Universidad de la Republica

Licenciatura en Biología Humana

Informe de Pasantía de Grado

2024

Análisis Bioinformático de un Panel Genómico (NGS) para detectar variantes patogénicas del gen del factor VIII de la coagulación.

Autor: Karen Patricia Viera Tejera

Tutor: M.Sc. Yasser Vega

PDU Diversidad Genética Humana, CUT UDELAR, Cenur noreste

Orientador de Pasantía: Dra. Lucia Spangenberg

Unidad de Bioinformática, Instituto Pasteur de Montevideo

RESUMEN

La hemofilia A es una coagulopatía hereditaria ligada al cromosoma X mas frecuente, es causada por mutaciones deletéreas en el gen del factor VIII (FVIII) de coagulación (*F8*). Existen más de 3500 mutaciones diferentes reportadas en el gen *F8* asociadas con los fenotípos, leves, moderados y severos. Afecta principalmente a los varones, sin embargo las mujeres portadoras pueden manifestar fenotipos desde leves a severo. Su diagnóstico constituye una etapa crítica en la atención, ya que involucra decisiones médicas que tendrán un impacto directo en la evolución del cuidado individual.

En este trabajo se desarrolló un pipeline de análisis bioinformático propio para detectar variantes patogénicas en el gen F8 a partir de datos de NGS generados de un panel de genes asociados a las coagulopatías hereditarias relevantes.Fueron aplicandas diferentes herramientas como BWA, Samtools, picard, GATK y annovar, para el mapeo, filtrado, llamada de variantes, y anotación de loas variantes funcionales, y sus existencia en diferentes bases de datos.

Como resultado se detecto una variante patogénica que consiste en la inserción de dos pares de bases (GA) en el exón 14*,* la cual cambia el marco de lectura generando un codón stop anticipado que genera una proteína truncada lo que explica el fenotipo severo del individuo. El análisis con la herramienta annovar mostró que la variante aún no ha sido reportada en ninguna base de datos. Este flujo de trabajo bioinformático permitió la detección de la variante patogénica esperada, así como de diferentes SNPs e indels en el gen del factor VIII de la coagulación, lo que posibilita que se realice de forma rápida y confiable el diagnóstico genético molecular de los pacientes con hemofilia y se detecten las portadoras lo que posibilita el asesoramiento genético.

Palabras Claves: Bioinformática, Flujo de trabajo, Coagulopatías, Hemofilia A, Factor VIII

Introducción

Según la Federación Mundial de Hemofilia (WFH) en Uruguay existen más de 235 pacientes con hemofilia, sin embargo, no se han realizado estudios a nivel nacional para estudiar las mutaciones causales, existe solamente un reporte del estudio de la Inversión del intrón 22 y del Intrón 1 en la región noreste en pacientes con hemofilia A severa (Vega., el al 2020). Es importante resaltar que dado el gran número de mutaciones reportadas en el gen F8 su estudio molecular es complejo y laborioso, por lo que puede ser beneficiosa la aplicación de tecnologías de secuenciación de alto rendimiento (NGS), en este sentido, actualmente en el CENUR Noreste - UdelaR junto a la Plataforma Genómica de la Facultad de Ciencias - UdelaR, se está desarrollando un protocolo de NGS para detectar las variantes patogénicas del gen F8 en familias con hemofilia A de todo el país.

En este trabajo se busca desarrollar un pipeline de análisis bioinformático propio para detectar variantes patogénicas en el gen F8 a partir de datos de NGS generados de un panel de genes asociados a las coagulopatías hereditarias relevantes y busca validar y mejorar los resultados obtenidos anteriormente con programas como Geneious prime.

Contexto biolígico de la enfermedad

La hemostasia se define como un conjunto de procesos biológicos cuya función es preservar la fluidez sanguínea y la integridad del sistema vascular, evitando y deteniendo la pérdida de sangre tras una lesión. Una vez alcanzado este objetivo, es crucial garantizar la eliminación del tapón hemostático para restaurar el flujo sanguíneo. Un equilibrio adecuado en este sistema contribuirá a limitar tanto el sangrado como la formación de trombos patológicos.La primera línea de defensa para detener la hemorragia se basa en la formación del tapón o trombo plaquetario, en el cual participan las plaquetas y el endotelio vascular. Simultáneamente, las proteínas del plasma activan la coagulación, dando lugar a la generación de fibrina y la formación de un trombo estable. De manera coordinada, se activan mecanismos anticoagulantes que previenen la obstrucción del sistema vascular debido a la propagación del coágulo. Finalmente, el sistema de fibrinólisis se encarga de disolver el coágulo una vez que la lesión ha sido reparada (Moraleda J, 2017).

Tanto la hemorragia como el sangrado excesivo pueden originarse por enfermedades, ya sean congénitas o adquiridas, que afectan a los vasos sanguíneos, plaquetas o factores procoagulantes. Además, la disminución de los anticoagulantes fisiológicos o la alteración

de la fibrinólisis pueden conducir al desarrollo de trombos patológicos. La hemostasia se desglosa en dos componentes principales: la hemostasia primaria, que depende de las plaquetas y los vasos sanguíneos, y la hemostasia secundaria, que se apoya en las proteínas de la coagulación (Roberts & Monroe, 2004).

La hemostasia primaria es el resultado de interacciones complejas entre proteínas adhesivas de la pared vascular y plaquetas, generando un trombo blanco rico en plaquetas. Fisiológicamente, las paredes vasculares mantienen propiedades antitrombóticas a través del endotelio, que sintetiza inhibidores de la activación plaquetaria, de la coagulación y activadores de la fibrinólisis. Sin embargo, agresiones externas transforman estas propiedades en protrombóticas, caracterizadas por la liberación de factores activadores de plaquetas, exposición de fosfolípidos aniónicos, y la presencia de componentes trombogénicos como el colágeno y el factor de Von Willebrand tras una lesión. La vasoconstricción de las arteriolas es la primera respuesta al daño vascular, reduciendo la pérdida de sangre mediante la liberación de serotonina y tromboxano A2 por parte de las plaquetas y el endotelio (Ruggeri, 2002).

Por otra parte la hemostasia secundaria o coagulación es un conjunto de reacciones bioquímicas que conducen a la transformación del fibrinógeno (soluble) en fibrina (insoluble), lo que da estabilidad al trombo tras la lesión de un vaso. En el proceso de la coagulación intervienen una serie de complejos enzimáticos en los que, además de la enzima y el sustrato, es necesaria la presencia de cofactores proteicos, fosfolípidos y calcio, que interaccionan entre sí para acelerar la velocidad de la reacción y aumentar su eficacia. Pero las reacciones procoagulantes que conducen a la formación de fibrina deben estar en un perfecto equilibrio con: 1) reacciones limitantes anticoagulantes que impidan la acción incontrolada de los factores de coagulación activados y eviten una coagulación generalizada, y 2) reacciones fibrinolíticas que se encarguen de eliminar la fibrina cuando ya no sea necesaria y de restablecer el flujo sanguíneo. Estos procesos son dinámicos y están estrictamente regulados, y su alteración puede ocasionar episodios tanto hemorrágicos como trombóticos (Moraleda J, 2017).

Todos los factores de la coagulación excepto el FvW (Factor de von Willebrand) se sintetizan en el hígado y circulan en la sangre periférica, salvo el factor tisular, que se encuentra en las membranas de ciertas células. Los factores V, XI y XIII también se encuentran en plaquetas. Los factores II, VII, IX y X, y las proteínas C y S necesitan de la vitamina K para que sean completamente funcionantes. Esta vitamina participa en la

gammacarboxilación de los residuos de ácido glutámico de estas proenzimas, lo que permite (a través del calcio) la unión de estos factores a los fosfolípidos de las superficies celulares. De esta manera, las reacciones entre factores, que en fase líquida serían muy poco eficientes, se realizan sobre superficies fosfolipídicas (membranas de plaquetas y, en menor medida, de células endoteliales) creando complejos enzimáticos que aumentan la eficiencia de la reacción (Moraleda J, 2017).

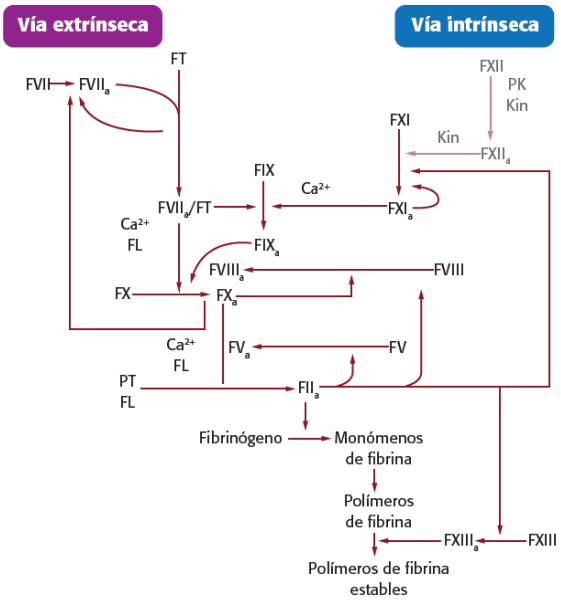

Figura 1. *Representación de la cascada de coagulación de iniciación de la formación del coágulo mediada por factor tisular; interacciones entre las vías y papel de la trombina en el mantenimiento de la casacasa por mecanismos de retroactivación de factores de coagulación. (tomado de Moraleda J, 2017)*

Una falla en unos de los componentes de esta cascada puede llegar a causar una coagulopatía.

Las coagulopatías constituyen un grupo heterogéneo de enfermedades que cursan con diátesis hemorrágica, y que son producidas por alteraciones de las proteínas plasmáticas de la hemostasia primaria (por ejemplo el factor de von Willebrand [FvW]), de la coagulación o de la fibrinólisis. Se clasifican en congénitas o hereditarias y adquiridas. Las coagulopatías congénitas pueden ser consecuencia del defecto selectivo de un factor, o raramente , una combinación de 2 mas defectos. Estas alteraciones de la hemostasia peden ser de tipo cuantitativo o cualitativo (moléculas disfuncionales o variantes) (Sarmiento *et al*., 2001).

Las coagulopatías hereditarias mas frecuentes son la hemofilia A (déficit del factor VIII de la coagulación), la hemofilia B (déficit del factor IX de a coagulación) y la enfermedad de Von Willebrand (EVW). La hemofilia A (HA) y la hemofilia B (HB) son enfermedades recesivas ligadas al cromosoma X que afectan a 1 en 5.000 y 1 en 30.000 varones respectivamente, producidas por mutaciones en los genes *F8 y F9* (Peyvandi *et al*., 2016).

La HA se diagnostica en función de los antecedentes de hemorragia y las pruebas de coagulación anormales en los pacientes y sus familias. Las anomalías en la prueba de coagulación incluyen tiempo de tromboplastina parcial activada (TTPA) prolongado y actividad disminuida del FVIII. La HA se clasifica en tres fenotipos según los niveles de actividad del FVIII en plasma: grave (<1% de la actividad normal), moderada (1 a 5%) y leve (5 a 40%). El fenotipo grave ocupa la mayoría de los casos (60%), moderado en el 15% y leve en el 25% de todos los casos [4]. Sin embargo, sólo aproximadamente el 30% de las mujeres heterocigotas HA tienen una actividad del FVIII inferior al 40%, lo que no es fiable para el diagnóstico (Wang *et a*l., 2022).

La HA es causada por mutaciones deletéreas en el gen del factor VIII (FVIII) de coagulación (*F8*), casi la mitad de las HA severas están asociadas a inversiones del *F8,* la inversión del intrón 22 (Inv22) (Lakich *et al*., 1993) y del intrón 1 (Inv1). (Bagnall *et al*., 2002). En el intrón 22 del F8 se extiende una secuencia de 9,5 kb (int22h-1) que posee dos copias extragénicas (ints22h-2 y-3) a una distancia de alrededor de 500 kb a 600 kb del *F8.* Estas copias promueven la recombinación intracromosómica principalmente durante la gametogénesis masculina. Por su parte, el intrón 1 posee un segmento de 1041 pb (int1h-1) con una copia extragénica (int1h-2) a unas 140 kb hacia el telómero Xq, cuya recobinación puede generar la Inv1 (Bagnall *et al.*, 2002).

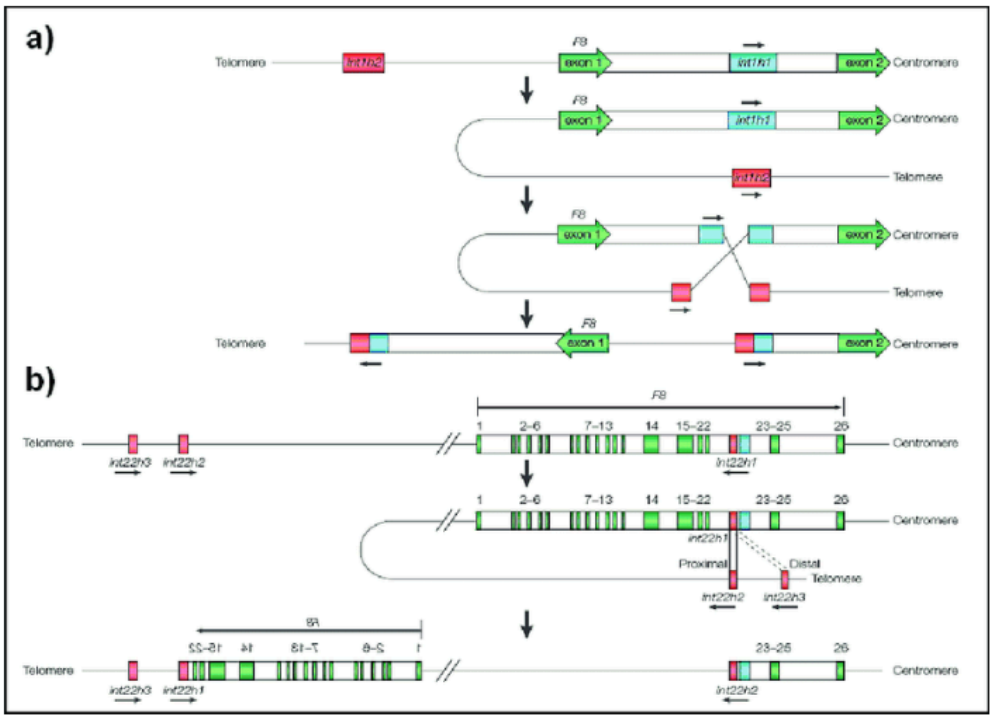

Figura 2. *Inversión del intrón 1 (a) y del intrón 2 (b) del gen factor VIII, por recombinación homóloga intracromosómica defectuosa (tomado de Rodriguez et al., 2009).*

Por otra parte, existen más de 3500 mutaciones diferentes reportadas en el gen *F8* asociadas con los fenotípos, leves, moderados y severos (Base de datos de variantes del factor VIII de EAHAD, 2022).

El patrón de herencia de la hemofilia, una enfermedad ligada al cromosoma X, es recesivo. Esto significa que la mayoría de los hombres afectados manifiestan los síntomas de la enfermedad, mientras que las mujeres parientes pueden ser heterocigotas para la mutación, a menudo denominadas portadoras de hemofilia (Rodríguez-Martorell *et al*., 2009). Las mujeres con ese gen anómalo tienen una posibilidad del 50% de transmitir esa característica genética a sus hijos varones, lo que provocará que la enfermedad se exprese en los hijos que reciban ese gen.

Los hombres con hemofilia transmiten su cromosoma Y a sus hijos varones, pero no transmiten el gen defectuoso a sus hijos varones (Franchini & Mannucci., 2017).

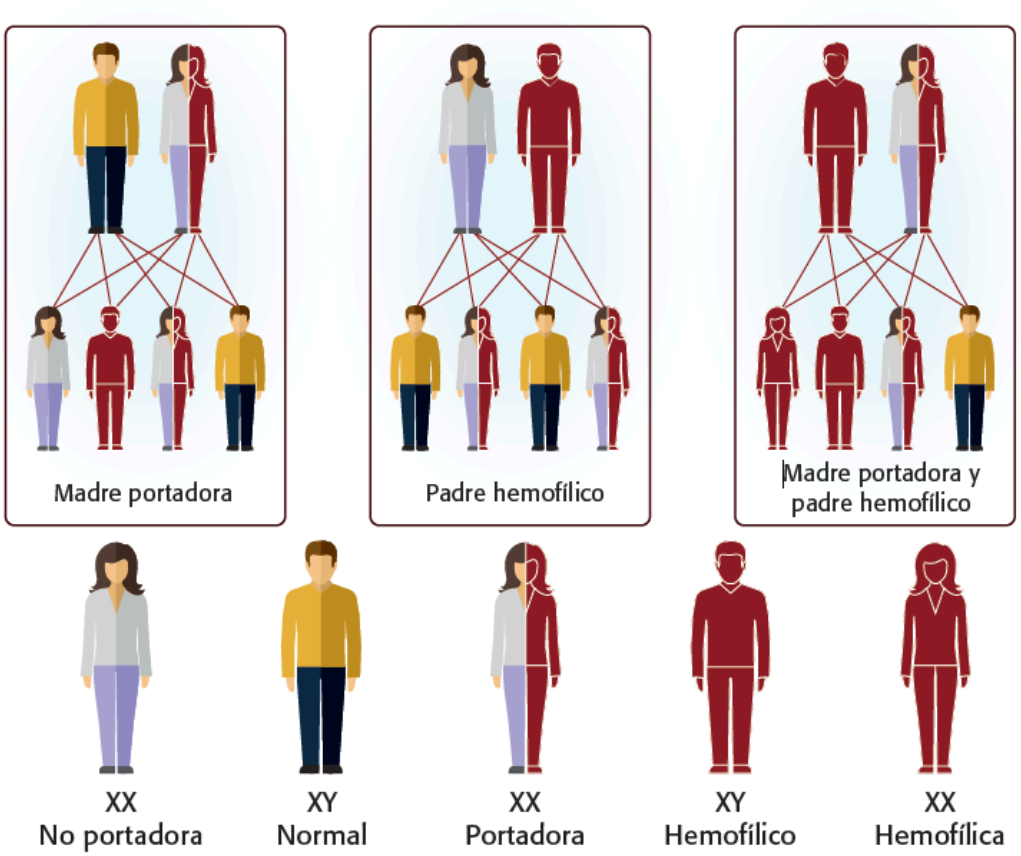

Figura 3. *Patrón de herencia de la hemofilia A y B (ligada al cromosoma x) (tomado de Moraleda J, 2017).*

Estado del arte del diagnóstico molecular y Análisis Bioinformático.

El diagnóstico constituye una etapa crítica en la atención de un paciente, ya que involucra decisiones médicas que tendrán un impacto directo en la evolución del cuidado individual. En este sentido, el respaldo diagnóstico emerge como un servicio fundamental para el equipo médico y, consecuentemente, para la institución que lo proporciona (Somak, *et al,* 2018).

La bioinformática desempeña un papel crucial en el diagnóstico al procesar datos masivos, como los generados por la secuenciación masiva NGS (por sus siglas en inglés, next generation sequence) permitiendo detectar alteraciones genéticas que impactan directamente en la comprensión y el manejo de las enfermedades redundando en la atención al paciente. Sanger fue una de las primeras técnicas de secuenciación basada en un método de didesoxinucleótidos y usada como tecnología convencional. Fue desarrollada por el bioquímico británico Frederick Sanger a mediados del siglo XX, y permite analizar la secuencia de nucleótidos que componen una molécula de ADN. En la actualidad, la secuenciación de Sanger sigue utilizándose en los laboratorios, pero presenta ciertas limitaciones, (Secuenciación por Sanger, 2010) funciona a bajo

rendimiento y su costo es elevado por lo que sólo es adecuada para validaciones y ensayos (Carneiro et al., 2019). Debido a esto, se han desarrollado diversas técnicas de secuenciación de nueva generación, mucho más rápidas y económicas.

NGS es una estrategia de secuenciación de alto rendimiento que permite generar millones de secuencias, tanto de genes de interés como exomas e incluso genomas completos en una única corrida de secuenciación.

NGS se vuelve cada vez más costo-efectiva, ya que desde hace varios años viene bajando los precios exponencialmente (Somak. *et al.*, 2018). Además, la secuenciación masiva tiene el potencial de detectar todos los tipos de variación genómica en un único experimento, incluyendo variantes de nucleótido único o mutaciones puntuales, pequeñas inserciones y deleciones, y también variantes estructurales tanto equilibradas (inversiones y traslocaciones) como desequilibradas (deleciones o duplicaciones) (Rodriguez & Armengol, 2012).

La secuenciación de segunda generación consiste en plataformas que producen gran cantidad de lecturas cortas de secuencias de ADN (Imelfort & Edwards, 2009)

En esta generación destaca la empresa Illumina Inc. (San Diego, Ca, USA), actualmente dominante en el mercado, la cual utiliza el método de secuenciación por síntesis (sequencing by synthesis, SBS), permitiendo la lectura paralela de millones de fragmentos por medio de la detección de las bases individuales a medida que se incorporan a las cadenas de ADN en crecimiento. La ADN polimerasa cataliza la incorporación de desoxirribonucleótidos trifosfato (dNTP) marcados con fluorescencia a una hebra molde de ADN durante ciclos secuenciales de síntesis. Durante cada ciclo los nucleótidos se identifican mediante excitación del fluoróforo (Sequencing Technology | Sequencing by Synthesis, 2024), creando una imagen de un terminador reversible marcado fluorescentemente a medida que se añade cada dNTP, y luego se separa para el ingreso de la siguiente base.

La tecnología de secuenciación por síntesis (SBS) permite secuenciar ambos extremos de fragmentos de ADN provenientes de una biblioteca genómica. Al secuenciar los dos extremos de cada fragmento, se obtienen lecturas de secuencia de alta calidad. La clave de la secuenciación de extremos emparejados (paired-end sequencing) es que se conoce la distancia aproximada entre cada par de lecturas generadas. Esto ayuda a mejorar la alineación de las lecturas al genoma de referencia y facilita el ensamblaje de novo de genomas. Al tener pares de lecturas alineadas, se puede detectar con mayor precisión variantes estructurales, como inserciones, deleciones, inversiones o traslocaciones(Sequencing Technology | Sequencing by Synthesis, 2024)

En resumen, la secuenciación de extremos emparejados mejora la calidad, la alineación y el ensamblaje de las secuencias genómicas, lo que a su vez permite un análisis más preciso de variantes y estructuras genéticas complejas.
Al finalizar, se obtiene una secuenciación base por base con datos de alta precisión y calidad, para una gran variedad de usos (Rojas et al,. 2024)

Por otro lado está IonTorrent (actualmenteThermo Fisher Scientific, Waltham, Ma, USA) que en 2010 comercializó su sistema PGM (Personal Genome Machiné), que incorporaba la tecnología de semiconductores y no dependía del uso de fluorescencia, sino de los cambios en pH que se producen cuando se libera un protón al incorporarse a una molécula de ADN (López, 2016).

Análisis Bioinformático para la detección de variantes

Los algoritmos de bioinformática que se ejecutan en una secuencia predefinida para procesar los datos de NGS se denominan colectivamente flujo de trabajo de bioinformática de NGS.
Un flujo de trabajo de bioinformática guía y procesa progresivamente datos de secuencias masivas y sus metadatos asociados a través de una serie de transformaciones utilizando múltiples componentes de software, bases de datos y entornos operativos (Somak. *et al.*, 2018) y consta de los siguientes pasos principales:

1) Alineación de secuencias
2) Detección de Haplotipos
3) Filtrado de variantes
4) Anotación de variantes
5) Priorización de variante

1) **Alineamiento de secuencias**

La alineación de las secuencias es el proceso de determinar dónde se alinea cada secuencia corta de ADN leída con un genoma de referencia.

Este proceso computacional intensivo asigna una puntuación de calidad a cada una de estas pequeñas lecturas de secuencia. Esta puntuación indica qué tan confiable es el proceso de alineación de esa lectura. También proporciona información sobre la ubicación de cada lectura en el genoma de referencia. Esto permite calcular que proporción de las lecturas se pudieron alinear con el genoma (lecturas mapeadas) y cuántas veces se leyó cada región del genoma (profundidad o cobertura de la secuenciación) (Canal-Alonso, et al., 2021)

Exiten algoritmos para efectuar este proceso de alineación. Esto algoritmos deben contemplar aspectos como la capacidad de alinear millones de secuencias, considerar que la correlación puede no ser única cuando las lecturas son cortas, y también hay que considerar que la secuenciación no es perfecta y existe una pequeña probabilidad de que una base sea incorrecta, además de las variantes reales de secuencia. También hay que tener presente que hay regiones del DNA que son difíciles de secuenciar, como los homopolímeros o las regiones con alto contenido de nucleótidos GC, y esto puede aumentar la probabilidad de error en la secuenciación. Todo esto hace que el proceso de alineamiento sea complejo (lópez, 2016)

Burrows-Wheeler Aligner BWA es un paquete de programas para mapear secuencias cortas respecto a un gran genoma de referencia.(Li & Durbin, 2009)

Para cada alineación, BWA calcula una puntuación de calidad de mapeo. También admite el mapeo de extremos emparejados. Primero encuentra las posiciones de todos los aciertos positivos, los clasifica según las coordenadas cromosómicas y luego realiza un escaneo lineal de todos los aciertos potenciales para emparejar los dos extremos.En el emparejamiento, BWA procesa pares de lecturas en un lote. En cada lote, BWA carga el índice BWA completo en la memoria, genera la coordenada cromosómica para cada aparición, estima la distribución del tamaño de inserción a partir de pares leídos con ambos extremos mapeados con una calidad de mapeo superior a 20 y luego los empareja. Después de eso, BWA borra el índice BWT de la memoria, carga la secuencia de referencia codificada de 2 bits y realiza la alineación Smith-Waterman para lecturas no asignadas cuyas relaciones de posición pueden alinearse de manera confiable. La alineación Smith-Waterman rescata algunas lecturas con diferencias excesivas. (Li & Durbin, 2009)

En general los software de alineamiento generan un archivo en formato SAM (Sequence Alignment Map), que tiene el formato binario equivalente BAM (Li & Durbin 2009). Se trata de un formato genérico que contiene las lecturas y el alineamiento correspondiente

respecto a la secuencia de referencia, los archivos (SAM/BAM) contienen mucha información, tanto de las lecturas mapeadas como de las no mapeadas, e incluso se pueden recuperar las lecturas originales a partir de estos archivos. Las propiedades de las lecturas y de su mapeo, es decir, las propiedades que indican si la lectura se ha mapado correctamente, si su pareja ha mapeado, o si está duplicada, entre otros. Es necesario realizar un filtrado de aquellos alineamientos que no nos interesen según el objetivo del análisis, como por ejemplo lecturas que mapan en más de una posición puede que no nos interesen, o lecturas cuya pareja alinean en otro cromosoma, entre otras opciones que variarán según el experimento (López, 2016)

Para manipular archivos SAM y BAM se puede utilizar SAMtools que es una herramienta de bioinformática , que permite realizar diversas operaciones de análisis sobre los datos de alineación y lo separa de los análisis posteriores, permitiendo un enfoque genérico para el análisis de genómica datos de secuenciación. (Li et al,. 2010)

Apartir de los archivos generados con SAMtools y utilizando la suite de herramientas Picard que sirve para la manipulación, limpieza y validación de los archivos de alineación, se agregan grupo de lectura y se identifica y marcan los duplicados. (Broad Institute, 2019)

2) **Detección de Haplotipos**

La detección de haplotipos se refiere al proceso de determinar las variantes genéticas que se heredan juntas en un cromosoma. Un haplotipo es una secuencia específica de alelos o variantes de polimorfismos de nucleótido único (SNPs) pequeñas inserciones y eliminaciones (indeles), alteraciones del número de copias y grandes alteraciones estructurales (inserciones, inversiones y translocaciones), que se encuentran en un cromosoma y se heredan como una unidad (Canal-Alonso, et al., 2021)

Las herramientas de llamada de variantes SNP (polimorfismo de un solo nucleótido) e indel (inserción/deleción) se utilizan para identificar variaciones en las secuencias de ADN a nivel de un solo nucleótido. Estas herramientas se pueden dividir en dos categorías: métodos heurísticos y métodos probabilísticos. Los métodos heurísticos, utilizan múltiples fuentes de información sobre la calidad de los datos para asignar variantes y también pueden utilizar pruebas estadísticas, como la prueba de Fisher, para comparar las variantes con distribuciones teóricas. Los métodos probabilísticos, como SAMtools y GATK, se basan en enfoques bayesianos que optimizan la probabilidad de identificar genotipos. (Canal-Alonso, et al., 2021)

GATK, que fue desarrollado por el Broad Institute, y es una sofisticada suite de herramientas diseñada para el análisis de datos de secuenciación de genomas, con un enfoque particular en la identificación de variantes genómicas. En el contexto de la

búsqueda de variantes, GATK ha establecido estándares en la industria con sus buenas prácticas, empleando un enfoque basado en haplotipos (Canal-Alonso, et al., 2021)

3) Filtrado de Variantes

Es el proceso mediante el cual se identifican y eliminan las variantes que son artefactos falsos positivos generados por el método de secuenciación NGS. Estas variantes son marcadas o filtradas del archivo VCF original en función de diversos metadatos asociados a la alineación de secuencias y las llamadas de variantes, como la calidad de mapeo, calidad de llamada de base, sesgos en la cadena, entre otros (Canal-Alonso et al., 2021)
Se filtran SNPs e indels según algunos parámetros como la profundidad de secuenciación, (QD) el sesgo de hebra (Strand Bias - FS, FisherStrand)que mide si hay un desequilibrio en la distribución de las lecturas en las hebras de ADN. Un valor alto puede indicar una posible variante falsa debido a sesgos en la secuenciación (Broad Institud, 2024).
También existen otras estrategias basadas en modelos bayesiandos y gaussianos para modelar la incertidumbre y la variabilidad en los datos genómicos y mejorar la precisión en la identificación de variantes relevantes.

4) Anotación de variantes

La anotación de variantes es el proceso de asignar significado biológico a los resultados obtenidos de la llamada de variantes. Implica buscar información sobre variantes en varias bases de datos y recursos en línea, como dbSNP. (Sherry, et al., 1999) o el proyecto 1000 Genomas (Auton et al.,2015) y utilizando métricas polifenol (Adzhubei et al., 2013) para evaluar el impacto clínico potencial de una variante. Estas métricas proporcionan una puntuación de predicción basada en la anotación de la variante y clasifican la variante según su posible impacto 3 clínico. Las variantes se pueden clasificar como patógenas, neutras, posiblemente benignas o variantes de significado incierto (VUS), según el nivel de confianza en su importancia clínica (Roy et al., 2018) . Las variantes también se pueden definir según el efecto que tienen en la cadena de proteínas, como ser sinónimos o no sinónimos, o causar una mutación por cambio de marco. Las variantes no sinónimas dan como resultado un cambio en la secuencia de la proteína, lo que puede tener consecuencias funcionales, mientras que las variantes sinónimas no dan como resultado un cambio neto en la secuencia de la proteína debido a la degeneración del código genético. Las mutaciones de cambio de marco, que son causadas por la ganancia o pérdida de nucleótidos, pueden alterar la lectura normal de la secuencia de ADN y dar como resultado una secuencia de proteínas completamente diferente (Canal-Alonso et al., 2021)

5) **Priorización de Variante**

Es el proceso mediante el cual se utilizan anotaciones de variantes para identificar y separar las variantes que son clínicamente insignificantes (por ejemplo, variantes sinónimas, intrónicas profundas y polimorfismos benignos establecidos) de aquellas que pueden tener importancia clínica conocida o desconocida. Facilitando su revisión e interpretación para una toma de decisiones clínicas precisa (Roy et al., 2018))

OBJETIVO GENERAL

Desarrollar un flujo de trabajo bioinformático para la detección de mutaciones en el gen *F8* a partir de datos de un panel de NGS.

BJETIVOS ESPECÍFICOS

- Identificación de variantes en el gen F8 a partir de datos de NGS en pacientes de hemofilia.
- Priorizar las variantes encontradas y asociarlas con el fenotipo del paciente.

METODOLOGÍA

Se analizaron los archivos fastq (R1 y R2) obtenidos de un panel de genes que incluye los genes, *F8, F9, F5, F7 y FVW* (generados en la plataforma genómica de la facultad de ciencia, Udelar) de un individuo con hemofília A severa y que posee una variante patogénica en el gen *F8*. Para esto utilizamos un flujo de trabajo ya creado por el instituto Pasteur de Uruguay para la obtención de variantes de línea germinal corta (SNP e INDELS), basado en buenas prácticas de GATK. (Spangenberg *et al*., 2022).

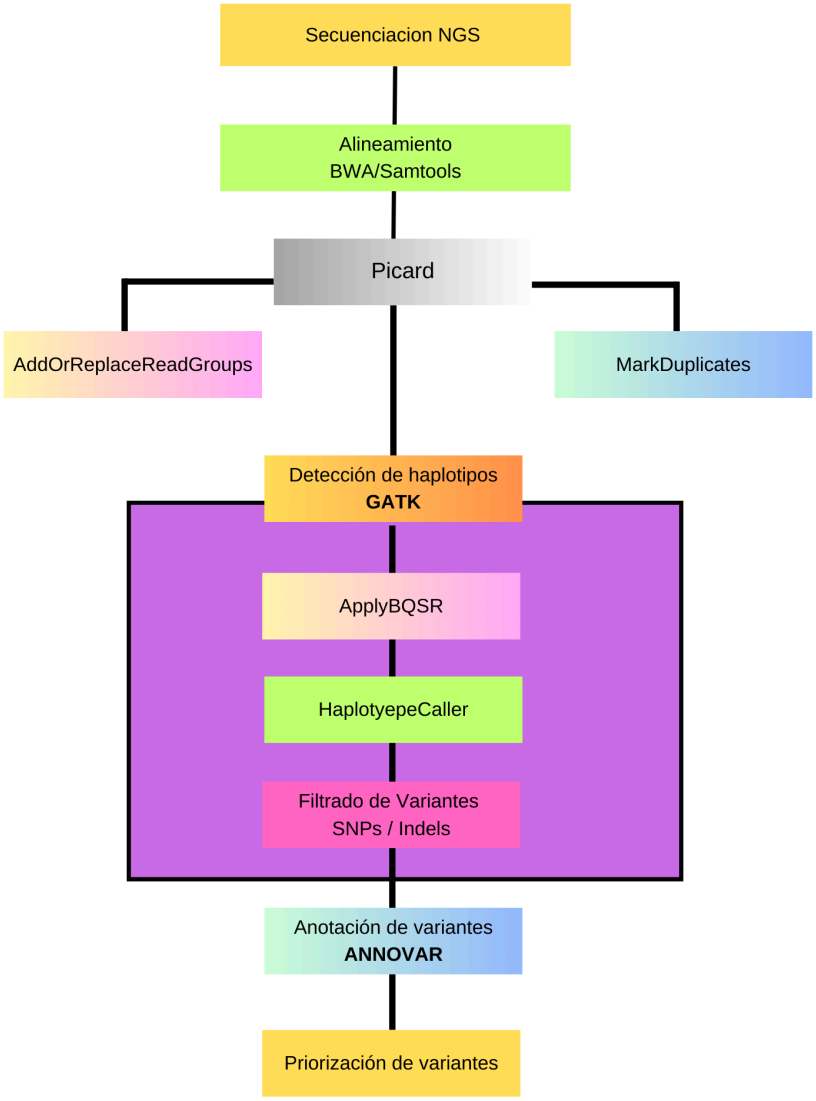

***Figura 4**. Esquema del algoritmo de trabajo aplicado. (Diseño propio)*

El script de la instalación de las herramientas utilizadas estará adjunto al anexo.
Como genoma de referencia utilizamos la versión **hg37.fa,** (ya que su disponibilidad y uso extendido facilitan la comparación de los resultados) el cual fue situado en el directorio bwa_index, asignando la siguiente ruta:
~/short_germline_variants-main/bundle/hg37/bwa_index

Dentro de este directorio fue indexado el genoma de referencia

```
bwa index /home/patricia/short_germline_variants-main/bundle/hg37/bwa_index/hg37.fasta
```

Esto generó los siguientes archivos hg37.fa.amb - hg37.fa.ann - hg37.fa.bwt - hg37.fa.pac - hg37.fa.sa, los cuales permiten a BWA realizar búsquedas eficientes y alineamientos rápidos durante el proceso de mapeo. Cuando se usa BWA para alinear secuencias de lecturas contra un genoma de referencia, el programa utiliza estos archivos para ubicar rápidamente las posiciones potenciales de alineamiento en el genoma.

1) **Alineamiento**

Implica colocar las lecturas en sus posiciones correctas dentro del genoma, para ello utilizamos la herramienta bwa aln (Burrows-Wheeler Aligner), para alinear la secuencia de estudio en formato fastq (Lyb-20HAS-S46_L001_R1_001 y Lyb-20HAS-S46_L001_R2_001) con el genoma de referencia (hg37).

```
bwa aln /home/patricia/short_germline_variants-main/bundle/hg38/bwa_index/hg37.fasta
/home/patricia/short_germline_variants-main/rawdata/Lyb-20HAS_S46_L001_R1_001.fastq >
/home/patricia/short_germline_variants-main/pipeline/bwa_results/Lyb-20HAS_S46_L001_R1.sai &

bwa aln /home/patricia/short_germline_variants-main/bundle/hg38/bwa_index/hg37.fasta
/home/patricia/short_germline_variants-main/rawdata/Lyb-20HAS_S46_L001_R2_001.fastq >
/home/patricia/short_germline_variants-main/pipeline/bwa_results/Lyb-20HAS_S46_L001_R2.sai &

wait
```

Se realizó el emparejamiento de las lecturas previamente alineadas en formato .sai (generedas en el paso anterior) lo que generó un archivo SAM que contiene las alineaciones de las lecturas contra el genoma de referencia.

```
bwa sampe /home/patricia/short_germline_variants-main/bundle/hg37/bwa_index/hg37.fasta
Lyb-20HAS_S46_L001_R1.sai Lyb-20HAS_S46_L001_R2.sai
/home/patricia/short_germline_variants-main/rawdata/Lyb-20HAS_S46_L001_R1_001.fastq
/home/patricia/short_germline_variants-main/rawdata/Lyb-20HAS_S46_L001_R2_001.fastq >
Lyb-20HAS46.sam
```

Fueron filtradas las alineaciones para mantener solo aquellas que tienen valor de calidad de mapeo mayor que 1, lo que generalmente indica una buena alineación. El resultado se guardó en unarchivo .unique.sam

```
samtools view  -Sh -q 1 Lyb-20HAS46.sam > Lyb-20HAS46_unique.sam
```

Posteriormente fueron filtradas las alineaciones del archivo unique.sam para mantener solo aquellas que son pares correctamente alineados.

Este resultado se almacenó en un archivo properpair.sam

```
samtools view -Sh -Sf 0x0002 Lyb-20HAS46_unique.sam >Lyb-20HAS46_properpair.sam
```

Se convertió el archivo .SAM en archivo .BAM que es una versión binaria comprimida e

indexada del formato .SAM, lo que lo hace mas eficiente para almacenar grandes cantidades de datos de alineamiento. Se utilizó el siguiente comando:

```
samtools view -bS Lyb-20HAS46_properpair.sam > Lyb-20HAS46.bam
```

Luego, se ordenó el archivo .BAM generado:

```
samtools sort Lyb-20HAS46.bam -o Lyb-20HAS46_sorted.bam
```

Para verificar cuantos reads mapearon

```
samtools view -c -F 4 Lyb-20HAS_sorted.bam
```

En este caso mapearon 373834 reads.

Se utilizó la el suite de herramientas Picard,que se utiliza para manipular datos y formatos de secuenciación de alto rendimiento como SAM/BAM/CRAM y VCF (Broad Institud, 2024).

AddOrReplaceReadGroups : Herramienta que se utiliza en el preprocesamiento de datos de secuenciación para agregar o reemplazar grupos de lecturas en archivos de datos de secuenciación. Tiene como funcionalidad asignar o modificar la información de los grupos de lecturas, como la biblioteca de origen, la plataforma de secuenciación, el identificador de la muestra, entre otros (Broad Institute, 2024).

```
sudo java -jar /home/patricia/short_germline_variants-main/picard-3.1.1/picard/build/libs/picard.jar
AddOrReplaceReadGroups
I=/home/patricia/short_germline_variants-main/pipeline/bwa_results/Lyb-20HAS_sorted.bam
O=/home/patricia/short_germline_variants-main/pipeline/picard_results/Lyb-20HAS.replaced RGID=ID_1
RGLB=LIBRARY_1 RGPL=ILLUMINA RGPU=UNIT_1 RGSM=SAMPLE
```

MarkDuplicates: se utiliza para identificar y marcar duplicados de lecturas en archivos de datos de secuenciación. Los duplicados pueden surgir durante el proceso de secuenciación y pueden afectar la precisión de los resultados (Broad Institute, 2024)

```
sudo java -jar /home/patricia/short_germline_variants-main/picard-3.1.1/picard/build/libs/picard.jar
MarkDuplicates -I
/home/patricia/short_germline_variants-main/pipeline/picard_results/Lyb-20HAS.replaced -O
/home/patricia/short_germline_variants-main/pipeline/picard_results/Lyb-20HAS.marked.bam -M
Lyb-20HAS.metrics
```

Para el siguiente paso fue descargado del servidor Ftp del Broat Institud el directorio bundle del genoma de referencia hg37. Dentro de este directorio encontraremos recursos genómicos utilizados en el análisis genéticos y genómicos, como por ejemplo:
dbnsp_137.hg37.vcf un archivo vcf (Variant Call Format) con variantes genética de dbSNP (Single Nucleotide Polymorphism Database)
Mills_and_1000G_gold_standard.indels.hg38.vcf otro archivo vcf que contiene inserciones y deleciones identificadas por el proyecto 1000 Genomas y el proyecto de indes de Mills en el genoma humano.

Descargado con el siguiente comando:

```
wget --ftp-user=gsapubftp-anonymous --ftp-password=contraseña -r
ftp://ftp.broadinstitute.org/bundle/hg37
```

2) Detección de Haplotipos

Para este analisis se utilizará herramientas bioinformáticas de GATK cuyo algoritmo utiliza un modelo probabilístico para determinar la probabilidad de que haya diferentes variantes en una posición dada, basándose en las lecturas observadas (Canal-Alonso, et al., 2021)

BaseRecalibrator: filtra y recalibra variantes basadas en múltiples métricas para asegurar que las variantes finales sean de alta calidad, tomando como archivo de entradael archivo .marked.bam.

```
sudo /home/patricia/miniconda3/bin/python /home/patricia/short_germline_variants-main/gatk-4.5.0.0/gatk
--java-options "-Xmx16g" BaseRecalibrator -R
/home/patricia/short_germline_variants-main/bundle/hg37/bwa_index/hg37.fasta -I
/home/patricia/short_germline_variants-main/pipeline/picard_results/Lyb-20HAS.marked.bam -O
/home/patricia/short_germline_variants-main/pipeline/gatk_results/20HAS.recalibrated.tab --known-sites
/home/patricia/short_germline_variants-main/bundle/hg37/1000G_phase1.snps.high_confidence.hg37.vcf
--known-sites
/home/patricia/short_germline_variants-main/bundle/hg37/Mills_and_1000G_gold_standard.indels.hg37.vcf
--known-sites /home/patricia/short_germline_variants-main/bundle/hg37/dbsnp_137.hg37.vcf
```

ApplyBQSR: mejorar la precisión de las llamadas de variantes y la calidad de los datos de secuenciación. Toma como entrada y recalibrando la tabla generada en el paso anterior.

```
sudo /home/patricia/miniconda3/bin/python /home/patricia/short_germline_variants-main/gatk-4.5.0.0/gatk
--java-options "-Xmx16g" ApplyBQSR -R
/home/patricia/short_germline_variants-main/bundle/hg37/bwa_index/hg37.fasta -I
/home/patricia/short_germline_variants-main/pipeline/picard_results/Lyb-20HAS.marked.bam --bqsr-recal-file
/home/patricia/short_germline_variants-main/pipeline/gatk_results/20HAS.recalibrated.tab -O
/home/patricia/short_germline_variants-main/pipeline/gatk_results/20HAS_S46.printed.bam
```

HaplotypeCaller: Algoritmo que utiliza GATK para la detección de variantes y se basa en considerar la evidencia de variantes en un contexto más amplio del genoma, en lugar de examinar cada posición de forma aislada. Esto permite una identificación más precisa de los haplotipos y, en consecuencia, una detección más exacta de las variantes genéticas. Este algoritmo utiliza un modelo probabilístico para determinar la probabilidad de que haya diferentes variantes en una posición dada, basándose en las lecturas observadas. Las variantes propuestas se califican en función de su calidad, y solo se retienen aquellas que superan ciertos umbrales para un análisis posterior. (Canal-Alonso, et al., 2021)
Proporciona como salida un archivo VCF (por su sigla en inglés, Variant Call Format)

```
sudo /home/patricia/miniconda3/bin/python /home/patricia/short_germline_variants-main/gatk-4.5.0.0/gatk
--java-options "-Xmx16g" HaplotypeCaller -R
/home/patricia/short_germline_variants-main/bundle/hg37/bwa_index/hg37.fasta  -I
/home/patricia/short_germline_variants-main/pipeline/gatk_results/20HAS_S46.printed.bam -O
/home/patricia/short_germline_variants-main/pipeline/gatk_results/20HAS_S46.raw_variants.vcf
-stand-call-conf 30 --min-pruning 3
```

3) **Filtrado de Variantes**

Se trata de la identificación y exclusión de variantes que se consideran errores falsos positivos. Se filtran tanto los SNPs como indels en base algunos parámetros como la profundidad de la variante , (QD - Quality by Depth) la probabilidad que la variante sea un error de mapeo (FS - Fisher Strand) y la calidad de asignación de las lecturas (MQ - Mapping Quality).

Selección de SNPs

```
sudo  /home/patricia/miniconda3/bin/python
/home/patricia/short_germline_variants-main/gatk-4.5.0.0/gatk --java-options "-Xmx4g" SelectVariants -R
/home/patricia/short_germline_variants-main/bundle/hg37/bwa_index/hg37.fasta -V
20HAS_S46.raw_variants.vcf --select-type-to-include SNP -O 20HAS_S46.raw_snps.vcf
```

Selección de INDELs

```
sudo /home/patricia/miniconda3/bin/python   /home/patricia/short_germline_variants-main/gatk-4.5.0.0/gatk
--java-options "-Xmx4g" SelectVariants -R
/home/patricia/short_germline_variants-main/bundle/hg37/bwa_index/hg37.fasta -V
20HAS_S46.raw_variants.vcf  --select-type-to-include INDEL -O 20HAS_S46.raw_indels.vcf
```

Filtrar SNPs

```
 sudo  /home/patricia/miniconda3/bin/python
/home/patricia/short_germline_variants-main/gatk-4.5.0.0/gatk --java-options "-Xmx4g" VariantFiltration -R
/home/patricia/short_germline_variants-main/bundle/hg37/bwa_index/hg37.fasta  -V
20HAS_S46.raw_snps.vcf --filter-expression "DP < 100" --filter-name "DPFilter" --filter-expression "QD <
2.0" --filter-name "QDFilter" --filter-expression "FS > 60.0" --filter-name "FSFilter" --filter-expression "MQ <
40.0" --filter-name "MQFilter"  --filter-expression "MQRankSum <-12.5" --filter-name "MQRankSumFilter"
--filter-expression "ReadPosRankSum <-8.0" --filter-name "ReadPosRankSumFilter" -O
20HAS_S46.filtered_snps.vcf
```

Filtrar INDELs

```
sudo /home/patricia/miniconda3/bin/python
/home/patricia/short_germline_variants-main/gatk-4.5.0.0/gatk --java-options "-Xmx4g" VariantFiltration -R
/home/patricia/short_germline_variants-main/bundle/hg37/bwa_index/hg37.fasta  -V
20HAS_S46.raw_indels.vcf --filter-expression "QD < 2.0" --filter-name "QDFilter" --filter-expression "FS >
200.0" --filter-name "FSFilter" --filter-expression "ReadPosRankSum < -20.0" --filter-name
"ReadPosRankSumFilter" -O 20HAS_S46.filtered_indels.vcf
```

Combinar SNPs e INDELs filtrados

```
sudo /home/patricia/miniconda3/bin/python
/home/patricia/short_germline_variants-main/gatk-4.5.0.0/gatk --java-options "-Xmx4g" MergeVcfs -I
20HAS_S46.filtered_snps.vcf -I 20HAS_S46.filtered_indels.vcf -O
/home/patricia/short_germline_variants-main/pipeline/final_vcfs/20HAS_S46.filtered_merged.vcf
```

4) Anotación

La anotación de variantes realiza consultas en múltiples bases de datos de secuencias y variantes para caracterizar cada variante. Para ello existen varias herramientas disponibles para la anotación funcional, incluída ANNOVAR, que utiliza información actualizada para anotar funcionalmente variantes genéticas detectadas en diversos genomas. Dada una lista de variantes con cromosoma, posición inicial, posición final, nucleótido de referencia y nucleótidos observados (Annovar, 2010).

Algunas bases de datos que se utilizó:

-Proyecto 1000 Genomas (1000 Genomes Project) de agosto de 2015. Tiene como objetivo crear un catálogo detallado de variaciones genéticas en la población humana. La base de datos contiene información sobre variantes genéticas comunes y raras encontradas en secuencias de ADN de miles de individuos de diversas poblaciones alrededor del mundo (Auton et al., 2015).

-Datos de frecuencia de alelos del exoma ExAC 65000 para ALL, AFR (africano), AMR (americano mixto), EAS (asiático oriental), FIN (finlandés), NFE (europeo no finlandés), OTH (otros), SAS (sur asiático)) . versión 0.3 (ExAC Browser, n.d.).

-Proyecto gnomAD (Genome Aggregation Database). Proporciona datos detallados sobre la frecuencia de alelos observada en diferentes poblaciones para variantes genéticas específicas, lo que puede ser útil para evaluar la rareza o la frecuencia de una variante en la población general (Collins et al., 2020).

Procedemos a la descarga de la herramienta y su instalación:

```
sudo wget http://www.openbioinformatics.org/annovar/download/0wgxR2rIVP/annovar.latest.tar.gz
```

Descomprimimos el archivo con permisos de superusuario

```
sudo tar -zxvf annovar.latest.tar.gz
```

Ejecutamos el Annovar

```
sudo perl /home/patricia/short_germline_variants-main/tools/annovar/table_annovar.pl
/home/patricia/short_germline_variants-main/pipeline/final_vcfs/20HAS_S46.filtered_merged.vcf
/home/patricia/short_germline_variants-main/tools/annovar/humandb/ -buildver hg37 -outfile
/home/patricia/short_germline_variants-main/pipeline/annovar/myanno.$NAME -remove -protocol
refGene,cytoBand,genomicSuperDups,gwasCatalog,clinvar_20200316,esp6500siv2_all,1000g2015aug_al
l,1000g2015aug_afr,1000g2015aug_eas,1000g2015aug_eur,snp138,exac03,avsnp150,dbnsfp35a,regsnpi
ntron,gwava,gnomad211_genome,intervar_20180118,kaviar_20150923,eigen -operation
g,r,r,r,f,f,f,f,f,f,f,f,f,f,f,f,f,f,f,f -nastring . -vcfinput
```

Annovar generó un archivo .txt con las anotaciones genéticas en formato annovar, cuyo archivo fue denominado filtered_Lyb20HAS_annovar.hg37_multianno.txt. El cual se utilizó en el siguiente scrip de Rstudio que esta diseñado para analizar variantes genéticas utilizando el paquete “VariantAnnotation” y otros paquete

```
setwd("/home/patricia/short_germline_variants-main/pipeline/annovar/Lyb20HAS/")
mart<-read.delim("/home/patricia/short_germline_variants-main/pipeline/grch37_p13_feb2016_mart_export.txt",header=T,sep="
")
sLyb20HAS<-read.delim2("/home/patricia/short_germline_variants-main/pipeline/annovar/filtered_Lyb20HAS_annovar.hg37_mu
ltianno.txt",skip=1,h=F)
nom<-c("Chr","Start","End","Ref","Alt","Func.refGene","Gene.refGene","GeneDetail.refGene","ExonicFunc.refGene",
"AAChange.refGene","cytoBand","genomicSuperDups","esp6500siv2_all","x1000g2015aug_all","x1000g2015aug_afr","x1000g2
015aug_eas", "x1000g2015aug_eur","snp138", "SIFT_score", "SIFT_pred",
"Polyphen2_HDIV_score","Polyphen2_HDIV_pred","Polyphen2_HVAR_score", "Polyphen2_HVAR_pred", "LRT_score",
"LRT_pred", "MutationTaster_score" , "MutationTaster_pred", "MutationAssessor_score", "MutationAssessor_pred",
"FATHMM_score", "FATHMM_pred","RadialSVM_score", "RadialSVM_pred", "LR_score", "LR_pred", "VEST3_score" ,
"CADD_raw","CADD_phred" , "GERP++_RS" , "phyloP46way_placental" , "phyloP100way_vertebrate"
,"SiPhy_29way_logOdds" , "CLNALLELID", "CLNDN", "CLNDISDB","CLNREVSTAT","CLNSIG" ,
"Kaviar_AF","Kaviar_AC","Kaviar_AN","ExAC_ALL" ,"ExAC_AFR","ExAC_AMR","ExAC_EAS",
"ExAC_FIN","ExAC_NFE","ExAC_OTH","ExAC_SAS","gnomAD_genome_ALL","gnomAD_genome_AFR","gnomAD_genome_A
MR","gnomAD_genome_ASJ","gnomAD_genome_EAS","gnomAD_genome_FIN","gnomAD_genome_NFE","gnomAD_genome
_OTH","Otherinfo", "QUAL1",
"num","chr.vcf","dp.vcf","algo.vcf","ref.vcf","alt.vcf","qual.vcf","pass.vcf","info.vcf","format.vcf","sLyb20HAS")
colnames(sLyb20HAS)<-nom sLyb20HAS.genotipos<-substr(sLyb20HAS$`sLyb20HAS`,1,3)
sLyb20HAS$`sLyb20HAS`<-sLyb20HAS.genotipos
colnames(sLyb20HAS)[80]<-"Zygosity"
omim<-read.delim("/home/patricia/short_germline_variants-main/scritp/omim/mim2gene.txt",h=F)
colnames(omim)<-c("MIM","type","Entrez","Gene","EnsemblID")
morbid<-read.delim("/home/patricia/short_germline_variants-main/scritp/omim/morbidmap.txt",comment.char="#",h=F)
colnames(morbid)<-c("Phenotype","Genes","MIM","Cyto")
omim.m<-merge(omim,morbid,by.x="MIM",by.y="MIM",all.x=T) options(scipen = 999)
sLyb20HAS.omim<-merge(sLyb20HAS,omim.m,by.x="Gene.refGene",by.y="Gene",sort=F,all.x=T) # freq max sin contar kaviar
options(scipen = 999)
a<-apply(sLyb20HAS.omim[,c(13:17,52:67)],2,as.numeric) sLyb20HAS.omim[,c(13:17,52:67)]<-a
mm<-unlist(apply(a,1,function(x){ if( sum(is.na(x))==dim(a)[2]){return(NA)}else{max(x[!is.na(x)])}}))
sLyb20HAS.omim$freq.max<-mm col.order<-c(2,1,3:11,18,80,44:48,81,85,88,13:17,49:67,19:43,12,68:78,82:84,86)
sLyb20HAS.or<-sLyb20HAS.omim[,col.order]head(sLyb20HAS.or)
```

```
############
# HET 0.5% #
############
sLyb20HAS.het<-(sLyb20HAS.or[ sLyb20HAS.or$Zygosity!="1/1"& (as.numeric(as.character(sLyb20HAS.or$freq.max))<0.005 |
is.na(as.numeric(as.character(sLyb20HAS.or$freq.max)))) & (as.numeric(as.character(sLyb20HAS.or$freq.max))<0.005 |
is.na(as.numeric(as.character(sLyb20HAS.or$freq.max)))),])
nonsyn<-c("frameshift deletion", "frameshift insertion","nonsynonymous SNV","stopgain","stoploss")
sLyb20HAS.het.NS<-sLyb20HAS.het[sLyb20HAS.het$ExonicFunc.refGene%in%nonsyn,]
sLyb20HAS.het.spl<-sLyb20HAS.het[sLyb20HAS.het$Func.refGene=="splicing",]
sLyb20HAS.het.NSspl<-rbind(sLyb20HAS.het.NS,sLyb20HAS.het.spl)
m.sLyb20HAS.het<-merge(sLyb20HAS.het.NSspl,unique(mart[,c("Associated.Gene.Name","Description")]),by.x="Gene.refGen
e",by.y="Associated.Gene.Name",all.x=T,sort=F)
m.sLyb20HAS.het.or<-(m.sLyb20HAS.het[,c(2:4,1,5:dim(m.sLyb20HAS.het)[2])])
m.sLyb20HAS.het.or.anot<-cbind(m.sLyb20HAS.het.or[,1:20],"Comentarios de trabajo"="","Comentarios a
informe"="",m.sLyb20HAS.het.or[,21:dim(m.sLyb20HAS.het.or)[2]])
write.table(m.sLyb20HAS.het.or.anot,file="/home/patricia/short_germline_variants-main/pipeline/final_tables/Lyb20HAS/hetNSs
pl_onlygenes_freq_sLyb20HAS.txt",sep="\t",quote=F, col.names=T, row.names=F)
############
# HOM 1% #
############
sLyb20HAS.hom<-(sLyb20HAS.or[ sLyb20HAS.or$Zygosity=="1/1"& (as.numeric(as.character(sLyb20HAS.or$freq.max))<0.01
| is.na(as.numeric(as.character(sLyb20HAS.or$freq.max)))) & (as.numeric(as.character(sLyb20HAS.or$freq.max))<0.01 |
is.na(as.numeric(as.character(sLyb20HAS.or$freq.max)))),])
nonsyn<-c("frameshift deletion", "frameshift insertion","nonsynonymous SNV","stopgain","stoploss")
sLyb20HAS.hom.NS<-sLyb20HAS.hom[sLyb20HAS.hom$ExonicFunc.refGene%in%nonsyn,]
sLyb20HAS.hom.spl<-sLyb20HAS.hom[sLyb20HAS.hom$Func.refGene=="splicing",]
sLyb20HAS.hom.NSspl<-rbind(sLyb20HAS.hom.NS,sLyb20HAS.hom.spl)
m.sLyb20HAS.hom<-merge(sLyb20HAS.hom.NSspl,unique(mart[,c("Associated.Gene.Name","Description")]),by.x="Gene.refG
ene",by.y="Associated.Gene.Name",all.x=T,sort=F)
m.sLyb20HAS.hom.or<-(m.sLyb20HAS.hom[,c(2:4,1,5:dim(m.sLyb20HAS.hom)[2])])
m.sLyb20HAS.hom.or.anot<-cbind(m.sLyb20HAS.hom.or[,1:20],"Comentarios de trabajo"="","Comentarios a
informe"="",m.sLyb20HAS.hom.or[,21:dim(m.sLyb20HAS.hom.or)[2]])
write.table(m.sLyb20HAS.hom.or.anot,file="home/patricia/short_germline_variants-main/pipeline/final_tables/Lyb20HAS/homN
Sspl_onlygenes_freq_sLyb20HAS.txt",sep="\t",quote=F, col.names=T, row.names=
write.table(m.sLyb20HAS.hom.or.anot,file="home/patricia/short_germline_variants-main/pipeline/final_tables/Lyb20HAS/homN
Sspl_onlygenes_freq_sLyb20HAS.txt",sep="\t",quote=F, col.names=T, row.names=F)
```

RESULTADOS

Se mapearon 389900 reads del panel secuenciado en la muestra contra el genoma de referencia de los cuales 381981 son unique o sea son reads de buena calidad y 373834 son pares correctamente alineados.

Fueron encontradas 88 variantes donde solamente 48 pasaron los filtros de calidad, es decir tienen una calidad mayor a 200.

En la anotación se logró identificar una variante patogénica para esta muestra es la inserción de dos pares de bases GA en la posicón 154159650 del cromosoma X, dicha variante se encuentra en una región codificante, específicamente en el exón 14 del gen *F8,* es una mutación del tipo frameshift insertion, osea, una inserción que no es múltiplo de tres lo que hace que se corra el marco de lectura provocando un codón de stop anticipado que causa que la proteína resultante (factor VIII de la coagulación) no sea funcional.

A continuación se observa la tabla de variantes tipo SNPs e indels encontradas con el flujo de trabajo aplicado con su parámetro de calidad cruzados con las bases de datos.

Tabla 1. Variantes heterocigotas 0,5%

Chr	Start	End	Gene.refGene	Ref	Alt	Func.refGene	ExonicFunc.refGene	AAChange.refGene	Phenotype	esp6500siv2_all	x1000g2015aug_all	x1000g2015aug_afr	x1000g2015aug_eas	x1000g2015aug_eur	gnomAD_genome_AMR	gnomAD_genome_ASJ	gnomAD_genome_EAS	gnomAD_genome_FIN	gnomAD_genome_NFE	gnomAD_genome_OTH	SIFT_score
1	169510675	169510675	F5	A	C	exonic	nonsynonymous SNV	F5:NM_000130:exon13:c.T6530G:p.I2177S	Factor V deficiency, 227400 (3)	NA	NA	NA	NA	NA	0	0	0	0.0001		0.74	
1	169510675	169510675	F5	A	C	exonic	nonsynonymous SNV	F5:NM_000130:exon13:c.T6530G:p.I2177S	{Stroke, ischemic, susceptibility to}, 601367 (3)	NA	NA	NA	NA	NA	0	0	0	0.0001		0.74	
1	169510675	169510675	F5	A	C	exonic	nonsynonymous SNV	F5:NM_000130:exon13:c.T6530G:p.I2177S	{Pregnancy loss, recurrent, x susceptibility to, 1},	NA	NA	NA	NA	NA	0	0	0	0.0001		0.74	
1	169510675	169510675	F5	A	C	exonic	nonsynonymous SNV	F5:NM_000130:exon13:c.T6530G:p.I2177S	{Thrombophilia, susceptibility to, due to factor V	NA	NA	NA	NA	NA	0	0	0	0.0001		0.74	
1	169510675	169510675	F5	A	C	exonic	nonsynonymous SNV	F5:NM_000130:exon13:c.T6530G:p.I2177S	{Budd-Chiari syndrome}, 600880 (3)	NA	NA	NA	NA	NA	0	0	0	0.0001		0.74	
X	154157393	154157393	F8	T	G	exonic	nonsynonymous SNV	F8:NM_000132:exon14:c.A6672C:p.N1558H	Hemophilia A, 306700 (3)	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	0.56
X	154157393	154157393	F8	T	G	exonic	nonsynonymous SNV	F8:NM_000132:exon14:c.A6672C:p.N1558H	Thrombophilia 13, X-linked, due to factor VIII defect, 301071 (3)	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	0.56

Tabla 2. Variantes homocigotas 1%

Chr	Inicio	End	Gen. REF	Ref	Alt	Func.	ExonicFunc. refGene	AAChange.refGene	Phenotype	gnomAD_genome_ALL	gnomAD_genome_FIN	gnomAD_genome_NFE	gnomAD_genome_OTH	SIFT_score	SIFT_pred	Polyphen2_HDIV_score
10	8.8E+07	88418424	OPN4	T	C	exonic	nonsynonymous SNV	OPN4:NM_033282:exon4:c.T608C:p.L203P,OPN4:NM_001030015:exon5:c.T641C:p.L214P	NA	.	.	.	NA	NA	NA	NA
X	1.5E+08	154159650	F8	-	GA	exonic	frameshift insertion	F8:NM_000132:exon14:c.2414_2415insTC:p.S805fs	Hemophilia A, 306700 (3)	.	.	.	NA	NA	NA	NA
X	1.5E+08	154159650	F8	-	GA	exonic	frameshift insertion	F8:NM_000132:exon14:c.2414_2415insTC:p.S805fs	Thrombophilia 13, X-linked, due to factor VIII defect, 301071 (3)	.	.	.	NA	NA	NA	NA

Discusión

En este flujo de trabajo creado para la detección de variantes geneticas causante de coagulopatias, se aplicaron las diferentes herramientas basado en las buenas práctica de bioinformática, BWA y samtools para alinear los reads de este panel al genoma de referencia, Picard y GATK para la llamada variante de nucleótido único (SNV) e inserciones y deleciones cortas. Con este análisis se ha llegado al resultados, donde se mapearon 389900 reads los cuales fueron fitrados quedando 373834 reads de buena calidad y correctamente alineados, de los cuales se logro identificar 88 variantes, de las cuales 44 pasaron por los filtros de calidad obtiendo una puntuación de calidad mayor a 200. Con la anotación se logró identificar una variante patogénica en el cromosoma X,, dicha mutación se encuentra en el gen *F8,* precisamente en el exón 14, donde se detecta una inserción de dos pares de base, una guanina (G) y una adenina (A), como muestra la tabla (3), la cual asociamos con el fenotipo del individuo. Es importante resaltar que dicha variante patogénica no ha sido reportada aún en ninguna base de datos, CytoBand, GenomicSuperDups, Esp6500siv2, 1000g2015aug, Ljb26, PolyPhen-2, Exac03, Gnomad_genome, kaviar_20150923.

En contraste con este trabajo, existen otros trabajos con el mismo propósito aplicando otras herramientas como es el caso del trabajo realizado por Villarreal-Martineza y colaboradores en el 2020, donde los datos brutos se analizaron utilizando el software Torrent suite v5.0.4. (Life technologies, California, EE. UU.). El análisis de cobertura se realizó utilizando un complemento de análisis de cobertura v5.0.2.0 y se detectaron variantes mediante Complemento Variant Caller v5.0.2.1 (Life Technologies, CA, EE. UU.). Se utilizó Reporter 5.0 para la anotación y clasificación de variantes. Los reads secuenciados se alinearon con el genoma de referencia GRCh37. En otro trabajo cuyo objetivo era analizar mutaciones en el gen *F8* en 485 familias con hemofilia A y realizar el diagnóstico prenatal en China, se utilizó secuenciación de próxima generación (NGS). Los datos resultantes de la secuenciación se analizaron utilizando el software de análisis de datos de secuenciación Ion Torrent de Torrent Suite (Life Technologies, Carlsbad, CA, EE. UU.) para generar lecturas de secuencia. Finalmente, se recogieron y analizaron todas las variaciones genéticas resultantes. (Yin et al., 2020). Estos resultados fueron similares a nuestro trabajo ya que los métodos analíticos implementados llevaron a los resultados obtenidos, identificando 478 variantes patogénicas en 485 pacientes con HA.

Por otro lado encontramos MutAid (Pandey et al., 2016) que es una herramienta integrada diseñada para la predicción y análisis de efectos funcionales de mutaciones genéticas. Lo cual potencialmente, podría ser una herramienta complementaria en este flujo de trabajo; debido a su funcionalidad pragmatica la cual nos permite la identificación eficaz de variantes. También existen softwares con la misma finalidad, por ejemplo Geneious Prime está repleto de herramientas bioinformáticas fundamentales, que incluyen ensamblaje, alineación, construcción de árboles, clonación, diseño de cebadores y análisis de variantes para datos de secuencias de Sanger y NGS . Un software bastante práctico pero con un alto costo de licencia, en contraste con este flujo de trabajo está basado en Gatk cuyas prácticas incluyen varias etapas del análisis, desde el procesamiento de los datos hasta la llamada de variantes y su filtrado.

Conclusión

En este trabajo probamos un flujo de trabajo característico para variantes de línea germinal corta (SNPs e Indels) utilizando herramientas de alneación como BWA para alinear el panel de genes de la muestra con el genoma de referencia utilizado, en este caso el GRCh37 (hg37), tambien la utilización de herramientas de Gatk que nos garantizan la calidad del filtrado.

El flujo de trabajo bioinformático permitió la detección de la variante patogénica esperada, así como de diferentes SNPs e indels en el gen del factor VIII de la coagulación, lo que posibilita que se realice de forma rápida y confiable el diagnóstico genético molecular de los pacientes con hemofilia y se detecten las portadoras lo que posibilita el asesoramiento genético.

ANEXO

Instalación BWA

Para la instalación de BWA se descargó el archivo desde el siguiente link: https://sourceforge.net/projects/bio-bwa/files/ en la terminal de linux descomprimimos el archivo:

```
tar -xvf bwakit-0.7.12_x64-linux.tar.bz2
```

Esto generará un directorio bwa-0.7.12 que contiene los archivos necesarios para copilar el binario bwa. Para ello se acedió al directorio y se inició la copilación

```
cd bwa-0.7.12
make
```

Instalación de Samtools

Se descargó la herramienta desde este link https://sourceforge.net/projects/samtools/files/ luego se descomprimió el archivo

```
tar -xvf samtools-0.1.2.tar.bz2
```

Esto produce un directorio llamado que samtools-0.1.2 contiene los archivos necesarios para compilar el binario de Samtools. Se compiló usando el siguiente comando:

```
cd samtools-0.1.2
make
```

Instalación de Picard

Fue descargado desde el link https://github.com/broadinstitute/picard/releases/ , y lo descomprimido con el siguiente comando:

```
tar xjf picard-tools-2.4.1.zip
```

Esto produce un directorio llamado picard-tools-2.4.1que contiene los archivos jar de Picard. Las herramientas Picard se distribuyen como un ejecutable Java precompilado (archivo jar), por lo que no es necesario compilarlas.

Se contruyó un building picard, clonando el repositorio con el siguiente comando:

```
git clone https://github.com/broadinstitute/picard.git
```

Luego se contruyó un jar Picard ejecutable y completamente empaquetado con todas las dependencias incluidas, con el siguiente comando:

```
./gradlew shadowJar
```

Instalación de GATK

Se instaló el GATK usando el siguiente comando

```
./gradlew bundle(crea gatk-VERSION.zip en build/)
```

Las herramientas GATK se distribuyen como un único ejecutable Java precompilado, por lo que no es necesario compilarlos.

BIBLIOGRAFIA

Auton, A., Abecasis, G. R., Altshuler, D., Durbin, R., Bentley, D., Chakravarti, A., Clark, A. G., Donnelly, P., Eichler, E. E., Flicek, P., Gabriel, S., Gibbs, R. A., Green, E. D., Hurles, M. E., Knoppers, B. M., Korbel, J. O., Lander, E. S., Lee, C., Lehrach, H., . . . Schloss, J. A. (2015). A global reference for human genetic variation. Nature, 526(7571), 68–74. https://doi.org/10.1038/nature15393

Aznar, José & Batlle, Javier & Bermejo, Nuria & Blázquez, A & Galmes, B & García-Frade, LJ & Iruín, G & López-Cabarcos, C & Lucía, JF & Moreno, M & Sedano, C & Simón, Miguel & Soriano, Vincent & Turnés, J & Liras, Antonio. (2009). Recomendaciones sobre Portadoras en Hemofilia.

Bagnall R, Waseem N, Green P, Giannelli F. Recurrent inversion breaking intron 1 of the factor VIII gene is a frequent cause of severe hemophilia A. Blood 2002; 99(1):168-74.

Base de datos de variantes del factor VIII de EAHAD . (Dakota del Norte). https://f8-db.eahad.org/

Bentley, DR, Balasubramanian, S., Swerdlow, HP, Smith, GP, Milton, J., Brown, CG, Hall, KP, Evers, DJ, Barnes, CL, Bignell, HR, Boutell, JM, Bryant, J. , Carter, RJ, Keira Cheetham, R., Cox, AJ, Ellis, DJ, Flatbush, MR, Gormley, NA, Humphray, SJ, ... Smith, AJ (2008). "Secuenciación precisa del genoma humano completo mediante química terminadora reversible" . Naturaleza , 456 (7218), 53-59. https://doi.org/10.1038/nature07517

Broadinstitute.org. Retrieved January 15, 2024, from https://gatk.broadinstitute.org/hc/en-us

Canal-Alonso, Á., Egido, N., Jiménez, P., Prieto Tejedor, J., & Corchado Rodríguez, J. M. (2022). Análisis de datos NGS: una revisión de las principales herramientas y marcos de trabajo para el descubrimiento de variantes.

Chun S, Fay JC. Identification of deleterious mutations within three human genomes. Genome Res. 2009 Nov;19(11):1553-61. doi: 10.1101/gr.092619.109. Epub 2009 Aug 24. PMID: 19706752; PMCID: PMC2775563.

Clarke, J., Wu, H.C., Jayasinghe, L, Patel, A, Reid, S., Bayley, H. 2009. Continuous base identification for single-molecule nanopore DNA sequencing. Nat. Nanotechno! 4: 265-270.

Collins RL, Brand H., Karczewski KJ. Una referencia de variación estructural para la genética médica y de poblaciones. Naturaleza 581 , 444–451 (2020). https://doi.org/10.1038/s41586-020-2287-8

ExAC browser. (n.d.). http://exac.broadinstitute.org/

Franchini, M., & Mannucci, P. M. (2017). Hemophilia A in the third millennium. Blood Reviews, 31(3), 176-183. doi: 10.1016/j.blre.2016.11.003

Fu W, O'Connor TD, Jun G, Kang HM, Abecasis G, Leal SM, Gabriel S, Rieder MJ, Altshuler D, Shendure J, Nickerson DA, Bamshad MJ; NHLBI Exome Sequencing Project. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature. 2013 Oct 24;493(7431):216-20. doi: 10.1038/nature11690. Epub 2012 Nov 28. PMID: 23187742; PMCID: PMC3564980.

Glusman G, Caballero J, Mauldin DE, Hood L, Roach JC. Kaviar: an accessible system for testing SNV novelty. Bioinformatics. 2011 Nov 1;27(21):3216-7. doi: 10.1093/bioinformatics/btr540. Epub 2011 Sep 12. PMID: 21908401; PMCID: PMC3198576.

Heng Li, Richard Durbin, Alineación de lectura breve rápida y precisa con la transformada de Burrows-Wheeler, Bioinformática, volumen 25, número 14, julio de 2009, páginas 1754–1760, https://doi.org/10.1093/bioinformatics/btp324

Hoffman, M., & Monroe, D. M. (2001). A cell-based model of hemostasis. Thrombosis and Haemostasis, 85(6), 958–965

I.Adzhubei, DM Jordan y SR Sunyaev, Predicción del efecto funcional de mutaciones sin sentido humanas utilizando PolyPhen-2, no. SUPL.76. 2013. doi: 10.1002/0471142905.hg0720s76.

Ju, J., Kim D.H., Bi, L, Meng, Q., Bai, X., Li, Z., Li, X., Marma, M.S., Shi, S., Wu, J., Edwards, J.R., Romu, A.,Turro, N.J. 2006. Four-color DNA seguencing by synthesis using cleavable fluorescent nucleotide reversible terminators. Proc. Nati. Acad. Sci. U.S.A. 103:19635-1964

"Kit de herramientas de Picard". 2019. Broad Institute, Repositorio GitHub. https://broadinstitute.github.io/picard/

Lakich D, Kazazian HJr, Antonarakis S, Gitschier J. Inversions disrupting the factor VIII gene are a common cause of severe haemophilia A. Nat Genet 1993; 5(3):236-41.

Li H. y Durbin R. (2010) Alineación de lectura larga rápida y precisa con la transformada de Burrows-Wheeler. Bioinformática, Epub. [PMID: 20080505]

Macfarlane, R. G. (1964). AN ENZYME CASCADE IN THE BLOOD CLOTTING MECHANISM, AND ITS FUNCTION AS A BIOCHEMICAL AMPLIFIER. Nature, 202, 498–499.

McVey J, Rallapalli P, Kemball-Cook G et al. The European Association for Haemophilia and Allied Disorders (EAHAD) Coagulation Factor Variant Databases: Important resources for haemostasis clinicians and researchers. Haemophilia 2020; 26(2): 306-313.

Moraleda Jiménez, J. M. (2017). Pregrado de Hematología 4ta Edición. Luzan5.

Pandey RV, Pabinger S, Kriegner A, Weinhäusel A (2016) MutAid: Canal integrado basado en Sanger y NGS para la identificación, validación y anotación de mutaciones en genética molecular humana. MÁS UNO 11(2): e0147697. https://doi.org/10.1371/journal.pone.0147697

Pascual L. A., (2012) Automatización de los procesos de alineamiento y búsqueda de variantes en secuencias de ADN.

Peyvandi, F., Garagiola, I., & Young, G. (2016). The past and future of haemophilia: diagnosis, treatments, and its complications. The Lancet, 388(10040), 187-197. doi: 10.1016/S0140-6736(15)01123-X

Ritchie, G., Dunham, I., Zeggini, E. et al. Anotación funcional de variantes de secuencia no codificantes. Métodos Nat 11 , 294–296 (2014). https://doi.org/10.1038/nmeth.2832

Roberts, H. R., & Monroe, D. M. (2004). Hemostasis and thrombosis. In: Hoffman R, Benz Jr EJ, Shattil SJ, et al., editors. Hematology: Basic Principles and Practice. Churchill Livingstone; p. 1747-1764. doi: 10.1016/B978-0-443-06631-9.50083-4

Rodríguez-Santiago, B., & Armengol, L. (2012). Tecnologías de secuenciación de nueva generación en diagnóstico genético pre-y postnatal. Diagnóstico prenatal , 23 (2), 56-66.

Rodríguez-Martorell FJ, Mingot ME, Palomo A, Núñez R, Pérez-Garrido R, Villar A, Tizzano EF, Alonso C, Altisent C, Aznar JA, Batlle J, Bermejo N, Blázquez A, Galmes B, García-Frade LJ, Iruín G, López-Cabarcos C, Lucía JF, Moreno M, Sedano C, Simón MA, Soriano V, Turnés J, Liras A. (2009) Recomendaciones sobre portadoras en hemofilia. Ediciones de la Real Fundación Victoria Eugenia y Federación Española de Hemofilia MADRID. ISBN: 978-84-692-9583-0 Nº de Registro: 10/10413

Rojas-Villalta, D., Benavides-Villegas, D., Angulo-Hidalgo, B., Muñoz-Solorzano, L., & Consumi-Tubito, C. (2024). Caracterización de las tecnologías de secuenciación genética de segunda y tercera generación. Revista Tecnología En Marcha, 37(2), Pág. 70–81. https://doi.org/10.18845/tm.v37i2.6494

Roy, S., Coldren, C., Karunamurthy, A., Kip, N. S., Klee, E. W., Lincoln, S. E., ... & Carter, A. B. (2018). Standards and guidelines for validating next-generation sequencing bioinformatics pipelines: a joint recommendation of the Association for Molecular Pathology and the College of American Pathologists. The Journal of Molecular Diagnostics, 20(1), 4-27.

Ruggeri, Z. M. (2002). Platelets in atherothrombosis. Nature Medicine, 8(11), 1227-1234. doi: 10.1038/nm1102-1227

Sequencing Technology | Sequencing by synthesis. (n.d.). https://www.illumina.com/science/technology/next-generation-sequencing/sequencing-technology.html

Shaffer LG, McGowan-Jordan J, Schmid M (2016) Karger, Basel. An International System for Human Cytogenomic Nomenclature.

Sharp AJ, Locke DP, McGrath SD, Cheng Z, Bailey JA, Vallente RU, Pertz LM, Clark RA, Schwartz S, Segraves R, Oseroff VV, Albertson DG, Pinkel D, Eichler EE. Segmental duplications and copy-number variation in the human genome. Am J Hum Genet. 2005 Nov;77(6):78-88. doi: 10.1086/498651. PMID: 16252243; PMCID: PMC1271386.

Somak Roy, Christopher Coldren, Arivarasan Karunamurthy, Nefize S. Kip, Eric W. Klee, Stephen E. Lincoln, Annette Leon, Mrudula Pullambhatla, Robyn L. Temple-Smolkin, Karl V. Voelkerding,Chen Wang, Alexis B. Carter (2018) Standards and Guidelines for Validating Next-Generation Sequencing Bioinformatics PipelinesA Joint Recommendation of the Association for Molecular. Pathology and the College of American Pathologists

Spangenberg L, Urquiola L, Simoes C, (2022). Short_germline_variants. Repositorio Gitlab https://gitlab.com/genomics_ubi/short_germline_variants

ST Sherry, M. Ward y K.Sirotkin, 'dbSNP: base de datos para polimorfismos de nucleótidos únicos y otras clases de variación genética menor', Genome Res, vol. 9, núm. 8, págs. 677–679, agosto de 1999, doi: 10.1101/GR.9.8.677.

Van der Auwera G, Carneiro M, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, Banks E, Garimella KV, Altshuler D, Gabriel S, DePristo MA. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinformatics. 2013 Sep 4;43:11.10.1-33. doi: 10.1002/0471250953.bi1110s43. PMID: 24474296; PMCID: PMC4243306.

Vega Y, Faguaga M, Aberlleyro M, De Brasi C, Bandinelli E, Sans M, Hidalgo P, (2020) Inversión de los intrones 1 y 22 del F8 en pacientes con hemofilia A severa y portadoras del noreste de Uruguay. Archivos de Pediatria Del Uruguay, 91(1). https://doi.org/10.31134/ap.91.2.3

Villarreal-Martíneza L, Ibarra-Ramirezb M, Calvo-Anguianob G, De Jesús Lugo-Trampeb J, Luna-Záizarc H, Martínez-de-Villarrealb L, Meléndez-Arandad L, Jaloma-Cruzd A, (2020) Molecular genetic diagnosis by next-generation sequencing in a cohort of Mexican patients with haemophilia and report of novel variants

Viso Sarmiento M, López Fernández M, Noya Pereira M.S, Batlle Fonrodona J, Coagulopatías. Criterios diagnósticos y tratamiento,Medicine - Programa de Formación Médica Continuada Acreditado,Volume 8, Issue 53, 2001, Pages 2809-2816, ISSN 0304-5412,https://doi.org/10.1016/S0304-5412(01)70531-8.(https://www.sciencedirect.com/science/article/pii/S0304541201705318)

Wang, J., Gu, J., Chen, H., Wu, Q., Xiong, L., Qiao, B., Zhang, Y., Xiao, H., & Tong, Y. (2022). A Novel Deletion Mutation of the F8 Gene for Hemophilia A. Diagnostics (Basel, Switzerland), 12(11), 2876. https://doi.org/10.3390/diagnostics12112876

Wang K, Li M, Hakonarson H. ANNOVAR: Anotación funcional de variantes genéticas a partir de datos de secuenciación de próxima generación Nucleic Acids Research , 38:e164, 2010

World Federation of Hemophilia. Annual Global Survey 2021. October 2022. Link: http://www1.wfh.org/publications/files/pdf-1731.pdf.

Yin, F., Li, Q., Shi, P., Ning, L., Kong, X., & Guo, R. (2020). Mutation analysis in the F8 gene in 485 families with haemophilia A and prenatal diagnosis in China. Haemophilia, 27(1). https://doi.org/10.1111/hae.14206