PROYECTO FINAL DE BIOINGENIERÍA

Validación y desarrollo de un flujo de análisis NGS de variantes para contribuir al manejo terapéutico de pacientes oncológicos

Presentado por

MARÍA TERESITA LAGUINGE
CARLA ANDREA MOYA

Tutores

BIOING. MSC DAIANA GANIEWICH

Departamento de Bioingeniería
INSTITUTO TECNOLÓGICO DE BUENOS AIRES

MARZO 2023

RESUMEN / ABSTRACT

Introducción:

El advenimiento de la secuenciación masiva (NGS) ha hecho posible interrogar el genoma de células de origen germinal y somático que amplían la información de datos que determinan el diagnóstico, el pronóstico, el tratamiento del cáncer; y permiten establecer el asesoramiento genético de familiares en riesgo. Para analizar los datos de NGS, es necesario utilizar flujos de análisis bioinformáticos, los cuales necesariamente deben ser validados para su utilización segura tanto en investigación como en el ámbito clínico.

Materiales y métodos:

Se utilizó un ámbito de testeo en *AWS* para las herramientas que fueron luego empaquetadas en *CWL* e implementadas en *CGC* para su disponibilización. Para el desarrollo de los flujos de análisis se utilizaron herramientas de *GATK* y *GRAF* para el alineamiento, llamado y filtrado de variantes; e *InterVar* y *CancerVar* para la posterior anotación de estas. Para realizar la validación del flujo de análisis germinal se utilizaron muestras *gold standard WES* y *WGS* de *GIAB* y se constataron los valores de las métricas de sensibilidad, precisión y valor F1, utilizando la herramienta *Hap.py*. Se realizó la clasificación de patogenicidad de las variantes y su relevancia clínica utilizando InterVar y CancerVar. Finalmente, se utilizó *R* y *Python* para el desarrollo de *scripts* customizados que permitieran la visualización sencilla de las variantes y creación del informe médico.

Resultados:

Se desarrollaron y validaron flujos de análisis germinales, obteniendo el mayor valor F1 tanto para muestras *WES* como *WGS* en aquel que implementaba las herramientas de *GRAF*. Además, se realizo un análisis de costo y tiempo, obteniendo el menor tiempo de ejecución para aquel que utilizaba *GRAF* tanto para el alineamiento como para el llamado y filtrado de variantes. Para las variantes somáticas, se logró también desarrollar un flujo de análisis en CWL conformando un solo flujo de análisis que obtuviera variantes SNP e INDEL de origen germinal y somático que posteriormente se disponibilizó en *CGC* habilitando un contexto mas amigable para usuarios no expertos. Por último, se logró añadir a los flujos de análisis la anotación de las variantes y clasificación de patogenicidad siguiendo las guías de ACMG y AMP con herramientas de código abierto y la elaboración de un informe que presente los datos de manera mas amigable.

Conclusión:

Se logró satisfactoriamente desarrollar flujo de análisis bioinformáticos con herramientas de código abierto para estudios NGS somático y germinal, el cuál también fue validado. Estos flujos integran el procesamiento bioinformático técnico, su anotación con datos biológicos/fisiopatológicos/clínicos y su clasificación por ACMG y AMP. Además, se logró generar un informe y disponibilizar el flujo de análisis en CGC facilitando el análisis de variantes NGS para usuarios sin experiencia en bioinformática. Se dejan las bases sentadas para que en futuros proyectos se pueda validar con mayor cantidad de muestras el flujo de análisis germinal y con un conjunto de validación generado

en el laboratorio, la detección de variantes somáticas.

AGRADECIMIENTOS

Queremos agradecer a todas aquellas personas que se involucraron de una manera u otra en la realización de este proyecto. En particular, a nuestra tutora Daiana Ganiewich por todos los conocimientos y herramientas que nos proporcionó a lo largo de estos dos años. También queremos agradecer a nuestras familias y amigos por su compañía y apoyo incondicional. Por último y no menos importante, nos agradecemos ambas autoras mutuamente por la determinación, esfuerzo, compañerismo y cariño que demostramos durante esta etapa.

TABLA DE CONTENIDOS

ÍNDICE DE FIGURAS

FIGURA **Página**

ÍNDICE DE TABLAS

TABLA **Página**

GLOSARIO

ACMG *American College of Medical Genetics and Genomics* - Colegio Americano de Genética Médica y Genómica. 12

ADN Ácido desoxirribonucleico. Moléculas que contienen la información genética y la transmiten de una generación a otra. 2

AMP *Association for Molecular Pathology* - Asociación de Patología Molecular. 12

ARN Ácido ribonucleico. Uno de los dos tipos de ácido nucleico que elaboran las células que cumple funciones variadas según el tipo. 2

ARNm ARN mensajero. molécula de ácido nucleico cuya traducción posibilita la síntesis de proteínas. 3

AWS *Amazon Web Services*. Es un proveedor de servicios en la nube. 20

CGC *Cancer Genomics Cloud*. Es una plataforma en la nube que permite el análisis, el almacenamiento y el cálculo de grandes conjuntos de datos sobre el cáncer. 22

Cobertura horizontal regiones de captura. Son las secciones del genoma que están representadas en el conjunto de datos al menos con cierta profundidad.. 12

CWL *Common Workflow Language*. Es un lenguaje para definir e instrumentar flujos de trabajo informáticos. 21

Cáncer esporádico Aquel causado por las mutaciones somáticas. 8

Cáncer hereditario Aquel causado por las mutaciones germinales. 8

EC2 Servidores virtuales en la nube brindados por AWS. 20

Flujo de análisis Es un conjunto de herramientas ejecutadas en un orden definido para recopilar, organizar y analizar datos de secuenciación genética y datos biológicos relacionados. 16

Frecuencia alélica incidencia de una variante genética en una población. 12

GATK *Genome Analysis Toolkit*. Es un referente en la industria que brinda recomendaciones y herramientas bioinformáticas. 19

Germinales Aquellas variantes o mutaciones que están presentes en todas las células del individuo debido a que se generan en la concepción y se pueden transmitir a la descendencia. 3

GIAB *Genome in a bottle*. Consorcio Genoma en una Botella dirigido por NIST que publicó un conjunto de datos de referencia de alta calidad *gold standard* de SNP e INDELs utilizando diversas tecnologías de secuenciación.. 19

GRAF Flujos de trabajo bioinformáticos y herramientas proporcionados por *Seven Bridges* para el análisis de datos de secuenciación de nueva generación. 18

Hap.py Herramienta de validación diseñada para comparar llamados de variantes de conjuntos gold standard y un archivo VCF de prueba. 20

Herramientas bioinformáticas bloque principal de un flujo bioinformático. Ejecuta un algoritmo particular para transformar uno o más datos de entrada y devolver el resultado procesado. 16

INDEL Inserción o deleción de uno o más nucleótidos de la cadena de ADN. 4

Linux Sistema operativo de acceso libre y gratuito. 21

Medicina de precisión Estrategia emergente para el tratamiento y la prevención médica, que tiene en cuenta la variabilidad individual en genes, ambiente y forma de vida. 9

NGS *Next Generation Sequencing* - Secuenciación masiva en paralelo. 9

Profundidad de lectura medida de la integridad general del conjunto de datos NGS que describe el número de veces que se ha leído un nucleótido dado. 12

S3 Servicio de almacenamiento de datos brindado por AWS. 20

SB *Seven Bridges*. Es una empresa especializada en software que brinda herramientas y datos bioinformáticos. 22

SNP *Single Nucleotide Polimorphysm* - Polimorfismos de nucleótido único en la cadena de ADN. 4

Somáticas Aquellas variantes o mutaciones que ocurren de manera esporádica después de la concepción, no son hereditarias y están presentes solo en algunas células del organismo. 3

tasa TS/TV Tasa de transiciones entre transversiones. Es una de las métricas utilizadas para el control de calidad de muestras. Algunos estudios indican que los SNP en las regiones del exoma deberían tener una relación Ts/Tv de alrededor de 3, enfatizando que el incremento en la relación Ts/Tv generalmente indica una mejor calidad. 57

Transiciones SNP en que se cambia un nucleótido de purina por otra purina o un nucleótido de pirimidina por otro de pirimidina. 4

Transversiones SNP en que se sustituye una purina por una pirimidina o viceversa. 4

VUS Variantes de significado incierto. 13

WES *Whole Exome Sequencing* - Secuenciación de exoma completo. 10

WGS *Whole Genome Sequencing* - Secuenciación de genoma completo. 10

Parte I

Introducción

CAPÍTULO 1

BREVE INTRODUCCIÓN BIOLÓGICA

Todos los seres vivos comparten una cualidad particular: almacenan información genética utilizando las mismas macromoléculas, ADN y ARN. Ambas están conformadas por un conjunto de subunidades llamados nucleótidos (que se juntan de a pares, formando lo que se conoce como “pares de bases”), y forman parte de una de las cuatro clases de macromoléculas principales consideradas cruciales para la vida, junto con las proteínas, los lípidos y los carbohidratos. [1]

En 1965, se mostró que el ADN se transcribe en ARN y luego se traduce en proteína, y que la tasa de transcripción está controlada por un circuito de retroalimentación en el que la proteína regula la actividad del complejo transcripcional. Esto es lo que se conoce como el Dogma Central de la biología molecular (ver figura 1.1). Este ha sido modificado a lo largo de los años ya que se entiende que el flujo de información biológica es mucho más complejo de lo que se pensaba hasta ese momento. [2]

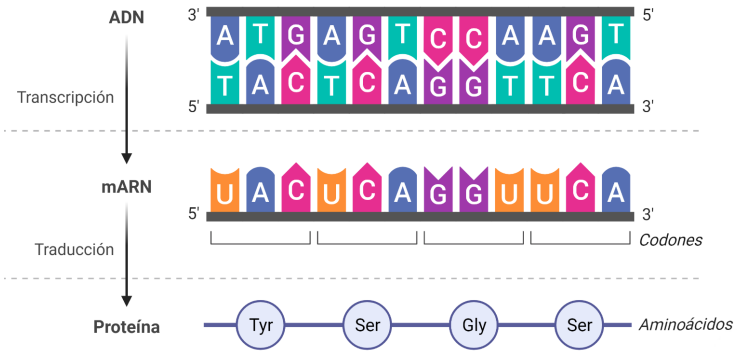


Figura 1.1: *Dogma central de la biología: El ADN se transcribe a ARN mensajero (ARNm) y este se traduce a proteína (aquí no se están considerando los diferentes tránscriptos que puede tener el mismo gen)*

El genoma humano está compuesto por 3.272.116.950 pares de bases, es decir, 6.544.233.900 nucleótidos. Estos nucleótidos son la adenina (A), guanina (G), citosina (C) y timina (T), que se disponen de cierta manera para generar instrucciones específicas, y son capaces de codificar así, proteínas determinadas. Esta información genética se encuentra altamente conservada entre especies, y particularmente entre humanos la similitud es del 99.9%. [3] Esto muestra que la manera en la que estas bases se disponen no es aleatoria y está íntimamente relacionada con el correcto funcionamiento del organismo. Pero, ¿qué sucede cuando estas no están dispuestas como deberían? ¿esto afecta de alguna manera a la creación de la proteína?

Estos cambios en los nucleótidos son conocidos como variantes. Las variantes a nivel genético, siguiendo el dogma central, pueden modificar el ARNm que a su vez, contiene la información para la creación de una proteína que también podría ser modificada por esta variante, incluso afectando el funcionamiento de esta. Este trabajo se centrará fundamentalmente en analizar las alteraciones que ocurren en las moléculas de ADN, con el fin de estudiar el impacto que tienen en el fenotipo, es decir, se estudiará el ADN desde la obtención de la secuencia hasta la detección de variantes, su clasificación y, de ser conocida, su patogenicidad asociada.

1.1. Las variantes y las mutaciones

Las variantes son alteraciones genéticas de uno o más nucleótidos, y pueden ocurrir por errores en el ciclo celular, cambios debidos a que la maquinaria celular no es perfecta en cada proceso de división celular, o por una exposición a agentes que dañen el ADN, entre otros. En caso de que las alteraciones en la secuencia de ADN de la célula sean patogénicas, se denominan mutaciones. [4, 5]

Las variantes y las mutaciones se pueden clasificar [5, 6] según el tipo de célula en la que se originan en:

- Germinales: están presentes en todas las células de un individuo, ya que se generan en la concepción y se pueden transmitir a la descendencia. Se denominan germinales debido a que están presentes en las células germinales[1] que dan origen al individuo, como se muestra en la imagen de la izquierda de la figura 1.2. [5, 6]
- Somáticas: son aquellas que ocurren de manera esporádica luego de la concepción, no son hereditarias y están presentes solo en algunas células del organismo. Éstas no se pasan de generación en generación y pueden ser causadas por diversos motivos como factores ambientales o errores durante la división celular. [5, 6]

[1] Aquellas que se encargan de la formación de los gametos, los óvulos y los espermatozoides.

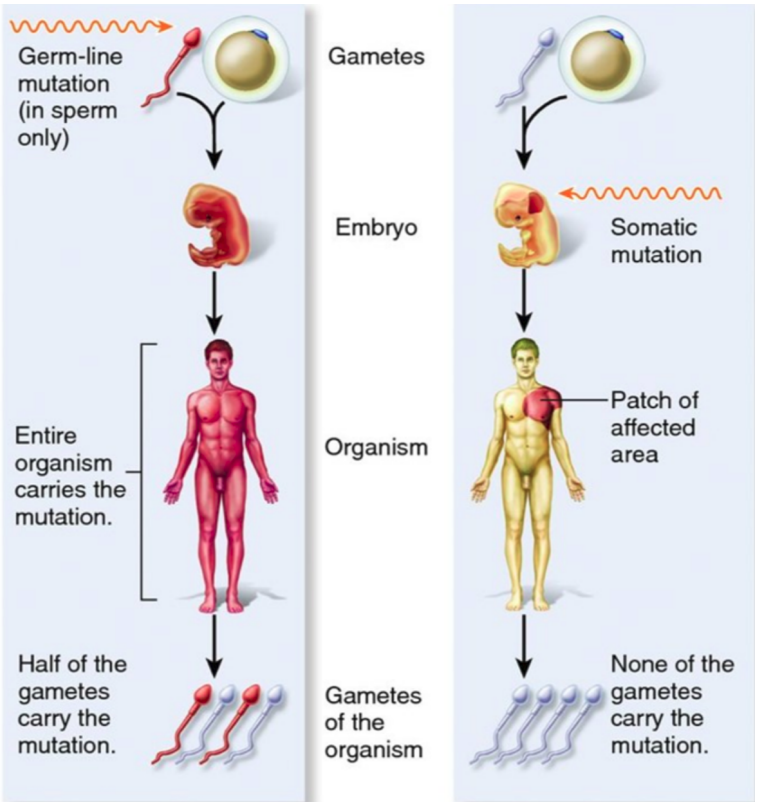


Figura 1.2: *Gráfico ilustrativo para variantes y mutaciones somáticas (a la derecha) y germinales (a la izquierda). Imagen adaptada de [7]*

A su vez, las variantes y mutaciones pueden clasificarse según el tipo de alteración en:

- Polimorfismos de nucleótido único (SNP, del inglés *Single Nucleotide Polimorphysm*): son aquellas que comprenden el cambio de solamente un nucleótido por otro. Estas pueden subdividirse en:

  1. Transiciones: Se cambia un nucleótido de purina por otra purina[2] (A o G) o un nucleótido de pirimidina[3] (C o T) por otro de pirimidina. Un ejemplo sería un cambio de A por G o un cambio de C por T.
  2. Transversiones: Se sustituye una purina por una pirimidina o viceversa. Un ejemplo sería un cambio de A por C o un cambio de A por T.

- Inserciones y deleciones (INDEL): implican la inserción o deleción de uno o más nucleótidos de la cadena. A menos que estas se den en la parte codificante como un múltiplo de 3,

[2]Uno de los dos compuestos químicos del ADN cuya estructura consta de dos anillos fusionados, uno de seis átomos y el otro de cinco. LA A y G, son derivados de una purina.

[3]Uno de los dos compuestos químicos del ADN cuya estructura consta de un anillo (estructura similar al benceno pero con dos átomos de nitrógeno que sustituyen al carbono en las posiciones 1 y 3). LA C y T, son derivados de una pirimidina.

producen cambios en el marco de lectura[4] cambiando completamente la cadena aminoacídica final.

- Variantes estructurales: es un reordenamiento de porciones más grandes del genoma (más de 50 pares de bases). Pueden ser deleciones, ganancias, duplicaciones, inserciones, inversiones o translocaciones (ver figura 1.3). [6, 8, 9]

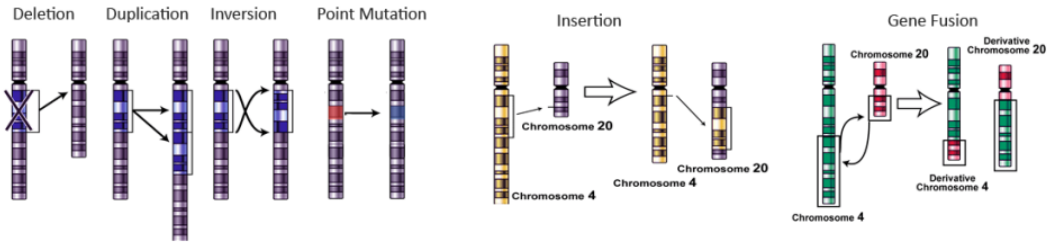


Figura 1.3: *Gráfico ilustrativo para variantes estructurales. Imagen adaptada de [10].*

Además, si las variantes SNP ocurren en una región exónica, es decir la región del ADN que codifica a proteína, pueden ocurrir distintos cambios, según las consecuencias que tengan a nivel aminoacídico se clasifican en:

- Sinónimas o silenciosas: La variante no produce un cambio en el aminoácido generado, esto se debe a la redundancia del código genético que hace que varios codones diferentes puedan codificar para el mismo aminoácido.
- No sinónimas o de pérdida de sentido: La sustitución de nucleótidos conduce a una sustitución de aminoácidos. Esto puede o no dar como resultado una variante patógena dependiendo del efecto de la sustitución de aminoácidos en la función y estructura de la proteína.
- Cambio en el marco de lectura: Una INDEL altera el largo de la secuencia, generando así una diferencia en cómo se leen los codones. Generalmente, da como resultado un codón de terminación prematuro y, en consecuencia, produce un truncamiento de la proteína. [6]

Otra clasificación importante que tienen las variantes es según su presencia en los alelos[5] es:

- Homocigotas: se refiere a tener la variante presente en ambos alelos heredados.
- Heterocigotas: se refiere a tener la variante presente en solo uno de los dos alelos heredados. [11–13]

[4] Un marco de lectura es una de las posibles formas en que se puede dividir una secuencia de nucleótidos de ADN o ARN para formar un grupo de tripletes consecutivos no solapados

[5] Los alelos son formas variantes de un gen que se encuentran en la misma posición, o locus genético, en un cromosoma.

Cabe aclarar que el concepto de variantes heterocigotas u homocigotas es únicamente válido en el contexto germinal En la figura 1.4 se muestra la diferencia entre estas.

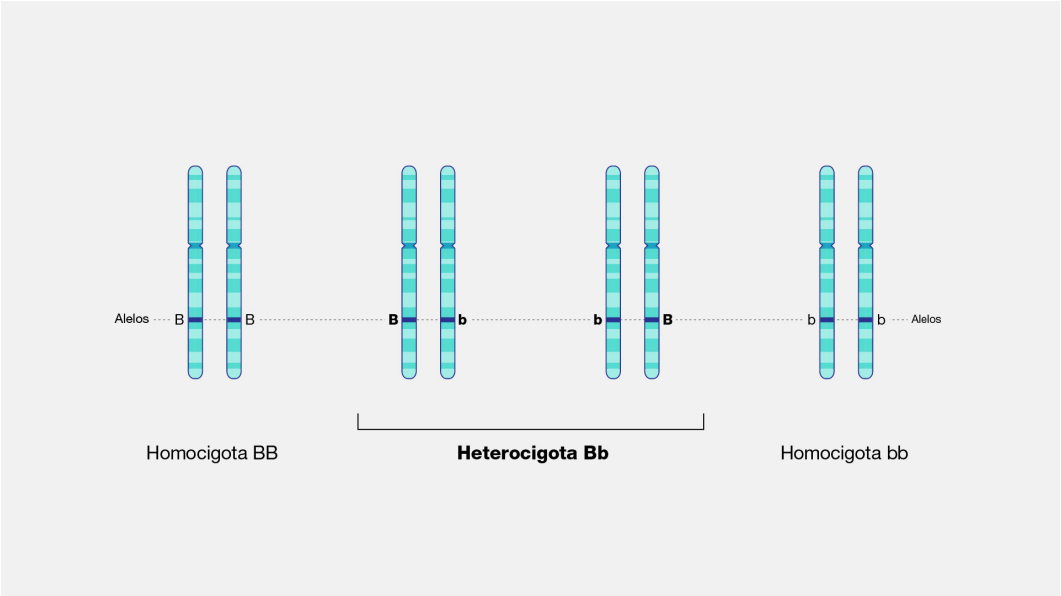


Figura 1.4: *Ilustración de heterocigosis y homocigosis para un alelo dado (Bb) [11]*

CAPÍTULO 2

CÁNCER

El cáncer es un término que describe el conjunto de enfermedades en las que existe un crecimiento descontrolado de células anormales que pueden diseminarse a otras partes del organismo. [14, 15]

Normalmente, las células humanas crecen y se multiplican (a través de un proceso llamado división celular) para formar nuevas células a medida que el organismo las necesita, y cuando las células envejecen o se dañan, mueren y nuevas células toman su lugar.

El proceso de división celular no es siempre perfecto, y la maquinaria celular no es capaz de eliminar a células anormales o dañadas que terminan creciendo y multiplicándose cuando no deberían. Estas células pueden formar tumores (masas de tejido) que pueden ser cancerosos (malignos) o no cancerosos (benignos). A través de la metástasis, los tumores cancerosos invaden tejidos cercanos y pueden diseminarse a lugares distantes del cuerpo para formar nuevos tumores.

Los tumores benignos no se diseminan ni invaden los tejidos cercanos. Cuando se extirpan, a diferencia de los cancerosos, estos generalmente no vuelven a crecer. [16, 17]

El cáncer es una enfermedad que tiene varios niveles de dimensionalidad, los cuales le proporcionan una complejidad importante. Las características más notables que se dan en esta patología se describen a continuación y se pueden observar en la figura 2.1.

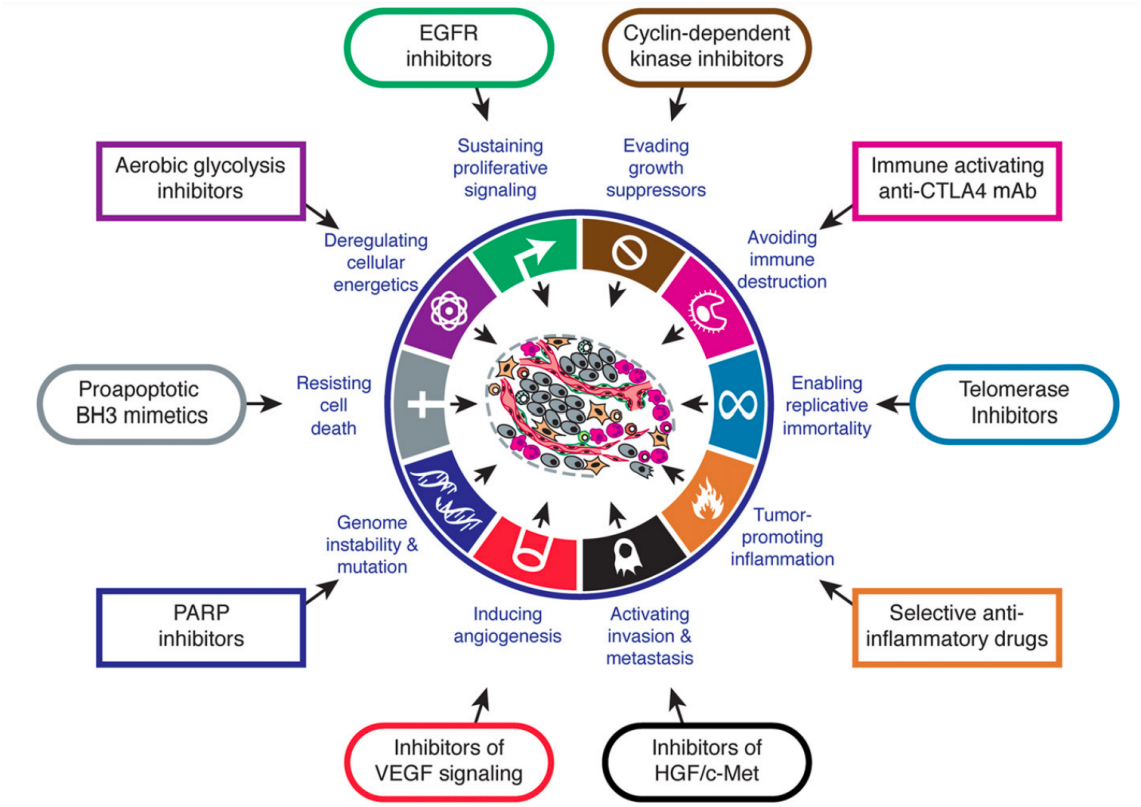


Figura 2.1: *Carácterísticas fundamentles del cáncer y sus posibles tratamientos. Imagen adaptada de [16]*

2.1. Cancer hereditario y esporádico

El cáncer puede ser hereditario o esporádico según el origen de las mutaciones causantes presentes en el organismo:

- Cáncer hereditario: causado por las mutaciones germinales.
- Cáncer esporádico: causado por las mutaciones somáticas.

CAPÍTULO 3

MEDICINA DE PRECISIÓN

La Medicina de precisión, también conocida como medicina estratificada se define por el Instituto Nacional de salud (NIH) como “una estrategia emergente para el tratamiento y la prevención, que tiene en cuenta la variabilidad individual en genes, ambiente y forma de vida"[18]. Esta forma de pensar a la medicina se opone al paradigma tradicional de que un tratamiento sirve para todos por igual y desafía el concepto de "paciente promedio"; considerando la variabilidad entre individuos. Además, tiene como objetivo guiar las decisiones de atención médica hacia el tratamiento más efectivo para un paciente determinado y, por lo tanto, mejorar la calidad de la atención; al tiempo que reduce la necesidad de pruebas de diagnóstico y terapias innecesarias. [19]

La medicina de precisión permitió contribuir a la atención médica en múltiples aspectos incluyendo la predicción del riesgo de transmitir trastornos genéticos a la descendencia, la detección temprana y precisa de enfermedades, la orientación de terapias para enfermedades crónicas y el descubrimiento y la recomendación de marcadores genómicos de eficacia, identificación de eventos adversos y dosificación de tratamientos personalizados para uso clínico que maximicen la seguridad y la eficiencia terapéutica. [19, 20]

3.1. Secuenciación de Nueva Generación (Next Generation Sequencing - NGS)

La secuenciación es un proceso que permite conocer el código genético a partir de una muestra biológica que contenga ADN o ARN. La secuenciación de nueva generación es una tecnología que permite la secuenciación simultánea de millones de secuencias nucleotídicas. Las ventajas de NGS en comparación con los métodos de secuenciación tradicionales incluyen un mayor rendimiento

al permitir la multiplexación de muestras (procesar más de una muestra a la vez), una mayor sensibilidad en la detección de variantes de baja frecuencia, un tiempo de respuesta más rápido para grandes volúmenes de muestras y un menor costo comparado a otras técnicas (como Sanger). [21]

La secuenciación de nueva generación forma parte de la secuenciación de segunda generación (siendo Sanger la de primera generación), y funciona detectando la incorporación de nucleótidos marcados con fluorescencia a unas sondas incorporadas en una placa. Así, evita la necesidad de separar el ADN en un gel (que debe previamente amplificarse por PCR). Según las sondas presentes en la placa, la secuenciación puede ser del genoma completo (del inglés, *Whole Genome Sequencing* o WGS), del exoma (del inglés, *Whole Exome Sequencing* o WES. Aquí se analiza solo la parte codificante del genoma) o de un panel de genes (un conjunto de genes específicos). Cada uno de estos tres análisis producirá datos con tamaños de ≈100 GB para WGS, ≈10 GB para WES, y dependiendo de la cantidad de genes analizados, los paneles pueden variar desde cientos de MB a ≈10 GB en el caso de los paneles. Se debe considerar también que cuanto más profundo sea el análisis, o sea, cuantas más copias se lean y secuencien, los datos resultantes resultarán más complejos y pesados. Esto es un gran beneficio y perjuicio de los datos de secuenciación NGS, dado que se brindan una gran cantidad de datos e información, pero estos generan enormes costo y trabajo bioinformáticos.

La empresa Illumina, fue una de las primeras fabricantes de dispositivos de nueva generación, y hoy domina el mercado actual con varios equipos como MiSeq, GAIIx, HiSeq [22], NovaSeq (entre otras). [21] Durante la última década, las tecnologías NGS han seguido evolucionando, aumentando la capacidad de secuenciación, incorporando innovaciones para abordar las complejidades de los genomas, así como también abaratando los costos del proceso. De hecho, a momento de escribir este trabajo, el costo de secuenciar un genoma completo en un NovaSeq 6000 de Illumina es de $200 [21] (unas 5 veces menor que hace 8 años [23]), siendo más costoso el almacenamiento posterior de los datos de secuenciación obtenidos. Esto permite cada vez más el uso de la secuenciación como una herramienta clínica. [21]

Sin embargo, a medida que surgen nuevas tecnologías, los problemas existentes se exacerban o surgen nuevos problemas. Las plataformas NGS proporcionan grandes cantidades de datos complejos, las tasas de error asociadas (0,1–15%) son más altas y las longitudes de lectura generalmente más cortas (35–700 pb para enfoques de lectura corta) que las de las plataformas de secuenciación Sanger tradicionales (≈1000 pb). Las técnicas de NGS requieren, por lo tanto, individuos altamente capacitados en bioinformática, idóneos a momento de realizar un examen cuidadoso de los datos y resultados, particularmente para el descubrimiento de variantes y aplicaciones clínicas. [23]

3.1.1. El genoma de referencia

Los genomas de referencia son construcciones genéticas que se han hecho con el objetivo de tener una referencia que puede usarse como punto de comparación. Han habido distintas construcciones del genoma a lo largo de los años. Los más utilizados son el hg19 (publicado en 2009), el cual comprende una representación única de múltiples genomas, y la última versión del genoma publicada, el GRCh *Build* 38 (GRCh38, publicado en 2013) que proporciona secuencias alternativas para algunas regiones genómicas cuya variabilidad impide una representación adecuada mediante una única referencia. [24]

Por otro lado, también han habido mejoras en la representatividad del genoma de referencia con respecto a las distintas poblaciones humanas. Podría decirse que las mejoras más significativas se han realizado en la representación de los llamados haplotipos alternativos[1], los cuales luego son útiles a momento de predecir frecuencias poblacionales.

La existencia de las diferentes construcciones del genoma, agravada por el advenimiento de dos flujos paralelos de evolución[2] ha causado mucha confusión y abundantes errores a lo largo de los años. Parte del problema es que muchas herramientas bioinformáticas no imponen el uso constante de una referencia específica. Por estas razones, en la bioinformática, se han utilizado las últimos dos versiones del genoma de referencia para la realización de este trabajo; fundamentalmente el GRCh38 dado que es una representación más precisa y es la versión más reciente aceptada por la comunidad.

Por otra parte, en 2019 también se introdujeron los genomas de referencia gráficos [25]. Estos portan haplotipos variables que no están presenten en el genoma de referencia común lo que hace que tengan el potencial no solo de aumentar el número de lecturas alineadas y resolver haplotipos, sino también de construir una mejor representación de la diversidad entre las distintas poblaciones, es decir, entre individuos que presentan una variación genética común debido a su origen, en relación a individuos con otro origen.

3.1.2. Parámetros importantes en un análisis NGS

Hay ciertos factores que se consideran convenientes de mencionar respecto a un análisis de secuenciación masiva:

- Parámetros de calidad de secuenciación (altamente dependiente de la complejidad de las regiones del genoma):

    1. Calidad de mapeo: Informa la probabilidad de que una lectura no se haya alineado al genoma de referencia correctamente.

[1] Regiones que a veces son dramáticamente diferentes en diferentes poblaciones

[2] Llamados hg* o b*, publicados por diferentes grupos con diferentes convenciones de nomenclatura, y algunas diferencias en las secuencias.

2. Calidad de llamada de la base nucleotídica: Informa la probabilidad de que la base alineada haya sido detectada como una variante cuando no lo sea.

- Frecuencia alélica: La proporción de cada alelo hallado en una posición de la secuencia. Si se hallan para una posición dos alelos en igual proporción, el paciente es heterocigoto para ese alelo; si solo hay un alelo será homocigoto.

- Profundidad de lectura o cobertura vertical: Describe el número de veces que se ha leído un nucleótido dado en el análisis del genoma. Para un experimento NGS, la profundidad de lectura promedio es una medida de la integridad general del conjunto de datos. Dado que el proceso de fragmentación del genoma secuenciado genera lecturas aleatoriamente, se deben generar una gran cantidad de fragmentos para saber que todas las áreas estarán representadas por superposiciones y podrán alinearse correctamente con el genoma de referencia dado. Por lo tanto, las profundidades de lectura promedio deben ser bastante altas para volver a ensamblar con precisión lecturas contiguas largas (se busca un promedio cercano a 30, o 30X, pero es sabido que algunas regiones del genoma estarán subrepresentadas y algunas se leerán con mayor profundidad).

- Frecuencia alélica poblacional: La frecuencia alélica poblacional consiste en la proporción de cada alelo en un locus dado en una población específica. La suma de las frecuencias alélicas en una población siempre es 1 (o 100 %). Esta frecuencia es de particular interés en cuanto a la transmisión de los genes y su genotipo en una población. [26]

- Cobertura horizontal: Comprende las secciones del genoma que están representadas en el conjunto de datos al menos con cierta profundidad, a estas secciones también se las conoce como regiones de captura y pueden comprender de uno o un par de genes al genoma en su totalidad.

3.2. Colegio Americano de Genética Médica y Genómica (American College of Medical Genetics and Genomics - ACMG)

Debido a la complejidad creciente de los datos de secuenciación, han habido nuevos desafíos en la interpretación de secuencias. En este contexto, el ACMG convocó un grupo de trabajo en 2013 compuesto por representantes del Colegio Americano de Genética Médica y Genómica (ACMG), la Asociación de Patología Molecular (AMP) y el Colegio de Patólogos Estadounidenses para revisar los estándares y las pautas para la interpretación estandarizada de variantes genéticas. Así, en el año 2015, se desarrolló una guía para la interpretación de variantes.

Esta guía se aplica principalmente a la variedad de pruebas genéticas que se utilizan en los laboratorios clínicos, incluidos el genotipado[3], los genes individuales, los paneles, los exomas y los

[3]Determinación del genotipo de una variante en el ADN específico de un organismo biológico

genomas.

Al momento de clasificar a las variantes, se tienen en cuenta 28 criterios para valorar la implicación clínica de cada una: características típicas de evidencia de variantes, como pueden ser la frecuencia poblacional, análisis de casos y controles, estudios funcionales, predicciones computacionales, datos alélicos, estudios de segregación y observaciones de novo. Cada criterio tiene un código que se compone por un peso (muy fuerte/*very strong*, fuerte/*strong*, moderada/*moderate*, de soporte/*supporting*) y una dirección (patogénica, probablemente patogénica, importancia incierta, probablemente benigna y benigna); y la conclusión final de la patogenicidad asignada a cada variante particular siguiendo la guía ACMG surge de un cálculo específico que pondera la relevancia de cada evidencia (ver figura 3.1).

	Benign		Pathogenic			
	Strong	**Supporting**	**Supporting**	**Moderate**	**Strong**	**Very Strong**
Population Data	MAF is too high for disorder *BA1/BS1* **OR** observation in controls inconsistent with disease penetrance *BS2*			Absent in population databases *PM2*	Prevalence in affecteds statistically increased over controls *PS4*	
Computational And Predictive Data		Multiple lines of computational evidence suggest no impact on gene /gene product *BP4* Missense in gene where only truncating cause disease *BP1* Silent variant with non predicted splice impact *BP7*	Multiple lines of computational evidence support a deleterious effect on the gene /gene product *PP3*	Novel missense change at an amino acid residue where a different pathogenic missense change has been seen before *PM5* Protein length changing variant *PM4*	Same amino acid change as an established pathogenic variant *PS1*	Predicted null variant in a gene where LOF is a known mechanism of disease *PVS1*
Functional Data	Well-established functional studies show no deleterious effect *BS3*		Missense in gene with low rate of benign missense variants and path. missenses common *PP2*	Mutational hot spot or well-studied functional domain without benign variation *PM1*	Well-established functional studies show a deleterious effect *PS3*	
Segregation Data	Non-segregation with disease *BS4*		Co-segregation with disease in multiple affected family members *PP1*	Increased segregation data →		
De novo Data				*De novo* (without paternity & maternity confirmed) *PM6*	*De novo* (paternity & maternity confirmed) *PS2*	
Allelic Data		Observed in *trans* with a dominant variant *BP2* Observed in *cis* with a pathogenic variant *BP2*		For recessive disorders, detected in *trans* with a pathogenic variant *PM3*		
Other Database		Reputable source w/out shared data = benign *BP6*	Reputable source = pathogenic *PP5*			
Other Data		Found in case with an alternate cause *BP5*	Patient's phenotype or FH highly specific for gene *PP4*			

Figura 3.1: Pesos y direccionalidad de las variantes genéticas que se tienen en cuenta por ACMG a momento de realizar la predicción de patogenicidad. [27]

La variante se clasifica entonces según su grado de patogenicidad en:

- Patogénica
- Probablemente patogénica
- Variante de significado incierto (VUS)

- Probablemente benigna
- Benigna [28]

3.3. Asociación de Patología Molecular (Association for Molecular Pathology - AMP)

En el año 2015, la Asociación de Patología Molecular convocó a un grupo de trabajo multidisciplinario encargado de evaluar el estado actual de las pruebas de cáncer basadas en NGS y establecer convenciones estandarizadas de clasificación, anotación, interpretación y notificación de consenso para variantes de secuencias somáticas en base a los criterios establecidos por ACMG. Las variantes somáticas incluyen SNPs, INDELs, genes de fusión resultantes de reordenamientos genómicos y CNVs. A diferencia de la interpretación de las variaciones de la secuencia de la línea germinal, que se centra en la patogenicidad de una variante para una enfermedad específica o la causalidad de la enfermedad, la interpretación de las variantes somáticas debe centrarse en su impacto en la atención clínica. Una variante puede considerarse un biomarcador que afecta la atención clínica si predice sensibilidad, resistencia o toxicidad a una terapia específica, altera la función del gen, que puede ser el objetivo de medicamentos aprobados o en investigación, sirve como criterio de inclusión para ensayos clínicos, influye en el pronóstico de la enfermedad, ayuda a establecer un diagnóstico de cáncer o garantiza la implementación de medidas de vigilancia para la detección temprana del cáncer. Los impactos clínicos deberían, por lo tanto, incluir acciones terapéuticas, pronósticas, diagnósticas y preventivas. [29]

El impacto clínico de una variante dada debe determinarse de acuerdo con la evidencia actualmente disponible. La evidencia utilizada para la categorización de variantes se puede sopesar de manera diferente en función de su importancia en la toma de decisiones clínicas. AMP propuso entonces, un sistema de cuatro niveles de clasificación: [29]

- Variantes de Nivel I: Variantes con fuerte importancia. Comprenden variantes que presentan terapias aprobadas por la FDA o estudios bien fundamentados con consenso de expertos en el campo.
- Variantes de Nivel II: Variantes con posible importancia clínica. Comprenden variantes que presentan terapias aprobadas por la FDA para diferentes tipos de tumores o terapias en investigación, o que presenten múltiples estudios pequeños publicados con cierto consenso, o ensayos preclínicos/informes de casos sin consenso.
- Variantes de Nivel III: Variantes con de significado clínico desconocido. Comprenden variantes que no se han observado en una frecuencia significativa de alelos en las bases de datos de subpoblaciones generales o específicas, o en las bases de datos de variantes pan-cáncer o

específicas de tumores; o bien no hay evidencia publicada convincente de su asociación con el cáncer.

- Variantes de Nivel IV: Variantes consideradas benignas o probablemente benignas. Son variantes observadas a una frecuencia alélica significativa en las bases de datos de subpoblaciones generales o específicas, o bien variantes para las cuales no existe evidencia publicada o asociación con el cáncer.

CAPÍTULO 4

BIOINFORMÁTICA: DESARROLLO DE UN FLUJO DE ANÁLISIS DE VARIANTES GENÓMICAS

La Bioinformática es un campo de las ciencias computacionales que permite investigar, desarrollar y aplicar herramientas informáticas para recopilar, almacenar y analizar y diseminar datos e información biológicos, incluyendo datos de NGS. [30, 31]

4.1. Flujos de análisis de detección de mutaciones germinales y somáticas

En la informática, un Flujo de análisis o *pipeline*, consiste en una serie de procesos o pasos secuenciales que permiten actuar sobre o transformar un conjunto de datos. [32, 33]

En el contexto de la genómica y la bioinformática, un flujo de análisis está compuesto por un conjunto de algoritmos que utilizan Herramientas bioinformáticas, matemáticas y estadísticas ejecutadas en un orden definido para recopilar, organizar y analizar datos de secuenciación genética y datos biológicos relacionados. [27, 34]

Dentro de los flujos de análisis más frecuentemente usados, se encuentran los flujos para WES y WGS tanto para variantes germinales como somáticas, cada una con sus particulares propias de las condiciones de la secuenciación.

Entre las diferencias más relevantes, encontramos que mientras que el estudio bioinformático proveniente de un análisis WES solo observa las regiones exómicas del genoma, con un análisis WGS se pueden evaluar todos los nucleótidos de un genoma individual y permitir la detección de variantes tanto en regiones codificantes como no codificantes. Por lo tanto, el flujo de análisis tiene que desarrollarse considerando estas diferencias.

Por otro lado, independientemente de si el análisis proviene de WES o WGS, la mayoría de los laboratorios clínicos que realizan diagnósticos de trastornos genéticos se enfocan en dos tipos de variantes: SNPs e INDELs. [35–37]

El flujo de trabajo general de genética clínica o del procesamiento y análisis de los estudios de NGS se describe en la figura 4.1:

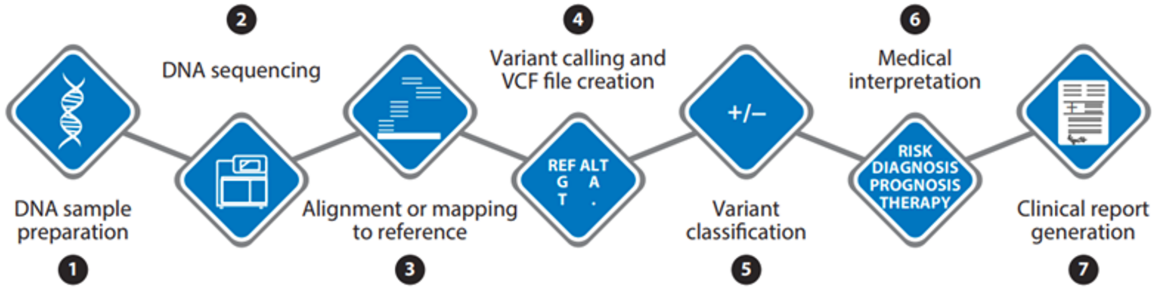


Figura 4.1: *Flujo de análisis general para la análisis genómico en el contexto de investigación (figura adaptada de [38])*

Las primeras dos etapas, de preparación de la muestra de ADN y de secuenciación, se hacen en un laboratorio de genómica a través de la técnica de NGS explicada anteriormente en la sección 3.1. Aquí, se pueden secuenciar tanto muestras germinales como somáticas. A cada nucleótido secuenciado en fragmentos cortos de ADN (o lecturas) se le asigna una puntuación de calidad específica en la plataforma de secuenciación [38], lo que servirá para el seguimiento del análisis de calidad del proceso.

A continuación, interviene la bioinformática según el análisis que se desee ejecutar. En la figura 4.2, se describen los pasos a seguir a momento de diseñar un flujo de análisis para variantes germinales que comprenden SNPs e INDELs.

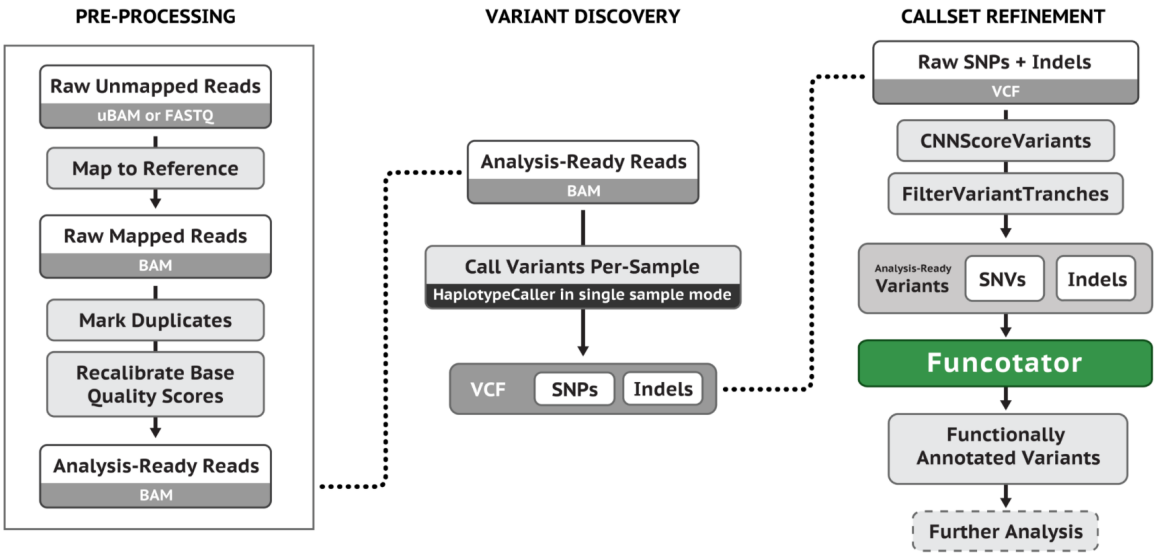


Figura 4.2: *Flujo de análisis general para la análisis germinal, desde la salida del secuenciador (archivos crudos FASTQ) hasta la anotación de variantes (figura adaptada de [39]).*

Una vez obtenidos los archivos crudos de secuenciación (cuyo formato es FASTQ) y realizado su preprocesamiento (con herramientas como Cutadapt y Samtools, capaces de eliminar secuencias no deseadas generadas en el proceso de secuenciación y obtener archivos con un índice necesarios para el alineamiento), se generan las partes de la secuencia que deberán alinearse con la referencia. En este proceso, se compara al genoma de interés con el genoma de referencia y se "aparean" las diferentes lecturas o fragmentos de ADN que se hayan producido durante la preparación de librerías. Como resultado del alineamiento se obtiene un archivo BAM (del inglés *Binary Sequence Alignment Map*). El genoma de referencia empleado es el "estándar" para una especie dada, lo que permite identificar genotipos; en específico, en genomas diploides permite determinar si un individuo para un alelo determinado es homocigoto o heterocigoto.[1] [40]

A pesar de que se considera BWA-MEM como el *gold standard* para alineación de lecturas cortas, se han desarrollado y se usan comúnmente varios otros alineadores como Bowtie2, Isaac (Illumina Inc. EE. UU.) y Novoalign (Novocraft Technologies, EE.UU). [27, 35, 40–42] Con la incorporación de los genomas de referencia gráficos, se han desarrollado otras herramientas de alineamiento como GRAF aligner que han demostrado tener mejores resultados tanto en métricas como en eficiencia computacional. [43]

Una vez que las lecturas están alineadas con el genoma de referencia se genera el llamado de variantes, el cual resulta en la generación de un archivo en formato VCF. Se han desarrollado múltiples herramientas para identificar las variantes SNP e INDEL, siendo los mayores desafíos de esta etapa el llamado en regiones de alta variabilidad y regiones repetitivas del genoma. [40, 41, 44]

[1] En el análisis somático, el genoma de referencia proviene de un tejido relacionado del mismo individuo. En este caso se busca determinar mosaiquismo entre células.

En este paso, se usan herramientas como Haplotype Caller, Varscan, FreeBayes, GRAF variant caller, Strelka2, SAMTtools (para el caso de análisis germinal) y Mutect2, FreeBayes y Strelka2 (para el caso somático), seguido por herramientas de filtrado para eliminar errores generados en la secuenciación y seleccionar únicamente las variantes reales[2] . [35, 40, 41]

Una vez obtenido el resultado del llamado de variantes, se procede a la clasificación y anotación de las variantes. La anotación de variantes es el proceso por el cual se le asigna información funcional a las variantes encontradas. Para la clasificación de las variantes, se pueden utilizar diferentes criterios, según si la muestra es del tipo somática o germinal, para determinar el grado de patogenicidad de las variantes listadas en el VCF[3]. [45, 46]

Existen distintos proyectos que comprenden varias de estas herramientas, como aquellas pertenecientes al *Genome Analysis Toolkit* (GATK)[4], de código libre o las GRAF (de *Cancer Genomics Cloud*), que permiten procesar el análisis desde el archivo crudo hasta el llamado de variantes, aunque no incluyen la anotación biológica/fisiopatológica/clínica necesaria para la toma de decisión terapéutica. Por otro lado, los anotadores como ANNOVAR, SNPEffect, etc., son herramientas que comúnmente podrían llegar a integrarse en estos flujos de análisis pero no incluyen la predicción de patogenicidad de acuerdo a las recomendaciones de ACMG para las variantes germinales o AMP para las somáticas. Según el conocimiento disponible hasta el momento, no existen flujos de análisis de variantes de código libre que hagan predicción de patogenicidad de acuerdo a las recomendaciones idóneas. Sin embargo, los algoritmos de InterVar y CancerVar se basan en ANNOVAR; y son capaces de anotar un archivo de variantes incluyendo la predicción de patogenicidad de ACMG y AMP respectivamente.

Los últimos pasos del análisis bioinformático corresponden a la interpretación médica por parte de un genetista de las variantes y la generación de un reporte o informe. [38, 46]

4.2. Validación

Desarrollar archivos de variantes *gold standard* para la validación de flujos de análisis ha sido un desafío en la bioinformática principalmente a la falta de evidencia ortogonal y a que los instrumentos de secuenciación todavía presentan tasas de error inevitables. En 2015, el Consorcio Genoma en una Botella (GIAB), dirigido por el Instituto Nacional de Estándares y Tecnología (NIST) , publicó un conjunto de datos de referencia de alta calidad que permiten utilizarse como *gold standard* de SNPs e INDELs empleando diversas tecnologías de secuenciación. Este conjunto

[2]Herramientas como CNN Score Variants acompañada de Filter Variant Tranches (caso germinal) y Filter Mutect (caso somático) se usan para filtrar por ciertos criterios que disponga el usuario como por ejemplo; frecuencia poblacional, frecuencia alélica, etc

[3]Para más información sobre los tipos de archivos, ver la sección B.1 del anexo.

[4]Hoy en día es el estándar de la industria para identificar SNP e INDELs en datos de ADN de línea germinal, viene acompañado de recomendaciones completas para el diseño de flujos de análisis. Pertenece a *Broad Institute* de MIT y Harvard

ha sido usado ampliamente por la comunidad bioinformática para el desarrollo, comparación y validación de flujos de análisis. [36, 47]

La herramienta Hap.py [48] está diseñada para comparar llamados de variantes *gold standard* (proporcionados por GIAB) con archivos VCF de prueba. Hap.py informa la sensibilidad y precisión del análisis a partir del cálculo de verdaderos positivos, falsos positivos y falsos negativos.

Los verdaderos positivos se calculan como aquellas variantes que fueron encontradas tanto en la referencia como en la muestra; los falsos positivos, como las variantes encontradas en la muestra y que no están en la referencia, y por último; los falsos negativos, como las variantes encontradas en la referencia y no en la muestra.

Para la validación de flujos de análisis de detección de variantes, se debe tener en cuenta que no se definen verdaderos negativos o especificidad porque estos no son aplicables a la secuenciación del genoma; dado que, por ejemplo, hay una cantidad infinita de posibles INDELs existentes, por lo que hay una cantidad infinita de verdaderos negativos para cualquier ensayo. [49]

4.3. Infraestructura

4.3.1. Computación en la nube y Amazon Web Services

La computación en la nube comprende la distribución de recursos informáticos bajo demanda a través de Internet mediante un esquema de pago por uso. En lugar de comprar, poseer y mantener servidores y centros de datos físicos, se puede acceder a servicios tecnológicos, como capacidad informática, almacenamiento y bases de datos, en función de las necesidades a través de un proveedor de la nube como *Amazon Web Services* (AWS). [50]

Algunos beneficios de la computación en la nube son la agilidad y disponibilidad de una variedad de servicios especializados. A su vez, es una opción flexible ya que no es necesario aprovisionar recursos en exceso con antelación para gestionar niveles pico de actividad computacional a futuro, en cambio, aprovisiona la cantidad de recursos que realmente se necesitan. Es decir, puede ajustar la escala de estos recursos para aumentar o disminuir la capacidad instantáneamente a medida que cambien las necesidades computacionales. [50]

AWS es uno de los servicios de computación en la nube más confiables, y algunos de los servicios útiles son:

- EC2 (*Amazon elastic compute cloud*): servidores virtuales en la nube con más de 500 instancias y la posibilidad de elegir el procesador, almacenamiento, redes, sistema operativo y modelo de compra más reciente lo cual permite ajustar las necesidades a la carga de trabajo. [51]

- S3 (*Amazon Simple Storage Service*): servicio de almacenamiento de objetos que ofrece escalabilidad, disponibilidad de datos, alta seguridad y rendimiento. [52]

- IAM (*Identity and Access Management*): servicio de administración de manera segura de las identidades y el acceso a los recursos y servicios de AWS. [53]

4.3.2. Common Workflow Language (CWL) y Docker

CWL es un lenguaje para definir e instrumentar flujos de trabajo, o sea, procesos o herramientas encadenados que permiten automatizar análisis genómicos complejos como la alineación de un genoma o exoma frente a una referencia, la llamada de variantes de ADN, la secuenciación de ARN, entre otras. Este lenguaje además permite empaquetar estas herramientas de manera muy simple, y convertirlas en un flujo de análisis rápidamente ya que permite tomar la salida de cada proceso como la entrada del siguiente. A su vez, posee una interfaz gráfica (Rabix Composer [54]) en donde se pueden diseñar, escribir y construir los flujos de análisis arrastrando cada herramienta creada sobre la pantalla principal de Rabix. Por ejemplo, a continuación se muestra el comando del *"Hello World"* empaquetado en CWL.

hello_world.cwl #

```yaml
cwlVersion: v1.2

# What type of CWL process we have in this document.
class: CommandLineTool
# This CommandLineTool executes the linux "echo" command-line tool.
baseCommand: echo

# The inputs for this process.
inputs:
  message:
    type: string
    # A default value that can be overridden, e.g. --message "Hola mundo"
    default: "Hello World"
    # Bind this message value as an argument to "echo".
    inputBinding:
      position: 1
outputs: []
```

Figura 4.3: *Código ejemplo donde se muestra un contenedor para la herramienta de línea de comandos echo. Ejecutar esta herramienta con los valores de entrada predeterminados producirá el mismo resultado que ejecutar* $echo\ "Hello\ World"$ *en la línea de comando de Linux. [55])*

Otra ventaja que brinda CWL es que permite integrarse fácilmente con Docker[5] [56]. De hecho, este es uno de los beneficios más importantes de CWL, dado que muchos softwares, incluyendo GATK, ya están disponibles en Docker, lo que permite construir herramientas y flujos de análisis de manera sencilla y con actualización automática.

[5]Docker es esencialmente un conjunto de herramientas que posibilita a los desarrolladores crear, implementar, ejecutar, actualizar y detener contenedores mediante comandos simples y automatización a través de una única API.

4.3.3. Seven Bridges: Cancer Genomics Cloud

Seven Bridges (SB) es una empresa especializada en software, datos biomédicos y su análisis para impulsar la investigación en salud pública y privada. SB brinda herramientas bioinformáticas, incluido el acceso a conjuntos de datos, flujos de trabajo analíticos y algoritmos, infraestructura de computación en la nube y soporte científico, que facilitan el camino desde los datos experimentales sin procesar hasta los tratamientos y diagnósticos. [57]

CGC, impulsada por SB y financiada por el Instituto Nacional del Cáncer, es una plataforma en la nube sin fines de lucro que permite el análisis, el almacenamiento y el cálculo de grandes conjuntos de datos sobre el cáncer. CGC utiliza AWS como apliación servidor y proporciona un portal intuitivo para acceder y analizar datos sobre el cáncer. Cualquier usuario con una cuenta puede acceder a petabytes de datos sobre el cáncer, compartirlos, analizarlos y usar el poder computacional de la nube sin necesidad de saber programar. [58] Sin embargo, si se desea desarrollar herramientas personalizadas, es necesario implementarlas en CWL.

CAPÍTULO 5

OBJETIVOS

El objetivo general de este trabajo es desarrollar un flujo de análisis automatizado y validado para la interpretación de variantes (de acuerdo a guías y recomendaciones internacionales) obtenidas a partir de muestras germinales y tumorales.

5.1. Objetivos Específicos de mínima

- Realizar una búsqueda bibliográfica de bases de datos de variantes y algoritmos de predicción de patogenicidad
- Realizar un análisis de sensibilidad y especificidad del flujo de análisis germinal ya disponible en el laboratorio tanto para análisis de exoma completo como de genoma completo, e investigar opciones de mejora para dicho flujo.
- Desarrollar un flujo de análisis en CWL incorporando las herramientas de interpretación estudiadas e integrarlo con el pipeline anterior

5.2. Objetivo Específico de máxima

- Analizar alternativas para la presentación en forma amigable del informe de NGS.

CAPÍTULO 6

HIPÓTESIS

Nuestra hipótesis es que la complejidad genómica humana, que se traduce en la complejidad para encontrar biomarcadores diagnósticos, pronósticos y de respuesta a tratamiento, puede interpretarse más fácil y rápidamente mediante la incorporación de flujos de análisis automatizados para datos NGS y específicamente diseñados para la patología de interés.

Parte II

Materiales y Métodos

CAPÍTULO 7

ARCHIVOS Y MUESTRAS UTILIZADOS

En este capitulo se mencionan todas las muestras utilizadas para la validación de los análisis de variantes WES y WGS y los archivos necesarios para ejecutar los distintos análisis.

Para los análisis de exoma se utilizaron:

- Muestras de exoma publicadas en GIAB (versión 4.2.1):
  - NA12878 (HG001) [59]
  - NA24385 (HG002) [60]
  - NA24631 (HG005) [1][61]
- Las secuencias de adaptadores de kits Illumina TruSeq e Illumina TruSeq CD[2], dado que con estos se produjo la secuenciación de las muestras anteriormente listadas
- Genomas de referencia:
  - Genoma de referencia lineal GRCh38/ GRCh37 (publicados el 12-04-2021 y el 07-01-2021 respectivamente) [62], [63]
  - Genoma de referencia gráfico GRCh38 para aquellos flujos de análisis que incorporen herramientas de GRAF (publicado el 17-04-2020) [64]

[1] Las últimas dos muestras de exoma se descargaron a partir de un enlace publicado en BCBIO (versión única) debido a su dificultad de obtención directa con GIAB. BCIO es un kit de herramientas de Python que proporciona flujos, recomendaciones y guías de mejores prácticas para el análisis de datos de secuenciación masiva.

[2] Hebra forward: .AGATCGGAAGAGCACACGTCTGAACTCCAGTCA"
Hebra reverse: .AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT

- Archivos VCF para la anotación de variantes conocidas (versión 0 de las referencias publicadas por el Broad Institute, tanto para hg19 como GRCh38):
    - Homo_sapiens_assembly19.dbsnp138.vcf.gz (publicado el 07-12-2019) [65]
    - Mills_and_1000G_gold_standard.indels.b37.vcf.gz (publicado el 07-12-2019) [65]
    - 1000G_phase1.snps.high_confidence.b37.annotated.vcf.gz (publicado el 07-12-2019) [65]
    - 1000G_omni2.5.b37.vcf.gz (publicado el 21-07-2016) [65]
    - phase1.snps.high_confidence.b37.vcf.gz (publicado el 24-04-2018) [65]
    - Homo_sapiens_assembly19.vcf (publicado el 13-08-2020) [65]
    - Homo_sapiens_assembly38.dbsnp138.vcf.gz (publicado el 07-04-2021) [66]
    - Mills_and_1000G_gold_standard.indels.hg38.annotated.vcf.gz (publicado el 07-12-2019) [66]
    - 1000G_phase1.snps.high_confidence.hg38.annotated.vcf.gz (publicado el 07-12-2019) [66]
    - 1000G_omni2.5.hg38.vcf.gz (publicado el 07-12-2019) [66]
    - hapmap_3.3.hg38.vcf.gz (publicado el 07-12-2019) [66]
    - phase1.snps.high_confidence.hg38.vcf.gz (publicado el 07-12-2019) [67]
    - Homo_sapiens_assembly38.dbsnp.annotated.vcf.gz (publicado el 21-07-2016) [66]
- Archivos BED:
    - Exome-Agilent-SureSelect-v05-GRCh38.bed para NA24385 y NA24631 (publicado el 25-11-2015) [68]
    - Exome-IDT-xGen-hg38.bed para NA12878 (publicado en 2015)[69]
- Archivos VCF para realizar la validación:
    - HG001_GRCh38_1_22_v4.2.1_benchmark.vcf.gz (publicado el 20-09-2021) [70]
    - HG002_GRCh38_1_22_v4.2.1_annotated.vcf.gz (publicado el 12-07-2021) [71]
    - HG005_GRCh38_1_22_v4.2.1_benchmark.vcf.gz (publicado el 20-09-2021) [72]

Para los análisis de genoma completo:

- Muestra de genoma completo (archivos FASTQ, publicados el 13-02-2020 [73])
- Genoma de referencia lineal GRCh38 (publicado el 12-04-2021) [62]
- Genoma de referencia gráfico GRCh38 (publicado el 17-04-2020) [64]

- Archivo VCF para realizar la validación (publicado el 20-09-2021) [70]
- Archivo BED: HG001_GRCh38_1_22_v4.2.1_benchmark.bed (publicado el 20-09-2021) [70]

CAPÍTULO 8

HERRAMIENTAS DEL FLUJO DE ANÁLISIS

El procesamiento de los archivos crudos (FASTQs) mediante dos pasos principales: alineamiento y llamado de variantes. En ambos pasos, se utilizó un conjunto de herramientas distintas que luego fueron comparadas entre sí.

8.1. Alineamiento

Primer conjunto de herramientas: Cutadapt [74–76] (versión 4.2) para la eliminación de adaptadores y secuencias de baja calidad, BWA-MEM [77–80] (versión 0.7.1) para el alineamiento de las lecturas, Samtools [81] (versión 1.16) para el ordenamiento de las lecturas y Mark Duplicates [82] (versión 4.0.1.1) para marcar secuencias duplicadas.

Segundo conjunto de herramientas: GRAF Aligner (utilizando los parámetros Generate index, Mark duplicates, Sort output y Trim adapters en "True") para el alineamiento directamente desde el FASTQ y la eliminación de adaptadores y secuencias de baja calidad [83] (versión 18-06-2020).

8.2. Llamado de variantes

Primer conjunto de herramientas germinal: GATK (versión 4.2.6.1) [39]: Haplotype Caller [84], CNN Score Variants [85], Filter Variant Tranches [86] (para el método de filtrado de redes neuronales) ; y por último, SelectVariants [87] y VariantFiltration [88] (para el método de Hard Filtering).

Segundo conjunto de herramientas germinal GRAF (revisión 3): GRAF Variant Caller (publicado el 13-05-2022, [89]) y GRAF Variant Filtration (publicado el 18-06-2020 [90]).

Conjunto de herramientas somático (GATK 4.2.6.1): Mutect2 [91] y FilterMutectCalls [92].

Para más información ver la sección B.2.

8.3. Herramientas de comparación

Se utilizó Hap.py [48, 49] (versión 0.3.15) que es una herramienta que permite comparar un VCF de prueba contra un gold standard para obtener valores de sensibilidad, precisión y valor F1.

Hap.py [48] informa recuentos de:

- Verdaderos positivos (TP): variantes/genotipos que coinciden en los archivos verdad y consulta.
- Falsos positivos (FP): variantes que tienen genotipos o alelos que no coinciden, así como llamadas de variante de consulta en regiones que un conjunto de verdad llamaría regiones hom-ref seguras.
- Falsos negativos (FN): variantes presentes en el conjunto de verdad, pero perdidas en la consulta.

A partir de estos conteos, Hap.py es capaz de calcular:

- Sensibilidad[1] = $\frac{TP}{TP+FN}$
- Precisión = $\frac{TP}{TP+FP}$
- $F1_{Score}$ = $\frac{2*Precision*Recall}{Precision+Recall}$

Para la creación de gráficos y cálculos personalizados se utilizó Python (versión 3.7).

8.4. Anotación de variantes y creación de los informes

Se utilizó InterVar [29] (versión 13-06-2022) para las variantes germinales y CancerVar [93] (versión 10-05-2022) para las somáticas. Para la creación del archivo con la anotación de variantes más amigable y el informe para el solicitante se utilizó Python (versión 3.7) y RStudio (versión 4.2.2). Se incluyeron como parámetros de calidad la profundidad de lectura media y la tasa TSTV[2] halladas.

[1] Recall

[2] La tasa TS/TV comprende a la proporción de transiciones a transversiones. Esta relación suele ser ≈2 en análisis de genoma completos, y ≈3 para exomas. Se incluye fundamentalmente para ratificar de manera global el análisis bioinformático realizado

CAPÍTULO 9

INFRAESTRUCTURA

En este capitulo se describen todas las herramientas de infraestructura necesarias para el desarrollo de este trabajo.

- Se utilizo Common Workflow Language (CWL, versión 1.0) para el empaquetado de herramientas y desarrollo de los flujos de análisis.
- Instancias EC2 en AWS: t3.2xlarge con sistema operativo Amazon Linux Machine.
- S3 para el almacenamiento de archivos.
- La nube de CGC para la implementación de los flujos de análisis en un contexto más amigable para usuarios no expertos.

Parte III

Resultados

CAPÍTULO 10

DISEÑO DE LOS FLUJOS DE ANÁLISIS

El diseño de los flujos de análisis consistió de dos etapas: la primera en AWS, en donde se contaba con un espacio de prueba con mayor flexibilidad según los requerimientos; y la segunda en CGC en donde si bien las capacidades eran más limitadas, fue la plataforma elegida para disponibilizar los flujos de análisis en un entorno más amigable, incluso para usuarios no expertos.

10.1. Diseño en AWS

En esta primera etapa de diseño, se utilizaron los recursos de AWS para contar con un espacio de pruebas en donde fuese posible familiarizarse con las herramientas bioinformáticas y testearlas previo a ser empaquetadas en CWL. Para ello, se utilizaron instancias EC2 a través de la línea de comando en donde se corrían las herramientas probando distintos argumentos y corroborando su correcto funcionamiento (por ejemplo, verificando que los archivos resultantes tuvieran el tamaño y formato correctos y que su contenido fuera acorde a lo esperado). A su vez, se utilizó S3 para almacenar todos los archivos de entrada y salida requeridos e IAM a través del cual la tutora fue progresivamente otorgándonos más permisos de uso en la plataforma a medida que contábamos con más experiencia.

Posteriormente, CWL fue utilizado para empaquetar cada una de las herramientas necesarias, así como para programar la anidación entre ellas dando lugar a la creación de flujos de análisis[1].

Una vez que se testearon todos los pasos y su correcto funcionamiento en el entorno de AWS, se procedió a disponibilizarlo en CGC.

[1] Para ver el código de todas las herramientas y flujos de análisis, ver el link https://drive.google.com/drive/folders/1D5mvd0Zau2UbIUAKV8L5jTfY5GjhDYqf?usp=share_link

10.2. Diseño en CGC

En una segunda etapa, se procedió a disponibilizar las herramientas y flujos de análisis en CGC. Para esto se transfirieron los códigos y datos de la instancia EC2 a la plataforma, así como también los archivos presentes en S3, necesarios para la ejecución de los programas y flujos de análisis. CGC también permitió modificar y añadir herramientas públicas al flujo de análisis.

CAPÍTULO 11

FLUJOS DE ANÁLISIS GERMINALES

Para las variantes germinales, se diseñaron y disponibilizaron en CGC, según lo explica la sección 10, 5 flujos de análisis combinando las herramientas mencionadas anteriormente (ver sección 8). En la figura 11.0 se muestran los 5 flujos de análisis resultantes para las variantes germinales WES y WGS.

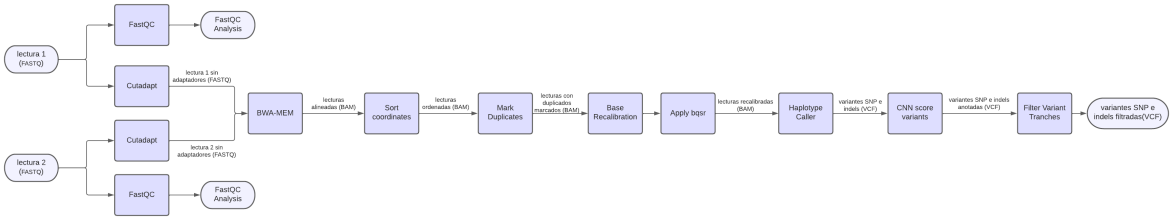
En el flujo diagramado en 11.0a, se puede observar el análisis inicial brindado por el laboratorio. Se nota que los archivos crudos de secuenciación (de entrada) pasan una etapa comprendiendo a las herramientas *FastQC* (para el análisis de calidad) y *Cutadapt* (para la eliminación de adaptadores), antes de alinearse con el genoma de referencia con *BWA-MEM*. Aquí se produce un archivo de salida con formato BAM que es la entrada de la herramienta *Sort Coordinate* (que ordena el archivo BAM por coordenadas cromosómicas), que luego produce la entrada de *Mark Duplicates* (donde se marcan las lecturas que se hayan duplicado durante la secuenciación). A continuación, se procede a las herramientas *Base Recalibration* y *Apply bqsr* que preparan al archivo BAM para la etapa posterior: el llamado de variantes. Esta etapa contiene a la herramienta *Haplotype Caller* (que identifica SNPs e INDELs) que produce como salida un archivo VCF. Por último, se produce el filtrado de las variantes con *CNN Score Variants* (que asigna puntuaciones a las variantes según la calidad del llamado) y *Filter Variant Tranches*, que añade una marca o *tag* a cada variante según si considere que fue o no falsamente llamada durante la etapa anterior.

El flujo que se muestra en 11.0b, se distingue del anterior pues la etapa de filtrado se produce con la herramienta *Hard Filtering*, que también asigna un *tag* a cada variante.

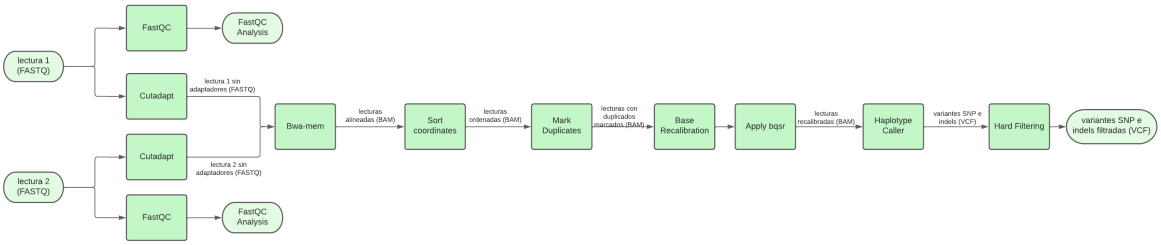
En la gráfica 11.0c, se muestra el flujo de análisis de Seven Bridges que utiliza las herramientas de GRAF, a diferencia de los flujos anteriores que usaban herramientas de GATK. En el caso de este flujo, los archivos crudos de secuenciación pasan por *GRAF Aligner*, que alinea las lecturas al genoma de referencia; y cuya salida es la entrada de *GRAF Variant Caller*, que produce un VCF con las variantes llamadas; y finalmente, *GRAF Variant Filter*, que produce el filtrado añadiendo un

tag a cada variante como resultado.

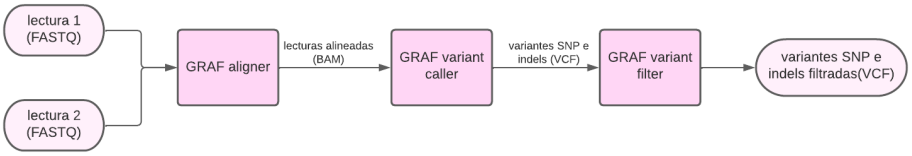
Las figuras 11.0d y 11.0e muestran combinaciones de los flujos 11.0a y 11.0b. El flujo 11.0d produce el alineamiento con *GRAF Aligner* y el resto de las etapas permanecen iguales que las de 11.0a; mientras que el flujo 11.0e usa el alineamiento de 11.0a con el llamado y filtrado de GRAF.



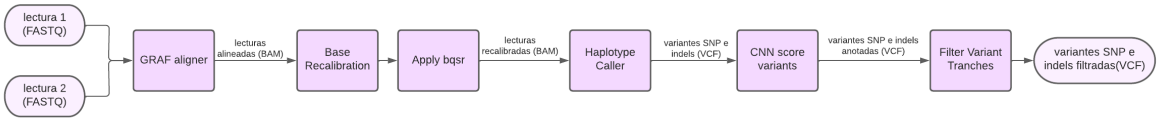
((a)) *Flujo de análisis de variantes germinales inicial con las herramientas utilizadas en el laboratorio.*



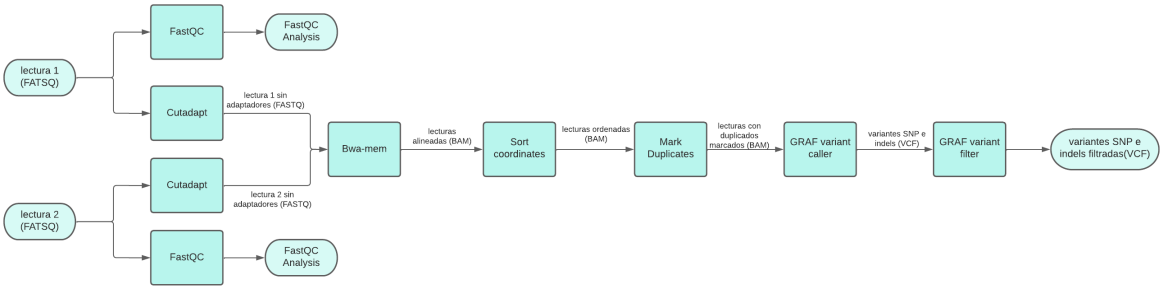
((b)) *Flujo de análisis inicial con el filtrado modificado con Hard Filtering.*



((c)) *Flujo de análisis de Seven Bridges que utiliza las herramientas GRAF.*



((d)) *Flujo de análisis que utiliza la herramienta de GRAF para alinear y el Haplotype Caller y Filter Variant Tranches.*



((e)) *Flujo de análisis compuesto por la herramienta de BWA-MEM para alinear y las herramientas GRAF para llamar y filtrar variantes.*

Figura 11.0: *Flujos de análisis diseñados para el análisis germinal de exoma. (a) Flujo de análisis de variantes germinales inicial, (b) Flujo de análisis de variantes germinales inicial modificado con el filtro Hard Filtering, (c) Flujo de análisis de variantes germinales GRAF de Seven Bridges, (d) Flujo de análisis de variantes germinales GRAF modificado con Haplotype Caller para el llamado de variantes y Filter Variant Tranches para el filtrado y, (e) Flujo de análisis de variantes germinales BWA-MEM GRAF que utiliza BWA-MEM para el alineamiento y GRAF para el llamado y filtrado de variantes.*

11.1. Análisis de Exoma

Para obtener el flujo de análisis óptimo para variantes germinales de WES, se analizaron cada uno de los flujos y/o partes de los flujos de la figura 11.0, priorizando los valores de sensibilidad y precisión; así como también el tiempo de ejecución (ya que este impacta directamente en el precio de análisis[1]).

Es importante tener en cuenta que las muestras WES NA24361 y NA24385, que resultaron difíciles de conseguir, se obtuvieron siguiendo los pasos definidos por BCBIO [94]. Los archivos crudos (FASTQ) se generaron a partir de los archivos de alineamiento (BAM) cuya última actualización fue en noviembre del 2015 [95].

11.1.1. Análisis de sensibilidad y precisión

Se corrieron los flujos de análisis de la figura 11.0 para las muestras de WES NA12878, NA24385 y NA24631 y se calculó el promedio de la sensibilidad, precisión y valor F1 para las variantes SNP e INDEL filtradas y no filtradas con Hap.py; obteniendo los resultados que se muestran en la tabla 11.1.

De esta tabla podemos observar que los flujos de análisis que utilizan las herramientas de GRAF tanto para el llamado de variantes como para su posterior filtrado, fueron los que presentaron un mayor valor F1, sensibilidad y precisión.

En particular, el flujo de análisis de la figura 11.0(c) que utiliza la herramienta de GRAF también para el alineamiento es el de mayor valor F1 alcanzando un valor promedio de 72,35% para INDELs filtradas y un 92,25% para SNPs filtradas. Este presentó una diferencia menor al 1% con el segundo flujo de análisis con mejores resultados (ver figura 11.0(e)) que utilizó también GRAF para el llamado y filtrado de variantes pero *BWA-MEM* para el alineamiento.

El flujo de análisis inicial del laboratorio fue el que presentó peores valores de F1 con una diferencia aproximada del 3% para las variantes INDEL, 1% para las variantes SNP sin filtrar y hasta un 12% para las variantes SNP filtradas.

Además, se obtuvo una amplia diferencia en los valores de las métricas entre SNP e INDEL, alcanzando el 20% de diferencia en todos los flujos de análisis.

[1] Se decidió no hacer mayor hincapié en un análisis de costos pues escapa a los objetivos de este trabajo. Sin embargo, se presentan los costos de cada análisis en particular en el informe y en los datos adjuntos.

Alineador	Llamado de variantes	Filtrado	Flujo de análisis	Tipo de variantes	Filtradas	Sensibilidad	Precisión	Valor F1
BWA-MEM	Haplotype Caller	Filter Variant Tranches	Inicial (11.0(a))	INDEL	No	79,42% (57,81% - 92,36%)	61,38% (51,89% - 68,84%)	69,06% (54,69% - 77,29%)
				INDEL	Sí	76,99% (56,90% - 90,23%)	62,46% (53,77% - 69,61%)	68,75% (55,29% - 76,07%)
				SNP	No	94,22% (85,61% - 98,69%)	87,15% (82,81% - 89,61%)	90,53% (84,18% - 93,93%)
				SNP	Sí	72,05% (68,12% - 78,48%)	89,42% (85,89% - 91,38%)	79,75% (75,98% - 84,27%)
		Hard Filtering	Inicial modificado (11.0(b))	INDEL	No	79,19% (57,53% - 92,15%)	62,33% (52,62% - 70,05%)	69,56% (54,97% - 77,97%)
				INDEL	Sí	78,97% (57,26% - 91,96%)	64,69% (56,23% - 72,02%)	70,85% (56,74% - 79,08%)
				SNP	No	94,21% (85,58% - 98,69%)	88,55% (83,87% - 91,23%)	91,27% (84,72% - 94,81%)
				SNP	Sí	88,62% (79,20% - 93,84%)	90,71% (87,50% - 92,57%)	89,59% (83,14% - 93,20%)
	GRAF	GRAF	BWA-MEM GRAF (11.0(e))	INDEL	No	80,16% (57,44% - 92,95%)	66,03% (60,54% - 72,76%)	71,93% (58,95% - 80,50%)
				INDEL	Sí	80,11% (57,44% - 92,87%)	66,48% (61,44% - 72,98%)	72,16% (59,37% - 80,61%)
				SNP	No	94,20% (85,48% - 98,72%)	89,53% (85,89% - 91,60%)	91,77% (85,68% - 95,03%)
				SNP	Sí	93,65% (84,67% - 98,24%)	90,86% (87,74% - 92,66%)	92,19% (86,17% - 95,37%)
GRAF	GRAF	GRAF	GRAF (11.0(c))	INDEL	No	80,88% (58,63% - 93,42%)	65,84% (60,55% - 72,37%)	72,14% (59,58% - 80,46%)
				INDEL	Sí	80,83% (58,63% - 93,33%)	66,27% (61,49% - 72,56%)	72,35% (60,02% - 80,55%)
				SNP	No	94,45% (85,77% - 98,93%)	89,53% (85,97% - 91,44%)	91,89% (85,87% - 95,04%)
				SNP	Sí	94,16% (85,38% - 98,70%)	90,50% (87,39% - 92,19%)	92,25% (86,37% - 95,33%)
	Haplotype Caller	Filter Variant Tranches	GRAF modificado (11.0(d))	INDEL	No	79,84% (58,36% - 92,58%)	62,73% (53,54% - 70,24%)	70,06% (55,84% - 78,35%)
				INDEL	Sí	77,27% (56,90% - 90,56%)	63,37% (55,07% - 70,28%)	69,38% (55,97% - 76,67%)
				SNP	No	94,43% (85,85% - 98,88%)	88,86% (84,33% - 91,30%)	91,54% (85,08% - 94,94%)
				SNP	Sí	72,11% (68,01% - 78,56%)	90,08% (86,67% - 91,83%)	80,05% (76,21% - 84,64%)

Cuadro 11.1: *Promedio de la sensibilidad, precisión y el valor F1 obtenido a partir de Hap.py para la muestras NA12878, NA24385 y NA24631 ejecutando los distintos flujos de análisis para exoma.*

Al graficar el promedio de todas las métricas por los tipos de variantes se obtiene el mapa de calor en la figura 11.1. Los colores más claros corresponden a valores más bajos y los colores más oscuros, más altos.

Tanto para INDELs como para SNPs, observamos que los colores más oscuros se observan en los flujos de análisis que utilizaron las herramientas de GRAF para el llamado y filtrado de variantes que corresponden a los mayores valores de precisión, sensibilidad y puntaje F1. En particular, se ve que el que presentó colores más claros y, por tanto, métricas más bajas, es el flujo de análisis inicial del laboratorio que utilizaba las herramientas *BWA-MEM*, *Haplotype Caller*, *CNN Score Variants* y *Filter Variant Tranches*.

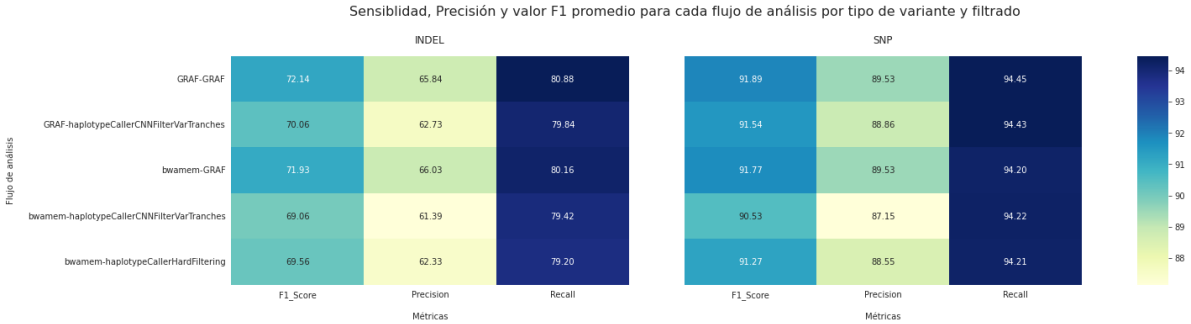


Figura 11.1: *Sensibilidad, Precisión y Valor F1 promedio para muestras WES para cada flujo de análisis por tipo de variante.*

11.1.2. Análisis de costo y tiempo

Luego de haber ejecutado todos los flujos de análisis de la figura 11.0 para las muestras de exoma completo de NA12878, NA24385 y NA24631; se calcularon el tiempo y costo promedios (ver tabla 11.2). Además, se le agregó el tipo de instancias utilizadas para correr cada uno de los flujos ya que esta interviene directamente en el costo y tiempo empleados.

Como se muestra en la tabla 11.2, el flujo de análisis que utiliza las herramientas de GRAF para el alineamiento, llamado y filtrado de variantes fue el de menor tiempo de ejecución tardando un promedio de 20 minutos, y menor costo con un promedio de 0,22 dólares; alcanzando una diferencia de hasta 12 veces menor con el que tardó mayor cantidad de tiempo (el flujo de análisis inicial del laboratorio).

Flujo de análisis	Alineador	Llamado de variantes	Filtrado	Tiempo total	Costo (US$)	Instancia utilizada
Inicial (11.0(a))	BWA-MEM (1h 40m)	Haplotype Caller (1h 40min)	Filter Variant Tranches (44min)	4h 5min	2,55	c4.2xlarge (1024GB)
Inicial modificado (11.0(b))	BWA-MEM (1h 40m)	Haplotype Caller (1h 40min)	Hard Filtering (3min)	3h 25min	1,82	c4.2xlarge (1024GB)
BWA-MEM GRAF (11.0(e))	BWA-MEM (1h 40m)	GRAF (3min)	GRAF (22seg)	1h 45min	1,09	c4.2xlarge (1024GB) c5.18xlarge (700GB)
GRAF (11.0(c))	GRAF (15min)	GRAF (3min)	GRAF (22seg)	20min	0,22	c5.18xlarge (700GB)
GRAF modificado (11.0(d))	GRAF (15min)	Haplotype Caller (1h 40min)	Filter Variant Tranches (44min)	2h 40min	2,55	c4.2xlarge (1024GB) c5.18xlarge (700GB)

Cuadro 11.2: *Promedio del costo en dólares y tiempo demorado en ejecutar los distintos flujos de análisis diseñados de la figura 11.0 para las muestras NA12878, NA24385 y NA24631 al generar el análisis de exoma completo. En la última columna se indican las instancias utilizadas ya que estas impactan directamente en el costo y tiempo de cada uno de los flujos.*

11.1.3. Optimización

Análisis de los alineadores

Se obtuvieron la sensibilidad, precisión y valor F1 promedio para todos los flujos de análisis que utilicen el mismo alineador para las variantes WES sin filtrar para evitar sesgos debido a que los resultados obtenidos con variantes filtradas son muy dependientes del filtrado de variantes (ver figura 11.2).

Se observa un gran contraste en los colores del mapa de calor que muestran las métricas promedio para los distintos alineadores, presentando un color más oscuro y por tanto un mayor valor en Precisión, sensibilidad y valor F1 para los flujos de análisis que utilizan el alineador GRAF. Sin embargo, si se observan los números en los mapas de calor, vemos que hay una diferencia menor al 1% en las métricas obtenidas entre alineadores.

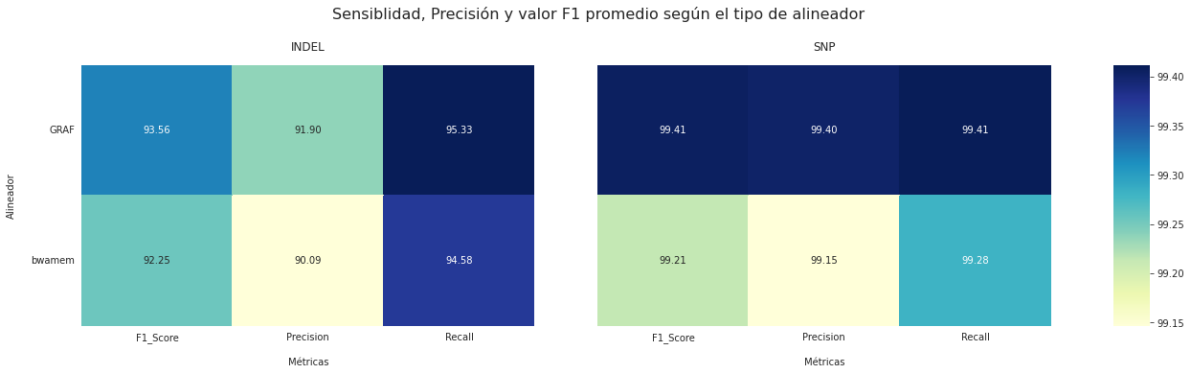


Figura 11.2: *Sensibilidad, Precisión y Valor F1 promedio para muestras WES para el alineador GRAF y bwamem.*

Análisis del llamado de variantes

Para obtener la herramienta de llamado de variantes que resulte mejor, se obtuvieron las métricas promedio de las variantes WES sin filtrar para todos los flujos de análisis que utilizan el mismo llamado de variantes en la figura 11.3. Al igual que en el análisis de alineadores, se utilizaron únicamente las variantes sin filtrar debido a que, como se mensionó anteriormente, la diferencia en resultados para variantes filtradas es muy dependiente del filtrado de variantes.

Al igual que en el caso de los alineadores, se obtuvieron colores más oscuros y, por lo tanto, mayores valores en métricas tanto para INDELs como SNPs, para los flujos de análisis que utilizaron GRAF para el llamado de variantes. Para las SNPs, la diferencia entre las métricas para los alineadores es menor al 1%, mientras que para las INDELs, se obtuvo una diferencia aproximada del 3% en el valor F1 para los distintos alineadores.

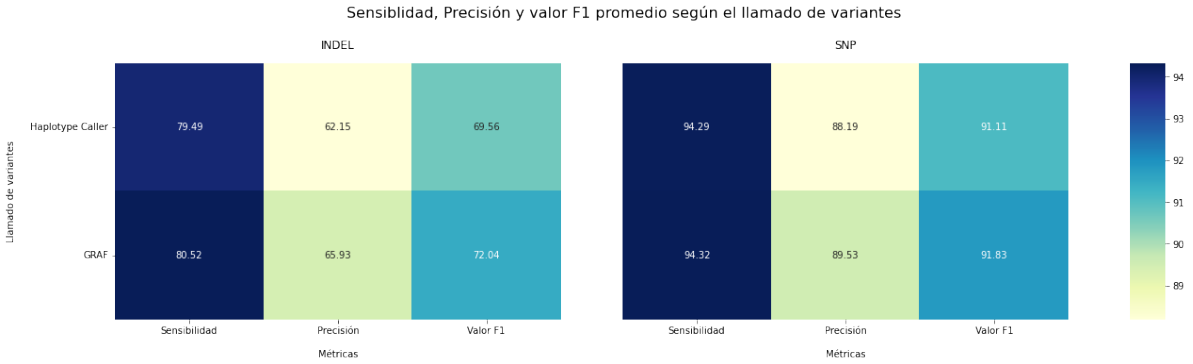


Figura 11.3: *Sensibilidad, Precisión y Valor F1 promedio para muestras WES para los llamados de variantes con Haplotype Caller y GRAF.*

Análisis del filtrado de variantes

Para elegir el filtrado óptimo, se tomo el promedio de las métricas obtenidas para todas las variantes SNP e INDEL. Promediando los resultados de sensibilidad, precisión y valor F1 obtenidos para todos los flujos de análisis que utilizan el mismo filtrado de variantes exómicas, se obtiene el mapa de calor en la figura 11.4.

Para los flujos de análisis que utilizaron GRAF para el filtrado de variantes, se obtuvieron los mayores valores de métricas, alcanzando una diferencia aproximada en el valor F1 del 3% para INDELs filtradas y de hasta el 12% para SNPs filtradas comparado a otros tipos de filtrados. Particularmente, se obtuvo el peor valor de F1 para las SNPs filtradas con las herramientas de *CNN score variants* y *Filter Variant Tranches*.

Además, se obtuvo mucho contraste entre los colores y por lo tanto, resultados en las métricas entre INDELs y SNPs, alcanzando una diferencia de casi el 15%.

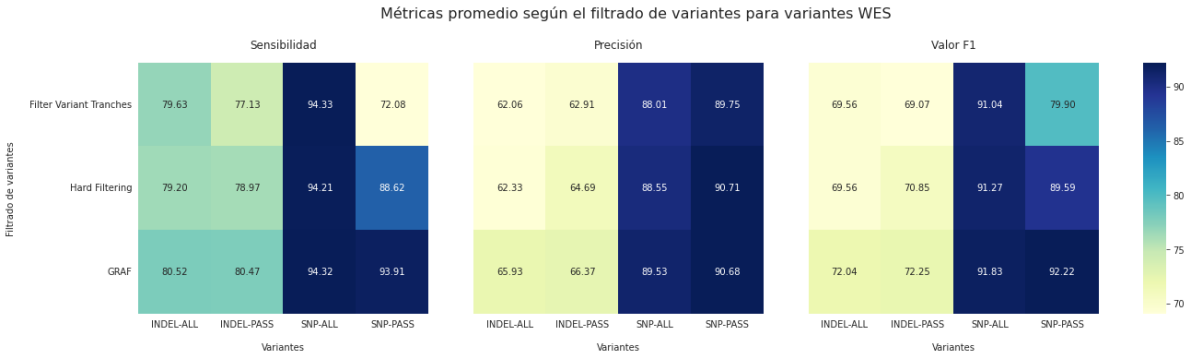


Figura 11.4: *Sensibilidad, Precisión y Valor F1 promedio para muestras WES para los filtrados de variantes con CNN Filter Variant Score en conjunto con Filter Variant Tranches, Hard Filtering y GRAF.*

11.2. Análisis de Genoma completo

Al igual que con las variantes germinales WES, se analizaron los valores de sensibilidad, precisión y el tiempo de ejecución para cada uno de los flujos y/o partes de los flujos de la figura

11.0 para lograr encontrar el análisis óptimo para las variantes WGS.

11.2.1. Análisis de sensibilidad y Precisión

En la tabla 11.3, se calculó el promedio de la sensibilidad, precisión y valor F1 para las variantes SNP e INDEL filtradas y no filtradas con Hap.py para la muestra NA12878 de genoma completo.

Para la muestra de WGS, se obtuvieron los valores más altos de F1 para el flujo de análisis que utiliza las herramientas de GRAF para el alineamiento, llamado y filtrado de variantes, alcanzando un valor del 99,51% para las SNPs filtradas y un 96,98% para las INDELs filtradas. En este caso, la mayor diferencia en el valor F1 se encuentra para las INDELs, con una diferencia aproximada que alcanza un valor del 7% para las INDELs obtenidas con el flujo de análisis inicial en comparación con aquellas obtenidas con el flujo de análisis de GRAF.

Al igual que en el caso de las muestras de WES, también se observa una diferencia entre los resultados para INDELs en comparación con las SNPs, alcanzando una diferencia aproximada máxima del 9% para los flujos de análisis que utilizan *Haplotype Caller* para el llamado de variantes y del 3% para los flujos de análisis que utilizan GRAF para el llamado de variantes.

Alineador	Llamado de variantes	Filtrado	Flujo de análisis	Tipo de variantes	Filtradas	Sensibilidad	Precisión	Valor F1
BWA-MEM	Haplotype Caller	Filter Variant Tranches	Inicial (11.0(a))	INDEL	No	93,43%	86,73%	89,95%
				INDEL	Sí	93,32%	87,28%	90,20%
				SNP	No	99,29%	98,98%	99,13%
				SNP	Sí	99,28%	99,18%	99,23%
		Hard Filtering	Inicial modificado (11.0(b))	INDEL	No	93,23%	86,72%	89,86%
				INDEL	Sí	93,21%	89,55%	91,34%
				SNP	No	99,29%	98,98%	99,13%
				SNP	Sí	98,13%	99,54%	98,82%
	GRAF	GRAF	BWA-MEM GRAF (11.0(e))	INDEL	No	97,06%	96,82%	96,95%
				INDEL	Sí	97,05%	96,87%	96,96%
				SNP	No	99,27%	99,47%	99,37%
				SNP	Sí	99,21%	99,62%	99,41%
GRAF	GRAF	GRAF	GRAF (11.0(c))	INDEL	No	97,14%	96,78%	96,96%
				INDEL	Sí	97,14%	96,82%	96,98%
				SNP	No	99,42%	99,50%	99,46%
				SNP	Sí	99,38%	99,63%	99,51%
	Haplotype Caller	Filter Variant Tranches	GRAF modificado (11.0(d))	INDEL	No	93,52%	87,02%	90,15%
				INDEL	Sí	93,43%	87,41%	90,32%
				SNP	No	99,41%	99,30%	99,35%
				SNP	Sí	99,40%	99,45%	99,43%

Cuadro 11.3: *Promedio de la sensibilidad, precisión y el valor F1 obtenidos a partir de Hap.py para la muestras NA12878 ejecutando los distintos flujos de análisis diseñados (ver figura 11.0) para análisis de genoma completo.*

Tomando el promedio de las métricas, se graficó un mapa de calor como el de la figura 11.5. Observamos mucho contraste en los mapas de color entre las variantes INDEL y SNP. Particularmente para las variantes INDEL, se observa la mayor diferencia entre flujos de análisis alcanzando casi un 97% de valor F1 en los flujos de análisis que utilizan a GRAF para el llamado y filtrado de variantes en comparación con un promedio del 90% para el resto de los flujos de análisis que utilizan *Haplotype Caller*. Para las variantes SNP, la diferencia entre flujos de análisis es menor al 1% alcanzando un valor F1 mayor al 99% para todos los casos, siendo el de mayor valor el flujo de análisis que utiliza todas las herrammientas de GRAF.

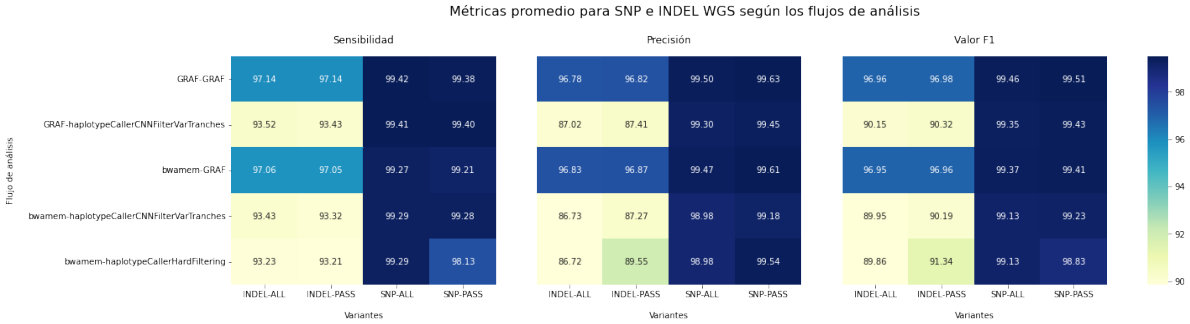


Figura 11.5: *Mapa de calor de la sensibilidad, Precisión y valor F1 promedio para cada uno de los flujos de análisis para SNP e INDELs WGS.*

11.2.2. Análisis de costo y tiempo

Luego de haber ejecutado todos los flujos de análisis de la figura 11.0 para la muestra de genoma completo NA12878, se calcularon el tiempo y costo promedios (ver tabla 11.2). Además, se le agregó el tipo de instancias utilizadas para correr cada uno de los flujos de análisis ya que esta interviene directamente en el costo y tiempo empleados.

El flujo de análisis que utiliza GRAF tanto para el alineamiento como para el llamado y filtrado de variantes, fue el que obtuvo un menor tiempo de ejecución (4h 5min) superando al resto de los flujos de análisis ampliamente, presentando un tiempo de seis veces menor con el segundo flujo de análisis que presentó menor tiempo de ejecución hasta más de 16 veces menor, en comparación a aquel que llevo mayor tiempo de ejecución (el flujo de análisis inicial del laboratorio)

Flujo de análisis	Alineador	Llamado de variantes	Filtrado	Tiempo total	Costo (US$)	Instancia utilizada
Inicial (11.0(a))	BWA-MEM (24h 36min)	Haplotype Caller (18h 10min)	Filter Variant Tranches (24h 30min)	67h 15min	33,01	c4.2xlarge (1024GB)
Inicial modificado (11.0(b))	BWA-MEM (24h 36min)	Haplotype Caller (18h 10min)	Hard Filtering (10min)	42h 55min	16,82	c4.2xlarge (1024GB)
BWA-MEM GRAF (11.0(e))	BWA-MEM (24h 36min)	GRAF (16min)	GRAF (2min)	24h 55min	4,54	c4.2xlarge (1024GB) c5.18xlarge (700GB)
GRAF (11.0(c))	GRAF (3h 45m)	GRAF (16min)	GRAF (2min)	4h 5min	10,75	c5.18xlarge (700GB)
GRAF modificado (11.0(d))	GRAF (3h 45m)	Haplotype Caller (18h 10min)	Filter Variant Tranches (24h 30min)	46h 25min	29,50	c4.2xlarge (1024GB) c5.18xlarge (700GB)

Cuadro 11.4: *Costo en dólares y tiempo demorado en ejecutar los distintos flujos de análisis diseñados de la figura 11.0 para la muestra WGS NA12878. En la última columna se indican las instancias utilizadas ya que estas impactan directamente en el costo y tiempo de cada uno de los flujos de análisis.*

11.2.3. Optimización

Análisis de los alineadores

Se obtuvieron la sensibilidad, precisión y valor F1 promedios para todos los flujos de análisis que utilizasen el mismo alineador para las variantes WGS sin filtrar (Ver la figura 11.6). A partir de los mapas de calor, no se nota una gran diferencia entre ambos alineadores debido a que presentan una diferencia en métricas menor al 1%. Sin embargo, para ambos todos los flujos de análisis, independientemente del alineador, observamos una diferencia mayor al 20% en el valor F1.

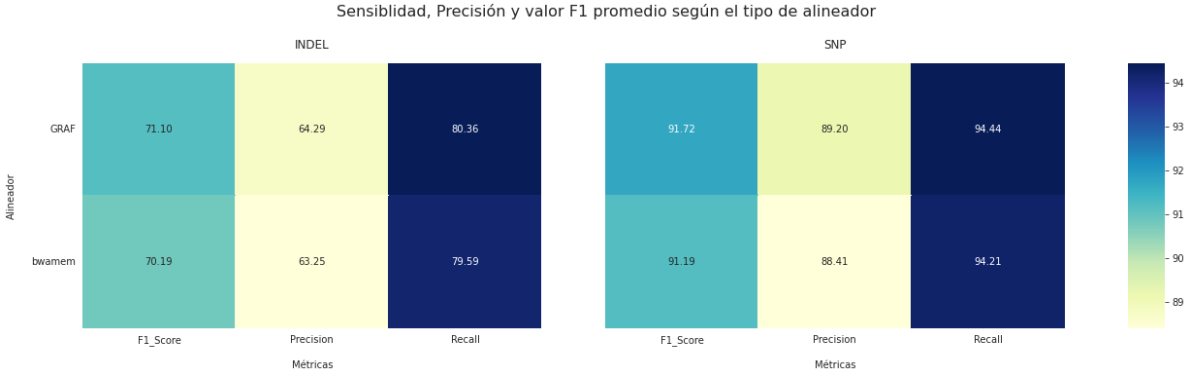


Figura 11.6: *Sensibilidad, Precisión y Valor F1 promedio para la muestra NA12878 WGS para el alineador GRAF y bwamem.*

Análisis del llamado de variantes

A momento de comparar las distintas herramientas de llamado de variantes, se obtuvieron las métricas promedio de las variantes WGS sin filtrar para todos los flujos de análisis que utilizasen la misma herramienta de llamado. Como se muestra en la figura 11.3, hay bastante contraste entre los flujos de análisis que utilizan *Haplotype Caller* para el llamado de variantes en comparación con aquellos que utilizan GRAF.

Particulamente para INDELs, el valor F1 promedio para los flujos de análisis que utilizan GRAF para el llamado de variantes es de un 96,95% mientras que para aquellos que utilizan *Haplotype Caller*, un casi 90%. Vemos una diferencia de casi el 7% para INDELs entre llamados de variantes. En el caso de las SNPs, la diferencia entre llamados de variantes es menor al 1%, siendo aquel que utiliza GRAF como llamado de variantes el de mayor valor F1, alcanzando un 99,42%.

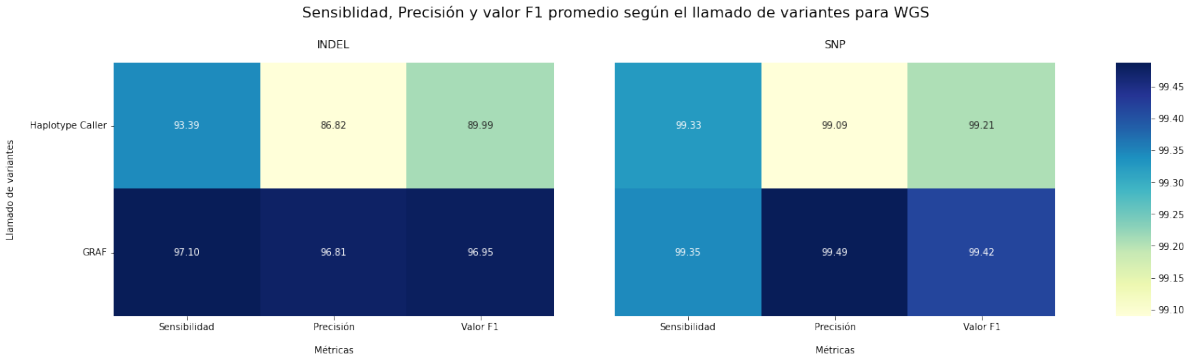


Figura 11.7: *Sensibilidad, precisión y valor F1 promedio para la muestra WGS para los llamados de variantes con Haplotype Caller y GRAF.*

Análisis del filtrado de variantes

Se analizaron las distintas herramientas de filtrado para la variantes WGS, promediando los resultados de sensibilidad, precisión y valor F1 obtenidos para todos los flujos de análisis que utilizasen el mismo filtrado de variantes WGS. De esta manera, se obtiene el mapa de calor de la figura 11.8.

No observamos grandes diferencias entre los distintos tipos de filtrados. En todos los filtros se observa una disminución de sensibilidad menor o cercana al 1% pero con un aumento también cercano al 1% a excepción del flujo de análisis que utilicó *Hard Filtering* en que se observo un aumento aproximado del 3%.

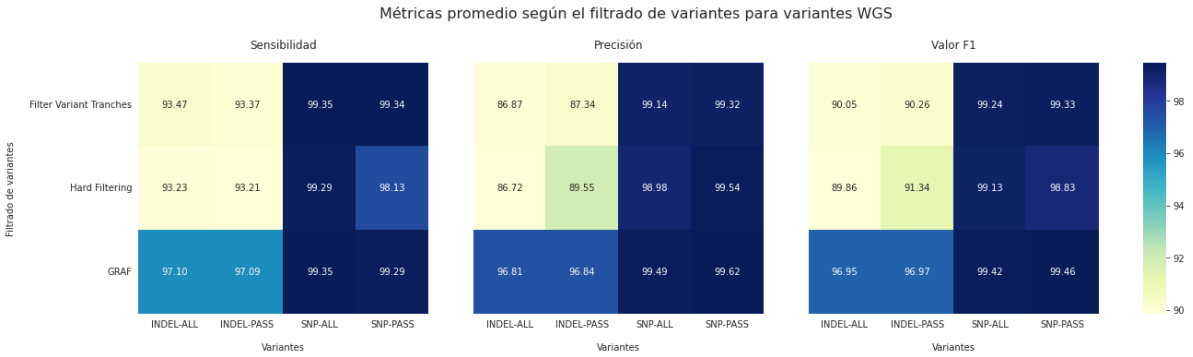


Figura 11.8: *Sensibilidad, Precisión y Valor F1 promedio para la muestra WGS para los filtrados de variantes con CNN Filter Variant Score en conjunto con Filter Variant Tranches, Hard Filtering y GRAF.*

11.3. Diseño final elegido del flujo de análisis germinal

Al contemplar las métricas resultantes para las variantes SNP e INDEL de los distintos flujos de análisis germinal, se diseñó un último flujo utilizando la combinación de herramientas en las que se observaron la mayor sensibilidad y precisión. A este, se le agregaron dos pasos de control de calidad: Fastqc para controlar la calidad del archivo de entrada FASTQ y Bamqc para el archivo alineado BAM (ver imagen 11.9).

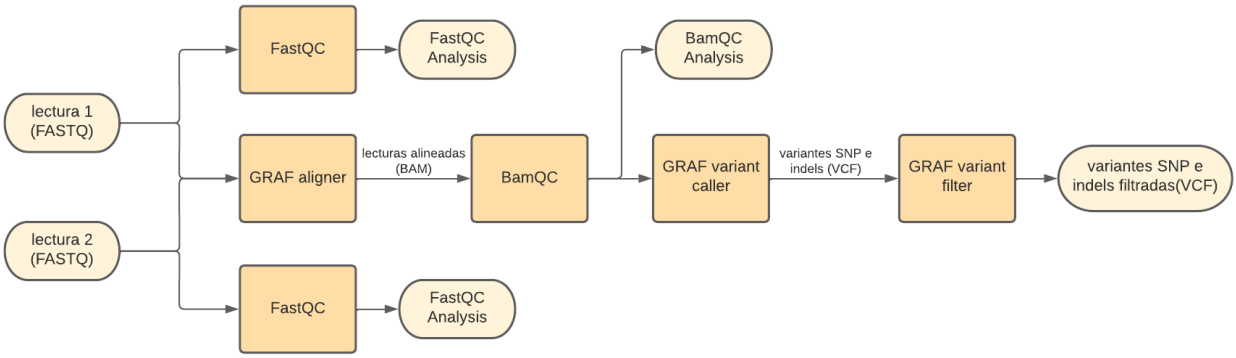


Figura 11.9: *Flujo de análisis germinal diseñado a partir del que tuvo mejor sensibilidad y precisión para las variantes SNP e INDEL obtenidas de WES.*

11.3.1. Archivo de salida resultante

A partir del flujo de análisis 11.9 se obtiene un archivo VCF con las variantes germinales. En la figura 11.10 se muestra una sección del archivo VCF final obtenido, donde podemos ver la posición de las variantes encontradas (cromosoma, y posición dentro del mismo), el ID de la variante (o RSID, si existe), el nucleótido de referencia, el alternativo, las puntuaciones de calidad, si la variante pasó o no el filtrado, en *INFO*, *FORMAT* y *SAMPLE* se informa sobre la profundidad de lectura, otras puntuaciones de calidad, frecuencias poblacionales para diferentes bases de datos, entre otras. Para mayor información sobre las características obtenidas en los VCF ver sección C.1 en el apéndice.

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	SAMPLE
chr1	65797	.	T	C	11.18	PASS	DP=2;AD=0,2	GT:PL:GQ:AD:DP:IL:URC	1/1:52,5,0:5:0,2:2:0,368:2
chr1	65872	.	T	G	87.4	PASS	DP=5;AD=0,5	GT:PL:GQ:AD:DP:IL:URC	1/1:130,14,0:14:0,5:5:0,346:5
chr1	817514	.	T	C	341.78	PASS	BaseQRankS	GT:PL:GQ:AD:DP:IL:URC	1/1:363,8,0:8:3,13:16:121,287:13
chr1	826893	.	G	A	7203.66	PASS	DP=265;AD=(	GT:PL:GQ:AD:DP:IL:URC	1/1:7226,782,0:99:0,265:265:0,208:262
chr1	827209	.	G	C	681.88	PASS	DP=21;AD=0,	GT:PL:GQ:AD:DP:IL:URC	1/1:704,62,0:62:0,21:21:0,234:21
chr1	827212	.	C	G	681.88	PASS	DP=19;AD=0,	GT:PL:GQ:AD:DP:IL:URC	1/1:704,62,0:62:0,19:19:0,234:19
chr1	827221	.	T	C	577.89	PASS	DP=17;AD=0,	GT:PL:GQ:AD:DP:IL:URC	1/1:600,50,0:50:0,17:17:0,234:17
chr1	856883	.	A	G	2887.37	PASS	DP=103;AD=(	GT:PL:GQ:AD:DP:IL:URC	1/1:2910,303,0:99:0,103:103:0,218:103
chr1	857100	.	C	T	6345.68	PASS	DP=232;AD=(	GT:PL:GQ:AD:DP:IL:URC	1/1:6368,679,0:99:0,232:232:0,196:228
chr1	930939	.	G	A	1335.72	PASS	DP=50;AD=0,	GT:PL:GQ:AD:DP:IL:URC	1/1:1358,144,0:99:0,50:50:0,174:49
chr1	931131	.	C	CCCCT	1017.39	PASS	DP=40;AD=0,	GT:PL:GQ:AD:DP:IL:URC	1/1:1040,96,0:96:0,40:40:0,171:37
chr1	941119	.	A	G	610.17	PASS	DP=23;AD=0,	GT:PL:GQ:AD:DP:IL:URC	1/1:633,68,0:68:0,23:23:0,209:22
chr1	942335	.	C	G	416.4	PASS	DP=16;AD=0,	GT:PL:GQ:AD:DP:IL:URC	1/1:439,47,0:47:0,16:16:0,173:16
chr1	942451	.	T	C	648.84	PASS	DP=29;AD=0,	GT:PL:GQ:AD:DP:IL:URC	1/1:671,77,0:77:0,29:29:0,141:26
chr1	942934	.	G	C	1530.02	PASS	BaseQRankS	GT:PL:GQ:AD:DP:IL:URC	0/1:1549,0,2240:99:109,91:200:173,156:86
chr1	944858	.	A	G	6571.02	PASS	BaseQRankS	GT:PL:GQ:AD:DP:IL:URC	1/1:6593,736,0:99:1,255:256:170,175:242
chr1	946247	.	G	A	2976.84	PASS	DP=111;AD=(	GT:PL:GQ:AD:DP:IL:URC	1/1:2999,320,0:99:0,111:111:0,174:107
chr1	948245	.	A	G	2344.52	PASS	DP=92;AD=0,	GT:PL:GQ:AD:DP:IL:URC	1/1:2367,274,0:99:0,92:92:0,174:89

Figura 11.10: *Sección del archivo CSV obtenido a partir del flujo de análisis de la figura 11.9 con información acerca de las variantes germinales.*

CAPÍTULO 12

DISEÑO DEL FLUJO DE ANÁLISIS SOMÁTICO

Para la primera etapa de alineamiento del FASTQ de variantes somáticas, se utilizaron muchas herramientas iguales que en el flujo de análisis germinal ya que hasta después del alineamiento, el origen de la muestra no altera el flujo bioinformático. En segundo lugar, para el llamado y filtrado de variantes somáticas, se eligió utilizar las herramientas Mutect2 [91] y Filter Mutect2 [92] respectivamente, después de realizar una revisión bibliográfica de las herramientas disponibles.

Para el llamado de variantes con Mutect2 se eligió utilizar el modo tumor-normal que recibe como entrada el BAM con las variantes de una muestra de tejido germinal obtenido del flujo de análisis que se muestra en la figura 11.9 y la compara con el de la muestra somática para obtener solamente las variantes somáticas.

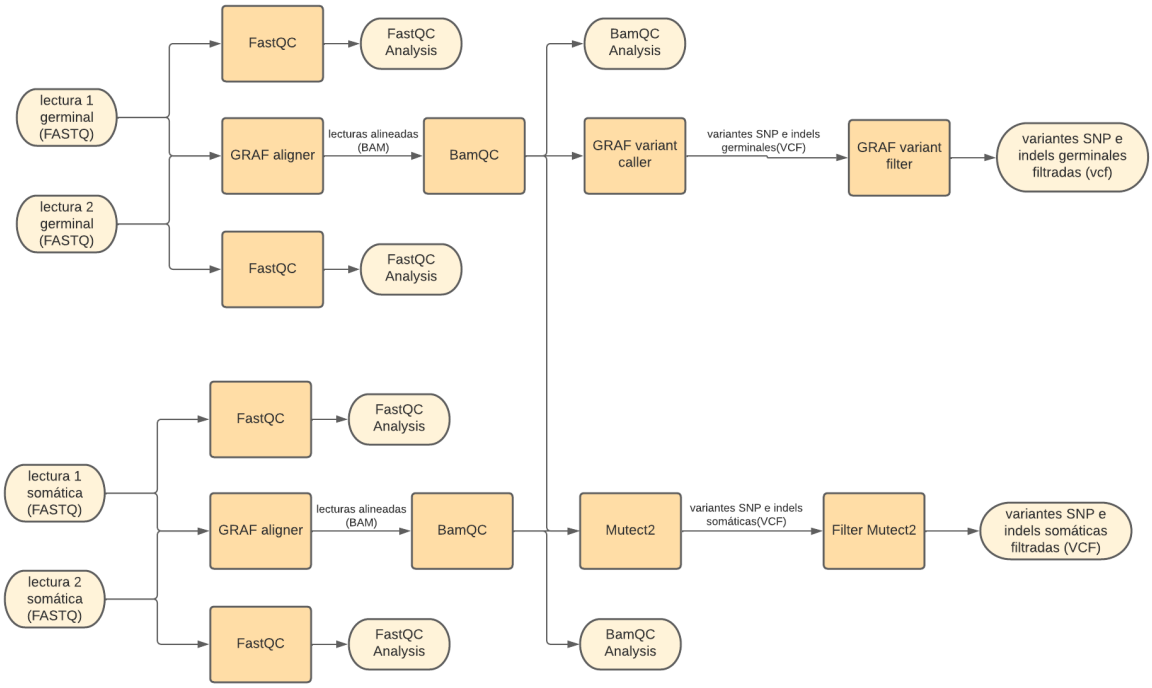


Figura 12.1: *Flujo de análisis para las variantes somáticas y germinales.*

12.1. Archivo de salida resultante

A partir del flujo de análisis 12.1 se obtiene un archivo VCF somático. En la figura 12.2 se muestra una sección del archivo VCF final obtenido, donde podemos ver la posición de las variantes encontradas (cromosoma, y posición dentro del mismo), el ID de la variante (si existe), el nucleótido de referencia, el alternativo, las puntuaciones de calidad, si la variante pasó o no el filtrado, en *INFO*, *FORMAT*, *HCC1143* y *HCC1143B* se informa sobre la profundidad de lectura, otras puntuaciones de calidad, frecuencias poblacionales para diferentes bases de datos, entre otras.

Se puede notar que, a diferencia de 11.10, este archivo CSV se diferencia en el campo posterior a la columna *FORMAT*; donde vemos que en el caso germinal, solo figura una columna posterior con información de la muestra (con el nombre de *SAMPLE*) mientras que en el somático, hay dos columnas posteriores (nombradas *HCC1143* y *HCC1143B*) que presentan la información resultante de ambos tipos de análisis; es decir, la primera corresponde a los resultados del análisis germinal, y la segunda al somático, a fines de poder hacer una comparación mas informativa.

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	HCC1143	HCC1143B
1	861433	.	C	T	.	weak_evider	CONTQ=93;D	GT:AD:AF:DP	0/0:37,0:0.02	0/1:28,2:0.093:30:15,2:13,0:0,28,0,2
1	880095	.	C	A	.	PASS	CONTQ=93;D	GT:AD:AF:DP	0/0:30,0:0.03	0/1:15,2:0.158:17:8,0:7,2:10,5,2,0
1	894248	.	C	A	.	PASS	CONTQ=93;D	GT:AD:AF:DP	0/0:32,0:0.02	0/1:23,2:0.111:25:12,0:11,2:16,7,2,0
1	897449	.	C	T	.	PASS	CONTQ=93;D	GT:AD:AF:DP	0/0:16,0:0.05	0/1:11,2:0.199:13:5,2:6,0:2,9,0,2
1	907757	.	G	T	.	weak_evider	CONTQ=93;D	GT:AD:AF:DP	0/0:10,0:0.07	0/1:18,3:0.156:21:8,2:9,0:9,9,0,3
1	980614	.	C	A	.	PASS	CONTQ=93;D	GT:AD:AF:DP	0/0:44,0:0.02	0/1:26,2:0.100:28:11,0:15,2:8,18,2,0
1	989981	.	T	C	.	weak_evider	CONTQ=93;D	GT:AD:AF:DP	0/0:10,0:0.08	0/1:1,1:0.500:2:0,1:1,0:0,1,0,1
1	1102472	.	C	A	.	PASS	CONTQ=93;D	GT:AD:AF:DP	0/0:20,0:0.04	0/1:17,2:0.145:19:4,0:10,2:13,4,2,0
1	1118411	.	G	T	.	weak_evider	CONTQ=93;D	GT:AD:AF:DP	0/0:18,0:0.04	0/1:20,2:0.123:22:8,2:12,0:4,16,0,2
1	1119553	.	G	T	.	PASS	CONTQ=93;D	GT:AD:AF:DP	0/0:46,0:0.02	0/1:47,3:0.077:50:22,1:24,2:0,47,0,3
1	1139744	.	C	A	.	weak_evider	CONTQ=93;D	GT:AD:AF:DP	0/0:19,0:0.04	0/1:24,2:0.106:26:12,0:12,2:20,4,2,0
1	1225718	.	C	A	.	PASS	CONTQ=93;D	GT:AD:AF:DP	0/0:64,0:0.01	0/1:67,3:0.055:70:32,0:35,3:43,24,3,0
1	1226883	.	C	T	.	PASS	CONTQ=93;D	GT:AD:AF:DP	0/0:15,0:0.05	0/1:13,3:0.219:16:4,0:9,3:6,7,3,0
1	1230794	.	C	A	.	PASS	CONTQ=93;D	GT:AD:AF:DP	0/0:36,0:0.02	0/1:30,2:0.088:32:15,0:14,2:26,4,2,0
1	1235964	.	C	A	.	PASS	CONTQ=93;D	GT:AD:AF:DP	0/0:55,0:0.01	0/1:16,2:0.150:18:5,0:11,2:12,4,2,0
1	1247157	.	G	T	.	PASS	CONTQ=93;D	GT:AD:AF:DP	0/0:15,0:0.05	0/1:15,2:0.155:17:6,2:9,0:13,2,0,2

Figura 12.2: *Sección del archivo CSV obtenido a partir del flujo de análisis de la figura 11.9 con información acerca de las variantes germinales.*

CAPÍTULO 13

ANOTACIÓN DE VARIANTES

Luego de diseñar el flujo de análisis que se muestra en la figura 12.1 y realizar una búsqueda bibliográfica de las herramientas disponibles para la anotación de variantes, se agregaron los pasos de InterVar y CancerVar, capaces de incluir la clasificación ACMG y AMP a las variantes germinales y somáticas respectivamente.

Debido a que ambas herramientas devolvían como resultado tres archivos TXT, se programó un script en Python para unificarlos y generar un único archivo (CSV) resultante tanto para CancerVar como para InterVar. Cada uno de estos archivos contiene la información genómica, clínica, poblacional, fisiopatológica, entre otras, de manera unificada para las variantes, además de la clasificación ACMG o AMP según el análisis ejecutado.

Estos scripts fueron añadidos al flujo de análisis, lo cual permitió no solo facilitar en su totalidad la información relevante sino que lo hace mediante un archivo mas amigable para el profesional interesado (ver figura 13.1).

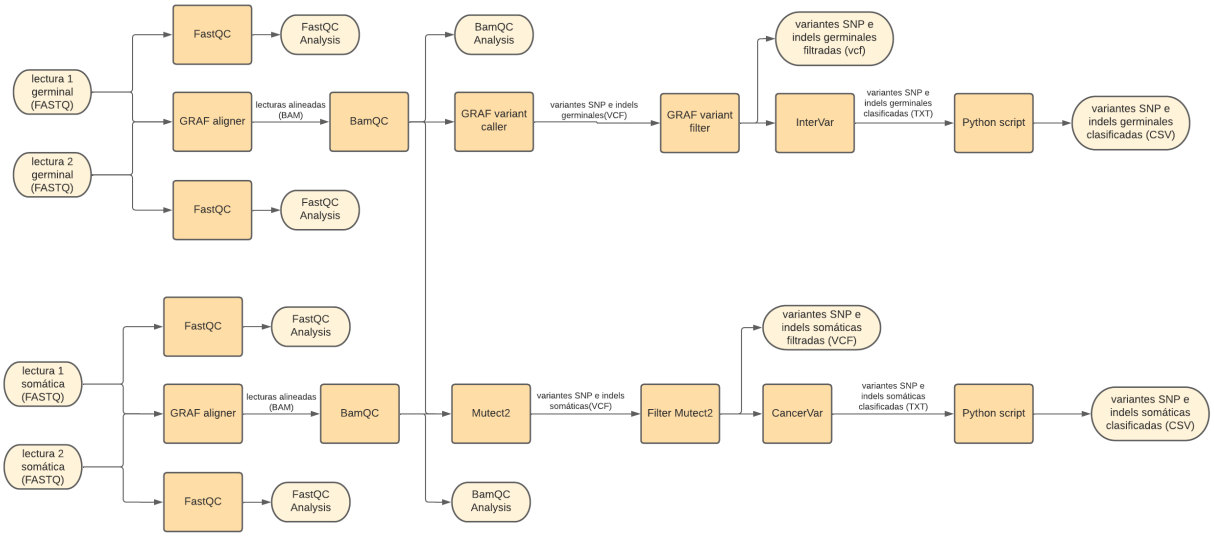


Figura 13.1: *Flujo de análisis para las variantes somáticas y germinales con los pasos de anotación de variantes y scripts incluidos.*

En la figura 13.1(a) se ve un ejemplo del CSV generado luego de este paso (columnas seleccionadas) y en 13.1(b) su equivalente para somático.

Chromosome	Start	End	Ref	Alt	Gene.refGene	Func.refGene	ExonicFunc.refGene	Gene.ensGene	avsnp147	AAChange.ensGene	AAChange.refGene	clinvar: Clinvar	InterVar: InterVar and Evidence	Freq_gnomAD_genome_ALL
chr1	65797	65797	T	C	FAM138A	intergenic	.	OR4F5	rs759525926	.	.	clinvar: UNK	InterVar: Benign PVS1=0 PS=[0, 0, 0,	0.1935
chr1	65797	65797	T	C	OR4F5	intergenic	.	OR4F5	rs759525926	.	.	clinvar: UNK	InterVar: Benign PVS1=0 PS=[0, 0, 0,	0.1935
chr1	65872	65872	T	G	FAM138A	intergenic	.	OR4F5	rs796239852	.	.	clinvar: UNK	InterVar: Benign PVS1=0 PS=[0, 0, 0,	0.1785
chr1	65872	65872	T	G	OR4F5	intergenic	.	OR4F5	rs796239852	.	.	clinvar: UNK	InterVar: Benign PVS1=0 PS=[0, 0, 0,	0.1785
chr1	817514	817514	T	C	FAM87B	ncRNA_exonic	.	FAM87B	rs3131971	.	.	clinvar: UNK	InterVar: Benign PVS1=0 PS=[0, 0, 0,	0.7294
chr1	826577	826577	-	T	LINC00115	ncRNA_exonic	.	LINC00115	rs59038458	.	.	clinvar: UNK	InterVar: Benign PVS1=0 PS=[0, 0, 0,	0.6940
chr1	826893	826893	G	A	LINC00115	ncRNA_exonic	.	LINC00115	rs3115849	.	.	clinvar: UNK	InterVar: Benign PVS1=0 PS=[0, 0, 0,	0.7256
chr1	827209	827209	G	C	LINC00115	ncRNA_exonic	.	LINC00115	rs3115848	.	.	clinvar: UNK	InterVar: Benign PVS1=0 PS=[0, 0, 0,	0.7305
chr1	827212	827212	C	G	LINC00115	ncRNA_exonic	.	LINC00115	rs3131950	.	.	clinvar: UNK	InterVar: Benign PVS1=0 PS=[0, 0, 0,	0.7317
chr1	827221	827221	T	C	LINC00115	ncRNA_exonic	.	LINC00115	rs3131949	.	.	clinvar: UNK	InterVar: Benign PVS1=0 PS=[0, 0, 0,	0.7335
chr1	827252	827252	T	A	LINC00115	ncRNA_exonic	.	LINC00115	rs3131948	.	.	clinvar: UNK	InterVar: Benign PVS1=0 PS=[0, 0, 0,	0.7445
chr1	841742	841742	A	T	LINC01128	ncRNA_exonic	.	LINC01128	rs2980319	.	.	clinvar: UNK	InterVar: Benign PVS1=0 PS=[0, 0, 0,	0.7383
chr1	856883	856883	A	G	LINC01128	ncRNA_exonic	.	LINC01128	rs1044922	.	.	clinvar: UNK	InterVar: Benign PVS1=0 PS=[0, 0, 0,	0.8409
chr1	857100	857100	C	T	LINC01128	ncRNA_exonic	.	LINC01128	rs2905036	.	.	clinvar: UNK	InterVar: Benign PVS1=0 PS=[0, 0, 0,	0.9749
chr1	926250	926250	G	A	SAMD11	intronic	.	SAMD11	rs2879816	.	.	clinvar: UNK	InterVar: Benign PVS1=0 PS=[0, 0, 0,	0.6144
chr1	930939	930939	G	A	SAMD11	intronic	.	SAMD11	rs9988021	.	.	clinvar: UNK	InterVar: Benign PVS1=0 PS=[0, 0, 0,	0.9616
chr1	931131	931131	-	CCCT	SAMD11	intronic	.	SAMD11	rs375757231	.	.	clinvar: UNK	InterVar: Benign PVS1=0 PS=[0, 0, 0,	0.5449
chr1	935954	935954	G	T	SAMD11	intronic	.	SAMD11	rs4072383	.	.	clinvar: UNK	InterVar: Benign PVS1=0 PS=[0, 0, 0,	0.5628
chr1	941119	941119	A	G	SAMD11	intronic	.	SAMD11	rs4372192	.	.	clinvar: UNK	InterVar: Benign PVS1=0 PS=[0, 0, 0,	0.9284
chr1	942335	942335	C	G	SAMD11	intronic	.	SAMD11	rs6605066	.	.	clinvar: UNK	InterVar: Benign PVS1=0 PS=[0, 0, 0,	0.8942

((a)) *Sección del archivo CSV obtenido a partir del flujo de análisis de la figura 13.1 con información acerca de las variantes germinales. En la imagen podemos ver la posición de las variantes encontradas (cromosoma, posición inicial y final), el nucleótido de referencia, el alternativo, el gen en la cual se encontró, el tipo de variante, su función exónica (en caso de conocerse), el ID de RS de mutación, el cambio de aminoácido con nomenclatura HGVS (en caso de conocerse), el veredicto de ClinVar, el veredicto de InterVar y la frecuencia alélica de la población general informada por gnomAD.*

Chr	Start	End	Ref	Alt	Region	Variant type	HGVS (AAChange)	RS ID	Gene symbol	AMP Veredict	Ensembl transcript id	Clinvar	ExAC exome frequencies of all variants
1	69428	69428	T	G	exonic	nonsynonymous SNV	OR4F5:NM_001005484:exon1:c.T338G:p.F113C	rs140739101	OR4F5	1#Tier_IV_benign	ENST00000641515.2_4	UNK	0.0246
1	69453	69453	G	A	exonic	synonymous SNV	OR4F5:NM_001005484:exon1:c.G363A:p.K121K	rs2854682	OR4F5	1#Tier_IV_benign	ENST00000641515.2_4	UNK	8.639e-05
1	989981	989981	T	C	intronic	.	.	rs114308080	AGRN	0#Tier_IV_benign	['.']	UNK	0.0022
1	1423267	1423267	A	G	exonic	synonymous SNV	ATAD3B:NM_001317238:exon10:c.A1101G:p.A367A,	rs819977	ATAD3B	2#Tier_IV_benign	ENST00000673477.1_2	UNK	0.2539
1	1848109	1848109	G	C	intronic	.	.	rs2803296	CALML6	3#Tier_III_Uncertain	['.']	UNK	.
1	1848121	1848121	G	C	intronic	.	.	rs28472657	CALML6	1#Tier_IV_benign	['.']	UNK	.
1	1890704	1890704	-	T	intronic	.	.	.	CFAP74	1#Tier_IV_benign	['.']	UNK	.
1	2538516	2538516	G	A	intronic	.	.	.	MMEL1	1#Tier_IV_benign	['.']	UNK	.
1	3394526	3394526	G	T	exonic	nonsynonymous SNV	ARHGEF16:NM_014448:exon11:c.G1561T:p.A521S	.	ARHGEF16	0#Tier_IV_benign	ENST00000378371.6_3,ENST00	UNK	.
1	3527709	3527709	G	T	exonic	nonsynonymous SNV	MEGF6:NM_001409:exon1:c.C124A:p.P42T	.	MEGF6	0#Tier_IV_benign	ENST00000356575.9_5	UNK	.
1	3527717	3527717	G	T	exonic	nonsynonymous SNV	MEGF6:NM_001409:exon1:c.C116A:p.P39Q	.	MEGF6	0#Tier_IV_benign	ENST00000356575.9_5	UNK	.
1	6529183	6529185	TCC	-	exonic	nonframeshift deletion	PLEKHG5:NM_001042664:exon19:c.2166_2168del:p.	rs113541584	PLEKHG5	0#Tier_IV_benign	ENST00000675123.1_1,ENST00	Benign/Likely_benign	0.1130
1	6727725	6727725	G	T	intronic	.	.	.	DNAJC11	1#Tier_IV_benign	['.']	UNK	.
1	7796641	7796641	A	G	intronic	.	.	rs2071985	CAMTA1	3#Tier_III_Uncertain	['.']	UNK	0.5321
1	7804871	7804871	T	C	intronic	.	.	rs72853877	CAMTA1	1#Tier_IV_benign	['.']	UNK	0.0044

((b)) *Sección del archivo CSV obtenido a partir del flujo de análisis de la figura 13.1 con las columnas que contienen información acerca de las variantes somáticas. En la imagen podemos ver la posición de las variantes encontradas (cromosoma, posición inicial y final y región), el nucleótido de referencia, el alternativo, el tipo de variante, el cambio de aminoácido con nomenclatura HGVS, el ID de RS de mutación, el símbolo del gen, el veredicto de AMP, el ID del transcripto en Ensembl, el veredicto en ClinVar y la frecuencia exómica de las variantes en ExAC.*

Figura 13.1: *Archivos CSV de salida con algunas de las columnas que contienen información relevante acerca de las variantes germinales y somáticas identificadas (ver anexo sección C.1.1 para mayor detalle acerca de la información obtenida sobre las variantes en cada uno de los CSV).*

Con estas herramientas se obtuvo, entre otras:

- Localización de la variante en el genoma, calidad de la lectura, características propias del alineamiento, cigocidad, frecuencia alelica, entre otras.
- Datos acerca del gen en el que se encuentra la variante (identificador de distintas bases de datos, símbolo del gen, etc.)
- Información acerca si pasó los filtros de GRAF para las variantes germinales y Mutect2 para las somáticas.
- Tipo de variante y cambio generado en aminoácidos si corresponde
- Identificador de la variante de distintas bases de datos: rsID, snp142, entre otras.
- Clasificación de la variante según ACMG para las variantes germinales y también AMP para las somáticas. También se muestra la clasificación de cada uno de los criterios del veredicto final.

- Clasificación de la variante en ClinVar.
- Frecuencia de las variantes en bases de datos poblacionales: exac, gnomad, entre otras.
- Predicción funcional in sílico de las variantes segun los distintos predictores: SIFT, PolyPhen2 HDIV, PolyPhen2 HVAR, LRT, MutationTaster, MutationAssessor, FATHMM, MetaSVM, MetaLR, VEST, CADD, GERP++, DANN, fitCons, PhyloP, dbscsnv11, entre otras.
- Información sobre las variantes específicas de enfermedades de distintas bases de datos: ClinVar (mencionada anteriormente), COSMIC, icgc21, OMIM, entre otras.

En el anexo, sección C.1.1 se detalla la información de las columnas más relevantes de estos archivos.

En conclusión, se logró cumplir con un objetivo que, a nuestro leal saber, no se encuentra actualmente disponible: automatizar la anotación ACMG/AMP y la generación de los informes médicos, con herramientas de código libre.

CAPÍTULO 14

INFORME FINAL

Con el objetivo de enviar la información mas relevante al profesional de la salud que lo haya solicitado, se generaron dos informes, uno para los análisis somáticos y otro para germinales. Estos informes contienen de una manera resumida y simple de leer toda la información técnica esencial junto con la predicción de patogenicidad de las variantes halladas en el paciente.

Se informan las herramientas utilizadas; la tasa TS/TV [1](como parámetro de calidad general) y profundidad de lectura media[2] obtenida.

Para el armado del informe, se partió de los archivos resultantes CSV de las variantes germinales y somáticas de los flujos de análisis que se muestran en las figuras 11.9 y 12.1.

Estos informes siempre son enviados al profesional de la salud solicitante junto con dicho archivo CSV que contiene todas las variantes y mayor información de las mismas.

14.1. Informe germinal

Para elaborar el informe, se decidió mostrar las variantes probablemente patogénicas y patogénicas (ver figura 14.3)[3]. En específico, se muestra para cada una de estas variantes:

- Chr: El cromosoma en el cual la variante fue hallada
- Start: Inicio de la posición cromosómica donde fue hallada la variante
- End: Fin de la posición cromosómica donde fue hallada la variante

[1]Se espera una tasa ≈2 en un análisis WGS y ≈3 en un WES.

[2]Se espera una profundidad de lectura mayor que 30X.

[3]De todas maneras, se adjunta el archivo final CSV con la información de todas las variantes encontradas en caso de que el médico necesite analizarlas

- Ref: El nucleótido hallado en la referencia
- Alt: El nucleótido hallado en la muestra
- Profundidad: La profundidad de lectura para esa variante
- RSID: Identificador de dbSNP de la variante
- Gen: Nombre del gen en el cual se halló la variante
- Veredicto ACMG: En este caso se muestran solamente probablemente patogénicas y patogénicas

Además de la tabla, se informan la tasa TS/TV y la profundidad de lectura media del análisis.

Chr	Start	End	Ref	Alt	Profundidad	RSID	Gen	Veredicto ACMG
chr4	186274193	186274193	G	T	339.0	rs121965063	F11	Patogénica
chr3	10046723	10046726	AGTA	-	221.0	rs369823368	FANCD2	Posiblemente patogénica
chr5	147399167	147399167	C	-	114.0	.	DPYSL3	Posiblemente patogénica
chr7	55173087	55173087	G	A	109.0	rs150423237	EGFR	Posiblemente patogénica
chr11	47580858	47580858	G	C	146.0	rs771848158	NDUFS3	Posiblemente patogénica
chrX	136874452	136874452	-	CATAACT	60.0	.	RBMX	Posiblemente patogénica

Figura 14.1: *Datos germinales que se informarían en el caso del paciente NA24385*

Por otra parte, se incluyo un gráfico de torta con las proporciones de los distintos tipos de variantes halladas utilizando la clasificación de ACMG (ver figura 14.2).

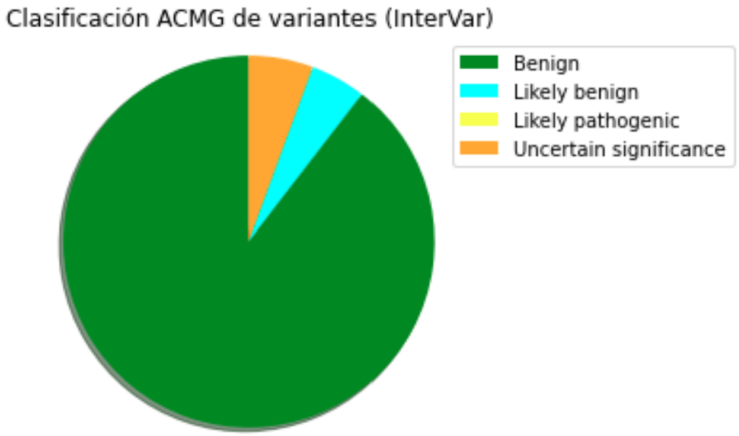


Figura 14.2: *Proporción de variantes germinales hallada en el paciente NA24385*

14.2. Informe somático

Para el informe de las variantes somáticas, se muestra la información de las variantes con fuerte y posible significancia (ver figura 14.3). Se muestra en el informe para cada una de las variantes:

- Chr: El cromosoma en el cual la variante fue hallada
- Start: Inicio de la posición cromosómica donde fue hallada la variante
- End: Fin de la posición cromosómica donde fue hallada la variante
- Ref: El nucleótido hallado en la referencia
- Alt: El nucleótido hallado en la muestra
- Profundidad: La profundidad de lectura para esa variante
- RSID: Identificador de dbsnp de la variante hallada
- Gen: Nombre del gen en el cual se halló la variante
- Veredicto AMP: En este caso se muestran solamente las variantes de posible y fuerte significancias.

Además de la tabla, se informa la profundidad de lectura media del análisis.

Chr	Start	End	Ref	Alt	Profundidad	RSID	Gen	Veredicto AMP
2	241723205	241723205	G	T	31	.	KIF1A	Posible importancia
17	42338084	42338084	C	T	132	rs28929480	SLC4A1	Fuerte importancia

Figura 14.3: *Datos somáticos que se informarían en el caso del paciente NA24385*

Se pueden observar ejemplos de informe germinal y somático en Drive

Parte IV

Discusión

CAPÍTULO 15

ELECCIÓN DEL FLUJO DE ANÁLISIS GERMINAL

Para analizar las variantes germinales se eligió el flujo de análisis que utiliza las herramientas de GRAF para el alineamiento, el llamado y filtrado de variantes ya que, a pesar de que no presentaba una diferencia significativa en sensibilidad, precisión o valor F1; fue el que tuvo mejores resultados tanto para muestras WES como para WGS. Además, era el más optimizado computacionalmente mostrando un tiempo de ejecución hasta 12 veces menor comparativamente con los otros flujos.

Si analizamos con mayor detenimiento los resultados de las métricas en la figura 11.1 y 11.5 para los distintos flujos de análisis, notamos que la mayor diferencia se encuentra en precisión y valor F1 para INDELs en los flujos que utilizan distintos llamados de variantes. Particularmente, aquellos que utilizan GRAF para el llamado de variantes son los que obtienen mayores valores de sensibilidad, precisión y valor F1. Esto es consistente con los resultados vistos en otras publicaciones científicas que sostienen que el llamado de variantes tiene más influencia en la detección de SNPs e INDELs que los alineadores. [96]

Para todos los pasos del flujo de análisis: alineamiento, llamado y filtrado de variantes, notamos una diferencia significativa en tiempo de ejecución para las herramientas GRAF (ver tablas 11.2 y 11.4), lo cual deriva en un impacto sustancial en el costo de cada análisis. [43]. Esto se debe a que GRAF está diseñado y optimizado para ser computacionalmente eficiente. [97] En este caso, la herramienta es capaz de independizarnos del tipo de instancia utilizada debido a que elije la más conveniente para el análisis a través de un cálculo aproximado de los requisitos de recursos; configurando aquella que sea más adecuada para el análisis a ejecutar. [98] Esta fue una de las principales razones para elegir GRAF para el flujo de análisis.

En cuanto al análisis de sensibilidad, precisión y valor F1, tanto para analizar el alineamiento como el llamado de variantes, se tomaron todas las variantes sin filtrar para evitar sesgos, ya que

los resultados obtenidos con variantes filtradas son muy dependientes del filtrado de variantes.

Si analizamos el alineamiento, observamos de las figuras 11.2 y 11.6 que tanto para WES como para WGS, no hay una diferencia significativa entre los alineadores. La mayor diferencia entre ambos alineadores es de aproximadamente un 1% para todas las métricas. Este pequeño porcentaje superior en sensibilidad, precisión y valor F1 se podría atribuir a la implementación de los genomas de referencia gráficos por parte de GRAF. Esto permite que en el alineamiento, se consideren haplotipos alternativos y variaciones estructurales que no están contempladas en el genoma de referencia lineal usual. [99] Esto genera que el alineamiento con GRAF, en comparación con el de BWA-MEM, presente altas tasas de mapeo incluso con INDELs largos. [43] En este caso en particular, debido a las muestras de GIAB utilizadas (que funcionan como referencia para pequeñas variantes y no contienen variantes estructurales de gran tamaño [47]), no observamos una diferencia significativa. [43]

Sin embargo, a pesar de que en términos de alineamiento GRAF no parece ser significativamente superior; como mencionamos anteriormente, el tiempo de ejecución es varias veces menor haciendo que los costos asociados también disminuyan (ver tablas 11.2 y 11.4). Esto generó que se elija a GRAF como el alineador a incorporar en el flujo de análisis final.

Respecto al llamado de variantes, observando las figuras 11.3 y 11.7, llegamos a la conclusión que aquellos flujos que utilizan GRAF para el llamado de variantes, obtienen mayores valores de sensibilidad, precisión y valor F1. Para las variantes SNP WES notamos una diferencia menor al 1%, mientras que para las INDEL WES, una diferencia de más del 3% en precisión y de casi un 3% para el valor F1. Para la muestra WGS, al igual que lo observado en las muestras WES, de la figura 11.7 vemos que la mayor diferencia entre métricas se encuentra para los INDELs ($\approx$4% en sensibilidad, $\approx$10% en precisión y $\approx$7% en valor F1). Por otra parte, para las variantes de tipo SNP, la diferencia entre los resultados de los diferentes llamados de variantes no es mayor al 1%.

Tanto para WES como para WGS, notamos que la diferencia en las métricas para las variantes SNP podría no responder a un motivo particular y ser más bien aleatoria. Sin embargo, para INDELs, el llamado de variantes de GRAF tiene mayor precisión que el de Haplotype Caller. Esto podría deberse a que la esta métrica es el porcentaje de casos positivos detectados y, al incorporar haplotipos alternativos y variaciones estructurales, disminuyen los falsos positivos, o variantes encontradas en la muestra que no están en la referencia. [43] Particularmente, creemos que el hecho de que GRAF utilice genomas gráficos, hace que presente una ventaja significativa para las variantes INDEL que son inherentemente más difíciles de detectar, debido a que incorpora estas variantes estructurales en el genoma de referencia. [43, 97]

Por otro lado, vemos que la diferencia en las métricas es consistentemente mayor en la muestra WGS que WES. Esto podría deberse a que la muestra WGS presenta una mayor cobertura horizontal. El rendimiento en el llamado de variantes podría deberse a regiones no incluidas en las muestras exónicas, como por ejemplo; las regiones UTR, no codificantes, [42] donde el uso del genoma gráfico permite detectar mayor cantidad de variantes. [43]

Por último, en cuanto al tiempo de ejecución del llamado de variantes, al igual que en el alineamiento, observamos de las tablas 11.2 y 11.4 que Haplotype Caller toma más de 30 veces el tiempo que GRAF para las muestras WES y la diferencia aumenta aún más para la muestra WGS. Como mencionamos anteriormente, esto podría deberse a que una de las prioridades durante el desarrollo de herramientas GRAF fue asegurar que los algoritmos sean computacionalmente más eficientes. [97]

En lo que refiere al filtrado, en la figura 11.4, vemos que utilizando el filtro Hard Filtering, la precisión mejora un ≈2% tanto para variantes SNP e INDEL de WES a costa de una disminución en sensibilidad de menos del 1% para INDEL y del ≈6% para SNP. En cambio, con CNN Score Variants + Filter Variant Tranches, se obtiene una mejora en la precisión menor del 1% tanto para INDEL como para SNP filtradas a costa de una disminución del ≈3% para las variantes INDEL y de un ≈20% para las variantes SNP. Teniendo en cuenta los efectos de ambos filtros en las distintas métricas, podemos concluir que el filtro de Hard Filtering resulta mejor para el análisis WES. Por otro lado, si analizamos los resultados obtenidos con el filtrado de GRAF vemos que las variantes filtradas SNP e INDEL presentan una mejora menor al 1% en precisión a costa de una disminución menor al 1% en sensibilidad para los dos tipos de variantes filtradas.

Por otra parte para la muestra WGS, como se observa en la figura 11.8, no encontramos gran diferencia entre los distintos tipos de filtrado. En todos los filtros, excepto el de Hard Filtering para los INDEL, se observa un aumento en precisión menor o cercano al 1% a costa de una disminución en sensibilidad menor o cercana al 1%. En el caso de los INDEL, utilizando Hard filtering se obtiene un aumento aproximado del 3%.

La diferencia entre CNN Score Variants Score + Filter Variant Tranches y Hard Filtering se podría deber a que no se recomienda utilizar el filtrado de variantes basados en modelos de aprendizaje automático (como CNN de GATK) en los datos WES, a menos que el usuario tenga un modelo interno pre-entrenado diseñado específicamente para ese tipo de datos. Esto se debe a que el aprendizaje automático resulta más beneficioso en análisis donde el perfil de cobertura es más uniforme (genoma completo), y; aunque se podría pensar que el mismo algoritmo se puede usar para WES, hay resultados que demuestran que son incompatibles y solo deben aplicarse a WGS. [42, 100]

Si priorizamos encontrar la mayor cantidad de variantes, es importante tener una disminución en la precisión para las variantes filtradas. Además, debido a la mayor eficiencia computacional [97], el tiempo de filtrado de GRAF resultó ampliamente menor que el resto (ver tabla 11.2).

Estas fueron las razones principales por las cuales se eligió GRAF para realizar el filtrado de variantes. Además, como se recomienda utilizar el filtrado de variantes en función de la herramienta que produzca el llamado; concluimos elegir usar el filtrado de GRAF en conjunto con su llamado de variantes. [97]

Esta pequeña diferencia en métricas entre los flujos de análisis son consistentes con lo encontrado en la revisión bibliográfica debido a que las herramientas elaboradas por GATK se

han convertido en estándar para la comunidad bioinformática en el análisis de datos NGS y son ampliamente utilizadas en laboratorios de diagnóstico genómico. [101–104]

Si comparamos estos resultados con lo obtenido en otras publicaciones, se menciona que los flujos de análisis que utilizan genomas de referencia gráficos resultan mejores que aquellos que utilizan los lineales. [105] El genoma de referencia gráfico tiene en cuenta una mayor cantidad de variantes para las distintas poblaciones capturando la diversidad genética poblacional, lo que podría explicar los valores de precisión obtenidos[1]. [43, 106]

En [105] también, mencionan específicamente a GRAF como uno de los algoritmos de mejor rendimiento junto con los que utilizan las herramientas de DRAGEN (no estudiadas en este trabajo, dada su reciente publicación en octubre de 2022). En el futuro se podría probar este nuevo conjunto de herramientas, bajo el nombre de DRAGEN-GATK (que ganaron el último desafío NIST y al igual que GRAF, utilizan genomas gráficos de referencia), que parecerían presentar mejores resultados para WGS. [105, 107, 108] Sin embargo, deberían confirmarse estos resultados para muestras WES y WGS.

En cuanto al flujo de análisis GRAF, si analizamos los resultados del valor F1 y lo comparamos con lo obtenido en la última publicación del desafío *FDA Truth Challenge V2: Calling variants from short- and long-reads in difficult-to-map regions* [105], publicado en Mayo del 2022, vemos que nuestros valores de métricas resultan similares, aunque comparativamente menores a los obtenidos por otros equipos. En esta publicación se muestra que el flujo de análisis con GRAF alcanza un promedio aproximado del 99% para el valor F1 tomando en cuenta tanto las variantes SNP como las INDELs; en cambio, tomando nuestros resultados se alcanza a un promedio de 98% con GRAF para la muestra WGS NA12878. La diferencia podría deberse a que en esta versión del desafío, se seleccionan los ganadores en función del promedio del valor F1 para los SNP e INDEL de las muestras WGS HG003 (NA24149) y HG004 (NA24143) de la versión 4.2 de GIAB, mientras que en nuestro caso usamos la muestra NA12878 de la misma versión. A su vez, cabe destacar que muchos son equipos de empresas privadas como Google o Sentieon; o co-financiadas, como GATK.

Además, hemos notado que los resultados son muy dependientes del archivo BED utilizado. Encontramos una diferencia de ≈2.6% en sensibilidad, y ≈15% en precisión y valor F1 al cambiar solamente este archivo para un mismo análisis (ver apéndice sección D.2). Esto podría justificar la diferencia en los resultados al comparar con otros trabajos [35, 42, 96, 109], donde se encuentran para muestras WES un valor F1 mayor a 96% para INDELs y 99% para SNPs; contra 72,2% aproximadamente y 94,5% respectivamente obtenidos en este trabajo. Al ser publicaciones de hace 3-4 años, usaron la versión 3.3.1 de GIAB (del 14 de octubre de 2016) con sus archivos BED del momento; mientras que en nuestro caso, utilizamos la versión 4.2.1 (del 29 de septiembre del 2021) con los archivos BED actualizados. Esta última versión del VCF para realizar la validación, incluye variantes en regiones más “difíciles de mapear”, como segmentos duplicados y las regiones

[1] Además, es posible agregar variantes específicas de la población, logrando mejores resultados en sensiblidad. [106]

altamente polimórficas (como las pertenecientes al complejo mayor de histocompatibilidad) [105], lo cual puede justificar las métricas más bajas.

Sin embargo, a pesar de que supera los objetivos de este trabajo, para futuras modificaciones se podría ejecutar el flujo de análisis para las muestras WES en [110] siguiendo los pasos del artículo [109] para comparar resultados.

En cuanto al uso práctico, todas las herramientas GRAF se encuentran inmersas en el contexto de CGC. Esto hace que sean más fáciles de utilizar y que estén periódicamente mantenidas y actualizadas por un grupo de bioinformáticos de la plataforma. En cambio, con las otras herramientas de GATK, se debe saber programar y el usuario debe encargarse de estar al día con las últimas versiones de cada herramienta. Sin embargo, cabe destacar una desventaja que surge de usar GRAF: no se sabe con exactitud qué es lo que está ejecutando cada herramienta (por ejemplo, cómo se produce la alineación), con lo cual funciona como una caja negra. Esto genera que el usuario no tenga libertad de hacer cambios particulares (salvo que existan parámetros que GRAF permita cambiar para una herramienta particular, como por ejemplo "Trim adapters.[en] "True/False"para el alineamiento) en el análisis o en las distintas herramientas si así lo desea. [102, 111]

Por último, es importante destacar que los flujos de análisis de GRAF generan resultados satisfactorios tanto para muestras WES como WGS. Hasta el momento, en el laboratorio solicitante de este trabajo, se usaban flujos que contaban con BWA-MEM para la alineación y Haplotype Caller para el llamado de variantes; y que eran seguidos de Hard Filtering o CNN Score Variants+Filter Variant Tranches según si la muestra era de WGS o WES, respectivamente. Con este trabajo entonces, se generó un nuevo flujo de análisis que el laboratorio puede usar independientemente de si la muestra pertenece a exoma o genoma completo, además de que se realizaron los análisis de sensibilidad y especificidad requeridos, se implementaron mejoras a los flujos, y se automatizaron la anotación y generación de los informes, cumpliendo con todos los objetivos propuestos.

Entre algunas de las limitaciones que surgieron a momento de realizar la validación para las variantes germinales, caben diferenciar las limitaciones generales de los proyectos bioinformáticos, inherentes a los datos de secuenciación; y los particulares que surgieron en la elaboración de este trabajo.

En cuanto a los primeros, cabe destacar a que pesar de que los archivos utilizados para validación (GIAB v4.2.1) incluyen regiones más desafiantes para la detección de variantes en comparación con otras versiones, siguen excluyendo regiones genómicas complicadas para la detección de variantes como segmentos altamente similares de duplicación, ADN satélite como centrómeros, INDELs de tamaño mediano mayor a 15 pb y variantes estructurales y de número de copia. Esto podría hacer también que se sobrestime la precisión y que el flujos de análisis se comporten distinto con otras muestras. Asimismo, en las regiones donde haya duplicaciones u otras variantes estructurales complejas, se podrían dar una cantidad significativa de falsos positivos. [112]. También es importante aclarar que se debe realizar una inspección manual de datos de

secuencia en un navegador de genoma para un subconjunto de falsos positivos y falsos negativos para una comprensión precisa de estadísticas como la sensibilidad y la precisión. [112]

En cuanto a los segundos, los resultados en las métricas se obtuvieron solamente para las muestras WES NA12878, NA24385 y NA24631 y de la muestra WGS NA12878; pero los flujos de análisis podrían funcionar de manera diferente cuando se confrontan con otras muestras. Para una evaluación de rendimiento más completa, se podría realizar en el futuro una validación con mayor cantidad de muestras y hacer experimentos adicionales. [113] Por otra parte, para mejorar los modelos y el desempeño en muestras de diferentes orígenes étnicos se debería incluir un conjunto más diverso en las muestras de GIAB, especialmente de ascendencia africana, hispana o mixta. [42]

CAPÍTULO 16

FLUJO DE ANÁLISIS DE VARIANTES SOMÁTICAS

La detección de mutaciones somáticas es importante ya que permite definir en muchos casos el diagnostico, pronostico y el mejor tratamiento de un paciente. De hecho, la mayoría de las terapias dirigidas contra el cáncer se dirigen a mutaciones somáticas específicas. [114]

En este trabajo, se desarrolló el flujo de análisis somático en CWL y se realizaron algoritmos *in silico* para examinar esta clase de variantes. Para llevar a cabo estos objetivos, se utilizó el flujo de análisis germinal hasta el llamado de variantes (no inclusive) y se incorporó Mutect2 [91] y su correspondiente herramienta de filtrado (Filter Mutect2 [92]). Estas herramientas se basan en la determinación de las variantes somáticas comparando las muestras tumoral y normal del paciente. Las variantes germinales se detectarán en ambas muestras, con lo cual estas se descartan en el análisis somático obteniendo así las variantes tumorales únicamente. [91]

A diferencia del análisis germinal, no existe hasta el momento una metodología estandarizada para la validación de flujos de análisis somáticos. Por esta razón, la elección de herramientas en nuestro caso, fue determinada siguiendo las recomendaciones de instituciones de referencia como el Broad Institute del MIT y Harvard [39], otros laboratorios europeos de análisis genómicos [104]; y por último mediante una revisión bibliográfica de las herramientas de llamado de variantes somáticas. En estas, se observó que tanto Strelka2 como Mutect2 generaban los mejores resultados, pero Mutect2 era superior en muestras fijadas en parafina [115] (que corresponde a como suelen llegar a los laboratorios las muestras somáticas [104]). Específicamente, en otro estudio ([116]) observaron que en cuanto a análisis de datos con mayor frecuencia de mutación ($\geq$ 20%), Strelka2 presentaba mayor precisión y Mutect2 un mayor valor F1; y a frecuencias de mutación menores del 10%, Strelka2 presentaba menores valores F1 que Mutect2. A momento de elegir entre ambas herramientas entonces, consideramos conveniente trabajar con Mutect2, lo cual coincide también con lo encontrado en la publicación *"La detección de mutaciones oncogénicas y*

clínicamente procesables en genomas de cáncer depende críticamente de las herramientas de llamado de variantes" [117].

Durante la escritura de este trabajo se publicó un algoritmo de Seven Bridges equivalente al análisis GRAF germinal, pero somático. Este, llama a las variantes somáticas a través de la utilización de un genoma gráfico personal a partir de las variantes germinales y utilizando un genoma de referencia gráfico para mapear con precisión la muestra tumoral. El genoma gráfico personal se usa para identificar a las variantes somáticas con una mayor probabilidad previa de ser un variante de la línea germinal durante el genotipado. Después de este paso, se aplica Hard Filtering para distinguir y eliminar llamadas somáticas falsas positivas. [111] Considerando los resultados alentadores obtenidos en el contexto germinal, seria conveniente considerar la incorporación de esta herramienta en un futuro.

Por último, se plantea también como proyecto en el laboratorio generar un conjunto de validación para obtener la herramienta óptima para la detección de variantes somáticas.

CAPÍTULO 17

ANOTACIÓN DE LAS VARIANTES

Para la anotación de variantes germinales se utilizó Intervar que asigna patogenicidad a las variantes según los criterios ACMG y; para las somáticas, CancerVar que implementa las reglas de AMP.

Las reglas de ACMG son fundamentales para la correcta asignación de patogenicidad ya que contemplan integralmente la importancia de las variantes a través de datos de bases poblacionales como frecuencias o efectos en las distintas poblaciones, relevancia clínica, biológica, fisiopatológica y datos funcionales e in silico. [28, 118] Al igual que ACMG, utilizar los criterios de AMP para la clasificación de variantes somáticas es imprescindible debido a que consideran los biomarcadores designados por la Administración de Drogas y Alimentos de los EE. UU. (FDA), pautas profesionales, ensayos clínicos, función de las variantes, bases de datos de población, bases de datos de variantes, algoritmos predictivos y literatura publicada. [29, 119]

Actualmente, ACMG y AMP son los dos criterios de asignación de patogenicidad mayormente adoptados por la comunidad científica para la clasificación de variantes. [119] De hecho, en laboratorios de Europa [104] y en Varsome, que es un estándar a nivel clínico, se utiliza ACMG para la clasificación de variantes germinales y AMP para la clasificación de variantes somáticas [120]

Se eligió trabajar con InterVar y con CancerVar dado que son herramientas de código abierto que siguen las recomendaciones internacionales de ACMG y AMP. Ambas son capaces de asignar patogenicidad o nivel de significancia según corresponda, así como también los criterios utilizados para la interpretación de las variantes. [29, 121]

Una de los grandes desafíos que encontramos en este trabajo fue incorporar InterVar/CancerVar al flujo de análisis ya que son dos herramientas de código abierto con escasa documentación y una implementación que no está tan optimizada. A modo de ejemplo, una de las dificultades fue integrar los tres archivos que devuelven las herramientas como salida en uno. Sin embargo, luego de una

curva de aprendizaje de la herramienta, pudimos observar el beneficio de asignar la patogenicidad correctamente y obtener una gran cantidad de información relevante sobre las variantes sin tener que pagar por este servicio.

A pesar de que no existe un *gold standard* para anotación[1], lo que hace que se requiera de un genetista para el análisis de resultados; es fundamental incluir la anotación debido a que permite obtener información integral y unificada sobre las variantes de manera amigable. Esto agiliza y facilita el proceso para el profesional que requiera los datos ya que le permite obtener la información sin tener que manipular bases de datos complejas, ni tener conocimientos en bioinformática. Incluso, en muchos casos, los anotadores elegidos pueden agregar información de interés personalizado para el profesional a las variantes detectadas. [122–124] Para CancerVar, se incluye la relación de la variante con la patología; de esta manera se apoya al médico genetista u oncólogo a buscar un tratamiento idóneo para su paciente. [121]

Por otro lado, con respecto a las herramientas de código abierto disponibles, no hay publicaciones que hablen del procesamiento e identificación técnica de variantes germinales y somáticas en conjunto con la anotación y clasificación de estas. Este paso es fundamental para la interpretación de variantes y toma de decisiones. [125] Esta es una de las grandes innovaciones de este trabajo: el agregado de las herramientas InterVar, CanCervar y los scripts que generan archivos de salidas más amigables e informes en la misma plataforma que el resto de las herramientas. Esto permite que se pueda obtener no solo la información técnica sobre las variantes, sino también con la clasificación ACMG/AMP, información sobre estadística poblacional, relevancia clínica, biológica y fisiopatológica desde CGC.

Cabe destacar que comercialmente, existen empresas bioinformáticas que dan esa información como Varsome [126] o Sophia Genetics [127]; que es un estándar a nivel clínico, dado que tiene sus análisis optimizados para datos de secuenciación Illumina, y ambas cobran $100 USD o más por el procesamiento bioinformático de las variantes genómicas. En nuestro país, la empresa Bitgenia [128] también brinda análisis de variantes germinales y se basa, como en este trabajo, en Intervar para el análisis de las variantes. En este trabajo el costo de los análisis fue menor a $1 USD para WES, y $\approx$ $15 USD para WGS.

[1] Esto se debe a la amplia cantidad de variantes que se describen día a día, así como también a las nuevas significancias dadas a las variantes de significado incierto.

CAPÍTULO 18

ACCESO A LOS FLUJOS DE ANÁLISIS Y RESULTADOS

Uno de los grandes avances en este trabajo fue poner a disposición los flujos de análisis germinal y somático en la plataforma CGC. Esto permite que un usuario sin experiencia en bioinformática pueda analizar muestras de manera sencilla. No solo es más fácil acceder y analizar datos[1], sino también administrar los recursos de IT al estar conectado con la nube de AWS para optimizar el procesamiento, asignar recursos informáticos y de almacenamiento a pedido y satisfacer las necesidades de los análisis en constante crecimiento. [129, 130]

Además, CGC permite escalar de manera sencilla y flexible la cantidad y los tipos de instancias informáticas permitiendo la optimización de los flujos de análisis para que se ejecuten en paralelo y permitiendo que no haya límites para la cantidad de ejecuciones o la cantidad de datos a analizar en la plataforma. [130] Además, a través de funcionalidades como la utilización de *Spot/Preemptible Instances*[2] y *Memoization (WorkReuse)*[3] puede lograr una optimización significativa del tiempo y costos. [131, 132] En este trabajo en particular, luego de la incorporación de CGC, nos fue posible ejecutar 5 flujos de análisis al mismo tiempo a través de 5 clics; algo que resultaba anteriormente impensado.

Por otro lado, también se diseñó un informe para visualizar y analizar los datos más fundamentales. En cuanto a la patogenicidad de las variantes informadas, se elige incluir las patogénicas y probablemente patogénicas de acuerdo a lo recomendado en diversas guías internacionales [29, 38]. En muchos casos, esta información es crucial para el diagnostico/pronóstico/tratamiento del expectante paciente, ya que depende de esta para el tratamiento de su patología.

[1] En comparación con los flujos de análisis que se corren en instancias de AWS a través de la línea de comando en Linux

[2] Instancias que aprovechan la capacidad de EC2 no utilizada en la nube de AWS adecuada para cargas de trabajo flexibles en el tiempo que están disponibles con un descuento de hasta el 90%

[3] Funcionalidad por la que CGC reutiliza resultados ya existentes de ejecuciones anteriores

El principal motivo de no informar variantes de significado incierto está relacionado al impacto negativo que pueden generar en el paciente, dado lo sensible y poco concluyentes/informativas que son este tipo de variantes, considerando que los genetistas son incapaces de dar un consejo inequívoco [133]. Además, hay una gran cantidad de VUS que se encuentran en cada individuo particular lo cual extendería enormemente el informe; y las guías AMP [29] recomiendan informes breves, simples y directos; ya que suponen que los datos que están en la página 2 o más allá tienen una alta probabilidad de ser pasados por alto por el médico tratante. De todas maneras, las variantes de significado incierto estarán presentes en el archivo digital final que se le envía al profesional, dado que son importantes de considerar en ciertos casos donde se cuenta con datos familiares o historia clínica del paciente. Por ejemplo, si una mujer se presenta al estudio con antecedentes familiares de cáncer de mama, y se obtienen variantes de significado incierto en los genes BRCA1/2, estas deberán analizarse en mayor medida por el médico genetista e irán incluidas en el informe final tras su decisión. [104, 134]

Es importante tener en cuenta que al momento de revisar los resultados obtenidos en el flujo de análisis, siempre debe haber un genetista experto que certifique el informe final. Por eso, se eligió que el flujo de análisis no elimine ninguna de las variantes llamadas, simplemente se agrega si pasaron o no los filtros. Si bien en la mayoría de los casos, las variantes filtradas son artefactos, según los resultados obtenidos en las secciones 11.1 y 11.2, puede llegar a suceder que algunas de las variantes filtradas sea una variante real. Por lo tanto, se decidió que lo más conveniente seria entregar un archivo que tenga todas las variantes y en una columna se mencione si la variante pasa todos los filtros o no. Dado que estos resultados tienen el objetivo de ayudar al médico genetista a definir el informe final, de esta manera se facilita la información para un procedimiento estandarizado. Este comúnmente comenzará analizando únicamente las variantes que pasaron los filtros pero, a su vez permite analizar las filtradas en caso de que no haya un resultado determinante entre las primeras. Cabe aclarar que estas variantes al no haber pasado los filtros, para poder tener una seguridad en la información, deberían ser validadas con una técnica ortogonal como por ejemplo Sanger. [23, 28]

Parte V

Conclusión

En el marco del proyecto final de la carrera de Bioingeniería, nos es fundamental destacar que la bioinformática es una disciplina que está en pleno auge y que, cada vez más, se necesitan individuos idóneos capaces de entender, manipular y extraer información útil a partir de datos biológicos para los profesionales de la salud que los requieran. Fundamentalmente, cuando se parte de datos de NGS, los datos albergan una gran complejidad y magnitud, además de las expectativas de un diagnóstico del paciente (y sus familias y allegados) al que pertenecen. Durante la elaboración de este trabajo aprendimos conceptos fundamentales de genómica, cáncer y manejo de datos de NGS, lo cual que nos ha abierto las puertas a un mar de posibilidades y nuevos horizontes para nuestro presente (pues ya ambas autoras nos involucramos laboralmente en este mundo) y futuro.

Durante la elaboración de este trabajo, logramos desarrollar un flujo de análisis bioinformático para estudios NGS somático y germinal. Ambos se desarrollaron en CWL, AWS y CGC.

Logramos a su vez, cumplir con el objetivo de máxima, y generar un informe amigable para el usuario no bioinformático donde se resumen los hallazgos más importantes jerarquizados de acuerdo a su relevancia/patogenicidad.

A diferencia de todos los flujos de análisis de código libre conocidos hasta el momento, pudimos integrar de manera satisfactoria tres pasos del procesamiento de variantes de análisis NGS: el procesamiento bioinformático técnico, la anotación con datos biológicos/fisiopatológicos/clínicos y la clasificación por ACMG y AMP.

Para el caso de las variantes germinales, realizamos la validación de análisis obteniendo resultados satisfactorios tanto para los parámetros internos del laboratorio como comparativamente a flujos de análisis comerciales. De esta manera se logró obtener un flujo de análisis que tenga las mejores métricas y sea óptimo y eficiente computacionalmente tanto para variantes en análisis de WES como de WGS, siendo el elegido aquel que utilizaba GRAF para el alineamiento, llamado y filtrado de variantes. En particular se justifico dicha elección, a partir del análisis de sensibilidad, precisión y valor F1 en las distintas combinaciones de herramientas.

En cuanto al análisis somático, logramos generar un flujo de análisis capaz de llamar y anotar variantes puramente somáticas (distinguiéndolas de las germinales). Aunque escapa a los objetivos de este trabajo, se dejan sentadas las bases para que, en un futuro, se estudien los parámetros de sensibilidad y precisión para su posible implementación clínica en el laboratorio.

Por último, se disponibilizó el flujo de análisis en CGC y los resultados en un informe y un archivo de salida completo con información relevante para la determinación diagnóstica o terapéutica, a saber, datos biológicos, fisiopatológicos, poblacionales y clínicos, entre otros. Esto permite que una persona sin experiencia en bioinformática pueda ejecutar de manera sencilla el flujo de análisis de variantes somáticas y germinales, tanto para muestras WES como WGS. A su vez, permite obtener los resultados de manera amigable y facilitarle el proceso a los profesionales de la salud para la toma de decisión diagnóstica o terapéutica que en ultima instancia, tiene un beneficio para la vida de los pacientes.

BIBLIOGRAFÍA

[1] M. Dhar, "What is rna? | live science," *Live Science*, 10 2020. [Online]. Available: https://www.livescience.com/what-is-RNA.html

[2] M. Jafari, N. Ansari-Pour, S. Azimzadeh, and M. Mirzaie, "A logic-based dynamic modeling approach to explicate the evolution of the central dogma of molecular biology," *PLoS ONE*, vol. 12, 12 2017. [Online]. Available: /pmc/articles/PMC5739447//pmc/articles/PMC5739447/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC5739447/

[3] S.-K. Low, H. Zembutsu, and Y. Nakamura, "Breast cancer: The translation of big genomic data to cancer precision medicine," 2017.

[4] "Definition of variant - nci dictionary of cancer terms - nci." [Online]. Available: https://www.cancer.gov/publications/dictionaries/cancer-terms/def/variant

[5] N. L. of Medicine, "What is a gene variant and how do variants occur?: Medlineplus genetics," 3 2021. [Online]. Available: https://medlineplus.gov/genetics/understanding/mutationsanddisorders/genemutation/

[6] G. I. of Medical Research, "Types of variants | garvan institute of medical research," 2021. [Online]. Available: https://www.garvan.org.au/research/kinghorn-centre-for-clinical-genomics/learn-about-genomics/for-gp/genetics-refresher-1/types-of-variants

[7] Z. Klaassen, "Siu virtual congress 2020: Germline and somatic mutations in advanced prostate cancer: Actionable targets," 10 2020. [Online]. Available: https://www.urotoday.com/conference-highlights/siu-2020/siu-2020-gu-malignancies-prostate/125093-siu-virtual-congress-2020-germline-and-somatic-mutations-in-advanced-prostate-cancer-actionable-html

[8] "Transitions vs transversions," 2021. [Online]. Available: https://www.mun.ca/biology/scarr/Transitions_vs_Transversions.html

[9] M. Arabnejad and B. A. Dawkins, "Definition of transitions and transversions. nucleotides a and g... | download scientific diagram," 11 2018. [Online]. Available: https://www.researchgate.net/figure/

Definition-of-transitions-and-transversions-Nucleotides-A-and-G-orange-circles-are-in_fig2_328727667

[10] "Genomic alterations : Changes in the humman genome | com.pl.it®," 2020. [Online]. Available: https://complit.gr/genomic-alterations/

[11] NIH, "Heterocigoto," 8 2022. [Online]. Available: https://www.genome.gov/es/genetics-glossary/Heterocigoto#

[12] "Homozygous," 12 2022. [Online]. Available: https://www.genome.gov/genetics-glossary/homozygous

[13] "Heterozygous," 12 2022. [Online]. Available: https://www.genome.gov/genetics-glossary/heterozygous

[14] "Definición de cáncer - diccionario de cáncer del nci - nci." [Online]. Available: https://www.cancer.gov/espanol/publicaciones/diccionarios/diccionario-cancer/def/cancer

[15] "Cáncer: Medlineplus enciclopedia médica." [Online]. Available: https://medlineplus.gov/spanish/ency/article/001289.htm

[16] D. Hanahan and R. A. Weinberg, "Hallmarks of cancer: The next generation," *Cell*, vol. 144, pp. 646–674, 3 2011. [Online]. Available: http://www.cell.com/article/S0092867411001279/fulltexthttp://www.cell.com/article/S0092867411001279/abstracthttps://www.cell.com/cell/abstract/S0092-8674(11)00127-9

[17] "What is cancer? - nci," 5 2021. [Online]. Available: https://www.cancer.gov/about-cancer/understanding/what-is-cancer#definition

[18] P. Garrido, A. Aldaz, R. Vera, M. Calleja, E. Álava, M. Martín, M. X., and J. Palacios, "Proposal for the creation of a national strategy for precision medicine in cancer: a position statement of SEOM, SEAP, and SEFH," *Clin Transl Oncol*, pp. 1–5, 2017.

[19] G. S. Ginsburg and K. A. Phillips, "Health aff (millwood)," vol. 37, pp. 694–701, 2018.

[20] K. B. Johnson, W. Q. Wei, D. Weeraratne, M. E. Frisse, K. Misulis, K. Rhee, J. Zhao, and J. L. Snowdon, "Precision medicine, ai, and the future of personalized health care," *Clinical and Translational Science*, vol. 14, p. 86, 1 2021. [Online]. Available: /pmc/articles/PMC7877825//pmc/articles/PMC7877825/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC7877825/

[21] A. Carracedo, "Técnicas de secuenciación - máster de genética y genómica," Oct 2022.

[22] L. J. Steinbock and A. Radenovic, "The emergence of nanopores in next-generation sequencing," *Nanotechnology*, vol. 26, p. 074003, 2 2015. [Online]. Available: https://iopscience.iop.org/article/10.1088/0957-4484/26/7/074003https://iopscience.iop.org/article/10.1088/0957-4484/26/7/074003/meta

[23] S. Goodwin, J. D. McPherson, and W. R. McCombie, "Coming of age: ten years of next-generation sequencing technologies," *Nature Reviews Genetics 2016 17:6*, vol. 17, pp. 333–351, 5 2016. [Online]. Available: https://www.nature.com/articles/nrg.2016.49

[24] GenBank, "Human genome version 38 faq general questions," 2022. [Online]. Available: http://genome.ucsc.edu/cgi-bin/hgLiftOver

[25] D. Kim, J. M. Paggi, C. Park, C. Bennett, and S. L. Salzberg, "Graph-based genome alignment and genotyping with hisat2 and hisat-genotype," *Nature Biotechnology 2019 37:8*, vol. 37, pp. 907–915, 8 2019. [Online]. Available: https://www.nature.com/articles/s41587-019-0201-4

[26] N. Education, "allele frequency | learn science at scitable," 2014. [Online]. Available: https://www.nature.com/scitable/definition/allele-frequency-298/

[27] S. Roy, C. Coldren, A. Karunamurthy, N. S. Kip, E. W. Klee, S. E. Lincoln, A. Leon, M. Pullambhatla, R. L. Temple-Smolkin, K. V. Voelkerding, C. Wang, and A. B. Carter, "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*, vol. 20, pp. 4–27, 1 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1525157817303732#:~:text=A%20bioinformatics%20pipeline%20is%20composed,entirely%20developed%20by%20the%20laboratory.

[28] S. Richards, N. Aziz, S. Bale, D. Bick, S. Das, J. Gastier-Foster, W. W. Grody, M. Hegde, E. Lyon, E. Spector, K. Voelkerding, and H. L. Rehm, "Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the american college of medical genetics and genomics and the association for molecular pathology," *Genetics in medicine : official journal of the American College of Medical Genetics*, vol. 17, pp. 405–424, 5 2015. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/25741868/

[29] M. M. Li, M. Datto, E. J. Duncavage, S. Kulkarni, N. I. Lindeman, S. Roy, A. M. Tsimberidou, yy L Cindy Vnencak-Jones, zz J Daynna Wolff, xx Anas Younes, M. N. Nikiforova, M. Basehore, C. Coldren, L. Cook, J. Crow, B. Funke, M. Hameed, L. Jennings, A. Karunamurthy, A. Kim, B. Krock, M. Lowery-Nordberg, M. Miller, B. Pinsky, M. Routbort, R. Schmidt, and D. Viswanatha, "Standards and guidelines for the interpretation and reporting of sequence variants in cancer,"

The Journal of Molecular Diagnostics, vol. 19, pp. 4–23, 2017. [Online]. Available: http://dx.doi.org/10.1016/j.jmoldx.2016.10.002jmd.amjpathol.org

[30] NIH, "Definiciones: Bioinformática," 12 2022. [Online]. Available: https://www.genome.gov/es/genetics-glossary/Bioinformatica

[31] "¿qué es la bioinformática y qué aplicaciones tiene en biomedicina?" 7 2020. [Online]. Available: https://www.isciii.es/InformacionCiudadanos/DivulgacionCulturaCientifica/DivulgacionISCIII/Paginas/Divulgacion/Bioinformatica.aspx

[32] D. D. Arce, "Diseño de una arquitectura en pipeline para la descarga y análisis de secuencias de promotores en solanum lycopersicum," 2016. [Online]. Available: https://rephip.unr.edu.ar/bitstream/handle/2133/12340/Trabajo%20Final%20Pistilli.pdf?sequence=3&isAllowed=y

[33] L. V. Cabrera and M. de J. Pérez Jiménez, "Técnicas inteligentes en bioinformática: Introducción a la bioninformática secuenciación," 2016. [Online]. Available: https://www.cs.us.es/cursos/tib-2015/temas/TIB-SecuenciacionEnsamblado-2015-2016.pdf

[34] S. Roy, "Next-generation sequencing bioinformatics pipelines | aacc.org," 3 2020. [Online]. Available: https://www.aacc.org/cln/articles/2020/march/next-generation-sequencing-bioinformatics-pipelines

[35] J. Chen, X. Li, H. Zhong, Y. Meng, and H. Du, "Systematic comparison of germline variant calling pipelines cross multiple next-generation sequencers," *Scientific Reports 2019 9:1*, vol. 9, pp. 1–13, 6 2019. [Online]. Available: https://www.nature.com/articles/s41598-019-45835-3

[36] S. Zhao, O. Agafonov, A. Azab, T. Stokowy, and E. Hovig, "Accuracy and efficiency of germline variant calling pipelines for human genome data," *Scientific Reports 2020 10:1*, vol. 10, pp. 1–12, 11 2020. [Online]. Available: https://www.nature.com/articles/s41598-020-77218-4

[37] Z. Ahmed, E. G. Renart, D. Mishra, and S. Zeeshan, "Jwes: a new pipeline for whole genome/exome sequence data processing, management, and gene-variant discovery, annotation, prediction, and genotyping," *FEBS Open Bio*, vol. 11, p. 2441, 9 2021. [Online]. Available: /pmc/articles/PMC8409305//pmc/articles/PMC8409305/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC8409305/

[38] H. H. Giles, M. R. Hegde, E. Lyon, C. M. Stanley, I. D. Kerr, M. E. Garlapow, and J. M. Eggington, "The science and art of clinical genetic variant classification and its impact on test accuracy," *Annual review of genomics and human genetics*, vol. 22, pp. 285–307, 2021. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/33900788/

[39] BroadInstitute, “Gatk,” 2022. [Online]. Available: https://gatk.broadinstitute.org/hc/en-us

[40] L. G. de los Países Bálticos, “Bio knowledge lab | tu partner biotecnológico,” 2021. [Online]. Available: https://www.b-kl.eu/variant-calling/

[41] J. H. Lee, S. Kweon, and Y. R. Park, “Sharing genetic variants with the ngs pipeline is essential for effective genomic data sharing and reproducibility in health information exchange,” *Scientific Reports* |, vol. 11, 1 2021. [Online]. Available: https://doi.org/10.1038/s41598-021-82006-9

[42] Y. A. Barbitoff, R. Abasov, V. E. Tvorogova, A. S. Glotov, and A. V. Predeus, “Systematic benchmark of state-of-the-art variant calling pipelines identifies major factors affecting accuracy of coding sequence variant discovery,” *BMC Genomics*, vol. 23, pp. 1–17, 12 2022. [Online]. Available: https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-022-08365-3

[43] G. Rakocevic, V. Semenyuk, W.-P. Lee, J. Spencer, J. Browning, I. J. Johnson, V. Arsenijevic, J. Nadj, K. Ghose, M. C. Suciu, S.-G. Ji, G. Demir, L. Li, B. Toptaş, A. Dolgoborodov, B. Pollex, I. Spulber, I. Glotova, P. Kómár, A. L. Stachyra, Y. Li, M. Popovic, M. Källberg, A. Jain, and D. Kural, “Fast and accurate genomic analyses using genome graphs.” [Online]. Available: https://doi.org/10.1038/s41588-018-0316-4

[44] S. Zverinova and V. Guryev, “Variant calling: Considerations, practices, and developments,” vol. 8, 2022. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1002/humu.24311

[45] F. Borchert, A. Mock, A. Tomczak, J. Hügel, S. Alkarkoukly, A. Knurr, A.-L. Volckmar, A. Stenzinger, P. Schirmacher, J. Debus, D. Jäger, T. Longerich, S. Fröhling, R. Eils, N. Bougatf, U. Sax, and M.-P. Schapranow, “Knowledge bases and software support for variant interpretation in precision oncology,” *Briefings in Bioinformatics*, vol. 22, pp. 1–17, 2021. [Online]. Available: https://doi.org/10.1093/bib/bbab134

[46] K. Dixon, S. Young, Y. Shen, M. L. Thibodeau, A. Fok, E. Pleasance, E. Zhao, M. Jones, G. Aubert, L. Armstrong, A. Virani, D. Regier, K. Gelmon, D. Renouf, S. Chia, I. Bosdet, S. R. Rassekh, R. J. Deyell, S. Yip, A. Fisic, E. Titmuss, S. Abadi, S. J. Jones, S. Sun, A. Karsan, M. Marra, J. Laskin, H. Lim, and K. A. Schrader, “Establishing a framework for the clinical translation of germline findings in precision oncology,” *JNCI Cancer Spectrum*, vol. 4, 10 2020. [Online]. Available: /pmc/articles/PMC7583151//pmc/articles/PMC7583151/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC7583151/

[47] NIST, “Genome in a bottle | nist,” 11 2022. [Online]. Available: https://www.nist.gov/programs-projects/genome-bottle

[48] P. Krusche and D. Skola, “hap.py/happy.md at master · illumina/hap.py · github,” 10 2019. [Online]. Available: https://github.com/Illumina/hap.py/blob/master/doc/happy.md

[49] P. Krusche, L. Trigg, P. C. Boutros, C. E. Mason, F. M. D. L. Vega, B. L. Moore, M. Gonzalez-Porta, M. A. Eberle, Z. Tezak, S. Lababidi, R. Truty, G. Asimenos, B. Funke, M. Fleharty, B. A. Chapman, M. Salit, and J. M. Zook, “Best practices for benchmarking germline small-variant calls in human genomes,” *Nature Biotechnology 2019 37:5*, vol. 37, pp. 555–560, 3 2019. [Online]. Available: https://www.nature.com/articles/s41587-019-0054-x

[50] “Aws | informática en la nube. ventajas y beneficios,” 2023. [Online]. Available: https://aws.amazon.com/es/what-is-cloud-computing/

[51] “Aws | elastic compute cloud (ec2) de capacidad modificable en la nube,” 2023. [Online]. Available: https://aws.amazon.com/es/ec2/?nc2=h_ql_prod_fs_ec2

[52] “Aws | almacenamiento de datos seguro en la nube (s3),” 2023. [Online]. Available: https://aws.amazon.com/es/s3/?nc2=h_ql_prod_fs_s3

[53] “Administración de identidades | iam | aws,” 2023. [Online]. Available: https://aws.amazon.com/es/iam/?nc2=type_a

[54] Seven-Bridges, “Rabix: Power tools for the common workflow language,” 2019. [Online]. Available: https://rabix.io/

[55] C. Project-Team, “1.1. quick start — common workflow language user guide 0.1 documentation,” 2013. [Online]. Available: https://www.commonwl.org/user_guide/introduction/quick-start.html

[56] D. Inc., “Docker: Accelerated, containerized application development,” 2023. [Online]. Available: https://www.docker.com/

[57] “Seven bridges genomics - the biomedical data analysis company,” 2022. [Online]. Available: https://www.sevenbridges.com/

[58] “Cancer genomics cloud,” 2022. [Online]. Available: https://www.cancergenomicscloud.org/

[59] GIAB-NA12878, “Reference sample na12878.” [Online]. Available: https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/NA12878/

[60] GIAB-NA24385, “Reference sample na24385.” [Online]. Available: https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/AshkenazimTrio/HG002_NA24385_son/

[61] GIAB-NA24631, “Reference sample na24631.” [Online]. Available: https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/ChineseTrio/HG005_NA24631_son/

[62] GIAB-GRCh38, “Reference sample grch38,” 4 2021. [Online]. Available: https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/release/references/GRCh38/

[63] GIAB-GRCh37, “Reference sample grch37,” 4 2021. [Online]. Available: https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/release/references/GRCh37/

[64] GRAF, “Grch38.graf.pan_genome_reference.v1.vcf.gz.tbi files files public files,” 6 2020. [Online]. Available: https://cgc.sbgenomics.com/public/files/5eea52fdc80cb0e4af7ee42f/

[65] GCP-Database, “Google cloud console,” 2023. [Online]. Available: https://console.cloud.google.com/storage/browser/gcp-public-data--broad-references/hg19/v0?pageState=(%22StorageObjectListTable%22:(%22f%22:%22%255B%255D%22))&prefix=&forceOnObjectsSortingFiltering=false

[66] 1000Genomes, “Consola de google cloud,” 7 2016. [Online]. Available: https://console.cloud.google.com/storage/browser/genomics-public-data/resources/broad/hg38/v0

[67] NCBI-Data, “Index of /snp,” 4 2018. [Online]. Available: https://ftp.ncbi.nih.gov/snp/

[68] “Sureselect custom dna target enrichment probes | agilent.” [Online]. Available: https://www.agilent.com/en/product/next-generation-sequencing/hybridization-based-next-generation-sequencing-ngs/ngs-custom-target-enrichment-probes/ngs-custom-target-enrichment-probes-232874

[69] “Whole exome sequencing | idt.” [Online]. Available: https://eu.idtdna.com/pages/technology/next-generation-sequencing/dna-sequencing/targeted-sequencing/exome-sequencing

[70] Benchmark-GIAB, “Index of /giab/ftp/release/na12878_hg001/latest/grch38,” 9 2020. [Online]. Available: https://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/NA12878_HG001/latest/GRCh38/

[71] Benchmark-GIAB-NA24385, “Index of /giab/ftp/release/ashkenazimtrio/hg002_na24385_son/latest/grch38,” 12 2020. [Online]. Available: https://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/AshkenazimTrio/HG002_NA24385_son/latest/GRCh38/

[72] Benchmark-GIAB-NA24631, “Index of /giab/ftp/release/chinesetrio/hg005_na24631_son/latest/grch38,” 9 2021. [Online]. Available: https://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/ChineseTrio/HG005_NA24631_son/latest/GRCh38/

[73] C. Lin, “giab_data_indexes/sequence.index.na12878_illumina300x_wgs_09252015 at master · genome-in-a-bottle/giab_data_indexes · github,” 2 2020. [Online]. Available: https://github.com/genome-in-a-bottle/giab_data_indexes/blob/master/NA12878/sequence.index.NA12878_Illumina300X_wgs_09252015

[74] ecSeq Bioinformatics, “Trimming adapter sequences - is it necessary?” 8 2016. [Online]. Available: https://www.ecseq.com/support/ngs/trimming-adapter-sequences-is-it-necessary

[75] Illumina, “Adapter trimming: Why are adapter sequences trimmed from only the 3’ ends of reads?” 2022. [Online]. Available: https://support.illumina.com/bulletins/2016/04/adapter-trimming-why-are-adapter-sequences-trimmed-from-only-the--ends-of-reads.html

[76] M. Martin, “User guide — cutadapt 4.1 documentation,” 2022. [Online]. Available: https://cutadapt.readthedocs.io/en/stable/guide.html

[77] H. Li and R. Durbin, “Fast and accurate long-read alignment with burrows-wheeler transform,” *Bioinformatics*, vol. 26, pp. 589–595, 1 2010.

[78] J. C. Na, H. Kim, H. Park, T. Lecroq, M. Léonard, L. Mouchard, and K. Park, “Fm-index of alignment: A compressed index for similar strings,” *Theoretical Computer Science*, vol. 638, pp. 159–170, 7 2016.

[79] R. J. Musich, “A recent (2020) comparative analysis of genome aligners a recent (2020) comparative analysis of genome aligners shows hisat2 and bwa are among the best tools shows hisat2 and bwa are among the best tools.” [Online]. Available: https://scholarworks.rit.edu/theses

[80] BWA-User-Manual, “Bwa mem for single or paired end reads,” 2013. [Online]. Available: https://chipster.csc.fi/manual/bwa-mem.html

[81] Samtools, “Sequence alignment/map format specification,” 6 2021. [Online]. Available: https://github.com/samtools/hts-specs.

[82] “Duplicate sequences.” [Online]. Available: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/8%20Duplicate%20Sequences.html

[83] “Graf germline variant aligner apps tesis-tere-carli.” [Online]. Available: https://cgc.sbgenomics.com/u/cmoya/tesis-tere-carli/apps/graf-germline-variant-alignment/5

[84] B. Institute, “Haplotypecaller – gatk,” 1 2022. [Online]. Available: https://gatk.broadinstitute.org/hc/en-us/articles/360037225632-HaplotypeCaller

[85] B. Institute-CNNScoreVariants, “Cnnscorevariants – gatk,” 2020. [Online]. Available: https://gatk.broadinstitute.org/hc/en-us/articles/360037226672-CNNScoreVariants

[86] B. Institute-FilterVariantTranches, “Filtervarianttranches – gatk,” 2020. [Online]. Available: https://gatk.broadinstitute.org/hc/en-us/articles/360040098912-FilterVariantTranches

[87] BroadInstitute, “Variantfiltration – gatk,” 2019. [Online]. Available: https://gatk.broadinstitute.org/hc/en-us/articles/360036726871-VariantFiltration

[88] BroadInstitute-Hard-filtering, “Hard-filtering germline short variants – gatk,” 8 2022. [Online]. Available: https://gatk.broadinstitute.org/hc/en-us/articles/360035890471-Hard-filtering-germline-short-variants

[89] “Graf germline variant caller apps tesis-tere-carli.” [Online]. Available: https://cgc.sbgenomics.com/u/cmoya/tesis-tere-carli/apps/graf-germline-variant-caller/1

[90] “Graf germline variant detection workflow public apps.” [Online]. Available: https://cgc.sbgenomics.com/public/apps/admin/sbg-public-data/graf-germline-variant-detection-workflow-1-0

[91] B. Institute-Mutect2, “Mutect2 – gatk,” 2019. [Online]. Available: https://gatk.broadinstitute.org/hc/en-us/articles/360037593851-Mutect2

[92] BroadInstitute, “Filtermutectcalls – gatk,” 2021. [Online]. Available: https://gatk.broadinstitute.org/hc/en-us/articles/360036856831-FilterMutectCalls

[93] L. Quan, “Github - wglab/cancervar: Clinical interpretation of somatic mutations in cancer,” 5 2022. [Online]. Available: https://github.com/WGLab/CancerVar

[94] BCBIO, “Small germline variants — bcbio-nextgen 1.2.9 documentation,” 2021. [Online]. Available: https://bcbio-nextgen.readthedocs.io/en/latest/contents/germline_variants.html

[95] “Contents — bcbio-nextgen 1.2.9 documentation,” 2021. [Online]. Available: https://bcbio-nextgen.readthedocs.io/en/latest/

[96] M. Kumaran, U. Subramanian, and B. Devarajan, “Performance assessment of variant calling pipelines using human whole exome sequencing and simulated data,” *BMC Bioinformatics*, vol. 20, pp. 1–11, 6 2019. [Online]. Available: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-019-2928-9

[97] “Seven bridges graf | pan-genome graphs identify snps, indels and structural variants with higher accuracy.” [Online]. Available: https://www.sevenbridges.com/pan-genome-graphs/

[98] “Making efficient use of compute resources.” [Online]. Available: https://docs.cancergenomicscloud.org/changelog/making-efficient-use-of-compute-resources#section-when-being-scattered-is-a-very-good-thing-optimising-a-whole-genome-analysis

[99] “Aligning reads - seven bridges.” [Online]. Available: https://www.sevenbridges.com/graf/aligning-reads/

[100] S. Friedman, L. Gauthier, Y. Farjoun, and E. Banks, “Lean and deep models for more accurate filtering of snp and indel variant calls,” *Bioinformatics*, vol. 36, pp. 2060–2067, 4 2020. [Online]. Available: https://academic.oup.com/bioinformatics/article/36/7/2060/5674040

[101] “Bioinformatics pipeline: Dna-seq analysis - gdc docs.” [Online]. Available: https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/DNA_Seq_Variant_Calling_Pipeline/

[102] “About the gatk best practices – gatk.” [Online]. Available: https://gatk.broadinstitute.org/hc/en-us/articles/360035894711-About-the-GATK-Best-Practices

[103] “A first look at gatk4 on the seven bridges platform - seven bridges.” [Online]. Available: https://www.sevenbridges.com/first-look-gatk4/

[104] A. Carracedo, “Flujos de análisis genómicos y exómicos - laboratorio central,” Oct 2022.

[105] N. D. Olson, J. Wagner, J. McDaniel, S. H. Stephens, S. T. Westreich, A. G. Prasanna, E. Johanson, E. Boja, E. J. Maier, O. Serang, D. Jáspez, J. M. Lorenzo-Salazar, A. Muñoz-Barrera, L. A. Rubio-Rodríguez, C. Flores, K. Kyriakidis, A. Malousi, K. Shafin, T. Pesout, M. Jain, B. Paten, P.-C. Chang, A. Kolesnikov, M. Nattestad, G. Baid, S. Goel, H. Yang, A. Carroll, R. Eveleigh, M. Bourgey, G. Bourque, G. Li, C. Ma, L. Tang, Y. Du, S. Zhang, J. Morata, R. Tonda, G. Parra, J.-R. Trotta, C. Brueffer, S. Demirkaya-Budak, D. Kabakci-Zorlu, D. Turgut, Özem Kalay, G. Budak, K. Narcı, E. Arslan, R. Brown, I. J. Johnson, A. Dolgoborodov, V. Semenyuk, A. Jain, H. S. Tetikol, V. Jain, M. Ruehle, B. Lajoie, C. Roddey, S. Catreux, R. Mehio, M. U. Ahsan, Q. Liu, K. Wang, S. M. E. Sahraeian, L. T. Fang, M. Mohiyuddin, C. Hung, C. Jain, H. Feng, Z. Li, L. Chen, F. J. Sedlazeck, and J. M. Zook, “Precisionfda truth challenge v2: Calling variants from short and long reads in difficult-to-map regions,” *Cell Genomics*, vol. 2, p. 100129, 5 2022.

[106] H. S. Tetikol, D. Turgut, K. Narci, G. Budak, O. Kalay, E. Arslan, S. Demirkaya-Budak, A. Dolgoborodov, D. Kabakci-Zorlu, V. Semenyuk, A. Jain, and B. N. Davis-Dusenbery, “Pan-african genome demonstrates how population-specific genome graphs improve high-throughput sequencing data analysis,” *Nature Communications 2022 13:1*, vol. 13, pp. 1–11, 8 2022. [Online]. Available: https://www.nature.com/articles/s41467-022-31724-3

[107] “Dragen-gatk – gatk.” [Online]. Available: https://gatk.broadinstitute.org/hc/en-us/articles/360045944831-DRAGEN-GATK

[108] “Functional equivalence in dragen-gatk – gatk.” [Online]. Available: https://gatk.broadinstitute.org/hc/en-us/articles/4410456501915-Functional-equivalence-in-DRAGEN-GATK

[109] “Benchmarking state-of-the-art secondary variant calling pipelines | by yih-chii hwang, ph.d. | dnanexus science frontiers |

medium," 8 2023. [Online]. Available: https://medium.com/dnanexus/benchmarking-state-of-the-art-secondary-variant-calling-pipelines-5472ca6bace7

[110] "Srp047086 : Study : Sra archive : Ncbi." [Online]. Available: https://trace.ncbi.nlm.nih.gov/Traces/index.html?view=study&acc=SRP047086

[111] "Seven bridges graf user guide." [Online]. Available: https://hello.sevenbridges.com/hubfs/Graph%20Files/GRAF_Technical_Guide_v1062020.pdf

[112] "Truth challenge v2: Calling variants from short and long reads in difficult-to-map regions - precisionfda challenge." [Online]. Available: https://precision.fda.gov/challenges/10/results

[113] "Precisionfda truth challenge – precisionfda." [Online]. Available: https://52.20.174.113/challenges/truth/results

[114] J. Jin, Z. Chen, J. Liu, H. Du, and G. Zhang, "Towards an accurate and robust analysis pipeline for somatic mutation calling," *Frontiers in Genetics*, vol. 13, p. 3298, 11 2022.

[115] L. de Schaetzen van Brienen, M. Larmuseau, K. V. der Eecken, F. D. Ryck, P. Robbe, A. Schuh, J. Fostier, P. Ost, and K. Marchal, "Comparative analysis of somatic variant calling on matched ff and ffpe wgs samples," 2020. [Online]. Available: https://doi.org/10.1186/s12920-020-00746-5

[116] Z. Chen, Y. Yuan, X. Chen, J. Chen, S. Lin, X. Li, and H. Du, "Systematic comparison of somatic variant calling performance among different sequencing depth and mutation frequency," *Scientific Reports 2020 10:1*, vol. 10, pp. 1–9, 2 2020. [Online]. Available: https://www.nature.com/articles/s41598-020-60559-5

[117] C. A. Garcia-Prieto, F. M. I.-J. Enez, A. Valencia, and E. Porta-Pardo, "Detection of oncogenic and clinically actionable mutations in cancer genomes critically depends on variant calling tools," 5 2022. [Online]. Available: https://doi.org/10.1093/bioinformatics/btac306

[118] Q. Li and K. Wang, "Intervar: Clinical interpretation of genetic variants by the 2015 acmg-amp guidelines," *American Journal of Human Genetics*, vol. 100, pp. 267–280, 2 2017. [Online]. Available: http://www.cell.com/article/S0002929717300046/fulltexthttp://www.cell.com/article/S0002929717300046/abstracthttps://www.cell.com/ajhg/abstract/S0002-9297(17)30004-6

[119] M. M. Li, C. E. Cottrell, M. Pullambhatla, S. Roy, R. L. Temple-Smolkin, S. A. Turner, K. Wang, Y. Zhou, and C. L. Vnencak-Jones, "Assessments of somatic variant classification using the association for molecular pathology/american society of clinical oncology/college of american pathologists guidelines: A report from the association for molecular pathology," *Journal of Molecular Diagnostics*, vol. 25, pp. 69–86, 2 2023.

[Online]. Available: http://www.jmdjournal.org/article/S1525157822003397/fulltexthttp://www.jmdjournal.org/article/S1525157822003397/abstracthttps://www.jmdjournal.org/article/S1525-1578(22)00339-7/abstract

[120] "Introduction to varsome's acmg and amp classifiers." [Online]. Available: https://docs.varsome.com/en/introduction-to-varsome-acmg

[121] Q. Li, Z. Ren, K. Cao, M. M. Li, K. Wang, and Y. Zhou, "Cancervar: An artificial intelligence–empowered platform for clinical interpretation of somatic mutations in cancer," *Science Advances*, vol. 8, p. 1624, 5 2022. [Online]. Available: https://www.science.org/doi/10.1126/sciadv.abj1624

[122] S. Tuteja, S. Kadri, and K. L. Yap, "A performance evaluation study: Variant annotation tools - the enigma of clinical next generation sequencing (ngs) based genetic testing," *Journal of Pathology Informatics*, vol. 13, p. 100130, 1 2022.

[123] G. Nicora, S. Zucca, I. Limongelli, R. Bellazzi, and P. Magni, "A machine learning approach based on acmg/amp guidelines for genomic variant classification and prioritization," *Scientific Reports 2022 12:1*, vol. 12, pp. 1–12, 2 2022. [Online]. Available: https://www.nature.com/articles/s41598-022-06547-3

[124] Y. Liu, W. S. Yeung, P. C. Chiu, and D. Cao, "Computational approaches for predicting variant impact: An overview from resources, principles to applications," *Frontiers in Genetics*, vol. 13, p. 2361, 9 2022.

[125] A. Muñoz-Barrera, L. A. Rubio-Rodríguez, A. D. de Usera, D. Jáspez, J. M. Lorenzo-Salazar, R. González-Montelongo, V. García-Olivares, and C. Flores, "From samples to germline and somatic sequence variation: A focus on next-generation sequencing in melanoma research," *Life 2022, Vol. 12, Page 1939*, vol. 12, p. 1939, 11 2022. [Online]. Available: https://www.mdpi.com/2075-1729/12/11/1939/htmhttps://www.mdpi.com/2075-1729/12/11/1939

[126] Varsome, "Varsome clinical," 2022. [Online]. Available: https://landing.varsome.com/varsome-clinical

[127] SOPHiA-GENETICS, "Sophia genetics - where others see data we see answers - where others see data we see answers," 2023. [Online]. Available: https://www.sophiagenetics.com/

[128] BITGENIA, "Bitgenia | trascending genomics," 2023. [Online]. Available: https://www.bitgenia.com/

[129] C. G. Cloud, "Cancer genomics cloud." [Online]. Available: https://www.cancergenomicscloud.org/#

[130] SevenBridges, “The seven bridges platform - seven bridges.” [Online]. Available: https://www.sevenbridges.com/platform/

[131] “About memoization (workreuse).” [Online]. Available: https://docs.cancergenomicscloud.org/docs/about-memoization

[132] “About spot/preemptible instances.” [Online]. Available: https://docs.cancergenomicscloud.org/docs/about-spot-instances

[133] G. Watts and A. J. Newson, “To offer or request? disclosing variants of uncertain significance in prenatal testing,” *Bioethics*, vol. 35, p. 900, 11 2021. [Online]. Available: /pmc/articles/PMC9291284//pmc/articles/PMC9291284/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC9291284/

[134] A. Kwong, C. Y. S. Ho, V. Y. Shin, C. H. Au, T. L. Chan, and E. S. K. Ma, “How does re-classification of variants of unknown significance (vus) impact the management of patients at risk for hereditary breast cancer?” *BMC Medical Genomics*, vol. 15, 12 2022. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9158111/

[135] “Fasta format.” [Online]. Available: http://bioinformatics.intec.ugent.be/MotifSuite/fastaformat.php

[136] I. Illumina, “Fastq files explained,” 10 2021. [Online]. Available: https://support.illumina.com/bulletins/2016/04/fastq-files-explained.html

[137] I. I.-S. quality scores, “Sequencing quality scores,” 2022. [Online]. Available: https://www.illumina.com/science/technology/next-generation-sequencing/plan-experiments/quality-scores.html

[138] J. Niu, D. Denisko, and M. M. Hoffman, “The browser extensible data (bed) format,” *The Browser Extensible Data (BED) format*, 1 2022. [Online]. Available: https://genome.ucsc.edu/FAQ/FAQformat.html

[139] EMBL-EBI, “1000 genomes | a deep catalog of human genetic variation,” 9 2021. [Online]. Available: https://www.internationalgenome.org/

[140] Y. Ma, “Rmvpfbam: Removing primers from bam files based on amplicon-based next-generation sequencing and cloud computing when analyzing personal genome data,” *Scientific Programming*, vol. 2021, 2021.

[141] M. Martin, “Cutadapt removes adapter sequences from high-throughput sequencing reads,” *EMBnet.journal*, vol. 17, pp. 10–12, 5 2011. [Online]. Available: https://journal.embnet.org/index.php/embnetjournal/article/view/200/479https://journal.embnet.org/index.php/embnetjournal/article/view/200

[142] SevenBridgesGenomics, “Seven_bridges_graf_user_guide,” *Seven Bridges GRAF Tools Workflows USER GUIDE*, 2021.

[143] B. Institute-SortSam, “Sortsam (picard) – gatk,” 2019. [Online]. Available: https://gatk.broadinstitute.org/hc/en-us/articles/360036510732-SortSam-Picard-

[144] S. Andrews, “Babraham bioinformatics - fastqc a quality control tool for high throughput sequence data,” 8 2019. [Online]. Available: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

[145] BroadInstitute, “Baserecalibrator – gatk,” 4 2020. [Online]. Available: https://gatk.broadinstitute.org/hc/en-us/articles/360036898312-BaseRecalibrator

[146] BroadInstitute-ApplyBQSR, “Applybqsr – gatk,” 2020. [Online]. Available: https://gatk.broadinstitute.org/hc/en-us/articles/360037055712-ApplyBQSR

[147] E. M. B. L. EMBL-EBI, “Variant identification and analysis | human genetic variation,” 2022. [Online]. Available: https://www.ebi.ac.uk/training/online/courses/human-genetic-variation-introduction/variant-identification-and-analysis/

[148] D. Benjamin, T. Sato, L. Lichtenstein, and M. Shand, “gatk/mutect.pdf at master · broadinstitute/gatk · github,” 2 2021. [Online]. Available: https://github.com/broadinstitute/gatk/blob/master/docs/mutect/mutect.pdf

[149] D. J. McCarthy, P. Humburg, A. Kanapin, M. A. Rivas, K. Gaulton, J. B. Cazier, and P. Donnelly, “Choice of transcripts and software has a large effect on variant annotation,” *Genome Medicine*, vol. 6, pp. 1–16, 3 2014. [Online]. Available: https://genomemedicine.biomedcentral.com/articles/10.1186/gm543

[150] K. Wang, M. Li, and H. Hakonarson, “Annovar: functional annotation of genetic variants from high-throughput sequencing data,” *Nucleic acids research*, vol. 38, 7 2010. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/20601685/

[151] “Per base sequence content.” [Online]. Available: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/4%20Per%20Base%20Sequence%20Content.html

[152] “Wes garvan na12878 hg001 hiseq exome details.” [Online]. Available: https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/NA12878/Garvan_NA12878_HG001_HiSeq_Exome/Garvan_NA12878_HG001_HiSeq_Exome.README

[153] “Wxs of homo sapiens: Chinese trio son - sra - ncbi.” [Online]. Available: https://www.ncbi.nlm.nih.gov/sra?term=SRR2962693

[154] "Wxs of homo sapiens: Aj trio son - sra - ncbi." [Online]. Available: https://www.ncbi.nlm.nih.gov/sra?term=SRR2962669

[155] "Nextera rapid capture exomes." [Online]. Available: https://support.illumina.com/content/dam/illumina-marketing/documents/products/datasheets/datasheet_nextera_rapid_capture_exome.pdf

[156] "Sureselect human all exon v5." [Online]. Available: https://www.agilent.com/cs/library/datasheets/public/AllExondatasheet-5990-9857EN.pdf

[157] W. J. Huss, Q. Hu, S. T. Glenn, K. J. Gangavarapu, J. Wang, J. D. Luce, P. K. Quinn, E. A. Brese, F. Zhan, J. M. Conroy, G. Paragh, B. A. Foster, C. D. Morrison, S. Liu, and L. Wei, "Comparison of sureselect and nextera exome capture performance in single-cell sequencing." [Online]. Available: https://www.fluidigm.com/

[158] K. H. Wong, W. Ma, C. Y. Wei, E. C. Yeh, W. J. Lin, E. H. Wang, J. P. Su, F. J. Hsieh, H. J. Kao, H. H. Chen, S. K. Chow, E. Young, C. Chu, A. Poon, C. F. Yang, D. S. Lin, Y. F. Hu, J. Y. Wu, N. C. Lee, W. L. Hwu, D. Boffelli, D. Martin, M. Xiao, and P. Y. Kwok, "Towards a reference genome that captures global genetic diversity," *Nature Communications 2020 11:1*, vol. 11, pp. 1–11, 10 2020. [Online]. Available: https://www.nature.com/articles/s41467-020-19311-w

[159] A. Belkadi, A. Bolze, Y. Itan, A. Cobat, Q. B. Vincent, A. Antipenko, L. Shang, B. Boisson, J. L. Casanova, and L. Abel, "Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 112, pp. 5473–5478, 4 2015. [Online]. Available: https://www.pnas.org/doi/abs/10.1073/pnas.1418631112

[160] D. Shigemizu, A. Fujimoto, S. Akiyama, T. Abe, K. Nakano, K. A. Boroevich, Y. Yamamoto, M. Furuta, M. Kubo, H. Nakagawa, and T. Tsunoda, "A practical method to detect snvs and indels from whole genome and exome sequencing data," *Scientific Reports*, vol. 3, 2013. [Online]. Available: /pmc/articles/PMC3703611//pmc/articles/PMC3703611/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC3703611/

[161] "engenome." [Online]. Available: https://www.engenome.com/news/wes_wgs/

APÉNDICE A

APÉNDICE DE NGS

A.1. WES

Los estudios de secuenciación de exoma completo (WES, del inglés Whole Exome Sequencing) comprenden al análisis de las zonas codificantes del genoma, es decir, los exones. Las zonas codificantes se caracterizan por poder transcribirse y luego traducirse a proteínas. El estudio de estas zonas se ha vuelto muy importante debido a que una variante en ellas puede significar un cambio en la proteína resultante [21, 104].

El genoma contiene también regiones intrónicas que rodean a los exones, pero estas regiones no estarán presentes en el producto proteico final pues se ven descartadas en el proceso de formación de las proteínas.

APÉNDICE B

APÉNDICE DEL FLUJO DE ANÁLISIS

B.1. Formato de los archivos utilizados

A continuación se listarán y explicarán los formatos de los distintos archivos que se han utilizado durante la realización del trabajo.

B.1.1. Archivos FASTA o FA

En bioinformática, el formato FASTA es un formato basado en texto para representar secuencias de ADN, en el que los pares de bases se representan mediante un código de una sola letra [A,C,G,T,N] donde A=adenosina, C=citosina, G= Guanina, T=Timidina y N=cualquiera de A,C,G,T. El formato también permite que los nombres de las secuencias y los comentarios precedan a las secuencias (ver figura B.1).

```
1  >gi|568336023|gb|CM000663.2| Homo sapiens chromosome 1, GRCh38 reference primary assembly
2  CCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTA
3  ACCCTAACCCTAACCCTAACCCTAACCCAACCCTAACCCTAACCCTAACCCTAACCCTAACCCC
4  TAACCCTAACCCTAACCCTAACCCTAACCTAACCCTAACCCTAACCCTAACCCTAACCCTAACC
5  CTAACCCTAACCCCTAACCCTAACCCTAAACCCTAAACCCTAACCCTAACCCTAACCCTAACCC
6  CAACCCCAACCCCAACCCCAACCCCAACCCCAACCCTAACCCCTAACCCTAACCCTAACCCTAAC
7  CCTAACCCTAACCCTAACCCTAACCCTAACCCCTAACCCTAACCCTAACCCTAACCCTAACCC
8  TAACCCTAACCCCTAACCCTAACCCTAACCCTAACCCTCGCGGTACCCTCAGCCGGCCCGCCCGCCCGGG
9  TCTGACCTGAGGAGAACTGTGCTCCGCCTTCAGAGTACCACCGAAATCTGTGCAGAGGACAACGCAGCTC
10 CGCCCTCGCGGTGCTCTCCGGGTCTGTGCTGAGGAGAACGCAACTCCGCCGTTGCAAAGGCGCGCCGCGC
11 CGGCGCAGGCGCAGAGAGGCGCGCCGCGCCGGCGCAGGCGCAGAGAGGCGCGCCGCGCCGGCGCAGGCGC
12 AGAGAGGCGCGCCGCGCCGGCGCAGGCGCAGAGAGGCGCGCCGCGCCGGCGCAGGCGCAGAGAGGCGCGC
13 CGCGCCGGCGCAGGCGCAGACACATGCTAGCGCGTCGGGGTGGAGGCGTGGCGCAGGCGCAGAGAGGCGC
14 GCCGCGCCGGCGCAGGCGCAGAGACACATGCTACCGCGTCCAGGGGTGGAGGCGTGGCGCAGGCGCAGAG
15 AGGCGCACCGCGCCGGCGCAGGCGCAGAGACACATGCTAGCGCGTCCAGGGGTGGAGGCGTGGCGCAGGC
16 GCAGAGACGCAAGCCTACGGGCGGGGGTTGGGGGGGCGTGTGTTGCAGGAGCAAAGTCGCACGGCGCCGG
17 GCTGGGGCGGGGGGAGGGTGGCGCCGTGCACGCGCAGAAACTCACGTCACGGTGGCGCGGCGCAGAGACG
18 GGTAGAACCTCAGTAATCCGAAAAGCCGGGATCGACCGCCCCTTGCTTGCAGCCGGGCACTACAGGACCC
19 GCTTGCTCACGGTGCTGTGCCAGGGCGCCCCCTGCTGGCGACTAGGGCAACTGCAGGGCTCTCTTGCTTA
20 GAGTGGTGGCCAGCGCCCCCTGCTGGCGCCGGGGCACTGCAGGGCCCTCTTGCTTACTGTATAGTGGTGG
```

Figura B.1: *Una secuencia en formato FASTA comienza con una descripción de identificador de una sola línea, seguida de líneas de datos de secuencia de ADN. La línea de descripción del identificador se distingue de los datos de secuencia por un símbolo mayor que ('>') en la primera columna. La palabra que sigue al símbolo ».es el identificador de la secuencia, y el resto de la línea es una descripción (opcional) separada del identificador por un espacio en blanco o un tabulador. Los datos de la secuencia comienzan en la línea siguiente a la línea de texto y terminan si aparece otra línea que comienza con »"; esto indica el comienzo de otra secuencia.[135]*

Los genomas de referencia utilizados en este trabajo están en este tipo de archivos, o en su forma comprimida, .fasta.gz.

B.1.2. Archivos FASTQ o FQ

La tecnología de secuenciación de Illumina utiliza la química de generación de grupos y secuenciación por síntesis (sequencing by synthesis o SBS) para secuenciar millones o miles de millones de grupos en una celda de flujo, según la plataforma de secuenciación. Durante la química SBS, para cada grupo, el software de análisis en tiempo real (RTA, del inglés real time analysis) del instrumento realiza y almacena llamadas de base para cada ciclo de secuenciación. RTA almacena los datos de llamadas base en forma de archivos de llamadas base individuales (o BCL). Cuando se completa la secuenciación, las llamadas base en los archivos BCL deben convertirse en datos de secuencia. Este proceso se llama conversión de BCL a FASTQ.

Un archivo FASTQ es un archivo de texto que contiene los datos de secuencia de los grupos que pasan el filtro en una celda de flujo. Si las muestras se multiplexaron, el primer paso en la generación de archivos FASTQ es la demultiplexación. La demultiplexación asigna grupos a una muestra, en función de la/s secuencia/s de índice del grupo. Después de la demultiplexación, las secuencias ensambladas se escriben en archivos FASTQ por muestra. Si las muestras no se multiplexaron, el paso de demultiplexación no se produce y, para cada carril de celda de flujo, todos los grupos se asignan a una sola muestra. [136]

Para una ejecución de lectura única, se crea un archivo FASTQ de Lectura 1 (Read 1 o R1) para cada muestra por carril de la celda de flujo. Para una ejecución de extremo emparejado

(paired-end), se crea un archivo FASTQ R1 y uno de lectura 2 (Read 2 o R2) para cada muestra de cada carril. Los archivos FASTQ se comprimen y crean con la extensión .fastq.gz. [136]

Para cada grupo que pasa el filtro, se escribe una sola secuencia en el archivo R1 FASTQ de la muestra correspondiente y, para una ejecución de dos extremos, también se escribe una sola secuencia en el archivo R2 FASTQ de la muestra. En cuanto al aspecto del archivo final, se distingue la lectura y la misma consta de 4 líneas:

1. Un identificador de secuencia con información sobre el experimento de secuenciación y el grupo. El contenido exacto de esta línea varía según el software de conversión de BCL a FASTQ utilizado.

2. La secuencia detectada: A, C, T, G y N (para designar un nucleótido que no ha podido ser detectado con seguridad).

3. Un separador, que es simplemente un signo más (+).

4. Las puntuaciones de calidad de la llamada base. [136]

Calidad de los archivos

Las puntuaciones de calidad de secuenciación miden la probabilidad de que una base se llame incorrectamente. Con la tecnología SBS, a cada base de una lectura se le asigna una puntuación de calidad que se define mediante la siguiente ecuación:

$$Q = -10\log(e)$$

Donde e es la probabilidad estimada de que la llamada base sea incorrecta. [137]

Las puntuaciones Q más altas indican una menor probabilidad de error y las puntuaciones Q más bajas pueden dar lugar a que una parte significativa de las lecturas no se pueda utilizar. Por otra parte, esta últimas también pueden dar lugar a un aumento de las llamadas de variantes de falsos positivos, lo que da lugar a conclusiones inexactas. Una puntuación de calidad de 20 ($Q20$) representa una tasa de error de 1 en 100 (lo que significa que cada lectura de secuenciación de 100 pb puede contener un error), con una precisión de llamada correspondiente del 99%. [137]

Cuando la calidad de la secuenciación llegue a $Q30$, prácticamente todas las lecturas serán perfectas, sin errores ni ambigüedades. Es por eso que $Q30$ se considera un referente de calidad en la secuenciación de próxima generación (NGS). [137]

B.1.3. Archivos BAM

Un archivo BAM (.bam) es la versión binaria comprimida de un archivo SAM que se utiliza para representar secuencias alineadas de hasta 128 Mb (128.000.000 pb).

Los archivos BAM usan el formato de nombre de archivo de NombreDeMuestra_S#.bam, donde # es el número de muestra determinado por el orden en que se enumeran las muestras para la ejecución. [81]

Los archivos BAM contienen una sección de encabezado y una sección de alineación. El encabezado contiene información sobre todo el archivo, como el nombre de la muestra, la longitud de la muestra y el método de alineación. Las alineaciones en la sección de alineaciones están asociadas con información específica en la sección de encabezado, y aquí aprarecen el nombre de lectura, la secuencia de lectura, la calidad de lectura, la información de alineación y las etiquetas personalizadas. El nombre de lectura incluye el cromosoma, la coordenada de inicio, la calidad de la alineación y la cadena del descriptor de coincidencia. [81]

La sección de alineaciones incluye también la siguiente información para cada par de lectura:

- RG: grupo de lectura, que indica el número de lecturas para una muestra específica.
- BC: Etiqueta de código de barras, que indica el ID de la muestra demultiplexada asociada a la lectura.
- SM: calidad de alineación de un solo extremo.
- AS: calidad de alineación de extremos emparejados.
- NM: Editar etiqueta de distancia, que registra la distancia de Levenshtein [1] entre la lectura y la referencia.

Los archivos de índice BAM (BAI) actúan como una tabla de contenido externa y permite que los programas salten directamente a partes específicas del archivo BAM sin leer todas las secuencias. [81]

B.1.4. Archivos BED

El formato BED (.bed, del inglés Browser Extensible Data) es un formato de archivo de texto que se utiliza para almacenar regiones genómicas como coordenadas y anotaciones asociadas. Los datos se presentan en forma de columnas separadas por espacios o tabuladores. Este formato fue desarrollado durante el Proyecto Genoma Humano y luego adoptado por otros proyectos de secuenciación.

Una de las ventajas de este formato es la manipulación de coordenadas en lugar de secuencias de nucleótidos, lo que optimiza la potencia y el tiempo de cómputo a la hora de comparar todos o parte de los genomas. Además, su simplicidad facilita la manipulación y lectura (o análisis)

[1] La métrica de la distancia Levenshtein mide la diferencia entre dos cadenas. Es el número mínimo de ediciones de un solo carácter que se requieren para cambiar una cadena por otra.

de coordenadas o anotaciones utilizando lenguajes de procesamiento de texto y secuencias de comandos como Python, Ruby o Perl o herramientas más especializadas como BEDTools. [138]

B.1.5. Archivos VCF

El Formato de Llamado de Variantes (.vcf, del inglés Variant Call Format) es un archivo de texto que se usa en Bioinformática para almacenar variaciones de la secuencia de genes y su información. Este formato se ha desarrollado a la luz de los grandes proyectos de secuenciación del ADN y genotipado, como el Proyecto 1000 Genomas. [139]

Por otra parte, estos archivos están presentes también a la salida, en ellos se anotan las variantes encontradas y se apunta información adicional de estas.

B.2. Herramientas del flujo de análisis

El pipeline o flujo de análisis se divide en 5 pasos principales:

1. Procesamiento del archivo crudo de secuenciación FASTQ: TRIMMING, FASTQC
2. Alineamiento: BWA MEM y post alineamiento
3. Llamado de variantes
4. Anotación de variantes
5. Generación del reporte

A continuación se detallan las herramientas utilizadas para ejecutar cada paso. Las entradas y salidas de cada una figuran en las tablas B.2.2, B.2.3.2 y B.2.4.2.

B.2.1. Procesamiento del archivo crudo de secuenciación FASTQ

El objetivo de esta sección de pasos es obtener un archivo tipo BAM a partir de archivos FASTQ. Para esto, se utilizan las siguientes herramientas:

- Cutadapt
- BWA-MEM y Alineador GRAF
- Sort coordinate
- Mark duplicates
- Controles de calidad de archivos FASTQ y BAM

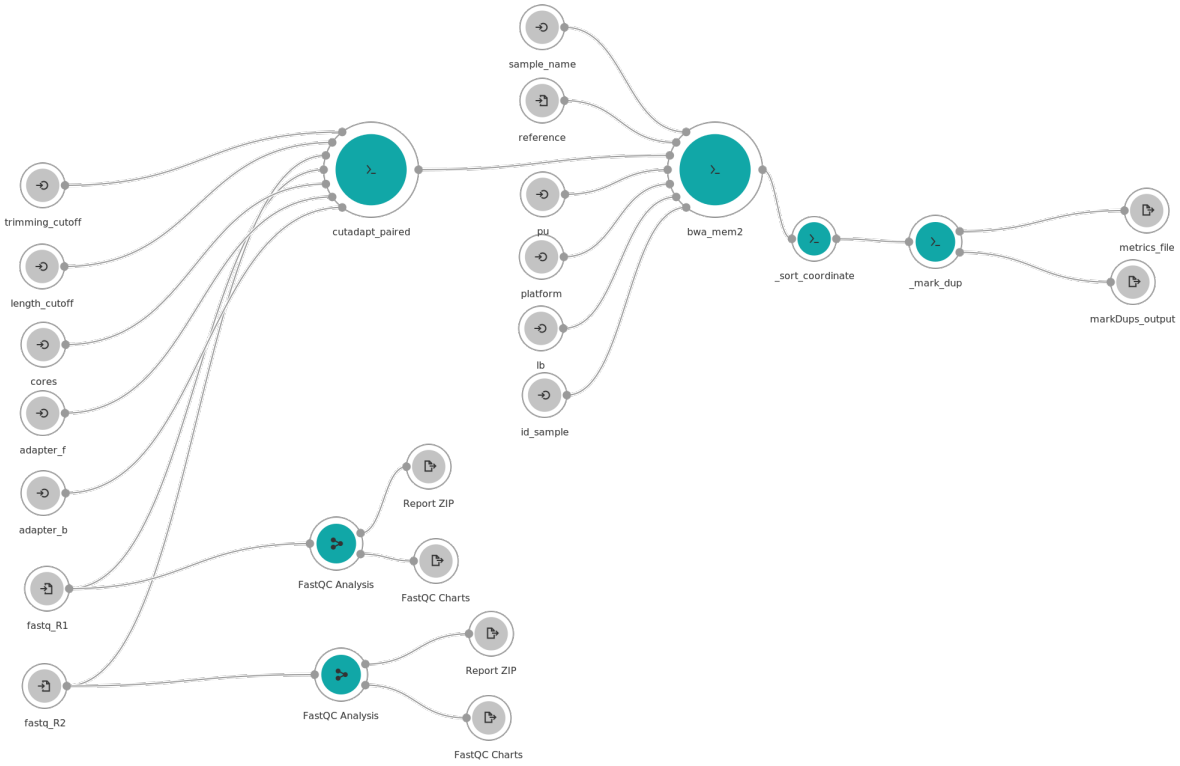


Figura B.2: *Sección "FASTQ to BAM"del procesamiento de los datos*

B.2.2. Resumen de entradas y salidas de cada herramienta

CUTADAPT-PAIRED	
INPUTS	OUTPUTS
■ 2 lecturas (FASTQ) ■ Adaptadores Forward ■ Adaptadores Reverse ■ Lecturas de la hebra forward (FASTQ) ■ Lecturas de la hebra reverse (FASTQ) ■ Calidad umbral ■ Longitud umbral	■ 2 lecturas cortadas (FASTQ)
RESUMEN	

Ante el archivo crudo obtenido de la secuenciación NGS, la herramienta elimina secuencias no deseadas (adaptadores, cebadores o primers y colas poli-A). Los adaptadores de los extremos 3' de ambas lecturas (reads) interfieren con los análisis río abajo como el alineamiento a la secuencia de referencia. Si no se eliminan los adaptadores, esto dará lugar a un mayor número de lecturas no alineadas, ya que las secuencias de los adaptadores son sintéticas y no ocurren naturalmente en la secuencia genómica, con lo cual no tienen similaridad con el genoma de referencia. [74, 140] Por otra parte, le herramienta Cutadapt también es capaz de demultiplexar las lecturas de cada muestra (permite identificar a que paciente le pertenece cada lectura secuenciada). [141]

BWA-MEM

INPUTS	OUTPUTS
■ 2 lecturas (FASTQ) ■ 2 lecturas (FASTQ) ■ Referencia (FASTA) ■ Plataforma ■ Librería	■ BAM alineado

RESUMEN

Permite mapear secuencias de baja divergencia contra un genoma de referencia grande, como el genoma humano[2]. Este software consta de tres algoritmos: BWA-backtrack, BWA-SW y BWA-MEM[3]. BWA-MEM generalmente es más utilizado para consultas de alta calidad, ya que es más rápido y más preciso; además de presentar un mejor rendimiento para lecturas de Illumina[4] [77]

[2]La herramienta BWA-MEM produce un alineamiento local. [77]

[3]El primer algoritmo está diseñado para lecturas de secuencia de Illumina de hasta 100 pb, mientras que los dos restantes para secuencias más largas oscilan entre 70 pb y 1 Mbp. BWA-MEM y BWA-SW comparten características similares, como soporte de lectura larga y alineación dividida

[4]La opción MEM significa máxima coincidencia exacta (Maximal Exact Match). Cuando BWA-MEM comienza su proceso de alineación para una lectura en particular, en primer lugar busca una subcadena larga que coincida exactamente con la referencia y que no se pueda extender a una coincidencia más larga en ninguno de los extremos, es decir, encuentra un valor máximo de coincidencia exacta.

GRAF ALIGNER

INPUTS	OUTPUTS
■ 2 lecturas (FASTQ) ■ Referencia (FASTA) ■ Plataforma ■ Librería ■ Adaptadores Forward ■ Adaptadores Reverse	■ BAM alineado

RESUMEN

Permite una alineación precisa y la llamada de variantes utilizando un genoma de referencia gráfico, este aumenta la representación lineal del genoma humano (GRCh37/hg19 o GRCh38) con información adicional sobre la diversidad genética de varias poblaciones humanas; conteniendo SNPs, INDELs y otras variaciones estructurales observadas con una frecuencia significativa en un gran número de poblaciones. [142]

SORT CORDINATE

INPUTS	OUTPUTS
■ Archivo a ordenar (BAM)	■ Archivo BAM ordenado por coordenadas

RESUMEN

Ordena el archivo BAM de entrada por coordenadas, nombre de consulta o alguna otra propiedad del registro del archivo. [143] Para un archivo BAM ordenado por coordenadas, las alineaciones de lectura se ordenan utilizando el diccionario de secuencia de referencia (FAI). En otras palabras, esto significa que las alineaciones se agrupan primero por ID de secuencia de referencia (es decir, todas las alineaciones de un cromosoma aparecen en un bloque) y dentro del bloque para cada secuencia de referencia, las alineaciones se ordenan por la posición de inicio en esta secuencia.

MARK DUPLICATES	
INPUTS	OUTPUTS
▪ Archivo BAM ordenado por coordenadas	▪ Archivo BAM con duplicados identificados para cada lectura ▪ Archivo de métricas (TXT)

RESUMEN

Localiza y etiqueta lecturas duplicadas en un archivo BAM, donde las lecturas duplicadas se definen como provenientes de un solo fragmento de ADN. Pueden surgir duplicados durante la preparación de la muestra[5]. Después de recopilar las lecturas duplicadas, la herramienta diferencia las lecturas primarias y duplicadas mediante un algoritmo que clasifica las lecturas por las sumas de sus puntuaciones de calidad de cada base.

FASTQC/BAMQC	
INPUTS	OUTPUTS
▪ Archivo FASTQ/BAM a analizar	▪ Archivo ZIP con resultados acerca de la calidad ▪ Archivo HTML con resultados acerca de la calidad

RESUMEN

Analiza datos de secuencias de archivos FASTQ, BAM o SAM, y produce un conjunto de métricas y gráficos que ayudan a identificar problemas técnicos con los datos. Más generalmente, FASTQC brinda una idea general de qué tan bien funcionó el experimento de secuenciación. [144]

B.2.3. Etapa pre-llamado de variantes

Para poder utilizar la herramienta Haplotype Caller para el llamado de variantes, descripta más adelante, se debe ejecutar un flujo de análisis anterior. En este, se parte del BAM al BAM recalibrado que funciona como entrada de Haplotype Caller. Este flujo comprende a las herramientas:

1. Base Recalibration
2. Apply BQSR

[5]Por ejemplo, construcción de bibliotecas mediante PCR. Las lecturas duplicadas también pueden resultar de un solo grupo de amplificación, detectado incorrectamente como múltiples grupos por el sensor óptico del instrumento de secuenciación. Estos artefactos de duplicación se denominan duplicados ópticos.

B.2.3.1. BAM to BQSR

Esta sección comprende los pasos ejecutados para llegar al archivo BAM recalibrado. En la imagen B.3 se puede notar el flujo de análisis utilizado.

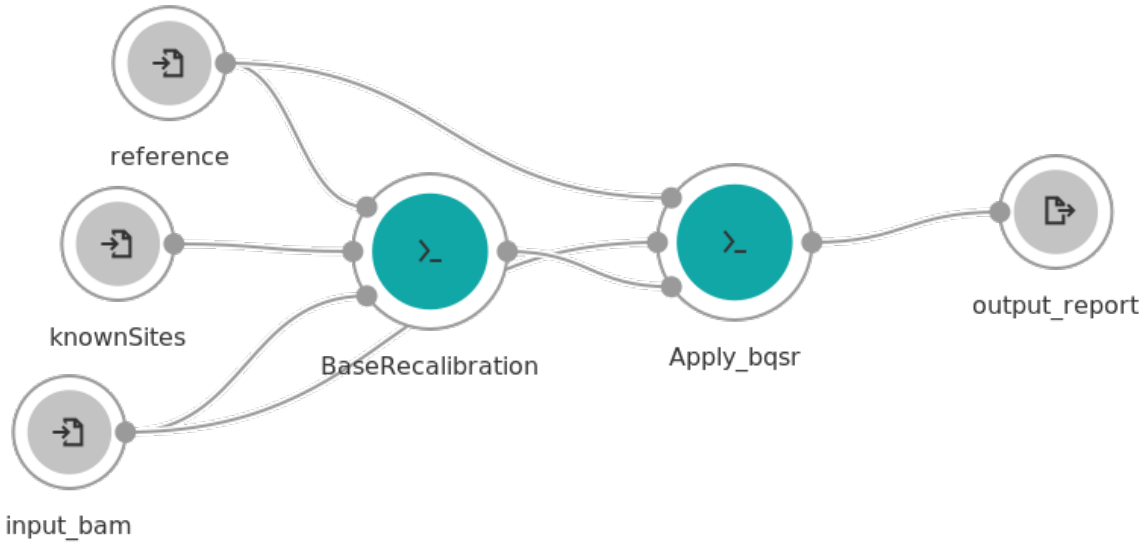


Figura B.3: *Sección BAM to BQSR del procesamiento de los datos*

B.2.3.2. Resumen de cada herramienta

BASE RECALIBRATION	
INPUTS	OUTPUTS
▪ Archivo BAM con duplicados identificados ▪ Genoma de referencia (FASTA) ▪ Archivos representen SNP y/o INDELs conocidos (VCF)	▪ Archivo BAM recalibrado ▪ Informe con tablas de recalibración

RESUMEN
Aplica aprendizaje automático para modelar errores de manera empírica y así ajustar los puntajes de calidad[6] a los nucleótidos alineados. [145]

[6]Por ejemplo se puede identificar que, para una ejecución determinada, siempre que se llame a dos nucleótidos A seguidos, la siguiente base llamada tenía una tasa de error un 1% más alta. Por lo tanto, cualquier llamada de base que venga después de AA en una lectura debería tener su puntaje de calidad reducido en un 1%. Esto se hace sobre varias

APPLY BQSR	
INPUTS	OUTPUTS
■ Archivo recalibrado (BAM+BAI) ■ Genoma de referencia ■ Archivo con tablas de recalibración	■ Archivo ajustado/recalibrado (BAM+BAI)

RESUMEN

Esta herramienta recalibra las calidades base de las lecturas de entrada según la tabla de recalibración producida por la herramienta BaseRecalibrator y genera un archivo BAM recalibrado. El objetivo de este procedimiento es corregir el sesgo sistemático que afecta la asignación de puntuaciones de calidad base por parte del secuenciador. [146]

B.2.4. Llamado de variantes

La llamada de variantes es el proceso mediante el cual se identifican variantes a partir de datos de secuencia.

El cuello de botella actual de la tecnología WGS y WES radica en los métodos de gestión de datos del sistema y el análisis complejo de llamadas de mutación en lugar de la secuenciación del genoma en sí, lo que plantea grandes desafíos para los investigadores genéticos. Una gran cantidad de pequeñas variantes genómicas, incluidos los SNPs y los INDELs, se detectan mediante diversas herramientas. [35]

B.2.4.1. Llamada de variantes somáticas y germinales

En la llamada de variante de línea germinal, el genoma de referencia es el estándar para la especie de interés. Esto permite identificar genotipos. Como la mayoría de los genomas son diploides, se espera ver que en cualquier locus dado, todas las lecturas tienen la misma base, lo que indica homocigosidad, o aproximadamente la mitad de todas las lecturas tienen una base y la mitad otra, lo que indica heterocigosidad. Una excepción a esto serían los cromosomas sexuales en los mamíferos machos [147]. El flujo de análisis realizado para el caso germinal se puede observar en la imagen B.4.

covariables diferentes (principalmente contexto de secuencia y posición en lectura, o ciclo) de una manera aditiva. Por lo tanto, es posible que la puntuación de calidad de la misma base aumente por una razón y disminuya por otra.

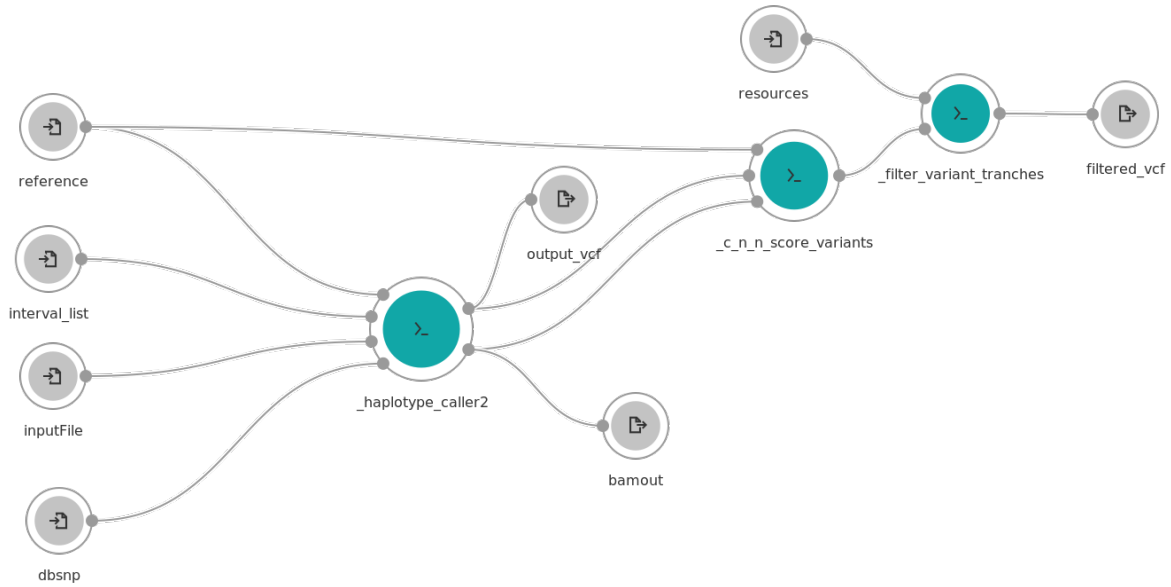


Figura B.4: *Flujo utilizado para el llamado de variantes germinal*

En la llamada variante somática, la referencia es un tejido relacionado del mismo individuo (además es de esperarse hallar mosaicismo entre células) [147] y por lo tanto, el llamado de variantes debe ser diferente.

B.2.4.2. Resumen de cada herramienta

HAPLOTYPE CALLER	
INPUTS	OUTPUTS
▪ Un archivo VCF con información de la línea germinal ▪ El archivo FASTA con el genoma de referencia utilizado ▪ Un archivo BED y un archivo VCF con información de la base de datos dbsnp, para SNPs	▪ Un archivo VCF con llamadas SNP e INDEL sin procesar y sin filtrar ▪ Un archivo BAM reensamblado

RESUMEN
Produce el llamado simultéaneo a los SNP e INDELs de la línea germinal a través del reensamblaje local de haplotipos. En otras palabras, cada vez que el programa encuentra una región que muestra signos de variación, descarta la información de mapeo existente y vuelve a ensamblar por completo

las lecturas en esa región. Esto permite que HaplotypeCaller sea más preciso al llamar a regiones que tradicionalmente son difíciles de llamar[7], y también hace que sea mucho mejor llamando INDELs que los llamadores basados en posición como UnifiedGenotyper [84]. Por otra parte, puede manejar organismos no diploides, así como datos de experimentos agrupados. Sin embargo, los algoritmos utilizados para calcular las probabilidades de variantes no se adaptan bien a las frecuencias alélicas extremas (en relación con la ploidía), por lo que utilizamos la herramienta Mutect2 para el descubrimiento de variantes somáticas.

CNN SCORE VARIANTS

INPUTS	OUTPUTS
■ Un archivo VCF con llamadas SNP e INDEL sin procesar y sin filtrar ■ Un archivo BAM (se usan los que genera Haplotype-Caller) ■ El archivo FASTA con el genoma de referencia utilizado	■ Archivo VCF anotado

RESUMEN

El nombre de este anotador proveniente de que utiliza una red neuronal convolucional (CNN) para generar la anotación de un VCF con puntuaciones. Esta herramienta transmite variantes y su contexto de referencia a un programa de Python, que evalúa una red neuronal preentrenada en cada variante. La red neuronal realiza circunvoluciones sobre la secuencia de referencia que rodea a la variante y combina esas características con un perceptrón multicapa en las anotaciones de la variante. Los modelos 2D convolucionan sobre lecturas alineadas, así como sobre la secuencia de referencia y las anotaciones de variantes. [85]

[7] Por ejemplo, cuando contienen diferentes tipos de variantes cercanas entre sí.

FILTER VARIANT TRANCHES

INPUTS	OUTPUTS
▪ Un archivo VCF (obtenido de CNN Score Variants) ▪ Archivos VCF con información de las bases de datos HapMap y dbsnp	▪ Archivo VCF filtrado

RESUMEN

Esta herramienta aplica el filtrado por tramos a un VCF en función de las puntuaciones de anotación en el campo INFO. La anotación en este caso proviene de CNNScoreVariants, ya que es la recomendada por GATK[8]. [86]. Los tramos se especifican en porcentaje de sensibilidad a las variantes en los archivos de entrada. El umbral de filtrado de tramo predeterminado para los SNP es 99,95 y para los INDEL es 99,4; ya que se ha visto que son los que maximizan la puntuación F1. [86]

HARD FILTERING

INPUTS	OUTPUTS
▪ Un archivo VCF con información de la línea germinal ▪ El archivo FASTA con el genoma de referencia utilizado	▪ Archivo VCF filtrado

RESUMEN

Esta herramienta es verdaderamente un flujo de análisis que está conformado por dos herramientas fundamentales:

- **GATK SelectVariants**

  Esta herramienta le permite seleccionar un subconjunto de variantes en función de varios criterios para facilitar ciertos análisis[9]. Los registros se filtran cambiando el valor en el campo FILTRO a algo que no sea APROBADO[10]. [**?**]

[8] Pero también podría hacerlo de otras tales como VQSR (VQSLOD) o cualquier otra herramienta de puntuación de variantes que agregue anotaciones numéricas en el campo INFO de un VCF.

[9] Como comparar y contrastar casos frente a controles, extraer loci variantes o no variantes que cumplen ciertos requisitos o solucionar algunos resultados inesperados, por nombrar algunos.

[10] Los registros filtrados se conservarán en la salida a menos que se solicite su eliminación en la línea de comando.

- **GATK VariantFiltration**

  GATK VariantFiltration se utiliza para filtrar variantes en un archivo VCF en función de las anotaciones dadas en ciertas columnas[11].

  Esta herramienta está diseñada para llamadas de variantes de "filtrado duro"[12] basadas en ciertos criterios. Los registros se filtran de forma estricta cambiando el valor en el campo FILTRO a algo que no sea APROBADO ("PASS"). Los registros filtrados se conservarán en la salida a menos que se solicite su eliminación en la línea de comando.

GRAF VARIANT CALLING AND FILTRATION

INPUTS	OUTPUTS
■ El archivo FASTA con el genoma de referencia utilizado ■ El archivo BAM con las secuencias alineadas ■ Un archivo VCF para la construcción del genoma de referencia gráfico ■ Un archivo BED con las regiones de interés	■ Archivo VCF filtrado

RESUMEN

Permite un análisis de variantes utilizando como base al genoma de referencia gráfico. En cuanto a la llamada de variantes, GRAF Variant Caller está diseñado para trabajar en conjunto con GRAF Aligner y detectar variantes pequeñas y estructurales con la ayuda de los datos disponibles sobre la variabilidad poblacional. [142]

Por otra parte, en cuanto a la filtración, se aplican criterios de filtrado duro a las variantes detectadas por GRAF Variant Caller.

[11] Como pueden ser INFO y FORMAT.

[12] El filtrado duro consiste en elegir umbrales específicos para una o más anotaciones y descartar cualquier variante que tenga valores de anotación por encima o por debajo de los umbrales establecidos.[88]

ANÁLISIS SOMÁTICO

MUTECT2

INPUTS	OUTPUTS
■ Un archivo VCF con datos germinales ■ Un archivo BED con las regiones de interés ■ Un archivo BAM obtenido del análisis germinal ■ Un archivo BAM obtenido del análisis somático ■ Un archivo FASTA conteniendo al genoma de referencia ■ Panel Of Normals en formato VCF	■ Archivo VCF con variantes sin filtrar ■ Archivo VCF con estadísticas ■ Archivo TAR con información de sesgos de orientación

RESUMEN

Produce el llamado de SNPs e INDELs a través del ensamblaje local de haplotipos, [91] utilizando la maquinaria basada en HaplotypeCaller[13]. Las variantes de la línea germinal son más sencillas, pues varían únicamente contra la referencia y asumen ploidía fija. Por otra parte, Mutect2 funciona principalmente contrastando la presencia o ausencia de evidencia de variación entre dos muestras (la tumoral y la normal emparejada) del mismo individuo.

Técnicamente hablando, las variantes somáticas son diferentes de la muestra control y de la referencia. Esto significa que si un sitio es variante en el control pero en la muestra somática coincide con el alelo de referencia, entonces no es una variante somática.

[13]Para profundizar en diferencias entre las dos herramientas, ver anexo B.3.

FILTER MUTECT	
INPUTS	OUTPUTS
■ Un archivo FASTA con el genoma de referencia utilizado ■ Un archivo VCF sin filtrar ■ Un archivo VCF con estadísticas	■ Archivo VCF filtrado

RESUMEN

Aplica filtros a la salida sin procesar de Mutect2. [92] Es necesario usar este filtrado ya que es capaz de aislar SNP e INDELs somáticos que cumplan alguna condición[14] que puede ser especificada por el usuario.

B.2.5. Anotación de variantes

La anotación de variantes es el proceso de asignar información funcional a las variantes de ADN. Hay muchos tipos diferentes de información que podrían asociarse con variantes, desde medidas de conservación de secuencias hasta predicciones sobre el efecto de una variante en la estructura y función de la proteína. [149]

ANNOVAR

Del inglés Annotate Variation (anotar variaciones), ANNOVAR es una herramienta de software impulsada por línea de comandos que se puede utilizar como una aplicación independiente. [118]

Se han desarrollado varias herramientas y bases de datos para ayudar a los laboratorios y médicos a comprender la importancia funcional de las variantes genéticas con respecto a sus efectos potenciales sobre los genes y las enfermedades. Por lo general, se dividen en varias categorías. Primero, varias herramientas de anotación, como ANNOVAR, SNPeff y VEP; entre otras [15], pueden predecir cómo las variantes genéticas afectan la estructura de la transcripción o las secuencias de codificación. Pueden clasificar las variantes en variantes intrónicas, intergénicas, de empalme y exónicas, y para las variantes exónicas, pueden calcular cómo se ven afectadas las secuencias de aminoácidos. En segundo lugar, para codificar variantes, una variedad de herramientas pueden predecir si la variante es perjudicial para la función o la estructura de la proteína mediante el uso de información evolutiva, el contexto dentro de la secuencia de la proteína y las propiedades bioquímicas. [150]

[14] Por ejemplo se puede filtrar por calidad de la lectura, longitud de la lectura, presencia o no en el PON, etc. [148]

[15] VAAST,6 SeattleSeq

B.2.5.1. Análisis germinal: Intervar

InterVar es un software controlado por línea de comandos escrito en Python y se puede usar como una aplicación independiente.

Funciona tomando archivos previamente anotados en formatos delimitados por tabuladores o archivos de entrada sin anotaciones en formato VCF o formato de entrada ANNOVAR, donde cada línea corresponde a una variante genética. Si los archivos de entrada no tienen anotaciones, InterVar llama a ANNOVAR para generar las anotaciones necesarias, [118] y así es capaz de asignar un criterio de patogenicidad a cada variante detectada.

B.2.5.2. Análisis somático: Cancervar

Esta herramienta es capaz de tomar los datos del llamado de variantes e interpretarlas siguiendo los criterios de AMP.

En primer lugar, la herramienta CancerVar toma archivos preanotados o archivos de entrada no anotados en formato VCF o formato de entrada ANNOVAR, donde cada línea corresponde a una variante genética; CancerVar llamará a ANNOVAR para generar las anotaciones necesarias. En el resultado, cada variante se asignará como:

- Nivel I: Fuerte
- Nivel II: Potencial
- Nivel III: Incierto
- Nivel IV: Benigno [93]

B.3. Puntos técnicos que resaltan las diferencias entre Mutect2 y Haplotype Caller

- Mutect2 no puede calcular la confianza de referencia, que es una característica de HaplotypeCaller que es clave para producir VCF. Como resultado, actualmente no hay forma de realizar llamadas conjuntas para el descubrimiento de variantes somáticas.
- Debido a que un conjunto de llamadas somáticas se basa en un solo individuo en lugar de una cohorte, las anotaciones en la columna INFO de un VCF de Mutect2 solo se refieren a los alelos ALT y no incluyen valores para el alelo REF. Esto difiere de un conjunto de llamadas de cohorte de línea germinal, en el que las anotaciones en el campo INFO generalmente se derivan de datos relacionados con todos los alelos observados, incluida la referencia.
- Mientras que HaplotypeCaller se basa en una suposición de ploidía fija para calcular las probabilidades de genotipo, que son la base de las probabilidades de genotipo (PL); Mutect2

permite variar la ploidía en forma de fracciones de alelos para cada variante. A menudo se observan fracciones de alelos variables dentro de una muestra de tumor debido a la pureza fraccional, los subclones múltiples[16] y la variación del número de copias.

- Mutect2 también se diferencia de HaplotypeCaller en que puede aplicar varios filtros previos a sitios y alelos según el uso de un recurso normal emparejado, un panel de normales (PoN) y una variante de población común que contiene frecuencias específicas de alelo.

[16] Los subclones múltiples son células tumorales en una muestra de tumor o biopsia que contienen una mutación particular.

APÉNDICE C

APÉNDICE DE RESULTADOS

C.1. Detalle sobre la información obtenida del llamado y filtrado de variantes

C.1.1. Selección de columnas importantes obtenidas de la anotación de variantes

En la tabla C.1 se muestra la información obtenida para cada una de las variantes en el CSV luego de correr el flujo de análisis.

Característica	Definición
Chr	Cromosoma en la que se encuentre la variante somática
Start	Posición inicial en el cromosoma
End	Posición final en el cromosoma
Ref	Alelo del genoma de referencia
Alt	Alelo alternativo que presenta la variante
Qual	La probabilidad de que exista un polimorfismo REF/ALT en este sitio dados los datos de secuenciación
FILTER	Si paso el filtro de Mutect2
CONTQ	Cualidades del alelo ALT de la escala de Phred que no se deben a la contaminación
DP	Profundidad aproximada de lectura
ECNT	Número de eventos en este haplotipo
GERMQ	Cualidades del alelo ALT de la escala de Phred para aquellas variantes que no son germinales
MBQ	Calidad media de la base nucleotídica

MFRL	Longitud media del fragmento
MMQ	Calidad media de mapeo
MPOS	Distancia media del final de la lectura
NALOD	Logaritmo negativo de base 10 de que sea artefacto en un tejido normal con la misma fracción alélica que el tumor.
NLOD	Logaritmo de base 10 de la probablidad de la proporción diploide heterocigota u homocigota del genotipo ALT
ACMG Veredict	Veredicto de ACMG
ACMG PVS1	Puntaje del criterio de clasificación: evidencia muy fuerte
ACMG PS1	Puntaje del criterio de clasificación: evidencia fuerte 1
ACMG PS2	Puntaje del criterio de clasificación: evidencia fuerte 2
ACMG PS3	Puntaje del criterio de clasificación: evidencia fuerte 3
ACMG PS4	Puntaje del criterio de clasificación: evidencia fuerte 4
ACMG PS5	Puntaje del criterio de clasificación: evidencia fuerte 5
ACMG PM1	Puntaje del criterio de clasificación: evidencia moderada 1
ACMG PM2	Puntaje del criterio de clasificación: evidencia moderada 2
ACMG PM3	Puntaje del criterio de clasificación: evidencia moderada 3
ACMG PM4	Puntaje del criterio de clasificación: evidencia moderada 4
ACMG PM5	Puntaje del criterio de clasificación: evidencia moderada 5
ACMG PM6	Puntaje del criterio de clasificación: evidencia moderada 6
ACMG PM7	Puntaje del criterio de clasificación: evidencia moderada 7
ACMG PP1	Puntaje del criterio de clasificación: evidencia patogénica "supporting"1
ACMG PP2	Puntaje del criterio de clasificación: evidencia patogénica "supporting"2
ACMG PP3	Puntaje del criterio de clasificación: evidencia patogénica "supporting"3
ACMG PP4	Puntaje del criterio de clasificación: evidencia patogénica "supporting"4
ACMG PP5	Puntaje del criterio de clasificación: evidencia patogénica "supporting"5
ACMG PP6	Puntaje del criterio de clasificación: evidencia patogénica "supporting"6
ACMG BA1	Puntaje del criterio de clasificación: evidencia benigna "standalone"
ACMG BS1	Puntaje del criterio de clasificación: evidencia benigna fuerte 1
ACMG BS2	Puntaje del criterio de clasificación: evidencia benigna fuerte 2
ACMG BS3	Puntaje del criterio de clasificación: evidencia benigna fuerte 3
ACMG BS4	Puntaje del criterio de clasificación: evidencia benigna fuerte 4
ACMG BS5	Puntaje del criterio de clasificación: evidencia benigna fuerte 5
ACMG BP1	Puntaje del criterio de clasificación: evidencia benigna "supporting"1
ACMG BP2	Puntaje del criterio de clasificación: evidencia benigna "supporting"2
ACMG BP3	Puntaje del criterio de clasificación: evidencia benigna "supporting"3
ACMG BP4	Puntaje del criterio de clasificación: evidencia benigna "supporting"4
ACMG BP5	Puntaje del criterio de clasificación: evidencia benigna "supporting"5

ACMG BP6	Puntaje del criterio de clasificación: evidencia benigna "supporting"6
ACMG BP7	Puntaje del criterio de clasificación: evidencia benigna "supporting"7
ACMG BP8	Puntaje del criterio de clasificación: evidencia benigna "supporting"8
PON	Sitio encontrado en la muestra germinal
POPAF	Logaritmo negativo de base 10 de las frecuencias alélicas de población de los alelos ALT
SEQQ	Calidad en escala de Phred de que los alelos no son errores de secuenciación
STRANDQ	Calidad en escala de Phred de que sea un artefacto de bias de la hebra
TLOD	Razón del Logaritmo de base 10 de la probabilidad de que una variante exista o no.
GT Normal sample	Genotipo de la muestra de tejido germinal
AD Normal sample	Produndidades alélicas para los alelos REF y ALT de la muestra de tejido germinal
AF Normal sample	Fracciones alélilcas del alelo ALT en el tumor de la muestra de tejido germinal
DP Normal sample	Produndidad aproximada de lectura de la muestra de tejido germinal (se filtran las lecturas con MQ=255 o con malos compañeros)
F1R2 Normal sample	Cantidad de lecturas en la orientación del par F1R2 que respaldan cada alelo de la muestra de tejido germinal
F2R1 Normal sample	Cantidad de lecturas en la orientación del par F2R1 que respaldan cada alelo de la muestra de tejido germinal
PGT Normal sample	Información de haplotipo de fase física de la muestra de tejido germinal, describiendo como los alelos ALT están en fase en relación entre sí.
PID Normal sample	Información de ID de fase física de la muestra de tejido germinal, donde cada ID único dentro de una muestra determinada (pero no entre muestras) conecta registros dentro de un grupo de fase
PS Normal sample	Conjunto de fase de la muestra de tejido germinal (típicamente la posición de la primera variante en el conjunto)
SB Normal sample	Estadísticas de componentes por muestra de tejido germinal que comprenden la prueba exacta de Fisher para detectar el sesgo de la hebra.
GT Tumor sample	Genotipo de la muestra de tejido tumoral
AD Tumor sample	Produndidades alélicas para los alelos REF y ALT de la muestra de tejido tumoral
AF Tumor sample	Fracciones alélilcas del alelo ALT en el tumor de la muestra de tejido tumoral
DP Tumor sample	Produndidad aproximada de lectura de la muestra de tejido tumoral (se filtran las lecturas con MQ=255)

F1R2 Tumor sample	Cantidad de lecturas en la orientación del par F1R2 que respaldan cada alelo de la muestra de tejido tumoral
F2R1 Tumor sample	Cantidad de lecturas en la orientación del par F2R1 que respaldan cada alelo de la muestra de tejido tumoral
PGT Tumor sample	Información de haplotipo de fase física de la muestra de tejido tumoral, describiendo como los alelos ALT están en fase en relación entre sí.
PID Tumor sample	Información de ID de fase física de la muestra de tejido tumoral, donde cada ID único dentro de una muestra determinada (pero no entre muestras) conecta registros dentro de un grupo de fase
PS Tumor sample	Conjunto de fase de la muestra de tejido tumoral (típicamente la posición de la primera variante en el conjunto)
SB Tumor sample	Estadísticas de componentes por muestra de tejido tumoral que comprenden la prueba exacta de Fisher para detectar el sesgo de la hebra.
Region	Region del gen en que se encuentra la variante
Variant type	Tipo de variante
HGVS (AAChange)	Cambio aminoacídico utilizando la nomenclatura HGVS
HGVS Ensembl (AAChange)	cambio aminoacídico utilizando la nomenclatura de Ensembl y HGVS
RS ID	Identificador RS para variantes SNP
Gene symbol	Símbolo del gen
Ensembl Gene	Identificador de Ensembl del gen
UCSC Gene	Identificador de UCSC del gen
Region UCSC Gene	Se indica la región del gen utilizando los identificadores de UCSC
Region Ensembl Gene	Se indica la región del gen utilizando los identificadores de Ensembl
Exonic variant function Ensembl	Función en Ensembl de la variante exónica
Exonic variant function UCSC	Función en UCSC de la variante exónica
GeneDetail RefSeq	Detalle del gen en RefSeq
GeneDetail Ensembl	Detalle del gen en Ensembl
GeneDetail UCSC	Detalle del gen en UCSC
AAChange UCSC	Cambio aminoacídico utilizando la nomenclatura UCSC
AMP Veredict	Veredicto de AMP
AMP CBP_1	Evidencia terapeútica: Aprobado por la FDA o en fase de investigación con pruebas sólidas
AMP CBP_2	Evidencia diagnóstica: En guía profesional o evidencia reportada con consenso

AMP CBP_3	Evidencia Prognóstica: En guía profesional o evidencia reportada con consenso
AMP CBP_4	Tipo de mutación: activación, LOF (missense, nonsense, indel, splicing), CNV, fusiones.
AMP CBP_5	Frecuencias de las variantes: principalmente mosaicas
AMP CBP_6	Potencialente germinal: Mayormente no mosaica
AMP CBP_7	Bases de datos poblacionales: Frecuencia alélica menor (MAF) ausente o extremadamente baja.
AMP CBP_8	Bases de datos germinales: puede estar presente en HGMD/ClinVar
AMP CBP_9	Bases de datos somáticas: principalmente presente en COSMIC, My Cancer Genome, TCGA
AMP CBP_10	Predictivo de SIFT, PolyPhen2, MutationTaster, CADD, MetaSVM, MetaLR, FATHMM, GERP++_RS
AMP CBP_11	Vías de señalización biológicasÑ implicado en vías de señalización patogénicas o asociadas con una enfermedad
AMP CBP_12	Publicaciones: Evidencia convincente de un estudio funcional, población u otro.
Zygocity	Cigocidad: el grado de similitud o disimilitud de las secuencias de ADN en segmentos de codificación específicos, o genes, en los cromosomas homólogos de un cigoto u óvulo fertilizado.
Transcript	Transcripto resultante del gen
Ensembl transcript id	Id de Ensembl del Transcripto resultante del gen
Clinvar	Clasificación de la variante de Clinvar
ExAC exome frequencies of all variants	Frecuencias Exomicas de todas las variantes de la base de datos ExAC
GnomAD exome frequencies of all variants	Frecuencias Exomicas de todas las variantes de la base de datos GnomAD
Esp6500siv2 exome frequencies of all variants	Frecuencias Exomicas de todas las variantes de la base de datos ExAC
1000g2015aug exome frequencies of all variants	Frecuencias Exomicas de todas las variantes de la base de datos 1000g2015aug
Pathway	Vías de señalización relacionadas con la variante
Therap_list	Lista de terapias dirigidas
Diag_list	Lista de diagnósticos
Prog_list	Lista de evidencia prognóstica
cosmic91 Id	Id de la mutación en la base de datos de mutaciones somáticas cosmic91
cosmic91 Ocurrence	ocurrencia en la base de datos de mutaciones somáticas

ICGC Id	Id de la variante en ICGC
ICGC Occurrence	Ocurrencia de la variante en ICGC
Allele frequency ExAC_ALL	Frecuencia alélica de la base datos ExAC considerando todas las razas.
Allele frequency ExAC_AFR	Frecuencia alélica para la población africana de la base datos ExAC
Allele frequency ExAC_AMR	Frecuencia alélica para la población americana de la base datos ExAC
Allele frequency ExAC_EAS	Frecuencia alélica en muestras de Asia oriental de la base datos ExAC
Allele frequency ExAC_FIN	Frecuencia alélica en muestras finlandesas de la base datos ExAC
Allele frequency ExAC_NFE	Frecuencia alélica de muestras europeas no finlandesas de la base datos ExAC
Allele frequency ExAC_OTH	Frecuencia alélica para la población del Sur de Asia de la base datos ExAC
Allele frequency ExAC_SAS	Frecuencia alélica de la base datos ExAC para otras poblaciones
CADD raw score	puntuación de CADD para la predicción funcional de una SNP. Mientras más grande el puntaje, mas probable que la SNP tenga un efecto dañino.
CADD scaled score	puntuación de CADD escalada.
SIFT score	Puntaje SIFT (va desde cero a uno). Mientras menor sea la puntuación, es más probable que tenga un efecto dañino.
SIFT categorical prediction	Prediccion categórica de SIFT, puede clasificarse en: dañina (D) o tolerada (T).
SIFT_pred	Predicción de SIFT: dañina (D) si el puntaje SIFT es menor a 0,05 y tolerada (T) si es mayor.
Polyphen2_HDIV score	Puntuación de Polyphen2 basada en HumDiv (hdiv$_prob$)
Polyphen2_HDIV pred	Predicción de Polyphen2 basada en el puntaje anterior. Puede ser: probablemente dañina (D), posiblemente dañina (P) o benigna (B)
Polyphen2_HVAR score	Puntuación de Polyphen2 basada en HumVar (hvar_prob).
Polyphen2_HVAR pred	Predicción de Polyphen2 basada en el puntaje anterior. Puede ser: probablemente dañina (D), posiblemente dañina (P) o benigna (B).
LRT_score	Puntaje del test de la prueba de la razón de similaridad (LRT): valores más bajos significa que la variante es más perjudicial.
LRT_pred	Predicción a partir del puntake LRT: puede ser dañina (D), neutral (N) o desconocida (U).
MutationTaster_score	puntaje de Mutation Taster: valores altos significa que es más perjudicial.

MutationTaster_pred	Predicción MutationTaster: puede ser causante automática de enfermedad (A), causante de enfermedad (D), polimorfismo (N) o polimorfismo automático (P).
MutationAssessor score	Puntaje de Mutation Assessor: presenta puntajes más altos cuando la variante es más perjudicial.
MutationAssessor pred	Predicción de Mutation Assesor del impacto funcional de las sustituciones de aminoácidos en las proteínas causados por las variantes: H (alto), M (medio), L (bajo) o N (neutral). H/M significa que es funcional y L/M, no funcional.
FATHMM_score	Puntaje de FATHMM: presenta valores más bajos cuando la variante es más prejudicial.
FATHMM_pred	Predicción de FATHMM, puede clasificarse en D o T.
PROVEAN_score	Puntaje de PROVEAN: presenta valores más altos cuando la variante es más perjudicial.
PROVEAN_pred	Predicción de PROVEAN para determinar si una sustitución aminoacídica tiene impacto en la función biológica de una proteína.
VEST3_score	Puntaje de VEST3. Mientras mayor sea la puntuación, la variante es más perjudicial.
MetaSVM_score	Puntaje de la máquina de vectores de soporte (SVM) basada en la puntuación de predicción en conjunto que incorpora 10 puntuaciones (SIFT, PolyPhen-2 HDIV, PolyPhen-2 HVAR, GERP++, MutationTaster, Mutation Assessor, FATHMM, LRT, SiPhy, PhyloP) y la frecuencia máxima observada en las poblaciones de 1000 genomas. Un valor mayor significa que es más probable que la SNV sea dañina.
MetaSVM_pred	Predicción de SVM : pueden clasificarse como toleradas (T) o dañinas (D).
MetaLR_score	Puntuación de regresión lineal (LR) basada en la puntuación de predicción en conjunto que incorpora 10 puntuaciones (SIFT, PolyPhen-2 HDIV, PolyPhen-2 HVAR, GERP++, MutationTaster, Mutation Assessor, FATHMM, LRT, SiPhy, PhyloP) y la frecuencia máxima observada en las poblaciones de 1000 genomas. Un valor mayor significa que es más probable que la SNV sea dañina.
MetaLR_pred	Predicción de LR : pueden clasificarse como toleradas (T) o dañinas (D).
DANN_score	Puntaje a partir de una red neuronal de aprendizaje profundo (DANN) para anotar la patogenecidad de las variantes.
fathmm-MKL coding score	Puntaje de fatmm de un algoritmo de múltiples núcleos (MKL) para predecir los efectos funcionales de las mutaciones sin sentido.
fathmm-MKL coding pred	Predicción de fathmm-MKL: pueden clasificarse como toleradas (T) o dañinas (D).

integrated fitCons score	Puntaje integrado de fitcons que representa la probabilidad de una variante de ser perjudicial, los puntajes más altos reflejan una mayor probablidad de que la variante cause la enfermedad.
integrated confidence value	valor de confianza integrado: las puntuaciones más altas significa que son más perjudiciales.
GERP++ score	Puntaje de GERP++. Puntajes más altos tienen mayor probablidad de que la mutación sea dañina.
phyloP7way vertebrate	Puntaje de phyloP7way basado en un modelo de longitudes de rama para 7 especies vertebradas. Cuanto mayor sea la puntuación, más conservado es el sitio. Puntajes más altos son más perjudiciales.
phyloP20way mammalian	Puntaje del modelo filogenético de Markov que usa 20 especies de mamíferos. Las puntuaciones más altas, quieren decir que las variantes son más perjudiciales.
phastCons7way vertebrate	puntaje de conservación de phastCons basado en las múltiples alineaciones de 7 vertebrados. Cuanto mayor sea la puntuación, más conservado el sitio.
phastCons20way mammalian	puntaje de conservación de phastCons basado en las múltiples alineaciones de 20 genomas de mamíferos (incluido el humano). Cuanto mayor sea la puntuación, más conservado el sitio.
SiPhy 29way logOdds	La distribución estacionaria estimada de A, C, G y T en el locus utilizando el algoritmo SiPhy basado en 29 genomas de mamíferos.
CLNALLELEID	Id del alelo en Clinvar
CLNDN	Enfermedades con la cual la variante está relacionada en ClinVar.
CLNDISDB	Nombre de la base de datos e identificador del nombre de la enfermedad de ClinVar.
CLNREVSTAT	Puntuación de confianza de clasificación.
CLNSIG	Clasificación de la significancia clínica patogenicidad obtenida de ClinVar.
gnomAD genome ALL	frecuencia alélicas en todas las muestras de gnomAD
gnomAD genome AFR	frecuencias alélicas en las muestras africanas/afroamericanas de gnomAD
gnomAD genome AMR	frecuencias alélicas en las muestras americanas de gnomAD
gnomAD genome ASJ	frecuencias alélicas en las muestras judío Ashkenazi de gnomAD
gnomAD genome EAS	frecuencias alélicas en las muestras de Asia oriental de gnomAD
gnomAD genome FIN	frecuencias alélicas en las muestras finlandesas de gnomAD
gnomAD genome NFE	frecuencias alélicas en las muestras europeas no finlandesas de gnomAD
gnomAD genome OTH	frecuencias alélicas en las otras muestras de gnomAD
dbscSNV ADA score	Puntuación de dbscSNV utilizando AdaBoost que predice la probabilidad de que la variante pueda afectar el empalmen (del inglés, splicing site)

dbscSNV RF score	Puntuación de dbscSNV utilizando Random Forest que predice la probabilidad de que la variante pueda afectar el empalmen (del inglés, splicing site)
Interpro domain	Dominio Interpro, base de datos para clasificar secuencias en familias de proteínas y predecir la presencia de dominios y sitios importantes.
OMIM	Enfermedad de OMIM relacionada con la variante.
Phenotype_MIM	Número de MIM de fenotipo relacionado con la variante de OMIM.
OrphaNumber	Número de Orpha de la enfermedad
Orpha	Enfermedad en Orpha relacionada con la variante.

Cuadro C.1: *Información obtenida para las variantes en el archivo CSV a partir del flujo de análisis*

C.2. WES contra WGS

Si graficamos las métricas promedio obtenidas para cada variante se obtiene el mapa de calor de la figura C.1

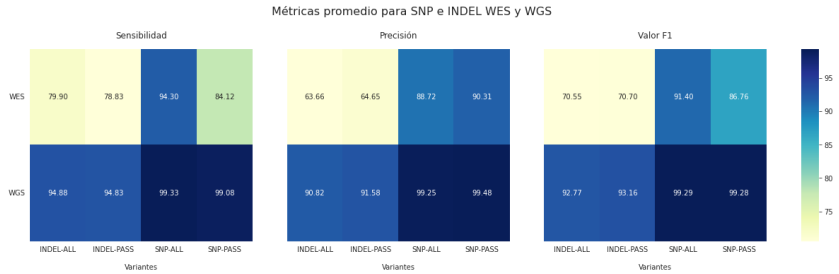


Figura C.1: *Sensibilidad, Precisión y Valor F1 promedio para las variantes SNP e INDEL de WES en comparación con WGS.*

En cuanto al tiempo de ejecución del análisis de WES en comparación con el de WGS, en la figura C.2

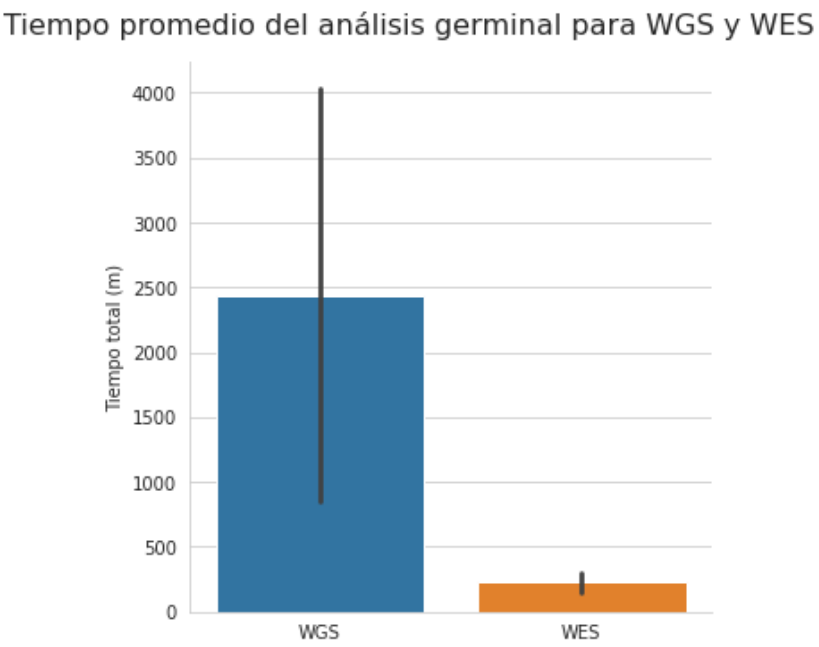
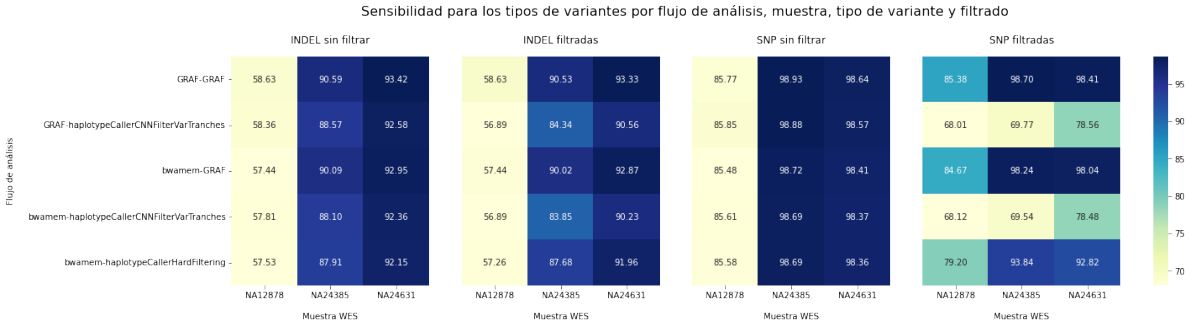


Figura C.2: *Tiempo de ejecución promedio junto con el intervalo de confianza para correr los flujos de análiis germinales para muestras WGS y WES*

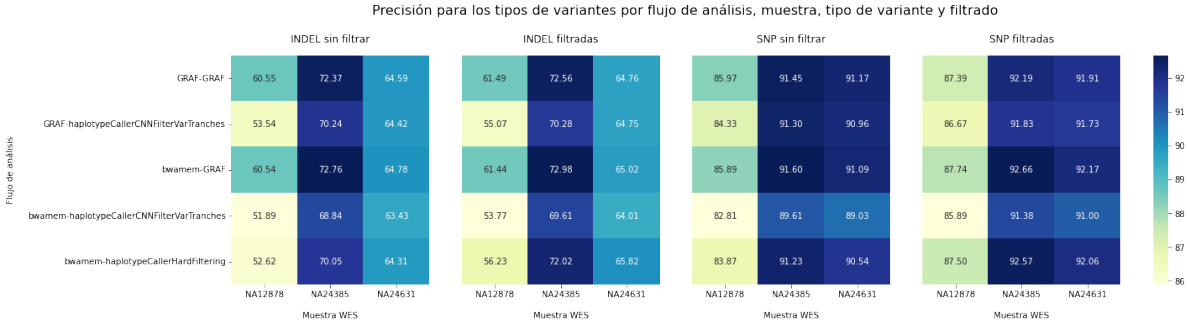
C.3. Diferencia entre muestras

A medida que se ejecutaban los flujos de análisis, observamos que los resultados para la muestra NA12878 eran comparativamente peores que para las otras dos muestras.

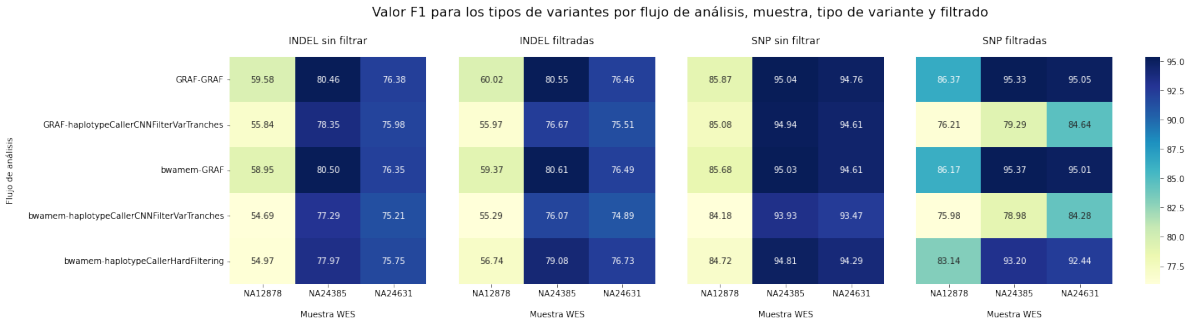
En la figura C.2 se muestra la sensibilidad (ver mapas de calor C.2(a)), precisión (ver mapas de calor C.2(b)) y valor F1 (ver mapas de calor C.2(c)) obtenido para cada flujo de análisis por muestra y tipo y filtrado de variante.



((a)) *Sensibilidad por tipo de variante y filtrado, flujo de análisis y muestra WES.*



((b)) *Precisión por tipo de variante y filtrado, flujo de análisis y muestra WES.*



((c)) *Valor F1 por tipo de variante y filtrado, flujo de análisis y muestra WES.*

Figura C.2: *Mapas de calor en que se visualiza de arriba a abajo: (a) la sensibilidad, (b) Precisión y, (c) Valor F1, para cada una de las muestras WES, en los flujos de análisis, por tipo de variante y filtrado. Se puede notar que para todos los flujos de análisis, la primera columna de la matriz de cada uno de los mapas de calor, es más clara lo que significa que tiene un menor valor de sensibilidad, precisión y valor F1.*

Tomando un promedio sensibilidad, precisión y valor F1 para todos los flujos de análisis, se obtienen los resultados mostrados en la tabla C.2 para cada muestra:

Promedio de las métricas obtenidas en cada flujo de análisis

Paciente	Tipo de variantes	Filtradas	Sensibilidad	Precisión	Valor F1
NA12878	INDEL	No	57,95%	55,83%	56,81%
	INDEL	Sí	57,42%	57,60%	57,48%
	SNP	No	85,66%	84,57%	85,11%
	SNP	Sí	77,07%	87,04%	81,58%
NA24631	INDEL	No	92,69%	64,31%	75,93%
	INDEL	Sí	91,79%	64,87%	76,02%
	SNP	No	98,47%	90,56%	94,35%
	SNP	Sí	89,26%	91,77%	90,28%
NA24385	INDEL	No	89,05%	70,85%	78,91%
	INDEL	Sí	87,28%	71,49%	78,60%
	SNP	No	98,78%	91,04%	94,75%
	SNP	Sí	86,02%	92,13%	88,44%

Cuadro C.2: *Cuadro comparativo con los promedios para cada muestra WES de cada una de las métricas tomando en cuenta los resultados obtenidos para todos los flujos de análisis.*

Tomando los resultados de la tabla C.2 y graficando esto mismo en un mapa de calor se obtiene:

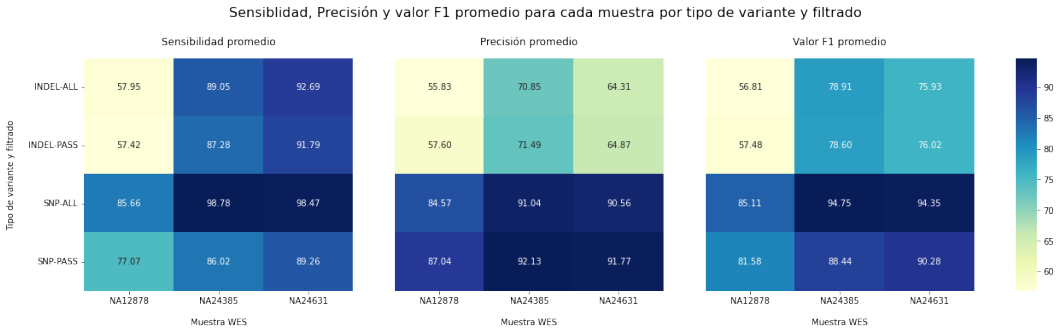


Figura C.3: *Mapa de calor que muestra el promedio de la Sensibilidad, Precisión y Valor F1 para todos los flujos de análisis por muestra WES y tipo de variante y filtrado.*

Si se grafican los resultados de la tabla C.2 en un grafico de barras junto con su intervalo de confianza se obtiene la figura C.4

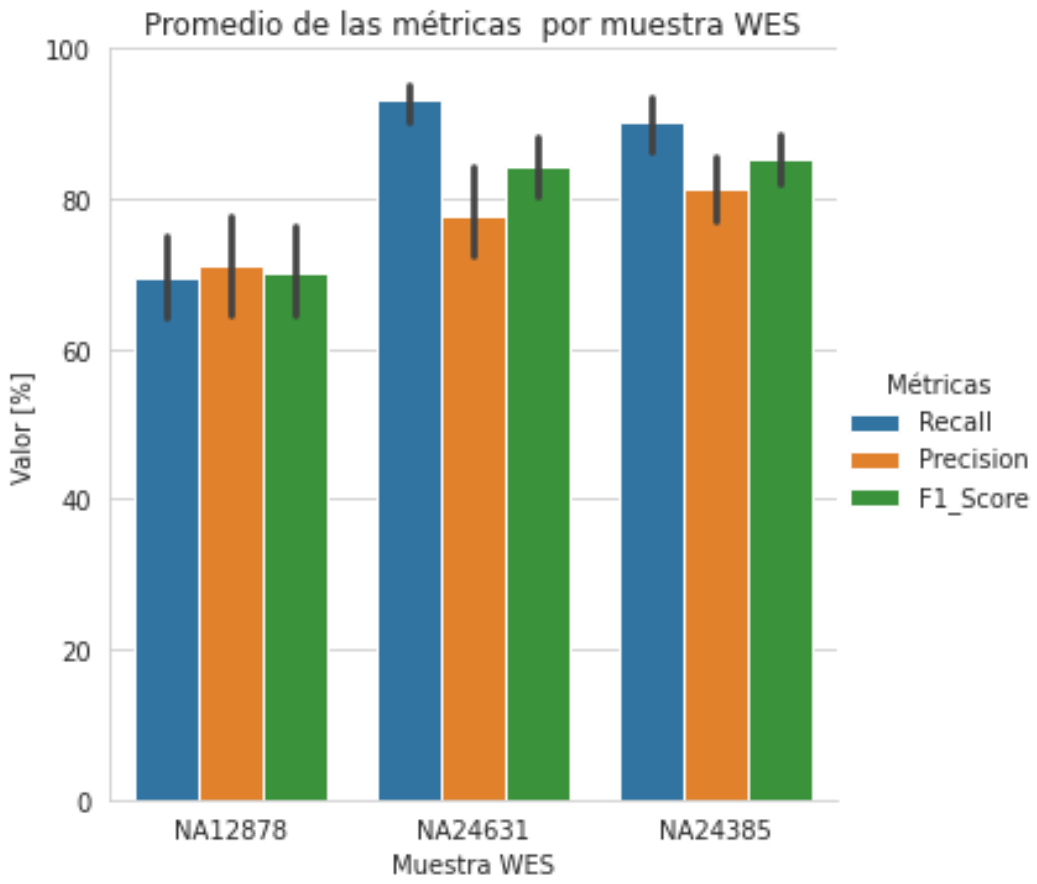
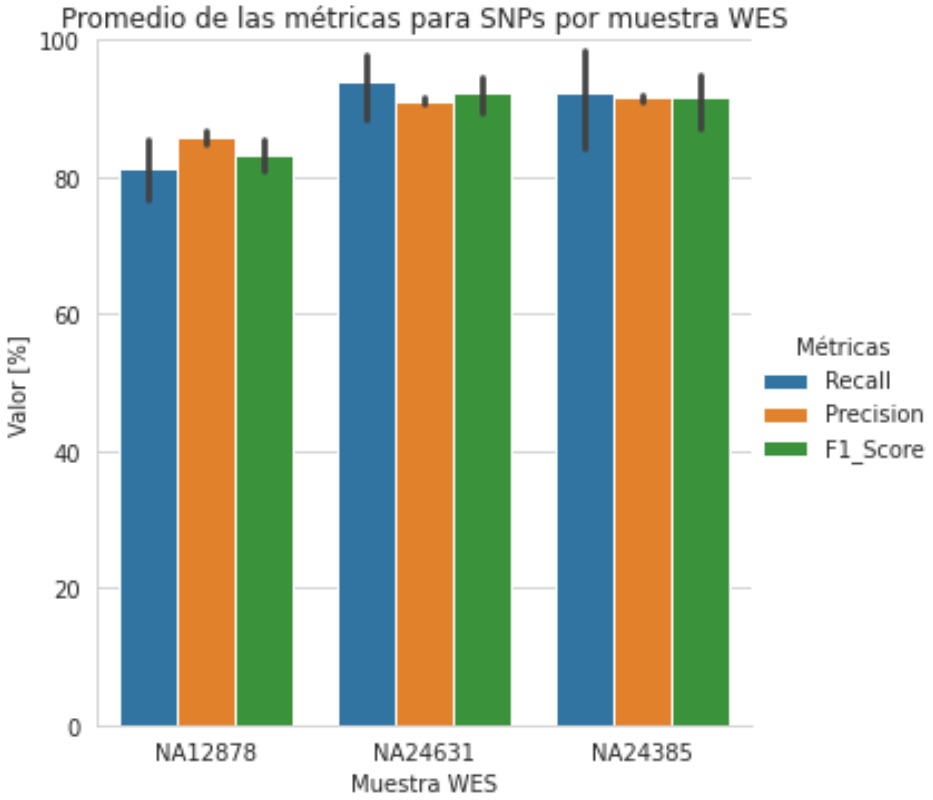
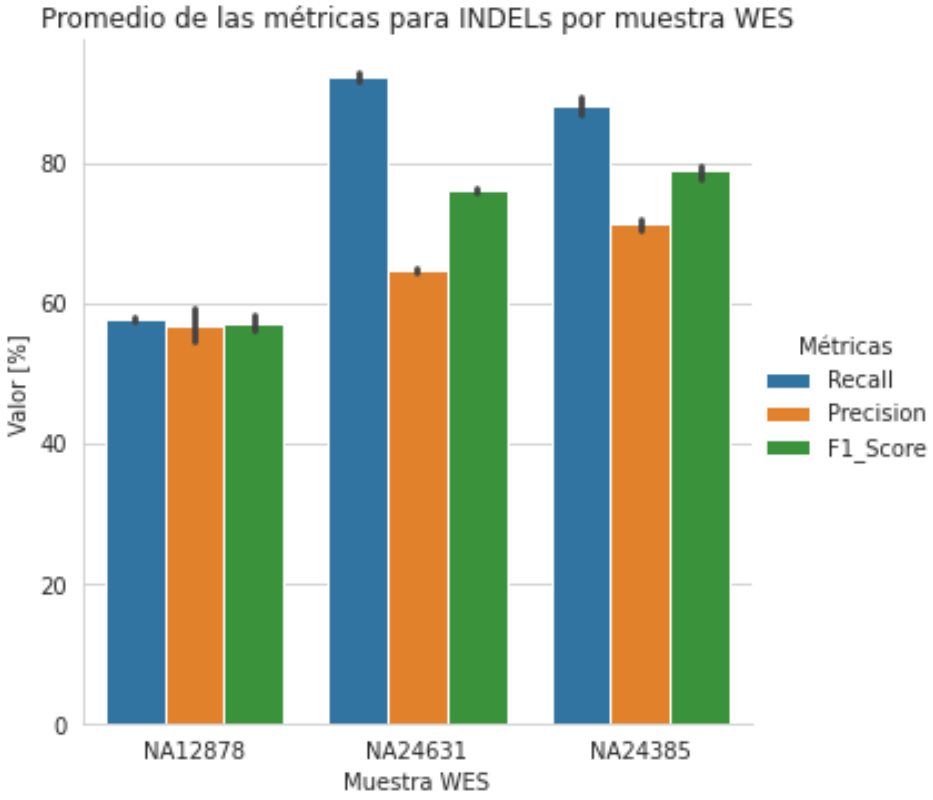


Figura C.4: *Gráfico de barras del promedio de la sensibilidad, precisión y valor F1 para cada muestra WES. También se puede notar el intervalo de confianza de cada una de las métricas.*

Más aún, al comparar los resultados entre SNPs e INDELs se obtienen las gráficas en la figura C.4



((a)) *Gráfico de barras del promedio junto con el intervalo de confianza de la sensibilidad, Precisión y valor F1 para las SNPs de las distintas muestras.*

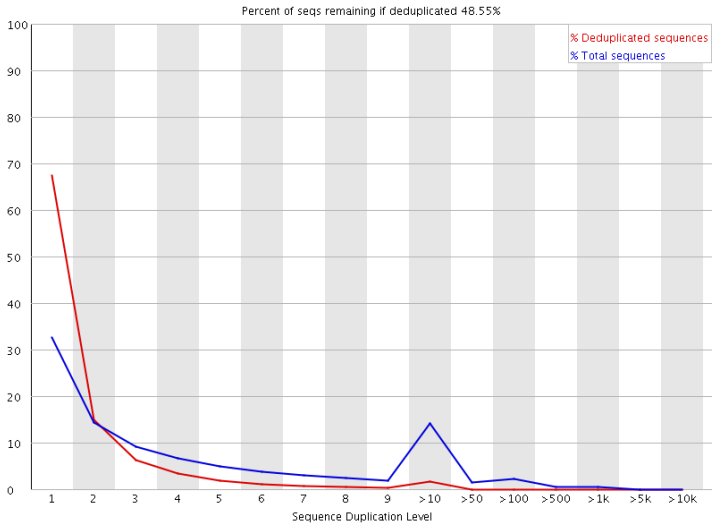


((b)) *Gráfico de barras del promedio junto con el intervalo de confianza de la sensibilidad, Precisión y valor F1 para las INDELs de las distintas muestras.*

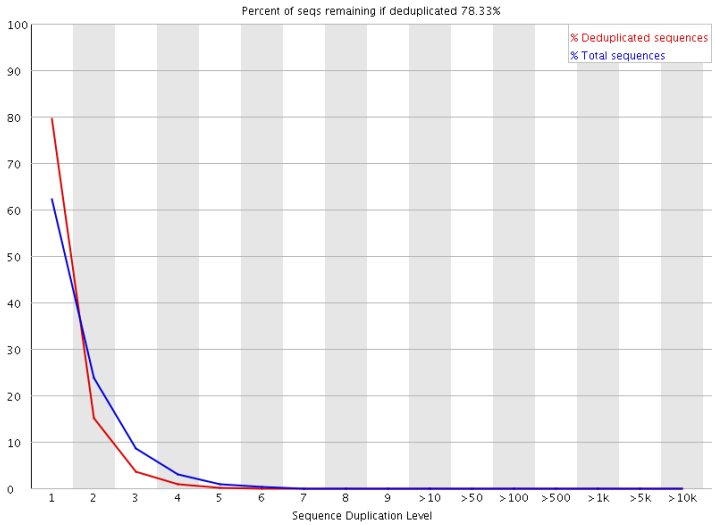
Figura C.4: *Gráficas del promedio junto con el intervalo de confianza de las métricas obtenidas para los SNPs (a) e INDELs (b) de las distintas muestras WES.*

En cuanto al análisis de la calidad de las muestras, en la figura C.5 podemos observar la tasa de desduplicación de las distintas muestras.

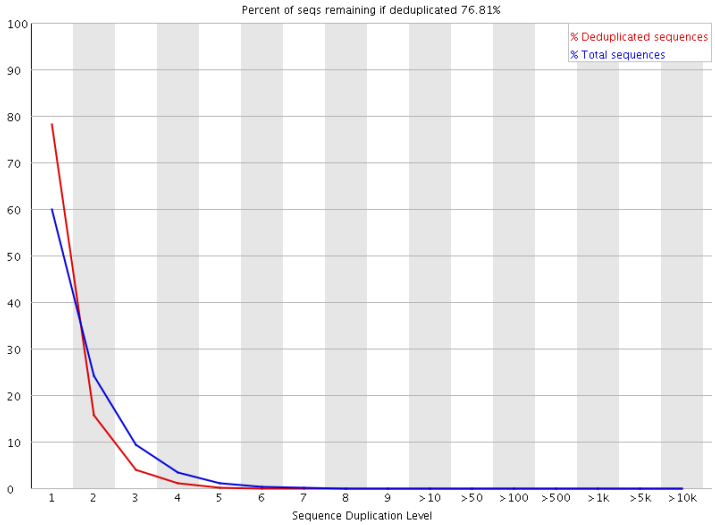
A través de la herramienta Fastqc también se obtuvo el contenido porcentual de cada base nucleotídica para las lecturas de las tres muestras de exoma como se muesra en la figura C.6



((a)) *Tasa de desduplicación para la muestra NA12878*

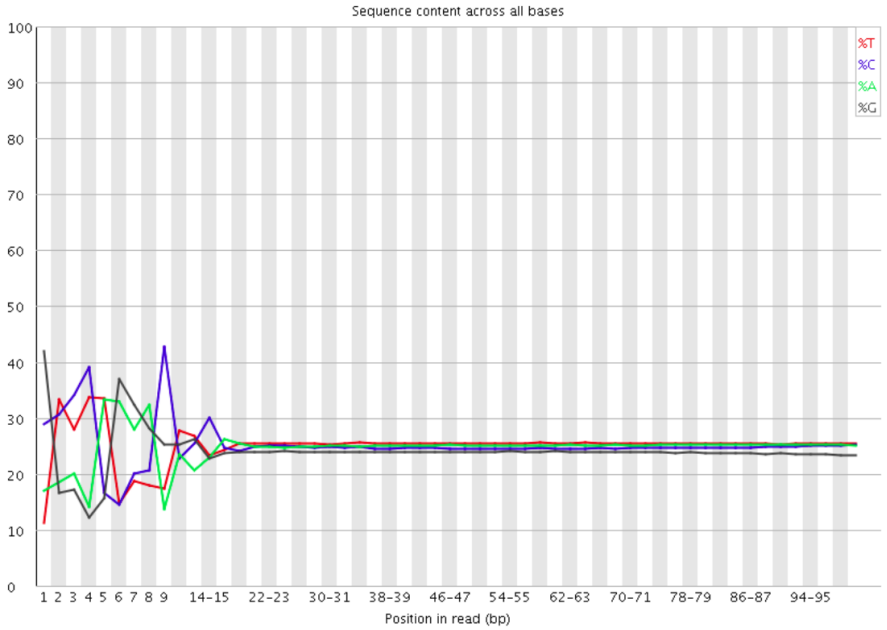


((b)) *Tasa de desduplicación para la muestra NA24631*

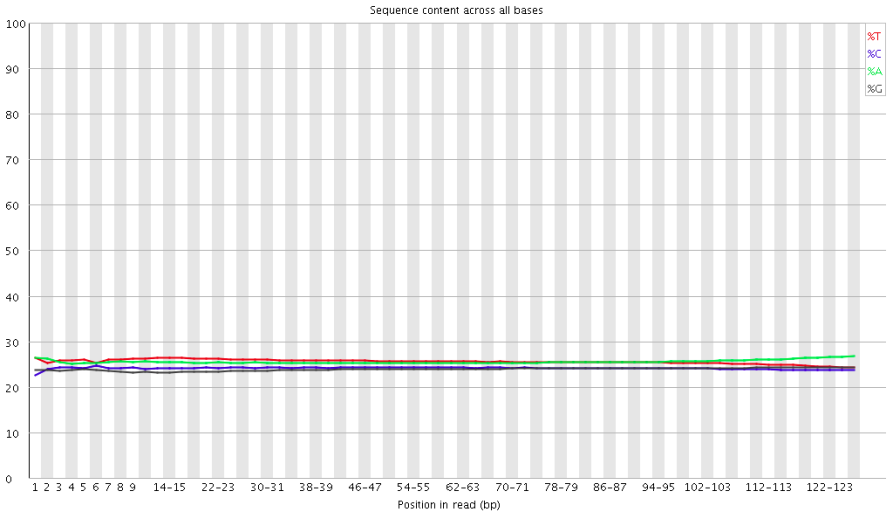


((c)) *Tasa de desduplicación para la muestra NA24385*

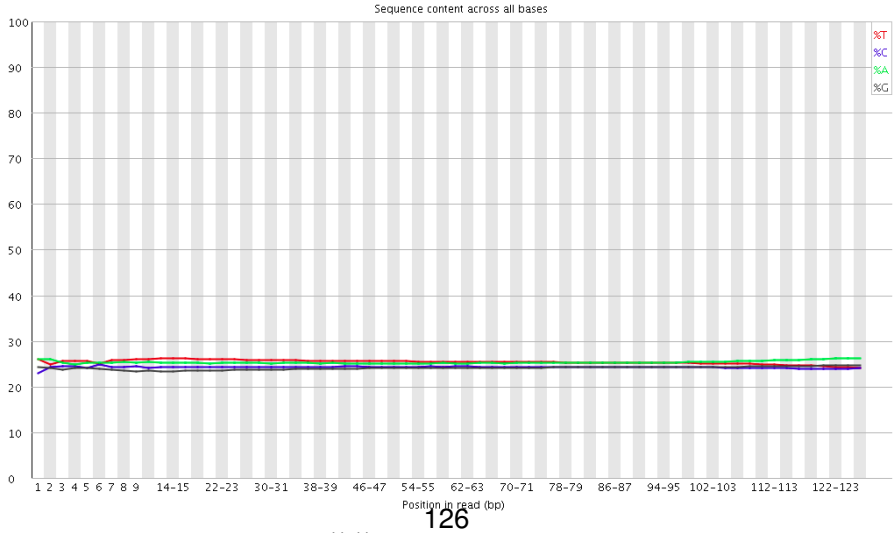
Figura C.5: *Gráficas obtenidas de la herramienta Fastqc para las lecturas de las tres muestras de exoma que muestran Tasa de desduplicación. La línea azul representa el conjunto completo de secuencias y muestra cómo se distribuyen sus niveles de duplicación y, el gráfico rojo, el conjunto deduplicado. En el título de cada uno de los gráficos se muestra el porcentaje genómico restante luego de quitar las secuencias duplicadas.*



((a)) *Muestra NA12878*



((b)) *Muestra NA24631*



((c)) *Muestra NA24385*

Figura C.6: *Contenido porcentual de cada base nucleotídica para las lecturas de las tres muestras de exoma analizadas*

APÉNDICE D

APÉNDICE DE LA DISCUSIÓN

D.1. Diferencias entre muestras

A medida que se ejecutaban los flujos de análisis, observamos que los resultados para la muestra NA12878 eran comparativamente peores que para las otras dos muestras.

Podemos observar de la figura C.2 que para la primer columna de todas las matrices, que es la que corresponde a la muestra WES NA12878, hay valores menores de sensibilidad, especificidad y F1, independientemente del flujo de análisis con el que se corra la muestra.

Observando los resultados obtenidos de la tabla C.2 y la figura C.3, se puede apreciar una clara diferencia tanto para sensibilidad, precisión y valor F1 en la muestra NA12878, con respecto de las otras dos. La diferencia entre la sensibilidad promedio entre las muestras con NA12878 va desde el 8,95% para las variantes SNP filtradas entre la muestra NA12878 y NA24385 alcanzando un máximo de diferencia del 34,74% para las INDEL sin filtrar entre la muestra NA12878 y NA24631. Para la precisión, la diferencia entre va desde el 4,73% para las SNP filtradas entre la muestra NA12878 y NA24631 hasta el 15,02% para las INDEL sin filtrar entre la muestra NA12878 y NA24385. Por último, en cuanto al valor F1, se notan diferencias promedio con valores desde 6,86% entre NA12878 y NA24385 para las SNP filtradas hasta el 22,10% para las INDEL sin filtrar entre las muestras NA12878 y NA24385. Esto también se puede notar en el gráfico de barras de la figura C.4.

De la gráfica C.4, podemos notar que para las variantes INDELs de NA12878, hay una diferencia promedio del 30,48% en sensibilidad, 14,46% en precisión y 21,61% en el valor F1 con respecto a la muestra WES NA24385. Con respecto a la muestra NA24631, hay una diferencia del 34,55%, 7,87% y 18,83% en sensibilidad, precisión y valor F1 respectivamente.

Para las variantes SNPs, también notamos grandes diferencias entre la muestra NA12878 con respecto a las dos muestras pero mucho menor que la que presentan las variantes INDELs.

Hay una diferencia del 11,03% en sensibilidad, 5,78% en precisión y del 8,25% en valor F1 con respecto a la muestra NA24385 y; del 12.50% en sensibilidad, 5,36% en precisión y 8,97% en el valor F1 con respecto a la muestra NA24631. Es casi más del doble de diferencia entre las SNPs y las INDELs.

Comparando parámetros devueltos por Fastqc se puede notar que el porcentaje genómico restante tras quitar la parte de secuencia duplicada disminuye a la mitad para NA12878, mientras que para las otras 2 queda más de un 75%[1]. Esto puede explicar la razón de estos resultados (ver figura C.5).

Además del porcentaje genómico de las secuencias deduplicadas, en las gráficas de la figura C.5, particularmente en la muestra NA12878, se observan picos en el trazo azul lo que sugiere que hay una gran cantidad de secuencias diferentes altamente duplicadas que podrían indicar un conjunto contaminante o una duplicación técnica muy severa. [82]

Por otra parte, para la muestra NA12878 también se puede notar que el contenido porcentual de ACGT, no se mantiene en un 25% a lo largo de las lecturas, a diferencia de las otras muestras que sí muestran un porcentaje constante de 25%. Específicamente se observa que las primeras bases de las lecturas son muy ruidosas (ver figura C.6, y por lo tanto poco confiables; ya que esto puede resultar en información errónea o perdida. [151]

Para comprender las causas de la diferencia de calidad entre muestras, analizamos más a fondo la preparación de librerías y secuenciación en la tabla D.1

Características de secuenciación	**NA12878** [152]	**NA24631** [153]	**NA24385** [154]
Secuenciador	Illumina HiSeq2500	Illumina HiSeq2500	Illumina HiSeq2500
Kit de secuenciación	Nextera Rapid Capture Exome and Expanded Exome [155]	Agilent SureSelect Human All Exon V5 kit [156]	Agilent SureSelect Human All Exon V5 kit [156]
Contenido objetivo del kit	Exones, regiones UTRs y miARN	Exones	Exones
Exones capturados	201,121	357,999	357,999

Cuadro D.1: *Características sobre la secuenciación de las muestras.*

De la tabla D.1 podemos notar que para las muestras NA24632 y NA24385 se utilizó un kit de secuenciación distinto que para la muestra NA12878. Esto tiene como consecuencia que para estas dos muestras el contenido objetivo sea solo los exones y que ese kit capture mayor cantidad de exones. Creemos la diferencia entre los kits de secuenciación puede ser una de las causas de la diferencia entre muestras. Se ha estudiado [157] que la secuenciación realizada con Agilent suele brindar mejores resultados que Nextera. Esto mismo se ha analizado por otros grupos de trabajo [46], llegando a la misma conclusión.

En segundo lugar, notamos una diferencia de más de un 10% en los tres casos al comparar resultados obtenidos entre análisis de INDELs y SNPs (ver figura C.3). Esto concuerda con lo

[1] Para más información en duplicación de secuencias ver https://www.biostars.org/p/107402/

esperado respecto a que los primeros, al ser variantes de más complejidad que los polimorfismos únicos, también resultan más difíciles de detectar por parte de los algoritmos. [42, 97] En la sección D.3 se hace mayor hincapié sobre la diferencia de los resultados entre los dos tipos de variantes.

También se puede notar una diferencia mayor en precisión y consecuentemente, también en el valor F1 entre aquellos flujos de análisis que utilizan a GRAF para el llamado y filtrado de variantes con respecto a los que contienen las herramientas de Haplotype Caller y Filter Variant Tranches o Hard Filtering para la muestra NA12878 que en el resto de las muestras(ver figuras C.2(b) y C.2(c)). Esto podría deberse a que Haplotype Caller utiliza el genoma de referencia lineal que no captura correctamente la diversidad de la población debido a que un 70% de sus secuencias son provenientes de un pequeño número de donantes de EEUU. [158] La muestra NA12878 proviene de residentes de Utah con descendencia del norte y oeste de Europa. Como GRAF utiliza un genoma de referencia gráfico, tiene en cuenta mayor cantidad de variantes para las distintas poblaciones capturando la diversidad genética poblacional, lo que podría aumentar la precisión. [43, 106]

D.2. Dependencia del BED

Se ha encontrado que los resultados de las métricas para la detección de variantes dependen mucho del archivo BED utilizado en el momento de realizar la validación. Un ejemplo se muestra en la tabla D.2

Archivo VCF	Archivo BED	Genoma de Referencia	Archivo VCF Gold Standard	Sensibilidad	Precisión	Valor F1
NA12878_GiaB_3.3.2_high_confidence_calls_GRCh38	GRCh38.GRAF.Genome_Intervals.v1.bed	GRCh38	HG001_GRCh38_1_22_v4.2	84,76%	85,18%	84,97%
NA12878_GiaB_3.3.2_high_confidence_calls_GRCh38	HG001_GRCh38_1_22_v4.2.1_benchmark.bed	GRCh38	HG001_GRCh38_1_22_v4.2	87,33%	99,99%	93,23%

Cuadro D.2: *Comparación de métricas para todas las variantes INDEL del mismo VCF generado por Seven Bridges modificando únicamente el BED para correr el Hap.py.*

D.3. Diferencia en resultados para INDELs contra SNPs

Analizando los gráficos de las tablas 11.1 y 11.3 , notamos que los valores de las métricas son menores en INDEL que en SNP independientemente de análisis utilizado. Observamos que el puntaje de las métricas es mayor en aquellos que utilizan GRAF para el llamado de variantes por lo explicado en la sección **??**. En general, las variantes INDEL son más dificiles de detectar debido a que por su naturaleza, son más complejas que las SNP, pudiendo se una combinación de inserción y eliminación. [42, 97, 159] Además, las variantes INDEL suelen encontrarse en regiones más complicadas de secuenciar lo que hace que el análisis de la secuenciación sea también más difícil, especialmente para los indeles heterocigóticos, donde requiere reconstruir manualmente los dos alelos a partir de largos tramos de picos superpuestos. [159]

Además, entre WES y WGS notamos que las métricas son mayores en WGS independientemente del tipo de variante. En la sección D.4 se discute más acerca de la diferencia entre ambas técnicas. Sin embargo, vemos que esta diferencia se acrecienta para las variantes INDEL.

Esto podría deberse a que WES presenta un bajo rendimiento para el llamado de variantes de INDELs cerca de los límites del exón. Por otro lado, está estudiado que el efecto del contenido de GC disminuye el rendimiento para el llamado de variantes. Se ha visto que el contenido de GC afecta mayormente al llamado de variantes INDEL que al de SNP, tanto en WES como en WGS. [160]

D.4. WES contra WGS

De la figura C.1, notamos que tanto para las variantes SNP como INDEL, las métricas dan mejor para WGS que para WES (aunque el tiempo de ejecución sea mayor para WGS debido a que los archivos de entrada son más grandes, ver figura C.2). Esto podría deberse a que en WGS se secuencia el genoma completo del paciente lo que garantiza una mejor uniformidad de lectura. [161] Además, algunas variantes se encuentran en regiones cerca de los límites del exón lo que hace que sean más difíciles de detectar utilizando la metodología WES. [160]

Particularmente para las variantes SNP, se ha observado que WGS es más preciso y eficiente que WES para identificar SNV verdaderos positivos en el exoma. [159] En cuanto a las INDEL, se nota todavía una mayor diferencia entre WES y WGS, alcanzo casi un 30% de diferencia en precisión y un 15% en sensibilidad. En la sección D.3 se explica con mayor detalle porque podría deberse.

A pesar de que este resultado no se podría generalizar ya que se corrieron los flujos de análisis solamente para una muestra WGS, otras publicaciones científicas notaron los mismos resultados al comparar WES con WGS para la detección de variantes SNP e INDEL. [159–161]

De la gráfica C.2, observamos que el análisis de variantes germinales para muestras WES tiene un tiempo de ejecución casi diez veces menor que para muestras WGS. Esta es una de las razones por las que se decidió trabajar con muestras WES y diseñar un flujo de análisis para estas muestras.