using docker, GATK, and WDL in Terra. Sebastopol, CA: O'Reilly Media; 2020.

4. Applied Biological Materials-abm. The beginner's guide to RNA-seq - #ResearchersAtWork webinar series [Internet]. Youtube; 2019 [citado el 30 de agosto de 2022]. Disponible en: https://www.youtube.com/watch?v=8lAVfKbRK3I

5. Green E. DNA sequencing [Internet]. Genome.gov. [citado el 30 de agosto de 2022]. Disponible en: https://www.genome.gov/genetics-glossary/DNA-Sequencing

6. Quality score encoding [Internet]. Illumina.com. [citado el 30 de agosto de 2022]. Disponible en: https://support.illumina.com/help/BaseSpace_OLH_009008/Content/Source/Informatics/BS/QualityScoreEncoding_swBS.htm

7. Bioinformatics [Internet]. Genome.gov. [citado el 30 de agosto de 2022]. Disponible en: https://www.genome.gov/genetics-glossary/Bioinformatics

8. Batini C. EMBL-EBI NGS bioinformatics. Introduction to Unix [Internet]. 2021 ene. Disponible en: https://drive.google.com/file/d/12wdjg6g1LVsv6AAgop-FuigP9JwyJHSC/view?usp=sharing

9. De Bo C. Linux for bioinformatics DAY 1 [Internet]. Vib.be. 2018 [citado el 30 de agosto de 2022]. Disponible en: https://data.bits.vib.be/pub/trainingen/Linux/Linux_01.pdf

10. Wikipedia contributors. Open source [Internet]. Wikipedia, The Free Encyclopedia. 2022. Disponible en: https://en.wikipedia.org/w/index.php?title=Open_source&oldid=1107319771

11. (解少俊) SX. Chapter 1 Why Linux? [Internet]. Github.io. 2021 [citado el 30 de agosto de 2022]. Disponible en: https://xie186.github.io/Novice2Expert4Bioinformatics/why-linux.html

12. Illumina. Intro to sequencing by synthesis: Industry-leading data quality [Internet]. Youtube; 2014 . Disponible en: https://www.youtube.com/watch?v=HMyCqWhwB8E

13. Home - Galaxy community hub [Internet]. Galaxyproject.org. [citado el 1 de septiembre de 2022]. Disponible en: https://galaxyproject.org/

14. Nguyen J. What are Phred Scores? [Internet]. Zymo Research Europe. 2021 [citado el 1 de septiembre de 2022]. Disponible en: https://www.zymoresearch.de/blogs/blog/what-are-phred-scores

15. Unknown. Trimmomatic Manual: V0.32 [Internet]. 2020. Disponible en: http://www.usadellab.org/cms/uploads/supplementary/Trimmomatic/TrimmomaticManual_V0.32.pdf

16. Babraham bioinformatics - FastQC A quality control tool for high throughput sequence data [Internet]. Babraham.ac.uk. Disponible en: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

17. Batin C. NGS data: from initial QC to a ready-to-analyse variant set [Internet]. 15-19th February 2021. Disponible en: https://docs.google.com/document/d/10RKw3w9iL0MeGvourEi69W9iT3NcHRAJYMmpUWFj7mE/edit

18. Batini C. EMBL-EBI NGS bioinformatics, alignment to a reference genome [Internet]. 2021 feb. Disponible en: https://docs.google.com/presentation/d/1p5z2S6usn9swWNEvExj9geOqnNBlSmnDpZRVtBcuDcY/edit#slide=id.p

19. GNU project. GNU Gzip [Internet]. 2009. Disponible en: https://www.gnu.org/software/gzip/

20. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754-1760. doi:10.1093/bioinformatics/btp324

21. Batini C. EMBL-EBI NGS bioinformatics. SAM/BAM/CRAM format and BAM refinement [Internet]. 2021 ene. Disponible en: https://docs.google.com/presentation/d/1bleyP3NTR4e7lRb68EITwykG91S5TFyIUy7gReCf7w/edit?usp=sharing

22. . Batini C. EMBL-EBI NGS bioinformatics. Variant Calling and filtering. [Internet]. 2021 ene. Disponible en: https://docs.google.com/presentation/d/1KKbBHIcIdP5EIyup8dW8UIZ8_ojRb5A_1BQDI99GHkc/edit?usp=sharing

23. Sun R. NGS GUI - Manuals [Internet]. Ucr.edu. Disponible en: http://manuals.bioinformatics.ucr.edu/home/gui-ngs-analysis

24. Sequence Alignment/Map Format Specification the SAM/BAM Format Specification working group [Internet]. Github.io. 2022 Disponible en: https://samtools.github.io/hts-specs/SAMv1.pdf

25. Morjaria S. Driver mutations in oncogenesis. Int J Mol Immuno Oncol [Internet]. 2021 [citado el 4 de septiembre de 2022];6(100):100–2. Disponible en: https://ijmio.com/driver-mutations-in-oncogenesis/

26. Cancer Genome Interpreter - Identification of therapeutically actionable genomic alterations in tumors [Internet]. Cancergenomeinterpreter.org. Disponible en: https://www.cancergenomeinterpreter.org/analysis?id=9cfa8ee976aebfc60541

27. Forbes SA, Beare D, Boutselakis H, Bamford S, Bindal N, Tate J, et al. COSMIC: somatic cancer genetics at high-resolution. Nucleic Acids Res [Internet]. 2017 ;45(D1):D777–83. Disponible en: https://cancer.sanger.ac.uk/cosmic/analyses

28. Lenet S, Polychronakos C. Endocrine abnormalities in cardiofaciocutaneous syndrome: a case of precocious puberty, hyperprolactinemia and diabetes insipidus. BMJ Case Rep [Internet]. 2019 ;12(6):e229032. Disponible en: http://dx.doi.org/10.1136/bcr-2018-229032

29. Martínez E. Nomenclatura de las Mutaciones [Internet]. 2021 sep. Disponible en: https://idoc.pub/documents/nomenclatura-de-las-mutaciones-vnd5vp92qjlx