

**ANÁLISIS DE LA VARIACIÓN GENÉTICA EN MUESTRAS DE EXOMAS
ASOCIADAS A PATOLOGÍAS EN PACIENTES COLOMBIANOS DE LA IPS
BIOTECGEN S.A.S.**

Allison Daian Redondo Aguilar

**Universidad El Bosque
Facultad de Ciencias
Programa de Biología
Bogotá D.C. 2022**

**ANÁLISIS DE LA VARIACIÓN GENÉTICA EN MUESTRAS DE EXOMAS
ASOCIADAS A PATOLOGÍAS EN PACIENTES COLOMBIANOS DE LA IPS
BIOTECGEN S.A.S**

Allison Daian Redondo Aguilar

Trabajo de grado para optar por el título de Biólogo

Director:

Daniel Hernán Mahecha López

Médico y MSc. en Biología Computacional, Universidad de los Andes

Codirectores:

Jorge Iván Díaz Riaño

Biólogo y MSc. en Biología Computacional, Universidad de los Andes

Silvia Lizeth Bustamante

MSc. Microbiología, Universidad Nacional de Colombia

**UNIVERSIDAD EL BOSQUE, FACULTAD DE CIENCIAS
PROGRAMA DE BIOLOGÍA
BOGOTÁ D.C. 2022**

Dedicatoria

A mi mamá Leydy Marcela Aguilar, mi mayor
motivación, apoyo, ejemplo de determinación,
disciplina, perseverancia y amor.

Agradecimientos

Agradezco en primer lugar a mi director Daniel Hernan Mahecha López, sin el cual no hubiese sido posible la realización de esta tesis. Agradezco profundamente su enseñanza paciente, compromiso y acompañamiento constante desde mi proceso de entrenamiento hasta la culminación del trabajo. A Biotecgen S.A.S. por la oportunidad de realizar este trabajo de grado. A la doctora Natali Iza por sus aportes en la revisión del trabajo y a mis codirectores Jorge Iván Díaz Riaño y Silvia Lizeth Bustamante.

En segundo lugar, agradezco a mis padres y hermanos quienes son mi motivación. Y en general a mí hermosa familia por su apoyo constante, por creer en mí y motivarme durante todo mi proceso académico.

Agradezco a Laura Yaneth Mazabel con quién tuve la fortuna de formar equipo, compartir durante todo el pregrado y quién fue un apoyo increíble tanto académico como emocional para mí. A mis amigos Juliana González y Gonzalo Ortiz quienes han sido incondicionales y a todas las hermosas personas entre amigos, familiares, docentes y conocidos que de manera muy especial me apoyaron; gracias a ustedes he logrado más de lo que alguna vez imaginé.

Nota de salvedad

“La Universidad el Bosque, no se hace responsable de los conceptos emitidos por el investigador en su trabajo, solo velará por el rigor científico, metodológico y ético de este en aras de la búsqueda de la verdad y la justicia”

Tabla de contenido

1. Introducción	1
2. Planteamiento del problema	2
3. Marco teórico	4
3.1. Secuenciación del exoma completo	4
3.2. Variantes	9
3.2.1. Variante de nucleótido simple	10
3.2.2. Indel	11
3.2.3. Frameshift	11
3.2.4. Variante sinónima	12
3.2.5. Missense	12
3.2.6. Nonsense	12
4. Pregunta de investigación	17
5. Justificación	18
6. Objetivos	19
6.1. Objetivos General	19
6.2. Objetivos específicos	19
7. Método	20
7.1. Control de calidad	20
7.2. Mapeo	21
7.3. Llamado y genotipado de variantes	21
7.4. Parentesco de las muestras	22
7.4.1. Dendograma	23
7.5. Frecuencia alélica	23
7.6. Equilibrio de Hardy-Weinberg	24
8. Resultados	28
8.1. Control de calidad	28
8.2. Alineamiento	28
8.3. Parentesco	31
8.4. Llamado de variantes	32
8.5. Tamaño de Indel	34
8.6. Frecuencia alélica y variantes en equilibrio de Hardy-Weinberg	35
8.7. Integración de la matriz de frecuencia alélica en VarSeq	39
9. Análisis y discusión	43
9.1. Control de calidad	43
9.2. Alineamiento	43
9.3. Parentesco	44

9.4. Proporción de la variación por clase funcional	44
9.5. Frecuencia alélica, variantes en equilibrio de Hardy-Weinberg y estructura poblacional	45
10. Conclusiones	47
11. Bibliografía	48
12. Anexos	53

Lista de tablas y figuras

Figura 1. *Método de captura en la secuenciación de próxima generación basada en hibridación*

Figura 2. *Variantes genéticas comunes. a nivel de nucleótidos. b. A nivel estructural. c. A nivel de SNV*

Figura 3. *Variante frameshift o de cambio de marco*

Figura 4. *Variante sinónima*

Figura 5. *Variante missense*

Figura 6. *Variante nonsense*

Figura 7. *Resumen del método de procesamiento de las muestras de secuenciación del exoma completo, cohorte 2019-2022*

Figura 8. *Control de calidad de profundidad*

Figura 9. *Métricas de calidad del alineamiento de las muestras de WES*

Figura 10. *Muestra de algunas familias reportadas por AKT de la cohorte 2019-2022 observadas en SplitsTree*

Figura 11. *Proporción de SNVs e Indels*

Figura 12. *Proporción de la variación por clase funcional*

Figura 13. *Tamaño de Indel*

Figura 14. *Media de SNPs identificados en el llamado de variantes por genotipo*

Figura 15. *Media de transiciones identificadas en el llamado de variantes*

Figura 16. *Media de transiciones/transversiones identificadas en el llamado de variantes total y por genotipo.*

Figura 17. *Distribución de la frecuencia del alelo menor (MAF) en el conjunto de variantes SNV*

Figura 18. *Análisis de la estructura poblacional con Admixture a. Gráfica del Análisis de Componentes Principales de las muestras analizadas.*

Figura 19. *Matriz de frecuencia alélica integrada en VarSeq*

Tabla 1. *Conteo de frameshift en algunos genes afectados*

Tabla 2. *Estimación de la proporción de subpoblaciones con Admixture para las 17 muestras colombianas (CLM) de 1000 genomas humanos*

Lista de anexos

Anexo 1. *Familias encontradas en el conjunto de muestras de WES de la cohorte 2019-2022*

Anexo 2. *Pasos para la integración a Varseq.*

Anexo 3. *Flujo de trabajo a detalle*

Anexo 4. *Formato de consentimiento informado*

Abreviaturas

ADN: *Ácido desoxirribonucleico*

AF: *Allele frequency*

ARN: *Ácido ribonucleico*

ARNm: *ARN mensajero*

BAM: *Binary Alignment Map*

ExAC: *Exome Aggregation Consortium*

HWE: *Hardy-Weinberg equilibrium*

Indel: *contracción para inserción o delección de nucleótidos*

NSG: *Next-Generation Sequencing*

pb: *pares de bases*

ORF: *Open Reading Frame*

PCR: *Polymerase Chain Reaction*

PTC: *Premature Termination Codon*

SNV: *Single nucleotide variant*

SSM: *Slipped strand mispairing*

WES: *Whole Exome Sequencing*

UTR: *Untranslated Region*

Resumen

La secuenciación de exomas humanos alrededor del mundo ha permitido establecer patrones de variación genética a escala global útiles para brindar una interpretación clínica de variantes. Sin embargo, las bases de datos públicas disponibles, actualmente no reflejan adecuadamente las frecuencias alélicas de poblaciones como la colombiana lo cual se convierte en una limitante importante a la hora de brindar una interpretación clínica de variantes apropiada al contexto específico de diagnóstico genético del país. Por este motivo, se propuso la identificación, la determinación de las frecuencias alélicas y el posterior análisis de variantes de 632 muestras de WES de Biotecgen S.A.S. mediante un flujo de trabajo de control de calidad, alineamiento y llamado de variantes. Se identificaron 1 881 670 SNVs bialélicos y 260 006 Indels. El conjunto presentó una mayor proporción de variantes raras ($MAF < 0,01$). El 88% de las variantes fueron SNVs. Los Indel estuvieron mayormente representados (63%) por deleciones con un tamaño menor a 6 bases. En la clasificación por consecuencia funcional la mayor proporción estuvo representada por missense (55.4%), seguida de sinónimas (43.7%) y nonsense (0.9%). La mayor proporción de loci se encontró en equilibrio de Hardy-Weinberg ($p > 0.05$). Finalmente, las frecuencias alélicas integradas a VarSeq se encuentran actualmente disponibles para los analistas de datos ómicos de Biotecgen S.A.S. proporcionando información relevante para la interpretación clínica de variantes y siendo potencialmente valiosas para futuros estudios de ascendencia genética y estructura poblacional en la cohorte de pacientes colombianos de Biotecgen S.A.S.

Palabras clave. Frecuencia-alélica, variante, SNVs, Indels

Abstract

The sequencing of human exomes around the world has made it possible to establish patterns of genetic variation at a global level that are useful to provide a clinical interpretation of variants. However, the public databases currently available do not adequately reflect the allele frequencies of populations such as the Colombian population, which becomes an important limitation when it comes to providing a clinical interpretation of variants appropriate to the specific context of genetic diagnosis in the country. For this reason, we proposed the identification, determination of allele frequencies and subsequent variant analysis of 632 WES samples from Biotecgen S.A.S. through a workflow of quality control, alignment and variant calling. A total of 1 881 670 biallelic SNVs and 260 006 Indels were identified. The set presented a higher proportion of rare variants ($MAF < 0,01$). In the classification by functional consequence the highest proportion was represented by missense (55.4%), followed by sinónimas (43.7%) and nonsense (0.9%). Eighty-five percent of the variants were SNVs. Indels were mostly represented (63%) by deletions smaller than 6 bases in size and the highest proportion of loci was found in Hardy-Weinberg equilibrium ($p > 0.05$). Finally, allele frequencies integrated to VarSeq are currently available to Biotecgen S.A.S. omics data analysts providing relevant information for the clinical interpretation of variants and being potentially valuable for future studies of genetic ancestry and Colombian population structure.

Keywords. *Allele-frequency, variant, SNVs, Indels*

1. Introducción

El desarrollo y el mejoramiento de las tecnologías de secuenciación de genomas y exomas humanos ha permitido establecer patrones de variación genética a escala global útiles en la interpretación clínica de variantes (Lek *et al.*, 2016). La secuenciación de próxima generación (NGS) ha permitido la secuenciación de millones de fragmentos de ADN en paralelo. Entre sus tecnologías, se encuentra la secuenciación del exoma completo (WES), en la cual se seleccionan las regiones de la secuencia de ADN codificantes de proteínas y brinda coberturas de incluso 100-200X (Rubio *et al.*, 2020; Barbitoff *et al.*, 2022).

Los genes codificantes representan entre el 1 y 2% del genoma. En Colombia, las investigaciones al respecto emplean bases de datos públicas cuya información está mayormente representada por pacientes caucásicos, lo que supone una limitante importante a la hora de brindar una interpretación clínica adecuada de variantes en el diagnóstico genético. A esto, se le suma el vacío en la caracterización de orígenes de ascendencia de individuos en poblaciones mixtas, como la afrodescendiente o la indígena (Conley *et al.*, 2017).

El presente estudio contiene la recopilación y el análisis de 632 muestras de secuenciación del exoma completo (WES) de pacientes colombianos de la IPS Biotecgen S.A.S. con sospecha de enfermedad genética de la cohorte 2019-2022, obtenidas mediante la tecnología de secuenciación por síntesis de Illumina con preparación de librerías mediante Agilent SureSelect Human All Exon V6.

Para poder realizar el análisis del conjunto, fue necesario llevar a cabo un flujo de trabajo de control de calidad, alineamiento y llamado de variantes. En el control de calidad se emplearon los software FASTQC y Trimmomatic, en el mapeo se hizo uso del genoma de referencia GRCh37 - hg19 del

Broad Institute y se emplearon los software BWA, SAMtools y GATK. El llamado y genotipado de variantes se llevó a cabo mediante GATK. Para los pasos adicionales de filtrado, anotación y cálculo de estadísticos se utilizó BCFtools, NGSEP y SNPeff. Posteriormente, se determinó el parentesco de las muestras con AKT y se obtuvieron los estadísticos de frecuencia alélica, equilibrio de Hardy-Weinberg y estructura poblacional analizados a continuación. Para finalmente, integrar las frecuencias alélicas obtenidas a la herramienta de visualización y análisis de variantes GoldenHelix VarSeq para la interpretación clínica de SNV e Indel en Biotecgen S.A.S.

2. Planteamiento del problema

En la actualidad ha sido posible la secuenciación de gran número de genomas y exomas de seres humanos alrededor del mundo, debido a la continua aparición y mejoramiento de tecnologías de secuenciación de alto rendimiento. Estas han permitido establecer patrones humanos de variación genética a nivel global (Lek *et al.*, 2016). Sin embargo, estos datos continúan siendo de difícil acceso por motivos éticos, prácticos y logísticos que varían en los diferentes países (Chande *et al.*, 2020; Conley *et al.*, 2017).

Las investigaciones al respecto se encuentran limitadas debido a que la mayor cantidad de información contenida en bases de datos, proviene de pacientes caucásicos. Lek *et al.*, (2016) realizaron un análisis y agregación de datos de secuencia de ADN de 60.706 humanos con ascendencia diversa para el descubrimiento de variantes en genes codificadores de proteínas, como parte del Exome Aggregation Consortium (ExAC). ExAC fue posteriormente integrada dentro de la base de datos GnomAD (Karczewski *et al.*, 2020) en la cual la población “latina” representa tan solo el 13% y corresponde a “estadounidenses mezclados” (Karczewski *et al.*, 2020). Por lo que se trata de una muestra no representativa de la población latinoamericana. Lo anterior es una limitante importante a la hora de brindar una interpretación clínica adecuada de variantes en el contexto específico de diagnóstico genético del país. A esto, se le suma el vacío en la caracterización de orígenes de ascendencia subcontinental de individuos en poblaciones mixtas, ya que la mayoría de estudios genéticos humanos de ascendencia en Colombia y latinoamérica se han centrado en poblaciones nativas americanas y mestizas con ascendencia europea y nativa americana (Chande *et al.*, 2020; Conley *et al.*, 2017). Poblaciones representativas en América Latina y el país, como la afrodescendiente, cuentan con pocos estudios de ascendencia genética. Esto restringe los aportes en investigación y epidemiología genética en estudios de asociación donde la existencia de subgrupos

o subestructuras poblacionales, pueden llevar a falsas asociaciones en un rasgo (Chande *et al.*, 2020; Córdoba *et al.*, 2012; Conley *et al.*, 2017).

3. Marco teórico

3.1. Secuenciación del exoma completo

La secuenciación de nueva generación (NGS) agrupa múltiples tecnologías de secuenciación de ADN. Las tecnologías de NGS se caracterizan por secuenciar millones de fragmentos de ADN en paralelo varias veces, brindando una gran profundidad que permite obtener datos precisos sobre la variación de secuencia del ADN. La NGS tiene la capacidad de realizar la secuenciación de genomas completos o regiones específicas como el conjunto de genes codificantes de proteína, es decir, la secuenciación del exoma completo (WES) (Behjati & Tarpey, 2013).

Se estima que existen 180.000 exones y 22.000 genes codificantes que representan aproximadamente entre el 1 y el 2% del genoma humano. La secuenciación del exoma completo es la aplicación de la tecnología de próxima generación diseñada para seleccionar específicamente las regiones de secuencia de ADN codificantes de proteínas (proceso conocido como enriquecimiento) y posteriormente secuenciarlas. Esta se remonta al año 2009, y se origina a partir de la combinación de enfoques de captura del exoma con tecnologías NGS. Gracias a la WES, ha sido posible determinar la variación de todos los exones de genes conocidos con una cobertura mayor al 95%, mejorando así el estudio de variantes como Indels y SNVs en tejidos enfermos y sanos. Por esta razón, se ha convertido en un enfoque sólido para la identificación y distinción del papel de más de 150 genes entre los que se agrupan los desencadenantes de enfermedades mendelianas comunes y raras (Ping, 2016; Robinson & Jäger, 2017; Mordoh, 2019; Rabbani, 2014).

3.1.1. Secuenciación por síntesis

En este estudio, se usaron datos obtenidos a partir de la tecnología de secuenciación por síntesis (Illumina®). Esta puede dividirse en cuatro pasos: 1. la preparación de la librería, 2. la generación de clusters y la amplificación en puente, 3. el proceso de secuenciación y 4. el análisis de todos los datos obtenidos (Illumina, 2010; Santamaría & Lezana, 2018).

3.1.1.1. Preparación de la librería

El procedimiento experimental de WES inicia con el enriquecimiento del ADN exónico mediante el método de captura elegido seguido por la construcción de una biblioteca de secuenciación (Ping, 2016; Santamaría & Lezana, 2018).

De manera general, en la preparación de la librería se lleva a cabo la tagmentación donde se emplean transposomas para realizar el marcaje con secuencias adaptadoras y la fragmentación del ADN en una misma reacción (Illumina, 2020). En la WES, se requiere la preparación inicial del ADN; el método de captura del exoma excluye las regiones intrónicas y las no traducidas (UTRs), enriqueciendo de manera selectiva la secuencia exónica para la NGS. Para este fin, existen múltiples kits de enriquecimiento que hacen la secuenciación de forma directa y sencilla como los comercializados por NimbleGen, Illumina® y Agilent (Illumina, 2010, Ping, 2016).

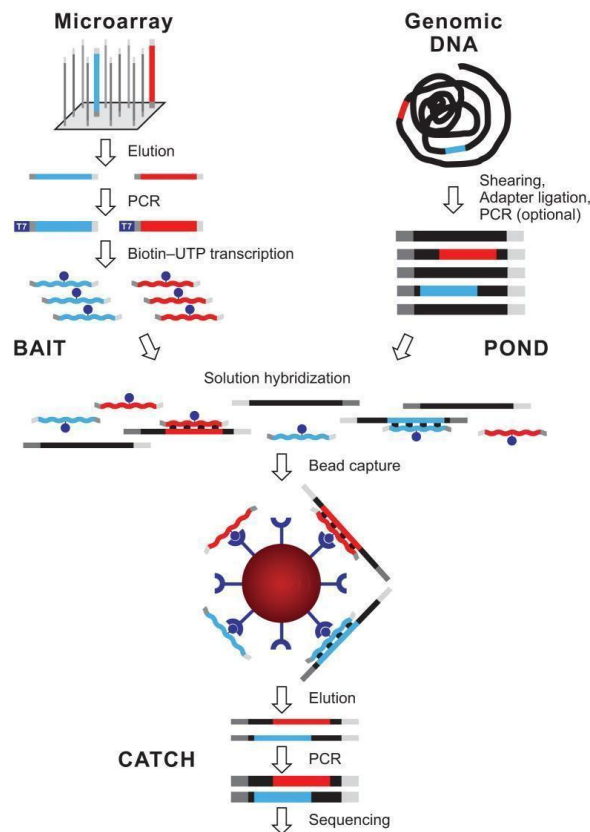
3.1.1.1.1. Método de captura

El método empleado para la captura del exoma fue Agilent SureSelect Human All Exon V6 en el cual, como bien describen Gnirke *et al.*, (2009). Este método aprovecha las ventajas tanto de la cinética favorable de la hibridación en solución como de la economía y flexibilidad que brinda la síntesis de oligonucleótidos en microarreglo. Lo anterior permite la preparación de grandes cantidades de “bait” partiendo de una única síntesis de matriz de oligonucleótidos que se puede almacenar y usar varias veces durante la ejecución de un proyecto de secuenciación dirigida.

La captura consiste en la síntesis de un conjunto de oligonucleótidos ultralargos de 200 nucleótidos en el microarreglo de Agilent que se escinden del soporte y se eluyen en un tubo (Metzker, 2010). Los oligonucleótidos comprenden una secuencia con 170 nucleótidos específicos de diana flanqueada con 15 bases de un primer universal en cada lado para conseguir la amplificación por reacción en cadena de la polimerasa (PCR). Este enfoque se basa en ácido ribonucleico (ARN) biotinilado en el que, en una segunda ronda de PCR se agrega un promotor T7 responsable de la incorporación de uridina-5'-trifosfato (UTP) en la secuencia de la sonda mediante transcripción *in vitro* que en presencia de biotina-UTP genera un cebo de hibridación de ARN para “pescar” las regiones de interés del ADN. El exceso de cebo de ARN monocatenario no autocomplementario impulsará la hibridación, seguido de esto, la captura emplea perlas magnéticas recubiertas de estreptavidina, se amplifica por PCR con primers universales y se analiza en el instrumento de NGS (Gnirke *et al.*, 2009; Mamanova *et al.*, 2010).

Figura 1

Preparación de sondas de captura de ARN biotinilado. La biblioteca de entrada de fragmentos de genoma completo se observa arriba a la derecha, el cebo se ve arriba a la izquierda y abajo la captura en la biblioteca de salida enriquecida seleccionada por híbridos. Se ilustran dos objetivos de secuenciación en rojo y azul cada uno con su cebo. Las hebras simples (líneas finas) y dobles (líneas gruesas). El adaptador universal está representado por el color gris. La hibridación se ve impulsada por el exceso de ARN monocatenario no autocomplementario representado con líneas onduladas



Nota. Gnirke A, Melnikov A, Maguire J, Rogov P, LeProust EM, Brockman W, Fennell T, Giannoukos G, Fisher S, Russ C, Gabriel S, Jaffe DB, Lander ES, Nusbaum C. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat Biotechnol. 2009 Feb;27(2):182-9. doi: 10.1038/nbt.1523. Epub 2009 Feb 1. PMID: 19182786; PMCID: PMC2663421.

3.1.1.2. La generación de clusters y la amplificación en puente

Posteriormente a ello, se lleva a cabo la amplificación con los primers específicos a las secuencias adaptadoras, lo que genera dos sitios de unión a primers diferentes. Adicionalmente se añaden los índices y los sitios complementarios a los sólidos de la celda de flujo (Illumina, 2010).

Para la generación de clusters se lleva a cabo la amplificación en puente donde los fragmentos de ADN se colocan en una celda de flujo, una superficie sólida de vidrio con carriles que contienen

nano pozos recubiertos por dos tipos de oligonucleótidos complementarios a los adaptadores de cada fragmento a secuenciar, lo que permitirá el anclaje de cada fragmento a uno de los pozos de la celda. En los pozos se amplifican los fragmentos de manera isotérmica y una polimerasa generará la secuencia complementaria del fragmento hibridado obteniendo una hebra complementaria a la muestra adherida a la celda. Esta molécula de cadena doble se desnaturaliza, el oligonucleótido original se remueve de la celda de flujo y la otra hebra se dobla uniéndose al segundo oligonucleótido anclado a la celda. Luego, una polimerasa realiza la extensión denominada amplificación en puente. Posteriormente, la estructura bicatenaria se desnaturaliza en una hebra sentido y una antisentido, las cuales serán cortadas y lavadas previamente a la secuenciación. Este proceso se repite múltiples veces generando un cluster de hebras (Illumina, 2010; Rubio *et al.*, 2020).

3.1.1.3. Secuenciación

La secuenciación inicia con la hibridación de un cebador universal (primer) a la cadena para, posteriormente, realizar un ciclo en el que se agregan una serie de soluciones sucesivas con un único nucleótido marcado con un fluoróforo diferente (para A, T, C y G). Estos emitirán para cada caso una longitud de onda distinta tras su adición a la cadena, indicando el tipo de nucleótido. El número de ciclos indicará la longitud de la secuencia. Este proceso se realiza con múltiples clústers emitiendo miles de señales al mismo tiempo. Terminado este proceso, se desnaturaliza la molécula bicatenaria y se retira la hebra generada, se inserta un segundo primer para la lectura de la secuencia índice, se lava el fragmento generado, se repite el proceso hasta la amplificación en puente, se obtienen las lecturas de extremo emparejado y se remueve esta vez, la hebra forward y se repite la secuenciación por síntesis con la hebra reverse (Illumina, 2010).

3.1.1.4. *Análisis bioinformático de los datos obtenidos*

De manera muy general, en el análisis bioinformático el secuenciador realiza un demultiplexado donde se clasifican las lecturas en función de los oligonucleótidos flanqueantes generando un fichero por muestra secuenciada. Posteriormente, se evalúa la calidad de las lecturas obtenidas con un software como FastQC o Prinseq-lite. Luego, se eliminan las secuencias contaminadas, las lecturas de baja calidad, los adaptadores u otras lecturas cuestionables generadas en la secuenciación. Este proceso de filtrado y limpieza se lleva a cabo con herramientas como Prinseq-lite o, como en este estudio, Trimmomatic. Posteriormente se lleva a cabo el mapeo de las lecturas al genoma de referencia, para luego eliminar posibles errores de secuenciación manteniendo la variabilidad. Se obtiene un archivo BAM para cada lectura, se detectan zonas no cubiertas por el genoma de referencia o con un exceso de lecturas alineadas y, finalmente, se lleva a cabo el análisis de variantes de las lecturas (Ping, 2016; López *et al.*, 2021).

3.2. *Variantes*

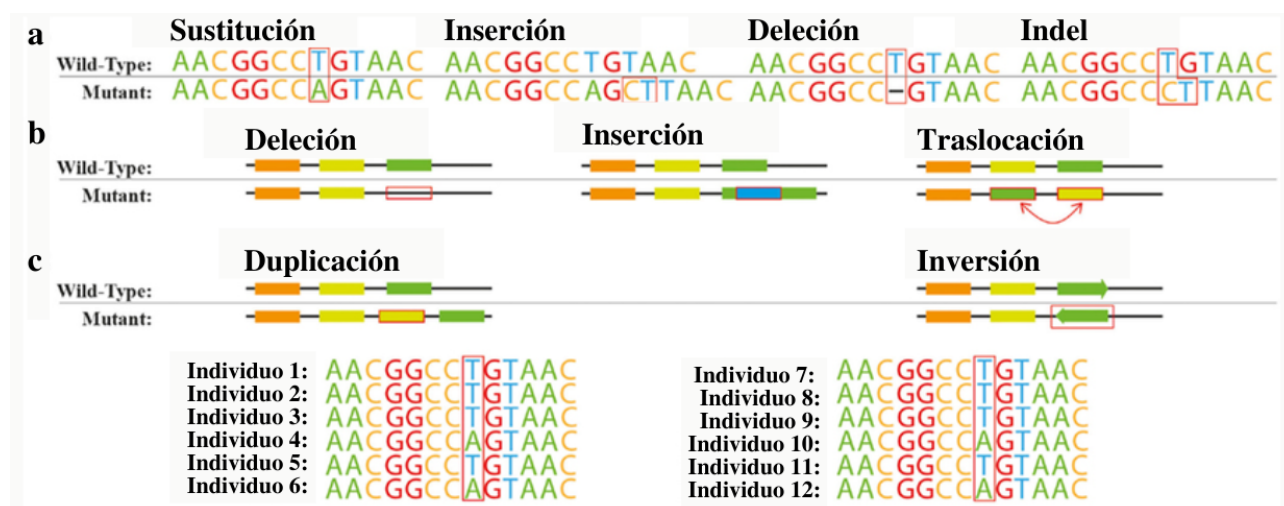
El genoma humano está conformado por un aproximado de 3200 millones de pares de bases contenidas en 23 pares de cromosomas formando una secuencia de nucleótidos casi idéntica entre dos individuos cualesquiera. Los procesos históricos de mutación y deriva genética han dado origen a variaciones en la secuencia, formas alternativas de un mismo gen (múltiples alelos) que se encuentran con una determinada proporción en la población (Lencz, 2022).

Las variantes se pueden clasificar según el tipo de alteración en: SNV, Indel o variación estructural para reordenamientos de más de 50 pares de bases. Estas últimas suelen ser producto de deleciones, duplicaciones, inversiones, inserciones o translocaciones (Ver la Figura 2). Por otro lado, las variantes que se ubican en regiones codificantes se han clasificado en función de su efecto en la estructura de las proteínas en: missense, sinónimas, de pérdida de codón de inicio y de pérdida de

función, las cuales incluyen las de cambio de marco de lectura y las variantes de parada (codón de parada prematuro y pérdida del codón de parada). Al igual que en Trudsø *et al.*, (2020), en este estudio se prefirió el término SNV sobre polimorfismos de único nucleótido (SNP) debido a que un SNV no depende de estar presente en más del 1% de la población. En otras palabras, se contemplan las variantes raras (frecuencia del alelo menor (MAF) <0.05) que a menudo son asociadas con enfermedades hereditarias.

Figura 2

Variantes genéticas comunes. a nivel de nucleótidos. b. A nivel estructural. c. A nivel de SNV



Nota. Cardoso, J.G.R., Andersen, M. R., Herrgård, M. J., Sonnenschein, N.(2015). Common genetic variations.[PNG]. Analysis of Genetic Variation and Potential Applications in Genome-Scale Metabolic Modeling.

https://www.researchgate.net/figure/Common-genetic-variations-Variations-at-the-A-nucleotide-level-and-B-structural_fig2_272676315

3.2.1. Variante de nucleótido simple

t: A type of variation affecting a single nucleotide in a DNA sequence, in which the nucleotide (for example, cytosine) is substituted with a different type of nucleotide

Cuando el cambio en la secuencia se da a nivel de único nucleótido se denomina variante de nucleótido simple (SNV), ocurre cuando el nucleótido es sustituido por otro tipo de nucleótido diferente (Eichler, 2019). Estas variantes de secuencia, que se encuentran en regiones intergénicas, intrónicas y exónicas de manera heterogénea en el genoma, suelen ocurrir de manera natural en individuos de una misma especie, por lo que han sido utilizadas como “huellas dactilares” de ADN para la diferenciación de individuos (Ping, 2016).

En los seres humanos se ha estimado una frecuencia media de 1 SNV por cada 500-1000 nucleótidos frente al genoma de referencia, lo que variará según el cromosoma y la región del mismo (Ding & Jin, 2009). Estas variantes determinan el mayor porcentaje de variabilidad genética humana (INCIFOR, 2021). Se dice que se originaron en diferentes puntos de la historia evolutiva humana y se estabilizaron en el genoma. Han permitido el estudio de orígenes de ascendencia, de rasgos complejos como la función metabólica, la altura y las bases genéticas de enfermedades comunes. Así mismo, las diversas alternativas de un determinado SNV pueden ser cruciales en la propensión al desarrollo de enfermedades, en la agresividad de las mismas o en las variaciones en la respuesta a medicamentos (INCIFOR, 2021; Kohlmeier, 2020; Lencz, 2022; Ping, 2016).

3.2.2. Indel

La extensión en pares de bases del genoma no es la misma para todas las personas; la secuencia de un ser humano puede variar siendo de una base a millones de bases más corta o más larga comparada con la de otro ser humano. Estas variaciones que se remontan varias generaciones e

incluso millones de años atrás se atribuyen a inserciones o deleciones de nucleótidos en la secuencia de ADN humana (Kohlmeier, 2020).

Las inserciones son variantes que implican la adición de uno o más nucleótidos en la secuencia de ADN. Por otro lado, las deleciones implican la pérdida de uno o más nucleótidos como se puede apreciar en la Figura 2. Ambas variantes suelen originarse por el fenómeno conocido como “replication slippage” o emparejamiento erróneo de la hebra deslizada (SSM), afectando desde un nucleótido hasta grandes regiones de un cromosoma. Se ha sugerido que las más comunes se han visto favorecidas por ventajas selectivas como la alineación con el nutritopo predominante (entorno nutricional al que está expuesta la población) y tienen la capacidad de moldear el genoma en una escala temporal grande (Biesecker, 2022; Ganguly, 2022; Kohlmeier, 2020; Savino *et al.*, 2022).

Es difícil atribuir los cambios en la secuencia a la adición o eliminación de un segmento en particular. Por ende, se hace referencia a ambas variantes mediante la contracción Indel. Estas se caracterizan por tener una longitud menor a 50pb (Eichler, 2019) y son de gran importancia ya que aportan información útil de ancestría y estructura poblacional. Sin embargo, su detección comparada con la de SNVs es más difícil debido a dificultades en el mapeo y a su baja frecuencia por lo que se requiere una mayor profundidad (Mordoh, 2019; Sehn, 2015).

3.2.3. *Frameshift*

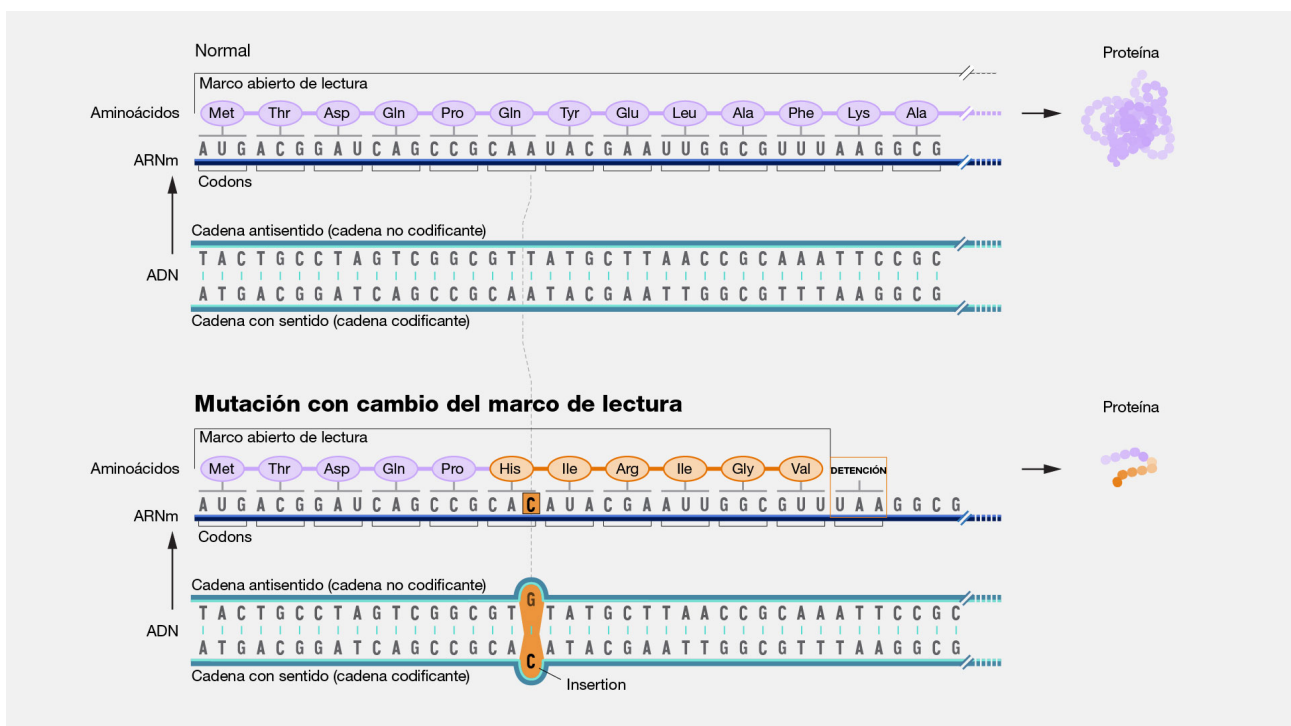
Cuando los nucleótidos insertados o eliminados de la secuencia no son múltiplos de tres la variante se denomina frameshift (de cambio de marco), ya que se ve afectado, como su nombre lo indica, el marco de lectura para la traducción de los triplete de nucleótidos. Por ejemplo, si el Indel es de una o dos bases, como se ilustra en la Figura 3.

Una lectura errónea da como resultado proteínas truncadas debido a que, tras la introducción de una mutación frameshift, de los 64 codones estándar, 44 pueden contribuir a la aparición de un codón de

parada “oculto”. Proteínas no funcionales pueden desencadenar una enfermedad (Adams, 2022; Kohlmeier, 2020; Savino *et al.*, 2022). Por otro lado, pares mutuamente compensatorios de frameshifts pueden contribuir a la recuperación de la funcionalidad al restaurar el marco original (por ejemplo, +1, -1 y +1 +2). De manera similar, se puede revertir el silenciamiento de un gen causado por un frameshift, volviendo a adquirir la versión funcional del gen afectado (Savino *et al.*, 2022).

Figura 3

Variante frameshift o de cambio de marco



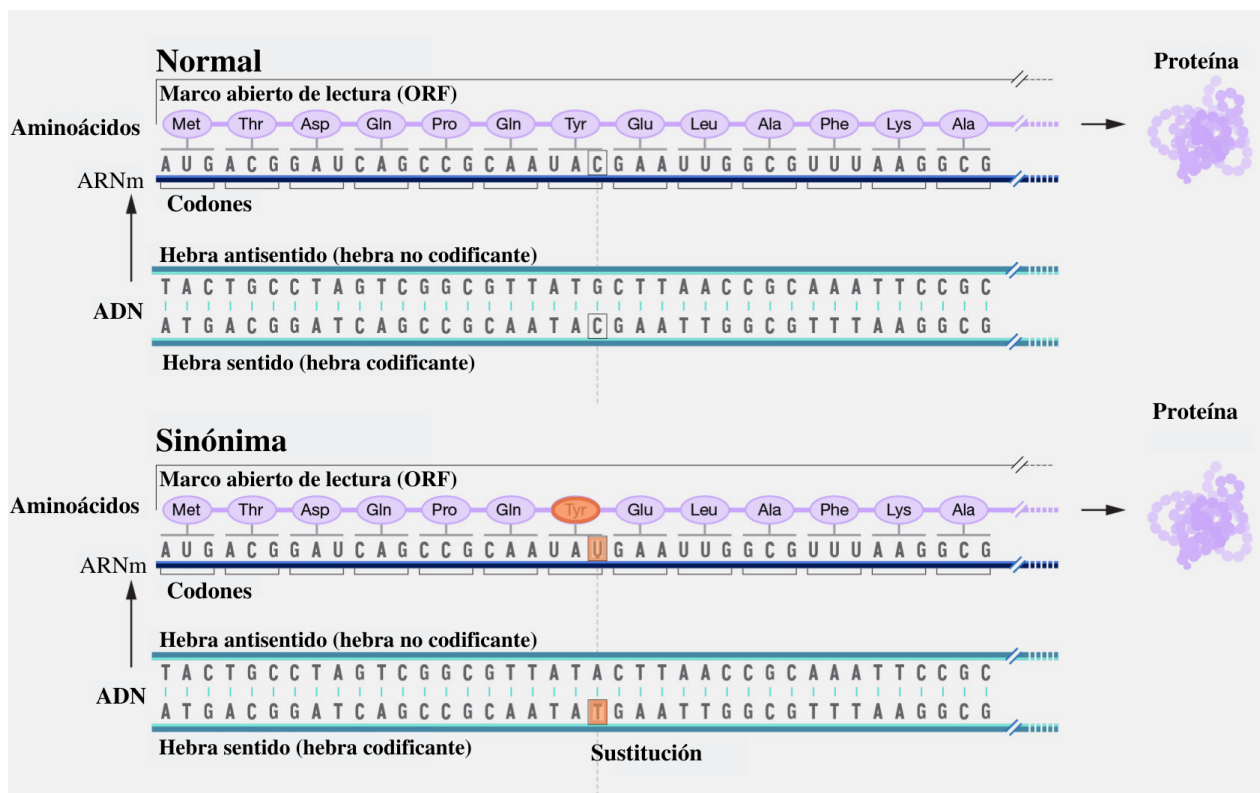
Nota. Leja, D. (2022). *Mutación con cambio del marco de lectura* [JPG]. National Human Genome Research Institute. Genome.gov. <https://www.genome.gov/genetics-glossary/Frameshift-Mutation>

3.2.4. Variante sinónima

Este tipo de variante se caracteriza porque, a pesar de alterar la secuencia de nucleótidos del ADN y ARNm. Gracias a que el código genético es degenerado, no producen una sustitución en el residuo de aminoácido de la proteína resultante, como se observa en la Figura 4. Por esta razón, son denominadas comúnmente como silenciosas (Henson & Resta, 2021). No obstante, como afirman Sharma *et al.*, (2019), en algunos casos estos cambios pueden llegar a afectar la transcripción, la maduración del ARNm y la traducción de ARNm alterando el fenotipo.

Figura 4

Variante sinónima



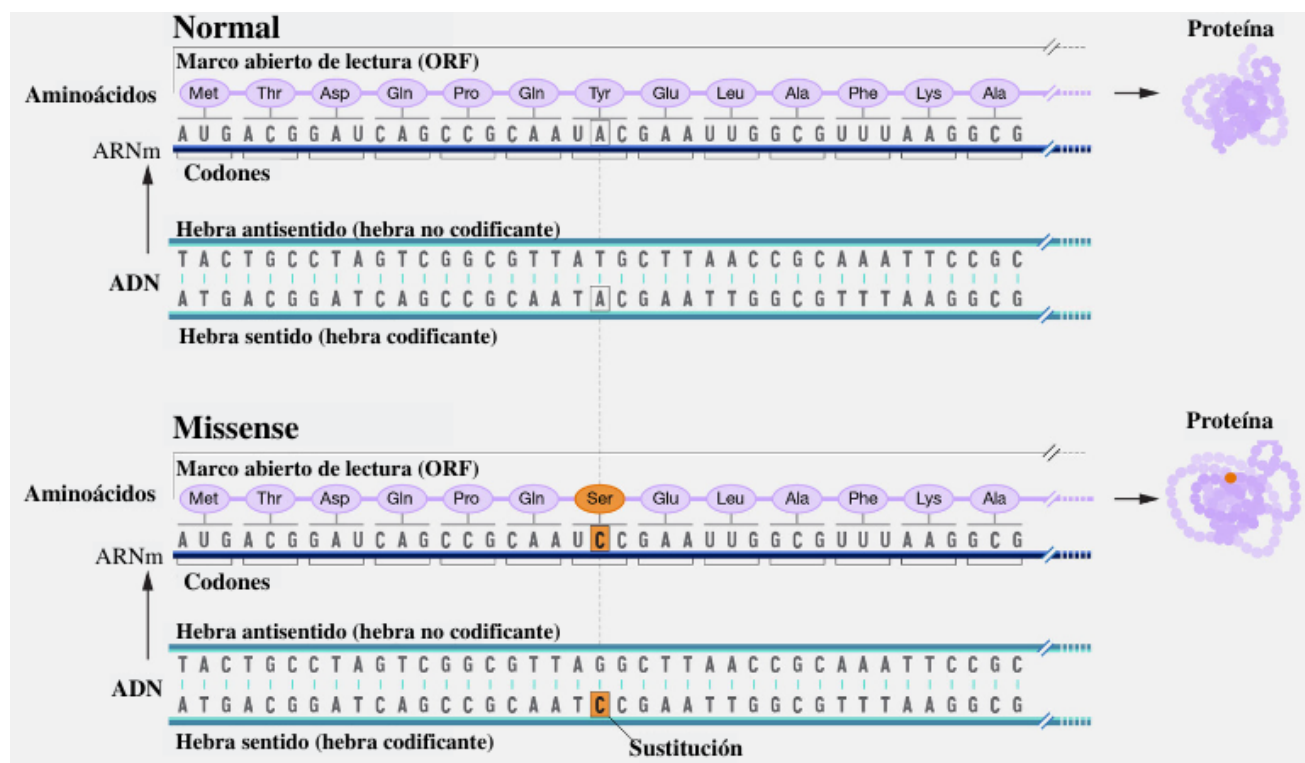
Nota. Adaptado de Missense mutation, de Leja, D. (2022). National Human Genome Research Institute. <https://www.genome.gov/genetics-glossary/Missense-Mutation>

3.2.5. Missense

En este tipo de variante, a diferencia de la anterior, la alteración de la secuencia de nucleótidos sí codifica para un residuo de aminoácido diferente en una posición particular en la secuencia de la proteína resultante, como se observa en la Figura 5 (Brody, 2022; Jepsen et al., 2020).

Figura 5

Variante missense



Nota. Leja, D. (2022). *Missense mutation* [JPG]. National Human Genome Research Institute. Genome.gov. <https://www.genome.gov/genetics-glossary/Missense-Mutation>

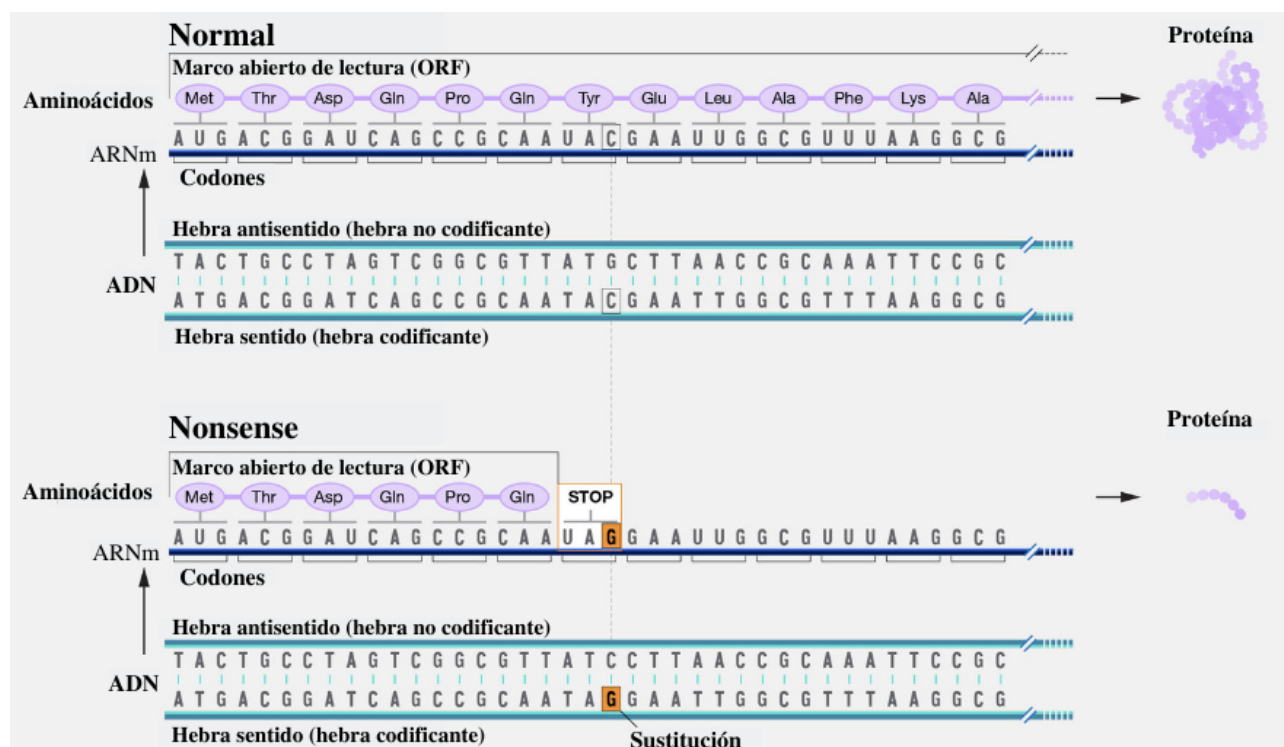
3.2.6. Nonsense

Como se aprecia en la Figura 6, cuando se presenta un codón de parada prematuro dentro del marco de lectura abierto (ORF), es decir, un codón de terminación prematura (PTC) finalizando la codificación de la proteína, la variante se denomina nonsense. Este tipo de variante agrupa a start-loss y stop-gained. Las variantes stop-gained son aquellas que se encuentran entre el codón de

iniciación y el de parada generando proteínas anormalmente cortas. Para el caso de las start-loss, la afectación ocurre específicamente en el codón de inicio (Benhabiles *et al.*, 2016; Guttman, 2013).

Figura 6

Variante nonsense



Nota. Leja, D. (2022). *Nonsense mutation* [JPG]. National Human Genome Research Institute. Genome.gov. <https://www.genome.gov/sites/default/files/media/images/tg/Nonsense-mutation.jpg>

La WES, es una alternativa eficiente a la secuenciación del genoma completo (WGS) gracias a la simplificación del análisis de variantes, el almacenamiento de datos y a su menor costo de secuenciación lo que ha facilitado el diagnóstico clínico y la construcción de perfiles personalizados de riesgo de enfermedad, al ser de gran utilidad en la detección de variantes con frecuencia baja como son las Indel cuyas alternativas pueden ser cruciales en la propensión al desarrollo de enfermedades (Bamshad *et al.*, 2011; Barbittoff *et al.*, 2020). Gracias a esto, cada vez son más las investigaciones respecto al estudio de estas variantes y la construcción y enriquecimiento de bases de datos, ya que esto proporciona información relevante sobre los patrones globales de variación

genética en los humanos. Como en Lek *et al.* (2016), quienes describen el llamado de variantes conjuntas a partir del análisis de 60.706 exomas superando en magnitud a un gran número de bases de datos, con el fin de demostrar la aplicación del conjunto de datos en el análisis de variación, el descubrimiento de recurrencias mutacionales generalizadas y la interpretación de variantes de significado clínico. El proyecto UK Biobank, por otro lado, es un estudio de cohorte prospectivo con datos genéticos y fenotípicos profundos de aproximadamente 500.000 individuos de todo el Reino Unido, que recopiló datos de todo el genoma de los participantes, con el fin de brindar oportunidades para el descubrimiento de nuevas asociaciones genéticas y de bases genéticas de rasgos complejos (Bycroft *et al.*, 2018). A su vez, Montgomery *et al.*, (2013), estudiaron: “El origen, la evolución y el impacto funcional de las variantes cortas de inserción-delección identificadas en 179 genomas humanos”.

Frente a los estudios que emplean la NGS, vale la pena mencionar las investigaciones de Trudsø *et al.*, (2020) y el de Hwang *et al.*, (2017), donde se exponen métodos computacionales para la detección rápida de variantes complejas y empalme en lecturas cortas.

A nivel latinoamericano, cabe mencionar el estudio titulado: "Reference exome data for a Northern Brazilian population", que señala un punto de referencia útil para el diagnóstico de enfermedades raras en Brasil (Weeks *et al.*, 2020). En México se realizó un estudio de variaciones con el fin de conocer sus impactos o si guardaban relaciones con eventos de mestizaje estimados a partir de la variación de los nativos amerindios mexicanos. En el caso nacional, el proyecto Chocogen busca caracterizar la ascendencia genética chocoana y observar la relación entre determinantes genéticos de salud y enfermedad en esta región del Pacífico colombiano (Chande *et al.*, 2020; Ballesteros, 2019).

3.2.7. Frecuencia alélica

El estudio de la estructura, la historia genética de la humanidad, el flujo de los alelos en diferentes poblaciones y generaciones y la determinación de los mejores métodos para la identificación de susceptibilidad genética frente a patologías comunes, ha sido posible gracias al conocimiento de los principios y métodos de la genética de poblaciones (Nussbaum *et al.*, 2016).

Comúnmente, para describir la estructura genética de una población se enumeran los tipos y frecuencias de genotipos y alelos. El conjunto génico de una población se puede describir en términos de la frecuencia genotípica o la frecuencia alélica (Pierce, 2015).

La frecuencia alélica hace referencia a la recurrencia de un alelo particular frente al resto de alelos de un locus. Las frecuencias alélicas reflejan la diversidad genética y permiten describir el conjunto génico en menos términos, ya que el número de alelos es menor al de los genotipos. Adicionalmente, a diferencia de la frecuencia genotípica, tiene continuidad de una generación a la siguiente. Sus variaciones en función del tiempo pueden revelar deriva genética o la introducción de nuevas variantes en la población. Como explica Pierce (2015), es posible calcular la frecuencia alélica a partir del número de genotipos de la siguiente forma:

$$\text{Frecuencia alélica} = \frac{\text{Número de copias de un alelo}}{\text{Número de copias del total de alelos del locus}}$$

También, es posible realizar el cálculo de frecuencia alélica para un locus con dos alelos (Pierce, 2015). En caso de un locus bialélico las frecuencias para “A” y “a” representadas por p y q se calculan mediante las siguientes ecuaciones:

$$p = f(A) = \frac{2n_{AA} + n_{Aa}}{2N}$$

$$q = f(a) = \frac{2n_{aa} + n_{Aa}}{2N}$$

donde f es la frecuencia, N es el número total de individuos y n_{AA} , n_{Aa} y n_{aa} son el número de individuos para “AA”, “Aa” y “aa”, respectivamente, donde la suma de las frecuencias es igual a 1.

Finalmente, también es posible el cálculo de la frecuencia alélica partiendo de las frecuencias genotípicas como se expone en Pierce (2015).

3.2.7. *Equilibrio de Hardy-Weinberg*

Un principio general de amplia aplicación y alcance que constituye la piedra angular de la genética de poblaciones es el HWE. Este fue formulado en 1908 por Godfrey Hardy y Wilhelm Weinberg; consta de dos componentes fundamentales. En primer lugar, plantea para una serie de condiciones ideales en la población, una relación simple entre frecuencias alélicas y genotípicas de un locus genético, la cual es resultado de una distribución binomial parametrizada por la frecuencia alélica (AF) para marcadores bialélicos. Cabe mencionar que, así la población no cumpla con la relación en el estado inicial, el emparejamiento aleatorio permitirá que se establezca el equilibrio. En segundo lugar, afirma que la proporción de los genotipos no cambiará si las AF no cambian a través de las generaciones. Estas condiciones contemplan una población grande donde los emparejamientos ocurren al azar, la tasa de mutación de novo es indiscernible, no existe selección contra un genotipo concreto y no se presenta inmigración de individuos con frecuencias alélicas significativamente diferentes a las de la población endógena (Hao & Storey, 2019; Nussbaum, McInnes & Willard, 2016).

La determinación del HWE suele ser un paso preliminar en los análisis genómicos; de hecho; sirve como una verificación de la calidad de los datos o verificación de los supuestos del modelo, ya que se espera que este equilibrio se mantenga en la mayoría de marcadores genéticos. Cuando esto no sucede una de las causas puede ser producto de una estructura de población, por lo cual en estudios como el de Hao & Storey (2019), afirman que es importante identificar las otras posibles causas de desviaciones en el HWE ya que la gran mayoría de poblaciones naturales que se han

estudiado presentan algún grado de estructura (Hao & Storeyructu, 2019; Nussbaum, McInnes y Willard, 2016).

Se presentará entonces un desequilibrio con el incumplimiento de los supuestos previamente mencionados, por ejemplo, cuando hay un bajo número de individuos en la población donde se ve alterada la frecuencia alélica, o cuando ocurre la adición de alelos nuevos debido a mutaciones o cuando la población presenta subgrupos donde el apareamiento no sucede al azar, como en el caso de los seres humanos a causa de fenómenos de estratificación, emparejamiento dirigido y la consanguinidad. En cuanto al mantenimiento de frecuencias alélicas constantes intervienen el efecto de la mutación, selección y *fitness*, la deriva genética, la migración y flujo génico (Nussbaum, McInnes y Willard, 2016).

La prueba de HWE es sumamente relevante en análisis de datos genéticos poblacionales, su aplicación práctica más importante en genética médica es el asesoramiento genético en casos de trastornos autosómicos recesivos (McInnes y Willard, 2016).

4. Pregunta de investigación

¿Cuáles son las frecuencias alélicas en un conjunto de variantes (SNVs e Indels) obtenidas mediante secuenciación de exoma completo en pacientes colombianos de una IPS de diagnóstico genético?

5. Justificación

Los datos de secuenciación de genomas y exomas de miles de humanos en los últimos años han permitido la obtención de gran número de datos que proporcionan información relevante de los patrones globales de variación genética humana aportando a la comprensión histórica, biológica y poblacional humana y al fortalecimiento de los recursos para la interpretación clínica de variantes. El filtro por frecuencia de las variantes de secuencia juega un papel clave en el descubrimiento de variantes asociadas a enfermedades genéticas raras, cuya interpretación es útil para el proceso de diagnóstico clínico para condiciones específicas, y a su vez, en la identificación de variantes comunes en nuestras poblaciones no descritas en bases de datos de referencia (Kobayashi *et al.*, 2017; Lek *et al.*, 2016; Weeks, 2020). En la actualidad, los conjuntos de datos de la variación de secuencias de ADN disponibles de población latinoamericana y, más específicamente, colombiana, representan una fracción pequeña de todas las muestras secuenciadas (Lek *et al.*, 2016; Ballesteros, 2019).

En Colombia, se han identificado alrededor de 1920 enfermedades huérfanas que se encuentran incluidas en la resolución 430 de 2013. Estas enfermedades se caracterizan por ser crónicas, debilitantes y peligrosas para la vida y por poseer una prevalencia menor a 1 en cada 5000 personas. Estas agrupan enfermedades raras, ultra huérfanas (prevalencia $<1/1000000$) y olvidadas. Según el Ministerio de Salud y Protección Social (2022), en una porción significativa de esta población, las enfermedades son producto de mutaciones genéticas que pueden ser heredadas y, debido a su baja recurrencia, presentar vacíos investigativos que se traducen en dificultad diagnóstica. Es ahí donde adquieren relevancia los estudios de identificación y distribución de variantes asociadas a estos rasgos patogénicos dentro y entre las poblaciones, ya que proporcionan información sobre la base genética para los rasgos relacionados con la salud (Chande *et al.*, 2020).

6. Objetivos

6.1. *Objetivos General*

Analizar la variación genética, a partir de los datos de secuenciación de exoma completo, de una cohorte de pacientes con sospecha de enfermedad mendeliana en la IPS Biotecgen S.A.S. para el periodo 2019-2022.

6.2. *Objetivos específicos*

Identificar las variantes en el conjunto de datos de secuenciación de exoma completo de 632 muestras de pacientes con sospecha de enfermedad mendeliana de Biotecgen S.A.S. mediante un flujo de trabajo de control de calidad, alineamiento y llamado de variantes.

Determinar las frecuencias alélicas de las variantes tipo SNV e Indel en el conjunto de variantes de la cohorte de pacientes con sospecha de enfermedad mendeliana de Biotecgen S.A.S.

Integrar la matriz de datos de frecuencia alélica de variantes SNV e Indel en las herramientas de visualización de datos de los analistas de datos ómicos de Biotecgen S.A.S.

7. Método

Se tuvieron a disposición 706 muestras de pacientes de entre 0 y 90 años con sospecha de enfermedad genética remitidos para análisis mediante secuenciación de exoma completo en la IPS Biotecgen S.A.S. Posterior al proceso de remoción de duplicados y muestras parentales se obtuvieron 177 muestras a las que se les realizó el flujo de trabajo completo (control de calidad, alineamiento y llamado de variantes), y 455 que Biotecgen S.A.S. proporcionó ya procesadas hasta el primer filtrado, para un total de 632 muestras analizadas. Se incluyeron muestras pertenecientes a la cohorte 2019-2022 y de secuenciación de exoma completo con una profundidad de secuenciación de 50X o superior que contaban con consentimiento informado donde se autorizaba el uso con fines investigativos de sus datos biológicos (Formato en el Anexo 40). No se tuvieron en cuenta los datos de pacientes que manifestaron desacuerdo en el consentimiento.

Se empleó la tecnología de secuenciación por síntesis de Illumina con preparación de librerías mediante Agilent SureSelect Human All Exon V6.

7.1. Control de calidad

Inicialmente para el procesamiento de las muestras, se hizo uso de dos scripts. El primero permitió la unión de los múltiples archivos pertenecientes a una misma muestra obteniendo al final únicamente dos archivos por muestra (forward and reverse). Este procesó todos los archivos de lectura emparejada mediante el comando *pigz* para paralelizar la generación de archivos FASTQ comprimidos en formato GZ, indicando un número determinado de procesadores (8), y posteriormente eliminar los *inputs*.

El control de calidad se realizó mediante la herramienta FASTQC V0.11.1 a partir de los archivos FASTQ para posteriormente realizar la limpieza de las secuencias. Para ello, se empleó

Trimmomatic V0.39, con el cual se removieron los adaptadores y se escanearon cada una de las lecturas con una slidewindow de tamaño 5 cortando cuando la calidad promedio por base fue <20 .

7.2. Mapeo

Con los archivos limpios se procedió a ejecutar el pipeline de mapeo y llamado de variantes. El *input* de esta fase fueron los archivos FASTQ limpios, y el genoma humano de referencia GRCh37 - hg19 del Broad Institute. El mapeo se realizó mediante el algoritmo mem de BWA (Burrows-Wheeler Aligner) paralelizando con 40 threads. Con el archivo resultante se llevó a cabo el ordenamiento por coordenadas, la selección de lecturas pareadas y el ajuste de las etiquetas de los grupos de lectura mediante SAMtools, obteniendo un archivo BAM. Finalmente, se realizó el marcaje de duplicados con GATK *MarkduplicatesSpark* con los parámetros por defecto cuyo *output* en formato BAM fue indexado con el comando *index* usando 40 threads.

7.3. Llamado y genotipado de variantes

En esta fase se realizó la recalibración del puntaje de calidad por base (BQSR) mediante GATK (Genomic Analysis Toolkit) V.4.2.5.0 haciendo uso de dos comandos: 1. *BaseRecalibrator* y 2. *ApplyBQSR*. Con el primero, se generó la tabla de calibración. Luego se usó *ApplyBQSR* con los parámetros por defecto para obtener archivos BAM con las recalibraciones de calidad por base de las lecturas respecto a la tabla de calibración generada previamente. El proceso se hizo de manera paralela para cada cromosoma para optimizar el tiempo de corrida. La unión de los archivos BAM de cada cromosoma y el cálculo de estadísticos de mapeo se llevó a cabo mediante SAMtools.

Habiendo realizado estos pasos, se hizo el llamado de variantes con GATK mediante el comando *HaplotypeCaller*, el cual en sus versiones recientes es capaz de llamar SNVs e Indels simultáneamente, paralelizando el proceso nuevamente por cromosomas. Después se realizó la

concatenación de los archivos VCF de cada cromosoma mediante BCFTools. Posteriormente se realizó el filtrado “suave” con inteligencia artificial mediante las herramientas *CNNscoreVariants* (Red neuronal convolucional) y *FilterVariantTranches* de GATK. Igualmente, se llevó a cabo un filtrado “duro” empleando BCFTools para seleccionar variantes con profundidades mayores a 10 y calidad de genotipado mayor a 30.

7.4. Parentesco de las muestras

En primer lugar, se empleó *bcftools merge* para la unión de todas las muestras en un único archivo multisample. Con la herramienta *+setGT* de BCFtools se imputaron los genotipos faltantes (./.) como genotipos de referencia (0/0).

A partir del archivo multisample indexado, se llevó a cabo el cálculo del Coeficiente de Parentesco (Kinship) en las muestras mediante AKT para generar un archivo de texto que proporciona un formato con 6 columnas con el ID, las secuencias idénticas por descendencia, el coeficiente de kinship. Con este *output* finalmente se empleó AKT para obtener el listado de muestras emparentadas y los duplicados en formato de texto. Las muestras entre parentales y duplicadas, fueron eliminadas (dejando solo un miembro de cada familia). Para ello, se recopilaban los nombres de las muestras a descartar y se generó el archivo multisample limpio propicio para los pasos posteriores.

Posteriormente, se empleó el anotador y predictor de los efectos de variantes genéticas SNPeff V5.1. usando el genoma de referencia GRCh37 - hg19 del Broad Institute. SNPeff facilita la anotación de las variantes por su impacto en: bajo, moderado, alto y modificador. Las variantes sinónimas son categorizadas como de bajo impacto, las missense como impacto moderado, y finalmente las nonsense como de alto impacto (Cingolani *et al.*, 2012).

7.4.1. Dendograma

El programa NGSEP permitió el cálculo de la matriz de distancia empleando el algoritmo básico IBS (identidad por estado) en el archivo que contenía las variantes (en formato VCF) por medio del comando *VCFDistanceMatrixCalculator*. Luego de esto, en el programa SplitsTree se generó una retícula a partir de la matriz y se ejecutó el algoritmo de *Neighbor joining* el cuál creó el dendograma permitiendo la visualización gráfica de las familias.

7.5. Frecuencia alélica

Teniendo como *input* el multisample limpio se empleó el comando *VCFSummaryStats* de NGSEP, y con el archivo resultante se obtuvo la gráfica de número de SNVs en función de la MAF como se observa en la Figura 17. Así mismo la determinación de la media de transiciones identificadas y la determinación de la relación transición/transversión por genotipo.

Adicional a lo anterior, con el objetivo de aportar al análisis de datos ómicos en Biotecgen SAS, se generó a partir del archivo multisample, una tabla mediante *BCFTools* en la que se extrajeron 5 columnas: cromosoma, posición, referencia, alelo alternativo y los valores de frecuencia alélica encontrados para cada una de las posiciones en las 632 muestras analizadas. Con la cual; se procedió con la integración de la matriz de datos de frecuencia alélica de variantes SNV e Indel en la herramienta de visualización de datos VarSeq V2.2.5. Para ello, se convirtió el archivo VCF en un archivo de base de datos de formato cerrado TSF de Golden Helix mediante la herramienta “Convert Data Source” de VarSeq y se seleccionaron los campos de interés (identificador, referencia, el conteo de alelos alternativos y la frecuencia alélica) y se especificó el genoma de referencia GRCh37 - hg19 del Broad Institute. El recurso generado se incluyó en el repositorio de la herramienta para su integración en las plantillas de análisis de los médicos genetistas. Los detalles completos sobre la integración de la matriz se describen en el Anexo 2.

7.6. Equilibrio de Hardy-Weinberg y MAF

Teniendo como *input* el multisample limpio se empleó el comando VCFDiversityStats de NGSEP el cuál determina la relación entre heterocigosis esperada y observada y reporta el χ^2 y la significancia para el HWE. Con este archivo fue posible determinar el porcentaje de posiciones en equilibrio de Hardy-Weinberg.

7.7. Estructura poblacional

Los datos crudos fueron procesados para obtener las variantes de las muestras, las cuales se analizaron en conjunto con el set de variantes previamente descrito de manera posterior al cálculo de las frecuencias alélicas y el HWE. El conjunto de variantes resultante se filtró para incluir únicamente variantes con frecuencias superiores a 0,5% y se transformó en formato Plink utilizando la herramienta Plink 1.07. Finalmente, se realizó un análisis de componentes principales con Plink 1.9 y se llevó a cabo un análisis de estructura poblacional con Admixture 1.3.0 probando valores k de subgrupos poblacionales de k=2, k= 3 y k=4 (Alexander *et al.*, 2022). Se agregaron 17 muestras colombianas del proyecto 1000 genomas humanos al archivo VCF del conjunto de datos: (ERR031954, ERR250276, ERR250292, SRR070799, SRR077359, SRR100659, SRR100685, SRR100694, SRR100697, SRR100706, SRR100852, SRR100859, SRR1517867, SRR709959, SRR710109, SRR764737, SRR765984).

Figura 7

Resumen del método de procesamiento de las muestras de secuenciación del exoma completo, cohorte 2019-2022



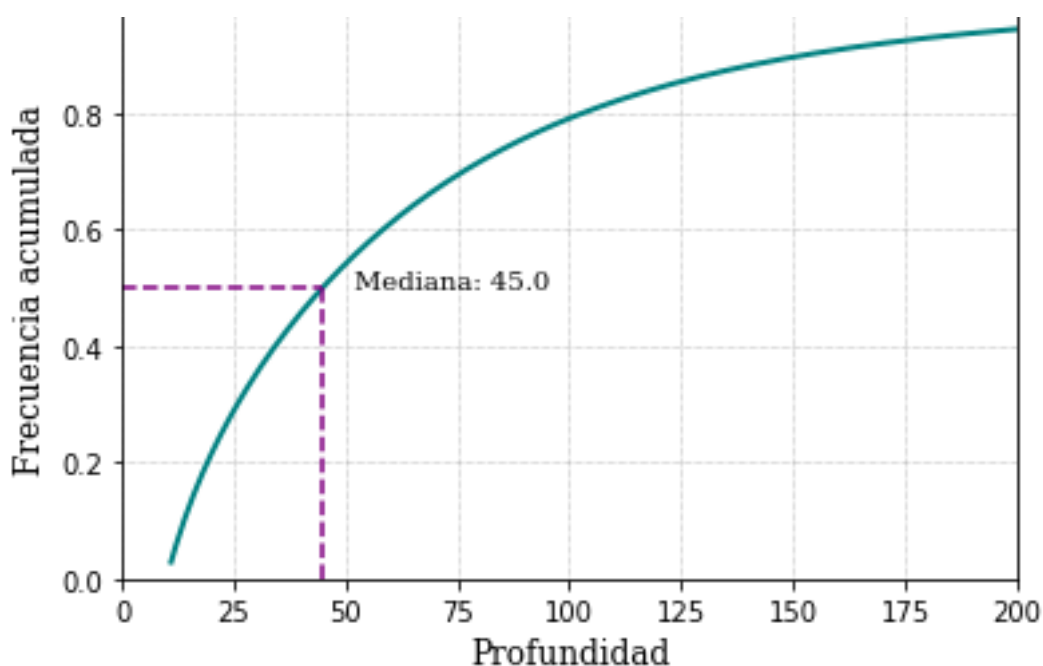
8. Resultados

8.1. Control de calidad

Como se mencionó previamente en la metodología, el trimming removió lecturas con calidades promedio por base menores a 20. La media de profundidad fue de 71,6X y como se observa en la Figura 8, la mediana de profundidad fue 45X.

Figura 8

Se observa la frecuencia acumulada de la profundidad de secuenciación de las variantes de todas las muestras de WES estudiadas junto con la mediana del conjunto



8.2. Alineamiento

En la Figura 9, se observa el resumen de las métricas de calidad de mapeo de secuencia para las 632 muestras de WES. Los valores de secuencias brutas totales presentaron una media de 70'498.044, el 50% de los valores se concentraron entre 50'512.850 y 81'705.064, esta presentó un máximo de 112'441.237 y un mínimo de 70'498.044 (Figura 9a). La métrica de lecturas correctamente

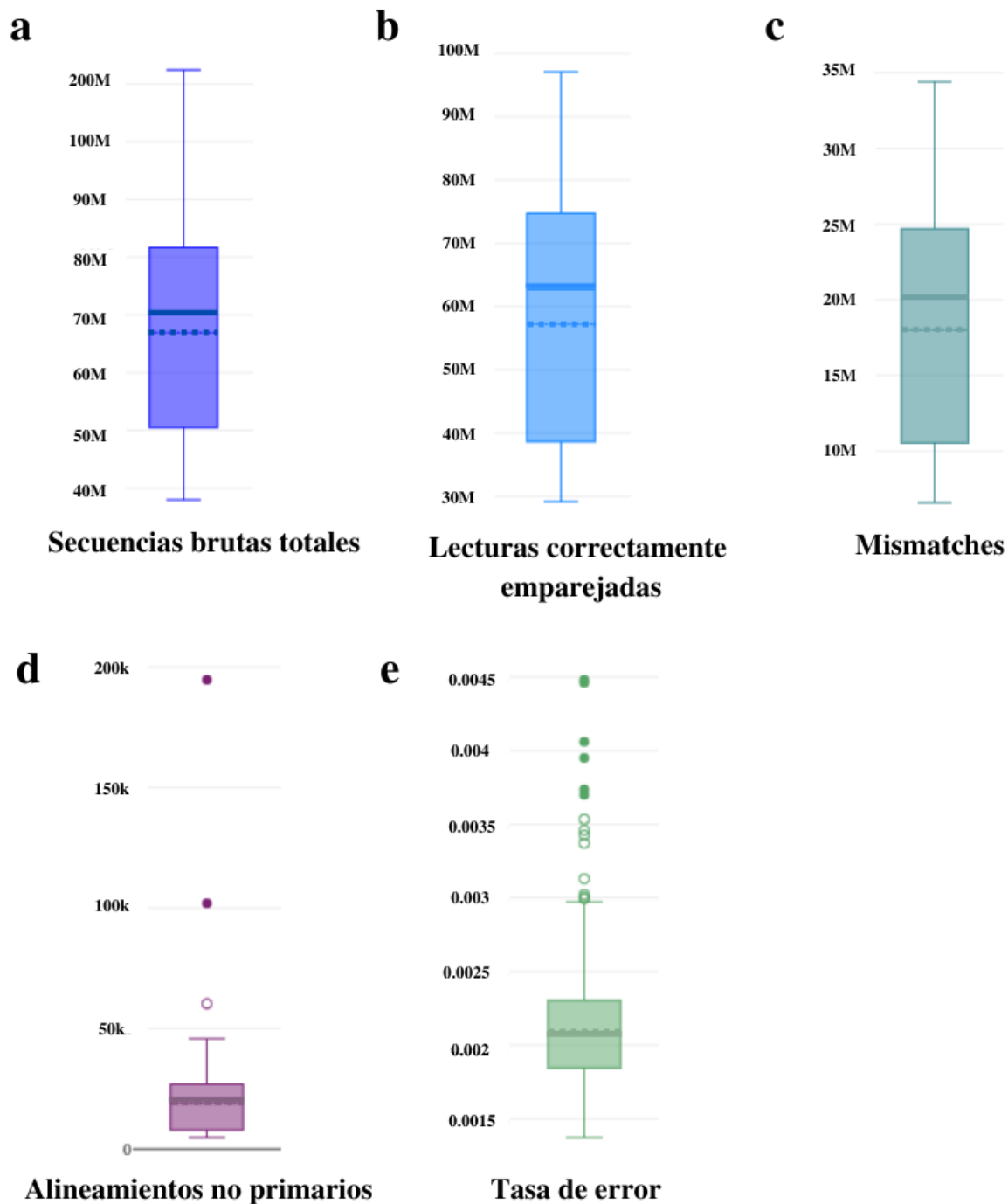
emparejadas presentó una media de 62'741.576, un rango intercuartil entre 38'615.444 y 74'719.778 y un máximo y mínimo de 97116380, 29155128 respectivamente, véase la Figura 9b. En la Figura 9c, se observan los mismatches, en esta métrica, el rango intercuartil estuvo entre 10'546.654 y 24'697.379, la media fue 20'148.716 con un mínimo de 6'593.616 y un máximo de 34'411.596. Cabe resaltar que, para las secuencias brutas totales, las lecturas correctamente emparejadas y los mismatches no presentaron valores atípicos.

Por otro lado, en la Figura 9d, los alineamientos no primarios presentaron un rango intercuartil entre 7.867 y 26.945, una media de 18.818, y un máximo y mínimo de 194.710 y 4.708 respectivamente. Los valores atípicos observados corresponden a las muestras B24924, D24990 y F24428.

En la Figura 9e, es posible observar la tasa de error de las métricas, la cual presentó un rango intercuartil entre 0,0018 y 0,0023, una media de 0,002. Los valores atípicos que se observan en la figura previamente citada corresponden a 14 muestras cuya tasa no superó el 0,0045.

Figura 9

Resumen de los estadísticos de las métricas de calidad obtenidas mediante SAMtools stats. a. Secuencias brutas totales. b. Lecturas correctamente emparejadas c. Mismatches d. Alineamientos no primarios y e. Tasa de error. La media y mediana se encuentran representadas por una línea continua y discontinua respectivamente



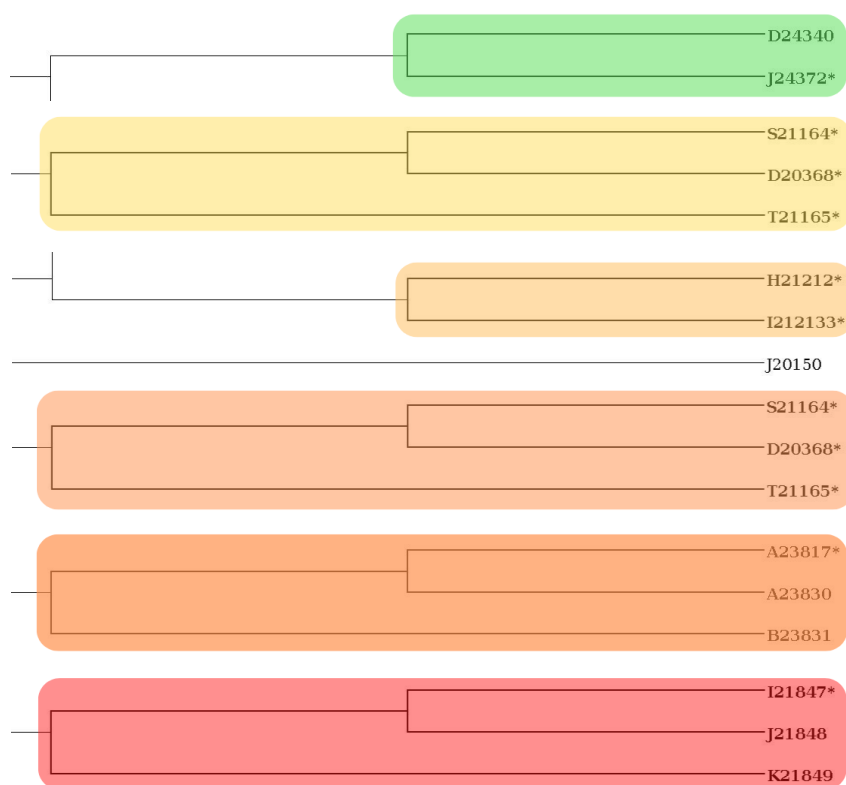
Finalmente, el filtrado “duro” seleccionó variantes con profundidades mayores a 10 y calidad de genotipado mayor a 30.

8.3. Parentesco

El software AKT reportó un total de 74 muestras entre parentales y duplicadas de las 706 totales las cuales fueron removidas, dejando únicamente la muestra asignada como hijo. En la Figura 10, es posible observar algunas de las familias reportadas de manera gráfica. Para ver el conjunto completo de familias de la cohorte 2019-2022 diríjase al Anexo 1.

Figura 10

Muestra de algunas familias reportadas por AKT de la cohorte 2019-2022 observadas en SplitsTree. Cada familia se encuentra diferenciada con un color. El asterisco () indica que la muestra es hija de las otras muestras del conjunto resaltado. Varios asteriscos dentro de una misma familia indican que las muestras son hermanas*



8.4. Llamado de variantes

El conjunto de 632 muestras presentó un total de 2 141 676 variantes: 1 881 670 SNVs bialélicos y 260 006 Indels cuya proporción puede apreciarse en la Figura 11. Por otro lado, SNPeff reportó un total de 494 912 variantes missense, 389 840 sinónimas y 8 231 nonsense.

Del total de variantes categorizadas por clase funcional, las missense fueron las que presentaron la mayor proporción, seguidas de las variantes sinónimas y, en menor medida, variantes nonsense, como se observa en la Figura 12.

Figura 11

Proporción de SNVs e Indels

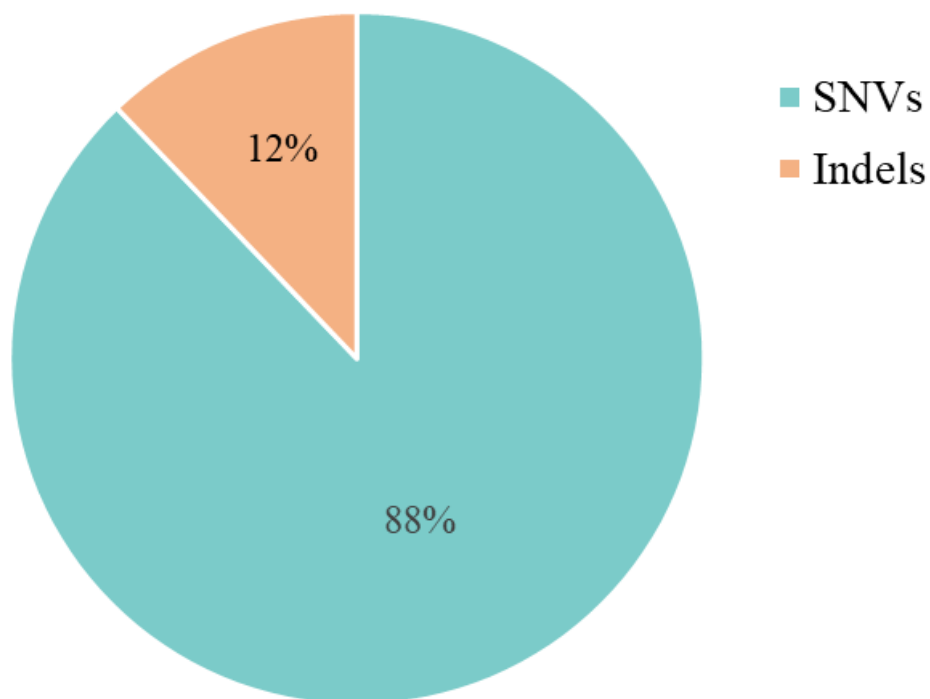
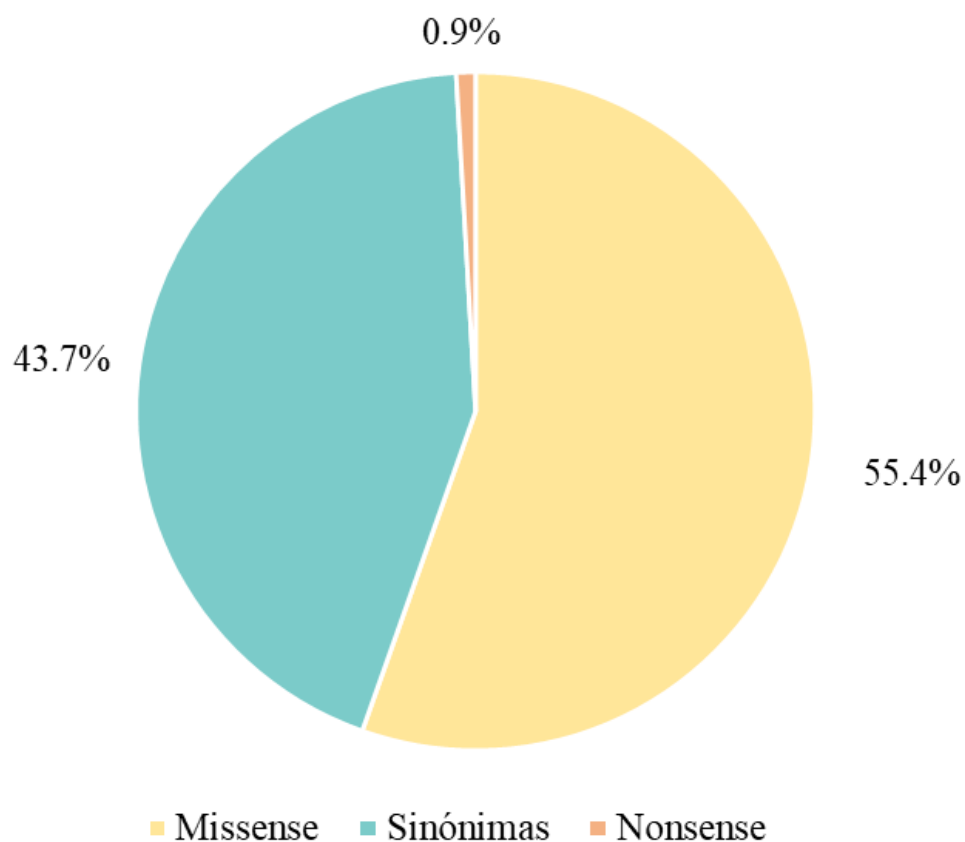


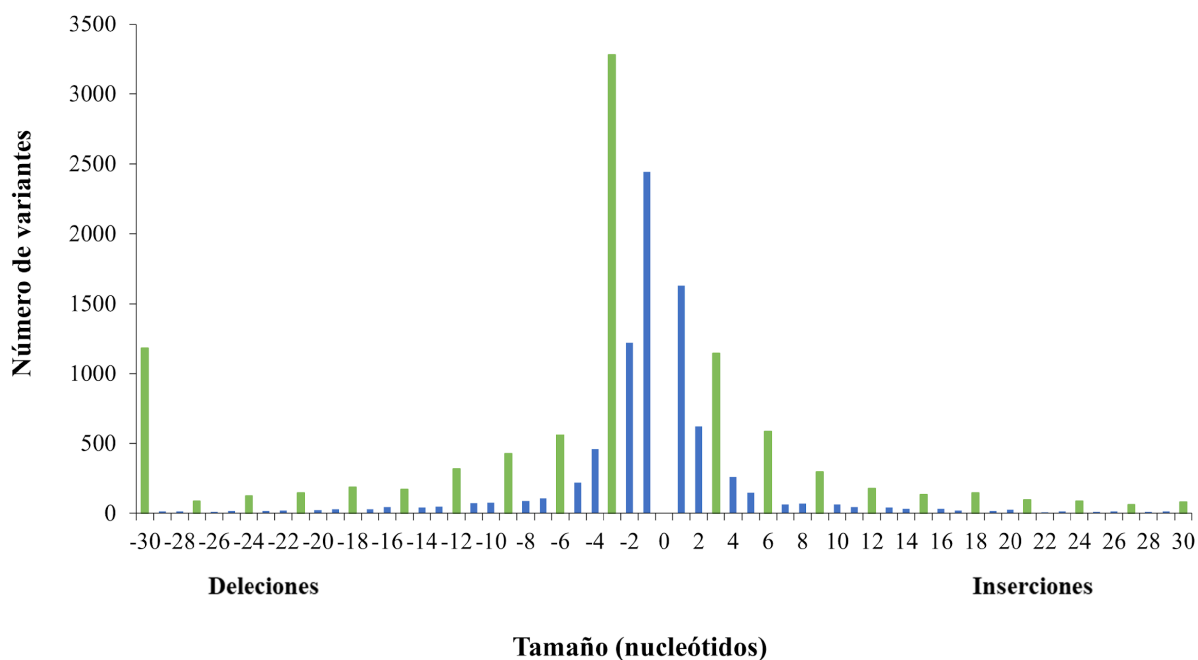
Figura 12

Proporción de la variación por clase funcional*8.5. Tamaño de Indel*

Se obtuvieron 260 006 Indels, las frameshift presentaron una frecuencia ligeramente menor comparadas con las in-frame que representaron el 52% de las variantes. En la Figura 13, se observan los picos en los múltiplos de 3 que reflejan una mayor proporción de variantes in-frame. Las deleciones de tres y treinta nucleótidos presentaron la mayor frecuencia con 3.282 y 1.183 variantes respectivamente. Por otro lado, la distribución por tamaño de las frameshift se concentró en las variantes de un nucleótido: 2.442 deleciones y 1.630 inserciones.

Figura 13

Tamaño de Indel. Se observa la distribución de frecuencias de los tamaños de Indel. En color verde se resaltan las variantes in-frame, las cuales generan picos en los múltiplos de 3. En azul se representan las variantes frameshift



Se encontró que 5323 genes se vieron afectados por Indels frameshift. En la Tabla 1, es posible observar algunos de los genes afectados por estas variantes y el conteo de estos Indel.

Tabla 1

Conteo de frameshift en algunos genes afectados

Nombre del gen	variants_effect_frameshift_variant
MUC4	254
MUC1	85
MUC19	85
KIAA1522	26
HR	28

HRC	28
HRCT1	28
F8	29
ZNF8	29
ZNF880	29
MUC16	30
C6	33
MUC6	33
TBP	33
ANKLE1	39
KL	39
ATXN3	42
TXN	42
HLA-DRB1	45
RB1	45
CCDC40	46
CDC40	46
HLA-DRB5	56

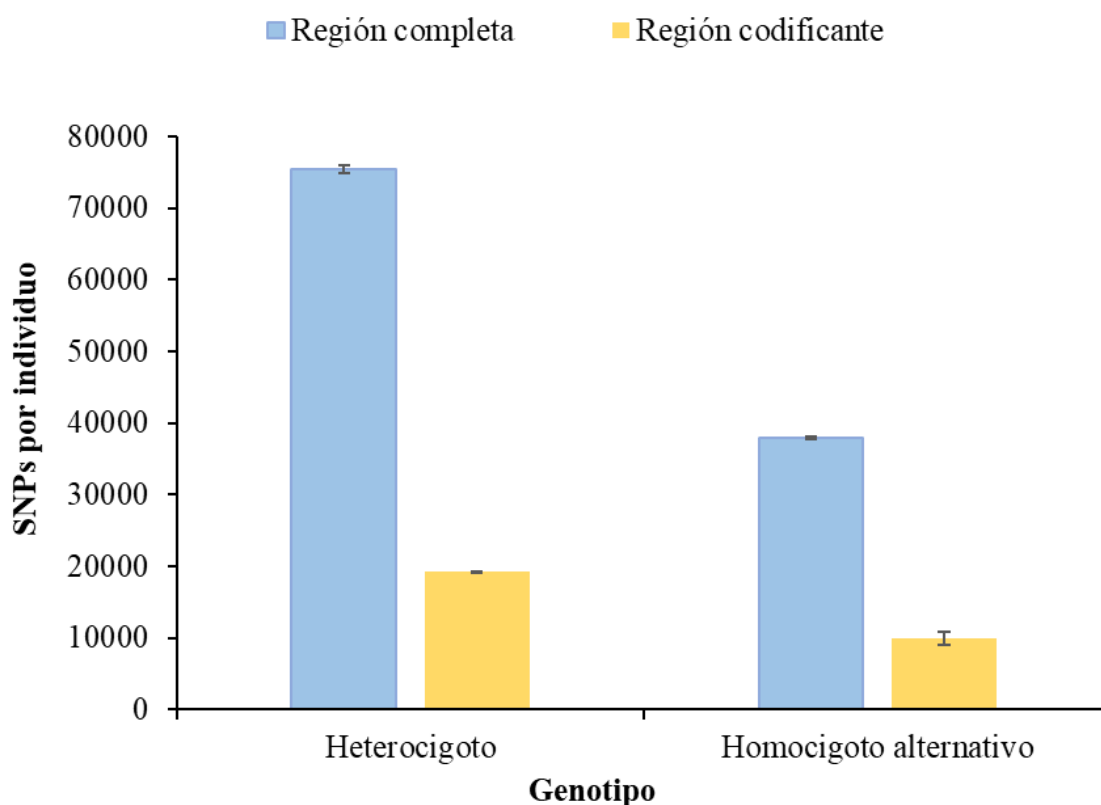
8.x. SNVs por genotipo

El total de llamadas de genotipo fue de 1 189 215 440. Se identificaron un total de 1 881 670 SNVs bialélicos. Posterior al filtrado, el total de llamadas de genotipo fue de 270 074 456 y se identificaron un total de 427 333 SNVs bialélicos.

Previo al filtrado se identificaron 75 401 SNVs heterocigotos y 37 929 homocigotos. Posterior al filtrado, se encontró que el individuo de la cohorte promedio tiene 19 222 SNVs heterocigotos y 9 944 homocigotos (Figura 14).

Figura 14

Media de SNPs identificados en el llamado de variantes por genotipo. Las barras amarillas representan la región de captura objetivo “target region” y en color azul se encuentra la media de las SNVs llamadas tanto dentro como fuera de la región objetivo. Las barras de error representan el error estándar

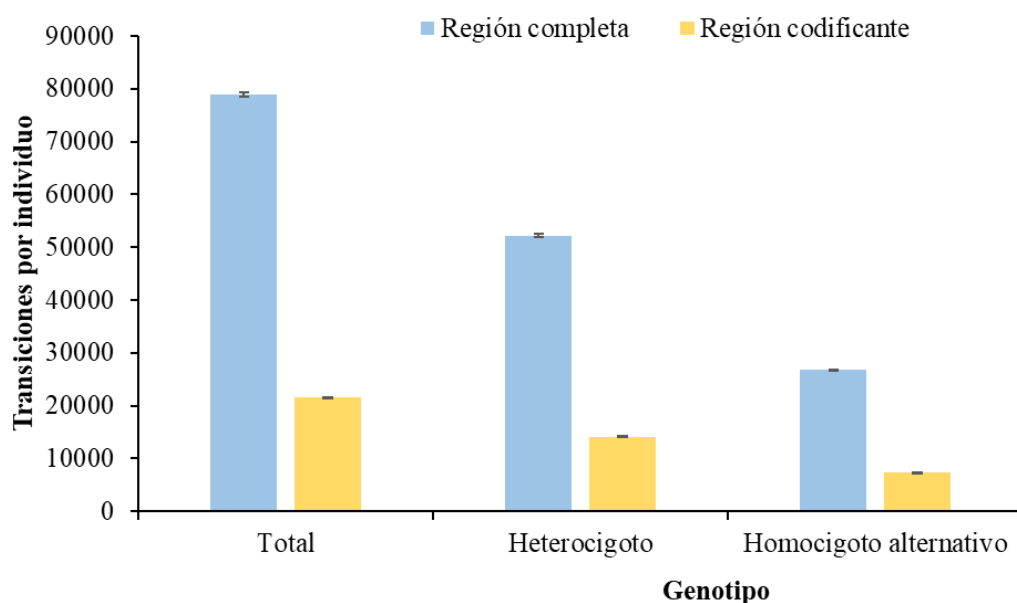


El total de llamadas de genotipo fue de 1 189 215 440. Se identificaron un total de 1 881 670 SNVs bialélicos. Posterior al filtrado, el total de llamadas de genotipo fue de 270 074 456 y se identificaron un total de 427 333 SNVs bialélicos.

Previo al filtrado se identificaron 75 401 SNVs heterocigotos y 37 929 homocigotos. Posterior al filtrado se encontró que el individuo de la cohorte promedio tiene 19 222 SNVs heterocigotos y 9 944 homocigotos en la región codificante (Figura 14).

Figura 15

Media de transiciones identificadas en el llamado de variantes. Las barras amarillas representan las transiciones en la región de captura objetivo “target region” y en azul se agrupan las transiciones tanto dentro como fuera de la región objetivo. En el eje x, se observan el total de transiciones y, las transiciones discriminadas por genotipo. Las barras de error representan el error estándar

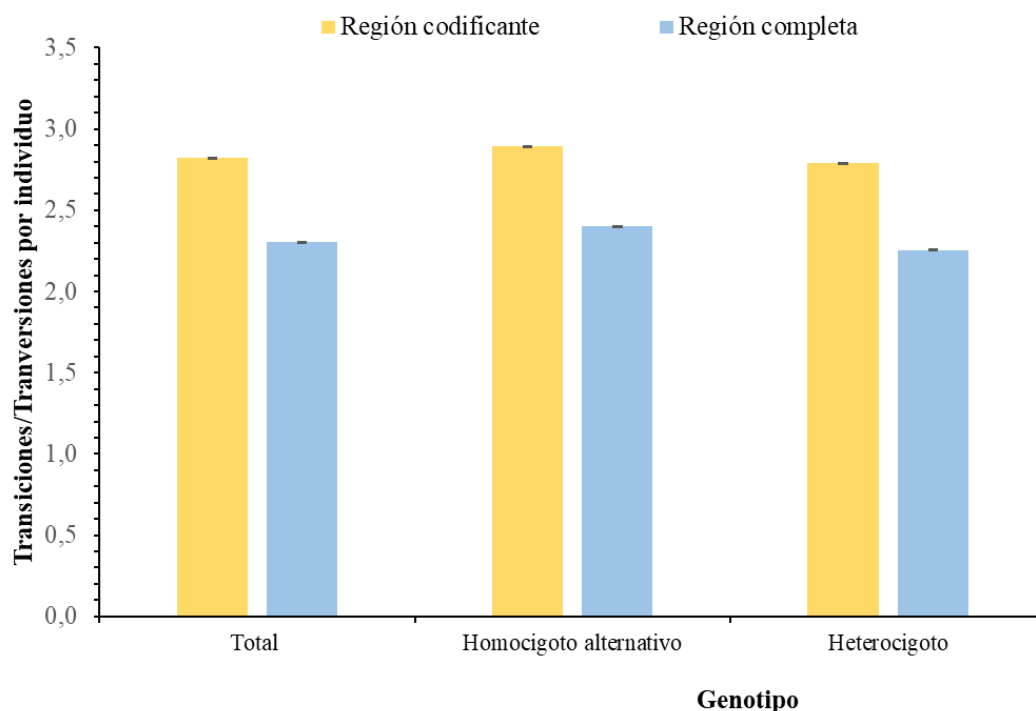


Previo al filtrado, la media de las transiciones fue de 78 918, 52 159 para el genotipo heterocigoto y 26 758 en el homocigoto. Posterior al filtrado, la media de transiciones se redujo notablemente a un total de 21 534 transiciones de las cuales, 14 146 fueron del heterocigoto y 7 388 del homocigoto como se observa en la Figura 15.

Figura 16

Media de transiciones/transversiones identificadas en el llamado de variantes total y por genotipo.

Las barras amarillas representan la región de captura objetivo “target region” y el azul abarca la región comprendida tanto dentro como fuera de la región objetivo (previa al filtrado). Las barras de error representan el error estándar.



Estudios previos reportan que las regiones del exoma deben tener valores de 3,0 para la relación transición/transversión (Tr/Tv). Una relación Tr/Tv mayor, indica generalmente una mejor calidad siempre y cuando esta no sea $> 4,0$. Si por otro lado, se observa un valor menor al esperado ($< 3,0$), para los SNP del exoma (SNVs con $MAF > 0.05$), un aumento en filtros como el de profundidad y la puntuación de calidad del genotipo contribuye al aumento de esta relación (Wang *et al.*, 2015).

Como es posible observar en la Figura 16, previo al filtrado la relación transición/transversión (Tr/Tv) presentó valores en un rango de 2,2 a 2,5 mientras que, dentro de los exones esta relación presentó valores entre 2.7 y 3. Estas métricas del conjunto de datos fueron consistentes con lo

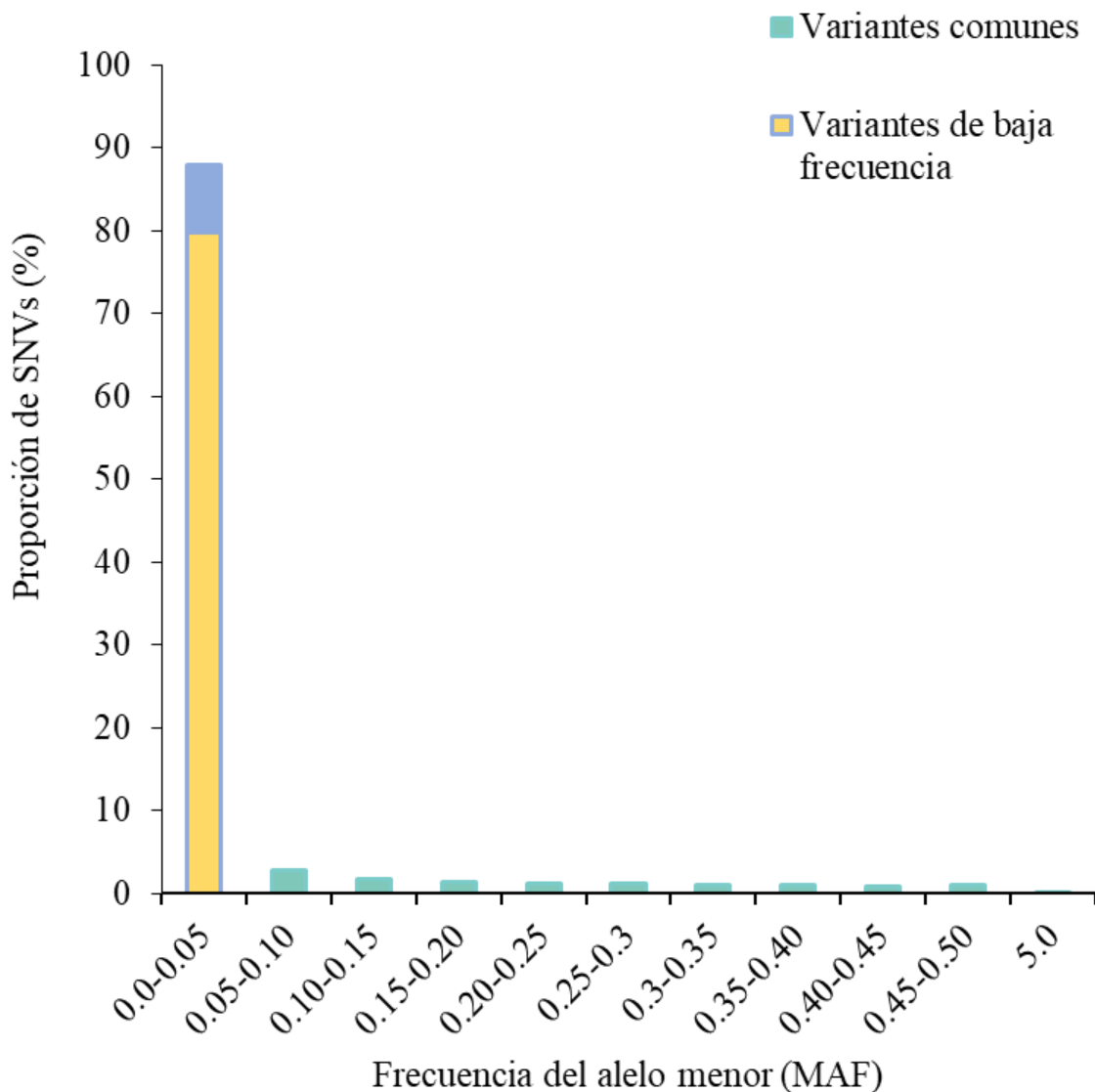
reportado en otras fuentes (Bainbridge *et al.*; 2011; Guo *et al.*, 2012; Wang *et al.*, 2015; Xu *et al.*, 2019).

8.6. Frecuencia alélica y variantes en equilibrio de Hardy-Weinberg

El 88% de los valores de MAF corresponden a SNVs de baja frecuencia, como se observa en la Figura 17. Se evidenció un 92% de posiciones en equilibrio de Hardy-Weinberg (2041270 posiciones con $p > 0.05$) y desviaciones en 187134 posiciones (2% restante).

Figura 17

Distribución de la frecuencia del alelo menor (MAF) en el conjunto de variantes SNV. En la primera barra se observa la proporción de SNVs de baja frecuencia ($MAF < 0.05$) y se diferencia la proporción de variantes raras ($MAF < 0.01$) en color amarillo. El 88% de SNVs fueron de baja frecuencia



8.6. Estructura poblacional

La Figura 16a representa el análisis por componentes principales del conjunto. Se observan dos agrupaciones de muestras. La Figura 16b representa las subpoblaciones encontradas por Admixture asumiendo 2, 3 y 4 subgrupos. En los tres casos, se observan dos subgrupos poblacionales dominantes. En el análisis con 3 grupos, aparece una proporción menor de asignación al grupo 3 y, con 4 grupos, se observa una proporción también limitada de asignación a los grupos 3 y 4. Es importante notar que por limitaciones en el uso de los datos no fue posible recuperar el origen

geográfico de las muestras. A pesar de esto, se analizaron tres muestras de 1000 genomas humanos en conjunto con los datos, las cuales arrojaron los puntajes estimados de proporción de Admixture detallados en la Tabla 2.

Tabla 2

Estimación de la proporción de subpoblaciones con Admixture para las 17 muestras colombianas (CLM) de 1000 genomas humanos

Estimación de proporción de Admixture por subpoblación (%)		
k=2		
Muestra	1	2
ERR031954	95,8721	0,041279
ERR250276	90,0675	0,099325
ERR250292	87,9637	0,120363
SRR070799	90,7456	0,092544
SRR077359	91,7188	0,082812
SRR100659	99,999	0,00001
SRR100685	99,999	0,00001
SRR100694	99,999	0,00001
SRR100697	99,999	0,00001
SRR100706	99,999	0,00001
SRR100852	99,999	0,00001
SRR100859	99,7867	0,002133
SRR1517867	98,0701	0,019299
SRR709959	92,1098	0,078902
SRR710109	91,5511	0,084489
SRR764737	90,8951	0,091049

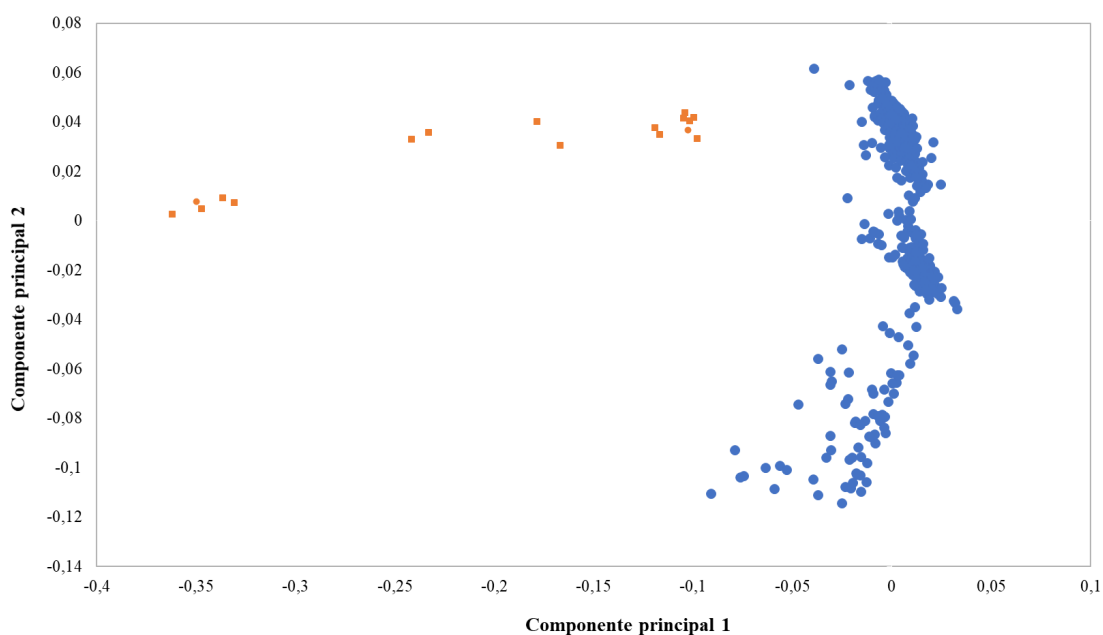
SRR765984

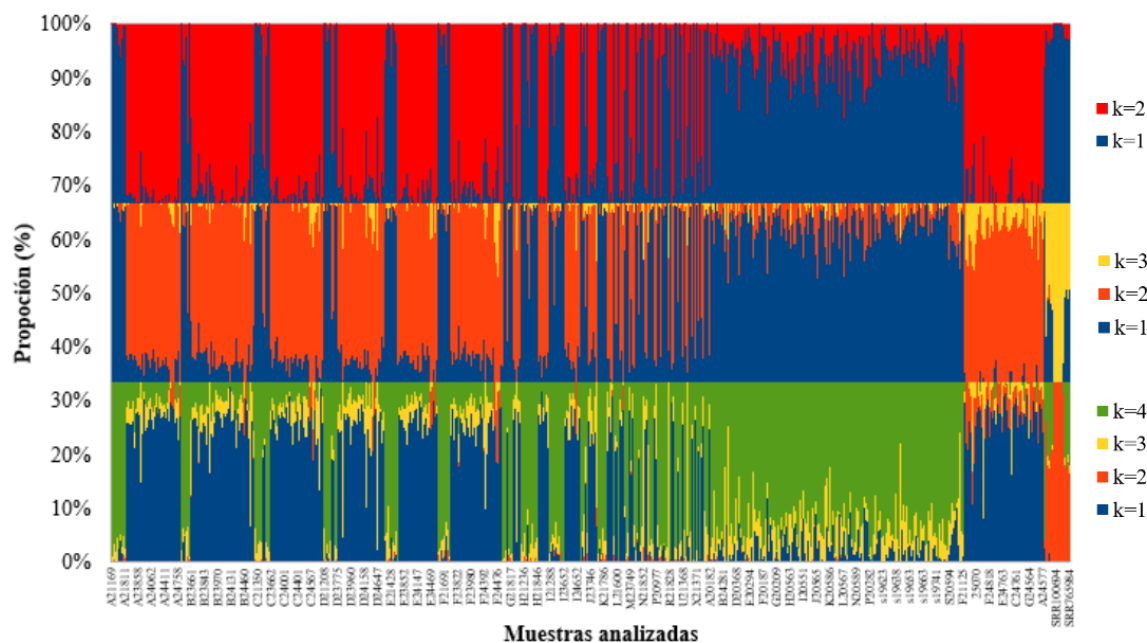
91,8428

0,081572

Figura 18

Análisis de la estructura poblacional con Admixture a. Gráfica del Análisis de Componentes Principales de las muestras analizadas. b. Representación de barras apiladas de los puntajes Q de cada muestra para cada uno de los grupos poblacionales asumidos por el modelo de Admixture. La gráfica superior representa los hallazgos con un supuesto de $k=2$, la del medio representa $k=3$ y la inferior representa $k=4$

**a.**



b.

8.7. Integración de la matriz de frecuencia alélica en VarSeq

A continuación, en la Figura 19, se aprecia el resultado de la integración de la matriz de frecuencias alélicas de las variantes en la herramienta de interpretación empleada en Biotecgen S.A.S.

Figura 19

Matriz de frecuencia alélica integrada en VarSeq

Variants: 728										Samples: 1 X										Coverage Regions:...										Log X										CNVs: 1,382 X																			
Variants										Filter Variants: 90261																				Variants: 728																													
Variant Info					Flags for...					90261										MERGED_31OCT22_AF										ACMG Sample Classifier for 90261																													
Chr...	Ref/Alt	EV	PF	SF	Zygosity	DP	VAF	GT	GQ	Filter	Identifier	Reference	Alternates	Alt Allele Counts (AC)	Alt Allele Freq (AF)	Ref/Alt	Gene Name	Gene Inheritance																																									
1289835	C/T				Heterozygous	207	0.492754	0/1	672	PASS	?	?	?	?	?	C/T	MXRAB	Default (Recess)																																									
2433809	G/A				Heterozygous	63	0.634921	0/1	180	PASS	?	?	A	11	0.00870253	G/A	PLCH2	Default (Recess)																																									
9640037	A/G				Heterozygous	203	0.35468	0/1	344	PASS	?	?	?	?	?	A/G	SLC25A33	Default (Recess)																																									
12854090	T/A				Heterozygous	102	0.343137	0/1	176	PASS	rs79698223	T	A	619	0.489715	T/A	PRAMEF1	Default (Recess)																																									
12907350	C/T				Heterozygous	92	0.304348	0/1	129	PASS	rs2359486	C	T	523	0.413766	C/T	HNRNPCL1	Default (Recess)																																									
16736538	C/T				Heterozygous	98	0.5	0/1	372	PASS	rs2011826...	C	T	4	0.00316456	C/T	SPATA21	Default (Recess)																																									
16918473	G/A				Heterozygous	299	0.264214	0/1	64	PASS	?	G	A	124	0.0981013	G/A	NBPFI	Default (Recess)																																									
17250924	C/T				Heterozygous	170	0.417647	0/1	418	PASS	rs1492497...	C	T	3	0.00237342	C/T	CROCC	Default (Recess)																																									
24932215	G/T				Heterozygous	260	0.5	0/1	860	PASS	?	G	T	2	0.00158228	G/T	NCMAP	Default (Recess)																																									
26608812	CCAG...				Heterozygous	?	0.512821	0/1	238	PASS	rs6672357...	GGGTCCA...	-,T.TCCGG...	4,133.3.1.720.112.13...	0.00316456.0.1052...	CCAGGACA...	UBXN11	Default (Recess)																																									
26608843	C/A				Heterozygous	33	0.545455	0/1	158	PASS	rs6661497...	TCCAGGAC...	T.TCCGG...	133.3.1.13.3.32.3.6.5...	0.105222.0.002373...	C/A	UBXN11	Default (Recess)																																									
26608852	G/A				Heterozygous	32	0.59375	0/1	149	PASS	rs7570948...	TCCAGGAC...	T.TCCGG...	133.3.1.6.51.0.404.72...	0.105222.0.002373...	G/A	UBXN11	Default (Recess)																																									
32049135	A/G				Heterozygous	231	0.402597	0/1	508	PASS	?	A	G	1	0.000791139	A/G	TINAGL1	Default (Recess)																																									
32085210	G/A				Heterozygous	150	0.486667	0/1	503	PASS	rs75232948	G	A	4	0.00316456	G/A	HCRT1	Default (Recess)																																									
40981165	G/A				Heterozygous	186	0.467742	0/1	566	PASS	?	G	A	2	0.00158228	G/A	EXOS	Default (Recess)																																									
43215930	C/T				Heterozygous	208	0.504808	0/1	691	PASS	rs11581921	C	T	50	0.039557	C/T	P3H1	Recess																																									
45444040	C/T				Heterozygous	109	0.431193	0/1	328	PASS	?	?	?	?	?	C/T	EIF2B3	Recess																																									
46193458	C/T				Heterozygous	41	0.414634	0/1	180	PASS	?	?	?	?	?	C/T	IPP	Default (Recess)																																									
55523033	A/G				Homozygous V...	101	1	1/1	301	PASS	rs509504	A	G	1248	0.987342	A/G	PCSK9	Domini																																									
59248085	G/C				Heterozygous	168	0.488095	0/1	540	PASS	?	G	C	4	0.00316456	G/C	JUN	Default (Recess)																																									
92944315	AG/-				Heterozygous	?	0.441176	/1	154	PASS	rs7239074...	AGAGAGA...	AGAGAGA...	84.24.312.18.36.173...	0.0664557.0.01898...	AG/-	GF1I	Domini																																									
44461749	C/T				Heterozygous	94	0.468085	0/1	330	PASS	rs77293072	C	T	15	0.0118671	C/T	ABCA4	Domini																																									
44505604	A/C				Heterozygous	167	0.580838	0/1	416	PASS	rs61750126	A	C	17	0.0134494	A/C	ABCA4	Domini																																									
48187098	T/C				Heterozygous	84	0.511905	0/1	335	PASS	rs2005629...	T	C	7	0.00553797	T/C	DPYD	Recess																																									
10086074	G/A				Heterozygous	229	0.39738	0/1	462	PASS	rs1826388...	G	A	9	0.00712025	G/A	GPR61	Default (Recess)																																									
15220601	C/A				Heterozygous	70	0.528571	0/1	283	PASS	?	C	A	1	0.000791139	C/A	AMPD1	Recess																																									
44598641	C/T				Heterozygous	5	0.4	0/1	66	PASS	rs1698687	C	T	10	0.00791139	C/T	NBPFI	Default (Recess)																																									

9. Análisis y discusión

9.1. Control de calidad

La media y mediana de profundidad presentaron valores acordes a lo esperado teniendo en cuenta los criterios preestablecidos para el procesamiento de las muestras, es decir, variantes con calidad de genotipado mayor a 30 y como se observa en la Figura 8, profundidades mayores a 10 donde la media de profundidad fue de 71.6X y la mediana de 45X.

9.2. Alineamiento

Las métricas de secuencias brutas totales, las lecturas correctamente emparejadas y los mismatches no presentaron valores atípicos. Por otro lado, los valores atípicos de alineamientos no primarios corresponden a las muestras: B24924, D24990 y F24428 cuya tasa de error de alineamiento fue baja y los valores para esta, se encontraron dentro de lo esperado.

Sumado a lo anterior, pese a que para la tasa de error se reportaron 14 valores atípicos; estos no superaron el 0,0045. Todo lo anterior, permite evidenciar que el mapeo se realizó de manera correcta y los valores obtenidos para las diferentes métricas fueron coherentes con lo esperado.

9.3. Parentesco

La determinación del parentesco mediante AKT, gracias a la cuál se removieron las 74 muestras que agrupaban parentales y duplicados, fue de suma importancia ya que permitió evitar sesgos en el análisis por una sobrerrepresentación en el conjunto estudiado. Las familias removidas se pueden observar en el Anexo 1.

9.4. Proporción de la variación por clase funcional

De acuerdo a lo esperado, el 88% de las variantes fueron SNVs lo que concuerda con múltiples autores como Kohlmeier (2020), INCIFOR (2021) y Ping (2016), quienes afirman estas variantes son las más comunes y determinan el mayor porcentaje de variabilidad genética de los organismos ya que como se menciona en Mordoh (2019), el 90% de la variación genética humana está representada por SNVs.

En “Analysis of protein-coding genetic variation in 60,706 humans”, Lek *et al.*, (2016) determinaron el número y distribución de las frecuencias de Indels por tamaño para 60706 exomas con origen europeo, del sur y sudeste asiático y en menor proporción, africano y americano mezclado reportando proporciones más bajas para variantes missense y nonsense atribuidas a presión selectiva. A su vez Sharma *et al.*, (2019), reportaron una proporción mayor para variantes sinónimas seguida de missense y nonsense. En el caso de este estudio, como se observó previamente en la Figura 12, la mayor proporción la presentaron las variantes missense (55.4%) las cuales se clasifican como variantes de impacto moderado, seguidas por las variantes sinónimas (43.7%) de bajo impacto, presentando proporciones cercanas entre sí y representando el 96% de todas las variantes analizadas lo que es coherente con Sharma *et al.*, (2019), quienes reportaron un patrón de recurrencia similar entre variantes sinónimas y missense.

Las variantes de alto impacto nonsense, al igual que en el estudio de Lek *et al.*, (2016) y Sharma *et al.*, (2019) tuvieron la más baja proporción, representando tan solo el 3% de las variantes identificadas. En este análisis, el 63% de los Indels presentaron un tamaño menor a 6 bases y el 90% presentaron longitudes menores a 30 bases siendo las deleciones más cortas (al igual que en el estudio de Lek *et al.*, (2016)), las variantes más comunes.

Por otro lado, la distribución por tamaño de las frameshift se concentró en mayor medida en las Indel de un nucleótido. Un total de 5323 genes se vieron afectados por Indels frameshift siendo MUC4, MUC1 y MUC19 los que presentaron el mayor número de frameshifts. En el estudio de

Oh *et al.*, (2015), también se encontró que genes MUC como el MUC4 se vieron alterados por frameshifts a los que atribuyeron presuntivamente la pérdida de la expresión.

9.5. Frecuencia alélica, variantes en equilibrio de Hardy-Weinberg y estructura poblacional

Colombia presenta patrones de mezcla bien diferenciados entre sus diferentes regiones, es el país latinoamericano con la mayor variabilidad interpoblacional con una alta ascendencia europea, nativa americana y africana (en la costa del Caribe y la del Pacífico) debida a la coexistencia con estos grupos durante más de 500 años (Ossa *et al.*, 2021). Esta alta variabilidad, hace sumamente relevante el estudio y la determinación de la frecuencia alélica de las variantes en pacientes lo que contribuye a brindar una interpretación clínica adecuada de estas variantes en el contexto específico de diagnóstico genético del país (Lek *et al.*, 2017; Ossa *et al.*, 2021).

La MAF mostró un mayor número de variantes de baja frecuencia ($MAF < 0,05$) con un porcentaje del 88% de los SNVs de los cuales el 78% corresponden a variantes raras. De acuerdo con el estudio del International HapMap Consortium (2005), la mayoría de SNVs observadas en la Encyclopedia of DNA Elements (ENCODE) son raras. El 46% poseía un $MAF < 0,05$. A su vez se ha apreciado que la distribución de frecuencias de alelos menores (MAF) está fuertemente sesgada hacia un exceso de variantes raras: más de un tercio de todos los SNVs tienen frecuencias por debajo del 5% según Gibson (2012).

Para los valores de equilibrio de Hardy-Weinberg, la mayor proporción de posiciones (92%), se encontraron en equilibrio ($p > 0.05$). Sin embargo, se detectaron algunas desviaciones que presentaron un $p < 0,05$. Ossa *et al.*, (2021), atribuyeron las desviaciones en el HWE a un exceso de homocigotos producto de la subestructura poblacional que presenta la población andina, chocoana y amazónica; lo cual es posible que se reflejase también en este análisis donde los puntajes $p < 0,05$,

podrían corresponder en su gran mayoría a posiciones en muestras provenientes en este caso, del eje cafetero y el norte del Valle del Cauca.

Por otro lado, las frecuencias alélicas integradas en VarSeq son de gran utilidad ya que los genetistas clínicos cuentan con una serie de recursos como son las guías del American College of Medical Genetics and Genomics, las cuales contienen una serie de criterios para la interpretación de variantes de secuencia, donde algunos requieren conocer las frecuencias poblacionales de las variantes para establecer la clasificación y ponderación, como es el caso de los criterios poblacionales: PM2, BA1, BS1 y BS2 (Richards *et al.*, 2015). Adicionalmente, se ha demostrado que las bases de datos disponibles como GnomAD, lanzada originalmente como ExAC en el 2014, actualmente no reflejan adecuadamente las frecuencias alélicas de poblaciones como la colombiana ya que el total de la población “latina” en esta base de datos representa tan solo el 13% y corresponde a “estadounidenses mezclados” (Karczewski *et al.*, 2020). Por lo tanto, la matriz de datos de frecuencia alélica de los pacientes de Biotecgen, que se encuentra actualmente disponible en la IPS cobra gran relevancia para complementar la información de la que disponen los genetistas de Biotecgen al momento de clasificar las variantes y es potencialmente valiosa para futuros estudios de ancestría y estructura poblacional en la cohorte de pacientes colombianos de Biotecgen S.A.S.

Finalmente, el análisis de Admixture permite observar diferentes subpoblaciones dentro de la cohorte analizada. Cómo es posible observar en la Tabla 2. Para un $k=2$ todas las muestras del Proyecto 1000 Genomas presentaron un porcentaje de asignación al subgrupo 2 del 90% o superior. Mientras que para un $k=3$, las muestras de 1000 genomas, presentaron porcentajes de asignación al subgrupo 3 que oscilaron entre un 40% y un 100%, mientras que las muestras de biotecgen presentaron mayores porcentajes de asignación a los subgrupos $k=2$ y $k=3$. Asumiendo 4 subgrupos se encontró que hubo un mayor porcentaje de asignación al sugrupo 4 en las muestras de Biotecgen

mientras que las muestras de población colombiana descargadas de 1000 genomas, presentaron en su gran mayoría porcentajes de asignación de 90% a la subpoblación 2.

Por limitaciones en el uso de metadatos asociados a las muestras analizadas para el presente estudio, no es posible conocer el origen geográfico de cada muestra de Biotecgen. Sin embargo, se observan dos subgrupos con un mayor porcentaje de asignación en la segunda gráfica correspondientes a ($k=1$ y $k=2$) similares a cuando se asumen dos subgrupos. Llama la atención que las 17 muestras de 1000 genomas humanos las cuales se observan en color naranja provienen de Antioquia en su mayoría, en la Figura 18, presentaron una alta dispersión. Según datos internos de la IPS Biotecgen S.A.S., un volumen importante de las muestras procesadas proviene de la región suroccidente del país (Valle del Cauca, Cauca y Nariño), por lo cual se podría plantear como hipótesis que las subpoblaciones observadas representan grupos con diferentes ancestrías, como podría ser un componente de origen afrocolombiano o nativo americano.

10. Conclusiones

El flujo de trabajo de control de calidad, alineamiento y llamado de variantes permitió llevar a cabo la identificación y posterior análisis de las variantes del conjunto de datos de WES de las 632 muestras de pacientes de Biotecgen S.A.S. Este conjunto de variantes de pacientes colombianos presentó frecuencias alélicas bajas donde la mayoría de las posiciones presentaron una $MAF < 0,01$ lo que nos indica una mayor proporción de variantes raras. La proporción por consecuencia funcional de las variantes indicó un 55.4% de variantes missense seguido de un 43.7% de variantes sinónimas y un 0.9% de variantes nonsense categorizadas con un impacto moderado, bajo y alto respectivamente. Las delecciones más cortas fueron las variantes más comunes para el tamaño de Indel donde el 63% de estos presentaron un tamaño menor a 6 bases y el 90% presentaron longitudes menores a 30 bases.

La mayor proporción de loci (92%) se encontró en equilibrio ($p > 0.05$). No obstante, se detectaron algunas desviaciones que representaron el 8% restante que podría atribuirse a un exceso de homocigotos por posición consecuencia de la subestructura poblacional. El análisis de estructura poblacional con Admixture permitió observar subpoblaciones en la cohorte. Si bien por limitaciones en el acceso a metadatos asociados a las muestras no es posible conocer el origen geográfico de las mismas, el análisis en conjunto con un mayor número de muestras de proyectos como 1000 Genomas será la base para estudios posteriores de genética de poblaciones en la cohorte de pacientes de Biotecgen S.A.S.

Finalmente, las frecuencias alélicas integradas en la herramienta de visualización y análisis de variantes VarSeq que se encuentran actualmente disponibles para los analistas de datos ómicos de Biotecgen S.A.S. proporcionan información sumamente relevante para la interpretación clínica de

variantes y son potencialmente valiosas para futuros estudios de ascendencia genética y estructura poblacional en Colombia.

11. Bibliografía

- Adams, D. (2022). *Frameshift mutation*. National Human Genome Research Institute. Genome.gov. <https://www.genome.gov/genetics-glossary/Frameshift-Mutation>
- Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome research*, 19(9), 1655–1664. <https://doi.org/10.1101/gr.094052.109>
- Arthur, R., Schulz-Trieglaff, O., Cox, A. J., O'Connell, J., (2017). AKT: kit de herramientas de ascendencia y parentesco, *Bioinformática*, volumen 33, número 1, 1 de enero de 2017, páginas 142–144, <https://doi.org/10.1093/bioinformática/btw576>
- Bainbridge, M. N., Wang, M., Wu, Y., Newsham, I., Muzny, D. M., Jefferies, J. L., ... & Gibbs, R. A. (2011). Targeted enrichment beyond the consensus coding DNA sequence exome reveals exons with higher variant densities. *Genome biology*, 12(7), 1-12.
- Ballesteros Villascán, J. (2019). *Tesis. Desarrollo de una herramienta bioinformática para el análisis poblacional de la variación genómica* [Benemérita Universidad Autónoma de Puebla]. <https://doi.org/10.16/CSS/JQUERY.DATATABLES.MIN.CSS>
- Bamshad, M. J., Ng, S. B., Bigham, A. W., Tabor, H. K., Emond, M. J., Nickerson, D. A., & Shendure, J. (2011). Exome sequencing as a tool for Mendelian disease gene discovery. *Nature Reviews Genetics*, 12(11), 745.
- Barbitoff, Y. A., Polev, D. E., Glotov, A. S., Serebryakova, E. A., Shcherbakova, I. V., Kiselev, A. M., ... & Predeus, A. V. (2020). Systematic dissection of biases in whole-exome and whole-genome sequencing reveals major determinants of coding sequence coverage. *Scientific reports*, 10(1), 1-13
- Barbitoff, Y. A., Abasov, R., Tvorogova, V. E., Glotov, A. S., & Predeus, A. V. (2022). Systematic benchmark of state-of-the-art variant calling pipelines identifies major factors affecting accuracy of coding sequence variant discovery. *BMC genomics*, 23(1), 1-17
- Behjati, S., & Tarpey, P. S. (2013). What is next generation sequencing?. *Archives of disease in childhood. Education and practice edition*, 98(6), 236–238. <https://doi.org/10.1136/archdischild-2013-304340>
- Benhabiles, H., Jia, J., & Lejeune, F. (2016). General aspects related to nonsense mutations. En H. Benhabiles, J. Jia, & F. Lejeune (Eds.), *Nonsense Mutation Correction in Human Diseases* (pp. 1–76). Elsevier.
- Biesecker, L. G. (2022). *Insertion*. National Human Genome Research Institute. Genome.gov. <https://www.genome.gov/genetics-glossary/Insertion>
- Brody, L. (2022). *Missense mutation*. National Human Genome Research Institute. <https://www.genome.gov/genetics-glossary/Deletion#>
- Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., Cortes, A., Welsh, S., Young, A., Effingham, M., McVean, G., Leslie, S., Allen, N., Donnelly, P., & Marchini, J. (2018). The UK Biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726), 203–209. <https://doi.org/10.1038/s41586-018-0579-z>
- Chande, A. T., Rishishwar, L., Ban, D., Nagar, S. D., Conley, A. B., Rowell, J., Valderrama-Aguirre, A. E., Medina-Rivas, M. A., & Jordan, I. K. (2020). The phenotypic consequences of genetic divergence between admixed latin american populations: Antioquia and Chocó, Colombia. *Genome Biology and Evolution*, 12(9), 1516–1527. <https://doi.org/10.1093/GBE/EVAA154>
- Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., ... & Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNVs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly*, 6(2), 80-92.

- Conley, A. B., Rishishwar, L., Norris, E. T., Valderrama-Aguirre, A., Mariño-Ramírez, L., Medina-Rivas, M. A., & Jordan, I. K. (2017). A comparative analysis of genetic ancestry and admixture in the Colombian populations of chocó and Medellín. *G3: Genes, Genomes, Genetics*, 7(10), 3435–3447. <https://doi.org/10.1534/g3.117.1118>
- Córdoba, L., García, J. J., Hoyos, L. S., Duque, C., Rojas, W., Carvajal, S., Escobar, L. F., Reyes, I., Cajas, N., Sánchez, A., García, F., Bedoya, G., & Ruiz-Linares, A. (2012). Composición genética de una población del suroccidente de Colombia. *Revista Colombiana de Antropología*, 48(1), 21–48. <https://doi.org/10.22380/2539472X.879>
- Ding, C., & Jin, S. (2009). High-throughput methods for SNP genotyping. *Single Nucleotide Polymorphisms*, 245-254
- Eichler, E. E. (2019). Genetic variation, comparative genomics, and the diagnosis of disease. *New England Journal of Medicine*, 381(1), 64-74
- Ganguly, P. (2022). *Deletion*. National Human Genome Research Institute. <https://www.genome.gov/genetics-glossary/Deletion#>
- Gibson G. (2012) Rare and common variants: twenty arguments. *Nat Rev Genet*. Jan 18;13(2):135-45. doi: 10.1038/nrg3118. PMID: 22251874; PMCID: PMC4408201
- Gnirke, A., Melnikov, A., Maguire, J., Rogov, P., LeProust, E. M., Brockman, W., Fennell, T., Giannoukos, G., Fisher, S., Russ, C., Gabriel, S., Jaffe, D. B., Lander, E. S., & Nusbaum, C. (2009). Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nature biotechnology*, 27(2), 182–189. <https://doi.org/10.1038/nbt.1523>
- Guo, Y., Long, J., He, J., Li, C. I., Cai, Q., Shu, X. O., ... & Li, C. (2012). Exome sequencing generates high quality data in non-target regions. *BMC genomics*, 13(1), 1-10
- Guttman, B. (2013). *Mutation, Nonsense*. Brenner's Encyclopedia of Genetics. Elsevier.
- Hao, W., Storey J.D. (2019). Extending Tests of Hardy–Weinberg Equilibrium to Structured Populations. *Genetics*, 213(3), 759. <https://doi.org/10.1534/genetics.119.302370>
- Henson, J. W. & Resta, R.G. (2021). *Diagnosis and Management of Hereditary Cancer*,. Academic Press. Tabla 2. Tabular-Based Clinical and Genetic Aspects. p. 13. <https://doi.org/10.1016/B978-0-323-90029-4.00012-2>
- Hwang, S., Kim, E., Lee, I., & Marcotte, E. M. (2015). Systematic comparison of variant calling pipelines using gold standard personal exome variants. *Scientific Reports*, 5. <https://doi.org/10.1038/srep17875>
- Illumina. (2010). Technology Spotlight: Illumina ® Sequencing. *Pub. No. 770-2007-002*. https://www.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf
- Illumina. (2020). Preparación de ADN de Illumina sin PCR, tagmentación. *Pub. No. 770-2020-003*. <https://www.illumina.com/content/dam/illumina-marketing/documents/products/datasheets/illumina-dna-pcr-free-data-sheet-770-2020-003-translations/illumina-dna-pcr-free-data-sheet-770-2020-003-esp.pdf>
- Instituto de Ciencias Forenses “Luís Concheiro” (INCIFOR). (2021). *Tipos de polimorfismos*. Universidad de Santiago de Compostela. https://www.usc.gal/gl/institutos/incifor/xeneticaforense_conceptos_tipospolimorfismos.html#Indels
- International HapMap Consortium. A haplotype map of the human genome. *Nature*. 2005 Oct 27;437(7063):1299-320. doi: 10.1038/nature04226. PMID: 16255080; PMCID: PMC1880871.
- Jepsen, M. M., Fowler, D. M., Hartmann-Petersen, R., Stein, A., & Lindorff-Larsen, K. (2020). Chapter 5 - Classifying disease-associated variants using measures of protein activity and stability. En A. L. Pey (Ed.), *Protein Homeostasis Diseases* (p. 91). Academic Press.

- Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., Collins, R. L., Laricchia, K. M., Ganna, A., Birnbaum, D. P., Gauthier, L. D., Brand, H., Solomonson, M., Watts, N. A., Rhodes, D., Singer-Berk, M., England, E. M., Seaby, E. G., Kosmicki, J. A., Walters, R. K., ... MacArthur, D. G. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, 581(7809), 434–443. <https://doi.org/10.1038/s41586-020-2308-7>
- Kobayashi, Y., Yang, S., Nykamp, K., Garcia, J., Lincoln, S. E., & Topper, S. E. (2017). Pathogenic variant burden in the ExAC database: An empirical approach to evaluating population data for clinical variant interpretation. *Genome Medicine*, 9(1), 1–14. <https://doi.org/10.1186/S13073-017-0403-7/FIGURES/3>
- Kohlmeier, M. (2020). Molecular biology of genetic variants. En R. D. E. Caterina, J. A. Martinez, & M. Kohlmeier (Eds.), *Principles of Nutrigenetics and Nutrigenomics* (pp. 11–16). Elsevier.
- Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., O'Donnell-Luria, A. H., Ware, J. S., Hill, A. J., Cummings, B. B., Tukiainen, T., Birnbaum, D. P., Kosmicki, J. A., Duncan, L. E., Estrada, K., Zhao, F., Zou, J., Pierce-Hoffman, E., Berghout, J., ... Williams, A. L. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 2016 536:7616, 536(7616), 285–291. <https://doi.org/10.1038/nature19057>
- Lencz, T., & Malhotra, A. K. (2022). Pharmacogenetics of antipsychotic-induced side effects. <https://doi.org/10.31887/DCNS.2009.11.4/Tlencz>, 11(4), 405–415. <https://doi.org/10.31887/DCNS.2009.11.4/TLENCZ>
- López, C., González, F., Carmona, T., & Oliver, A., (2021). Aplicaciones de las técnicas de secuenciación masiva en la Microbiología Clínica. Antonio Oliver Palomo (coordinador). Procedimientos en Microbiología Clínica. Cercenado, E., & Cantón, R. (editores). Sociedad Española de Enfermedades Infecciosas y Microbiología Clínica (SEIMC).
- Mamanova, L., Coffey, A. J., Scott, C. E., Kozarewa, I., Turner, E. H., Kumar, A., ... & Turner, D. J. (2010). Target-enrichment strategies for next-generation sequencing. *Nature methods*, 7(2), 111–118.
- Metzker, M. L. (2010). Sequencing technologies—the next generation. *Nature reviews genetics*, 11(1), 31–46
- Ministerio de Salud y Protección Social (MinSalud). (2022). *Enfermedades huérfanas*. 28 de noviembre de 2022, de <https://www.minsalud.gov.co/salud/publica/PENT/Paginas/enfermedades-huerfanas.aspx#:~:text=%C2%BFQu%C3%A9%20es%20una%20enfermedad%20ultra,1%2D9%20por%20100%20mil>.
- Montgomery, S. B., Goode, D. L., Kvikstad, E., Albers, C. A., Zhang, Z. D., Mu, X. J., Ananda, G., Howie, B., Karczewski, K. J., Smith, K. S., Anaya, V., Richardson, R., Davis, J., MacArthur, D. G., Sidow, A., Duret, L., Gerstein, M., Makova, K. D., Marchini, J., ... Lunter, G. (2013). The origin, evolution, and functional impact of short insertion-deletion variants identified in 179 human genomes. *Genome Research*, 23(5), 749–761. <https://doi.org/10.1101/GR.148718.112>
- Mordoh, A. (2019). Secuenciación masiva de ADN: la próxima generación | Dermatología Argentina. *Dermatología Argentina*, 25(1). <https://www.dermatolarg.org.ar/index.php/dermatolarg/article/view/1868>
- Nussbaum, R., McInnes, R., & Willard, H. (2016). Thompson & Thompson. Genética en Medicina - 8th Edition. <https://www.elsevier.com/books/thompson-and-thompson-genetica-en-medicina/nussbaum/978-84-458-2642-3>

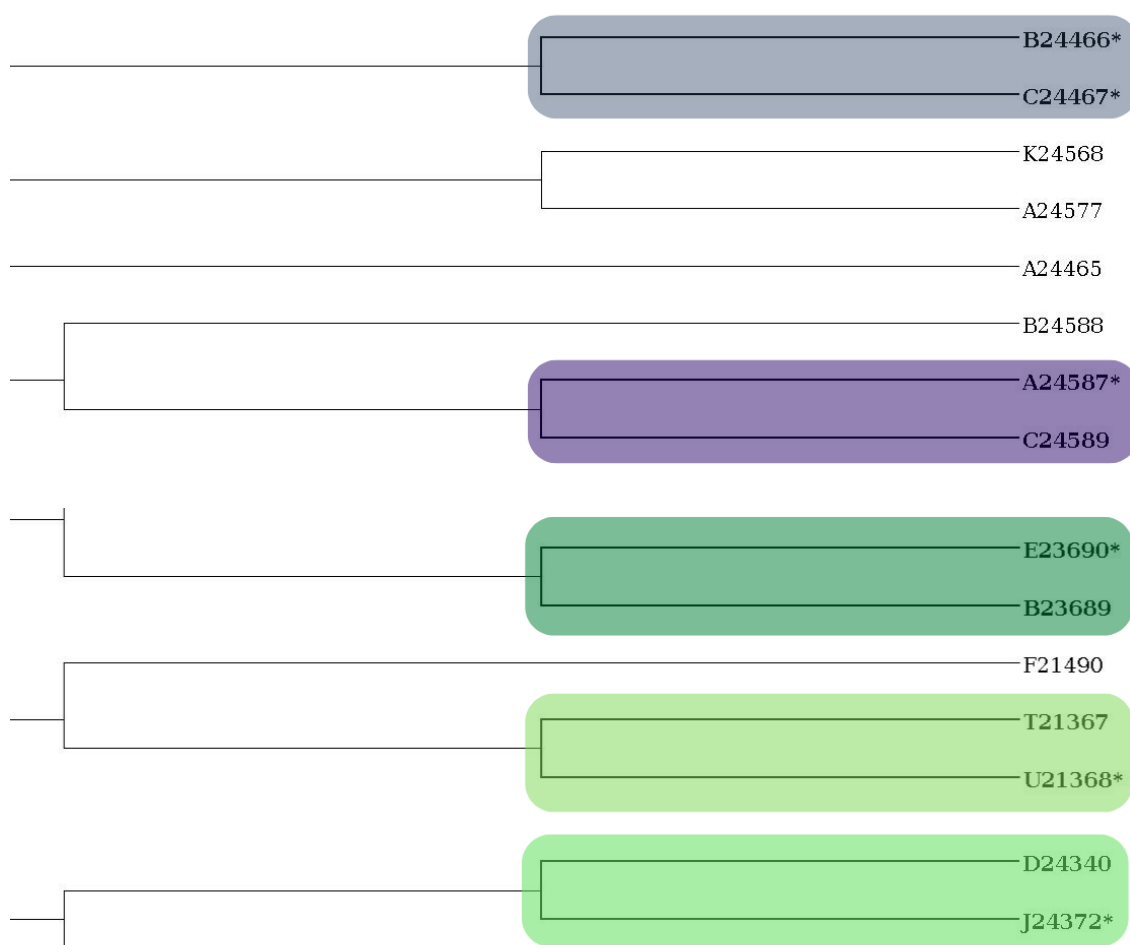
- Odell, E. W. (2021). Aneuploidy and loss of heterozygosity as risk markers for malignant transformation in oral mucosa. *Oral Diseases*, 27(8), 1993-2007.
- Oh, H. R., An, C. H., Yoo, N. J., & Lee, S. H. (2015). Frameshift mutations of MUC15 gene in gastric and its regional heterogeneity in gastric and colorectal cancers. *Pathology & Oncology Research*, 21(3), 713-718
- Ossa, H., Posada, Y., Trujillo, N., Martínez, B., Loiola, S., Simão, F., ... & Gusmão, L. (2021). Patterns of genetic diversity in Colombia for 38 indels used in human identification. *Forensic Science International: Genetics*, 53, 102495
- Pagel, K. A., Pejaver, V., Lin, G. N., Nam, H.-J., Mort, M., Cooper, D. N., Sebat, J., Iakoucheva, L. M., Mooney, S. D., & Radivojac, P. (2017). When loss-of-function is loss of function: assessing mutational signatures and impact of loss-of-function genetic variants. *Bioinformatics*. <https://doi.org/10.1093/bioinformatics/btx272>
- Pierce, B. A. (2015). *Genética: Un enfoque conceptual*. (Médica Panamericana 1 Ed., pp. 715-743).
<http://www.medicapanamericana.com.ezproxy.unbosque.edu.co/VisorEbookV2/Ebook/9788498357332?token=c356031c-161e-4cf8-bda8-60590062c559#{%22Pagina%22:%22VI%22,%22Vista%22:%22Buscador%22,%22Busqueda%22:%22problema%22%22}>.
- Ping, K. (2016). *Next-generation Sequencing and Sequence Data Analysis*. Bentham Science Publishers.
- Rabbani, B., Tekin, M., & Mahdih, N. (2014). The promise of whole-exome sequencing in medical genetics. *Journal of human genetics*, 59(1), 5-15.
- Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W. W., Hegde, M., Lyon, E., Spector, E., Voelkerding, K., & Rehm, H. L. (2015). 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* 17:5, 17(5), 405–423. <https://doi.org/10.1038/gim.2015.30>
- Robinson, P. N., Piro, R. M., & Jager, M. (2017). *Computational exome and genome analysis*. CRC Press.
- Rubio, S., Pacheco-Orozco, R., Milena, A., Perdomo, S., & García-Robles, R., (2020). Secuenciación de nueva generación (NGS) de ADN: presente y futuro en la práctica clínica. *Universitas Médica*, 61(2). <https://doi.org/10.11144/Javeriana.umed61-2.sngs>
- Santamaría, M., & Lezana, J. M. (2018). Aplicaciones clínicas de las técnicas actuales de biología molecular técnicas de secuenciación masiva (NGS). *Cont. Lab. Clin*, 37, 33–40. <https://www.seqc.es/download/tema/25/5627/786418279/826284/cms/tema-5-tecnicas-de-secuenciacion-masiva-ngs.pdf/>
- Savino, S., Desmet, T., & Franceus, J. (2022). Insertions and deletions in protein evolution and engineering. *Biotechnology Advances*, 60(108010), 108010. <https://doi.org/10.1016/j.biotechadv.2022.108010>
- Sharma, Y., Miladi, M., Dukare, S., Boulay, K., Caudron-Herger, M., Groß, M., Backofen, R., & Diederichs, S. (2019). A pan-cancer analysis of synonymous mutations. *Nature Communications*, 10(1), 2569. <https://doi.org/10.1038/s41467-019-10489-2>
- Sehn, J. K. (2015). Chapter 9 - Insertions and Deletions (Indels). En *Clinical Genomics* (pp. 129–130). Academic Press.
- Trudsø, L. C., Andersen, J. D., Jacobsen, S. B., Christiansen, S. L., Congost-Teixidor, C., Kampmann, M. L., & Morling, N. (2020). A comparative study of single nucleotide variant detection performance using three massively parallel sequencing methods. *PLOS ONE*, 15(9), e0239850. <https://doi.org/10.1371/JOURNAL.PONE.0239850>

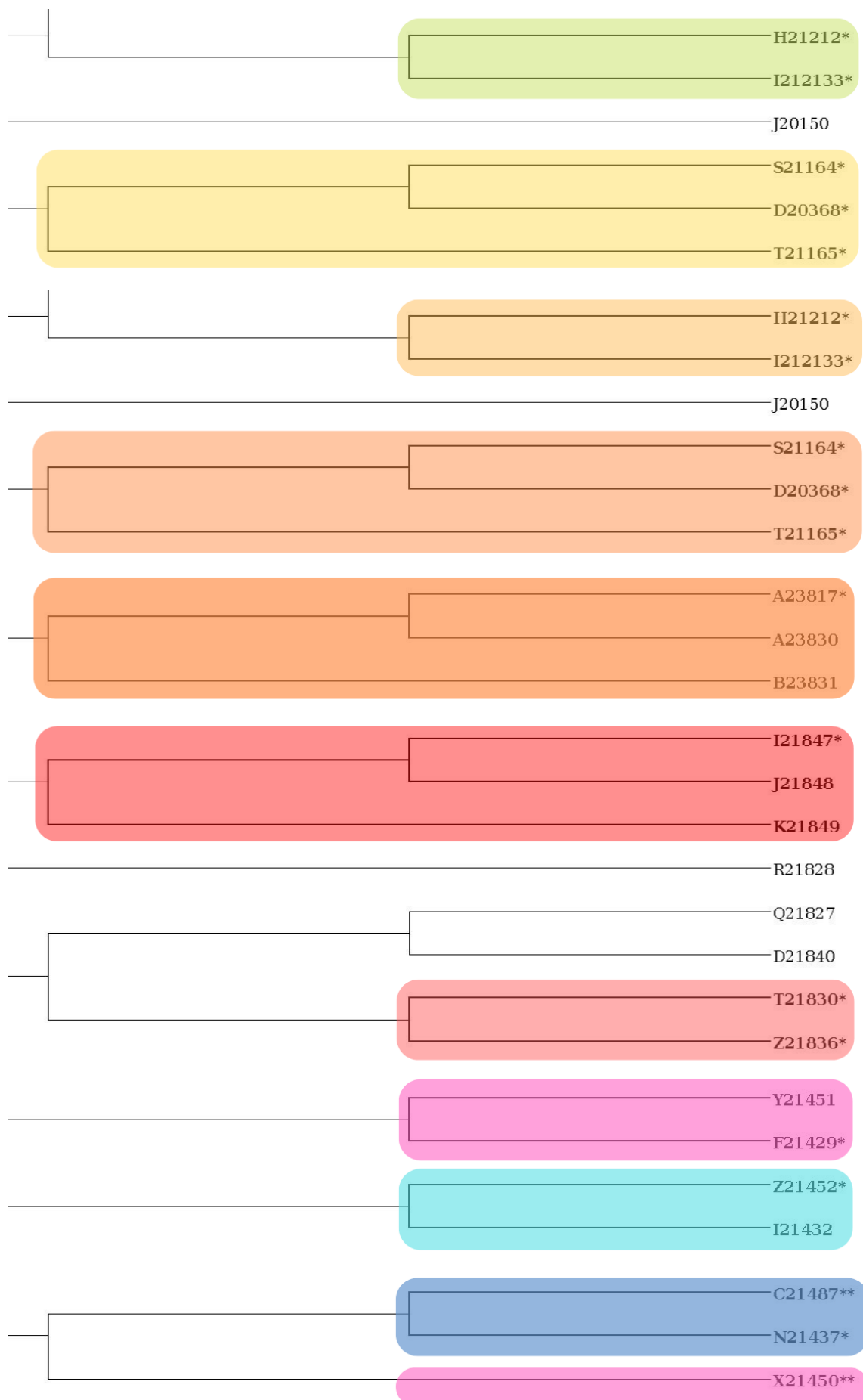
- Weeks, A. L., Francis, R. W., Neri, J. I. C. F., Costa, N. M. C., Arrais, N. M. R., Lassmann, T., Blackwell, J. M., & Jeronimo, S. M. B. (2020). Reference exome data for a Northern Brazilian population. *Scientific Data* 2020 7:1, 7(1), 1–5. <https://doi.org/10.1038/s41597-020-00703-y>
- Xu, Y., Lin, Z., Tang, C., Tang, Y., Cai, Y., Zhong, H., ... & Gao, Q. (2019). A new massively parallel nanoball sequencing platform for whole exome research. *BMC bioinformatics*, 20(1), 1-9
- Wang, J., Raskin, L., Samuels, D. C., Shyr, Y., & Guo, Y. (2015). Genome measures used for quality control are dependent on gene function and ancestry, *Bioinformatics*, 31(3), 318–323, <https://doi.org/10.1093/bioinformatics/btu668>

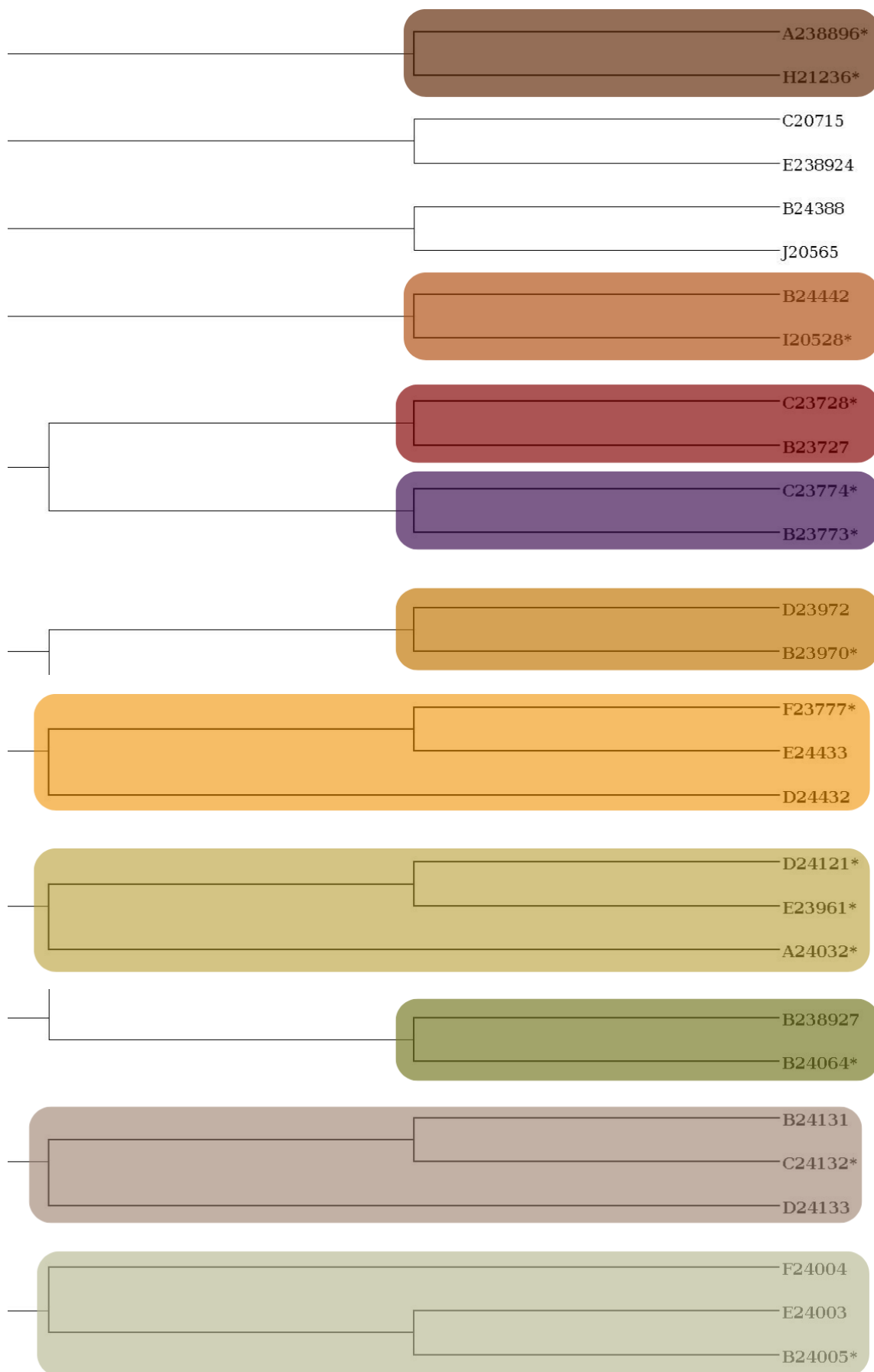
12. Anexos

Anexo 1

*Se resaltan las familias reportadas por el programa akt en el dendograma generado en SplitzTree mediante líneas gruesas. Cada familia se encuentra diferenciada con un color. El * indica que la muestra es hija de las otras muestras del conjunto resaltado. Varios asteriscos dentro de una misma familia representan que las muestras son hermanas. El ** representa la ausencia de una muestra reportada por AKT que no se ve reflejada en el dendograma*

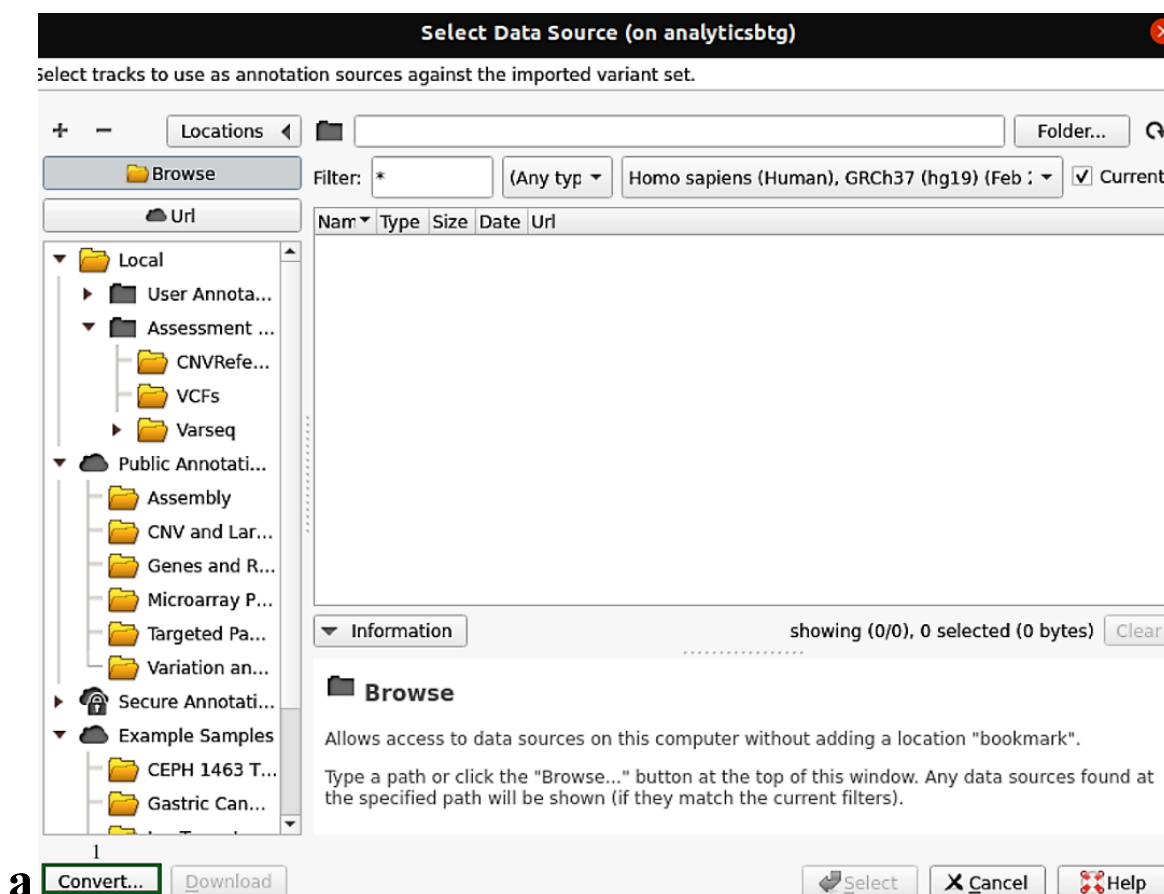






Anexo 2

Pasos para la integración a Varseq. a-f. Se evidencia la secuencia de pasos requeridos para la integración de la matriz de frecuencia alélica de variantes SNV e Indel. g. Se observa la información disponible en la herramienta de visualización para los analistas de datos ómicos de Biotecgen S.A.S.



Convert Data Source

1

① Define Input
② Scan Input
③ Change Options
④ Convert

Select one or more files to Convert.
Files must be of the same type to be converted together.

Select Files:

1

Add
Remove

Advanced Options

Help

2

< Back Next > Cancel

b**Convert Source Wizard (on analyticsbgtg)****Convert Data Source**

① Define Input
② Scan Input
③ Change Options
④ Convert

1

You can rename, drop, reorder and change the type of fields.
Note that when data fails to be coerced into a specified type, it will be set as missing.

Desired Plot Type: (Automatically Detect)
Detected Plot Type: Interval
Edit Output Fields:

Use	Input	Name	Type
<input type="checkbox"/>	Ref/Alt	Ref/Alt	String
<input checked="" type="checkbox"/>	Identifier	Identifier	String Array
<input checked="" type="checkbox"/>	Reference	Reference	String
<input checked="" type="checkbox"/>	Alternates	Alternates	String Array
<input type="checkbox"/>	Quality	Quality	Float
<input type="checkbox"/>	Filter	Filter	Categorical Array
<input checked="" type="checkbox"/>	Alt Allele Counts (AC)	Alt Allele Counts (AC)	Int Array
<input checked="" type="checkbox"/>	Alt Allele Freq (AF)	Alt Allele Freq (AF)	Float Array
<input type="checkbox"/>	#Alleles (AN)	#Alleles (AN)	Int
<input type="checkbox"/>	BaseQRankSum	BaseQRankSum	Float
<input type="checkbox"/>	CNN_1D	CNN_1D	Float

Preview: 1000 features read into preview ([Read More](#))

	Chr	Start	Stop	Identifier	Reference	Alternates	Alt Allele Counts (AC)	Alt Allele Freq (AF)
1	1	12783	12783	?	G	A	53	0.04193
2	1	12807	12807	?	C	T	8	0.00632
3	1	12839	12839	?	G	C	4	0.00316
4	1	12882	12882	?	C	G	5	0.00395
5	1	12948	12948	?	T	C	1	0.00079
6	1	12993	12993	?	G	A	1	0.00079
7	1	13012	13012	?	G	C,A	2,1	0...
8	1	13079	13079	?	C	G	16	0.01265
9	1	13091	13091	?	G	A	1	0.00079
10	1	13110	13110	?	G	A	47	0.03718
11	1	13116	13116	rs201725126...	T	G	520	0.41139

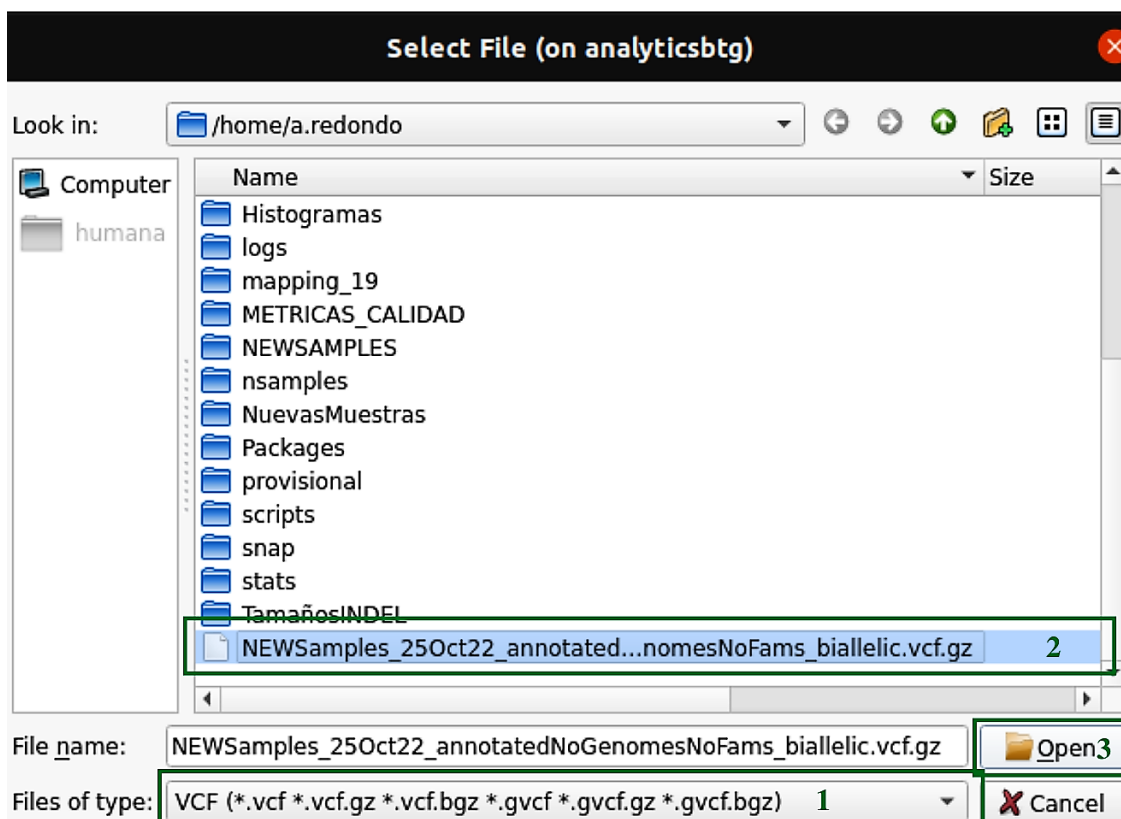
Advanced Options

Help

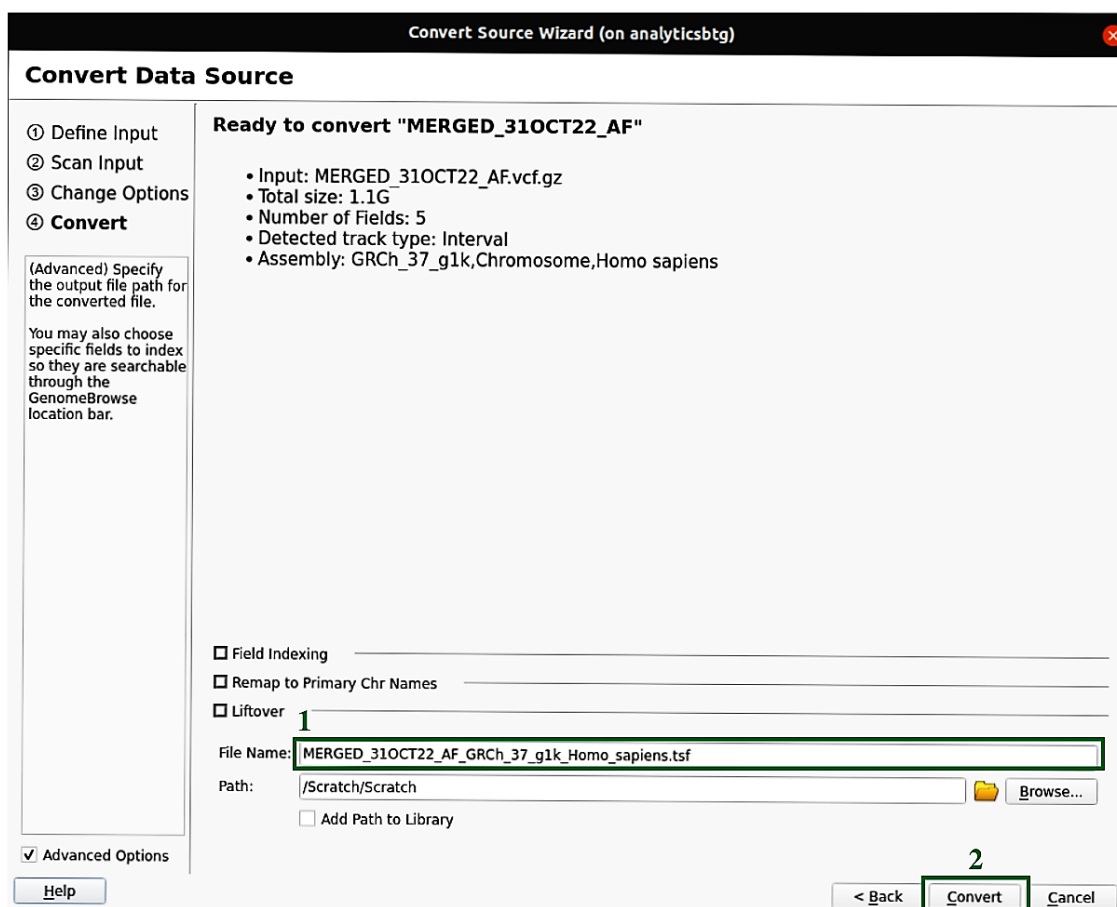
2

< Back Next > Cancel

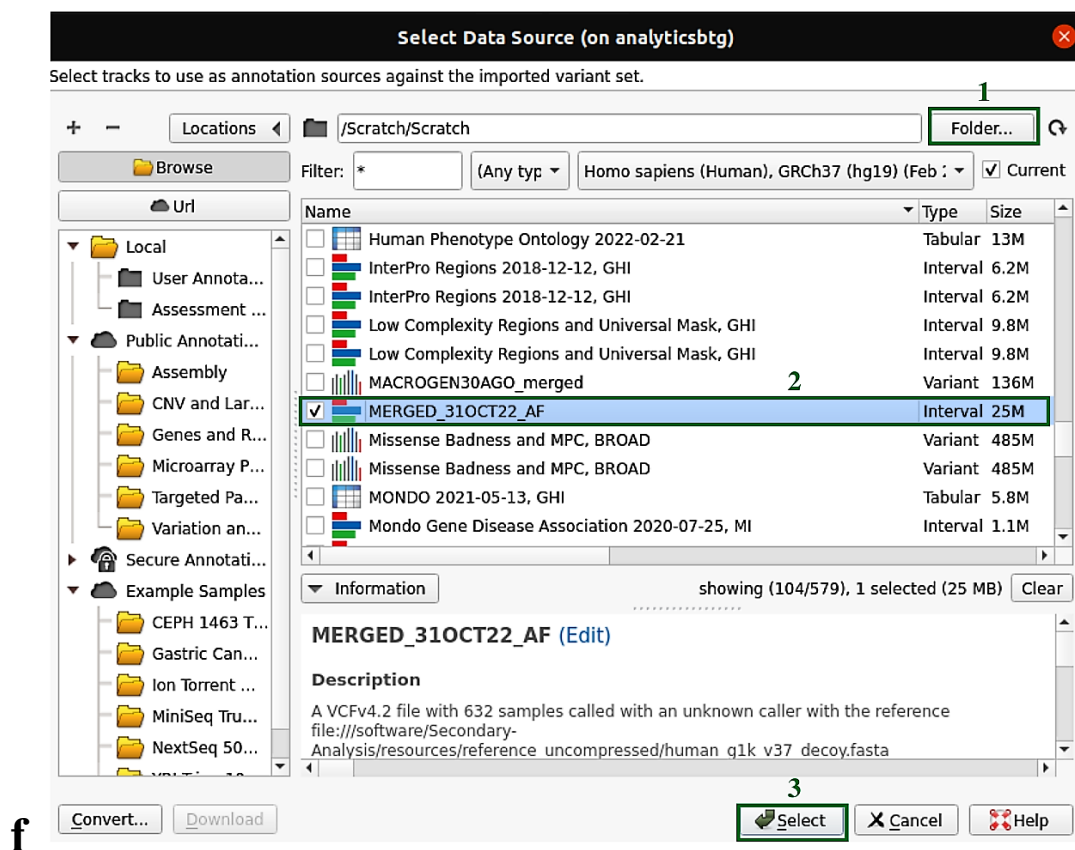
c



d



e



*80330_Exoma - Golden Helix VarSeq 2.2.5 (on analyticsbtg)

File View Tools Help

ACMG Guidelines X ACMG Guidelines X

Filter Variants X Filter CNVs X

Filter Variants: 29 X Samples: 1 X Log X CNVs: 1,959 X Coverage Regions: X

Filter Variants: 25876-80330 * a

Flags (Current) is missing 522

NOTip-value (Current) > 0.001 257

NOTIClassification is (Benign, Likely Benign) 201

NOTType is Low Complexity 169

NOTType is Low Complexity (Data source: Overlapping Regions Low Complexity Regions and Universal Mask, GHI) 40

Fenotipo is 17

ACMG Guidelines: Evaluation Genes Variants CNVs Phenotypes Report

Sample 25876-80330

CNVs to Evaluate in Sample 25876-80330

Sort By: Adjust Genomic Classification Status

Filter Out: Previously Classified Saved

Genes	Type / Effect	Score / Draft Classification	Record Sets	Annotations
SLC35A1 ex18 dup	Splice Acceptor	Gain		0.001 1.000
WDFY3	Duplicate (125bp)	Uncertain Significance (0.0)		0.001 1.000
WDFY3 ex11 dup	Protein Truncation	Gain		0.001 1.000
NIPBL	Duplicate (156bp)	Uncertain Significance (0.45)		0.001 1.000
NIPBL ex18 dup	Protein Truncation	Gain: 21:0.45		0.001 1.000
CSPPI	Duplicate (2.3kb)	Uncertain Significance (0.0)		0.001 1.000
CSPPI ex10-11 dup	Splice Acceptor	Gain		0.001 1.000
INTS8	Duplicate (179bp)	Uncertain Significance (0.0)		0.001 1.000
INTS8 ex2 dup	Protein Truncation	Gain		0.001 1.000
INTS8	Duplicate (112bp)	Uncertain Significance (0.0)		0.001 1.000
INTS8 ex7 dup	Splice Acceptor	Gain		0.001 1.000

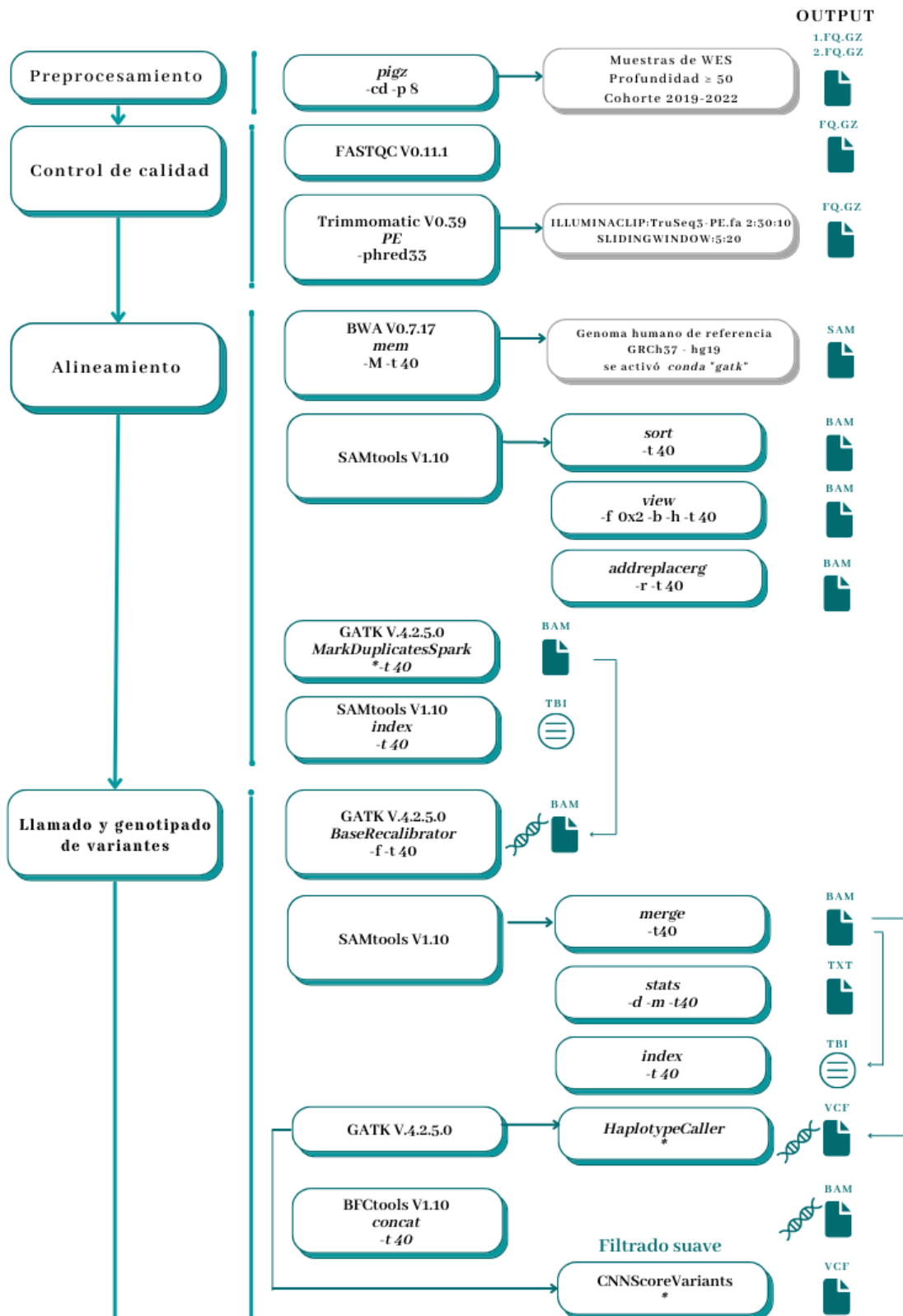
Variant Info	Flags for 25876-80330	Zygosity	Read Depths	Variant Allele Freq	GT	GO	Filter	Identifier	Reference	MERGED_31OCT22_AF	ACMG S...
1:114372222	T/C	Heterozygous	44	0.477273	0/1	219	PASS	rs2003035...	G	?	T/C
1:161467775	G/A	Heterozygous	73	0.326767	0/1	154	PASS		G	?	G/A
2:197061790	C/T	Heterozygous	76	0.555632	0/1	232	PASS		T	?	C/T
2:234229274	A/G	Heterozygous	188	0.457447	0/1	551	PASS		T	?	A/G
4:79400795	A/G	Heterozygous	153	0.54902	0/1	455	PASS		T	?	A/G
4:79458330	C/A	Heterozygous	113	0.477876	0/1	390	PASS		T	?	C/A
5:61467088	G/T	Heterozygous	61	0.406557	0/1	255	PASS		G	?	G/T
5:74807107	G/A	Heterozygous	154	0.461039	0/1	499	PASS	rs1433179...	A	?	G/A
6:32035554	C/T	Heterozygous	106	0.424528	0/1	311	PASS		T	?	C/T
6:32063665	G/C	Heterozygous	302	0.413907	0/1	671	PASS		T	?	G/C
6:52874263	A/C	Heterozygous	75	0.56	0/1	268	PASS		T	?	A/C
7:4823836	-G	Heterozygous	7	0.291339	0/1	280	PASS		T	?	-G

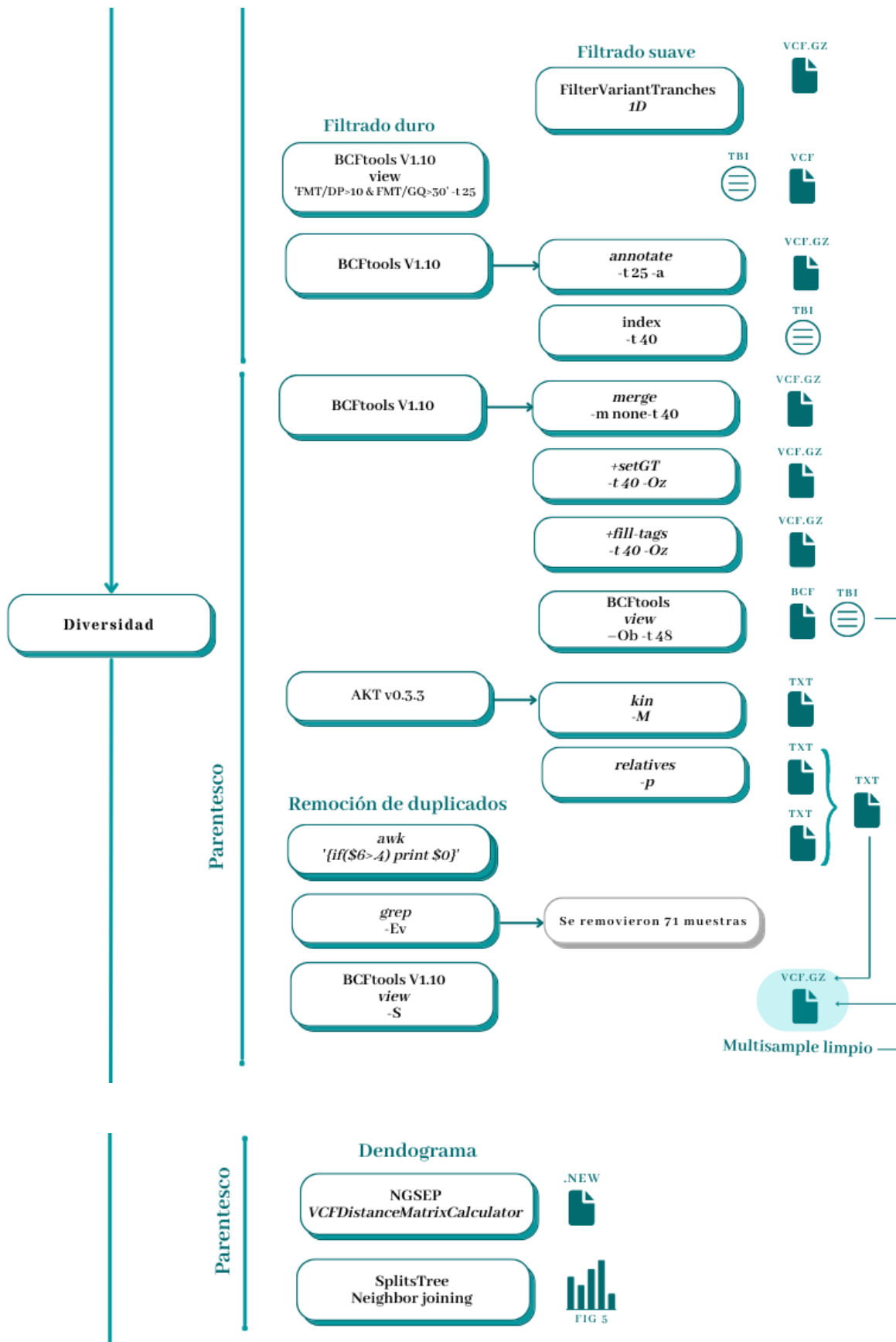
NOTType is Low Complexity: Overlapping Regions Low Complexity Regions and Universal Mask, GHI: Type

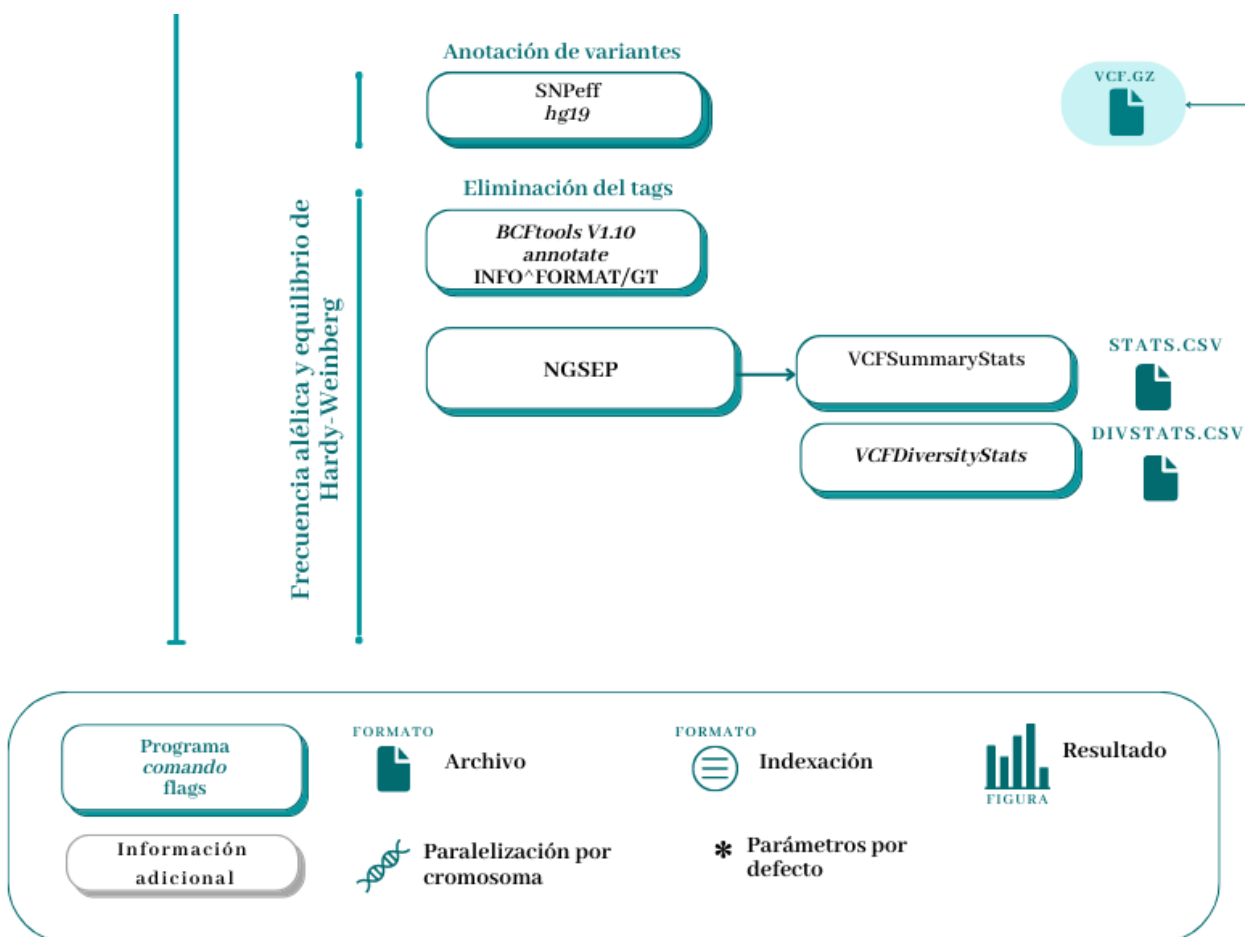
Navigation: (X: 105,286,479, 0) 1 20 bp

Anexo 3

Flujo de trabajo a detalle








Anexo 4

Formato de consentimiento informado

	BIOTECNOLOGIA Y GENETICA S.A.	Código: LH-FT-08
		Fecha de Aprobación: 01/09/2021
	FORMATO CONSENTIMIENTO INFORMADO PARA REALIZACIÓN DE ESTUDIOS MOLECULARES	Página: 2 de 3
		Versión: 05

Declaración de consentimiento del paciente o su acudiente.

Al firmar abajo, reconozco que:

- He leído este documento en su totalidad y entiendo sus implicaciones.
- Se me ha brindado la oportunidad de formular preguntas que han sido resueltas por el personal técnico de la institución.
- Accedo a someterme a las pruebas. Una vez recibidos los resultados haré entrega de éstos al médico que solicitó el examen. Discutiré los resultados así como el manejo adecuado y/o tratamiento médico con mi proveedor de atención médica.

Por favor marque su decisión con respecto a las siguientes anotaciones:

- I. En caso de no poder asistir por el resultado de la prueba realizada, acepto recibirlo al correo electrónico.

Si ☒ No ☐ No Aplica ☐

Indique su correo electrónico: [REDACTED]

- II. Es importante aclarar que en las pruebas genómicas ampliadas y las pruebas de secuenciación exómica podrían reportar variantes de riesgo para condiciones genéticas diferentes a las de su motivo de consulta; podrían reportar por ejemplo riesgo aumentado de sufrir enfermedades a edades avanzadas. Estas variantes patogénicas asociadas a condiciones diferentes serán solo reportadas si usted acepta. Teniendo en cuenta esto, autorizo que en mi resultado se publiquen variantes de riesgo o susceptibilidad a condiciones genéticas diferentes a las de mi motivo de consulta:

Si ☒ No ☐ Aplica ☐

- III. Autorizo la utilización anónima de la muestra y/o los resultados de la prueba en estudios de investigación posteriores y que pueden ayudar en el futuro a entender las causas de la entidad genética estudiada, sin recibir a cambio ningún tipo de beneficio económico.

Si ☒ No ☐ No Aplica ☐

NOMBRE DEL PACIENTE: [REDACTED] Doc. Identidad: [REDACTED]

NOMBRE DEL ACUDIENTE: _____ Doc. Identidad: _____

TELÉFONO DE CONTACTO: [REDACTED]

FIRMA PACIENTE O ACUDIENTE: [REDACTED]

Se firma este consentimiento en constancia de asistencia.