UNIVERSIDAD DE PANAMÁ

FACULTAD DE CIENCIAS NATURALES, EXACTAS Y TECNOLOGÍA

ESCUELA DE BIOLOGÍA

DEPARTAMENTO DE GENÉTICA Y BIOLOGÍA MOLECULAR

PERFIL GENÓMICO DE LA POBLACIÓN PANAMEÑA Y SU RELACIÓN CON ENFERMEDADES CRÓNICAS

Por

BEATRICE DI BIASE

Trabajo de Graduación presentado a la Escuela de Biología para optar por el título de Licenciada en Biología con orientación en Genética y Biología Molecular.

Ciudad de Panamá, República de Panamá

Mayo 26, 2023

UNIVERSIDAD DE PANAMÁ

FACULTAD DE CIENCIAS NATURALES, EXACTAS Y TECNOLOGÍA

ESCUELA DE BIOLOGÍA

Ciudad Universitaria 26 de mayo de 2023

Por este medio se hace constar que el proyecto de trabajo de graduación cuyo título es:

PERFIL GENÓMICO DE LA POBLACIÓN PANAMEÑA Y SU RELACIÓN CON ENFERMEDADES CRÓNICAS

Ha sido recomendado por el Departamento de Genética y Biología Molecular

El trabajo fue realizado por la estudiante:

Beatrice Di Biase CIP EC-51-14049

Edgardo Castro-Pérez, Ph.D.	Magaly de Chial, Ph.D.	Carlos Ramos, Ph.D.
Asesor Principal	Co-Asesora	Co-Asesor

Olga Chen, M.Sc.

Directora del Departamento de Genética y Biología Molecular

Aprobado por: _____ Fecha: _____

Profesor Ovidio Durán, Ph.D.

Director, Escuela de Biología

Se adjunta tesis

UNIVERSIDAD DE PANAMÁ

FACULTAD DE CIENCIAS NATURALES, EXACTAS Y TECNOLOGÍA

ESCUELA DE BIOLOGÍA

Programa: Biología

Subprograma: Genética y Biología Molecular

Línea de Investigación: Genética y Biología Molecular

Título: Perfil Genómico de la Población Panameña y su relación con Enfermedades Crónicas

Autor: Beatrice Di Biase _____

Firma del estudiante

Profesor Asesor: Edgardo Castro-Pérez, Ph.D. _____

Profesor Asesor

Duración: 1 año

Localidad: Ciudad de Panamá, Panamá

Fecha: 26 de mayo de 2023

UNIVERSIDAD DE PANAMÁ
FACULTAD DE CIENCIAS NATURALES, EXACTAS Y TECNOLOGÍA
ESCUELA DE BIOLOGÍA
DEPARTAMENTO DE GENÉTICA Y BIOLOGÍA MOLECULAR

PERFIL GENÓMICO DE LA POBLACIÓN PANAMEÑA Y SU RELACIÓN CON ENFERMEDADES CRÓNICAS

Asesor

Edgardo Castro-Pérez, Ph.D.

Estudiante

Beatrice Di Biase

Co-asesores

Carlos Ramos, Ph.D.

Magaly de Chial, Ph.D.

Mayo 26, 2023

AGRADECIMIENTOS

Al respaldo de la Vicerrectoría de Investigación y Posgrado (VIP) en el proyecto y a la Escuela de Biología, en especial, al Departamento de Genética y Biología Molecular por las facilidades brindadas.

A mi asesor, el Dr. Edgardo Castro Pérez, de manera especial y sincera por haberme confiado este trabajo y por su disposición durante el desarrollo de esta investigación. Su valiosa guía y apoyo fueron fundamentales para el logro de los resultados, así como para la ampliación de mis conocimientos y el potenciamiento de mis habilidades.

A los miembros del Comité Asesor, Dr. Carlos Ramos y Dra. Magaly de Chial por el tiempo dedicado a la revisión de la redacción final de esta tesis, sus comentarios y sugerencias fueron fundamentales.

A mis compañeras de laboratorio, Kathia Canto y Madián Poveda, por el productivo intercambio en los ámbitos comunes de nuestras respectivas investigaciones.

A mi familia, en particular, a mis padres y a mi hermano por su profundo amor y apoyo incondicional siempre.

ÍNDICE GENERAL

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

RESUMEN

Las enfermedades no transmisibles (ENT) representan las principales causas de muerte por enfermedades en Panamá. Evidencias indican que existen disparidades raciales y ancestrales en el riesgo de distintas ENT en la población del país, es decir, la prevalencia varía de acuerdo con el trasfondo ancestral dominante africano, europeo o amerindios. Esto sugiere que el componente genético de la población es un factor de riesgo altamente determinante en la distribución, susceptibilidad y muertes por ENT en el país. Sin embargo, no existen estudios biomédicos sobre la población panameña que abordan científicamente la genética de estas enfermedades principalmente porque se desconocen las variantes y los polimorfismos genéticos asociados con las mismas. Identificar estos polimorfismos es fundamental para comprender el riesgo y susceptibilidad, así como para mejorar el diagnóstico y desarrollo de mejores tratamientos. Para la identificación de estas variantes génicas en otras poblaciones del mundo se ha reportado el uso de estrategias genómicas. Desafortunadamente, no se han realizado estudios genómicos en la población panameña.

Estudios previos han demostrado que el componente ancestral de la población panameña es dominado por genes de origen Ngäbe con un 51%. Hipotetizamos que una significativa parte de los polimorfismos genéticos asociados a enfermedades de la población panameña puede tener un origen en esta población ancestral amerindia. En esta investigación analizamos el genoma de cuatro individuos Ngöbe con el objetivo de identificar variantes genéticas posiblemente asociadas a ENT en la población panameña. Aunque secuenciamos el genoma completo, en los análisis nos enfocamos en el exoma (parte codificante del genoma).

Realizamos secuenciación de nueva generación (NGS) y análisis bioinformáticos lo cual nos permitió identificar algunos polimorfismos candidatos. Los resultados demostraron que la mayor parte de las variantes posiblemente asociadas a ENT eran SNPs. Entre ellas se encontraron 15 variantes candidatas de las cuales nos enfocamos en el polimorfismo rs1801133, o 677C>T, del gen MTHFR. Este causa hiperhomocisteinemia y tiene asociación a cánceres y a enfermedades vasculares según distintos estudios. Ejecutamos el genotipado de esta variante en la población Ngäbe mediante PCR y secuenciación Sanger. Nuestros resultados demuestran que, en la población Ngäbe, esta variante tiene una frecuencia alélica de 15% para el alelo de tipo silvestre y de 85% para el alelo mutante, así como una frecuencia genotípica de 0% para el genotipo homocigoto silvestre, 31% para el genotipo heterocigotos y 69% para el genotipo homocigoto mutante.

ABSTRACT

Non-communicable diseases (NCDs) represent the leading causes of death by disease in Panama. Evidence indicates that there are racial and ancestral disparities in the risk of different NCDs in the country's population, i.e., prevalence varies according to the dominant African, European or Amerindian ancestral background. This suggests that the genetic component of the population is a highly determinant risk factor in the distribution, susceptibility and deaths from NCDs in the country. However, there are no biomedical studies on the Panamanian population that scientifically address the genetics of these diseases mainly because the genetic variants and polymorphisms associated with them are unknown. Identifying these polymorphisms is fundamental to understand risk and susceptibility, as well as to improve diagnosis and develop better treatments. For the identification of these gene variants in other populations around the world, the use of genomic strategies has been reported. Unfortunately, genomic studies have not been performed in the Panamanian population.

Previous studies have shown that the ancestral component of the Panamanian population is dominated by genes of Ngäbe origin with 51%. We hypothesize that a significant part of the genetic polymorphisms associated with diseases in the Panamanian population may have an origin in this ancestral Amerindian population. In this research we analyzed the genome of four Ngöbe individuals with the aim of identifying genetic variants possibly associated with NCDs in the Panamanian population. Although we sequenced the entire genome, we focused our analyses on the exome (coding part of the genome).

We performed next generation sequencing (NGS) and bioinformatics analysis which allowed us to identify some candidate polymorphisms. The results showed that most of the variants possibly associated with ENT were SNPs. Among them we found 15 candidate variants of which we focused on the rs1801133, or 677C>T, polymorphism of the MTHFR gene. This causes hyperhomocysteinemia and is associated with cancers and vascular diseases according to different studies. We performed genotyping of this variant in the Ngäbe population by PCR and Sanger sequencing. Our results show that, in the Ngäbe population, this variant has an allelic frequency of 15% for the wild-type allele and 85% for the mutant allele, as well as a genotypic frequency of 0% for the homozygous wild-type genotype, 31% for the heterozygous genotype and 69% for the homozygous mutant genotype.

INTRODUCCIÓN

El presente trabajo hace parte del Proyecto de Investigación "Identificación de Factores de Riesgo Genético a Enfermedades en el Genoma de la Población Panameña Mediante Análisis de Exoma" adjudicado al Dr. Edgardo Castro-Pérez y el Dr. Carlos Ramos y financiado por la Vicerrectoría de Investigación y Posgrado (VIP) de la Universidad de Panamá.

Las enfermedades crónicas, llamadas también enfermedades no transmisibles (ENT), son afecciones de larga duración y de lenta progresión (Kelly et al., 2016; *World Health Organization*, 2016). Las ENT poseen una etiología múltiple, desarrollo poco predecible, múltiples factores de riesgo y, con pocas excepciones, un origen no infeccioso (*World Health Organization*, 2005).

La OMS ha estimado que las enfermedades crónicas causan la muerte de 41 millones de personas cada año, un equivalente al 74% de todas las muertes a nivel global. Cuando no conducen al deceso del enfermo, muy frecuentemente causan deterioro de la calidad de vida, resultando particularmente inhabilitantes (World Health Organization, 2016). Consistente con estas tendencias mundiales, las cifras del Ministerio de Salud de Panamá (MINSA) muestran que la principal causa de muerte por enfermedades en el país es el cáncer, seguido por las enfermedades cardio y cerebrovasculares entre otras (Ministerio de Salud, 2018; G. Reyes, 2022). A estas cifras siguen la diabetes y la obesidad o enfermedades asociadas al metabolismo y, por último, otros tipos de enfermedades cardíacas. Aunque la diabetes, la obesidad y las enfermedades asociadas al metabolismo no estaban entre las primeras causas, es bien conocido que la diabetes y obesidad están vinculadas a una mayor amenaza de algunos tipos de cánceres y también constituye un factor de riesgo para otras enfermedades. El Estado ha hecho énfasis sobre este bloque de enfermedades que afectan al panameño para que se brinde mayor prevención y cuidado, debiendo así orientar los recursos del país para mejorar la prevención, el diagnóstico y el tratamiento.

Desafortunadamente, la disponibilidad de datos sobre estadísticas vitales con respecto a las enfermedades que afectan al panameño por provincia es escasa (Castro-Pérez, 2022). Sin embargo, los datos de algunas enfermedades son un poco más accesibles, por ejemplo: las cifras relativas a algunos cánceres y enfermedades cardio y cerebrovasculares. En particular, los

números indican que Panamá y Colón son las provincias que tienen por muy lejos la mayor incidencia de cáncer de próstata. Múltiples reportes indican que hay considerables disparidades raciales en el riesgo de cáncer de próstata, cuya incidencia es mucho mayor entre los afrodescendientes (Bock et al., 2009; Robbins et al., 2007; Zeigler-Johnson et al., 2008). En consecuencia con esta noción, los datos genéticos señalan que las provincias de Panamá y Colón son las provincias que muestran la mayor proporción de genes de origen africano (Arias et al., 2002; Castro-Pérez et al., 2016; Ramos et al., 2018). Con relación a los trastornos cerebrovasculares, la mayor incidencia de muertes causadas por estas enfermedades se da en las provincias de Herrera y Los Santos. Los reportes previos muestran que hay considerables disparidades raciales en el riesgo de enfermedades cardio y cerebrovasculares, cuyas incidencias son mucho mayores entre la población de ascendencia europea, seguida por los afrodescendientes (Cheng et al., 2010; Donnan et al., 2008; Hankey, 1999). Atendiendo estos datos, estudios genéticos apuntan que Herrera y Los Santos son provincias que comparten la mayor proporción de genes europeos y una cantidad moderadamente alta de genes africanos (Arias et al., 2002; Castro-Pérez et al., 2016; Ramos et al., 2018). En cuanto a los trastornos relacionados al síndrome metabólico (obesidad, diabetes, entre otros), las provincias de Bocas del Toro, Colón, Chiriquí y Panamá advierten una mayor prevalencia de obesidad (Sasson et al., 2014), las cuales son provincias dominadas por genes de origen amerindio (Arias et al., 2002; Castro-Pérez et al., 2016; Ramos et al., 2018). De forma interesante, varios reportes y estadísticas vitales del Centro de Control y Prevención de Enfermedades de Estados Unidos (CDC) indican que la diabetes, la obesidad y el síndrome metabólico presentan un mayor factor de riesgo e incidencia en poblaciones de origen amerindio (Mariscal Davy, 2021). Además, varios estudios indican que obesidad, diabetes y el síndrome metabólico son condiciones que se asocian con mayor riesgo a algunos cánceres, así como la hipertensión y otras condiciones crónicas (Mariscal Davy, 2021). Aunque estos reportes presentan de manera general los problemas de epidemiología genética de enfermedades no transmisibles del país, desafortunadamente las causas genéticas y moleculares para la disparidad relativa a estas enfermedades no han sido estudiadas en la población panameña.

Algunos estudios previos (Arias et al., 2002; Castro-Pérez et al., 2016; Ramos et al., 2018) han tratado de abordar el problema de una manera indirecta usando marcadores generales como STR (microsatélites: repeticiones cortas en tándem) e INDELs (marcadores de inserción/deleción) e

indican que, en general, la genética del panameño es muy heterogénea y que ha recibido aportes de tres poblaciones ancestrales: africanos, europeos y amerindios. Estos estudios también indican que hay un polimorfismo alto entre los panameños y que la distribución y contribución genética de las poblaciones ancestrales es diferente entre las provincias, pongamos por caso: Chiriquí, Coclé y Veraguas, que son dominadas por genes indígenas; Los Santos y Herrera, por genes europeos; mientras que Panamá y Colón tienen el mayor porcentaje de genes de origen africano y una proporción relativamente alta de genes amerindios seguidos de genes europeos. Los estudios con marcadores generales (STRs, INDELs) sugieren fuertemente que el componente genético ancestral del panameño tiene un papel significativo en la etiología y distribución de las enfermedades, pero no se han realizado estudios genéticos ni genómicos que aborden la epidemiología genética que identifique las variantes y polimorfismos genéticos específicos asociados a factores de riesgo en ENT en la población panameña. Identificar estas variantes genéticas es esencial para entender la etiología de enfermedades, los factores de riesgo y desarrollar nuevos tratamientos y políticas de prevención acordes a nuestra genética autóctona y resolver los problemas biomédicos del país.

Para abordar estos problemas, en esta investigación nos hemos enfocado en identificar por primera vez las variantes y los polimorfismos genéticos de importancia biomédica asociados a las enfermedades en la población panameña utilizando técnicas genómicas de nueva generación mediante secuenciación del genoma completo y del análisis del exoma en cuatro individuos panameños. Aunque hemos secuenciado el genoma completo, para los análisis e identificación de los polimorfismos asociados a enfermedades, nos hemos enfocado en el exoma, ya que las evidencias indican que aproximadamente el 80-90% de las enfermedades conocidas son asociadas a exones. El estudio del exoma consiste en analizar en detalle los exones, es decir, las regiones del ADN que se transcriben para generar RNAs mensajero y proteínas, que conforman el exoma, mientras excluye los intrones y las regiones intergénicas no codificantes. El exoma humano comprende alrededor de 180,000 exones, que conforman casi el 1.5% del total del genoma, es decir, unas 30 megabases de ADN (Sant Joan de Déu Barcelona Hospital, 2014).

Identificar factores y polimorfismos genéticos asociados al riesgo de enfermedades mediante análisis del exoma nos permitiría mejorar el diagnóstico genético, que consiste en detectar el gen y las mutaciones asociados a la enfermedad. Además, podríamos reconocer interacciones

entre polimorfismos de diferentes genes que causan de manera combinada un efecto sumatorio en riesgo o severidad en enfermedades de la población panameña, concretamente, la interacción y la coexistencia de variantes heteroalélicas que podrían ser muy raras o inexistentes en las poblaciones ancestrales y que generan un riesgo combinado a enfermedades en Panamá.

Esta estrategia genómica tiene muchas ventajas sobre estudios previos y otras estrategias tradicionales. en particular, si bien los estudios previos demuestran que es posible identificar genes y variantes genéticas de interés biomédico en la población realizando estudios puntuales, los cuales se enfocan en un marcador o región en particular; los estudios tienden a ser muy laboriosos y toman muchos años abordar unos pocos genes a la vez. Además, con cada estudio realizado se utiliza mucho material genético y con los años se requiere volver a recolectar muestras de ADN de estos o de nuevos voluntarios. En contraste, las técnicas genómicas de nueva generación permiten con una pequeña cantidad de ADN analizar en modo definitivo miles de genes del genoma completo en un gran número de individuos de la población a la vez en un tiempo muy corto.

El trabajo realizado representa un estudio piloto. Nos hemos enfocado en analizar el genoma de amerindios Ngöbe, el cual representa la población ancestral con mayor contribución a la genética del panameño. Estudios previos han demostrado que el trasfondo genético ancestral de la población panameña tiene un origen Ngöbe. Este representa aproximadamente el 51% del acervo genético de la población del país (Castro-Pérez et al., 2016; Ramos et. al., 2018). Dada la alta contribución ancestral de la población amerindia, se espera que gran parte de los genes asociados a enfermedades no transmisibles y problemas biomédicos pueden tener un origen en esta población ancestral (Arias, 1991), por lo tanto, esta investigación ha hipotetizado que, secuenciar el genoma de algunos individuos de esta población ancestral Ngöbe, nos conduciría a identificar variantes y polimorfismos genéticos posiblemente asociados a ENT en el panameño. Los resultados presentados han permitido identificar por primera vez variantes génicas candidatas asociadas a diferentes ENT. El trabajo se centra en la variante/polimorfismo del gen MTHFR, que se asocia a mayor riesgo a algunos tipos de cánceres. Los resultados presentados aquí representan los primeros estudios genómicos en la población panameña.

CAPÍTULO I: JUSTIFICACIÓN

Las estadísticas del MINSA evidencian que las principales causas de muerte de la población panameña están asociadas con ENT como cánceres, enfermedades cardiovasculares y cerebrovasculares, entre otras. Estas enfermedades, a su vez, se relacionan con otros trastornos que aumentan el riesgo de padecerlas como la hipertensión, la diabetes, la obesidad, el síndrome metabólico y estilos de vida no saludables. Estudios previos (Arias et al., 2002; Sasson et al., 2014; Castro-Pérez et al., 2016; Ramos et al., 2018; Mariscal Davy, 2021) sugieren que los patrones de incidencia y muertes por estas enfermedades entre las provincias del país no sólo son diferentes, sino también altamente consistentes con las enfermedades y mortalidad conexionadas a las poblaciones ancestrales que predominan en la genética de cada provincia, por ejemplo: las provincias cuya genética ancestral está dominada por genes africanos muestran una mayor incidencia y mortalidad por enfermedades altamente asociadas a poblaciones con trasfondo genético africano como el cáncer de próstata. De forma similar, las provincias con un trasfondo genético predominante europeo y relativamente alto en genes africanos muestran una mayor incidencia y mortalidad a enfermedades ligadas a estas poblaciones como cardiovasculares y cerebrovasculares. Asimismo, los trastornos unidos con mayor incidencia en poblaciones de origen amerindio como el síndrome metabólico (obesidad, diabetes, entre otros) muestran mayor incidencia en provincias cuya genética ancestral es dominada por genes de origen amerindio, en específico: Bocas del Toro, Chiriquí y Panamá Los datos acotan que el componente genético ancestral de la población es un factor de riesgo altamente determinante en la distribución, susceptibilidad y muertes de ENT en el país. Desafortunadamente, no existen estudios biomédicos en la población panameña que abordan científicamente la genética de estas enfermedades principalmente porque se desconocen las variantes y polimorfismos genéticos asociadas con las mismas. Identificar y estudiar estas variantes es fundamental para la comprensión de las enfermedades y los factores de riesgo genético en la población que permitan mejorar el diagnóstico, prognosis, así como en el desarrollo de nuevos tratamientos más específicos de acuerdo con nuestra genética. Adicionalmente, la alta heterogeneidad ancestral de la población dificulta estudiar las enfermedades de genética compleja por métodos tradicionales que involucran analizar unos pocos genes a la vez.

Para identificar estas mutaciones se han realizado numerosos estudios de secuenciación genómica en poblaciones del mundo, principalmente en poblaciones de origen europeo y más recientemente africanos (mayormente afroamericanos) y asiáticos (Gurdasani et al., 2019;

Sirugo et al., 2019). Sin embargo, este tipo de estudios, donde se secuencian los genomas, no se ha desarrollado en la población panameña. Más aún, a nivel de poblaciones latinoamericanas los datos genómicos son muy pocos o incluso inexistentes en el caso de poblaciones indígenas (Gurdasani et al., 2019; Sirugo et al., 2019). La cantidad de estudios genómicos realizados en el mundo se ha dado prevalentemente sobre poblaciones de origen europeo; en particular, más de la mitad de los estudios de asociación genómica completa (del inglés *Genome-Wide Association Studies* o GWAS) son de origen europeo con un 52%, mientras que, el 21%, son asiáticos, principalmente de China. Además, algunas poblaciones de origen africano, principalmente afroamericanos, representan el 9,56% de estos estudios (Popejoy & Fullerton, 2016). Las poblaciones hispánicas o latinoamericanas están entre las poblaciones con el menor número de estudios llegando tan solo al 5.12%. Con relación al número de individuos muestreados para todos los estudios, el 80% son de origen europeo, el 10.22% de origen asiático, el 2.03% de origen africano y apenas el 1.13% de origen hispano o latinoamericano (Sirugo et al., 2019). Considerando el bajo número de latinoamericanos muestreados, aquel de los amerindios es aún mucho menor.

Estos datos indican que la mayoría de los genomas humanos secuenciados no son representativos de nuestra población, ya que el panameño muestra alta heterogeneidad y polimorfismos con diferentes proporciones de genes de las tres poblaciones ancestrales.

Dada la poca diversidad de genomas humanos secuenciados, esta subrepresentación de poblaciones étnicamente diversas es un obstáculo para el total entendimiento de la arquitectura genética de las enfermedades humanas y exacerba las desigualdades en la salud de poblaciones más huérfanas en estudios genómicos que permitan comprender sus enfermedades, su biología y su evolución. Además, la escasez de diversidad étnica en los estudios genómicos humanos no nos permitiría llevar la investigación genética a la práctica clínica o a políticas de salud pública de manera completa y acertada. La secuenciación del genoma completo (WGS) se usa cada vez más para inferir las causas de enfermedades raras no diagnosticadas, los genomas de referencia de poblaciones con mayor diversidad étnica son, por lo tanto, particularmente importantes.

Ante lo expresado anteriormente, para este estudio genómico desarrollado, nos hemos centrado en la población ancestral amerindia Ngöbe por ser la de mayor contribución al trasfondo genético de la población del país con aproximadamente un 51%. Por lo tanto, hipotetizamos que

muchas variantes y polimorfismos génicos asociadas a ENT pueden tener un origen ancestral en esta población (Castro-Pérez et al., 2016; Ramos et al., 2018). Además, los estudios previos que caracterizaron la genética de amerindios Ngöbe de Panamá han permitido crear un banco de ADN de esta población accesible para las investigaciones (Castro et al., 2007; Jorge-Nebert et al. 2002). Las muestras son de ADN genéticamente puro de poblaciones aisladas no mezclado con poblaciones mestizas ni con otras poblaciones amerindias como Emberá, como lo demuestran reportes anteriores (Castro et al., 2007; Castro-Pérez et al., 2016; Jorge-Nebert et al. 2002). En consecuencia, son muestras muy valiosas. Su análisis se hace necesario ejecutarlo con urgencia puesto que, con el pasar del tiempo, podría degradarse y probablemente las poblaciones amerindias como los Ngöbes podrían llegar a mezclarse por sus migraciones a las ciudades. Esto disminuiría su identidad genética única y, por consiguiente, su utilidad biomédica y la cantidad de individuos Ngöbe no entrecruzados genéticamente con otras poblaciones podría ser inaccesible en las próximas décadas haciendo perder su información para siempre (Castro-Pérez & Ramos, 2020). En un aspecto antropológico, identificar el origen de los alelos de ciertas enfermedades ayudaría a entender nuestra evolución, orígenes y cómo estas se habrían propagado en la población y dilucidar si hay tendencias específicas tanto en grupos de la población, así como en los individuos. Dependiendo de los alelos que se encuentren, se podrá determinar a qué enfermedades genéticas están asociadas. Los resultados alcanzados aquí representan los primeros datos genómicos generados de la población panameña (en particular de amerindios Ngöbe) y uno de los primeros a nivel latinoamericano. Estos datos son la base para futuras investigaciones genéticas y genómicas de enfermedades no transmisibles en el país.

OBJETIVOS DE LA INVESTIGACIÓN

Objetivo general:

1. Identificar variantes genéticas posiblemente asociadas a enfermedades no transmisibles en la población panameña mediante secuenciación genómica y análisis bioinformáticos.

Objetivos específicos:

1. Secuenciar el genoma completo de 4 individuos Ngöbe mediante secuenciación de nueva generación.
2. Analizar mediante estrategias bioinformáticas los datos genómicos enfocados en el exoma para identificar candidatos a posibles variantes génicas asociadas a enfermedades crónicas en el país.
3. Determinar la frecuencia de una variante genética asociada a alguna enfermedad no transmisible en una muestra de 50 individuos de la población ancestral amerindia Ngöbe.

HIPÓTESIS DE TRABAJO

Secuenciar y analizar el genoma de individuos Ngöbe nos permitirá identificar variantes y polimorfismos génicos asociadas a enfermedades en la población panameña.

CAPÍTULO II: ANTECEDENTES

1. Origen demográfico de las poblaciones ancestrales del continente americano

1.1 Origen Ancestral de las Poblaciones Indígenas del Continente Americano

El origen de las poblaciones indígenas de América es una de las preguntas que más ha intrigado desde la llegada de los europeos. El desarrollo de las técnicas moleculares en genética moderna ha permitido comprender en mayor profundidad y certeza estas preguntas. En particular, el ADN antiguo es una herramienta para entender las dinámicas evolutivas de poblaciones actuales y su relación con las ancestrales extintas mediante el uso de remanentes de huesos. Por ejemplo, Posth et al. (2018) usó el ADN antiguo para dilucidar preguntas importantes sobre las poblaciones de América Central y del Sur, proporcionando una herramienta importante para entender la historia de nuestros ancestros y el origen de nuestra diversidad genética; y develando especialmente la historia compleja de los nativo-americanos y sus descendientes (O'Connor, 2018). Una vez más, en referencia de los estudios Posth et al. (2018) se han secuenciado 49 nuevos genomas de nativos americanos antiguos algunos de los cuales representan los genomas más antiguos secuenciados, de los cuales uno es de Belice, América Central y tiene 9,300 años y el otro es de Chile, América del Sur y data 10,900 años. Estos datos fueron analizados en conjunto con las secuencias más antiguas muestreadas de América del Norte; un niño de 12,800 años de la cultura Clovis (Posth et al., 2018; Rasmussen et al., 2014). Los resultados demostraron que las poblaciones de América del Sur derivan de uno de los mayores linajes antiguos en América del Norte (Scheib et al., 2018).

Existen fuertes argumentos que ocurrió una continuidad sustancial de la población luego de los poblamientos iniciales de las diferentes zonas de las Américas. Los componentes nativos americanos de las poblaciones de las tres regiones geográficas principales, Caribe, América Central y América del Sur se dividieron hace aproximadamente 12,000 años (Gravel et al., 2013; Harris et al., 2018) y las poblaciones cercanas tuvieron divergencias evolutivas profundas (Moreno-Estrada et al., 2014). Las evidencias sugieren que hubo dos olas de migración entre el Norte y Suramérica, con la primera llegada desde la cultura Clovis o un grupo relacionado, que fue sustituido hace aproximadamente 9,000 años. Según los datos arqueológicos, no hay pruebas de la cultura Clovis en América del Sur a pesar del vínculo genético durante el mismo lapso. La primera razón posible es que, incluso con la migración de los Clovis hacia Suramérica, no se trasladó su tecnología, por lo que no quedó traza arqueológica. La otra explicación es que, los

ancestros comunes de los Clovis y los antiguos suramericanos no poseían la tecnología Clovis, sino que la desarrollaron después de que su grupo hermano partió por América del Sur.

La investigación también ha arrojado luz sobre la hipótesis de la estructura de poblaciones antiguas y la semejanza de nativo-americanos con poblaciones austroasiáticas. Esta relación fue propuesta con base en las similitudes en la morfología craneal (Neves & Hubbe, 2005). Sin embargo, los estudios genéticos que incluyen el estudio de O'Connor (2018) no encuentran una correspondencia entre la sospechada morfología craneal autralo melanesiana y la estructura poblacional dentro de los nativos americanos. Una hipótesis alternativa postuló que una origen de población Y de antigua estructura, relacionada con los autralo-melanesios, se puede encontrar en algunas poblaciones amazónicas y representan una estructura original que ingresó por el estrecho de Bering (Skoglund et al., 2015). No obstante, en Posth et al., (2018) no encuentran ninguna evidencia genética adicional de población y usando un muestreo extensivo de nativos americanos ancestrales.

1.2 Estudios Genómicos en Amerindios Latinoamericanos

Aunque los reportes anteriores han contribuido a un mejor entendimiento ancestral y evolutivo del origen y relaciones entre poblaciones indígenas de América, algunos autores (Belbin et al., 2018). enfatizan la necesidad de más estudios genómicos en poblaciones indígenas y latinas con enfoque biomédico. Las iniciativas de la secuenciación genómica en estas poblaciones ayudarían a comprender no sólo su historia y orígenes, sino también las bases de las enfermedades genéticas y fomentaría iniciativas de medicina moderna y personalizada La mayoría de los estudios genómicos con enfoque biomédico se ha desarrollado en las poblaciones del viejo mundo con ascendencia europea (**Figura 1**), mientras que hay escasez en la información de la estructura genómica de poblaciones latinoamericanas mestizas que representan la mayoría de la población y más aún datos genómicos de los nativos americanos, que, aunque representan una minoría demográfica, sus genes representan más del 50% en varias poblaciones latinoamericanas (Castro-Pérez et al., 2016).

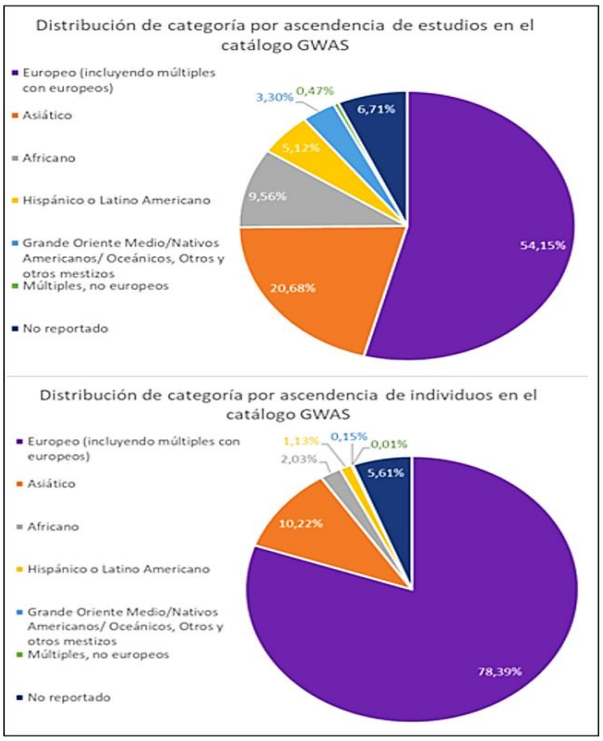


Figura 1. Resumen de los estudios GWAS por Ascendencia para estudios en el catálogo de GWAS a través de enero 2019. Se muestra la distribución de las categorías de ascendencias en porcentajes incluidos en el GWAS (https://www.ebi.ac.uk/gwas/home) basada en el número de estudios (arriba) y basada en el número total de individuos (abajo). Adaptado de (Sirugo et al., 2019).

Uno de los pocos estudios genómicos con enfoque biomédico en indígenas titulado *Whole Genome Sequence, Variant Discovery and Annotation in Mapuche-Huilliche Native South Americans* (Vidal et al., 2019) presenta la anotación y el descubrimiento de variantes de secuencias del genoma completo de alta calidad de un grupo de 11 indígenas Mapuche-Huilliche (HUI) del Sur de Chile. El estudio reporta aproximadamente 3.1 x 10^6 variantes de un solo nucleótido (SNVs) por individuo y se identificaron unos 403,383 (6.9%) eventos SNV novedosos. Los análisis de eventos genómicos a gran escala detectaron 680 variaciones en el número de copias (CNVs) y 4,514 variantes estructurales (SVs), incluyendo 398 y 1910 eventos novedosos, respectivamente.

A las variantes encontradas se les aplicaron distintos filtros para identificar variantes funcionales con potencial impacto deletéreo en los genomas HUI. Las variantes que pasaron estos filtros tomaron el nombre de Variantes con Potencial Impacto Funcional (VPFIs). Se encontraron múltiples VPFIs en genes funcionalmente enlazados que podrían modular la susceptibilidad no sólo de los Mapuche-Huilliche, sino también de la población chilena con mezcla ancestral de

estos amerindios a determinadas enfermedades no transmisibles (ENT) comunes, por ejemplo: el SNV rs1801133 en el gen MTHFR ha sido correlacionado con homocisteína plasmática alta (Jin et al., 2018). Curiosamente, el aumento en homocisteína (Hiperhomocisteinemia) está asociado a enfermedades cardiovasculares, hipertensión y neoplasmas (Kim et al., 2018), siendo estas las enfermedades no transmisibles comunes en Chile. La medición de homocisteína en las poblaciones Mapuche-Huilliche y el efecto étnico-específico para la variante rs1801133 en chilenos mixtos con ascendencia Mapuche-Huilliche pueden ser factores de riesgo por evaluar. A parte de esta variante, se reportaron otras que están asociadas a las 5 o 6 principales ENT que afectan a la población en Chile, lo cual tiene un impacto en el riesgo de enfermedades prevalentes en poblaciones chilenas y amerindias. Estos datos representan una fuente útil que contribuye a estudios basados en poblaciones y para el diseño de diagnósticos tempranos o herramientas de prevención para poblaciones nativo-americanas y latinas mixtas.

En otro estudio Ribeiro-dos-Santos et al. (2020), investigaron la variación del exoma completo de 58 individuos nativos americanos de 8 poblaciones localizadas en la parte oriental de la Cuenca del Amazonas y obtuvieron conocimientos sobre el poblamiento de América del Sur. La estructura genética de las poblaciones nativo americanas fue explorada y comparada con otras poblaciones del mundo con énfasis en las poblaciones nativo americanas, abarcando individuos contemporáneos y antiguos. El análisis PCA de todas las muestras de nativos americanos contemporáneos mostró que ellos se agrupaban según sus regiones geográficas. Cuando se realizó un análisis de mezcla genómica ancestral (*admixture*) sin supervisión, se encontraron unos claros componentes genéticos nativos americanos y nativos americanos antiguos con cinco componentes ancestrales (K=5), mientras que los otros tres eran africanos, europeos y asiáticos orientales. El componente nativo americano se divide ulteriormente en andino occidental y amazónico en el modelo con siete componentes ancestrales (K=7).

Datos de los análisis reportados en el estudio también sugieren que poblaciones del sudeste del Brasil y del norte de Argentina son más similares en su estructura genética a las poblaciones amazónicas que a las poblaciones andinas. Esto sugiere que la ocupación de la parte central de América del Sur involucró una ruta migratoria del norte de Brasil en vez de una ocupación del oeste. Los datos señalan que, con respecto a la ocupación y la expansión de los nativos americanos en Sudamérica, el proceso ocurrió en dos olas de migración separadas, lo más

probable a través de rutas del Pacífico y Atlántico con la parte sureste del continente siendo ocupada por migraciones desde la región Amazónica (Barbieri et al., 2019; Gnecchi-Ruscone et al., 2019; Reich et al., 2012; Tarazona-Santos et al., 2001; S. Wang et al., 2007).

1.3 Origen demográfico de nuestras poblaciones ancestrales en Panamá

La posición geográfica de Panamá ha determinado los patrones de migración de sus poblaciones ancestrales que dieron origen a su población actual. Estos patrones estaban asociados a la posición geográfica del istmo y a la función de este como país de tránsito en la época precolombina. Por un lado, desde que llegaron los españoles, Panamá se convirtió en un lugar caracterizado por un alto nivel de intercambio cultural de diferentes etnias debido al comercio (Jaén Suárez, 1998, 1978). La mayoría de los europeos que participaron en la conquista de América fueron hombres, por lo que su *cromosoma Y* estuvo bien representado en el lugar, lo cual contribuyó al cuello de botella poblacional de este linaje. Por otro lado, hay reportes indican que alrededor de 35,000 africanos fueron traídos a Panamá desde África Occidental en la época colonial, con el fin de reemplazar a los amerindios varones que habían sido esclavizados. Esta ola de genes africanos que llegaron al istmo ayudó a la supervivencia y recuperación de los varones amerindios y contribuyó más a los genes del *cromosoma Y* (Jaén Suárez, 1978). Hubo una segunda ola de genes africanos que se presentó del área afrocaribeña para la construcción del ferrocarril transístmico y una tercera migración de genes africanos llegó también desde las islas del Caribe para la construcción del Canal de Panamá.

En cuanto a los amerindios, en Panamá había muchos grupos pequeños de indígenas, que tenían una alta diversidad en cuanto a lenguas, según relatan las crónicas de los archivos de indias y datos etnohistóricos. Basados en estos reportes se estima que alrededor de 500,000 amerindios habitaban el territorio actual de Panamá en la época de la llegada de los españoles. Los dos grupos principales eran los Chibchas, estaban al Oeste de Panamá desde parte de la región de Coclé hasta Chiriquí y Bocas del Toro, y los Cueva, al Este, desde parte de Coclé hasta el Darién.

1.4 Estudios Genéticos en la Población Panameña

Actualmente, se estima que alrededor del 70% de la población panameña es mestiza (Arias et al., 2002; Castro-Pérez et al., 2016; Ramos et al., 2018). Aunque datos etnohistóricos indican que los amerindios, los europeos y los africanos fueron las principales poblaciones ancestrales que más contribuyeron al origen de la población panameña. Sin embargo, hasta hace algunos

años se desconocía la exacta proporción de genes que permanecieron de estas poblaciones ancestrales hasta nuestros días.

El primer estudio sobre la mezcla de genes de la población panameña se publicó en 2002 con el título "La mezcla racial de la población panameña" (Arias et al., 2002). En esta investigación se estudiaron 4,200 sujetos a lo largo de todo el país, de los cuales se analizaron marcadores genéticos clásicos de los grupos sanguíneos en los diferentes hospitales del país. Los sistemas genéticos utilizados para calcular la mezcla genética en el país y por provincia fueron con marcadores clásicos ABO y Rh, siguiendo un modelo trihíbrido, es decir, basado en las tres poblaciones ancestrales principales mencionadas anteriormente (Arias et al., 2002). En este estudio la estimación de la mezcla de cada provincia y del total del país se hizo empleando las frecuencias genéticas de los dos sistemas genéticos de las poblaciones ancestrales. Dado que no es posible determinar las frecuencias alélicas de las poblaciones ancestrales originales, se buscaron las frecuencias de sus representantes contemporáneos: para el componente blanco, se empleó la frecuencia de los españoles actuales (Planas et al., 1966); para el componente indígena se tomó la de los actuales indígenas que residen en Panamá y Costa Rica (Barrantes et al., 1990); y para el africano, se utilizó la de poblaciones del África occidental (Roychoudhury & Nei, 1988). Los análisis de mezcla ancestral se basan en que las poblaciones ancestrales tienen diferencia en sus frecuencias alélicas de estos marcadores genéticos. Los indígenas son todos de *tipo O*, mientras que hay más variación en los grupos africanos y en los europeos. El grupo B es más alto en los africanos (15%) y más bajo en los europeos (5,6%). La situación se invierte respecto al grupo A: es más elevado en la población blanca (28,6%) y más bajo en la negra (18,1%). En cambio, el grupo O es más balanceado (**Tabla 1**).

Tabla 1. Frecuencias ancestrales utilizadas en el estudio de Arias et al., (2002).

SISTEMA	ALELO	NEGRO	INDÍGENA	BLANCO
ABO	O	0.665	1.000	0.650
	A	0.181	0.000	0.286
	B	0.154	0.000	0.056
Rh*	+	0.7717	1.000	0.597
	-	0.2283	0.002	0.401

(Barrantes et al., 1990; Planas et al., 1966; Roychoudhury & Nei, 1988) *Las frecuencias del "alelo Rh+" son las sumatorias de las frecuencias de los genotipos que contienen el alelo de: CDE, CDe, cDE y cDe, mientras que las del "alelo Rh-" son las sumatorias de las frecuencias de los genotipos que no contienen este alelo: cde, Cde y cdE.

Estas diferencias alélicas nos ayudan a estimar las diferentes proporciones en contribución genéticas de las poblaciones ancestrales a la población actual. En la **Tabla 2**, se detallan las frecuencias génicas de los alelos correspondientes a cada sistema. Estos resultados se muestran por provincias de nacimiento de los sujetos, en la categoría de no clasificados y en el total de sujetos objeto de la muestra en toda la república. Por ejemplo, la frecuencia total de A es de 14.18%, la del grupo B, 6.62% y así sucesivamente. Esta diferencia se estima en cada provincia. Es importante señalar que estamos usando las frecuencias alélicas y no las fenotípicas, ya que estas indican simplemente el tipo sanguíneo de cada persona; ejemplo, una persona A puede ser homocigoto (AA) o heterocigoto (AO).

Tabla 2. Frecuencias génicas de alelos de los sistemas ABO y Rh por provincia de nacimiento.

	COCLÉ (n =233)	COLÓN (n = 146)	CHIRIQUÍ (n = 1312)	HERRERA (n = 323)	LOS SANTOS (n = 222)	PANAMÁ (n = 597)	VERAGUAS (n = 246)	NO CLASIFICADOS (n = 1123)[1]	TOTAL (n = 4202)
A	0.0922	0.1478	0.1395	0.1417	0.1996	0.1365	0.1188	0.1514	0.1418
B	0.0416	0.1047	0.0317	0.0943	0.1017	0.0968	0.0500	0.0803	0.0662
O	0.8662	0.7475	0.8287	0.7640	0.6987	0.7667	0.8311	0.7687	0.7920
Rh+	0.2173	0.8149	0.8028	0.7918	0.7878	0.7544	0.7450	0.7911	0.7862
Rh-*	0.7827	0.1851	0.1972	0.2082	0.2122	0.2456	0.2550	0.2089	0.2138

Las frecuencias del “alelo Rh+” son las sumatorias de las frecuencias de los genotipos que contienen el alelo D: CDE, CDe, cDE y cDe, mientras que las del “alelo Rh” son las sumatorias de las frecuencias de los genotipos que no contienen este alelo: cde, Cde y cdE. Adaptado de Arias et al., (2002).

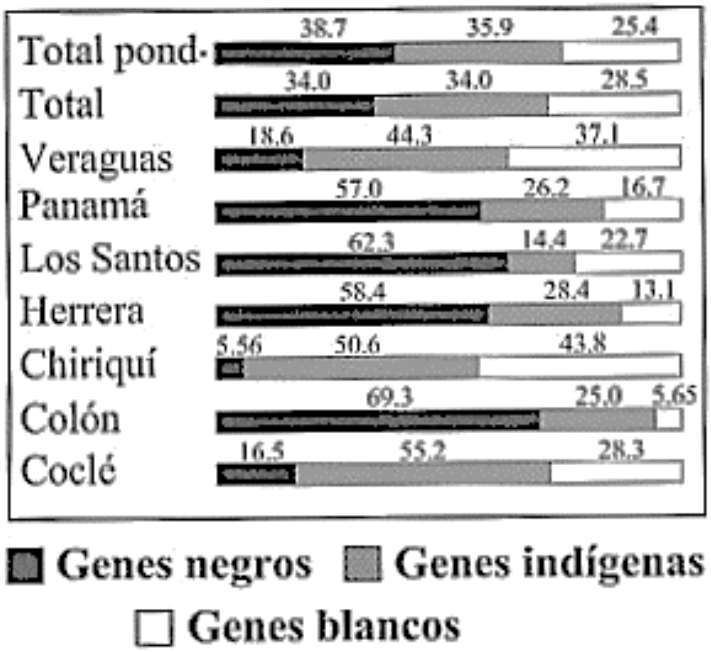


Figura 2. Diagrama de barras horizontales apiladas del porcentaje de mezcla racial por provincias y para el total del país. Se incluye además es total ponderado para el país. Genes "negros" se refiere a genes de origen africano y genes "blancos" se refiere a genes de origen europeo. Adaptado de Arias et al., (2002).

Este estudio demostró por primera vez patrones diferenciales y polimórficos entre las provincias del país con una población relativamente pequeña. Ha habido aislamiento en cada provincia, lo cual se refleja en el componente genético bastante diferente entre ellas. Hay varias explicaciones de este fenómeno en Panamá, a pesar de que este es un país pequeño con una población relativamente pequeña, que a finales de años 90's se acercaba a los 3 millones de habitantes. Una de las razones de este aislamiento en estas poblaciones es que, en parte, no fue hasta 1962 que se construyó el Puente de las Américas y no fue hasta entonces en 1967 que se construyó la Carretera Panamericana. Esto quiere decir que, la precedente falta de vías de comunicación desmotivaba la dinámica de la migración de manera que cada población se mantuvo relativamente aislada (Castro-Pérez, 2022). La construcción de la Carretera Interamericana y del Puente de las Américas causó también un movimiento migratorio y explosión demográfica del interior del país a la capital en especial en áreas como San Miguelito y las afueras del este y del oeste de la ciudad de Panamá. El estudio demostró, además, que la proporción de genes de origen africano, indígena y europeo en Panamá se distingue de los demás países latinoamericanos (**Tabla 3**), ya que no se ha observado en ellos una proporción como esta.

Tabla 3. Comparación de la proporción de genes en los países latinoamericanos

País	Negro	Indígena	Blanco
Argentina	N. C.	18.4	81.62
Chile (Santiago)	N.C.	31.0	69.0
Rep. Dominicana	43.0	17.0	40.0
Puerto Rico	37.0	18.0	45.0
Cuba	20.0	18.0	62.0
México**	6.1	56.2	37.7
Costa Rica	9.05	29.91	61.04
Jamaica	93.2	N. C.	6.8
Panamá	38.7	35.9	25.4

Adaptado de Arias et al., (2002).

En otra investigación Castro et al., (2007), se realizaron un estudio genético con nueve loci microsatélites del tipo de repeticiones cortas en tándem (STR, Short Tandem Repeats), Estos nueve marcadores (CSF1PO, TPOX, TH01, F13A01, FESFPS, VWA, D16S539, D7S820, y D13S317). Se analizaron en los amerindios Ngöbe y Emberá. El artículo se titula "*Genetic Polymorphism and Forensic Parameters of Nine Short Tandem Repeat Loci in Ngöbé and Emberá Amerindians of Panama*". Se determinó la diversidad genética de estas dos poblaciones. Mediante las frecuencias alélicas obtenidas, se compararon estadísticamente las diferencias entre ellas. También se calcularon parámetros poblacionales para evaluar su deficiencia alélica y de heterocigosidad. En esa misma línea, estuvieron incluidos parámetros forenses que permitieron cálculos certeros para hacer identificaciones genéticas en estas poblaciones.

En la **Tabla 4**, se muestran las frecuencias alélicas en cada locus STR, así como el porcentaje de homocigotos y heterocigotos. Ambas poblaciones comparten sus alelos con las frecuencias más elevadas en los siete loci, pero hay un locus que muestra diferencias marcadas: el locus TPOX, que demostró su mayor frecuencia en el alelo 11 para los Ngöbe y en el alelo 6 para los Emberá. De hecho, el alelo del locus TPOX presente en los Ngöbe está ausente en los Emberá y viceversa, lo cual sugiere que los alelos pueden ser distintivos para cada población. Estos resultados demostraron que, aunque ambas poblaciones conservan similitudes en la distribución

de sus frecuencias alélicas, tienen diferencias significativas que permiten la discriminación genética de las mismas, como también se confirmó en otro estudio (Castro-Pérez et al., 2016). También se determinaron parámetros forenses y de paternidad, los cuales demostraron que la mayor parte de los marcadores STR son informativos para objetivos forenses.

Tabla 4. Frecuencias alélicas de los nueve loci de STR en los amerindios Ngöbe (NG) y Emberá (EM) de Panamá.

	CSF1PO		*TPOX*		*TH01*		*F13A01*		*FESFPS*		*VWA*		*D16S539*		*D7S820*		*D13S317*	
Allele	*NG*	*EM*	*NG*	*EM*	*NG*	*EM*	*NG*	*EM*	*NG*	*EM*	*NG*	*EM*	*NG*	*EM*	*NG*	*EM*	*NG*	*EM*
3.2							0.48	0.38										
3.3								0.01										
4							0.31	0.11										
5	0.00	0.00		0.01	0.00	0.00	0.13	0.24					0.00	0.00				
6	0.00	0.00	0.00	0.54	0.86	0.00	0.07	0.02					0.00	0.00	0.00	0.00		
7	0.00	0.00	0.00	0.39	0.13	0.01	0.01	0.23	0.00	0.00			0.00	0.00	0.00	0.00	0.00	0.00
8	0.00	0.05	0.40	0.02	0.00	0.55	0.00	0.01	0.00	0.00			0.01	0.00	0.00	0.05	0.00	0.01
9	0.06	0.04	0.01	0.00	0.00	0.01	0.00	0.00	0.00	0.00			0.15	0.18	0.01	0.01	0.56	0.33
9.3				0.05	0.01	0.00												
10	0.03	0.27	0.00	0.00	0.00	0.01	0.00	0.00	0.01	0.10			0.37	0.37	0.18	0.25	0.04	0.13
11	0.33	0.14	0.48	0.00	0.00	0.20	0.00	0.00	0.77	0.65	0.00	0.00	0.25	0.19	0.43	0.45	0.05	0.10
12	0.46	0.43	0.11	0.00		0.23	0.00	0.00	0.21	0.14	0.00	0.00	0.22	0.22	0.37	0.21	0.10	0.22
13	0.13	0.07	0.01	0.00		0.01	0.00	0.00	0.01	0.11	0.00	0.00	0.00	0.03	0.01	0.04	0.17	0.14
14	0.00	0.00					0.00	0.00	0.00	0.00	0.06	0.06	0.00	0.01	0.00	0.00	0.09	0.07
15	0.00	0.00					0.00	0.00			0.09	0.09	0.00	0.00			0.00	0.00
16							0.00	0.00			0.60	0.56						
17											0.23	0.19						
18											0.01	0.11						
19											0.01	0.00						
20											0.00	0.00						
21											0.00	0.00						
Homozygosity (%)	23.3	28.4	39.3	50.5	75.8	38.2	51	32.6	64.6	41.3	54	42.2	43.9	27.1	31.6	36.5	43.9	15.6
Heterozygosity (%)	76.7	71.6	60.7	49.5	24.2	61.8	49	67.4	35.4	58.7	46	57.8	56.1	72.9	68.4	63.5	56.1	84.4
Total chromosomes	120	176	122	186	124	178	98	92	96	92	100	90	114	192	114	192	114	192

Adaptado de Castro-Pérez et al., (2016).

Tabla 5. Parámetros forenses y de paternidad de los nueve loci de STR in las poblaciones amerindias de Ngöbe y Emberá

	Ngöbé				*Emberá*			
Locus	*MP* [a]	*PD* [b]	*PE* [c]	*TPI* [d]	*MP*	*PD*	*PE*	*TPI*
CSF1PO	0.2	0.8	0.539	2.14	0.131	0.869	0.453	1.76
TPOX	0.255	0.745	0.299	1.27	0.279	0.721	0.183	0.99
TH01	0.602	0.398	0.042	0.66	0.23	0.77	0.313	1.31
D16S539	0.124	0.876	0.247	1.14	0.115	0.885	0.475	1.85
D7S820	0.202	0.798	0.404	1.58	0.154	0.846	0.336	1.37
D13S317	0.173	0.827	0.247	1.14	0.085	0.915	0.683	3.2
F13A01	0.184	0.816	0.179	0.98	0.133	0.867	0.389	1.53
FESFPS	0.465	0.535	0.088	0.77	0.254	0.746	0.276	1.21
VWA	0.245	0.755	0.155	0.93	0.172	0.828	0.265	1.18
Combined	–	0.9999	0.9338	2.58	–	0.9999	0.989	40.44

a. Matching probability.
b. Power of discrimination.
c. Power of exclusion.
d. Typical paternity index.

Adaptado de Castro-Pérez et al., (2016).

Otro estudio, publicado con el título "*Genetic Ancestry of the Panamanian Population: Polymorphic Structure, Chibchan Amerindian Genes; and Biological Perspectives on Diseases*" (Castro-Pérez et al., 2016), se analizaron los mestizos, que representan la mayoría de la población general del país. Se muestrearon 650 individuos, a los cuales se les analizaron 15 marcadores polimórficos STR. Se determinó la diversidad y estructura genética de la población a nivel general y por provincias. Adicionalmente, se abordó la pregunta del origen de la ascendencia de las tribus amerindias en la población. Se analizó también si las diferencias en la mezcla ancestral entre las provincias podrían ser asociadas a las diferencias en incidencia de ENT.

Para abordar estas preguntas se examinaron microsatélites polimórficos STR en las poblaciones ancestrales putativas Ngöbe (Chibchas) y Emberá (Chocoes) reportados previamente (Castro et al. 2007) junto con datos de estos mismos marcadores en poblaciones actuales de España (Camacho et al., 2007) y África occidental desde Angola (Beleza et al., 2004) y Guiné-Bissau (Gonçalves et al., 2002). Los datos de estas poblaciones fueron analizados en combinación con los datos STR de los 650 mestizos panameños. Esta estrategia de análisis revelaría la contribución genética de las poblaciones amerindias ancestrales putativas, así como las contribuciones europeas y africanas (**Figura 4**).

Esta investigación se enfocó con particular interés en identificar el origen de los genes amerindios de la población panameña, ya que, aunque datos etnohistóricos y genéticos reportaban de manera bastante aproximada que los españoles trajeron el componente europeo mientras que los africanos fueron traídos desde África del Oeste hacia Panamá en la época colonial (y más recientemente para la construcción del ferrocarril y del Canal de Panamá), el componente genético amerindio había sido menos claro en términos de si tenía origen tribal específico o si era una mezcla de diferentes amerindios, por ejemplo: reportes anteriores sugerían que el proceso de mestizaje involucró distintos grupos indígenas, pero no se entendía con claridad cuál fue su contribución específica en el trasfondo genético debido a la presencia de distintas tribus y lenguas que ocupaban Panamá. En este sentido, los reportes indican que, en el siglo XVI, en la parte del Oeste del país (desde Coclé a Chiriquí y Bocas del Toro), habitaban pueblos con lenguas diferentes afines a los Chibchas, quienes eran los ancestros de los actuales Ngäbe (**Figura 3**). En la parte oriental (desde Coclé hasta el Darién), habitaban los Cueva, quienes se extinguieron totalmente entre 30 ó 50 años después de la conquista. Los Chibchas estaban bien caracterizados lingüística y genéticamente, porque hoy en día, sobreviven los Ngöbe y otras tribus afines a Costa Rica y Panamá. Sin embargo, no era claro si los Cueva pertenecían al grupo Chibchan o al grupo Chocoe; a los cuales pertenecen los Emberá actuales, algunos autores consideraban que los Cueva pertenecían al grupo Chibchan; mientras que otros, consideraban que pertenecían a los Chocóes (Constenla, 1991; Loewen, 1963). Tampoco se sabía si los Cueva contribuyeron a la genética panameña actual o si se extinguieron antes de aportar significativamente al mestizaje. En este sentido, el estudio reportó que la población panameña muestra niveles altos de polimorfismo y mezcla ancestral significativamente diferente entre las provincias del país (**Figura 4**) con contribuciones relativamente elevados por parte de las tres poblaciones ancestrales: 24% africano, 25 % europeo y 51% amerindio, lo cual es consistente con estudios previos (Arias et al., 2002). Además, se demostró que el componente genético amerindio es de origen Chibchan, mostrando gran distribución dentro de las regiones del Oeste y del Este. Esto sugiere que los Cueva probablemente eran Chibchas o que si eran Chocoes se extinguieron antes de contribuir significativamente al proceso de mestizaje.

Figura 3. Arriba: Distribución pre-colombina de amerindios Chibchas y Chocóes y su relación con ciudades fundadas por los españoles durante la época colonial. Adaptado de Castro-Pérez et al. (2016). La mayor parte de estas ciudades fueron fundadas en lugares ocupados originalmente por asentamientos indígenas. Algunas de estas ciudades existen todavía y se convirtieron en las capitales de provincias panameñas actuales y en la ciudad capital del país. Abajo: Distribución actual de las tribus indígenas y provincias panameñas. Estas regiones urbanas/ciudades de provincia son las áreas principales donde se colectaron muestras de ADN. Este mapa fue adaptado de T. D. Arias et al. (1992); Barrantes et al. (1990); Jopling (1994) y Romoli (1987); y Castro-Pérez et al., (2016).

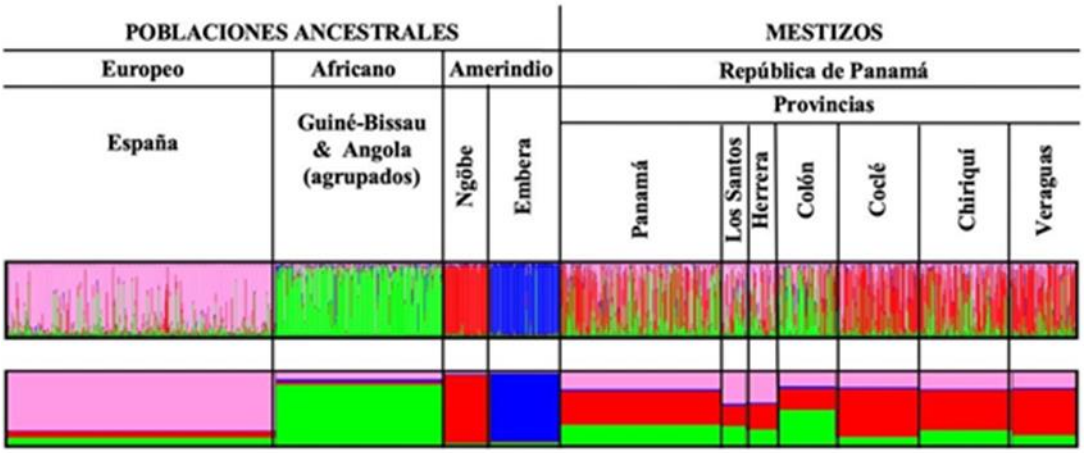


Figura 4. Estructura Genética Ancestral de la Población Panameña. Análisis genético de 15 Marcadores Polimórficos STR utilizando el Programa STRUCTURE (N = 650 individuos). Los Código de Colores corresponden a grupos de genes de las diferentes poblaciones ancestrales. Rosa = Genes Europeos; Verde = Genes Africanos; Rojo = Genes Indígenas. Arriba: Cada línea vertical representa un sujeto analizado con su mezcla genética ancestral individual. Abajo: Promedio de contribución de cada población ancestral por provincia. Adaptado de Castro-Pérez et al., (2016).

El estudio demostró que, en general, los mestizos panameños son altamente polimórficos y mezclados diferencialmente entre las provincias del país como se evidencia por los parámetros de diversidad y distancia genética calculadas y el análisis en STRUCTURE (**Figura 4**). Dadas las diferencias significativas en contribución ancestral entre las provincias, en el estudio se hipotetizó que estas diferencias podrían también estar asociadas a diferencias en la incidencia de ENT. Por lo tanto, se calcularon los parámetros de relevancia biomédica incluyendo datos forenses y epidemiológicos de las mayores enfermedades que afectan a la población y se focalizó en el cáncer de próstata y trastornos cerebrales y cardiovasculares.

Por un lado, los datos determinaron que el cáncer de próstata mostró una incidencia mayor en las provincias dominadas por genes africanos (Panamá y Colón). Con relación a este último, muchos de los datos que están reflejados en la barra de Panamá (**Figura 5A**) corresponden con mayor probabilidad a pacientes de Colón, considerando que muchos de ellos se atienden en Panamá, por lo que estos datos pueden ser una mezcla de ambas provincias. El hecho de que el cáncer de próstata tiene una mayor incidencia en personas con origen africano, refleja la idea que estas provincias con alto porcentaje de genes de ese origen, teóricamente, podrían ser más susceptibles o tener mayor riesgo a unas enfermedades que afectan con mayor incidencia a las poblaciones africanas ancestrales. De forma llamativa, múltiples reportes indican que hay

considerables disparidades en el riesgo de cáncer de próstata, que es desproporcionadamente mayor entre los afrodescendientes (Bock et al., 2009; Robbins et al., 2007; Zeigler-Johnson et al., 2008).

Por otro lado, los datos determinaron que los trastornos cerebro cardiovasculares mostraron una incidencia mayor en las provincias con mayor mezcla europea y genes africanos moderadamente altos (Herrera y Los Santos) (**Figura 5B**). Consistente con estos datos, algunos reportes previos indican que hay considerables disparidades en el riesgo de enfermedades cardio y cerebrovasculares, que es desproporcionadamente mayor entre personas de ascendencia europea y afrodescendientes (Cheng et al., 2010; Donnan et al., 2008; Hankey, 1999).

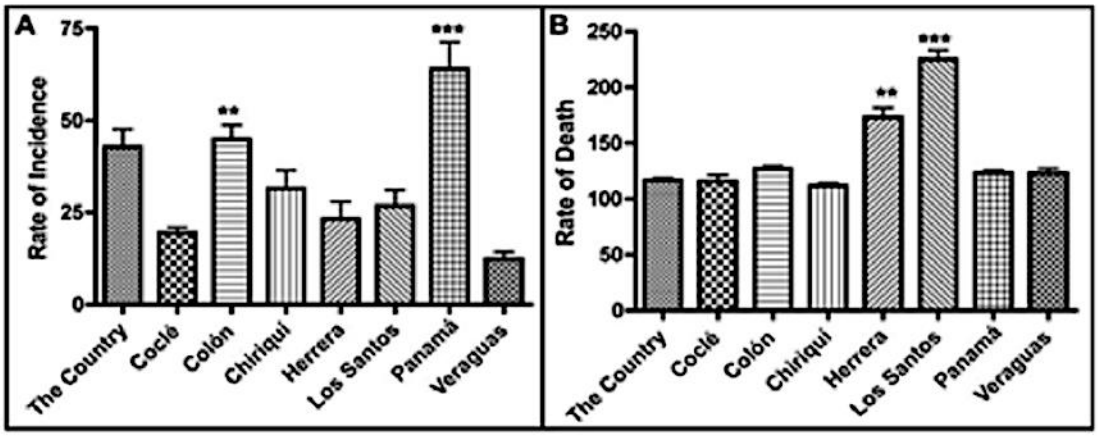


Figura 5. Incidencia de ENT por provincia con diferente trasfondo genético ancestral.A. Incidencia de cáncer de próstata en provincias por 100,000 habitantes varones. One-Way ANOVA y estadísticos de Barlett Post-correction determinaron que las provincias con la mayor proporción de genes africanos en su trasfondo genético (Colón y Panamá) muestran la mayor incidencia de cáncer de próstata; $P < 0.001$ para Panamá y $P < 0.01$ para Colón. B. Muertes causadas por enfermedades cardiovasculares y cerebrovasculares por provincia por 100,000 habitantes. One Way-ANOVA y estadísticos de Barlett Post-correction demostraron que Los Santos y Herrera, las provincias con mayor proporción de genes ancestrales europeos y moderadamente alta proporción de genes africanos, exhiben la mayor tasa de muertes por enfermedades cardio y cerebrovasculares; $P < 0.001$ para ambas provincias. Adaptado de Castro-Pérez et al., (2016).

Luego de este estudio, se realizó otro similar, pero con marcadores INDELs titulado "*Analysis of 30 INDEL Polymorphic Markers in the Panamanian Population*: *Gene Admixture Estimates, Population Structure and Forensic Parameters*" por parte de (Ramos et al., 2018). Para este estudio se seleccionaron 350 sujetos de origen mestizo, mitad varones mitad mujeres con padres y abuelos nacidos en Panamá, así como 30 marcadores InDel. El objetivo del estudio era

determinar los clústeres del país y su estructura genética, así como estimar los parámetros de aplicación forense y paternidad de estos marcadores. Los resultados fueron muy significativos con los anteriores y confirmaron los niveles de diversidad, polimorfismo y contribución ancestral trihíbrida (K=3) entre las provincias y la población general del país con 46% de genes indígena, 30% de contribución europea y 24% de origen africano (**Figura 6**).

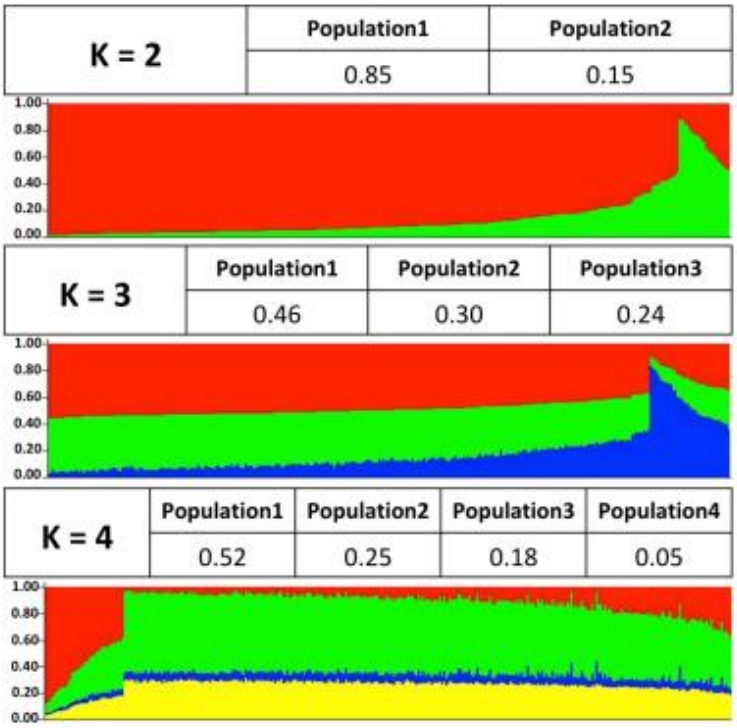


Figura 6. Modelos de mezcla y proporciones ancestrales de cada clúster (color) poblacional predefinido del país total estimados usando STRUCTURE. Structure harvester calculó los valores de los plots L(K) de las probabilidades Ln más altas. Adaptado de Ramos et al., (2018).

2. Marcadores genómicos

En los estudios precedentes, se usaron marcadores moleculares como STR e InDel, los cuales arrojaron información sobre la estructura genética de Panamá, pero la cantidad de genes que analizaron era mínima y de manera no específica. Para este estudio, en cambio, están involucrados marcadores genómicos que estudian una cantidad enorme de genes contemporáneamente y de manera específica, para poder así identificar múltiples variantes/polimorfismos de genes candidatos posiblemente asociados a enfermedades no transmisibles. Estos marcadores genómicos son las variantes en número de copias de genes *CNVs* (del inglés *copy number variations*). También usamos inserciones/deleciones, conocidas

también como *INDELs*; así como variantes estructurales del genoma conocidas como *SV* (del inglés *structural variations*) y finalmente usamos polimorfismos de nucleótidos simples, conocidas como *SNPs* (del inglés *single nucleotide polymorphisms*). A continuación, explicaremos brevemente cada uno de los marcadores mencionados.

2.1 Variaciones Estructurales (SV)

Tomando como guía, según Novogene (2022): "las variantes estructurales (SVs) son variantes genéticas de tamaño relativamente grande (>50 pb), incluyendo inversiones, duplicaciones, translocaciones, deleciones e inserciones. Las SVs podrían formar la base genética subyacente a las diferencias individuales, y tienen efecto potencial en la susceptibilidad a enfermedades y cánceres".

2.2 Inserciones/Deleciones (InDels)

Se añaden más descripciones respecto a las inserciones, como establece Novogene (2022): "Las inserciones/deleciones (InDels) son otro tipo de marcadores menores de 50 pb de longitud y constituyen otra clase de variantes genéticas en el genoma humano con posible asociación a polimorfismos de enfermedades. Los InDels que ocurren en las regiones codificantes o sitios de corte y empalme de exones pueden causar cambios en los transcritos y en las proteínas. Si el número de nucleótidos insertados o eliminados no corresponde a un codón o no es un múltiplo de tres, el marco de lectura completo se alteraría".

2.3 Polimorfismos de Nucleótidos Simples (SNPs)

Novogene (2022) expresa que: "los polimorfismos de nucleótidos simples (SNPs), también conocidos como variantes de nucleótidos simples (SNVs), constituyen la clase más amplia de variantes genéticas en el genoma. Un típico genoma humano tiene alrededor de 3.6 millones de SNPs".

2.4 Variación en Número de Copias (CNV)

Novogene (2022) atribuye las propiedades de la variación en número de copias como "el último grupo de marcadores analizados son las Variaciones en Número de Copias (CNVs), del inglés *Copy Number Variation*. Las CNVs son variantes genéticas que conducen a variaciones en el número de copias de fragmentos relativamente largos (más largos que 50 pb) entre individuos.

Hay dos tipos de CNVs: ganancia y pérdida de copias. Los CNVs podrían formar la base genética subyacentes de las diferencias individuales y de los cánceres".

Las variantes estructurales tienen una distribución en el genoma que no es al azar. Por lo general tienden a estar en regiones ricas de repeticiones comunes y duplicaciones de segmentos, pero el sesgo mayor ha sido observado en los últimos 5 Mpb de los brazos de los cromosomas (Audano et al., 2019). Por cuanto respecta el número de variaciones de copias, esta variante afecta del 4.8 al 9.7% del genoma humano (Zarrei et al., 2015). En cuanto a los polimorfismos de nucleótidos simples, hay uno cada 300 pb dentro del genoma humano. La mayor parte se da en la región intergénica, seguida por la intrónica (*Human Whole Genome Sequencing Project Demo Report (Disease)*, 2016). Las inserciones/deleciones se distribuyen a lo largo del genoma humano dentro de todos los cromosomas (Mullaney et al., 2010). También se encontró que la mayor parte de ellas se dan en la región intergénica, seguida por la intrónica (*Human Whole Genome Sequencing Project Demo Report (Disease)*, 2016).

3. Identificación del polimorfismo rs1801133 del gen MTHFR

La secuenciación del genoma y análisis de exoma en amerindios Ngöbe de Panamá nos condujo a la identificación de múltiples variantes génicas candidatas asociadas a enfermedades. Sin embargo, en esta investigación nos enfocamos en el polimorfismo rs1801133 del gen que codifica la enzima metilentetrahidrofolato reductasa (MTHFR) como un candidato vinculado a las enfermedades en la población panameña (explicado en detalle más adelante). Hasta la fecha, se han descrito más de 100 diferentes mutaciones clínicamente relevantes en el MTHFR, la mayor parte de las cuales son mutaciones con cambio de sentido (del inglés *missense mutation*) (n=70, >60%). Deficiencias enzimáticas menos severas, debidas al polimorfismo de nucleótido único del gen MTHFR, han sido relacionadas a varios trastornos comunes, entre los cuales, el más estudiado es el p.Ala222Val [c.665C>T en MN_005957 según el sistema de nomenclatura de *Human Genome Variation Society* (HGVS), comúnmente anotada como c.677C>T según la nomenclatura clásica (den Dunnen et al., 2016; *Human Genome Variation Sequence*, s. f.) identificado como un factor de riesgo para un agobiante número de trastornos multifactoriales, que incluyen enfermedades vasculares, neurológicas, varios tipos de cánceres, diabetes y pérdida del embarazo (Froese et al., 2018).

El metilentetrahidrofolato reductasa (MTHFR) es una enzima reguladora clave en el metabolismo del folato y de la homocisteína. La clonación de la secuencia codificante del MTHFR condujo a la identificación de las primeras mutaciones deletéreas en el MTHFR en pacientes con homocistinuria. El polimorfismo rs1801133 es una mutación puntual en la posición 665 de citosina a timina en la región codificante en el exón 5, que genera un nuevo codón, por lo que se sustituye la alanina por valina (A222V) en la enzima MTHFR (ClinVar, 2023). Sin embargo, este polimorfismo es comúnmente conocido como C677T debido a que la secuencia de referencia humana de ADNc determinada por (Goyette et al., 1994) estaba incompleta, por lo que la enumeración de los nucleótidos empezó desde el enlazador al inicio de la secuencia. En consecuencia, la posición de la sustitución de C a T resultó en el nucleótido 677 de la secuencia dentro del exón 4 (Rosenberg et al., 2002). Este principio de enumeración se usó para describir la variante termolábil por Frosst et al. (1995) y prácticamente en todas las publicaciones sucesivas (Leclerc et al., 2013), incluso en la actualidad. Por esta razón, en este trabajo este polimorfismo está referido preferiblemente como C677T.

Este polimorfismo de nucleótido único reduce la termoestabilidad de la enzima MTHFR debido a la actividad disminuída de la enzima a 37° o más. La actividad enzimática de MTHFR en sujetos homocigotos es menor en un 50-60% a 37° y 65% menor a 46° comparado con los controles no mutados normales (Kang et al., 1988; Rozen, 1997).

Esta variante se ha reconocido como la causa más común de la hiperhomocisteinemia y ha sido extensamente investigada como un factor de riesgo para diferentes trastornos multifactoriales asociados a alteraciones en el metabolismo de la homocisteína. La progresión experimental, desde el delineamiento del trastorno severo raro del metabolismo hasta las consecuencias menos deletéreas de la mutación leve del 677C→T, fue facilitada por la elucidación de información molecular sobre el MTHFR (Leclerc et al., 2013).

3.1 Generalidades del gen MTHFR

El gen MTHFR es un gen del *cromosoma 1* que codifica para la enzima metilentetrahidrofolato reductasa, localizado específicamente en 1p36.22. Posee una longitud de 2.2 kb y está compuesto por 13 exones (anteriormente contabilizados hasta 11) (Goyette et al., 1998; NCBI, 2023). La región promotora del gen MTHFR no tiene una caja TATA, pero contiene islas CpG,

múltiples sitios de unión potenciales para SP1 y sitios de unión para otros factores de transcripción (Gaughan et al., 2000).

El mayor producto del gen humano MTHFR es una proteína de 77 kDa, una segunda isoforma humana de aproximadamente 70 kDa. También se ha observado a través de Western blot. Un corte y empalme alternativo complejos en el extremo 5' del MTHFR fue reportado por Chan et al, (1999) y otros (Homberger et al, 2000). Posteriormente, se identificó el sitio de inicio de transcripción corriente arriba predicho del MTHFR, generando un mRNA producto del corte y empalme alternativo cuyo ADNc se logró clonar y expresar. La expresión de este ADNc produjo la isoforma más grande de 77 kDa (Tran et al., 2002).

Los datos demostraron que los transcritos de distinto tamaño de MTHFR son el resultado de sitios de inicio de transcripción alternativos y múltiples señales de poliadenilación. (Tran et al., 2002). Algunos transcritos que se originan en la región corriente arriba resultan del corte y empalme alternativo y no contienen el ATG presente en la isoforma larga. Se predice que estos transcritos traducen la proteína de 70 kDa, con una UTR 5' de aproximadamente 50 nucleótidos (Tran et al., 2002; van der Velden y Thomas, 1999).

3.2 Estructura de la enzima metilentetrahidrofolato reductasa

La MTHFR humana es una proteína multidominio de 656 aminoácidos. El dominio catalítico es conservado a través de la evolución (**Figura 7**). El dominio catalítico forma un barril TIM ($\beta_8\alpha_8$) y tiene residuos críticos para el ligamiento del cofactor FAD, el donante de electrones NADPH y el producto CH_3-THF. En las bacterias la enzima posee solo el dominio catalítico. En los eucariotas además existe un dominio regulador terminal C, conectado al dominio catalítico por una secuencia linker. Este dominio terminal C es capaz de ligar S-adenosilmetionina SAM, lo cual resulta en una inhibición alostérica de la actividad enzimática, un efecto que es muy lento y puede ser revertido por la unión de S-adenosilhomocisteína (SAH), la forma demetilada de SAM. La MTHFR humana contiene además una región de 35 aminoácidos rica en serina en la región N terminal. A partir de expresiones heterólogas en células de insectos y levaduras se ha observado diferentes patrones de fosforilación de esta región. Resultados similares se observaron mediante inmunoprecipitación en líneas de células cancerosas humanas. La fosforilación se ha asociado a la disminución moderada de la actividad catalítica y al aumento de la inhibición total mediado por el SAM (Froese et al., 2018; Yamada et al., 2001).

Dado que no hay interfaz directa entre el sitio activo del dominio catalítico y el dominio regulador, el ligamiento del SAM provoca la inhibición enzimática a través de un cambio conformacional propagado desde el dominio regulador hasta el catalítico. El efecto más probable de este cambio conformacional es la extensión de la región linker dado que crea múltiples contactos a ambos dominios, el regulador y el catalítico y forma parte del sitio de unión de SAM/SAH.

Experimentos de la expresión *in vitro* (Shan et al., 1999). sugieren que la presencia de la terminal C tiene un efecto inhibidor en la actividad MTHFR. Este hallazgo es consistente con la localización del dominio inhibido del SAM en la región terminal C. Datos in vivo en la levadura sugieren que el dominio del terminal C es crítico para el crecimiento celular

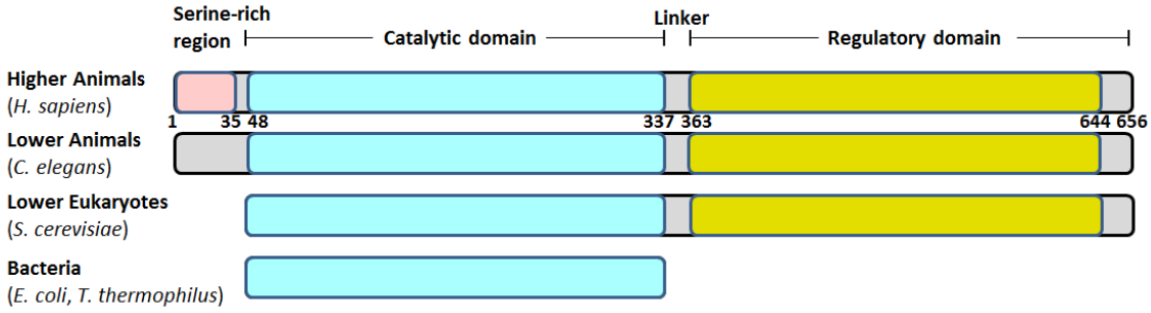


Figura 7. Organización de los dominios de los ortólogos de MTHFR a través de la evolución. Fuente: Froese et al. (2018).

3.3 Función y mecanismo de regulación de metilentetrahidrofolato reductasa

La metilentetrahidrofolato reductasa cumple un rol fundamental en el metabolismo del folato y de la homocisteína (**Figura 8**). El primero es el de portador celular mayor de unidades de un carbono. Es necesario para la síntesis de purinas y timidin monofosfato. El segundo rol es un producto azufrado intermediario que resulta del metabolismo de la metionina.

En el ciclo del folato, la proteína metilentetrahidrofolato reductasa cataliza la reducción irreversible de 5,10-metilentetrahidrofolato (CH_2-THF) a 5-metiltetrahidrofolato (CH_3-THF), que es un cosustrato para la remetilación de homocisteína a metionina. Esta reacción requiere de FAD como cofactor y NADPH como donador de electrones. El producto CH_3-THF es utilizado exclusivamente por la metionina sintasa, y solo la forma demetilada, o sea

tetrahidrofolato (THF), puede ser reciclada en el ciclo del folato. Debido a esto, MTHFR envía las unidades de un carbono ligadas al THF al ciclo de la metionina (Froese et al., 2018).

Dentro del ciclo de la metionina, la homocisteína es convertida en metionina por la 5-metiltetrahidrofolato homocisteína metiltransferasa mediante una reacción de transmetilación en presencia de la vitamina B12 y el ácido fólico. De esta forma, la proporción de la vitamina B12 y el ácido fólico respecto al nivel de homocisteína están inversamente relacionados y la suplementación de la vitamina B12 y el ácido fólico reduce el nivel de homocisteína plasmática.

La metilación de la homocisteína a metionina por la metionina sintasa produce un aminoácido esencial que puede ser usado para la síntesis de proteínas o convertido a S-adenosilmetionina (SAM), un donador importante para la metilación del ADN, ARN y proteínas, así como la metilación de una gran cantidad de compuestos. Estos dos ciclos intersecan a la enzima 5,10-metilentetrahidrofolato reductasa.

La homocisteína en cierta proporción también es convertida es cisteína mediante la ruta de trans-sulfuración con la ayuda de la vitamina B6.

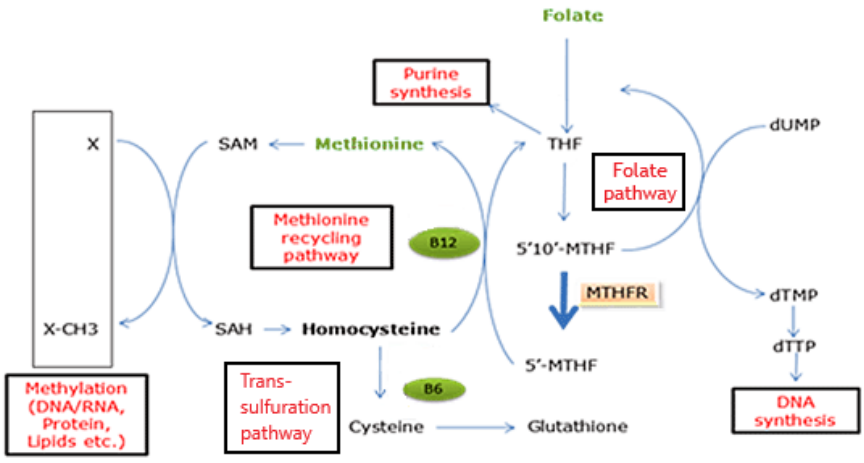


Figura 8. Ciclos de la metionina y del folato. Adaptado y modificado [cuadros complementarios] de: Patiño Vásquez (2014)

3.4 Enfermedades asociadas al gen MTHFR

3.4.1 Hiperhomocisteinemia

La hiperhomocisteinemia es una condición médica caracterizada por un nivel anormalmente elevado de homocisteína en la sangre, mayores de 15 µmol/L. En este sentido, el reporte inicial de Frosst et al (1995) demostró la asociación del genotipo mutante homocigoto (677TT) de

MTHFR con una leve hiperhomocisteinemia. Aun así, según Jacques et al. (1997) se observó que esta asociación estaba presente solo en individuos con niveles bajos de folato. En el estudio de 365 individuos del *Family Heart Study* del Instituto Nacional del Corazón, los Pulmones y la Sangre (NHLBI) de EE. UU., se encontró que el genotipo mutante homocigoto fue unido con niveles más altos de homocisteína plasmáticas. De acuerdo con esto, cuando el grupo fue subdividido con base en el folato plasmático, hubo incrementos más dramáticos en homocisteína plasmática en individuos que fueron mutantes homocigotos y tuvieron valores de folato plasmático debajo de la mediana. No hubo efecto del genotipo en niveles de homocisteína en individuos con valores de folato plasmático mayores de la mediana. Estos hallazgos sugirieron que la suplementación de folato debería ser efectiva en el tratamiento de la hiperhomocisteinemia en individuos con la mutación.

Los reportes (Rozen, 1997) indican que la incapacidad de la enzima MTHFR de catalizar la conversión de 5,10-metilentetrahidrofolato a 5-metiltetrahidrofolato genera un incremento en los niveles de homocisteína plasmática en los sujetos homocigotos mutantes. Según estos reportes estudios, los homocigotos mutantes tienen mayores niveles de homocisteína mientras que los sujetos heterocigotos tienen niveles de homocisteína un poco elevados comparados con los controles normales. Estudios bioquímicos de enzimas codificadas por genes no mutados y mutados han aportado un fundamento lógico para el efecto protectivo del folato en la hiperhomocisteinemia en individuos mutantes.

3.4.2 Homocistinuria debido a deficiencia de la actividad del N(5,10)-metilentetrahidrofolato reductasa

La homocistinuria debido a la deficiencia de la actividad del N(5,10)-metilentetrahidrofolato reductasa es un error congénito del metabolismo debido a un gen autosómico recesivo del metabolismo del folato. La severidad clínica es variable: oscila entre características neurológicas severos y ausencia de síntomas. Las características clínicas incluyen homocistinuria, homocisteinemia, retraso en el desarrollo, discapacidad intelectual severa, muerte perinatal, disturbios psiquiátricos y aparición tardía de trastornos neurodegenerativos.

Datos *in vitro* e *in vivo* sugieren que el polimorfismo 677C→T puede modular actividad enzimática hasta en pacientes con grave deficiencia de MTHFR. Esto ha sido demostrado en estudios *in vitro* donde la presencia del alelo valina del polimorfismo 677C→T (A→V) disminuyó la actividad enzimática de aproximadamente el 50%. Un fenómeno similar ha sido

observado en vivo donde el polimorfismo 677 aparenta contribuir a la termolabilidad del MTHFR en pacientes con graves deficiencias de MTHFR, aunque los reportes iniciales presumieron que la termolabilidad fuera a causa de una mutación deletérea. Aunque la presencia de mutaciones deletéreas conocidas puede ser un buen predictor de actividad enzimática, el efecto de ligeros polimorfismos contribuye a la complejidad del análisis genotipo-fenotipo (Liew & Gupta, 2015).

3.4.3 Anomalías congénitas

3.4.3.1 Defectos del tubo neural sensitivos al folato

Algunos estudios demuestran que la homocigosidad 677C-T confería un mayor riesgo para los defectos del tubo neural (Ou et al., 1996; Christensen et al., 1999). Esta variante influye en la progresión de la condición mediante la elevación de la homocisteína plasmática, pues la homocisteína o uno de sus metabolitos tienen efectos tóxicos en la vasculatura o en el desarrollo del embrión. Los defectos del tubo neural más comunes son espina bífida abierta (mielomeningocele) y anencefalia. Las mujeres con homocisteína plasmática elevada, bajo folato o baja vitamina B12 presentan mayor riesgo de tener un hijo con defectos del tubo neural (O'Leay et al. 2005). Motulsky (1996) citó la evidencia desde Centers for Disease and Control (anónimo, 1992) que el ácido fólico dado antes y durante las primeras 4 semanas de embarazo puede prevenir 50% o más de defectos del tubo neural.

3.4.3.2 Espina bífida

La espina bífida es una condición en que el tubo neural, que es un precursor del cerebro y la médula espinal, falla en cerrarse completamente durante las primeras semanas de desarrollo embrionario. Por lo tanto, cuando se forma la columna, los huesos de la columna dorsal no se cierran del todo alrededor de los nervios en desarrollo de la médula espinal. Parte de ella puede sobresalir a través de una abertura en la columna. Las personas que nacieron con este defecto pueden tener un meningocele, que es un saco lleno de fluido en la espalda cubierto de piel. Si dicho saco contiene parte de la médula espinal y su cubierta protectora, se conoce como mielomeningocele.

Los síntomas incluyen la pérdida del sentido debajo del nivel de la abertura, debilidad o parálisis de los pies o piernas, y problemas con el control de la vejiga y el intestino. Los individuos afectados padecen también de una acumulación de líquido alrededor del cerebro (hidrocefalia)

y problemas de aprendizaje. Muchas personas pueden vivir hasta la edad adulta mediante cirugías y otros tratamientos (*Spina bifida: MedlinePlus Genetics*, s. f.).

3.4.3.3 Anencefalia

La anencefalia es una condición que previene el desarrollo normal del cerebro y de los huesos del cráneo. Esta condición aparece cuando el tubo neural no logra cerrarse durante las primeras semanas del desarrollo embrionario. Como estas anormalidades del sistema nervioso son tan graves, casi todos los bebés con anencefalia mueren antes del nacimiento o pocas horas o días después del nacimiento. Como el tubo neural falla en cerrarse adecuadamente, el cerebro y la médula espinal en desarrollo están expuestas al líquido amniótico que envuelve al feto en el útero. Esta exposición hace que se degenere el tejido del sistema nervioso. En consecuencia, a las personas con anencefalia les faltan el telencéfalo y el cerebelo. Estas regiones del cerebro son necesarias para pensar, oír, ver, sentir emociones y tener movimientos coordinados. Además, los huesos del cráneo faltan o están incompletos (*Anencephaly: MedlinePlus Genetics*, s. f.).

3.4.4 Enfermedades vasculares

3.4.4.1 Accidente cerebrovascular isquémico

Considerando el abordaje de un accidente cerebrovascular, UniProt, s.f. expresa lo siguiente:

> Un accidente cerebrovascular es un evento neurológico agudo que lleva a la muerte del tejido neural del cerebro y resulta en la pérdida de funciones motoras, sensoriales y/o cognitivas. Los accidentes cerebrovasculares isquémicos, que resultan de oclusión vascular, se consideran unas enfermedades altamente complejas que consisten en unos grupos de trastornos heterogéneos con factores genéticos y ambientales múltiples.

3.4.5 Cánceres

Diversos estudios han demostrado que deficiencias de folato pueden causar cáncer. La disminución de la actividad MTHFR resulta en un decremento en metiltetrahidrofolato y un incremento en otras formas del folato, lo cual ha sido demostrado en linfocitos en individuos homocigotos mutantes (TT). La redistribución de folatos afecta la síntesis de timidina o purina, pues se dar la completa transformación de desoxiuridilato monofosfato a desoxitimidilato monofosfato y causar incorporación errónea de uracilos en el ADN, con consecuentes efectos nocivos en la síntesis o reparación de ADN. La desestabilización del ADN lleva a aberraciones cromosómicas y potencialmente a transformaciones malignas (McKinnon y Caldecott, 2007).

El polimorfismo 677C-T también puede contribuir a la formación de un cáncer mediante la disrupción en la síntesis de la metionina o S-adenosilmetionina con consecuentes efectos en la metilación. Una disrupción en la metilación también ocurre a causa de la conversión de la homocisteína a la S-adenosilhomocisteína, un inhibidor de varias metiltransferasas. Los individuos con el genotipo TT tienen una metilación en linfocitos decrementada; este disturbio depende del folato, ya que la metilación del ADN alterado es asociada a los cambios de la expresión genética, una ligera deficiencia de MTHFR resulta en la expresión de protooncogenes o potencial transformación maligna a través de este mecanismo.

3.4.5.1 Cáncer de próstata

El cáncer de próstata es una enfermedad que radica en la formación de células cancerosas en los tejidos de la próstata (Tratamiento del cáncer de próstata (PDQ®)–Versión para pacientes - *NCI*, s. f.). Puede no provocar síntomas en su primer estadío, pero sí en fases más avanzadas, incluyendo problemas para orinar, disminución en la fuerza del flujo de la orina, sangre en la orina, sangre en el semen, dolor de huesos, pérdida de peso involuntaria y disfunción eréctil. Por un lado, se ha encontrado una relación entre C677T y el cáncer de próstata en la población asiática (Küçükhüseyin et al., 2011; Wu et al., 2010; Zhang et al., 2012), así como en la ecuatoriana (López-Cortés et al., 2013) y otros (Safarinejad et al., 2010). Por otro lado, no se ha encontrado tal asociación en los varones de Irán (Fard-Esfahani et al., 2012), ni Algeria (Mouhoub-Terrab et al., 2022). Otro meta análisis afirma que este polimorfismo está ligado a una baja susceptibilidad para el cáncer de próstata, incluso puede tener efectos protectores contra el riesgo de cáncer de próstata (Bai et al., 2009). Un meta análisis sobre la población de India del Norte de Yadav et al. (2021) revela que hay asociación entre este polimorfismo y el cáncer de próstata en dicha población, pero no se encuentra en los otros estudios de poblaciones asiáticas y caucásicas utilizados en el meta análisis, a excepción de dos: Abedinzadeh et al. (2015) para la población asiática, y Chen et al. (2015) para la población del Este Asiático.

3.4.5.2 Cáncer de mama

El cáncer de mama es un tipo de cáncer que se forma en las células de las mamas. Los síntomas incluyen: un engrosamiento en la mama que no se siente igual al tejido que la rodea; una alteración en el tamaño, aspecto o forma de una mama y cambios en la piel sobre la mama. Por ejemplo, la formación de hoyuelos; la inversión reciente del pezón; desprendimiento de la piel, descamación, formación de costras y pelado de la areola o la piel de la mama; enrojecimiento o

foros chiquitos en la piel sobre la mama comparables a la piel de una naranja. El cáncer de mama tiende a comenzar en los conductos para producir leche, llamándose así carcinoma ductal invasivo, pero puede originarse también en los lobulillos, tomando el nombre de carcinoma lobulillar invasivo, o en otras células o tejido mamarios (Cáncer de mama - Síntomas y causas - *Mayo Clinic*, s. f.).

Por un lado, en un meta análisis de Petrone et al. (2021) se describió una asociación entre C677T y 677TT y el cáncer de mama en poblaciones de China (Lu Q et al., 2015), Jordania (Awwad et al., 2015), Irán (Hesari et al., 2019), Italia (Castiglia et al., 2019) y Marruecos (Diakite et al., 2012).

Por otro lado, un meta análisis de He et al. (2017) sostiene que tiende a haber un mayor riesgo a padecer de cáncer de mama en mujeres asiáticas y caucásicas heterocigotas (CT) y homocigotas (TT). Otro grupo evaluó la asociación del polimorfismo C677T con el cáncer de mama en las latinoamericanas y encontró un elevado riesgo para modelos genéticos alélicos mutantes recesivos (Meneses-Sánchez et al., 2019).

CAPÍTULO III: MATERIALES Y MÉTODOS

1. Obtención de muestras

Para los experimentos del proyecto utilizamos DNA genómico amerindio (Ngöbe) del banco de DNA del Departamento de Genética y Biología Molecular de la Universidad de Panamá. Estos DNAs fueron colectados hace más de 20 años en estudios liderados por el Dr. Tomás Arias y la Dra. Lucía Jorge, a partir de los cuales obtuvieron varias publicaciones (Arias et al., 1993; Jorge-Nebert et al., 2002; Petersen et al., 1991). Desafortunadamente, la información referente a estas muestras se perdió con los años, incluyendo el sexo de estas. Por lo tanto, con el objetivo de mantener un balance en la cantidad de genomas de hombre y mujer analizados fue necesario determinar el sexo de las muestras mediante PCR, además de verificar la concentración y calidad las mismas para escoger las de mejor calidad y descartar las que evidenciaban degradación. Basados en los análisis de control de calidad del DNA y el sexado de las muestras (que explicaremos a continuación) seleccionamos para la secuenciación genómica dos (2) muestras de DNA de hombres y dos (2) mujeres. Los análisis de secuenciación genómica nos permitieron identificar algunos polimorfismos SNPs candidatos posiblemente asociados a riesgo de ENT. Así, decidimos enfocarnos en un SNP del gen MTHFR para análisis de genotipaje poblacional mediante PCR y secuenciación Sanger. Para los análisis de genotipaje de alelos de SNPs utilizamos 50 muestras de DNA de la mejor calidad: mitad de mujeres y mitad de hombres de la población amerindia Ngöbe.

2. Análisis preliminar de ADN

2.1 Cuantificación y Calidad de ADN genómico

La concentración del ADN se determinó mediante cuantificación en NanoDrop (ThermoFisher), mientras que la calidad de este fue determinada mediante electroforesis de agarosa. Con el NanoDrop, se realizó la medición espectrofotométrica de la concentración del DNA y del nivel de contaminación por proteínas (260/280) y por otros contaminantes (260/230) de 31 muestras vertiendo 1 µL de blanco (amortiguador TE) y 1 µL de cada muestra. Para determinar que una muestra era de buena concentración para secuenciación genómica seleccionamos aquellas con valores iguales o mayores de 100ng/uL y una razón de absorbancia 260/280 entre 1.8 y 2.0. Las concentraciones obtenidas se muestran en la **Tabla 7**. Luego, se verificó la calidad del DNA mediante electroforesis en geles de agarosa al 1% en amortiguador TAE. La calidad de las muestras fue establecida con base en el peso molecular (banda de alrededor de 20Kb), en las

que no se observó degradación total o parcial, (también conocido como “smear”) fueron seleccionadas como muestras de buena calidad. Se utilizó como marcador de peso molecular un estándar de peso molecular (100 bp *Ladder*). Las muestras analizadas en gel de agarosa se muestran en la Figura 14. Los parámetros de concentración y calidad del DNA fueron los que requiere la compañía que realizó el servicio de secuenciación genómica.

2.2 Sexado de las muestras de ADN mediante PCR del *Cromosoma Y*

Las muestras de mejor calidad fueron escogidas para el sexado que se hizo mediante la amplificación por PCR de una región localizada en el cromosoma Y. Para identificar el linaje masculino de las muestras se utilizaron dos cebadores de oligonucleótidos: Y1 y Y2 que flanquean un fragmento de 170 pb de las repeticiones alfoides del *cromosoma Y* humano (YI: ATG ATA GAA CGG AAA TAT G; Y2: AGT AGA ATG CAA AGG GCT CC, (Witt & Erickson, 1989; Wolfe et al., 1985). Los oligonucleótidos (cebadores) fueron sintetizados por la compañía IDTDNA *Technologies* y purificados mediante precipitación estándar (*standard desalting*). La mezcla de reacción en la amplificación del *cromosoma Y* consistió en 15 µL de Mastermix (2X Blastaq PCR Taq), 3ul de cada primer (5 µM cada uno), 1 µL de ADN y 8 µL de agua libre de nucleasas en un volumen final de 30 µL. La PCR se realizó en un termociclador *Applied Biosystems* 2720 y las condiciones utilizadas para amplificar el *cromosoma Y* fueron 94°C durante 8 min, 30 ciclos a 94°C por 1 min, 55°C por 1 min, una extensión final de 72°C por 2 min, y cuando terminó estuvo a 12° por un tiempo indefinido. Los productos de amplificación de 170 pb fueron resueltos en electroforesis de agarosa al 1% en tampón (buffer/amortiguador) TAE 1X y teñidos con GelRed. Como control utilizamos un estándar de peso molecular (100bp *ladder*).

Utilizamos como control de amplificación el gen nuclear (PPARγ2). Los cebadores utilizados fueron forward 5'- AAG GAA TCG CTT TCC G-3' y un cebador reverse 5'- GCC AAT TCA AGC CCA GTC -3' reportados previamente (Priya et al., 2016). La mezcla de reacción en la amplificación del gen control consistió en 15 µL de Mastermix (2X Blastaq PCR Taq), 2 µL de ambos primers, 1 µL de ADN y 10 µL de agua libre de nucleasas en un volumen final de 30 µL. Las condiciones empleadas en el termociclador (Applied Biosystems 2720) para la amplificación del gen control fueron 94° durante 8 min, 35 ciclos a 94°C por 50s, 50°C por 50s, 72°C por 1 min, una extensión final a 72°C por 5min y al final 12°C por un tiempo indefinido. El producto de 270 pb fue resuelto por electroforesis en agarosa al 1% teñida con GelRed en

amortiguador TAE 1X. Se utilizó como marcador de peso molecular la 100 bp *ladde*r. La electroforesis en gel se realizó inicialmente a 60V los primeros 10-15min y luego se subió a 90-100V por 40-50 min. Los resultados de la amplificación fueron documentados con un Gel Doc™ EZ *System* (BioRad).

3.Secuenciación NGS en Illumina

Una vez se determinó el sexo de las muestras, se escogieron dos (2) muestras sexadas como hombres y dos (2) sexadas como mujer. Las muestras fueron preparadas para envío a Estados Unidos para los servicios de secuenciación del genoma completo y análisis bioinformático preliminar. El proceso de secuenciación de nueva generación (NGS) fue llevado a cabo por la compañía NOVOGENE, utilizando los *kits* de la empresa Illumina, fabricante de secuenciadores NGS.

3.1 Flujo de trabajo de la secuenciación de Illumina

Las etapas en el proceso de secuenciación Illumina en NOVOGENE constan de cuatro pasos: preparación de la muestra, preparación de la biblioteca (incluyendo generación de racimos (*clusters*), secuenciación y análisis de los datos (**Figura 9**).

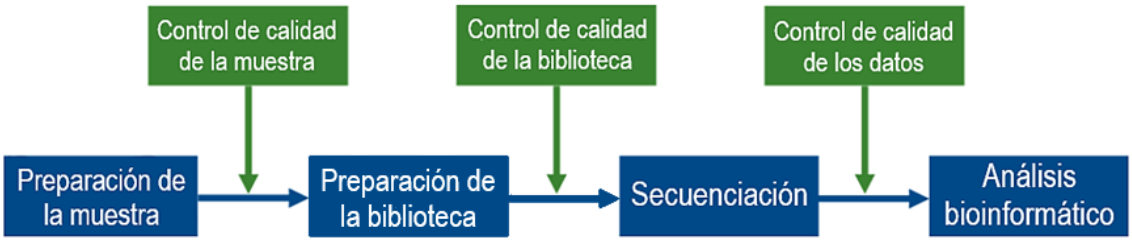


Figura 9. Flujo de trabajo de la secuenciación de Illumina por Novogene.

3.2 Rotulación de la muestra

Cada muestra fue rotulada con un código ciego de la misma sin ninguna información que identifique al donante, excepto que es humano y su sexo. Tampoco se revela que la muestra de ADN es de origen amerindio (Ngöbe).

3.2.1 Control de calidad de las muestras

La calidad de la muestra de ADN se evalúa mediante dos métodos. Uno de los métodos es la cuantificación de ADN a partir de la intensidad de fluorescencia de un tinte fluorescente que se une al ADN de cadena doble utilizando un fluorómetro Qubit. El otro método utiliza un

sistema automatizado de electroforesis “TapeStation” que integra en equipo un programa de procesamiento de la data, reactivos y dispositivos de cinta de pantalla específicos para DNA y RNA. Este equipo determina el tamaño, la cantidad y calidad de la muestra.

3.3 Preparación de la biblioteca

Brevemente, el ADN es fragmentado por endonucleasas a fin de generar extremos pegajosos compatibles con oligonucleótidos que son unidos con ADN ligasa a los extremos pegajosos generados en la muestra de ADN. Los fragmentos generados unidos a los oligonucleótidos son purificados utilizando columnas. Los fragmentos purificados constituyen la biblioteca genómica, la cual es usada en los *chips* de secuenciación de Illumina. En corto tiempo, las librerías generadas con un rango de hasta 6 muestras fueron desnaturalizadas a cadena sencilla de ADN e hibridadas con sondas de 95 bases de longitud marcadas con biotina. Luego, las muestras son enriquecidas con *beads* magnéticos de estreptavidina y precipitados mediante un magneto. Estos fragmentos enriquecidos son luego eluidos de los *beads* e hibridados en una segunda reacción de enriquecimiento. Los fragmentos son entonces amplificados *in situ* con *chips* de secuenciación produciendo así secuencias listas para exportación y análisis. Los fragmentos secuenciados tienen una longitud de entre 200 y 400 pares de bases (**Figura 10**) (**Figura 11**).

Figura 10. Flujo de trabajo para la construcción de la librería. El ADN genómico fue cortado al azar en fragmentos. A los fragmentos obtenidos se le emparejaron los extremos, añadió la cola poli-A y se ligaron posteriormente con un adaptador de Illumina. Estos fragmentos con adaptadores fueron amplificados por PCR, seleccionados por tamaño y purificados. Adaptado de Novogene (2022).

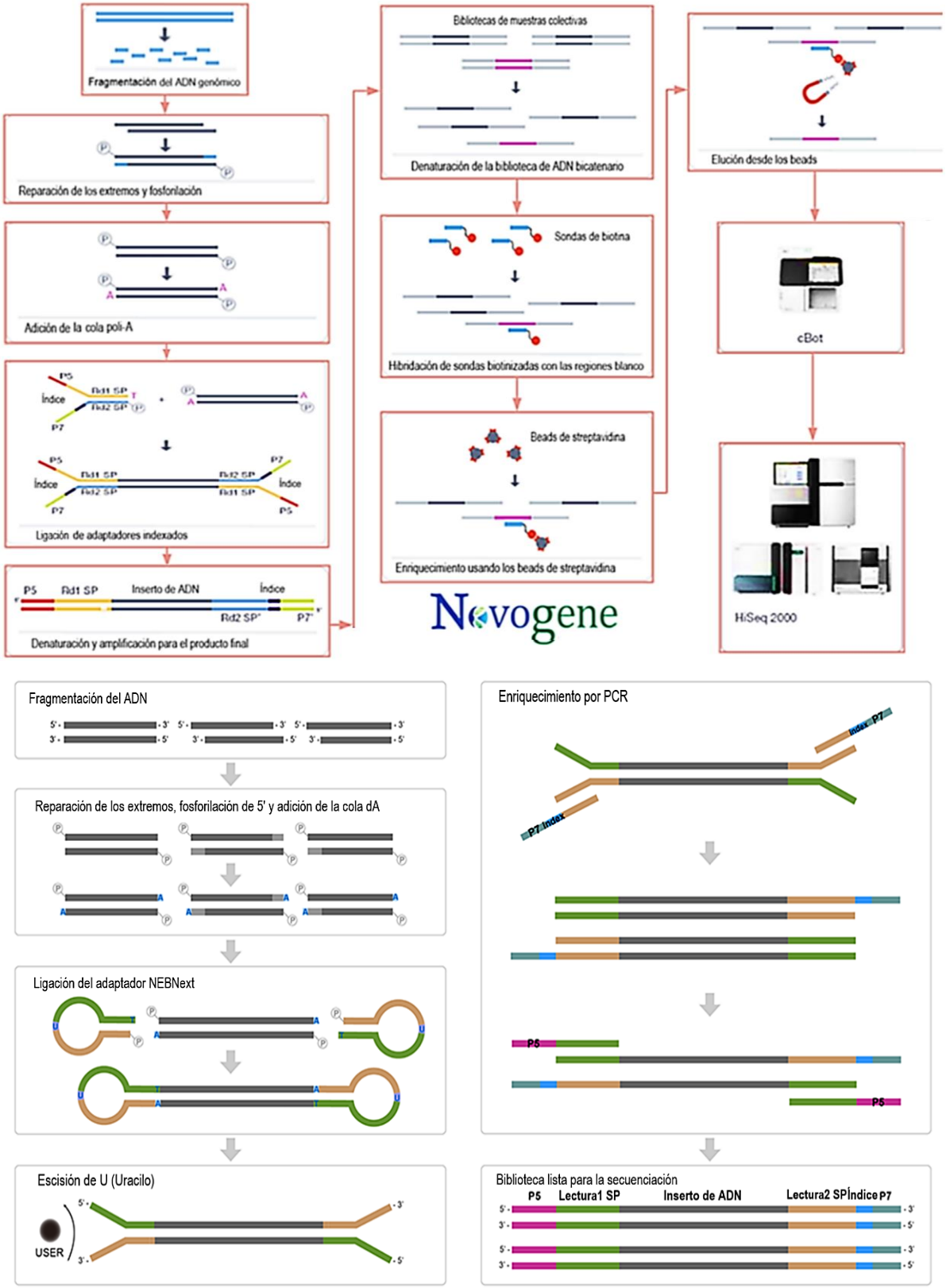


Figura 11. Flujo de trabajo de construcción de la biblioteca detallado. En relación con los resultados de: New England BioLabs Inc., (s. f.); Novogene, (2022).

3.4 Generación de racimos (*clusters*)

Los racimos (clusters) son un grupo de hebras de ADN agrupadas juntas que son sintetizadas y secuenciadas *in situ*. Cada *cluster* representa miles de copias de la misma hebra de ADN en un punto de 1-2 micrómetros de la celda de flujo (**Figura 12**). El proceso de generación de clústeres toma las cadenas sencillas de la biblioteca de ADN y produce una amplificación de alta fidelidad para producir por amplificación clonal, donde hay miles de copias que permiten amplificar señales para el subsiguiente proceso de secuenciación por síntesis (Westenberger, 2020).

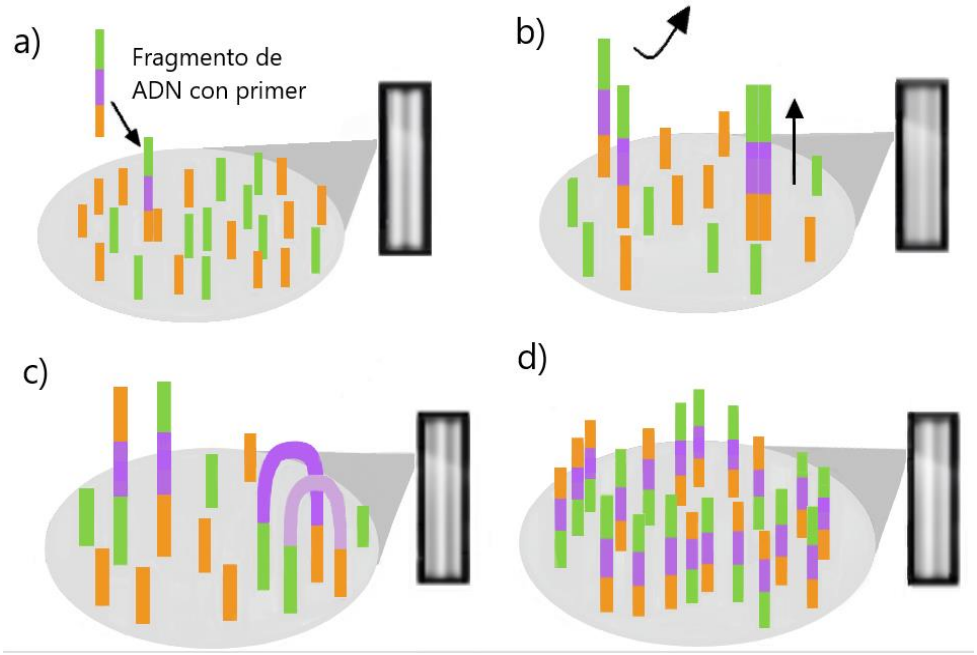


Figura 12. Generación de racimos. a) El fragmento de ADN se ancla a la celda de flujo por medio de la unión de sus adaptadores a los oligonucleótidos complementarios. b) La polimerasa genera una hebra reversa complementaria y la hebra original es retirada. c) La hebra reversa se pliega y queda en forma de puente, donde la polimerasa genera una hebra idéntica a la original. d) El proceso se repite masivamente (Adaptado y modificado [cambios de perspectivas para el complementar la generación de racimos] de Rubio et al., 2020).

La generación de racimos ocurre en las celdas de flujo, que son platinas de vidrio espeso con carriles, dentro los cuales hay canales fluidos y la superficie de la platina de vidrio está cubierta con una matriz en donde están incrustados los oligonucleótidos complementarios a los adaptadores de bibliotecas. Estas moléculas moldeadas de la biblioteca de ADN monocatenario se ligan al azar a unos oligonucleótidos de captura adaptadores sobre la superficie de la platina de vidrio. Luego, una polimerasa de alta fidelidad realiza una extensión 3’ para hacer una copia exacta de la hebra molde inicial, que después se une de forma covalente sobre la superficie de

la platina de vidrio. Una vez terminada la síntesis de la cadena complementaria, se realiza una denaturación y se remueve la cadena molde original, dejando solo una copia monocatenaria de la molécula de la biblioteca original. La molécula de hebra sencilla se pliega y forma un puente hibridizando con un cebador complementario adyacente. Después, el cebador hibridado se extiende mediante la polimerasa, formando así un puente bicatenario, y luego este último es denaturado, lo cual resulta en dos copias de moldes monocatenarios unidos por enlace covalente. Todo este proceso se repite mediante varias hibridaciones, extensiones y denaturaciones hasta la formación de miles de clusters idénticos a las moléculas de la librería de ADN de cadena sencilla (Westenberger, 2020).

3.5 Secuenciación

Una vez que se obtienen clusters amplificados clonalmente, para proceder con los siguientes pasos de secuenciación por síntesis, se realiza un proceso de linearización. Este consiste en cortar las hebras inversas (*reverse*), eliminarlas mediante lavado y dejar solamente clústeres con hebras *forward* (Westenberger, 2020). Luego, unos nucleótidos con etiquetas fluorescentes específicas para todos los tipos de nucleótido son puestos en la placa. Esos nucleótidos presentan una modificación química llamada terminadores reversibles, la cual evita la unión de múltiples nucleótidos marcados en cada sitio de reacción de modo que se localiza el que corresponde a cada punto en la secuencia, disminuyendo el riesgo de errores en la secuenciación. La identificación de las bases se da por la fluorescencia específica que emite cada vez al incorporarse. Se remueve la etiqueta antes de la colocación del siguiente nucleótido con el fin de evitar la emisión de la señal por dos bases a la vez. Cuando se termina la primera lectura, el fragmento resultante es retirado. Este paso se repite contemporáneamente con cada hebra del mismo clúster de forma paralela hasta la finalización de la secuenciación (Rubio et al., 2020).

4. Análisis bioinformático

El genoma humano comprende alrededor de 3.2 billones de bases, de las cuales casi el 1.2% corresponde al exoma y representa aproximadamente 180,000 exones y 30 millones de bases (Maróti et al., 2018). Los análisis preliminares fueron realizados por la misma empresa que secuenció los genomas, la empresa NOVOGENE con programas de análisis y plataformas de súper computadoras. Aunque secuenciamos el genoma completo de las muestras para los análisis, nos enfocamos solamente en el exoma porque se estima que aproximadamente. El 80-90% de las enfermedades conocidas son causadas a nivel de exones. La secuenciación y

análisis del exoma identifica variantes a través de un amplio rango de aplicaciones alcanza coberturas comprensivas de regiones codificantes y produce un *set* de datos más pequeños y manejables para un análisis de datos más veloz y fácil respecto a enfoques de genoma completo (*Whole Exome Sequencing | Detect exonic variants*, s. f.).

4.1 Secuencia ordenada de los análisis bioinformáticos

El análisis bioinformático incluye los siguientes puntos:

1) Control de calidad de los datos: filtración de las lecturas que contienen secuencias de los adaptadores o de baja calidad
2) Alineamiento con el genoma humano de referencia, estadísticas de la profundidad y cobertura de secuenciación
3) Llamada de SNP/InDel/SV/CNV, anotación y estadísticas
4) Llamada de SNP/InDdel/SV/CNV somático anotación y estadísticas (muestras emparejadas, p.ej.: tumor-normal)

El siguiente diagrama representa el flujograma de datos de control de calidad y análisis bioinformático que Novogene usó (**Figura 13**). Los análisis somáticos se realizaban solo cuando se proporcionaban las muestras emparejadas tumor normal.

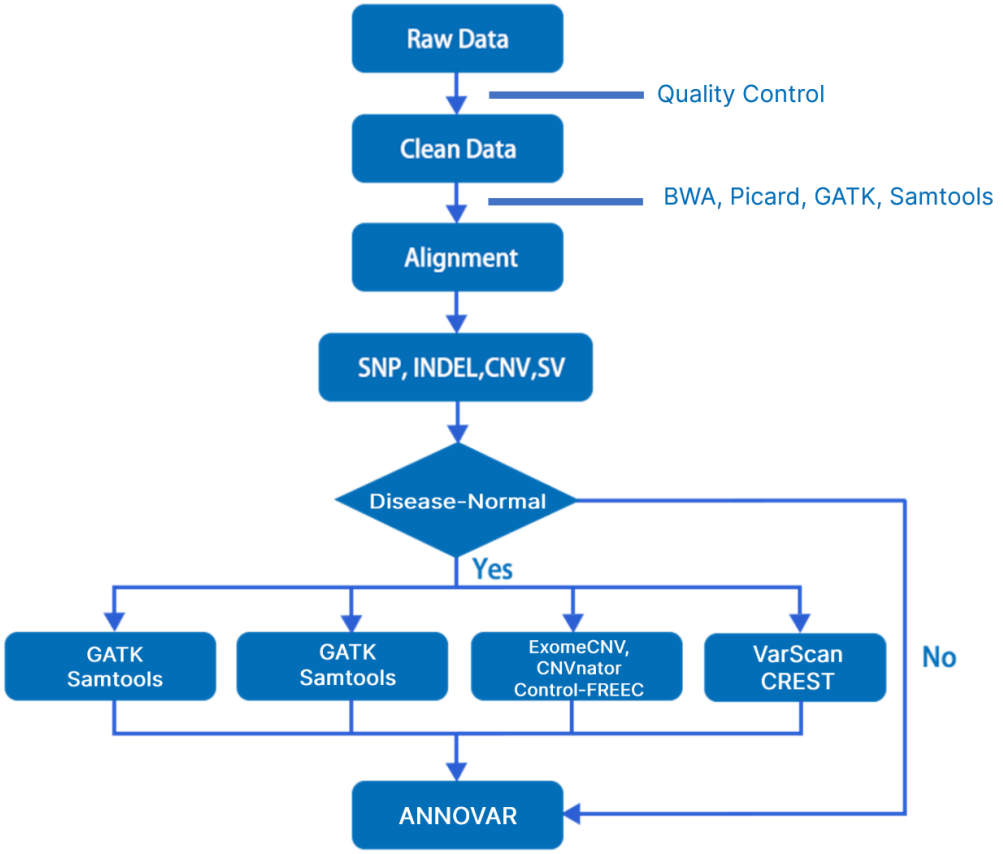


Figura 13. Flujograma donde se resumen los pasos de los análisis bioinformáticos. Los análisis de estos datos requieren varios pasos. En primer lugar, se exportaron los datos en formato FASTA para comprobar la calidad de las secuencias y descartar fragmentos de secuencias muy cortas con un umbral de corte mínimo de 150 bases usando FASTQ.

4.1.1 Datos crudos

Los datos originales de imágenes crudas, obtenidos a partir de plataformas de secuenciación de alto rendimiento (por ejemplo, la plataforma Illumina), se transforman en lecturas secuenciadas llamadas bases. Las lecturas secuenciadas se consideran datos en crudo o lecturas sin procesar, que se registran en un archivo FASTQ (fq) que contiene información sobre la secuencia (lecturas) y la correspondiente información sobre la calidad de la secuenciación.

Cada lectura en formato FASTQ se almacena en cuatro líneas, como sigue en este ejemplo:

@EAS139:136:FC706VJ:2:2104:15343:197393 1:Y:18: ATCACG

GCTCTTTGCCCTTCTCGTCGAAAATTGTCTCCTCATTCGAAACTTCTCTGT

+ (esta línea empieza con el carácter + y sigue opcionalmente con el identificador de secuencia)

@@CFFFDEHHHHFIJJJ@FHGIIIEHIIJBHHHIJJEGIIJJIGHCCF

La línea 1, que comienza con un carácter "@", va seguida de un identificador (ID) de la secuencia y de una descripción opcional (como una línea de título FASTA). La línea 2 contiene las lecturas de la secuencia en bruto. La línea 3 comienza con un carácter "+" y va seguida opcionalmente por el mismo identificador de secuencia (y cualquier descripción) nuevamente. La línea 4 codifica los valores de calidad para la secuencia de la línea 2, y contiene el mismo número de caracteres como bases en la secuencia.

El valor ASCII de cada carácter en la cuarta línea menos 33 es el valor de calidad de la base de secuenciación correspondiente en la segunda línea. Si la tasa de error de secuenciación se registra mediante "e" y la calidad de la base para la plataforma Illumina se expresa como Qphred, se obtendría la ecuación n.º 1 a continuación:

Ecuación 1: Qphred = -10log10(e)

4.1.2 Control de calidad de la secuenciación genómica

Los cuatro genomas secuenciados fueron examinados preliminarmente para verificar la calidad de secuenciación. Para ello, se realizó el control de calidad descartando las lecturas de extremos emparejados; si una de ellas contenía contaminación del adaptador, si más de 10 bases eran inciertas en cualquier lectura o si la proporción de baja calidad era mayor de 50 en cualquier lectura. Así, se determinó cuántas lecturas se secuenciaron y cuántas de ellas eran limpias; si salieron secuencias no específicas (“N” o con alguna otra ambigüedad); cuales lecturas tenían baja calidad; y cuáles secuencias eran enlazadas a los adaptadores, que son secuencias no específicas que causan ruido.

Se usaron parámetros estadísticos como el porcentaje de tasa de error y el porcentaje de contenido en GC. Factores como la plataforma de secuenciación, reactantes químicos y calidad de las muestras afectan a la tasa de errores de secuenciación y la calidad de las bases. Por cuanto respecta el porcentaje de contenido en GC, este tipo de evaluación sirve para controlar el potencial de la separación de AT/GC, sesgo de secuenciación y errores en la preparación de librerías.

Para garantizar el análisis posterior, se requiere que la calidad de la mayoría de las bases sea superior a Q20, que indica una tasa de error de secuenciación de 1/100; es decir, la probabilidad

de que la base llamada sea errada es de 1 cada 100 pb dentro de la lectura secuenciada, así como una tasa de secuenciación correcta del 99%; o sea, la probabilidad de que la base llamada sea la correcta es del 99%. La disminución de la calidad de las bases a lo largo de las lecturas es habitual, lo que constituye una circunstancia inherente a la secuenciación de nueva generación.

De acuerdo con la característica de secuenciación de la plataforma Illumina, para los datos de extremos emparejados, requerimos que el porcentaje medio de Q30 sea superior al 80%, y que la tasa de error sea inferior al 0,1%.

4.2 Alineamiento de Secuencias con el Genoma Humano de Referencia

Se mapearon las lecturas limpias de extremos emparejados y se alinearon con el genoma humano de referencia con *Burrows-Wheeler Aligner* (BWA). Los resultados del mapeo original se obtuvieron con el formato BAM. Los archivos BAM se ordenaron con SAMtools y las lecturas duplicadas se marcaron con Picard. Luego, se obtuvieron los archivos BAM finales y después se calcularon la cobertura y la profundidad de cobertura con base en BAM. Cuando se habla de cobertura, se refiere al porcentaje de bases del genoma de referencia que se secuencia al menos una vez en determinada cantidad de veces. En cambio, la profundidad es el número promedio de veces que cada base en el genoma es secuenciada en los fragmentos de ADN (Rubio et al., 2020).

Para cada uno de los cuatro genomas, fueron calculadas las estadísticas de las regiones mapeadas, de la profundidad y de la cobertura. La cobertura fue calculada para cuatro veces, diez veces y veinte veces. En esa misma línea, se calculó la profundidad para cada cromosoma.

4.3 Detección Preliminar de Mutaciones y Polimorfismos en la Línea Germinal

En esta etapa, se abordó la principal pregunta de esta investigación: ¿cuáles son las variantes y polimorfismos genéticos posiblemente asociadas a ENT en el panameño?, específicamente en los genomas de la población ancestral amerindia Ngöbe. Los análisis genómicos identificaron múltiples variantes/polimorfismos de genes candidatos posiblemente asociados a ENT. El reconocimiento de estos genes se realizó mediante el uso de varias estrategias (explicadas más adelante) para identificar marcadores moleculares que incluyen *SVs*, *CNVs*, *INDELs*. Finalmente, nos enfocamos de manera primordial en los *SNPs*. A continuación, explicaremos una descripción breve de la detección de cada uno de los marcadores mencionados (Novogene, 2022).

4.3.1 Variaciones Estructurales (SV)

Los análisis calcularon el número de los distintos tipos de SV, incluyendo duplicaciones, inversiones, translocaciones, deleciones e inserciones, así como el tamaño de ganancias y pérdidas (Novogene, 2022).

4.3.2 Inserciones/Deleciones (InDels)

Los análisis establecieron múltiples InDels candidatos, y si los mismos causaban algún cambio en el marco de lectura, así como si hubo ganancia o pérdida de codones de parada, o simplemente en qué regiones se dieron los cambios en las características generales de los InDels, en particular: su heterocigosidad o su homocigocidad (ver en el capítulo de resultados) (Novogene, 2022).

4.3.3 Polimorfismos de Nucleótidos Simples (SNPs)

Los análisis clasificaron y calcularon el número total y de cada tipo de SNPs, entre sinónimas, con cambio de sentido, con ganancia de codón de parada, con pérdida de codón de parada y desconocidas, y determinó en qué regiones están (ver en el capítulo de resultados) (Novogene, 2022).

4.3.4 Variación en Número de Copias (CNV)

Los análisis calcularon la cantidad total de CNVs, la cantidad que corresponde a ganancia o pérdida en términos de regiones analizadas o términos de tamaño o en número de regiones analizadas. También determinaron las regiones genómicas afectadas por los CNVs en cada muestra (ver en el capítulo de resultados) (Novogene, 2022).

4.4 Identificación y Anotación de las Variantes Candidatas

Los análisis preliminares resultaron en la identificación y anotación de enormes listas en MS *Excel* con miles de variantes génicas candidatas posiblemente asociadas a ENT. Para cada marcador molecular (CNV, INDEL, SV, SNP), se obtuvieron tablas de *Excel* con miles de variantes génicas candidatas preliminares que resultan del análisis simultáneo con miles de datos clínicos donde se han reportado estas variantes asociadas a ENT. Inicialmente no se sabía cuál de los cuatro marcadores resultarían en la identificación de variantes fusionadas a enfermedades, por lo que decidimos estudiarlas todas basados en varios criterios (explicados más adelante) que nos permitieron identificar unos pocos marcadores candidatos con mayor probabilidad y enfocarnos en SNPs.

Luego de la detección y anotación de las variantes genéticas candidatas, se realizaron las anotaciones de variantes con la herramienta ANNOVA (Wang et al., 2010) en muchos aspectos, incluyendo cambios en la codificación de proteínas, regiones genómicas afectadas por las variantes, frecuencia alélica, predicción de nocividad, etc. Las variantes genéticas de interés biomédico fueron identificadas usando GATK *HaplotypeCaller* para single nucleotide *polymorphysms* (SNPs)-polimorfismos de base sencilla e insertions/deletions (INDELs)-inserciones/deleciones. Las regiones de interés se enfatizaron creando intervalos utilizando Bedtools y la profundidad de cobertura de secuenciación de los intervalos fue determinada utilizando GATK DepthOfCoverage (profundidad de cobertura). Con las enormes tablas generadas para cada marcador molecular, realizamos los análisis más exhaustivos de miles de genes candidatos utilizando estrategias de *big data* con filtros especiales que refinaron la identificación precisa de los genes candidatos implicados en ENT. Esto fue mediante una búsqueda jerárquica de varias selecciones de genes utilizando diferentes criterios recomendados por la literatura para el tipo de análisis en varias etapas. En particular, implementamos los parámetros recomendados por las Guías y Estándares de la Asociación Médica Americana de Genética y Genómica Médica para el análisis e interpretación de variantes moleculares (Li et al., 2017; Richards et al., 2015) y usando criterios de bases de datos clínicas que recomiendan estas guías. Primero que todo, filtramos intrones y solo tomamos en cuenta exones, ya que este campo indica los cambios de aminoácidos como resultado de la variante exónica. Solo las variantes exónicas tienen información en este campo, es decir, cuando ‘Func’ (función) es ‘exónica’; este campo indica regiones transcritas a RNA (**Tabla 6**).

Tabla 6. Tabla de *Excel* con los criterios principales para elegir variantes candidatas

GeneNa	Func	Gene	ExonicFunc	AAChange	CLNDN	CLNSIG
MTHFR	exonic	NM_001330358,NM_005	missense SNV	MTHFR:NM_001330358:exon5:c.C788T:p.A263V,I	Neoplasm_of_stomach|Gastrointestin	drug_response
SDC3	exonic	NM_014654	missense SNV	SDC3:NM_014654:exon4:c.C986T:p.T329I	Obesity,_association_with	association
FAAH	exonic	NM_001441	missense SNV	FAAH:NM_001441:exon3:c.C385A:p.P129T	Polysubstance_abuse,_susceptibility_t	risk_factor
IL6R	exonic	NM_000565	missense SNV	IL6R:NM_000565:exon9:c.A1073C:p.D358A	Interleukin_6,_serum_level_of,_quant	association
FCGR2A	exonic	NM_001136219,NM_021	missense SNV	FCGR2A:NM_001136219:exon4:c.A500G:p.H167R	Lupus_nephritis,_susceptibility_to|Pse	drug_response
EPHX1	exonic	NM_000120,NM_00113(	missense SNV	EPHX1:NM_000120:exon3:c.T337C:p.Y113H,EPHX	EPOXIDE_HYDROLASE_1_POLYMORPI	drug_response
GCKR	exonic	NM_001486	missense SNV	GCKR:NM_001486:exon15:c.T1337C:p.L446P	Fasting_plasma_glucose_level_quantit	association
SLC11A1	exonic	NM_000578	missense SNV	SLC11A1:NM_000578:exon15:c.G1627A:p.D543N	Buruli_ulcer,_susceptibility_to	risk_factor
GHRL	exonic	NM_001134941,NM_001	missense SNV	GHRL:NM_001134946:exon2:c.A116T:p.Q39L,GH	Obesity	risk_factor
XPC	exonic	NM_004628	missense SNV	XPC:NM_004628:exon16:c.C2815A:p.Q939K	Xeroderma_pigmentosum|not_specifi	drug_response
ADD1	exonic	NM_001119,NM_00128(	missense SNV	ADD1:NM_001119:exon10:c.G1378T:p.G460W,AI	Hypertension,_salt-sensitive_essential	drug_response
TLR1	exonic	NM_003263	missense SNV	TLR1:NM_003263:exon4:c.A743G:p.N248S	Leprosy_5	risk_factor
UGT2B15	exonic	NM_001076	missense SNV	UGT2B15:NM_001076:exon1:c.T253G:p.Y85D	oxazepam_response_-_Other|lorazep	drug_response
ADRB2	exonic	NM_000024	missense SNV	ADRB2:NM_000024:exon1:c.G46A:p.G16R	salbutamol_response_-_Efficacy|salm	drug_response
PPARGC1B	exonic	NM_001172698,NM_001	missense SNV	PPARGC1B:NM_001172698:exon4:c.G490C:p.A16	Obesity,_variation_in	association
OR2J3	exonic	NM_001005216	missense SNV	OR2J3:NM_001005216:exon1:c.A337G:p.T113A	C3HEX,_ability_to_smell	Affects
OR2J3	exonic	NM_001005216	missense SNV	OR2J3:NM_001005216:exon1:c.G677A:p.R226Q	C3HEX,_ability_to_smell	Affects
PLA2G7	exonic	NM_001168357,NM_005	missense SNV	PLA2G7:NM_001168357:exon11:c.T1136C:p.V379	Asthma_and_atopy,_susceptibility_to	risk_factor

En la columna Func (función), se seleccionó *exonic*, pues es en los exones donde se encuentran los cambios de aminoácidos, que están evidenciados en la columna AAChange. Sucesivamente, se hizo un filtro en la columna CLNSIG para que resultaran solo las variantes con *affects*, *Affects_association*, *association*, *association_risk factor*, *drug response, pathogenic, pathogenic protective y risk factor*.

Esta información generada en las tablas fue contrastada con otros criterios de las bases de datos clínicas. Las principales bases de datos usadas fueron las siguientes:

1. Las bases de datos RefSeq y Gencode (páginas *webs* en los anexos) se usaron para encontrar las regiones genómicas afectadas por las variantes y cambios posibles en la proteína.
2. Se anotaron las características de las regiones genómicas afectadas por las variantes, como la banda, ARN pequeño, sitos diana regulatorios de microRNA mamífero conservados, regiones conservadas de vertebrados, sitios de unión de factor de transcripción, repeticiones, etc.
3. La predicción de la nocividad de mutaciones se realizó con los puntajes de SIFT, PolyPhen, MutationAssessor, LRT y CADD. Se usaron los puntajes de GERP++ para acceder a las conservaciones de mutaciones.
4. A fin de encontrar las frecuencias alélicas alternativas en poblaciones que se reportaron, se usaron bases de datos establecidas, como 1000 Human Genome, Exome Aggregation Consortoium (ExAC), Genome Aggregation Database (gnomAD) y exome sequencing Project (ESP).

5. Se usaron las bases de datos dbSNP, COSMIC (*Catalogue Of Somatic Mutations In Cancer*), OMIM (*Online Mendelian Inheritance In Man*), GWAS (*Genome Wide Association Studies*) Catalog y HGMD (*Human Gene Mutation Database*) para encontrar información reportada sobre la variante, como los mayores SNPs en GWAS y asociaciones a cáncer o enfermedad.
6. Se proporcionó información adicional sobre anotación funcional o de las vías aplicando las bases de datos incluyendo *Gene Ontology*, KEGG, Reactome, Biocarta y PID.
   Una descripción más detallada sobre la anotación de las variantes está en los anexos.

4.5 Selección de Variantes Génicas Candidatas Asociadas a ENT

Los criterios descritos arriba fueron analizados en el conjunto de los datos de las diferentes variantes/polimorfismos, identificados preliminarmente en las tablas de MS Excel para cada uno de los cuatro marcadores moleculares y mencionados con el objetivo de generar una lista más reducida con los genes/variantes de mayor significancia clínica y probabilidad de tener un rol en ENTs. Un criterio adicional que utilizamos es que la variante candidata estuviese presente en al menos 3 de los 4 genomas secuenciados. Esto nos indicaría una posible mayor probabilidad de que la misma estuviese en alta frecuencia en la población ancestral Ngäbe y, por lo tanto, la misma mostraría una frecuencia relativamente alta en el resto de la población mestiza, contribuyendo así de manera más significativa a ENT.

En primer lugar, para cada tipo de marcador molecular, buscamos sus exones en la columna Func. En segundo lugar, buscamos los cambios en aminoácidos que causaban en la columna AAChange, pues consideramos que eso indica que la cantidad o la función de la proteína podría verse afectada. Estos criterios de análisis iniciales nos permitieron determinar que los marcadores CNV y SV no eran muy informativos en términos de significancia clínica ni en variantes génicas posiblemente relacionadas a ENT, por lo tanto, decidimos descartarlos de los siguientes análisis y enfocarnos sólo en INDELs y SNPs.

Para los análisis de INDELs y SNPs filtramos los datos de la columna CLNSIG, que indicaba la significancia clínica de cada variante. Centramos nuestra atención en todas las variantes que tenían los valores de *Affects, Affects_association, association, association_risk factor, drug response, pathogenic, pathogenic protective y risk factor*, que son términos y criterios que se

usan en las guías para describir variantes identificadas en genes que causan trastornos mendelianos (Richards et al., 2015).

Revisamos todos los genes que estaban asociados a cánceres, diabetes y enfermedades cardiovasculares en la columna CLNDN. Para cada gen/variante polimórfica de esas categorías, nos enfocamos en los datos de las columnas como SIFT, Polyphen2_HDIV, Polyphen2_HVAR, LRT, MutationTaster, MutationAssessor y FATHMM para identificar cuáles y cuántas variantes funcionales o deletéreas resultaban conectadas a ENT. También investigamos esas variantes polimórficas en distintas bases de datos y páginas web, que comprendían dbSNP, ClinVar, OMIM, UniProt y Varsome para verificar si la información sobre su patogenicidad estaba actualizada.

Realizamos una selección de 28 genes candidatos entre SNPs e INDELs (**Tabla 16**).

Se descartaron los genes/variantes polimórficas que se repetían en al menos de tres de los genomas secuenciados. Luego, se investigaron los otros genes en distintos artículos científicos para ver cuáles estaban más asociados a enfermedades crónicas y al final se escogieron los genes mostrados en la **Tabla 17**.

Los análisis bioinformáticos realizados, las comparaciones con múltiples bases de datos, así como parámetros explicados arriba, resultaron en un listado de variantes/polimorfismos génicos candidatos que según estos criterios mostraron mayor probabilidad de ser asociados a diferentes ENT. Entre estas variantes nos enfocamos en el polimorfismo rs1801133 del gen MTHFR, enlazados a diferentes tipos de cánceres.

4.6 Identificación del SNP rs1801133 en el gen MTHFR como polimorfismo posiblemente asociado a cánceres en la población panameña

Los análisis bioinformáticos del exoma en el genoma de cuatro individuos Ngöbe descritos arriba nos condujo a la identificación de múltiples variantes génicas posiblemente asociadas a ENT. Sin embargo, nos enfocamos en genes asociados a cánceres; particularmente en el polimorfismo rs1801133del gen MTHFR como un candidato relacionado a enfermedades (cánceres) en la población panameña. Los cuatro genomas analizados mostraron la presencia de esta variante de manera consistente. Los análisis y revisión de diferentes bases de datos contrastadas indican que este polimorfismo es una categoría de variante previamente reportada en la literatura con posible rol en riesgo de varios cánceres. Por lo tanto, dirigimos nuestros esfuerzos en determinar la frecuencia de esta variante en la población amerindia panameña. La

estrategia bioinformática para identificar este SNP del gen MTHFR fue GATK (*Genome Analysis Toolkit*) para llamar polimorfismos de un solo nucleótido (SNP) de archivos BAM (*Binary Alignment Map*) y usamos ANNOVAR para variantes en conjunto con bases de datos clínicas descritas arriba. Luego de culminar esta etapa, nuestro siguiente objetivo fue clasificar la frecuencia alélica de esta variante en una muestra de cincuenta individuos no relacionados de la ancestral Ngäbe. Estudios previos han reportado que una muestra de entre 40 y 50 individuos de esta población son suficientes para tener un buen estimado de diversidad y frecuencias génicas, ya que es una población con baja diversidad genética (Castro et al., 2007). Por esta razón, el siguiente paso fue optimizar un método de genotipaje del SNP rs1801133 en esta muestra de la población Ngäbe.

5. Genotipaje del Polimorfismo rs1801133 del gen MTHFR en amerindios Ngöbe mediante PCR-secuenciación Sanger y análisis de restricción *in silico*.

Las variantes de algunos pocos genes candidatos fueron verificadas en la literatura para establecer los métodos de genotipación mediante PCR o secuenciación de ADN, particularmente para el polimorfismo rs1801133 de MTHFR. Se ordenaron los primeros y se realizaron PCR y electroforesis en agarosa en los laboratorios del Departamento de Genética y Biología Molecular de la VIP[1]. Estos productos de PCR fueron enviados para secuenciación Sanger y analizados en más detalle posteriormente (ver a continuación).

5.1 Amplificación de la Región del Gen MTHFR con el SNP rs1801133

Realizamos amplificación por PCR y secuenciación preliminar de varios SNPs de algunos genes candidatos y obtuvimos resultados significativos en el polimorfismo tipo SNP rs1801133 del gen MTHFR (metilentetrahidrofolato reductasa). El genotipado del polimorfismo rs1801133 se realizó combinando métodos de análisis moleculares de amplificación por PCR y secuenciación Sanger del producto de PCR junto con métodos de análisis de restricción *in silico* mediante polimorfismos de la longitud de los fragmentos de restricción (PCR-Secuenciación-RFLP). Esta estrategia fue adaptada por nosotros basada en métodos descritos previamente (Khalil et al., 2021). Para esto, simplificamos una región del exón 5 del gen MTHFR que contiene el sitio del SNP. La mezcla de reacción fue realizada en 30 µl que consistieron en 15 µl de master mix (BlasTaq™ 2X PCR MasterMix), 3 µl de cada primer (5µM cada primer), templado de ADN genómico de 2 µl y 7 µl de agua. Como control negativo, utilizamos una mezcla de reacción

[1] Vicerrectoría de Investigación y Posgrado de la Universidad de Panamá

con todos los componentes, excepto que el DNA templado fue reemplazado con agua. Para amplificar el fragmento de 198 pb de la región que contiene el SNP, se usaron los primers FW (5'-TGA AGG AGA AGG TGT CTG CGG GA-3') y RV (5'-AGG ACG GTG CGG TGA GAG TG-3'). Las condiciones de termociclador fueron las siguientes: 95°C por 10 min (desnaturalización inicial y activación de la *Hot Start* DNA Polimerasa), seguido de 35 ciclos a 94°C por 30 s, 62°C por 30 s yC 72° 30 s y una extensión final de 72°C por 10 min. Se usaron 3 µL de los productos de PCR para resolverlos y visualizarlos mediante electroforesis en gel de agarosa al 1% en tampón TAE 1X y teñidos con GelRed visualizando la banda del tamaño esperado de 198 pares de bases, la cual fue verificada que contaba con un estándar de peso molecular (100 bp PCR ladder).

La electroforesis en gel se realizó a 70 V por 10-15min y luego se subió a 100 V durante 40-45min. Los resultados fueron fotodocumentados digitalmente en un Gel Doc™ EZ *System* (BioRad). Estos productos de PCR fueron procesados y preparados para envío a la empresa Psomagen en los Estados Unidos para secuenciación Sanger. Se enviaron un total de 58 muestras de Ngöbe a la compañía Psomagen para secuenciación en *New York,* EE.UU.

6. Procesamiento y análisis de secuencias para el genotipaje

Una vez que recibimos las secuencias, se les realizó varios análisis para determinar el genotipaje y otros controles mediante los siguientes pasos:

- Alineamiento de ambas hebras de los primers FW y RV en *SEQUENCHER*
- Limpieza de secuencias en *SEQUENCHER*
- Obtención de secuencias consenso en *SEQUENCHER*
- Verificación del tamaño del fragmento en *SEQUENCHER*
- Verificar la identidad de la secuencia en *BLAST*
- Análisis de restricción in sílico con en *SEQUENCHER* o NEBCutter
- Alineamiento de los nucleótidos en MEGA
- Traducción a proteínas con un predictor de traducción de *ExPasy*
- Alineamiento de las proteínas en MEGA
- Verificación de la posición de los nucleótidos del SNP
- Verificación de la posición de los aminoácidos del SNP

6.1 Alineamiento de las dos hebras, limpieza, obtención secuencias consenso y verificación tamaño

Los cuatro primeros pasos se realizaron en el programa *SEQUENCHER* 4.1.4. (*Gene Codes Corporation*, Ann Arbor, MI USA). Los primers FW (5′-TGAAGGAGAAGGTGTCTGCGGGA-3′) y el RW (5′-AGGACGGTGCGGTGAGAGTG-3′) se importaron en el programa y se ensamblaron usando los parámetros del programa. Una vez ensambladas, se observa el tamaño del fragmento. La limpieza consistió en revisar los cromatogramas y según esto, corregir las bases ambiguas de la secuencia consensus, incluso eliminando bases que se encontraran al extremo de la secuencia de ser necesario. Al mismo tiempo, se buscaba la posición del SNP. Al final, se exportó la secuencia consensus para cada muestra en formato FASTA. Para los homocigotos se generó una sola secuencia consensus, mientras que, para los heterocigotos, se generaron y exportaban dos secuencias consensus: una por cada alelo, lo cual se verificaba mediante la presencia de picos dobles en los cromatogramas en la posición del SNP.

6.1.1 Búsqueda de secuencias de referencia nucleotídicas y proteínicas

La identidad molecular de todas las secuencias fue verificada con los datos de secuencias depositadas en Genbank (NCBI-NIH) utilizando un Blastn (*Basic Local Alignment Search Tools*), escogiendo una base de datos genómicos y transcritos humanos para identificar homologías con secuencias de referencia. Una vez encontradas estas secuencias nucleotídicas, las mismas fueron examinadas en *ExPasy* para obtener la predicción de su secuencia de aminoácidos en la proteína. Las secuencias nucleotídicas y de proteínas exportadas se alinearon con secuencias de referencia dentro de la base de datos del NCBI-GenBank.

6.2 Análisis de restricción *in sílico*

Realizamos el análisis de restricción para verificar la posición del SNP, ya que en esta posición ocurre un sitio de restricción de manera específica. El análisis de restricción se ejecutó en la plataforma NEBCutter en sílico o en programa *SEQUENCHER*. La enzima de restricción para este análisis se escogió con base en el artículo de (Khalil et al., 2021), que describe el uso de la enzima HinfI para hacer digestión del fragmento amplificado de 198 pb para el genotipaje del polimorfismo C677T de MTHFR. Esta enzima corta los productos de PCR de los homocigotos mutantes en fragmentos de 175 pb y 23 pb. Los homocigotos de tipo silvestre no tienen el sitio de restricción para la enzima y, por lo tanto, presentan una sola banda de 198 pb. En los heterocigotos se produjo un patrón de tres bandas; 198 pb, 175 pb y 23 pb. Estas predicciones

por análisis de restricción *in silico* junto con la posición del SNP en los cromatogramas permitieron identificar el genotipo de cada muestra. En la plataforma se sometían las secuencias exportadas, así como la secuencia de referencia nucleotídica para observar el sitio de restricción y el tamaño de los fragmentos que cortaba la enzima.

6.3 Alineamiento en de nucleótidos en MEGA, traducción a proteínas y alineamiento en MEGA

Las secuencias nucleotídicas exportadas se alinearon con las secuencias de referencia nucleotídicas en dos archivos distintos de MEGA. Sucesivamente, se tradujo cada secuencia con *ExPasy* y se introdujeron las traducciones en Blastp para verificar su identidad molecular a nivel de proteína que coincidiera o no con la secuencia de referencia proteínica para esa región del gen MTHFR. Luego de confirmar que eran iguales a las secuencias de referencia, se alinearon las secuencias traducidas a proteínas con las secuencias proteínicas.

6.4 Verificación de la posición del nucleótido (SNP) y del aminoácido

La verificación de la posición del nucleótido y del aminoácido se hizo buscando la posición de rs1801133 dentro de los gráficos de las secuencias de referencias nucleotídicas y proteínicas en NCBI. En esa misma línea, se investigó en Ensembl si la posición era la variante dentro del transcrito o dentro de la región CDS. Una vez encontradas las posiciones, se buscaban dentro de los alineamientos de MEGA para ver si los SNPs y los aminoácidos mutados se alineaban en las posiciones que resultaban en las secuencias de referencia del GenBank y Ensembl.

7. Análisis de la estructura poblacional

Ya verificado el genotipo de cada muestra, se realizó un análisis de la estructura poblacional a las 55 muestras mediante el programa PopGene. Se prepararon los datos para someterlos en el programa sustituyendo el alelo silvestre por A y el alelo mutante por B, de modo que el genotipo heterocigoto resultaba como AB y el genotipo mutante como BB. Luego, se importó el archivo y se seleccionaron las barras *Codominant Data* y *Diploid Data* para hacer el análisis, en el cual se calcularon las frecuencias alélicas de ambos alelos.

CAPÍTULO IV: RESULTADOS

1. Cuantificación y Verificación de Calidad de ADN genómico

Se cuantificaron 31 muestras mediante Nanodrop, de las cuales presentamos algunas muestras representativas en la **Tabla 7**. Todas las muestras de ADN mostraron una concentración mayor de 100 ng/µL, el cual era considerado el mínimo. La mayoría tuvo un valor entre 1.8 y 2.0 para la razón de 260/280 y un valor mínimo de 1.8 para la razón 260/230 (**Tabla 7**), lo cual estaba de acuerdo con los estándares de calidad requeridos por la compañía de secuenciación genómica.

Tabla 7. Concentración de muestras de DNA determinadas mediante NanoDrop

	Código	**ng/µL**	**260/280**	**260/230**
1	P11/2	420.5	1.81	1.91
2	732/2	199.7	1.82	1.27
3	P2/3	1008.7	1.80	2.06
4	P2	342.1	1.83	0.92
5	P6	242.8	1.83	0.74
6	736D	160.6	1.80	2.01
7	852	143.3	1.81	1.78
8	P13/2	990.6	1.82	1.96
9	P10/2	519.4	1.82	1.58
10	P6/2	899.4	1.79	1.91
11	P7/2	356.7	1.82	1.97
12	P8	421.9	1.84	1.01
13	P3	189.6	1.80	1.02
14	P5/2	511.8	1.82	1.24
15	P4/2	457.9	1.82	2.13
16	P12/3	649.0	1.84	1.43
17	P9/2	524.7	1.81	1.96
18	P9/3	995.2	1.84	2.18
19	P8/3	133.3	1.77	0.93
20	P7/3	721.9	1.68	1.55
21	P5/3	384.1	1.75	1.17

22	P3/3	167.3	1.77	1.35
23	P5	511.3	1.83	1.22
24	729/2	322.2	1.82	1.15
25	P2/2	402.0	1.81	1.40
26	P12/2	266.5	1.83	1.47
27	P3/2	360.4	1.81	1.28
28	P7	370.7	1.84	1.69
29	P8/2	688.4	1.84	1.94
30	P1	52.4	1.62	3.48
31	714/2	259.6	1.81	1.34

Las muestras de mejor concentración, determinada por NanoDrop, se escogieron para analizar la calidad de las mismas mediante electroforesis en agarosa. Esta técnica nos permitió verificar si había o no algo de degradación con DNA de bajo peso molecular o manchas de *smear*. Los análisis demostraron que la mayoría de las muestras tenían DNA de alto peso molecular con muy poca o ninguna señal de degradación (**Figura 14**).

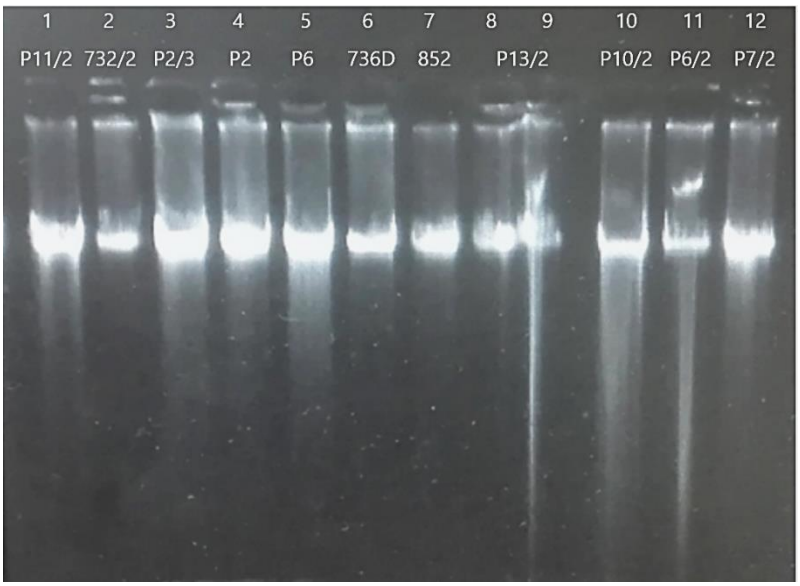


Figura 14. Electroforesis de DNA genómico en gel de agarosa. De izquierda a derecha: P11/2, 732/2, P2/3, P2, P6, 736D, 852, P13/2 (carriles 8 y 9), P10/2, P6/2 y P7/2.

2. Sexado de las muestras mediante PCR

El sexado de las muestras se realizó mediante identificación molecular del cromosoma Y por PCR amplificando un fragmento de 170 pb de las repeticiones alfoides de dicho cromosoma (Witt & Erickson, 1989; Wolfe et al., 1985). Un gen nuclear (PPARγ2), del cual se tenían cebadores en el laboratorio, fue utilizado como control. Los resultados se muestran en la **Figura 15** y en la **Tabla 8**, donde se evidencia de forma representativa la amplificación exitosa de ambos genes y la determinación del sexo de cada muestra respectivamente.

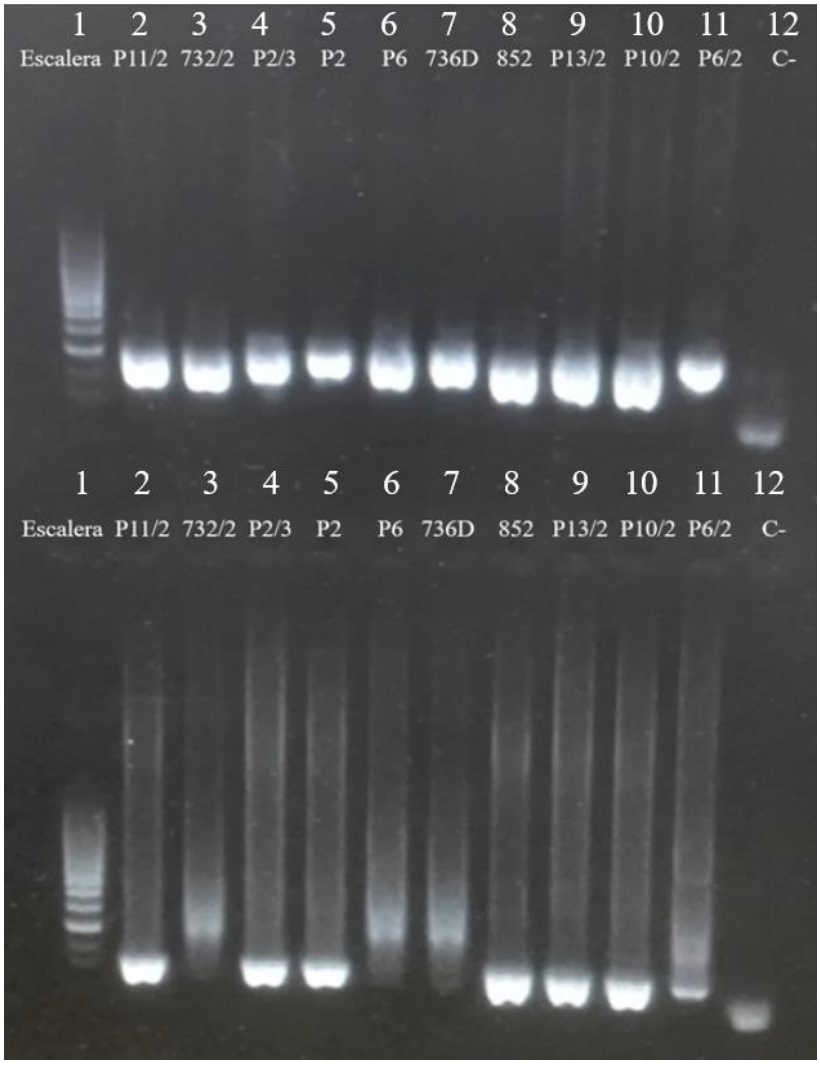


Figura 15. Electroforesis en gel de agarosa de PCR con gen nuclear control (arriba) y el fragmento del *cromosoma Y* (abajo). Arriba: gen control PPARγ2. De izquierda a derecha: P11/2, 732/2, P2/3, P2, P6, 736D, 852, P13/2, P10/2, P6/2 y Master Mix control negativo. Como se esperaba, todas las muestras son positivas excepto el Master Mix control-. Abajo: cromosoma Y. De izquierda a derecha: P11/2, 732/2, P2/3, P2, P6, 736D, 852, P13/2, P10/2, P6/2 y Master Mix control -. Las muestras P11/2, P2/3, P/2, 852, P13/2 y P10/2 resultaron ser todas positivas para muestras de varones; mientras que las muestras 732/2, P6, 736D y P6/2 resultaron ser femeninas.

Como se ejemplifica en la **Tabla 8,** el gen control se amplifica en todas las muestras y que se detectó el *cromosoma Y* en seis muestras, que representan los varones, mientras que las otras cuatro son mujeres.

Tabla 8. Electroforesis de agarosa 1% del PCR1 control y de *cromosoma Y.*

Arriba: amplificación del gen control PPARγ2

Muestra/Carril	**Código**	**Amplificación**
1	Ladder	N.A.
2	P11/2	+
3	732/2	+
4	P2/3	+
5	P2	+
6	P6	+
7	736D	+
8	852	+
9	P13/2	+
10	P10/2	+
11	P6/2	+
12	Master Mix control -	-

Abajo: amplificación del *cromosoma Y* y sexado

Muestra/Carril	**Código**	**Sexado**

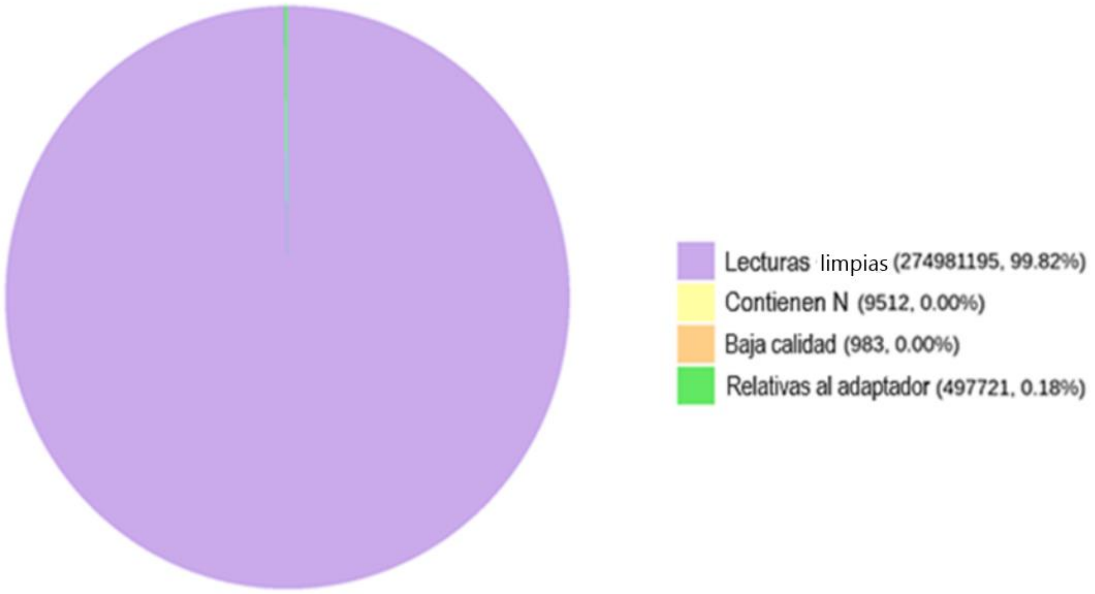
1	Ladder	PCR	Sexo
2	P11/2	+	Varón
3	732/2	-	Fémina
4	P2/3	+	Varón
5	P2	+	Varón
6	P6	-	Fémina
7	736D	-	Fémina
8	852	+	Varón
9	P13/2	+	Varón
10	P10/2	+	Varón
11	P6/2	-	Fémina
12	Master Mix control -	-	

3. Análisis bioinformático de secuencias genómicas

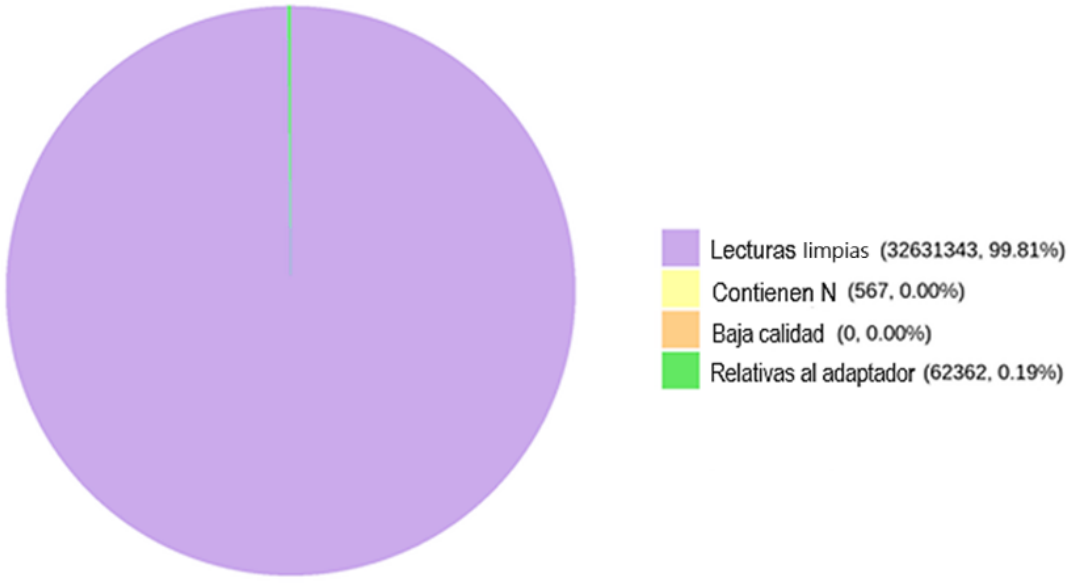
3.1 Control de calidad

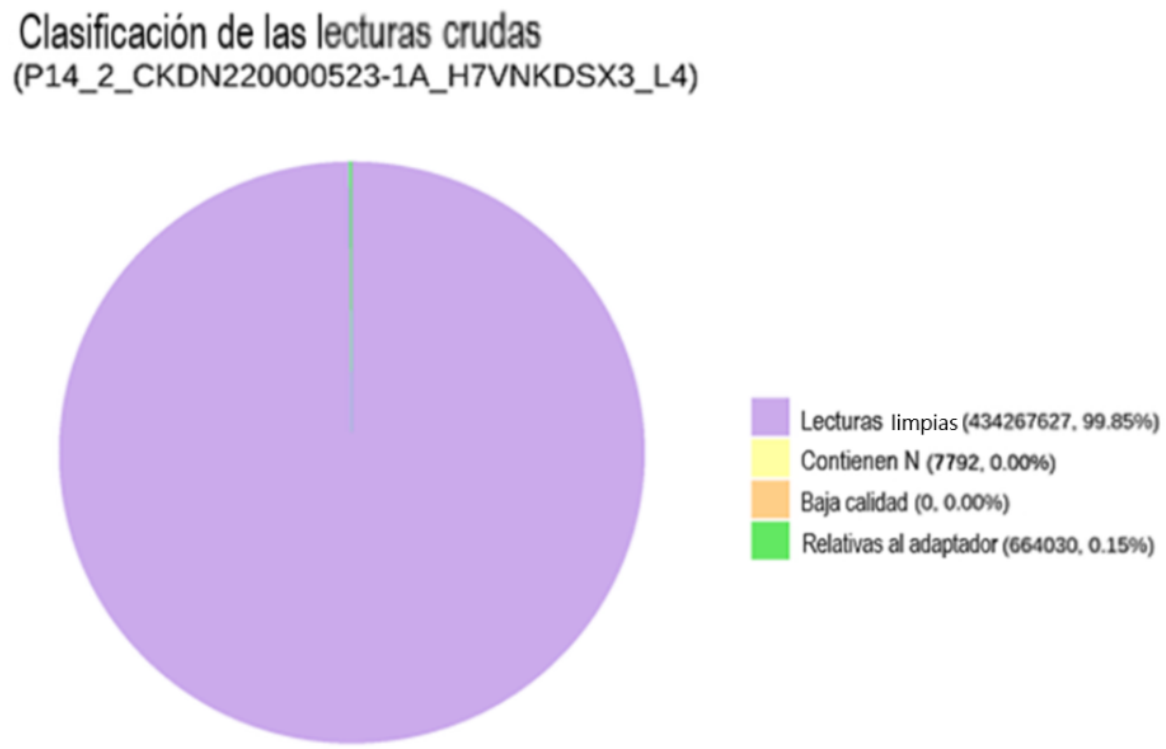
En primer lugar, se analizó la calidad de las secuencias genómicas. Se clasificaron las lecturas *reads* sin procesar, "crudas" (**Figura 16**), descartando las lecturas de extremos emparejados si una de ellas contenía contaminación del adaptador o si más de 10 bases eran inciertas en cualquier lectura o si la proporción de baja calidad era mayor de 50 en cualquier lectura (Novogene, 2022). Así, se determinó que en todas las muestras un porcentaje mayor de 99.80% de lecturas que se secuenciaron eran limpias, ninguna era no específica (N) ni baja calidad; y menos del 0.20% estaba asociada a los adaptadores.

Clasificación de las lecturas crudas
(P11_2_CKDN220000518-1A_H7VGKDSX3_L1)



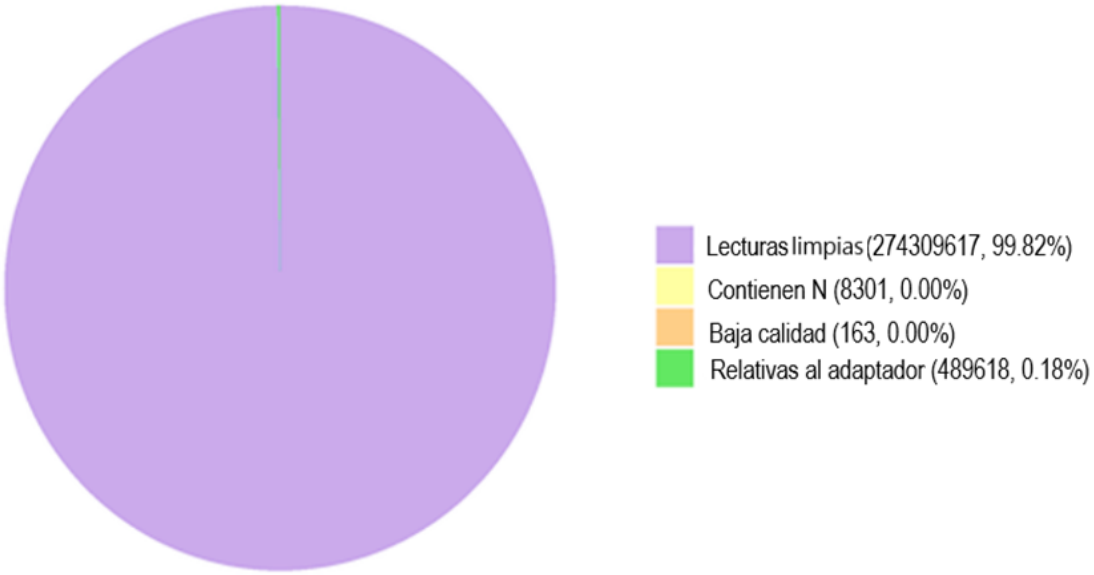
Clasificación de las lecturas crudas
(P11_2_CKDN220000518-1A_HGC2VDSX3_L2)




Clasificación de las lecturas crudas
(P14_2_CKDN220000523-1A_H7VNKDSX3_L4)
Lecturas limpias (434267627, 99.85%)
Contienen N (7792, 0.00%)
Baja calidad (0, 0.00%)
Relativas al adaptador (664030, 0.15%)


Clasificación de las lecturas crudas
(P713_CKDN220000526-1A_HGC2VDSX3_L2)
Lecturas limpias (29780048, 99.81%)
Contienen N (517, 0.00%)
Baja calidad (0, 0.00%)
Relativas al adaptador (56151, 0.19%)

Clasificación de las lecturas crudas
(P713_CKDN220000526-1A_H7VGKDSX3_L2)



Clasificación de las lecturas crudas
(P715_CKDN220000527-1A_H7VGKDSX3_L1)

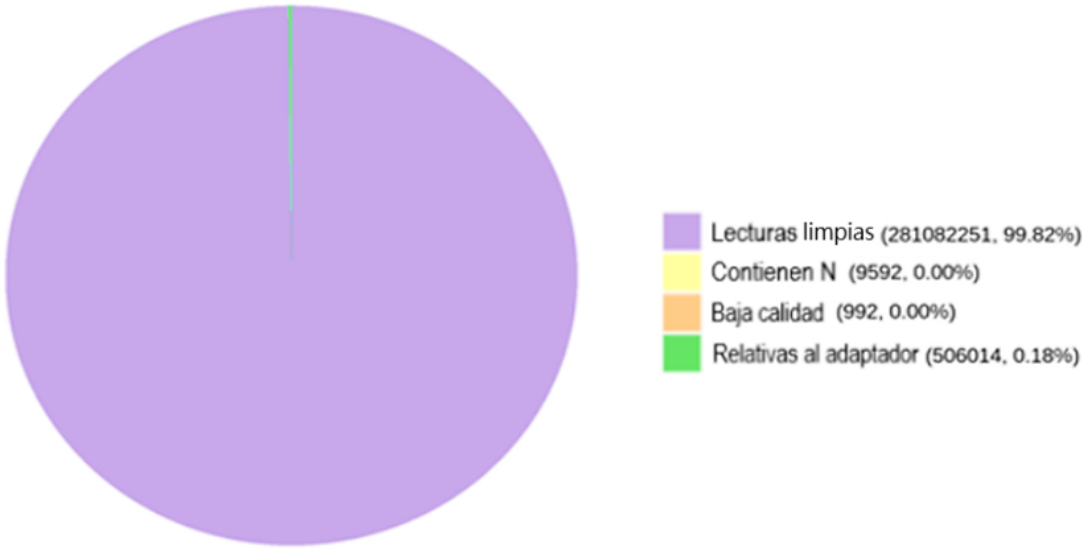
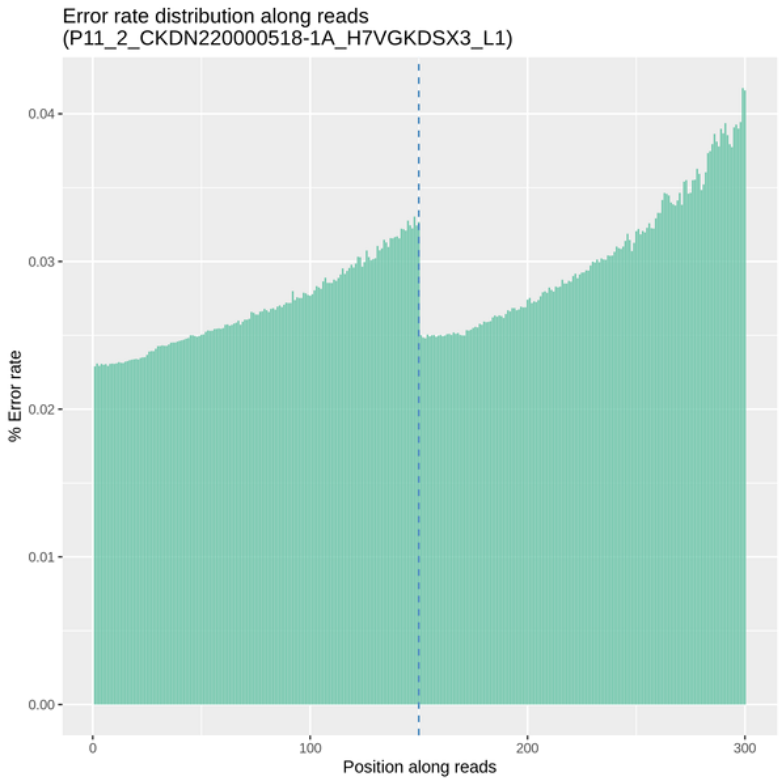
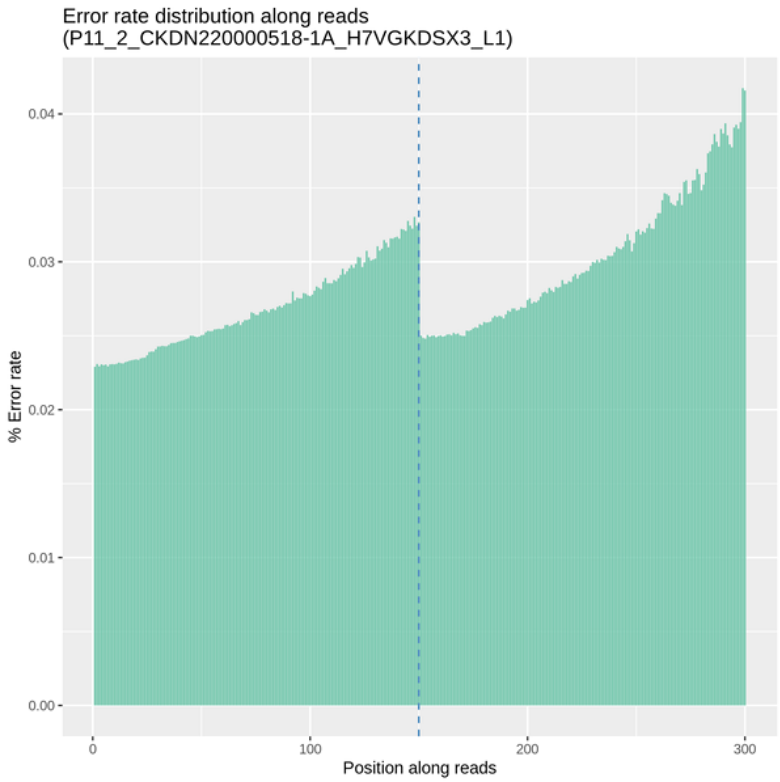
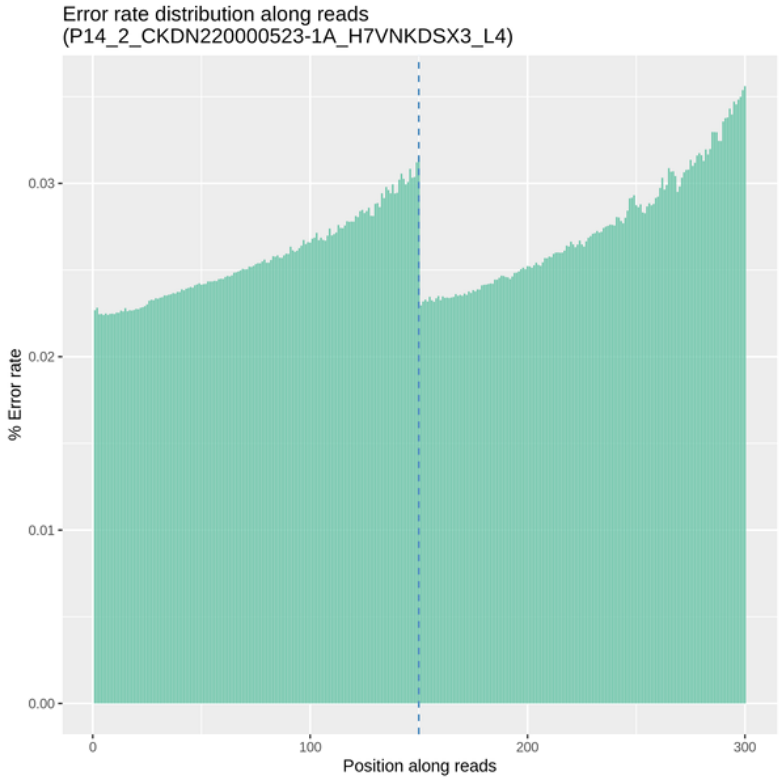


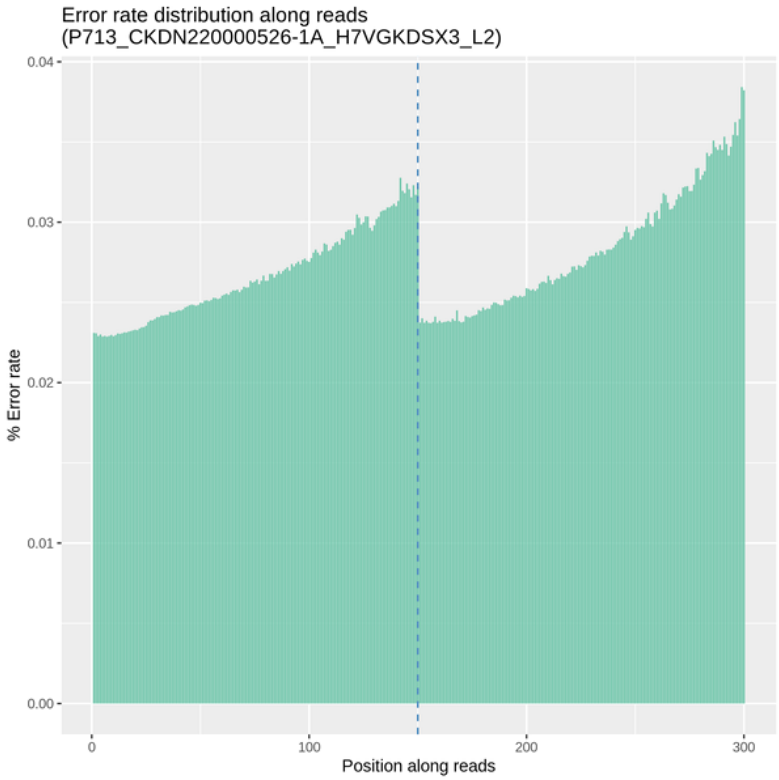
Figura 16. Clasificación sobre Calidad de datos crudos. El código de colores indica para violeta es el porcentaje de secuencias limpias (clean reads); amarillo indica el porcentaje de secuencias no específicas "N"; El anaranjado indica porcentaje de secuencias de baja calidad y el verde indica secuencias relacionadas con los adaptadores de las librerías.

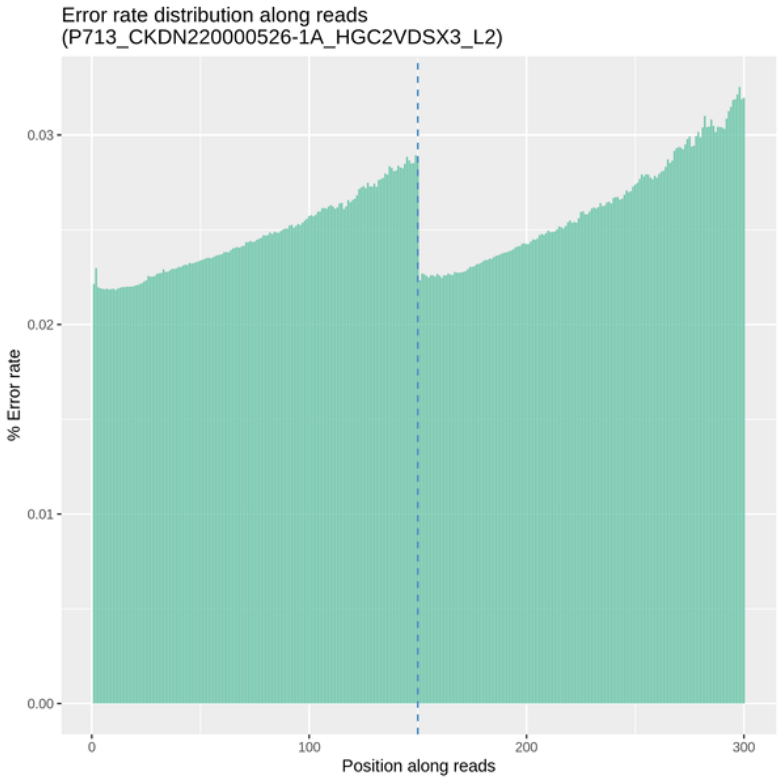
Las gráficas ilustran que los datos de los cuatro genomas secuencias (P11_2, P14_2, P713 y P715) son de alta calidad, ya que, el porcentaje de error es cerca de 0.03%. En particular, la mitad de los genomas (P11_2_HGC2VDSX3_L2, P713_HGC2VDSX3_L2, HHM3TDSX3_L3 y P715_HGC2VDSX3_L2) poseen menos del 0.03% en el error general y la otra mitad tiene un error poco más alto de 0.03%. En todos los genomas el porcentaje de secuencias no específicas o de baja calidad es menor de 0.00%.

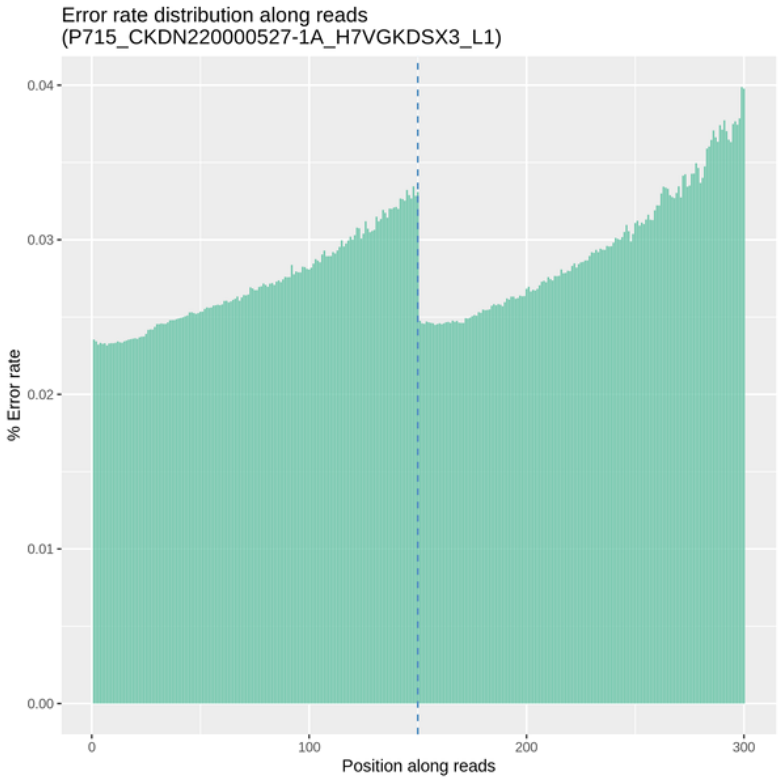


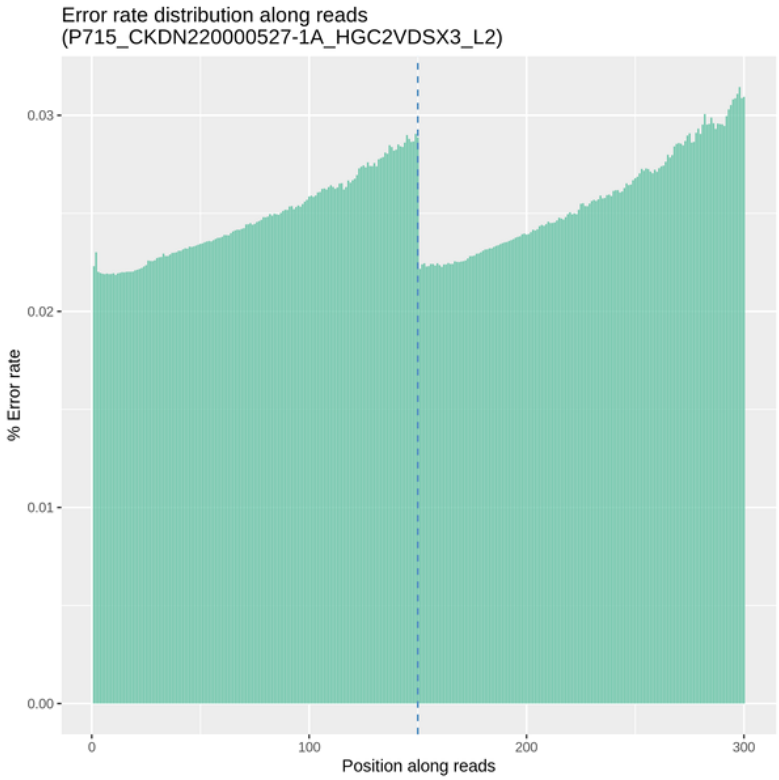

Error rate distribution along reads
(P11_2_CKDN220000518-1A_H7VGKDSX3_L1)
% Error rate
0.04
0.03
0.02
0.01
0.00
0
100
200
300
Position along reads


Error rate distribution along reads
(P14_2_CKDN220000523-1A_H7VNKDSX3_L4)
% Error rate
0.03
0.02
0.01
0.00
0
100
200
300
Position along reads


Error rate distribution along reads
(P713_CKDN220000526-1A_H7VGKDSX3_L2)
% Error rate
0.04
0.03
0.02
0.01
0.00
0
100
200
300
Position along reads


Error rate distribution along reads
(P713_CKDN220000526-1A_HGC2VDSX3_L2)
% Error rate
0.03
0.02
0.01
0.00
0
100
200
300
Position along reads


Error rate distribution along reads
(P715_CKDN220000527-1A_H7VGKDSX3_L1)
% Error rate
0.04
0.03
0.02
0.01
0.00
0
100
200
300
Position along reads


Error rate distribution along reads
(P715_CKDN220000527-1A_HGC2VDSX3_L2)
% Error rate
0.03
0.02
0.01
0.00
0
100
200
300
Position along reads

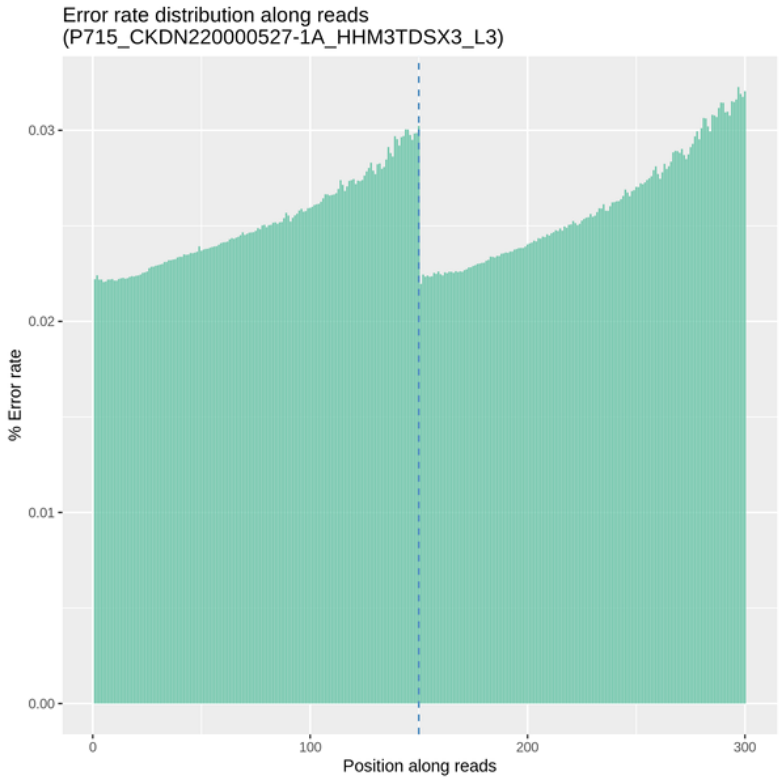
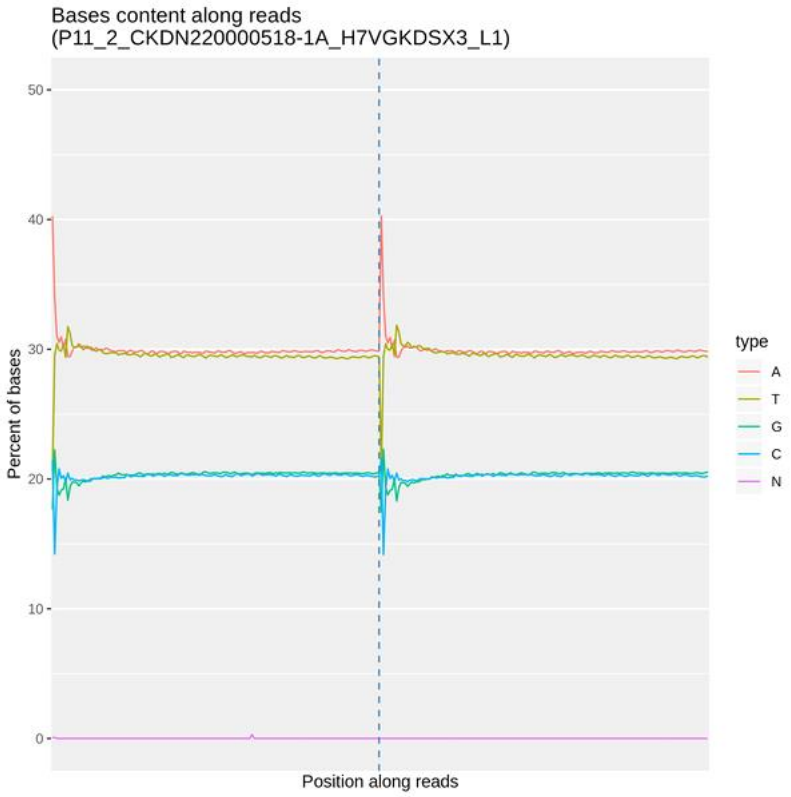
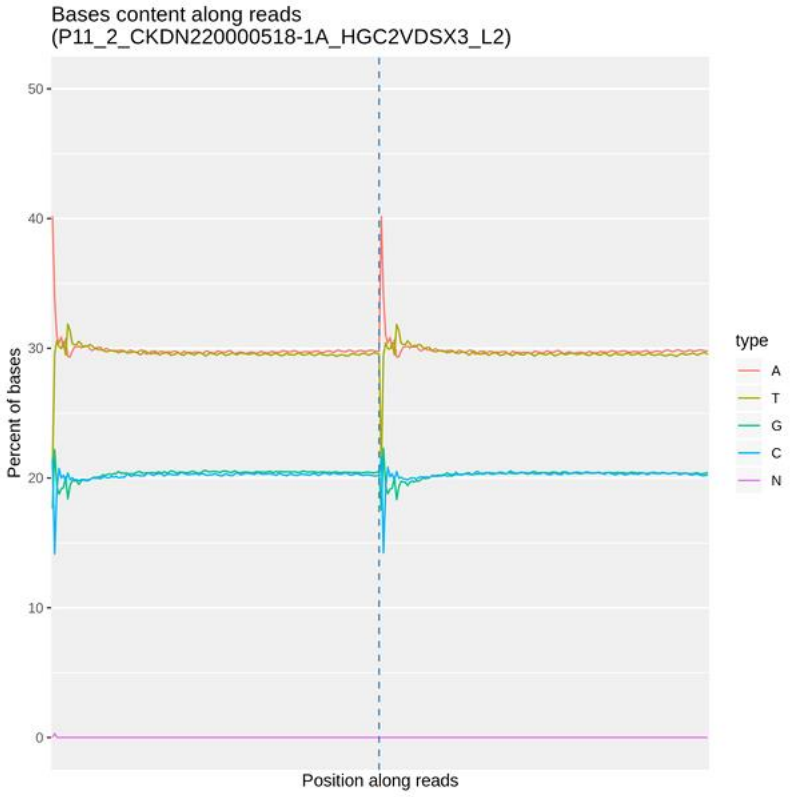
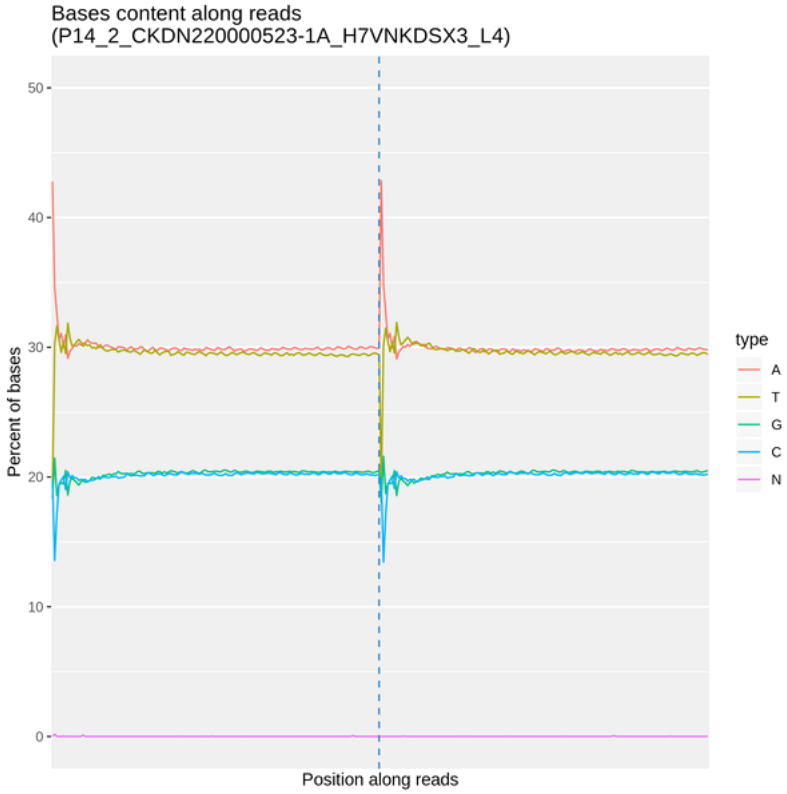
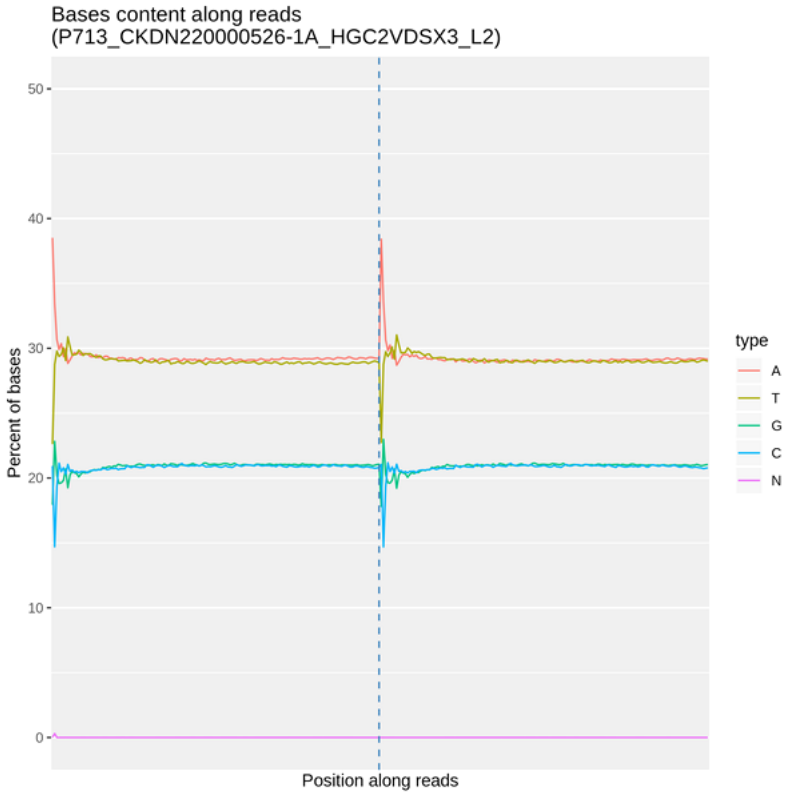


Figura 17. Distribución de la tasa de error de secuenciación. Distribución de la tasa de error de secuenciación se aplicó para detectar si hay bases anormales con alta tasa de error en las lecturas de los genomas secuenciados P11_2, P14_2, P713 y P715. El eje de x representa la posición en lecturas, de acuerdo con el tamaño en longitud de la lectura de los fragmentos (*reads*) y el eje de y. La tasa de error media de todas las bases de todas las lecturas en una posición. Se observa que el porcentaje de error aumenta en las secuencias a medida que se acercan al extremo más largo de hasta 300 bases. Aunque el porcentaje de error es aceptable, ya que es muy bajo (menos de 0.03% ó 0.04%), este efecto en el porcentaje de error es compensado aún más con la alta profundidad y cobertura de cada posición secuenciada.


Bases content along reads
(P11_2_CKDN220000518-1A_H7VGKDSX3_L1)
Percent of bases
Position along reads
type
A
T
G
C
N


Bases content along reads
(P11_2_CKDN220000518-1A_HGC2VDSX3_L2)
Percent of bases
Position along reads
type
A
T
G
C
N


Bases content along reads
(P14_2_CKDN220000523-1A_H7VNKDSX3_L4)
Percent of bases
50
40
30
20
10
0
Position along reads
type
A
T
G
C
N


Bases content along reads
(P713_CKDN220000526-1A_HGC2VDSX3_L2)
Percent of bases
50
40
30
20
10
0
Position along reads
type
A
T
G
C
N

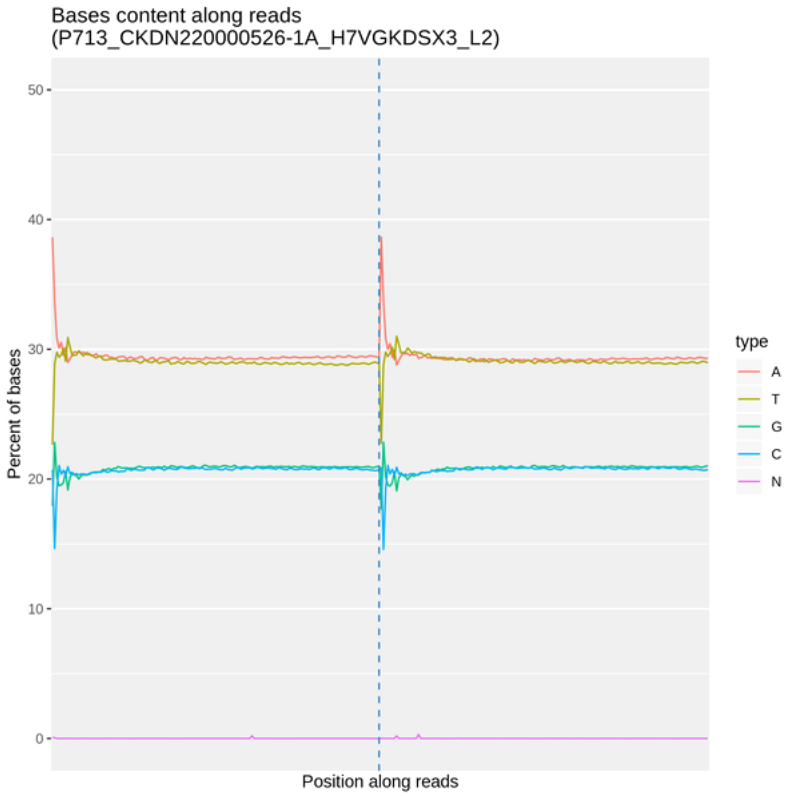
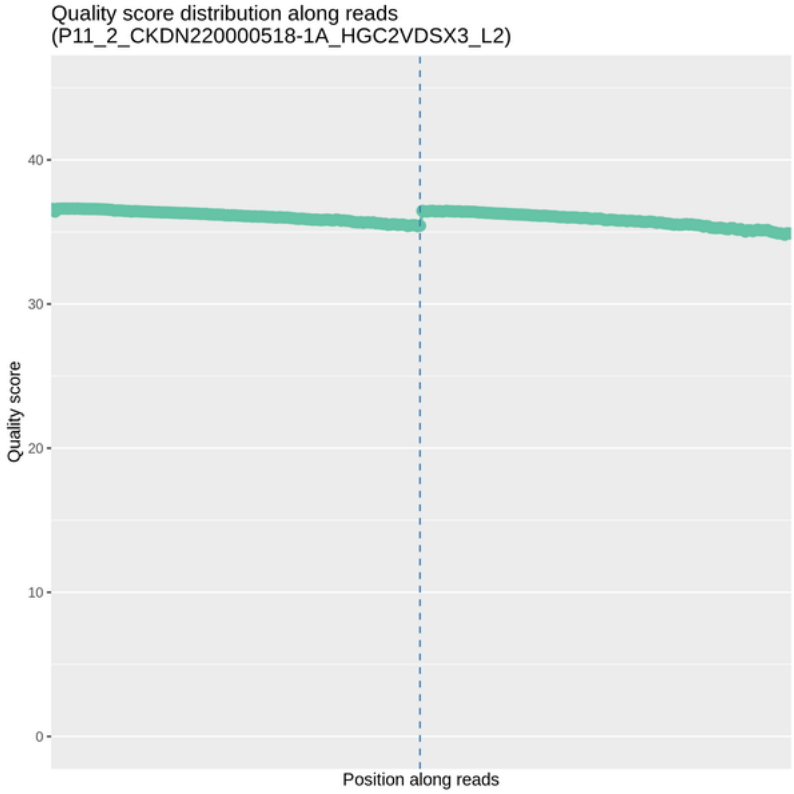
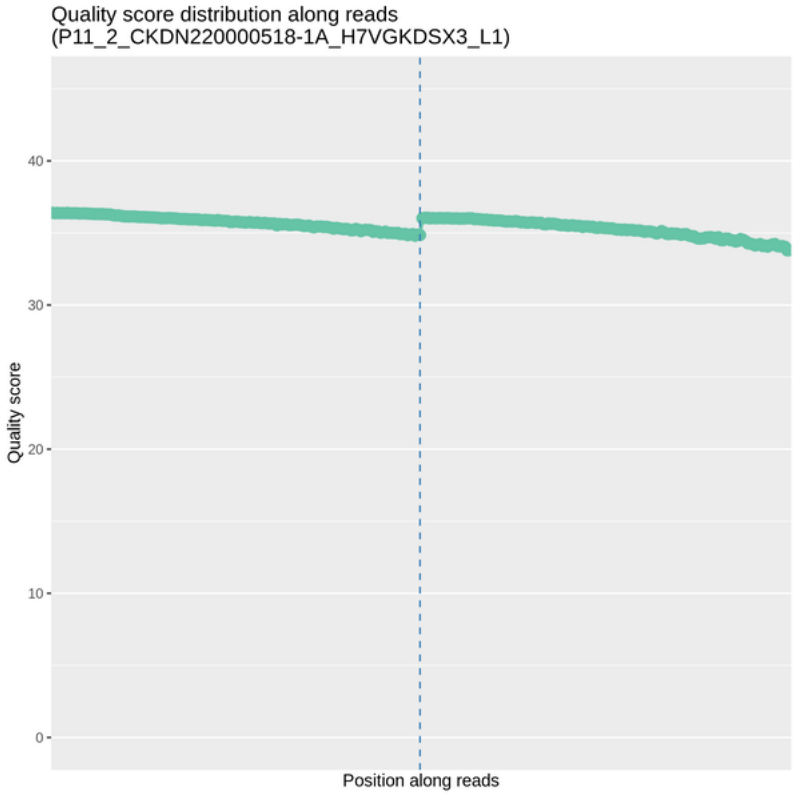
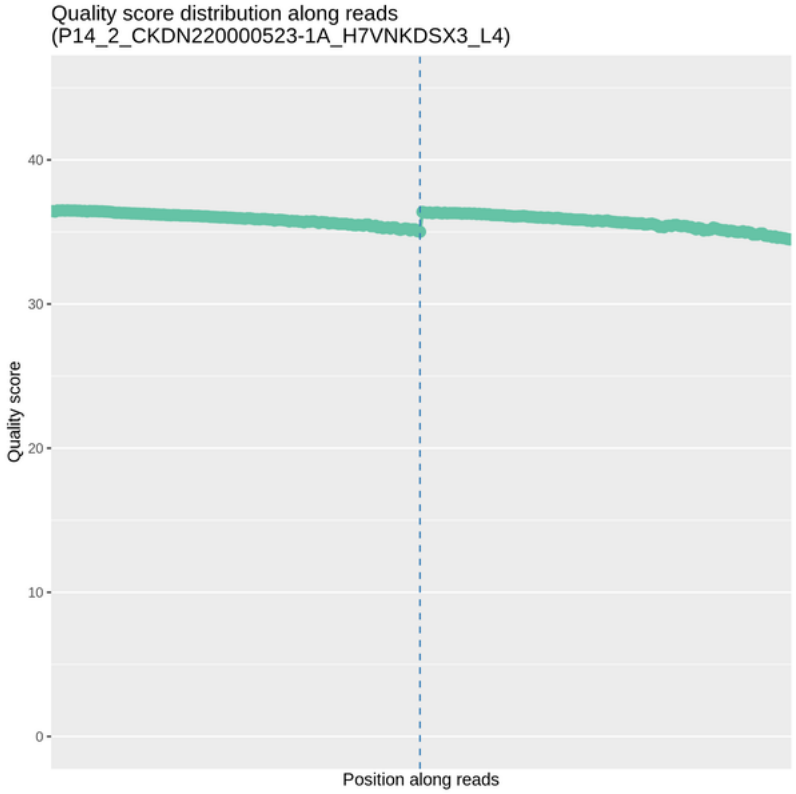


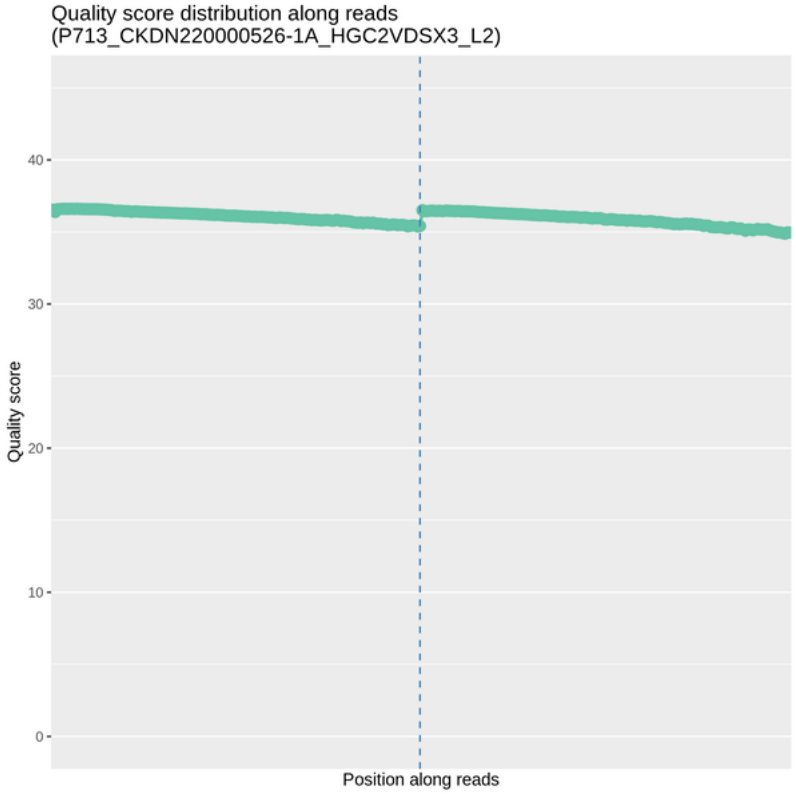
Figura 18. Distribución del porcentaje o contenido de guanina citosina (GC).El contenido en GC se ha determinado para los genomas secuenciados P11_2, P14_2, P713 y P715. El eje x representa la posición en lecturas; el eje y, el porcentaje de cada tipo de bases (A, T, G y C), que son indicadas a través de distintos colores. En las gráficas se observa el porcentaje de GC, que está alrededor del 40%, así como el porcentaje de AT, que está alrededor del 60% en todos los genomas secuenciados.

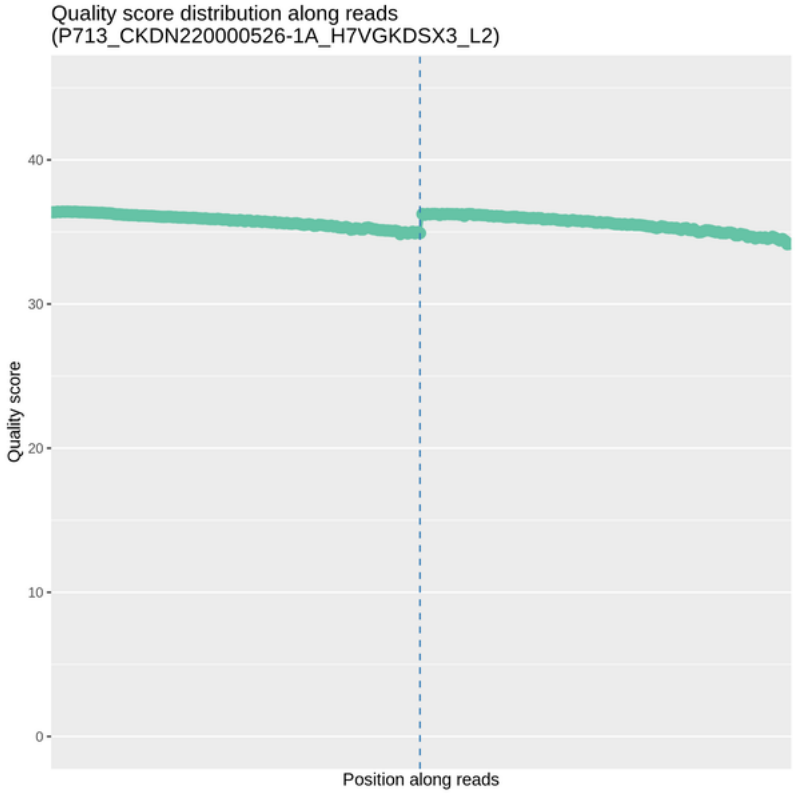
La **Figura 19** se muestra la distribución de la calidad de la secuenciación en los cuatro genomas. El puntaje de calidad se mantiene en un rango entre 30 y 40 y como se espera es más elevado al inicio de la secuencia, para luego declinar hasta el final.

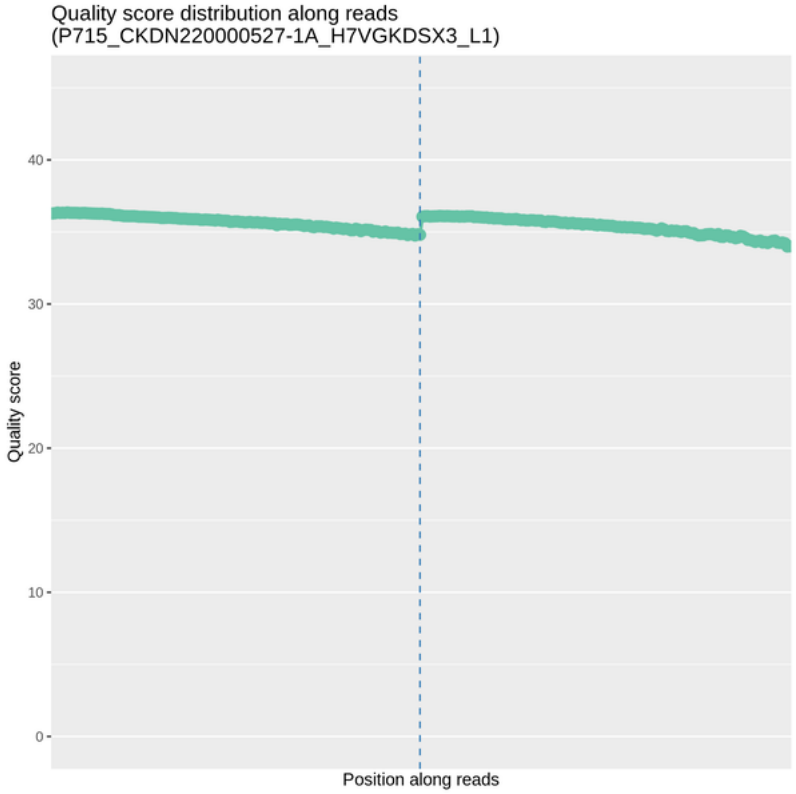

Quality score distribution along reads
(P11_2_CKDN220000518-1A_HGC2VDSX3_L2)
40
30
20
10
0
Quality score
Position along reads

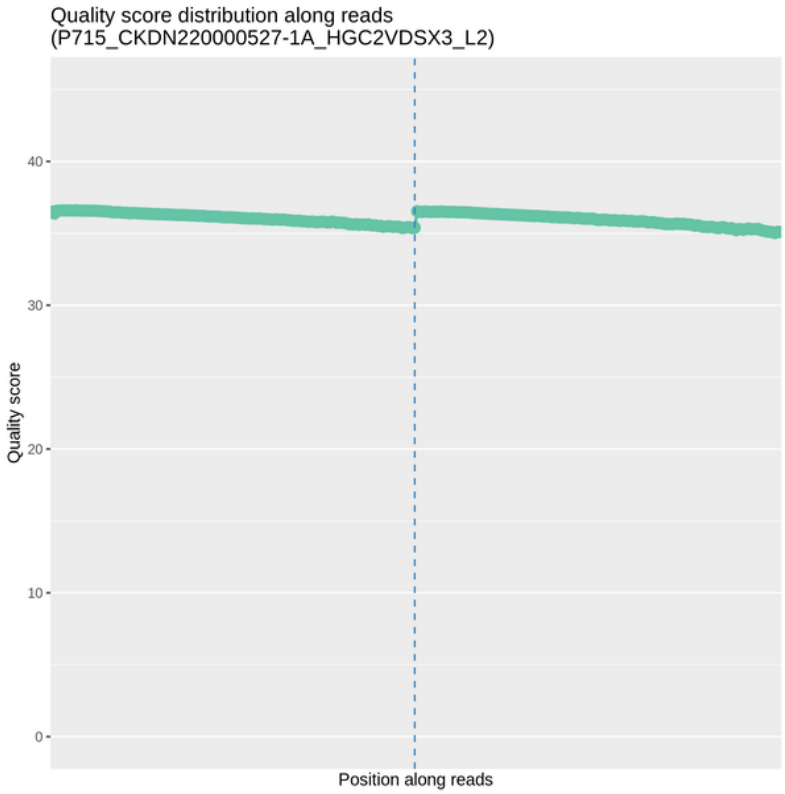

Quality score distribution along reads
(P11_2_CKDN220000518-1A_H7VGKDSX3_L1)
40
30
20
10
0
Quality score
Position along reads


Quality score distribution along reads
(P14_2_CKDN220000523-1A_H7VNKDSX3_L4)
40
30
20
10
0
Quality score
Position along reads


Quality score distribution along reads
(P713_CKDN220000526-1A_HGC2VDSX3_L2)
40
30
20
10
0
Quality score
Position along reads


Quality score distribution along reads
(P713_CKDN220000526-1A_H7VGKDSX3_L2)
40
30
20
10
0
Quality score
Position along reads


Quality score distribution along reads
(P715_CKDN220000527-1A_H7VGKDSX3_L1)
40
30
20
10
0
Quality score
Position along reads


Quality score distribution along reads
(P715_CKDN220000527-1A_HGC2VDSX3_L2)
40
30
20
10
0
Quality score
Position along reads

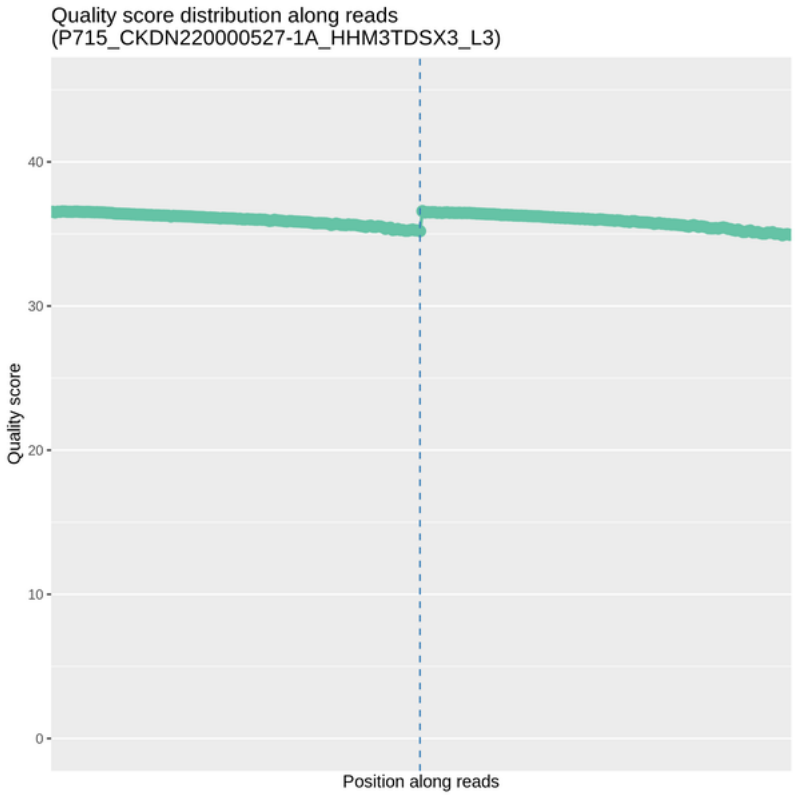


Figura 19. Distribución de calidad de secuenciación. La distribución de calidad de secuenciación se ha determinado para los genomas P11_2, P14_2, P713 y P715. El eje x representa la posición en lecturas. El eje y representa la puntuación media de la calidad de las bases de todas las lecturas en una posición.

3.1.1 Estadísticas de calidad de secuenciación

La **Tabla 9** prueba el resumen de la calidad de producción de datos. Los cuatro genomas mostraron un porcentaje efectivo de más del 99% y un porcentaje de error del 0.03%, valores de 93% para arriba para Q20, valores más elevados de 91% para Q30 y un porcentaje de alrededor de 40-41% para el contenido GC.

Tabla 9. Reseña de la calidad de producción de datos

Nombre de la muestra	NovoID	Celda de flujo/carril	Lecturas crudas	Datos crudos (G)	Efectivo (%)	Error(%)	Q20(%)	Q30(%)	GC(%)
P14_2	CKDN220000523-1A	H7VNKDSX3_L4	434939449	130,5	99,85	0,03	97,39	92,96	40,36
P11_2	CKDN220000518-1A	H7VGKDSX3_L1	275489411	92,5	99,82	0,03	96,63	91,3	40,5
P11_2	CKDN220000518-1A	HGC2VDSX3_L2	32694272	99,81	0,03	97,85	93,82	40,51	
P715	CKDN220000527-1A	H7VGKDSX3_L1	281598849	90,2	99,82	0,03	96,67	91,5	41,85
P715	CKDN220000527-1A	HGC2VDSX3_L2	16571148	99,81	0,03	97,9	94,1	42,34	
P715	CKDN220000527-1A	HHM3TDSX3_L3	2427938	99,83	0,03	97,73	93,86	42,22	
P713	CKDN220000526-1A	H7VGKDSX3_L2	274807699	91,4	99,82	0,03	97,01	92,21	41,5
P713	CKDN220000526-1A	HGC2VDSX3_L2	29836716	99,81	0,03	97,84	93,89	41,7	

Leyenda:

1) Nombre de la Muestra: código identificativo de la muestra

2) NovoID: código identificativo novo para la muestra.

3) Celda de flujo/carril: el código identificativo de la celda de flujo y el número del carril.

4) Lecturas crudas: El número de pares de lecturas de secuenciación; cuatro líneas serían consideradas como una unidad de acuerdo con el formato de FASTQ.

5) Datos crudos: Los datos originales de secuenciación en Giga bases (G).

6) Efectivo: El porcentaje de lecturas limpias en todas las lecturas crudas.

7) Error: La tasa de error promedio de todas las bases en la lectura 1 y la lectura 2; la tasa de error de una base se obtiene de la ecuación 1.

8) Q20: El porcentaje de bases con niveles de calidad de la escala Phred mayores de 20.

9) Q30: El porcentaje de bases con niveles de calidad de la escala Phred mayores de 30.

10) Contenido GC: El porcentaje de G y C en todas las bases.

3.2 Alineamiento de secuencias

En la tabla 10, se observan las estadísticas de las regiones mapeadas, de la cobertura y de la profundidad de los cuatro genomas secuenciados. La tabla muestra que las regiones mapeadas se han secuenciado exitosamente en más del 99% del mapeo. El parámetro de regiones mapeadas apropiadamente muestra tendencialmente un porcentaje ligeramente mayor al 98%. Los extremos emparejados mapeados resultaron todos alrededor del 88.80% y los extremos únicos mapeados resultaron entre 0.13 y 0.16% incluyendo estos últimos. La profundidad de cada región fue bastante alta con un promedio de 34 veces. La cobertura también fue alta, prácticamente más del 98% en todos los genomas secuenciados. La cobertura de, al menos 4 veces, fue alrededor del 98%; la de al menos 10 veces, más del 97%; y la de al menos 20 veces,

del 86% para arriba, demostrando una calidad muy buena de los datos obtenidos.

Tabla 10. Estadística de mapeo, cobertura y profundidad en cada muestra

Muestra	P14_2	P11_2	P713	P715
Total	868535254 (100%)	615225076 (100%)	608179330 (100%)	600091006 (100%)
Duplicado	202830273 (23.38%)	132671192 (21.59%)	99492618 (16.38%)	147596227 (24.63%)
Mapeadas	867379707 (99.87%)	614387945 (99.86%)	607406561 (99.87%)	599239005 (99.86%)
Mapeadas apropiadamente	852675676 (98.17%)	606687640 (98.61%)	598622640 (98.43%)	590931178 (98.47%)
Extremos emparejadosa mapeados	866826184 (99.80%)	613946614 (99.79%)	606996900 (99.81%)	598761800 (99.78%)
Extremos únicos mapeados	1107046 (0.13%)	882662 (0.14%)	819322 (0.13%)	954410 (0.16%)
Con pareja mapeada a un diferente cromosoma	9975908 (1.15%)	4748488 (0.77%)	5768464 (0.95%)	5129310 (0.85%)
Con pareja mapaeada a un diferente cromosoma (map Q>=5)	5667202 (0.65%)	2298553 (0.37%)	3077426 (0.51%)	2530333 (0.42%)
Profundidad de secuenciación promedio	43,3	30,88	30,51	30,04
Cobertura	98,96%	99,64%	98,93%	98,90%
Cobertura de al menos 4X	98,68%	99,25%	98,58%	98,54%
Cobertura de al menos 10X	98,26%	97,82%	97,94%	97,77%
Cobertura de al menos 20X	96,97%	87,23%	90,27%	86,72%

Leyenda:

1) Nombre de la muestra: Código identificativo de la muestra.
2) Total: El número total de lecturas limpias. Los índices abajo están calculados con base en las lecturas limpias.
3) Duplicado: El número de lecturas duplicadas (porcentaje: lecturas duplicadas/lecturas limpias).
4) Mapeadas: el número de lecturas que fueron contrastadas con el genoma de referencia (porcentaje).
5) Mapeadas apropiadamente: el número de lecturas que contrastan con el genoma de referencia y dentro del tamaño del inserto esperado (porcentaje).
6) Extremos emparejados mapeados: El número de lecturas con extremos emparejados que fueron contrastadas con el genoma de referencia (porcentaje).
7) Extremos únicos mapeados: El número de lecturas con extremos únicos que fueron contrastadas con el genoma de referencia (porcentaje).
8) Con pareja mapeada a un cromosoma diferente: el número de lecturas con lecturas emparejadas mapeadas a cromosomas diferentes (porcentaje).
9) Con pareja mapeada a un cromosoma diferente (map Q>=5): El número de lecturas con lecturas emparejadas mapeadas a cromosomas diferentes y al MAQ >5. Cobertura de

secuenciación promedio: la cobertura de secuenciación promedio en el genoma completo.

10) Profundidad de secuenciación promedio: la profundidad de secuenciación promedio en el genoma completo.

11) Cobertura: cobertura en todo el genoma.

12) Cobertura de al menos 4X: la cobertura en todo el genoma donde se consideran solo bases con profundidad >4X.

13) Cobertura de al menos 10X: la cobertura en todo el genoma donde se consideran solo bases con profundidad >10X.

14) Cobertura de al menos 20X: la cobertura en todo el genoma donde se consideran solo bases con profundidad >20X.

La **Tabla 11** muestra el porcentaje de cobertura de cada cromosoma en cada genoma, con un promedio de 0,990 para arriba en todos, a excepción del *cromosoma Y*, que mientras tiene una cobertura de 0,998 en el genoma P11_2, tiene una cobertura menor que 0,1 en los genomas restantes. Se mostró anteriormente en los resultados del sexado que P11_2 era un varón, por lo que el valor de su cobertura se da porque se logró cubrir el *cromosoma Y*. Esto significa también que el valor de los otros genomas indica que ellos pertenecen a mujeres.

Tabla 11. Cobertura de cada cromosoma por cada genoma secuenciado

chr	P11_2	P14_2	P713	P715
chr1	0,997	0,998	0,997	0,997
chr2	0,997	0,998	0,997	0,997
chr3	0,998	0,999	0,999	0,998
chr4	0,997	0,998	0,998	0,997
chr5	0,998	0,999	0,998	0,998
chr6	0,999	0,999	0,998	0,999
chr7	0,998	0,997	0,996	0,996
chr8	0,997	0,999	0,999	0,998
chr9	0,996	0,997	0,996	0,996
chr10	0,992	0,99	0,992	0,992
chr11	0,995	0,997	0,996	0,996
chr12	0,998	0,997	0,998	0,997
chr13	0,998	0,998	0,998	0,998
chr14	0,998	0,999	0,998	0,998
chr15	0,994	0,998	0,995	0,993
chr16	0,998	0,999	0,998	0,998
chr17	0,993	0,986	0,987	0,988
chr18	0,986	0,996	0,995	0,996
chr19	0,999	0,999	0,999	0,999
chr20	0,995	0,995	0,994	0,995
chr21	0,996	0,997	0,996	0,996
chr22	0,997	0,996	0,995	0,995
chrX	0,996	0,998	0,997	0,998
chrY	0,998	0,091	0,088	0,081

Las líneas indican los cromosomas, mientras que las columnas señalan los genomas.

3.3 Detección de variantes polimórficas en la línea germinal

3.3.1 Detección de SV *(Structural variants)*

En la **Tabla 12** se muestran las estadísticas del marcador de variantes estructurales (SV). La variante estructural más numerosa identificada fue la deleción (DEL), mientras que la menos numerosa fue la inserción.

Tabla 12. Resultados de los SV identificados en cada genoma.

Nombre de la muestra	DUP	INV	BND	DEL	INS
P11_2	944	554	854	3356	69
P14_2	1429	807	1506	4335	96
P713	1168	615	1057	3651	80
P715	1044	576	936	3444	65

Leyenda:

1) Nombre de la muestra: código identificativo de la muestra.
2) DUP: (Duplicación en tándem) El número de duplicaciones en tándem.
3) INV: (Inversión) El número de inversiones.
4) INS: (Inserción) El número de inserciones.
5) DEL: (Deleción) El número de deleciones.
6) BND: (Translocación) El número de translocaciones.

Se estimó el número total de cada tipo de SV, resultando ser los que predominaron las deleciones, seguidas por las duplicaciones y luego las translocaciones. Las inversiones y las duplicaciones fueron las menos numerosas (**Figura 20**).

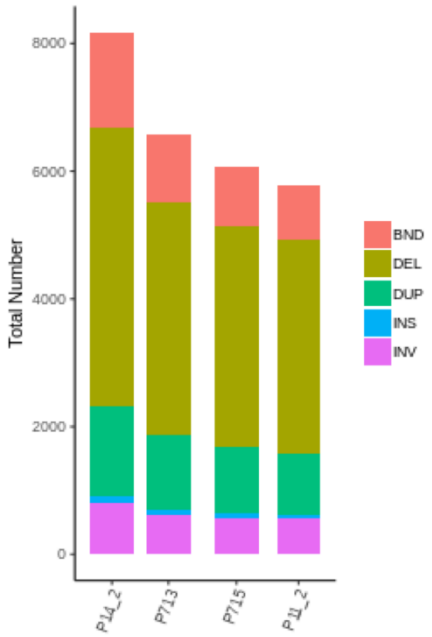


Figura 20. Número de diferentes tipos de SV en cada muestra. El eje X representa las muestras y el eje Y representa el número de cada tipo SV dentro de P14_2, P713, P715 y P11_2.

3.3.2 Detección de InDels *(Insertions/Deletions)*

En la **Tabla 13,** se muestra el número de InDels en distintas regiones genómicas, así como el cambio que ocasionaron. Se detectaron InDels que causaron una alteración en el marco de lectura. Otros causaron ganancia o pérdida de codones de parada. Aún más, otros, que se dieron en algunas regiones desconocidas o regiones que pertenecían a los UTR del RNA mensajero, que son regiones que no se traducen (3'UTR y 5'UTR) y otras regiones no codificantes de proteínas que son exónicas (*ncRNA-exonic*).

Tabla 13. InDels identificados en los genomas secuenciados.

Nombre de la muestra	P11_2	P14_2	P713	P715
CDS	614	677	663	656
Delección que desplaza el marco de lectura	142	153	147	157
Inserción que desplaza el marco de lectura	91	111	111	109
Delección que no desplaza el marco de lectura	203	219	199	201
Inserción que no desplaza el marco de lectura	155	169	175	169
Ganancia del codón de parada	10	11	14	11
Pérdida del codón de parada	1	2	0	2
Desconocido	22	20	25	17
intrónico	275900	317511	282697	273918
UTR3	6352	7275	6541	6389
UTR5	905	1056	986	971
Corte y empalme	47	49	46	49
ARNnc exónico	1794	1995	1847	1829
ARNnc intronico	45821	51990	46249	44965
ARNnc de corte y empalme	13	14	12	14
Aguas arriba	5127	5893	5478	5399
Aguas abajo	5778	6672	5862	5843
Intergénico	409664	465361	417243	406560
Total	752229	858740	767863	746836

Leyenda:

1) Nombre de la muestra: código identificativo de la muestra.
2) CDS: El número de InDels de la región codificante.

Deleción que desplaza el marco de lectura: una delección de uno o más nucleótidos que causan cambios en el marco de lectura en las secuencias codificantes de proteínas.

Inserción que desplaza el marco de lectura: una inserción de uno o más nucleótidos que causan cambios en el marco de lectura en las secuencias codificantes de proteínas.

Deleción que no desplaza el marco de lectura: una deleción que no causa cambios por desplazamiento del marco de lectura.

Inserción que no desplaza el marco de lectura: una inserción que no causa cambios por desplazamiento del marco de lectura.

Ganancia del codón de parada: una inserción o una deleción que conduce a la creación de un codón de parada en el sitio de la variante.

Pérdida del codón de parada: una inserción o una deleción que conduce a la eliminación de un codón de parada en el sitio de la variante.

Desconocido: función desconocida (debida a varios errores en la definición de la estructura del gen en el archivo de la base de datos).

3) Intrónico: El número de InDels en la región intrónica.
4) UTR3: el número de InDels en la región 3'UTR.
5) UTR5: el número de InDels en la región 5'UTR.
6) Corte y empalme: El número de InDels dentro de 2-pb de un sitio de unión de corte y empalme.
7) ARNnc exónico: el número de InDels en la región exónica de los ARNs no codificantes.
8) ARNnc intrónico: El número de InDels en la región intrónica de ARNs no codificante
9) ARNnc de corte y empalme: El número de InDels dentro de 2-pb de la unión de corte y empalme de ARNs no codificantes.
10) Aguas arriba: El número de InDels dentro de 1 kb del sitio de inicio de la transcripción.
11) Aguas abajo: El número de InDels dentro de 1 kb del sitio de terminación de la transcripción.
12) Intergénico: El número de InDels en la región intergénica.
13) Total: El número total de InDels.

En términos generales, el análisis nos da la información sobre las características de los InDels, como su heterocigosidad o su homocigocidad (**Tabla 14**).

Tabla 14. Característica de los InDels identificados en los genomas.

Nombre de la muestra	P11_2	P14_2	P713	P715
Total	752229	858740	767863	746836
Het	379297	517137	387292	370914
Hom	372932	341603	380571	375922
Porcentaje de dbSNP	665964 (88.53%)	749723 (87.31%)	677351 (88.21%)	660107 (88.39%)
novel	86265	109017	90512	86729

Leyenda:

1) Nombre de la muestra: código identificativo de la muestra.
2) Total: el número total de InDels.
3) Het: el número de heterocigotos.
4) Hom: el número de homocigotos.
5) Porcentaje de dbSNP: El número de InDels que han sido reportados en la base de datos de dbSNP divididos entre el número total de InDels.

6) Novedoso: El número de InDels que no han sido reportados en la base de datos de dbSNP.

En el **Figura 21**, se ve la misma información del **Tabla 13** de los Indels, pero manifiesta en manera más general si hay cambio en el marco de lectura, ganancia o pérdida de codones de parada o si su función es desconocida.

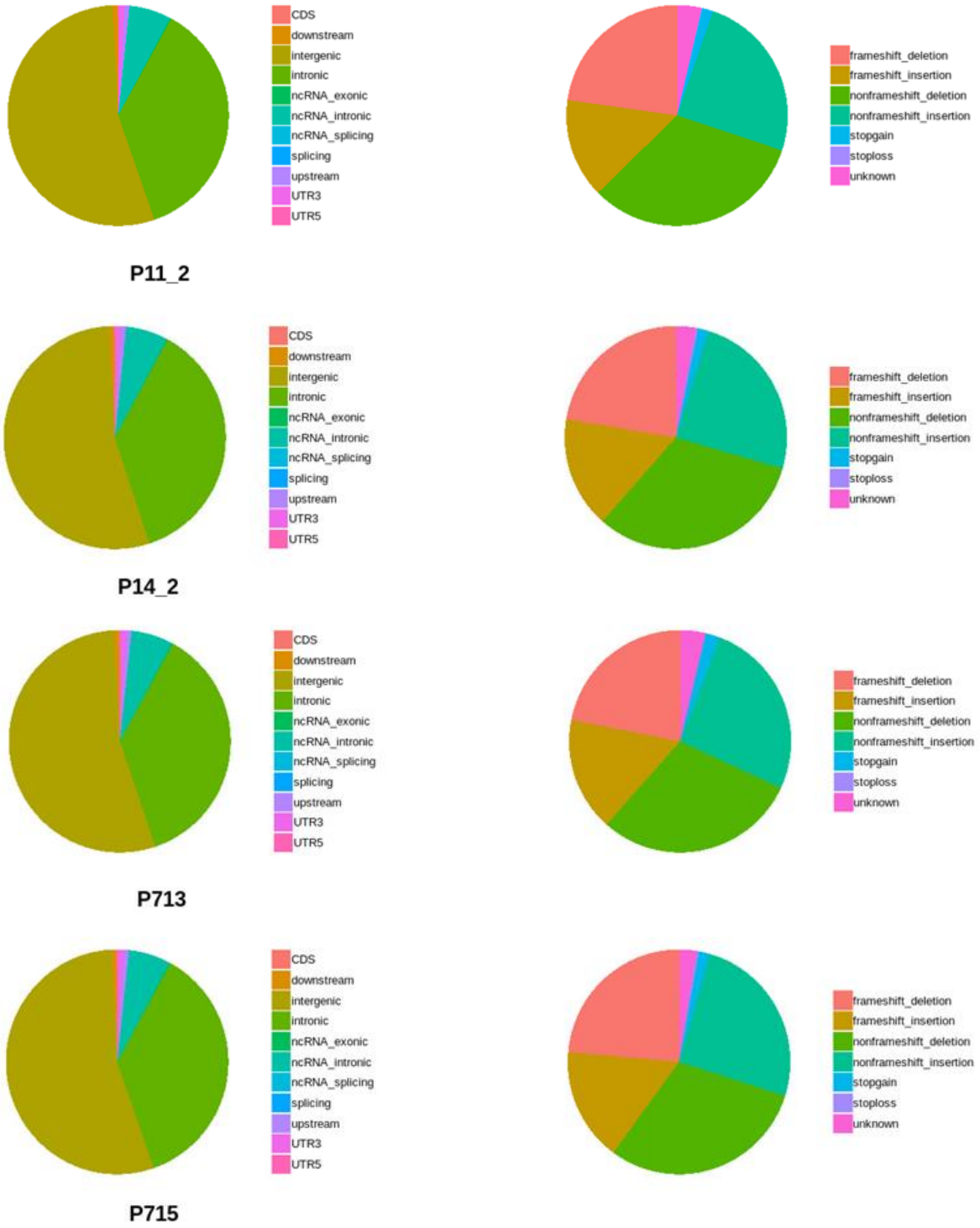


Figura 21. Clasificación de los InDels encontrados en los cuatro genomas. Número de diferentes tipos de InDels en regiones codificantes de P11_2, P14_2, P713 y P715.

Tabla 15. SNPs identificados en los genomas

Nombre de la muestra	P11_2	P14_2	P713	P715
CDS	22524	24297	22192	22395
SNP sinónimo	11442	12253	11195	11271
SNP con cambio de sentido	10716	11620	10627	10771
Ganancia del codón de parada	96	112	106	95
Pérdida del codón de parada	10	12	11	12
Desconocido	271	315	265	257
Intrónico	1191674	1300685	1198239	1183876
UTR3	24936	27177	25148	24895
UTR5	5690	6179	5735	5832
Corte y empalme	84	90	91	82
ARNnc exónico	13142	14183	13158	13047
ARNnc intrónico	208872	225666	207288	205610
ARNnc de corte y empalme	72	77	71	67
aguas arriba	21913	23724	21929	21932
aguas abajo	23101	25002	23081	23028
Intergénico	1997479	2158378	2006088	1991547
Total	3510451	3806520	3524031	3493351

Leyenda:

1) Nombre de la muestra: código identificativo de la muestra.

2) CDS: el número de los SNP en la región codificante.

SNP sinónimo: un cambio en un solo nucleótido que no causa un cambio en los aminoácidos

SNP con cambio de sentido: un cambio en un solo nucleótido que causa un cambio en los aminoácidos

Ganancia del codón de parada: un SNP no sinónimo que conduce a la creación del codón de parada en el sitio de la variante

Pérdida del codón de parada: un SNP no sinónimo que conduce a la eliminación del codón de parada en el sitio de la variante

Desconocido: función desconocida (debida a varios errores en la definición de la estructura del gen en el archivo de la base de datos).

3) Intrónico: El número de SNPs en la región intrónica.

4) UTR3: el número de los SNPs en la región de 3'UTR.

5) UTR 5: el número de los SNPs en la región de 5'UTR.

6) Corte y empalme: El número de SNPs dentro de 2 pb de un sitio de unión de corte y empalme.

7) ARNnc exónico: el número de los SNPs en la región exónica de los ARNs no codificantes.

8) ARNnc intrónico: El número de SNPs en la región intrónica de ARNs no codificantes.

9) ARNnc de corte y empalme: El número de SNPs dentro de 2-pb de la unión de corte y empalme de ARNs no codificantes.

10) Aguas arriba: El número de SNPs distante 1 kb del sitio de inicio de la transcripción.

11) Aguas abajo: El número de SNPs distante 1 kb del sitio de terminación de la transcripción.

12) Intergénico: El número de SNPs en la región intergénica.

13) Total: El número total de SNPs.

En la **Tabla 16** el total de SNP encontrado en cada genoma, cuántos son heterocigóticos, homocigóticos, si hay transiciones, transversiones y una ratio (razón/proporción) de transiciones/transversiones.

Tabla 16. Características de los SNPs.

Nombre de la muestra	P11_2	P14_2	P713	P715
Total	3510451	3806520	3524031	3493351
Het	1693698	2210294	1695538	1664789
Hom	1816753	1596226	1828493	1828562
Transición	2345485	2542387	2353632	2333850
Transversión	1164966	1264133	1170399	1159501
ts/tv	2,01	2,01	2,01	2,01
Porcentaje de dbSNP	3411527 (97.18%)	3707424 (97.40%)	3426531 (97.23%)	3398112 (97.27%)
novel	98924	99096	97500	95239
novel ts	53790	53511	52809	51546
novel tv	45134	45585	44691	43693
novel ts/tv	1,19	1,17	1,18	1,18

Leyenda:

1) Nombre de la muestra: código identificativo de la muestra.

2) Total: el número total de los SNPs.

3) Het: el número de los heterocigotos.

4) Hom: el número de los homocigotos.

5) Transición (ts): el número de transiciones.

6) Transversión (tv): el número de transversiones.

7) Ts/tv: el número de transiciones dividido por el número de transversiones.

8) Porcentaje de dbSNP: El número de SNPs que han sido reportados en la base de datos del dbSNP dividido entre el número total de SNPs.

9) Novedoso: El número de SNPs no reportado en la base de datos del dbSNP.

10) ts novedosas: El número de transiciones de SNPs que no han sido reportados en la base de datos del dbSNP.

11) tv novedosas: El número de transversiones de SNPs que no han sido reportados en la base de datos del dbSNP.

12) ts /tv novedosas: ts novedosas dividido entre tv novedosas.

La **Figura 22** muestra los conteos de SNPs para cada genoma, en todas sus variantes. Como se observa, los SNPs que evidencian un cambio de sentido constituyen poco más de la mitad, los sinónimos, prácticamente la otra mitad, los desconocidos, las ganancias y pérdidas de codones de terminación son minorías.

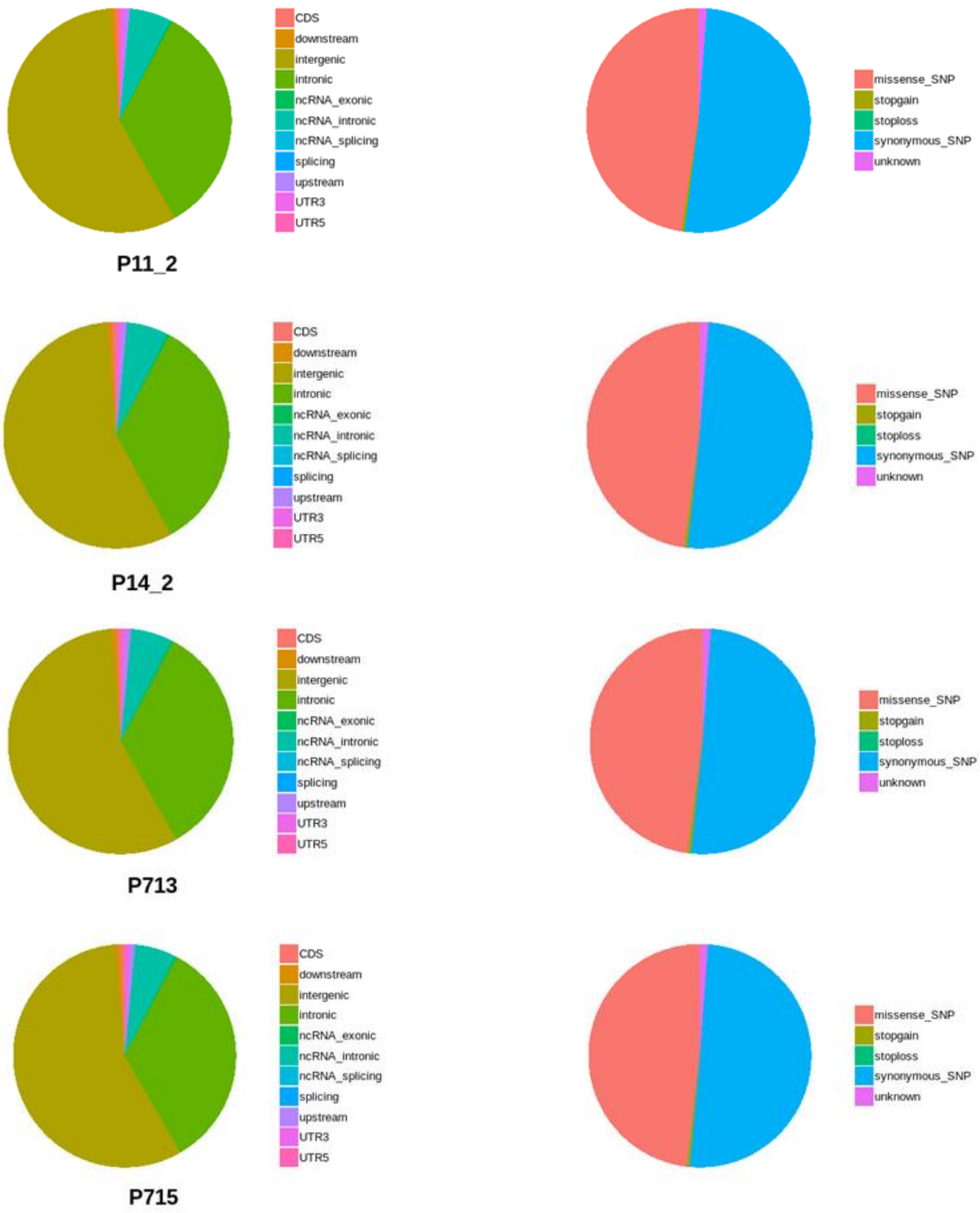


Figura 22. Clasificación de los SNPs encontrados en los cuatro genomas. Número de diferentes tipos de SNPs en los genomas P11-2, P14_2, P713 y P715.

3.3.4 Detección de CNV *(Copy Number Varation)*

La **Tabla 17** muestra la cantidad total de los marcadores "CNV" que representa variaciones en número de copias que se identificaron en cada genoma. Los CNVs corresponden a ganancia de copias en términos de las regiones analizadas, ganancia en tamaños, o ganancia o pérdida en término de tamaños o en número de regiones analizadas. El conteo de ganancias resulta ser en general mayor que el de las pérdidas y que el tamaño de ganancia también es mayor que el de la pérdida.

Tabla 17. Resultado de la detección de CNV

Nombre de la muestra	Conteo de ganancias	Tamaño de ganancias	Conteo de pérdidas	Tamaño de pérdidas	Conteo total	Tamaño total
P11_2	149	6039000	97	1579000	246	7618000
P14_2	244	6664000	146	1606000	390	8270000
P713	206	2670000	154	2210000	360	4880000
P715	192	3886000	134	1399000	326	5285000

Leyenda:

1) Nombre de la muestra: código identificativo de la muestra.
2) Conteo de ganancias: el número de ganancias.
3) Tamaño de ganancia: el tamaño total de las ganancias.
4) Conteo de pérdidas: el número de pérdidas.
5) Tamaño de pérdidas: el tamaño total de las pérdidas.
6) Conteo total: el número total de los CNV.
7) Tamaño total: el tamaño total de los CNV.

La **Figura 23** muestra el tamaño de regiones genómicas afectadas por los CNV en cada genoma, tanto en término de ganancias como en término de pérdidas de ADN. Resulta que, en general, las primeras son mayores que las segundas, lo cual nos indica diferencias en los genomas.

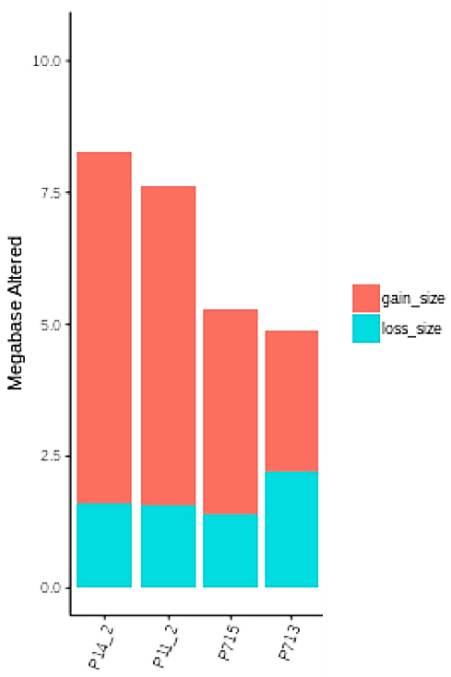
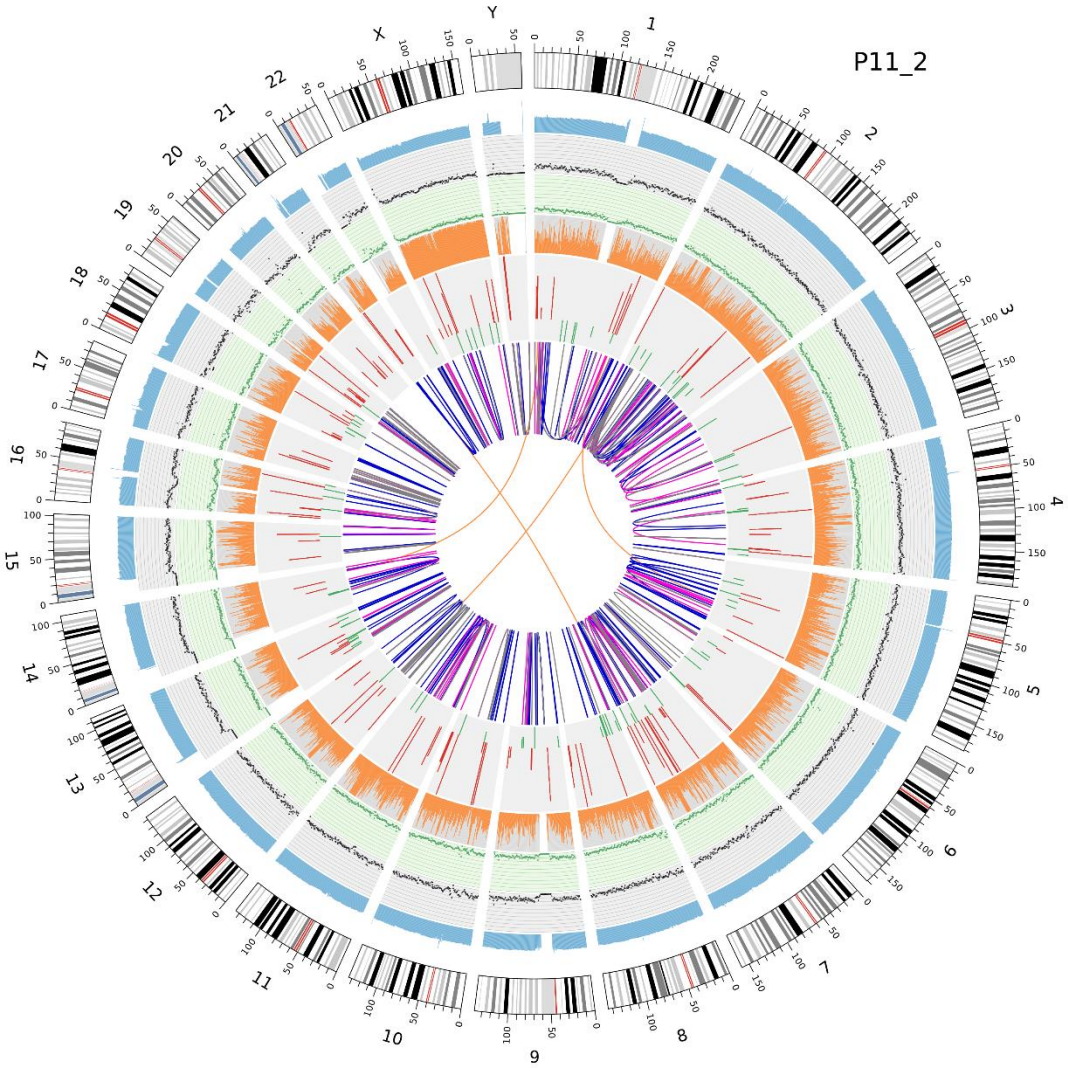
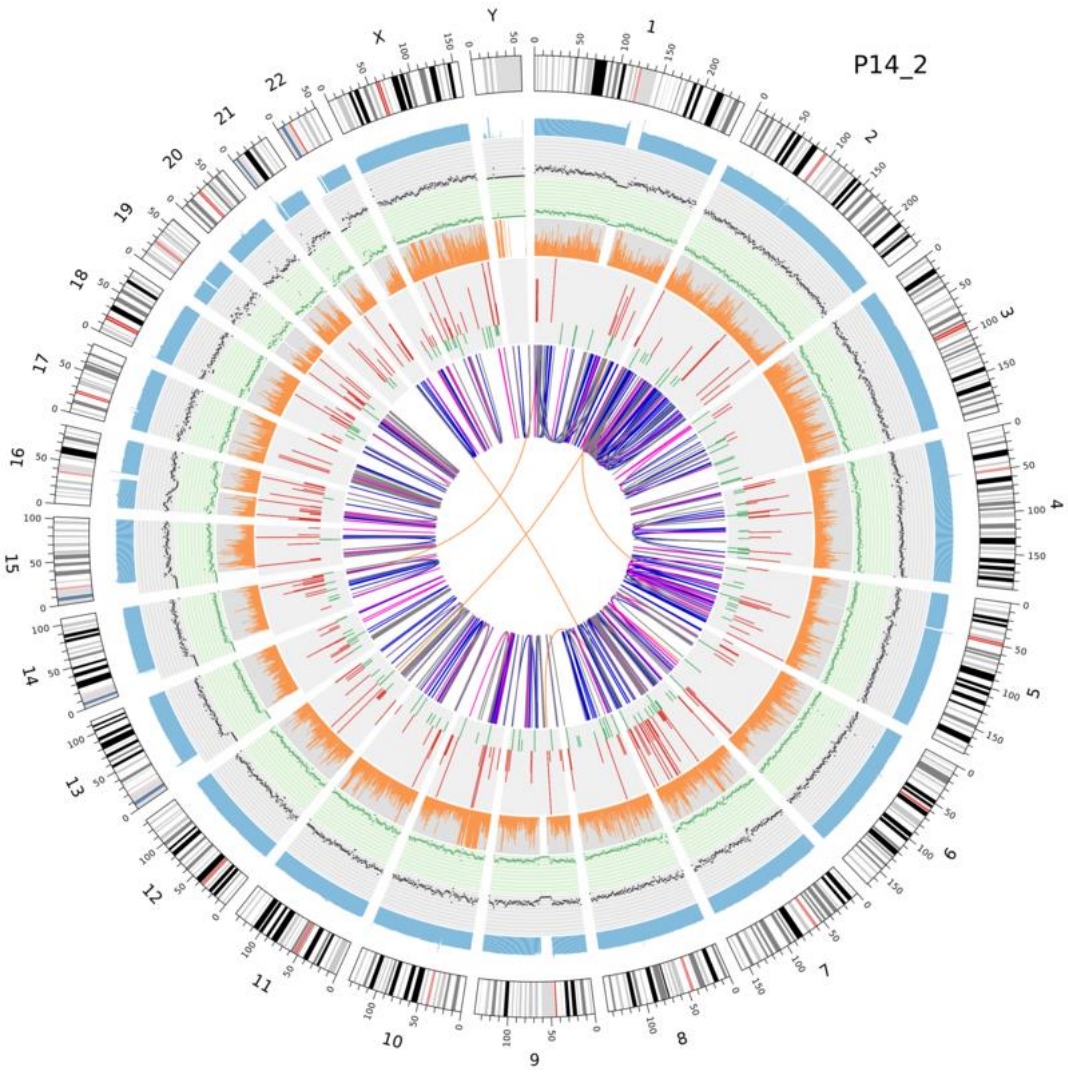


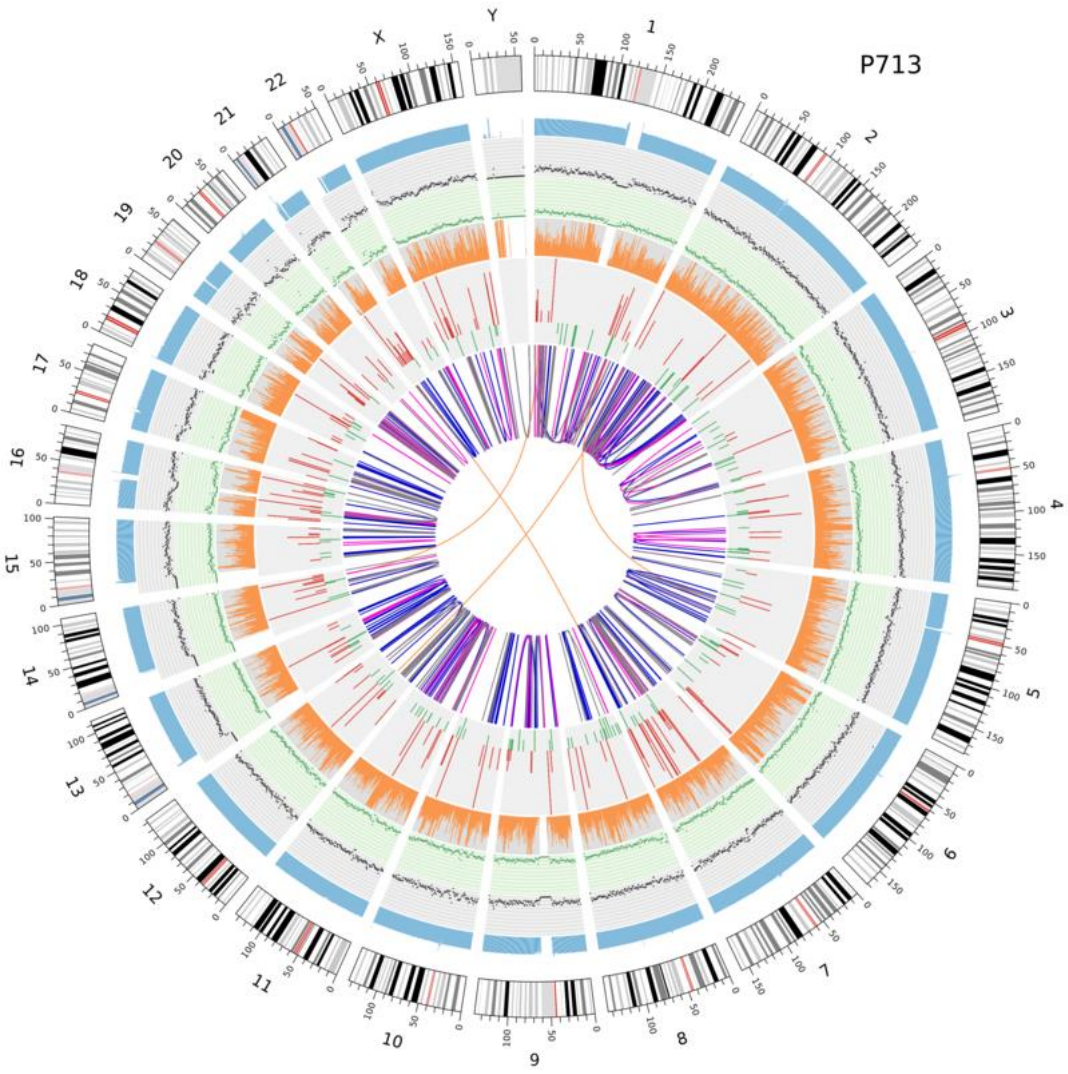
Figura 23. El tamaño de regiones genómicas afectadas por los CNV en cada genoma. El eje X representa el nombre de las muestras y el eje Y representa el tamaño total de las regiones genómicas afectadas por ganancias y pérdidas (Mb).

Los resultados de detección de CNV se visualiza también en las gráficas de "Circos" (**Figura 24**). Hay una gráfica de Circos para cada genoma, donde se observa en la leyenda del anillo más externo los 23 cromosomas numerados en orden alrededor de los anillos del Circos desde el cromosoma 1 hasta los cromosomas sexuales. En el segundo anillo, se ve en azul la cobertura que se logró en todos los cromosomas de cada genoma específico. Como el *cromosoma 1* es metacéntrico, la región donde no hubo cobertura corresponde con gran probabilidad al centrómero, caracterizado por una elevada cantidad de ADN repetitivo, el cual es más difícil de secuenciar. Es una región mayormente no codificante y, por lo tanto, menos relevante en términos de presencia de genes. Lo mismo se repite en los cromosomas 5, 9, 16 y 19, en los que la región en blanco representa una región de menor cobertura que corresponde al centrómero, solo que como los cromosomas 5 y 9 son submetacéntricos. Esta región no se encuentra en el centro. También las regiones con huecos pequeños de los cromosomas 21 y 22 son muy probablemente centrómeros o algunas regiones del telómero, el cual también es una región mayormente de DNA repetitivo. En el genoma P11_2, se dio una cobertura bastante alta del *cromosoma Y*, mientras que en los otros tres la cobertura fue más baja porque es conocido que este cromosoma presenta gran cantidad de DNA repetitivo, difícil de secuenciar. Sin embargo,

la muy baja cobertura en dos de las muestras secuenciadas confirma que las mismas corresponden a mujeres.




P14_2


P713

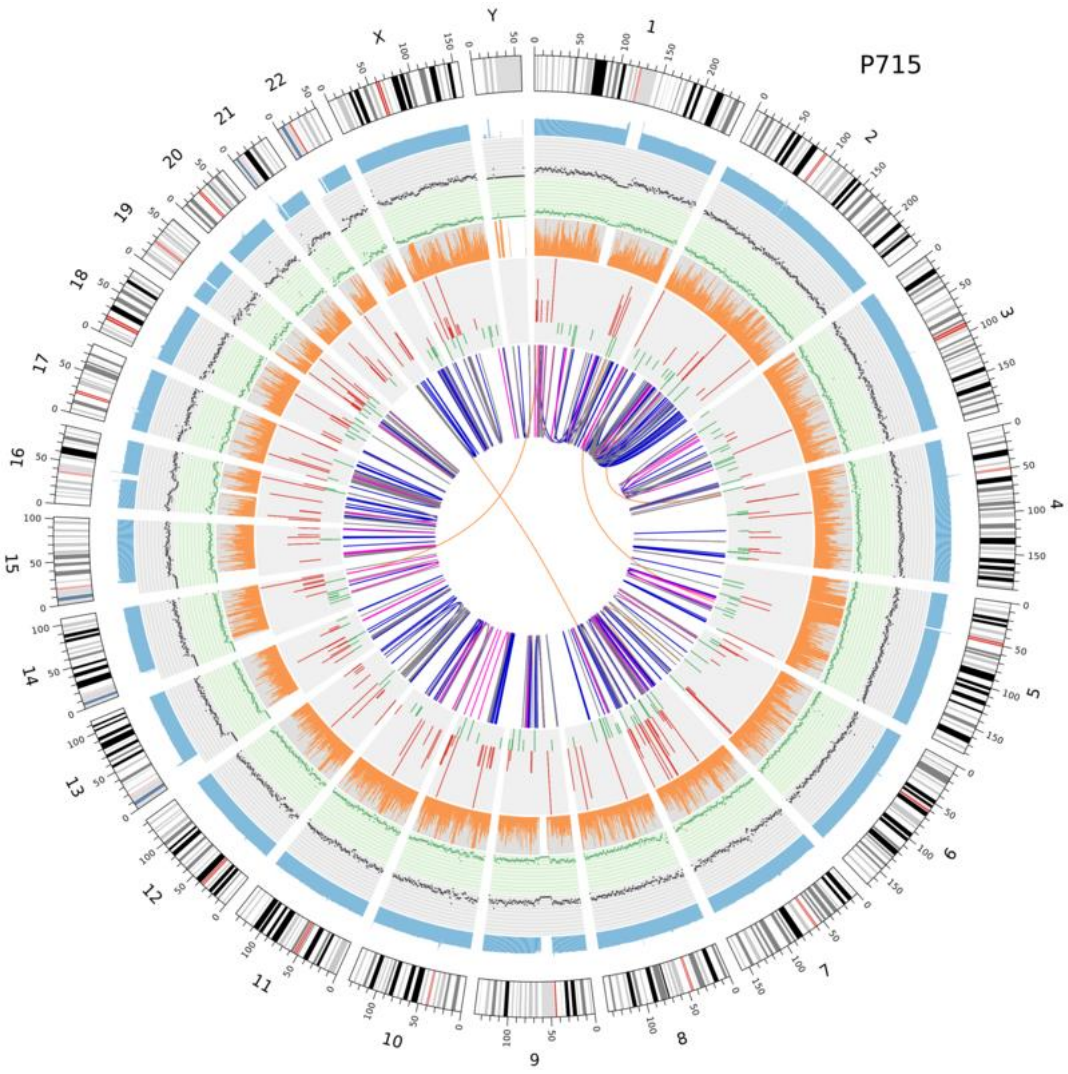


Figura 24. CNVs representados en gráfica de "Circos". La figura consta de siete anillos contados desde el exterior hacia el interior. A continuación, se define el significado de cada anillo:

1) El primer anillo, es decir, el círculo concéntrico más externo, es la información sobre el cromosoma.
2) El segundo anillo representa a manera de histograma la cobertura de lectura. Un histograma es la cobertura promedio de una región de 0.5 Mpb.
3) El tercer anillo representa a manera de gráfico de dispersión la densidad de InDel. Un punto negro es calculado como un número de InDel en un rango de 1 Mpb.
4) El cuarto anillo representa a manera de gráfico de dispersión la densidad de SNP. Un punto verde es calculado como un número de SNP en un rango de 1 Mpb.

5) El quinto anillo representa a manera de histograma la proporción de SNP homocigótico (anaranjado) y del SNP heterocigótico (gris). Un histograma es calculado de una región de 1 Mpb.
6) El sexto anillo representa la inferencia de CNV. Rojo significa ganancia y verde significa pérdida.
7) El séptimo anillo, o sea, el más central, representa la inferencia de SV en regiones exónicas y en las de corte y empalme. Si SV se llama usando *breakdancer* o crest, entonces CTX (anaranjado), INS (verde), DEL (gris), ITX (rosado) e INV (azul). Si SV se llama usando *delly*, entonces TRA (anaranjado), INS (verde), DEL (gris), DUP (rosado) e INV (azul).

4. Selección de variantes génicas candidatas

Se realizaron varios análisis para la búsqueda e identificación de variantes/polimorfismos génicos candidatos. Inicialmente, se seleccionaron 28 incluyendo SNPs e INDELs. Una variante del gen EOMES (Eomesodermina)fue encontrada entre los datos de los INDELs; NUDT15 fue encontrada entre los datos de ambos; y el resto de los genes fueron entre los datos de los SNPs.

Tabla 18. Variantes/polimosfirmos génicos candidatos seleccionados preliminarmente (SNPs e InDels).

SNP					
Cromosoma	**SNP-ID**	**SNP-Referencia**	**SNP-Mutado**	**Gen**	**Significancia Clínica (ENT)**
chr1		A	G	DPYD	Deficiencia de dihidropirimidina dehidrogenasa
chr1	rs1801133	C	T	MTHFR	Neoplasma estomacal, tumor del estroma gastrointestinal, deficiencia de MTHFR
chr1		G	A	SDC3	Obesidad
chr1		T	C	EPHX1	Polimorfismo epóxido hidrolasa
chr1		C	T	LRP8	Infarto del miocardio 1
chr2		C	A	TTN	no especificado
chr2		A	C	TTN	Distrofía muscular Limb-girdle tipo 2, cardiomiopatía dilatada 1G
chr2		T	C	TTN	no especificado
chr3		G	T	GHRL	Síndrome metabólico, Edad del inicio de la obesidad
chr3		T	A	GHRL	Obesidad
chr4		G	T	ADD1	Hipertensión
chr5		G	A	FGFR4	Progresión del cáncer y motilidad de las células tumorales
chr5		A	G	MTRR	Síndrome de Down, defectos del tubo neural, tumor del estroma gastrointestinal
chr6		A	G	SOD2	Complicaciones microvasculares de diabetes
chr6		C	A	LTA	Infarto del miocardio
chr7		C	A	PPP1R3A	Resistencia a la insulina
chr7		A	G	PRSS1	Pancreatitis hereditaria
chr8		G	A	EPHX2	Hipercolesterolemia familiar
chr8		C	T	SLC3OA8	Diabetes mellitus tipo 2
chr9		C	A	AQP7	Defectivo en la secreción de glicerol durante el ejercicio
chr10		T	C	STOX1	Preclampsia/eclampsia 4
chr11		A	G	GSTP1	Neoplasma colorrectal
chr12		G	A	OAS1	Diabetes mellitus tipo 1
chr13		C	T	NUDT15	Metabolismo pobre de las tiopurinas
chr13		C	T	IRS2	Diabetes tipo II
chr16		A	G	IL4R	Síndrome de Inmunodeficiencia Adquirida, Atopia
chr16		G	A	NQO1	Cáncer de pulmón, Leucemia post quimioterapia
chr17		C	G	TP53	Síndrome Li-Fraumeni, cánceres hereditarios
chr17		T	C	AKAP10	Defecto de conducción cardíaca
chrX		C	T	DMD	Distrofía muscular de Duchenne, Cardiomiopatia dilatada
INDEL					
Cromosoma	**INDEL-ID**	**INDEL Referencia**	**INDEL Mutado**	**Gen**	**Significancia Clínica (ENT)**
chr3		G	GCGGCGC	EOMES	No especificado
chr13		A	AGGAGTC	NUDT15	Pobre metabolismo de tiopurinas

Esta tabla resume la información sobre las variantes génicas candidatas identificadas incluyendo SNPs e INDELs en cuanto al cromosoma en

donde están localizadas, el ID del SNP, el alelo silvestre, el alelo mutante, el gen y las enfermedades a las cuales están asociadas para los primeros; el cromosoma donde están localizadas, el ID del INDEL, el alelo silvestre, el alelo mutante, el gen y las enfermedades a las cuales están asociadas para los segundos.

Tabla 19. Variantes génicas candidatas finales con su información (sólo SNPs)

Cromosoma	SNP-ID	SNP-Referenci	SNP-Mutado	Gen	Significancia Clínica (ENT)
chr1	rs1801133	C	T	MTHFR	Neoplasma estomacal, tumor del estroma gastrointestinal, deficiencia de MTHFR
chr1		G	A	SDC3	Obesidad
chr1		T	C	EPHX1	Polimorfismo epóxido hidrolasa
chr3		G	T	GHRL	Síndrome metabólico, Edad del inicio de la obesidad
chr3		T	A	GHRL	Obesidad
chr5		G	A	FGFR4	Progresión del cáncer y motilidad de las células tumorales
chr6		A	G	SOD2	Complicaciones microvasculares de diabetes
chr7		C	A	PPP1R3A	Resistencia a la insulina
chr8		C	T	SLC3OA8	Diabetes mellitus tipo 2
chr12		G	A	OAS1	Diabetes mellitus tipo 1
chr13		C	T	IRS2	Diabetes tipo II
chr16		A	G	IL4R	Síndrome de Inmunodeficiencia Adquirida, Atopia
chr16		G	A	NQO1	Cáncer de pulmón, Leucemia post quimioterapia
chr17		C	G	TP53	Síndrome Li-Fraumeni, cánceres hereditarios

Esta tabla resume la información sobre las variantes génicas candidatas en cuanto al cromosoma en donde están localizadas, el ID del SNP, el alelo silvestre, el alelo mutante, el gen y las enfermedades (ENT) a las cuales están vinculadas. En la primera fila se muestra el SNP del gen MTHFR, el cual fue seleccionado para estudios más amplios de genotipado en la población Ngöbe.

5. Amplificaciones del gen MTHFR

5.1 Primera amplificación del gen MTHFR (optimización 1)

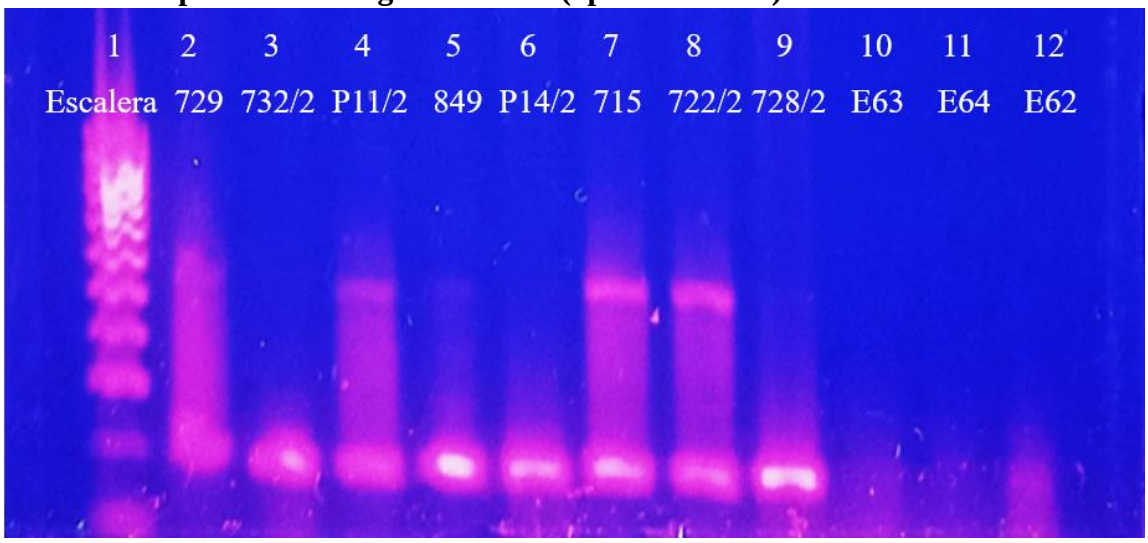


Figura 25. Electroforesis en gel de agarosa de la primera PCR de optimización del gen MTHFR en la región con la variante SNP rs1801133. De izquierda a derecha: 729, 732/2, P11/2, 849, P14/2, 715, 722/2, 728/2, E63, E64 y E62. Como se esperaba, la banda observada corresponde a unas 198pb.

5.2 Segunda amplificación del gen MTHFR (optimización 2)

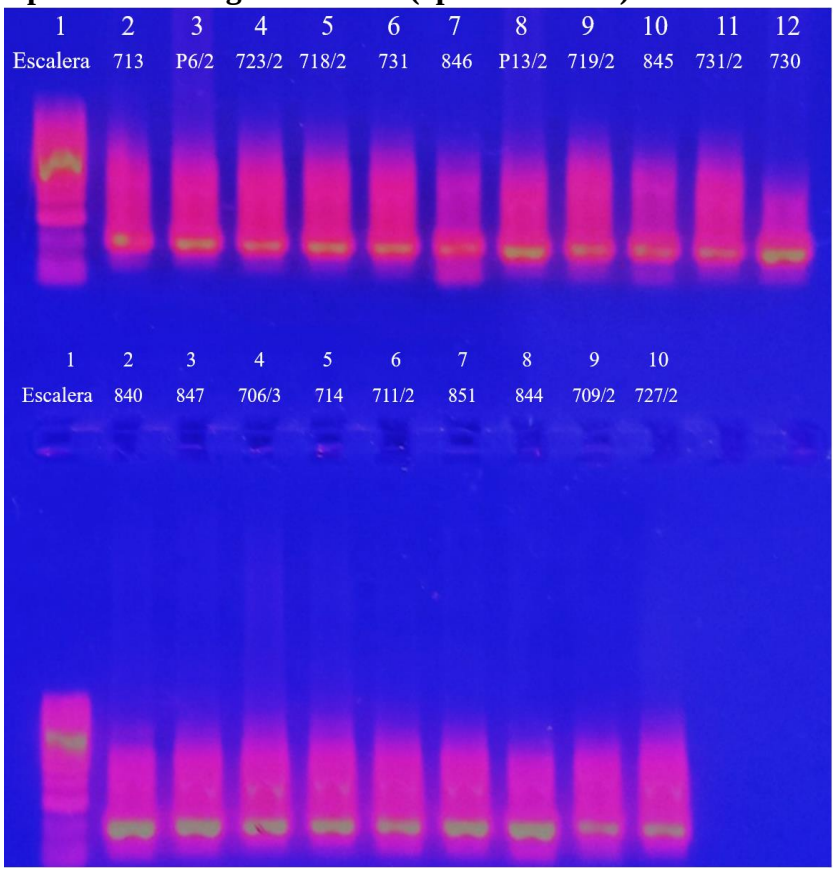


Figura 26. Electroforesis en gel de agarosa de un segundo grupo de muestras analizadas por PCR del gen MTHFR variante rs1801133. Arriba de izquierda a derecha: 713, P6/2, 723/2, 718/2, 731, 846, P13/2, 719/2, 845, 731/2 y 730. Debajo de izquierda a derecha: 840, 847, 706/3, 714, 711/2, 851, 844, 709/2 y 727/2.

5.3 Tercera amplificación del gen MTHFR (optimización final)

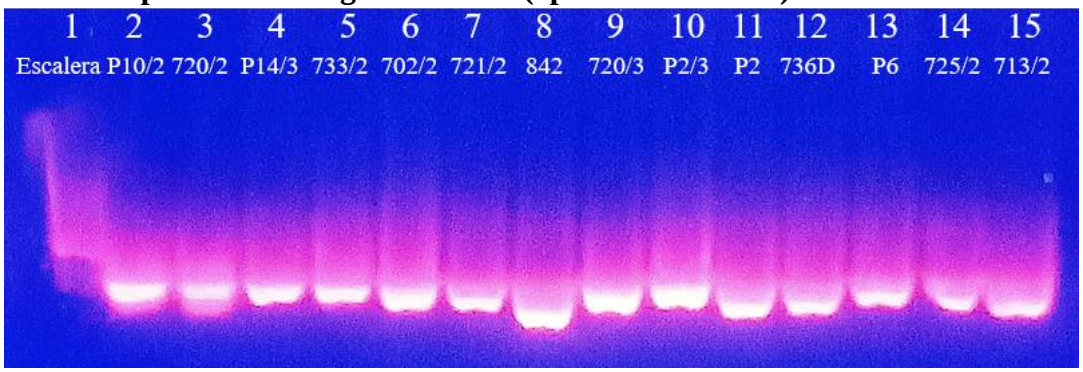


Figura 27. Electroforesis en gel de agarosa de un tercer grupo de muestras analizadas por PCR del gen MTHFR variante rs1801133. De izquierda a derecha: P10/2, 720/2, P14/3, 733/2, 702/2, 721/2, 842, 720/3, P2/3, P2, 736D, P6, 725/2 y 713/2.

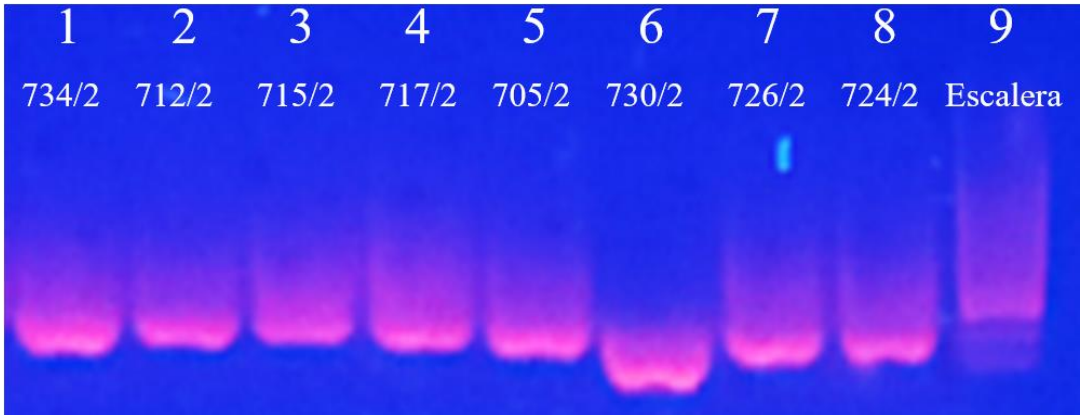


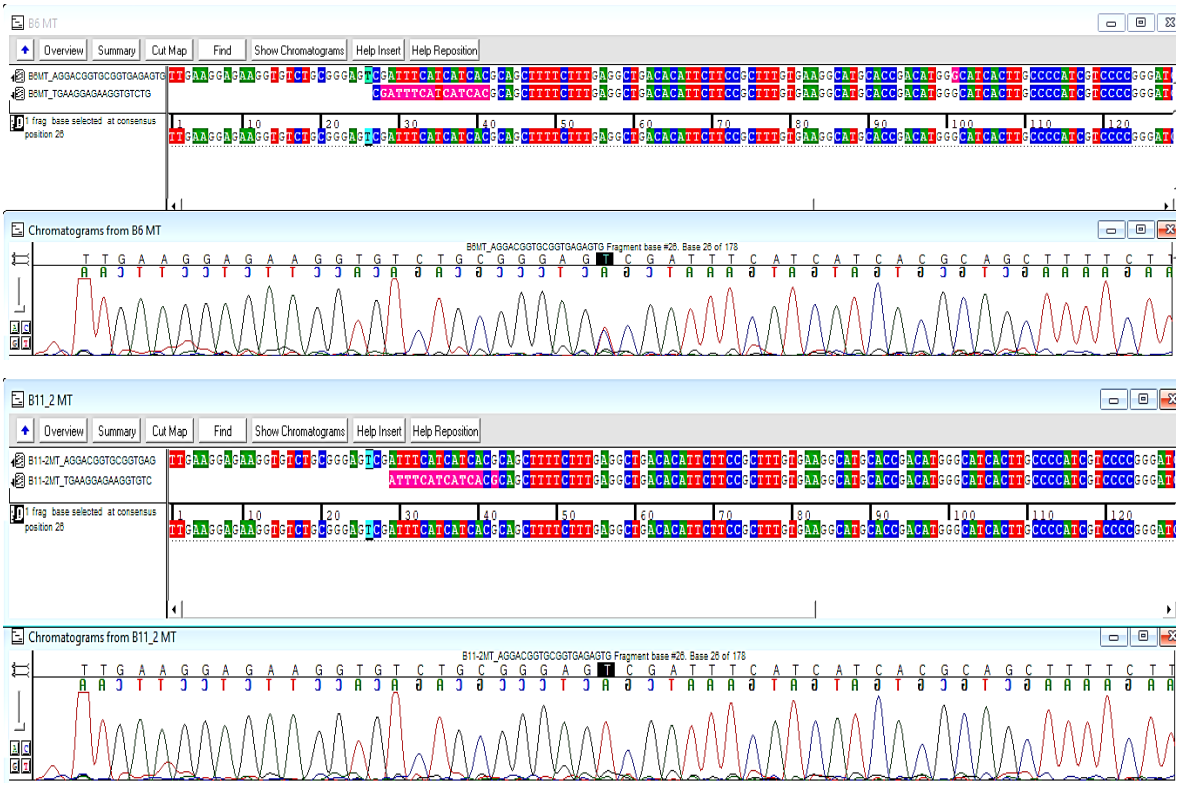
Figura 28. Electroforesis en gel de agarosa de un tercer grupo de muestras analizadas por PCR del gen MTHFR variante rs1801133. De izquierda a derecha: 734/2, 712/2, 715/2, 717/2, 705/2, 730/2, 726/2 y 724/2. Este grupo de muestras se corrió en otra gel, por eso se muestran aparte del grupo de la **Figura 27**.

6. Análisis de secuencias para el genotipaje

6.1 Alineamiento de hebras, limpieza, obtención de secuencias consenso y verificación tamaño

La confirmación del genotipo del polimorfismo en la posición del SNP rs1801133 se pudo verificar mediante visualización y análisis de las secuencias en el programa *Sequencher*. En los heterocigotos, apareció como un doble pico que representaba la base silvestre y la base mutada, la C y la T, mientras que en los homocigotos silvestres o mutantes se presentó como un único

pico correspondiente a la base respectiva C o T (**Figura 29**).



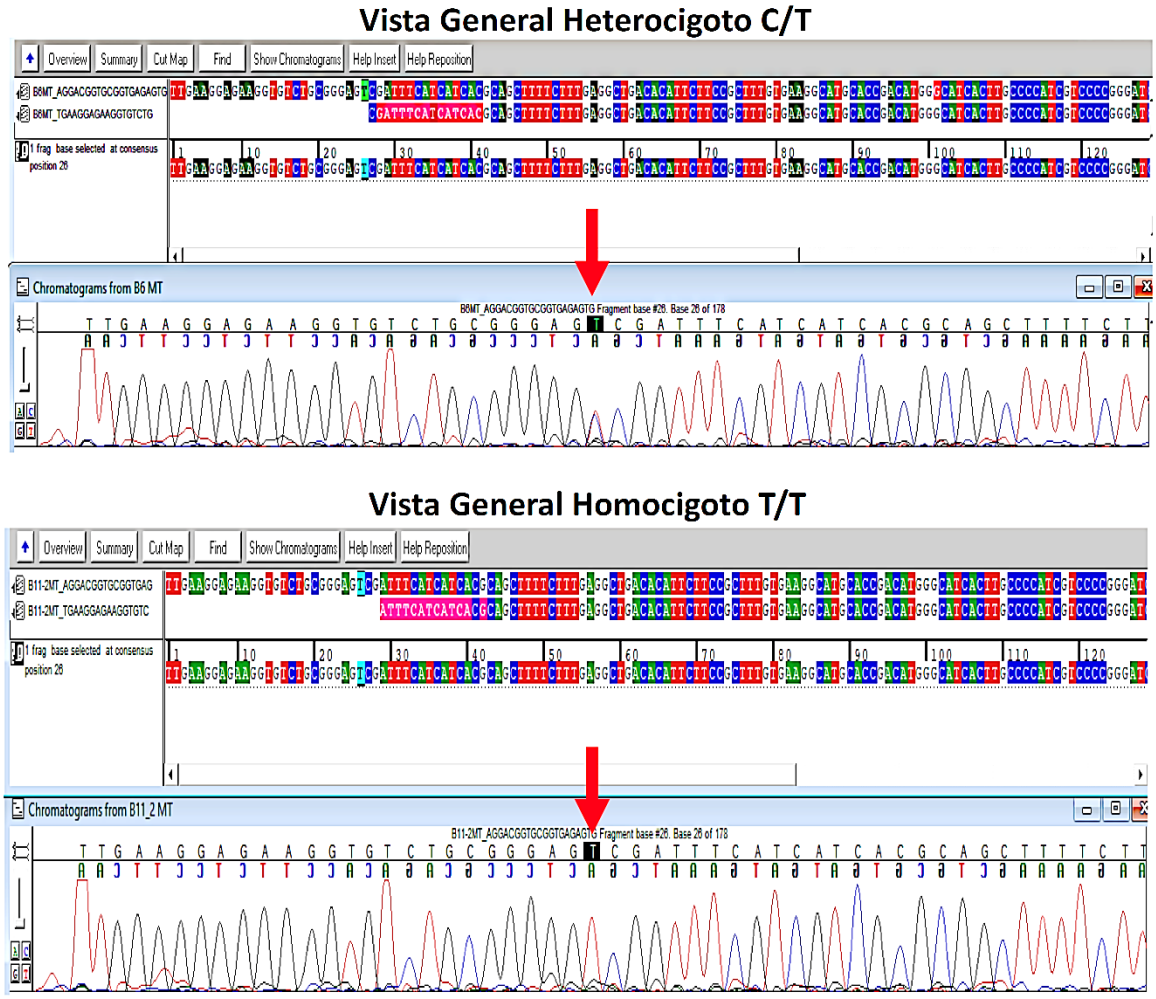


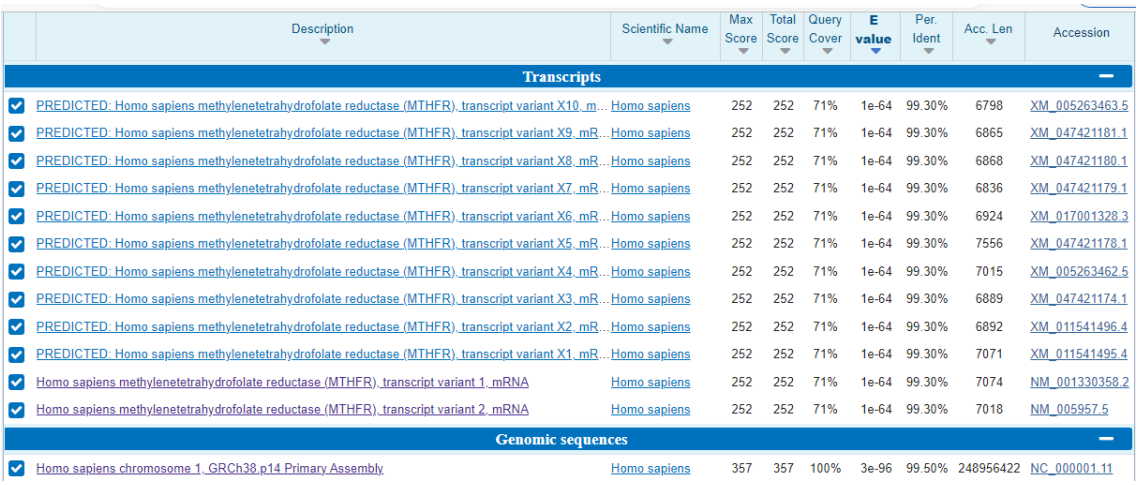
Figura 29. Imágenes representativas con la posición del SNP en Sequencher. Imágenes representativas confirmando el genotipo de la posición del SNP en el programa Sequencher evidenciado en dos de las 55 muestras. **Arriba**, heterocigoto C/T, se observa picos dobles en la posición del SNP (flecha). **Abajo**, homocigoto T/T, se observa un solo pico en la posición del SNP (flecha). No se han observado homocigotos C/C, por lo que no se presentan imágenes de esos.

El alineamiento de las hebras generadas con cada cebador (*forward*-FW y *reverse*-RV) permitió verificar la certeza molecular de cada posición de cada base y proseguir con la limpieza de las secuencias. De esta manera, se eliminaron los segmentos cortos de los extremos con errores o secuencias no específicas y se corrigieron bases con ambigüedades. Luego de la limpieza, se obtuvieron las secuencias consenso de cada muestra. Para los homocigotos, se obtuvo sólo una secuencia consenso porque tenían un solo pico correspondiente al alelo que era igual en ambas hebras FW y RV. Para los heterocigotos, se obtuvieron dos (2) secuencias consenso, una para cada alelo, correspondiente a cada pico doble en la posición del SNP, es decir, una secuencia

consenso para el alelo de tipo silvestre y la otra, para el alelo mutante. Estos análisis también permitieron descartar las secuencias de las muestras B2_3 (B2_3MT), B836 (B836MT) y B850 (B850MT) porque fueron de mala calidad con cromatogramas que mostraban demasiados picos ambiguos y superpuestos a lo largo de la secuencia, por lo que el análisis se continuó con las otras 55 muestras de excelente calidad.

6.2 Confirmación de la identidad molecular de las secuencias de ADN y proteínas en BLAST

La identidad molecular de las secuencias consenso obtenidas fue confirmada mediante análisis de BLAST con la base de datos NCBI-GenBank. Para efectos de comparación de los alineamientos de las secuencias obtenidas, también incluimos en los análisis secuencias de referencia de ADN y proteínas del Genbank correspondiente a la región de interés con el SNP. En esta base de datos, se confirmó que todas las secuencias obtenidas corresponden al gen MTHFR y las secuencias nucleotídicas con mayor porcentaje de similitud de al menos 99.30% (*Percentage of identity*) y un *E-value* de 1e-64 fueron NM_001330358.2 (*Homo sapiens methylenetetrahydrofolate reductase* (MTHFR), transcript variant 1, mRNA) y NM_005957.5, (*Homo sapiens methylenetetrahydrofolate reductase* (MTHFR), transcript variant 2, mRNA). No se tomó en cuenta las secuencias identificadas como *PREDICTED*, porque no eran secuencias de referencia del genoma humano. Luego, se buscó la secuencia de proteína utilizando el *protein ID* que aparecía en la descripción. La variante del transcripto 1 de MTHFR codifica para la proteína NP_001317287.1 (methylenetetrahydrofolate reductase (NADPH) isoform 1 ([*Homo sapiens*]), que es la isoforma más larga; mientras que la variante del transcripto 2 de MTHFR codifica para la proteína NP_005948.3 (methylenetetrahydrofolate reductase (NADPH) isoform 2 ([*Homo sapiens*]), que es la isoforma más corta (**Figura 30**). En ambas variantes, se presenta un cambio de C>T en el exón 5, así como un cambio correspondiente de alanina a valina, pero este resulta en diferentes posiciones: para la variante del transcripto 1, es un cambio de C>T en la posición 788 del CDS y de A por V en la posición 263; para la variante del transcripto 2, se da un cambio de C>T en la posición 665 del CDS y de A a V en la posición 222 (ClinVar, 2023). Los pasos siguientes del análisis del genotipaje ayudaron a verificar con cuál de las dos variantes de los transcritos se identifican las secuencias.



	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
Transcripts									
☑	PREDICTED: Homo sapiens methylenetetrahydrofolate reductase (MTHFR), transcript variant X10, m...	Homo sapiens	252	252	71%	1e-64	99.30%	6798	XM_005263463.5
☑	PREDICTED: Homo sapiens methylenetetrahydrofolate reductase (MTHFR), transcript variant X9, mR...	Homo sapiens	252	252	71%	1e-64	99.30%	6865	XM_047421181.1
☑	PREDICTED: Homo sapiens methylenetetrahydrofolate reductase (MTHFR), transcript variant X8, mR...	Homo sapiens	252	252	71%	1e-64	99.30%	6868	XM_047421180.1
☑	PREDICTED: Homo sapiens methylenetetrahydrofolate reductase (MTHFR), transcript variant X7, mR...	Homo sapiens	252	252	71%	1e-64	99.30%	6836	XM_047421179.1
☑	PREDICTED: Homo sapiens methylenetetrahydrofolate reductase (MTHFR), transcript variant X6, mR...	Homo sapiens	252	252	71%	1e-64	99.30%	6924	XM_017001328.3
☑	PREDICTED: Homo sapiens methylenetetrahydrofolate reductase (MTHFR), transcript variant X5, mR...	Homo sapiens	252	252	71%	1e-64	99.30%	7556	XM_047421178.1
☑	PREDICTED: Homo sapiens methylenetetrahydrofolate reductase (MTHFR), transcript variant X4, mR...	Homo sapiens	252	252	71%	1e-64	99.30%	7015	XM_005263462.5
☑	PREDICTED: Homo sapiens methylenetetrahydrofolate reductase (MTHFR), transcript variant X3, mR...	Homo sapiens	252	252	71%	1e-64	99.30%	6889	XM_047421174.1
☑	PREDICTED: Homo sapiens methylenetetrahydrofolate reductase (MTHFR), transcript variant X2, mR...	Homo sapiens	252	252	71%	1e-64	99.30%	6892	XM_011541496.4
☑	PREDICTED: Homo sapiens methylenetetrahydrofolate reductase (MTHFR), transcript variant X1, mR...	Homo sapiens	252	252	71%	1e-64	99.30%	7071	XM_011541495.4
☑	Homo sapiens methylenetetrahydrofolate reductase (MTHFR), transcript variant 1, mRNA	Homo sapiens	252	252	71%	1e-64	99.30%	7074	NM_001330358.2
☑	Homo sapiens methylenetetrahydrofolate reductase (MTHFR), transcript variant 2, mRNA	Homo sapiens	252	252	71%	1e-64	99.30%	7018	NM_005957.5
Genomic sequences									
☑	Homo sapiens chromosome 1, GRCh38.p14 Primary Assembly	Homo sapiens	357	357	100%	3e-96	99.50%	248956422	NC_000001.11

Figura 30. Identificación molecular de secuencias mediante análisis de BLAST en NCBI-*GenBank.* Las secuencias de referencia nucleotídicas en GenBank. Tanto la NM_001330358.2 como la NM_005957.5 tienen el mismo porcentaje de identidad, pero la primera es más larga y la segunda, más corta y codifican, respectivamente para la isoforma más larga y la más corta.

6.3 Análisis de restricción in sílico

Para los análisis de restricción las secuencias, se sometieron al sitio de NEBCutter usando la enzima de restricción HinfI, que no digirió las secuencias de tipo silvestre, pero realizó un corte en la posición 23 en las secuencias mutantes (**Figura 31**).

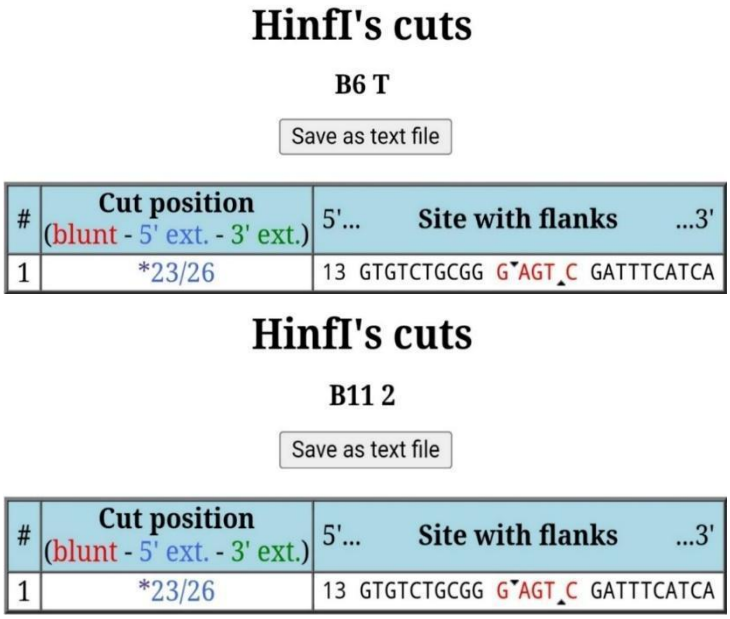


Figura 31. Sitio de corte de restricción de HinFI**.** Sitio de corte de restricción de HinFI en dos muestras representativas de los 55 productos de PCR secuenciados para genotipaje.

6.4 Predicción a proteínas y alineamiento de secuencias nucleotídicas y de aminoácidos

Las secuencias nucleotídicas se alinearon en MEGA con la secuencia de referencia de la variante de transcripción 2 en un file de MEGA (**Figura 32**).

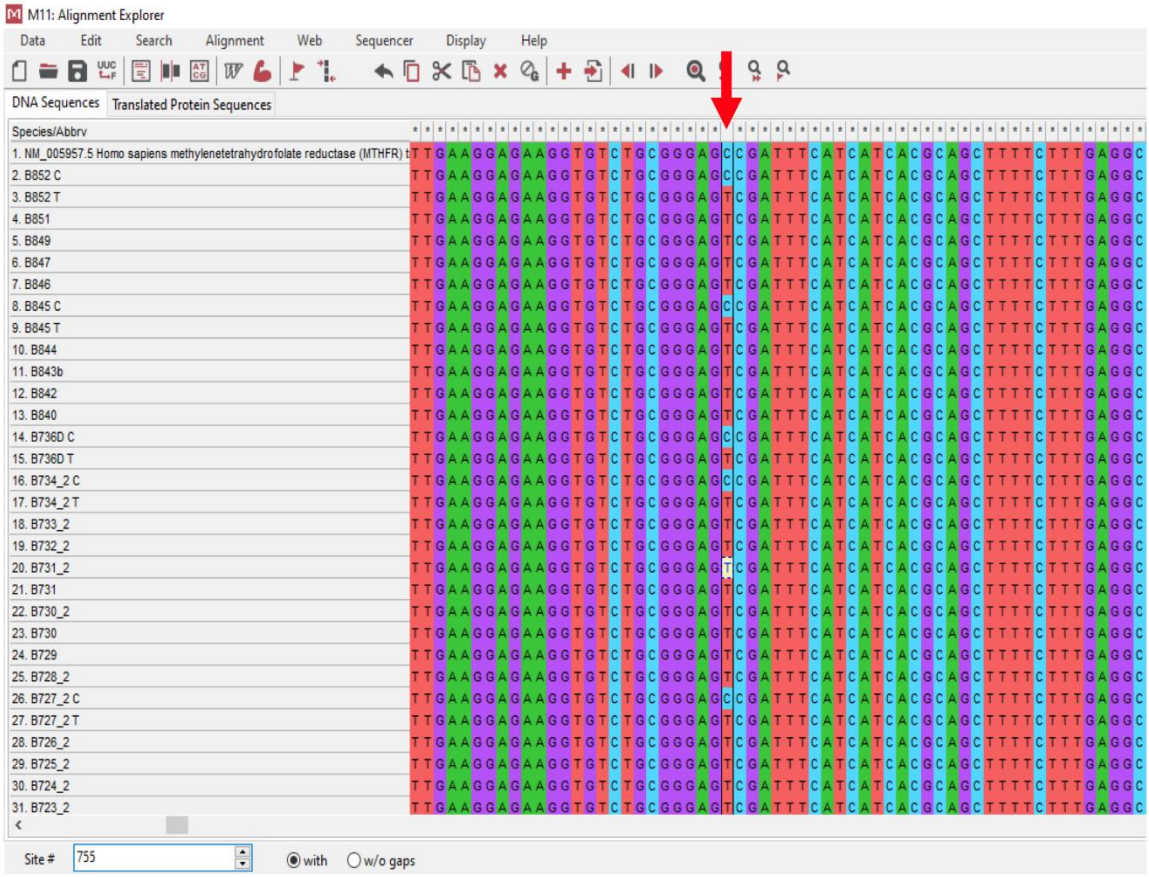


Figura 32. Alineamiento de las secuencias de ADN. Alineamiento de las secuencias de ADN de las muestras con la secuencia de referencia nucleotídica metilentetrahidrofolato reductasa variante 2 del Genbank (arriba, primera secuencia). Se evidencia la posición del SNP señalado con la flecha.

La posición del SNP dentro del alineamiento coincidió con aquella de la base de datos para la variante 2 (**Figura 33**).

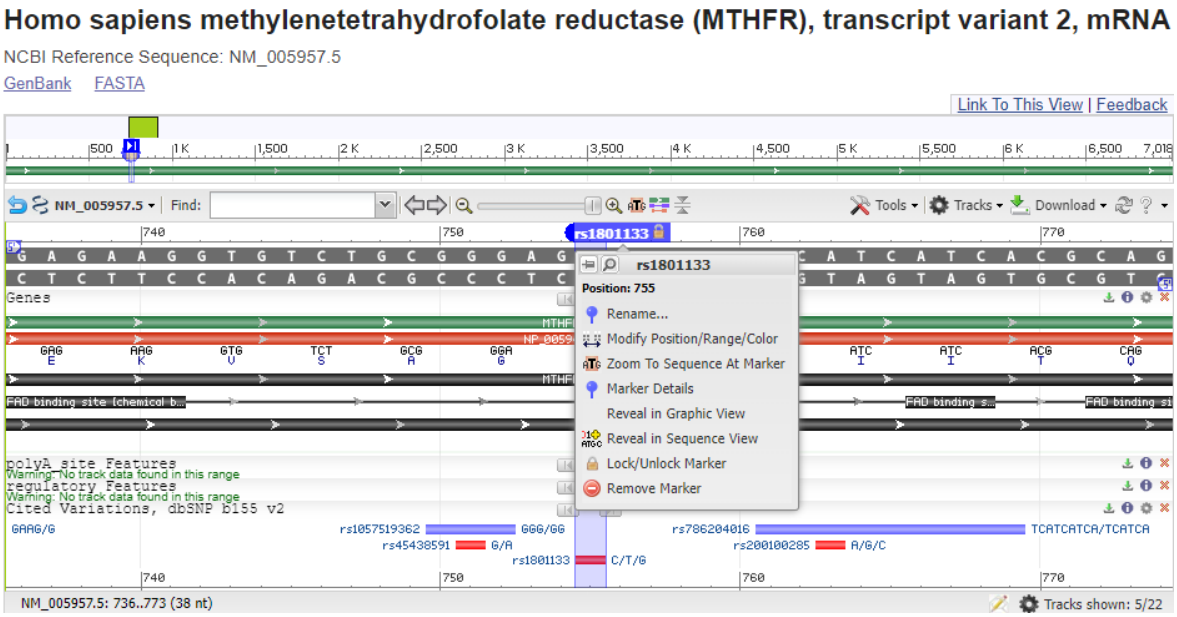


Figura 33. Posición del SNP dentro la variante 2 en GenBank. En Ensembl, se indica que 755 es la posición del SNP dentro del transcrito ENST00000376590.9, que según RefSeqMatch corresponde al transcrito NM_005957.5, y coincide con la posición 655, que es la del CDS (región codificante), que a su vez corresponde a la posición 222 del aminoácido (Ensembl, s. f.-b, s. f.-a).

Las secuencias obtenidas de las 55 muestras se procesaron para el análisis, con el objetivo de obtener la predicción de la secuencia de proteínas y así identificar la posición del cambio de aminoácidos utilizando la herramienta de *ExPasy*. La herramienta ExPasy muestra seis (6) posibles marcos de lectura, de los cuales se escogieron los que mostraban una secuencia de aminoácidos coherente sin cortes ni codones sin sentido, por ejemplo: codones de parada. La **Figura 34** muestra uno de los seis marcos de lectura que resultó ser la secuencia correcta, lo cual fue confirmado mediante BLAST protein (Blastp) donde se observó que las secuencias coincidían con la secuencia de referencia proteínica humana para methylenetetrahydrofolate reductase isoform 2 (NP_005948.3), cuyo SNP representa un cambio en la posición 222 de alanina a la valina, y se encontró tal coincidencia (**Figura 35**). Luego, se confirmó la posición del aminoácido correspondiente al SNP, donde se observa alanina para las secuencias con el alelo silvestre y una valina para aquellas con el alelo mutante en la misma posición. Todos los aminoácidos están dentro del ORF, pero el programa solamente resalta donde aparece la primera metionina, ya que por defecto asume que ese es el codón de inicio.

Figura 34. Predicción del marco de lectura de las secuencias traducidas a proteína. (Arriba) se observa el aminoácido silvestre alanina (A), resaltado con un círculo rojo. (Abajo),se observa el aminoácido mutante valina (V), resaltado con un círculo rojo.

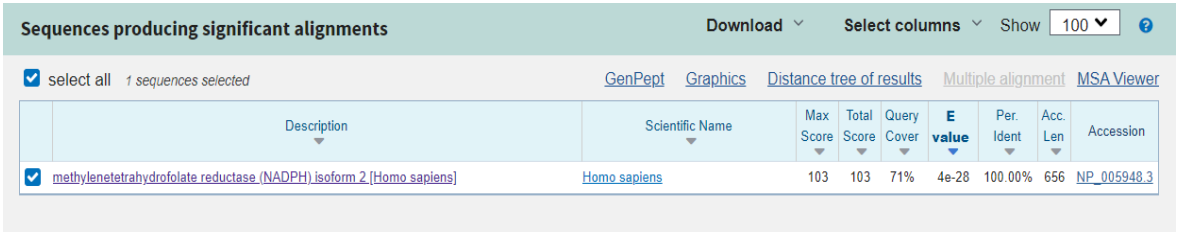


Figura 35. Análisis en Blastp de las secuencias de proteínas obtenidas confirmó la identidad molecular de las mismas.

Las 55 secuencias de proteínas obtenidas fueron alineadas junto con una secuencia de referencia del genoma humano en GenBank (NP_005948.3) methylenetetrahydrofolate reductase isoform 2 [*Homo sapiens*] correspondiente a la región secuenciada del gen MTHFR (**Figura 36**).

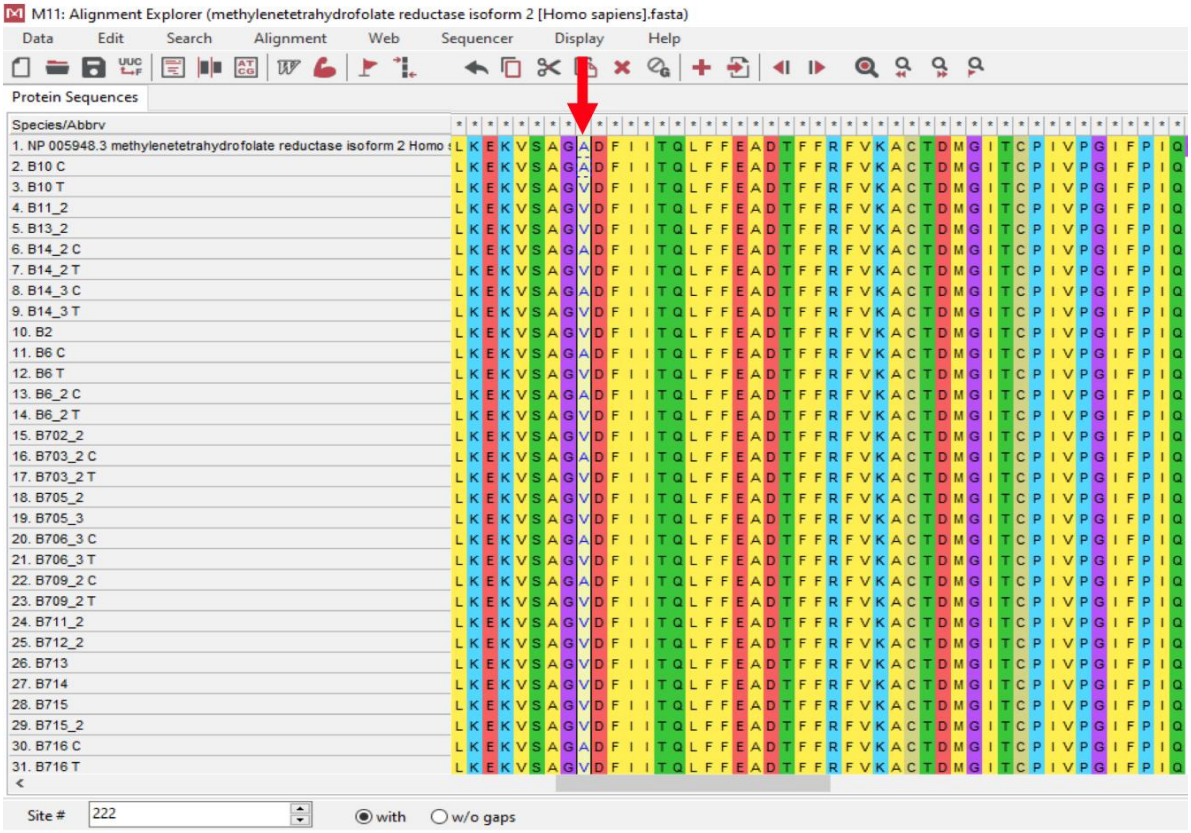


Figura 36. Alineamiento de las secuencias proteínicas. Alineamiento de las secuencias traducidas a proteínas con la secuencia de referencia proteínica metilentetrahidrofolato reductasa isoforma 2 (arriba, primera secuencia). Se observa (Flecha) el polimorfismo en la posición del aminoácido mutante (V, valina) y silvestre (A, alanina).

6.5 Verificación de la posición del nucleótido (SNP) y del aminoácido

La posición del SNP fue confirmada en la posición 755 del transcrito ENST00000376590.9 según Ensembl y el Graphics de la variante de transcripción 2 (NM_005957.5), que a su vez coincide con la 665 en el dbSNP, o 667 como citado en la literatura. Además, la posición del aminoácido coincide con aquellas encontradas en el Graphics de la isoforma 2, en el dbSNP y en la literatura (**Figura 37**).

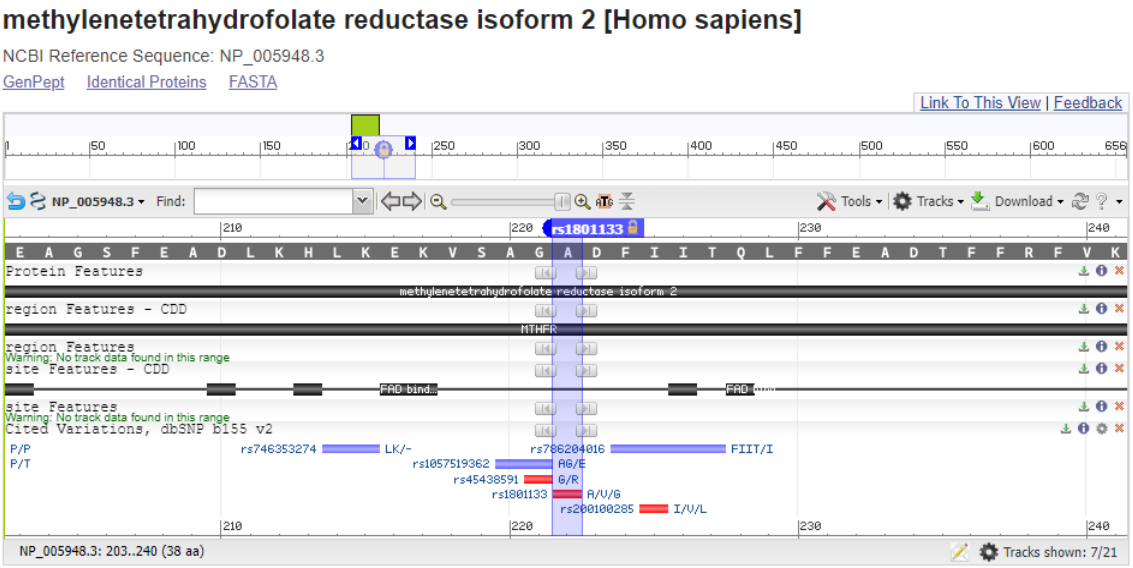


Figura 37. Posición del SNP dentro la isoforma 2 en GenBank

6.6 Genotipo de cada secuencia.

La secuenciación y análisis de las 55 muestras permitió determinar el genotipo de cada individuo, los cuales se muestran en la **Tabla 20**.

Tabla 20. Genotipo de todas las secuencias

Genotipaje de las secuencias	
Muestra/Código	**Genotipo**
B6_2MT	CT
B719_2MT	CT
B845MT	CT
B706_3MT	CT
B709_2MT	CT
B727_2MT	CT
B10_2MT	CT
B14_3MT	CT
B736DMT	CT
B6MT	CT
B734_2MT	CT
B717_2MT	CT
B717MT	CT

B703_2MT	CT
B852MT	CT
B716MT	CT
B14_2MT	CT
B715MT	TT
B722_2MT	TT
B728_2MT	TT
B713MT	TT
B723_2MT	TT
B718_2MT	TT
B731MT	TT
B846MT	TT
B13_2MT	TT
B731_2MT	TT
B730MT	TT
B840MT	TT
B847MT	TT
B714MT	TT
B711_2MT	TT
B851MT	TT
B844MT	TT
B720_2MT	TT
B733_2MT	TT
B702_2MT	TT
B721_2MT	TT
B842MT	TT
B720_3MT	TT
B2MT	TT
B725_2MT	TT
B713_2MT	TT

B712_2MT	TT
B715_2MT	TT
B705_2MT	TT
B730_2MT	TT
B726_2MT	TT
B724_2MT	TT
B705_3MT	TT
B843bMT	TT
B729MT	TT
B732_2MT	TT
B11_2MT	TT
B849MT	TT

7. Parámetros de genética poblacional de los genotipos

El análisis de genética poblacional con el programa PopGene permitió cuantificar la frecuencia alélica y genotípica de las muestras. El formato para los análisis con este programa requiere que cada alelo reciba un código con letras, para lo cual escogimos **A** para el silvestre y **B** para el mutante. Los resultados demostraron una frecuencia alélica de 0.15 (15%) para el alelo de tipo silvestre [Citosina (C)/Alanina (A)] y de 0.85 (85%) para el alelo mutante [Timina (T)/Valina (V)] (**Tabla 21 y Figura 38**). El número de genotipos observados fue de 0 individuos homocigotos silvestres, 17 individuos heterocigotos y 38 individuos homocigotos mutantes (**Tabla 22**), para los cuales las frecuencias genotípicas fueron respectivamente de 0%, 31% y 69% (**Figura 39**).

Tabla 21. Frecuencias alélicas de C (A) y T (B)

Allele Frequency of population 1

Allele \ Locus	MTHFR
Allele A	0.1545
Allele B	0.8455

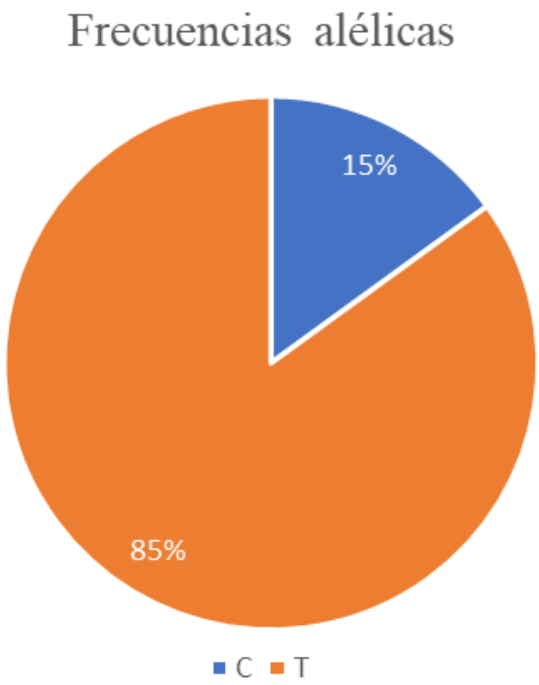


Figura 38. Frecuencias alélicas. Frecuencias alélicas de los alelos C (silvestre, azul) y T (mutante, anaranjado) para el polimorfismo SNP rs1801133.

Tabla 22. Número de genotipos CC (A, A), CT (B, A) y TT (B, B)

Genotypes	Obs. (O)
(A, A)	0
(B, A)	17
(B, B)	38

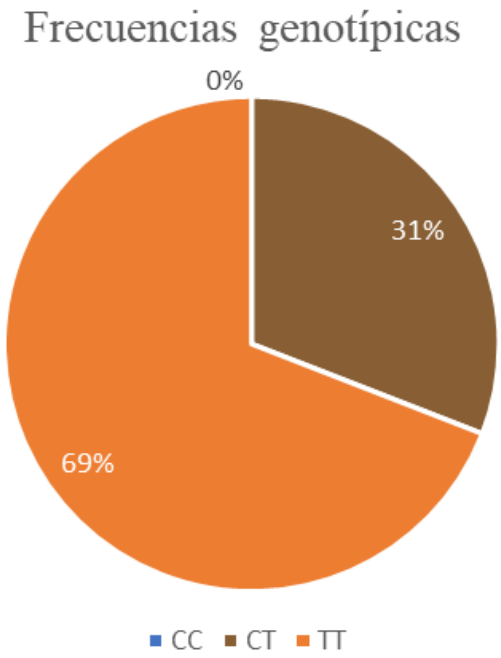


Figura 39. Frecuencias genotípicas. Frecuencias genotípicas para los genotipos identificados CT (heterocigotos, marrón) y TT (homocigotos mutantes, anaranjado) en las muestras para el polimorfismo rs1801133. No se observaron homocigotos CC (silvestre).

Las frecuencias alélicas y genotípicas fueron contrastadas con estos mismos parámetros poblacionales en otras etnias y otros países a fin de realizar una comparación general.

En primer lugar, en un estudio sobre los grupos afroamericanos, caucásicos e hispánicos (Graydon et al., 2019), se encontró que el porcentaje de 677T (mutante) es más elevado en los hispánicos (42%), más bajo en los afroamericanos (16%) e intermedio en los caucásicos (32%), (**Tabla 23**).

Tabla 23. Frecuencias de alelos de afroamericanos, caucásicos e hispánicos en EE. UU.

Distribución de los alelos de MTHFR para el locus 677 por etnia (%)		
	C	T
Afroamericanos	84	16
Caucásicos	68	32
Hispánicos	58	42

Adaptado de Graydon et al. (2019).

Adicionalmente, se determinó que la distribución de la frecuencia genotípica para el homocigoto mutante entre los tres grupos es mayor para los hispánicos. Los reportes indican que esa

distribución es mayor que la de la frecuencia genotípica del homocigoto silvestre y del heterocigoto dentro de los hispánicos (**Tabla 24**).

Tabla 24. Frecuencias de genotipos de afroamericanos, caucásicos e hispánicos en EE. UU.

Distribución de la frecuencia de los genotipos por etnia (%)			
	677CC	**677CT**	**677TT**
Afroamericanos	46.4	23.0	8.1
Caucásicos	29.8	39.4	29.1
Hispánicos	23.8	37.6	62.8

Adaptado de Graydon et al. (2019)

En segundo lugar, se encontraron frecuencias alélicas y números genotípicos similares a los del presente estudio en un estudio de metaanálisis de la población de China (Wang et al., 2016) y en otro estudio de metaanálisis de México y otros países de América Central (Reyes et al., 2021). Las **Tablas 25** y **26** muestran ambas una frecuencia alélica mayor para el alelo mutante y tendencialmente un número genotípico del homocigoto mutante mayor que el homocigoto silvestre. La diferencia entre una y otra es que en las poblaciones de China el genotipo heterocigoto es en su totalidad mayor a los otros genotipos, mientras que en las poblaciones nativas de América Central hay veces en que el genotipo homocigoto mutante supera tanto al genotipo homocigoto silvestre y aquello heterocigoto, lo cual ocurre también en las muestras del presente estudio.

Tabla 25. Frecuencias alélicas de los grupos étnicos en distintas localizaciones en China donde el genotipo CC era el menos numeroso.

Autor y año publicación	**Localización/Región**	**Grupo étnico**	**Genotipos**			**Frecuencias alélicas**	
			CC	CT	TT	C	T
Zhang CS, 2005	Shandong	Han	11	42	33	0.37	0.63
Li AF, 2007	Henan	Han	163	173	164	0.4999	0.501
Chen YX, 2010	Shanxi	Han	6	24	20	0.36	0.64

He YX, 2012	Henan	Han	198	493	402	0.41	0.59
Cong YY, 2012	Shandong	Han	130	457	454	0.34	0.66
Zhang YL, 2012	Shandong	Han	138	398	289	0.41	0.59
Chen HB, 2012	Shanxi	Han	10	31	22	0.40	0.60
Xiu X, 2013	Shandong	Han	442	1354	1138	0.38	0.62
Chen YX, 2013	Shanxi	Han	32	97	63	0.42	0.58
Yan Q, 2014	Shanxi	Han	497	1313	860	0.43	0.57
Xing JF, 2014	Henan	Han	57	207	158	0.38	0.62
Li JH, 2015	Hebei	Han	220	617	430	0.42	0.58
Chen HL, 2014	Guangxi	Han	82	271	211	0.39	0.61
Ma LM	Heilongjiang	Han	78	240	137	0.44	0.56
Tang HY, 2014	Shandong	Han	107	373	307	0.37	0.63
Lu GR, 2014	Shandong	Han	201	625	526	0.38	0.62
Jiao FY, 2014	Shandong	Han	93	261	175	0.42	0.58
Gao X, 2014	Hebei	Han	158	429	273	0.43	0.57
Cui HL, 2015	Henan	Han	201	542	510	0.38	0.62

Adaptada de (Wang et al., 2016)

Tabla 26. Frecuencias alélicas de los grupos étnicos en distintas localizaciones en México y países de Centroamérica donde el genotipo CC era el menos numeroso.

Población	País	Genotipos			Frecuencias alélicas	
		CC	CT	TT	C	T
Bribrí	Costa Rica	10	54	46	0.34	0.66
Cabecar	Costa Rica	5	58	84	0.2297	0.7703
Chinanteco	México	4	22	55	0.19	0.81
Oaxaca Chontal	México	10	16	18	0.49	0.59
Chorotega	Costa Rica	18	74	41	0.41	0.59
Chuj	México	1	4	12	0.29	0.71
Guatusos	Costa Rica	1	14	14	0.28	0.72
Guaymí	Costa Rica	9	38	110	0.18	0.82
Huastescos	México	9	32	38	0.32	0.68
Huetar	Costa Rica	21	79	53	0.4	0.6
Huicholes	México	8	28	14	0.44	0.56
Jakaltecos	Guatemala	3	10	27	0.2	0.8
Kanjobal	Guatemala	2	9	18	0.22	0.78
Kaqchibel	Guatemala	0	11	25	0.15	0.85
Mam	México	4	17	24	0.28	0.72
Yucatán Maya	México	7	28	19	0.39	0.61
Mazatecos	México	4	21	34	0.25	0.75
Mixe	México	11	43	35	0.38	0.62
Mixtecas	México	12	50	62	0.3	0.7
Nahuas	México	7	58	70	0.27	0.73
Nahuatl1	México	7	25	20	0.37	0.63
Nahuatl2	México	6	26	12	0.43	0.57
Nahuatl3	México	3	25	24	0.3	0.7
Nahuatl4	México	4	23	17	0.35	0.65
Otomí	México	28	84	108	0.32	0.68

Popoluca	México	5	18	13	0.39	0.61
Purepecha 1	México	4	10	7	0.43	0.57
Purepecha 2	México	2	13	14	0.29	0.71
Tojolabal	México	6	28	12	0.53	0.57
Trikís	México	0	14	75	0.08	0.92
Zapotecas	México	7	9	26	0.27	0.73

Adaptada de Reyes et al (2021)

CAPÍTULO V: DISCUSIÓN

1. La calidad de los datos de secuenciación genómica NGS fue excelente

Todas las muestras cuantificadas y validadas tenían en su totalidad una concentración mayor de 100 ng/µL y un valor entre 1.8 y 2.0 tanto para 260/280 y 260/230, por lo que todas las muestras eran de buena calidad.

La Secuenciación de Nueva Generación arrojó lecturas limpias que resultaron más del 99.80% para todos los genomas, mientras que la contaminación relativa al adaptador fue menos del 20% y no hubo ninguna lectura con N o de baja calidad. En esa misma línea, los cuatro genomas mostraron un porcentaje efectivo de más del 99% y un porcentaje de error del 0.03% y para Q30 se obtuvieron valores superiores al 91%. También se logró una secuenciación exitosa de más del 98% para las lecturas mapeadas apropiadamente con una cobertura tendencialmente mayor del 98% y una profundidad promedio de 34 veces, por consiguiente, los datos obtenidos demostraron buena calidad.

Respecto a las mutaciones en la línea germinal, las más numerosas fueron los SNPs, seguidas por los InDels, los SV y por último los CNV. Entre los primeros, prevalecían los SNP sinónimos seguidos por los SNP con cambio de sentido. En este último grupo se identificaron por primera vez en genomas humanos panameños las variantes posiblemente asociadas a enfermedades no transmisibles en la población.

2. Las variantes polimórficas identificadas están asociadas con enfermedades no transmisibles en la población panameña

Los datos elaborados por la Contraloría General de la República, mediante el Instituto Nacional de Estadística y Censo (INEC) sobre las defunciones ocurridas en el año 2021, indican como principales causas de muerte: las enfermedades del sistema circulatorio, de las cuales las principales son las isquémicas del corazón y las cerebrovasculares; los tumores malignos, de los cuales los principales son mama, próstata, el colorrectal, el de estómago; y las enfermedades endocrinas y metabólicas, de las cuales la principal es la diabetes mellitus (Ibarra H. & Carrión, 2021).

Los datos arrojados por la Secuenciación de Próxima Generación muestran que casi todas las variantes encontradas dentro del exoma que causan cambios significativos en los aminoácidos son SNPs, en especial aquellos con cambio de sentido, es decir, se cambia un aminoácido por

otro. Los SNPs identificados vinculados a las enfermedades crónicas más numerosos tienden a estar relacionados con el cáncer, enfermedades vasculares, diabetes y obesidad, lo cual es consistente con los patrones de enfermedades en nuestra población. Los polimorfismos más numerosos encontrados estaban relacionados con enfermedades vasculares, la primera causa de muerte en general de 2021. Entre los genes candidatos de la tabla, aparecen como polimorfismos asociados a enfermedades vasculares las variantes de MTHFR, EPHX1 (epóxido hidrolasa 1) y SOD2 (superóxido dismutasa 2), que eran las variantes que se repetían en todos los genomas secuenciados.

La segunda variante más numerosa encontrada era la asociada al cáncer, que ha sido la segunda causa principal de muerte en el año 2021. Los genes y sus variantes polimórficas identificadas como enlazadas a los cánceres incluyen los genes candidatos NQO1 (NAD(P)H deshidrogenasa [quinona] 1), FGFR4 (receptor 4 del factor de crecimiento de fibroblastos), MTHFR, TP53 (proteína tumoral p53) y EPHX1, de los cuales el primero se observó en mayor número de genomas respecto a los otros. Aunque se hayan encontrado menos variantes genéticas asociadas a los cánceres en cuanto a las vinculadas a enfermedades vasculares, las primeras han aparecido en más genomas respecto a las segundas.

La tercera variante más numerosa era la asociada a la diabetes, que ha sido la principal causa de muerte entre las enfermedades nutricionales, endocrinas y metabólicas. Fueron la tercera causa de muerte en el año 2021. Los genes candidatos cuyas variantes están asociadas a la diabetes son SLC30A8 (Portador de soluto Familia 30 Miembro 8), IRS2 (sustrato 2 del receptor de insulina), OAS1 (2'-5'-oligoadenilato sintetasa 1) y PPP1R3A (subunidad reguladora 3A de la proteína fosfatasa 1), entre las cuales la variante de IRS2 ha aparecido en más genomas respecto a las otras.

La cuarta variante más numerosas era la relacionada a la obesidad. Los genes candidatos cuyas variantes están asociadas a la obesidad son GHRL (Prepropéptido de grelina y obestatina) y SDC3 (Syndecan-3), siendo el primero el que apareció en más genomas que el segundo.

Entre los genes candidatos, el ILR4 (receptor de interleucina 4) es el único cuya variante no está asociada al tipo de enfermedades que causaron más muertes en el 2021, pues está conectada a la lenta progresión de la adquisición del Síndrome de Inmunodeficiencia Adquirida.

Las variantes que más se han encontrado en los distintos genomas están asociadas a enfermedades no transmisibles que tienen una gran coincidencia con las que causaron el mayor número de muertes en el año 2021, por lo que es posible que estas variantes contribuyan a la aparición de ellas. Sin embargo, más estudios genéticos y genómicos serán necesarios para confirmar estas asociaciones, por ejemplo: incluir en futuros análisis muestras de pacientes de diferentes enfermedades y comparar los genotipos con grupos control. Hasta la fecha no se han realizado estudios de este tipo en Panamá.

3. Asociación entre rs1801133 C677T y cáncer y enfermedades vasculares

Parte de la información sobre las enfermedades a las cuales estaban asociadas las variantes de los genes candidatos estaba presente entre los datos que resultaron de la NGS. Estas cifras son provenientes de bases de datos tales como dbSNP, COSMIC, OMIM, GWAS Catalog y HGMD. Los datos incluyen información reportada sobre la variante polimórfica, como los mayores SNPs en GWAS y asociaciones de cáncer y otras enfermedades. Como esta información es variable, según se vayan reportando nuevos polimorfismos y nuevas publicaciones sobre los mismos, confirmando o no, dichas asociaciones de enfermedades, o incluso asociaciones conflictivas entre diferentes poblaciones. Por lo tanto, siempre es necesario contrastar los datos obtenidos con las actualizaciones más recientes, pero la información no es siempre igual al 100% en todas las bases de datos.

La variante rs1801133 de MTHFR, según dbSNP, está clasificada de la siguiente forma dentro de la significancia clínica: interpretaciones conflictivas de patogenicidad para el polimorfismo de la MTHFR termolábil y homocistinuria debido a la deficiencia de la metilentetrahidrofolato reductasa; significancia incierta para el tumor estromal gastrointestinal, la trombofilia por el defecto de trombina y el accidente cerebrovascular. El mismo se atribuye posiblemente como benigno para los defectos del tubo neural sensitivo al folato; respuesta a drogas para la respuesta al metotrexato; sin significancia para el neoplasma estomacal (*rs1801133 RefSNP Report - dbSNP - NCBI*, 2022). Estas clasificaciones están también en ClinVar y GWASCatalog, pero ClinVar especifica que el polimorfismo de la MTHFR termolábil y la homocistinuria debido a la deficiencia de la metilentetrahidrofolato reductasa han sido reportadas tanto como benignas cuanto como patogénicas (ClinVar, 2023; *Variant: rs1801133*, s. f.). Según COSMIC, la variante está asociada al cáncer de mama, al cáncer gástrico, a la leucemia mieloide aguda, al

cáncer de pulmón, al adenocarcinoma de próstata y al carcinoma urotelial de la vejiga (Mutation *Overview Page* MTHFR_ENST00000376590 - p.A222V (*Substitution* - Missense), n.d.). Según OMIM, rs1801133 está asociada a la homocistinuria debido a la deficiencia de MTHFR, a la susceptibilidad a los defectos del tubo neural, a la susceptibilidad del tromboembolismo y a la susceptibilidad a enfermedades vasculares (OMIM, s. f.). Según HGMD, tiene asociación con enfermedades cardiovasculares (*HGMD® home page*, s. f.).

En relación con la predicción de la nocividad de la variante, resultó ser deletérea para SIFT (0.002), probablemente dañina para Polyphen2_HDIV (0.998), probablemente dañina para Polyphen2_HVAR (0.941), deletérea para LRT (0.000) y deletéreo para FATHMM (-4.03). En cambio, para *Mutation Taster* resulta polimorfismo automático, pero no deletéreo y para *MutationAssessor*, resulta en la categoría media, que significa que la variante es funcional; adicionalmente, para CADD resulta probablemente benigno.

4. Comparación de las frecuencias alélicas en diferentes poblaciones del mundo

El polimorfismo 677C→T podría constituir solamente un factor de riesgo moderado para algunos trastornos. Sin embargo, desde un punto de vista de la población mundial, representa una carga considerable ya que la variante es bastante común (Leclerc et al., 2013).

Según los análisis de genotipado poblacional realizado en esta tesis, no hay genotipos silvestres homocigotos, pero sí hay 17 genotipos heterocigotos y 38 genotipos homocigotos mutantes; mientras que la frecuencia del alelo silvestre (C) es de 0.1545 y la del alelo mutante (T), 0.8455. Esto contrasta con lo que dicen las bases de datos en dbSNP, puesto que allí resulta que, a nivel mundial, el alelo C tiende a ser mayor que el alelo T (ALFA: C=0.659842, T=0.340158; 1000Genomes: C=0.7546, T=0.2454; *Allele Frequency Aggregator:* C=0.659842; T=0.340158; ExAC C=0.696333, T=0.303667, gnomAD Exomes: C=0.685141, T=0.314859; TopMed: C=0.708757, T=0.291243) (*rs1801133 RefSNP Report - dbSNP - NCBI*, 2022).

Los datos demuestran que en general las poblaciones africanas poseen la menor frecuencia del alelo T: 0.090 (1000 *Genome Projects Phase* 3), 0.108 (gnomAD), 0.12 (ALFA). Los asiáticos exhiben una frecuencia mayor que los africanos: 0.339 (ALFA). Los europeos muestran la segunda frecuencia más elevada: 0.296 (1000 Genomes), 0.315 (gnomAD), 0.349 (ALFA). Los

latinoamericanos muestran la frecuencia más elevada: 0.474 (1000 Genomes) y 0.503 (gnomAD) (*rs1801133 (SNP) - Population genetics - Homo_sapiens - Ensembl genome browser 109*, s. f.).

En algunos casos, se demostró que las frecuencias del alelo mutante de algunas poblaciones latinas son incluso mayores que las del alelo silvestre, lo cual se acerca a los resultados encontrado por nosotros en la población Ngöbe. Según 1000 *Genomes Project* Phase 3, la población de Medellín, Colombia, posee un 0.543 % del alelo mutante. Se dan resultados similares para el alelo mutante en distintos estudios, como el de Graydon et al. (2019), donde se comparan las frecuencias genotípicas entre las poblaciones afro americanas, caucásicas e hispánicas, incluyendo estas últimas a todos los individuos quienes se identificaban como hispánicos o latinos independientemente de su linaje familiar o país de origen familiar. En este estudio se encontró que la distribución del alelo mutante es mayor en los hispánicos y menor en los afroamericanos; mientras que el alelo silvestre es mayor para los afroamericanos y menor para los hispánicos. En esos estudios se demostró que el genotipo homocigoto mutante en los hispánicos es el de mayor frecuencia. En cambio, el genotipo homocigoto mutante 677TT tiene distribución intermedia en la población caucásica y menor en la afroamericana. El genotipo predominante en la población caucásica es el heterocigoto 677CT, mientras que en las poblaciones afroamericanas sobresale el genotipo homocigoto silvestre 677CC. A pesar de determinar frecuencias etno geográficas en EE.UU.,estos estudios tienen algunas limitaciones, por ejemplo, el término “hispánico” se refería a individuos que se identificaban a sí mismos como hispánicos o latinos sin tener en cuenta su linaje genético ancestral o el país de origen familiar. Considerando la generalidad de los términos “hispánico” y “latino”, podrían no resultar claras las diferencias genéticas subyacentes entre los varios grupos étnicos, porque estos exhiben diferencias en ancestralidad, ya que se mostraron discrepancias en las frecuencias de 677C>T en varios grupos hispánicos/latinos (Graydon et al., 2019).

En algunos estudios de México y China, se demostró cómo en ciertas poblaciones de dichos países el alelo mutante incluso superaba al alelo silvestre en frecuencia. En Wang et al. (2016), se comparó la distribución del polimorfismo C677T de MTHFR entre los distintos grupos étnicos de distintas localidades de China. Para el grupo étnico Han, se encontraron frecuencias

de 677T mayores que 677C, así como genotipos 677CC en número menor respecto a los otros dos genotipos, mientras que el genotipo 677CT era el prevalente.

En Reyes at al. (2021), se estudió el polimorfismo C667T en los indígenas de México y países de América Central, como Guatemala y Costa Rica. Para varios grupos étnicos en México, Costa Rica y Guatemala, se encontraron frecuencias alélicas de 677T mayores que 677C, así como genotipos CC en menor cantidad respecto a los otros dos. En 18 poblaciones, el genotipo TT era el mayor de todos, mientras que en las otras lo era el CT. En el caso particular de los grupos indígenas Kaqchibel en Guatemala y Trikis en México, no se encontró para nada el genotipo homocigoto silvestre, mientras que el genotipo homocigoto mutante era el predominante, lo cual ocurre también con las muestras del presente estudio. Adicionalmente, la frecuencia del alelo mutante del presente estudio, 0.8455, se parece a la de Kaqchibel, que es de 0.85, y es poco inferior a la de Trikis, la cual es de 0.92. Estos reportes son consistentes con nuestros resultados en cuanto a la alta frecuencia del alelo mutante en la población Ngöbe.

Nuestros datos y los otros publicados previamente demuestran no sólo que es posible reportar poblaciones indígenas, sino también en poblaciones mestizas como las latinoamericanas, lo cual sugiere que el origen de estos alelos mutantes en las poblaciones mestizas podría ser derivado de las indígenas. La altísima frecuencia del alelo mutante en la población Ngöbe también podría estar relacionada con los fenómenos de deriva génica, como los de cuello de botella reportados previamente en indígenas durante los períodos de conquista y colonial (Castro-Pérez et al., 2016). El resultado de este cuello de botella se refleja en la baja diversidad genética, por ejemplo: frecuencias alélicas muy altas. En el caso particular de los amerindios Ngöbe, estudios previos utilizando marcadores STR también reportaron frecuencias alélicas muy altas de entre 60% y más de 80% para algunos loci (Castro et al., 2007). En ese estudio se reportó, en particular, que la población Ngöbe exhibe frecuencia de 86% para el alelo TH01*6, 77% para el alelo FESFPS*11 y 60% para el alelo VWA*16. Estas frecuencias altas reportadas en marcadores STR en la población Ngöbe se acerca a las frecuencias alélicas reportadas por nosotros en el polimorfismo rs1801133 del gen MTHFR.

CONCLUSIONES Y RECOMENDACIONES

CONCLUSIONES

- La estrategia de secuenciación genómica de nueva generación (NGS) nos permitió analizar exitosamente las variantes polimórficas de miles de genes en los cuatro genomas secuenciados de amerindios Ngöbe.
- La calidad de las secuencias genómicas obtenidas fue excelente, las cuales tuvieron casi la totalidad de lecturas limpias, un valor grande para Q30, una cobertura superior al 99% y una profundidad promedio muy elevada, demostrando una calidad excelente de los datos obtenidos.
- Los análisis genómicos nos condujeron a la identificación por primera vez de variantes genéticas de importancia biomédica que afectan a la población Ngöbe y presumiblemente a la población mestiza general derivada de esta población indígena ancestral.
- El tipo de variante polimórfica más abundante y asociada a posibles enfermedades no transmisibles son los SNPs, de los cuales predominan los sinónimos y los que muestran cambio de sentido, siendo estos últimos lo que más causan cambios asociados a enfermedades.
- Se identificaron múltiples candidatos de polimorfismos genéticos asociadas principalmente a enfermedades del sistema circulatorio, cánceres y enfermedades metabólicas, siendo los asociados a cáncer los que se han encontrado en más comúnmente.
- Para análisis más detallados nos enfocamos en determinar la frecuencia alélica de la variante polimórfica rs1801133, que corresponde a un cambio de base en la posición 677C>T del gen MTHFR en una muestra de 55 individuos la población ancestral amerindia Ngöbe. La frecuencia alélica de la variante silvestre resulta ser de 0.1545 y la del alelo mutante, de 0.8455. Esta frecuencia es consistente con algunos reportes en otras poblaciones.
- Los análisis sugieren que esta variante génica podría estar presente en la población mestiza del país y tener un origen ancestral Ngöbe.
- Estos datos sugieren que la variante mutante podría representar un factor de riesgo a algunas enfermedades no sólo en la población Ngöbe, sino también en la población del país.

RECOMENDACIONES Y DIRECCIONES FUTURAS

El estudio piloto presentado en esta tesis sobre los estudios genómicos en la población panameña ha conducido por primera vez a la identificación de variantes genéticas asociadas a enfermedades no transmisibles en el país. Los datos sugieren que los polimorfismos genéticos identificados podrían estar vinculados a varias enfermedades no transmisibles. Para confirmar estos resultados sugerimos varias estrategias:

- Aumentar el número de genomas secuenciados, no sólo de amerindios Ngöbes, sino también de otros grupos de la población con trasfondo africano, europeo y mestizo.
- Secuenciar y analizar el genoma de pacientes panameños que padezcan enfermedades específicas e incluir grupos control que permitan identificar los posibles factores de riesgo de manera más específica.
- En el caso particular del polimorfismo rs1801133 del gen MTHFR recomendamos expandir estos estudios con un grupo de pacientes que padezcan alguna de las enfermedades asociadas al mismo, en particular: cánceres o enfermedades vasculares, e incluir un grupo control de individuos sanos. Este modelo de estudio podría aplicarse también a estudios genómicos y a estudios involucrando polimorfismos específicos.

BIBLIOGRAFÍA

Anencephaly: MedlinePlus Genetics. (s. f.). Recuperado 30 de enero de 2023, de https://medlineplus.gov/genetics/condition/anencephaly/

Arias, T., Castro, E., Ruiz, E., Barrantes, R., & Jorge-Nerbert, L. (2002). La mezcla racial de la población panameña. *Revista Médica de Panamá - ISSN 2412-642X*, *27*, 5–17. https://doi.org/10.37980/IM.JOURNAL.RMDP.200225

Arias, T. D., Barrantes, R., Jorge, L. F., Azofeifa, J., Carles, M., & Cooke, R. G. (1992). ["Cholos de Coclé": determination of their racial mixture and genetic origins]. *Revista médica de Panamá*, *17*(3), 180–187.

Arias, T. D., Jorge, L. F., Griese, E.-U., Inaba, T., & Eichelbaum, M. (1993). Polymorphic N-acetyltransferase (NAT2) in Amerindian populations of Panama and Colombia: high frequencies of point mutation 857A, as found in allele S3/M3. *Pharmacogenetics*, *2*(6), 328–331.

Barbieri, C., Barquera, R., Arias, L., Sandoval, J. R., Acosta, O., Zurita, C., Aguilar-Campos, A., Tito-Alvarez, A. M., Serrano-Osuna, R., Gray, R. D., Mafessoni, F., Heggarty, P., Shimizu, K. K., Fujita, R., Stoneking, M., Pugach, I., & Fehren-Schmitz, L. (2019). The Current Genomic Landscape of Western South America: Andes, Amazonia, and Pacific Coast. *Molecular Biology and Evolution*, *36*(12), 2698–2713. https://doi.org/10.1093/molbev/msz174

Barrantes, R., Smouse, P. E., Mohrenweiser, H. W., Gershowitz, § Henry, Azofeifa, J., Arias, T. D., & Neel$, J. v. (1990). Microevolution in Lower Central America: Genetic Characterization of the Chibcha-speaking Groups of Costa Rica and Panama, and a Consensus Taxonomy Based on Genetic and Linguistic Affinity. En *Am. J. Hum. Genet* (Vol. 46).

Belbin, G. M., Nieves-Coló N, M. A., Kenny, E. E., Moreno-Estrada, A., & Gignoux, C. R. (2018). Genetic diversity in populations across Latin America: implications for population and medical genetic studies. *Current Opinion in Genetics & Development*, *53*, 98–104. https://doi.org/10.1016/j.cogede.2018.07.006

Beleza, S., Alves, C., Reis, F., Amorim, A., Carracedo, A., & Gusmão, L. (2004). 17 STR data (AmpF/STR Identifiler and Powerplex 16 System) from Cabinda (Angola). *Forensic Science International*, *141*(2–3), 193–196. https://doi.org/10.1016/j.forsciint.2004.01.008

Bock, C. H., Schwartz, A. G., Ruterbusch, J. J., Levin, A. M., Neslund-Dudas, C., Land, S. J., Wenzlaff, A. S., Reich, D., McKeigue, P., Chen, W., Heath, E. I., Powell, I. J., Kittles, R. A., & Rybicki, B. A. (2009). Results from a prostate cancer admixture mapping study in African-American men. *Human Genetics*, *126*(5), 637. https://doi.org/10.1007/s00439-009-0712-z

Camacho, M. v., Benito, C., & Figueiras, A. M. (2007). Allelic frequencies of the 15 STR loci included in the AmpFlSTR® Identifiler™ PCR Amplification Kit in an autochthonous sample from Spain. *Forensic Science International*, *173*(2–3), 241–245. https://doi.org/10.1016/J.FORSCIINT.2007.02.006

Cáncer de mama - Síntomas y causas - Mayo Clinic. (s. f.). Recuperado 1 de febrero de 2023, de https://www.mayoclinic.org/es-es/diseases-conditions/breast-cancer/symptoms-causes/syc-20352470

Castro, E., Trejos, D., Berovides-Alvarez, V., Arias, T., & Ramos, C. (2007). Genetic Polymorphism and Forensic Parameters of Nine Short Tandem Repeat Loci in Ngöbe and Emberá. *Human Biology*, *79*(5), 563–577.

Castro-Pérez, E. (2022, agosto 5). Estudios Genéticos y Genómicos de la Población Panameña. *Introducción a las ciencias ómicas: investigaciones y aplicaciones*.

Castro-Pérez, E., & Ramos, C. (2020). *Registro de Proyectos de Investigación*.

Castro-Pérez, E., Trejos, D. E., Hrbek, T., Setaluri, V., & Ramos, C. W. (2016). Genetic Ancestry of the Panamanian Population: Chibchan Amerindian Genes; and Biological Perspectives on Diseases. *The Internet Journal of Biological Anthropology*, *9*(1). https://doi.org/10.5580/IJBA.44045

Cheng, C. Y., Reich, D., Coresh, J., Boerwinkle, E., Patterson, N., Li, M., North, K. E., Tandon, A., Bailey-Wilson, J. E., Wilson, J. G., & Kao, W. H. L. (2010). Admixture Mapping of Obesity-related Traits in African Americans: The Atherosclerosis Risk in Communities (ARIC) Study. *Obesity*, *18*(3), 563–572. https://doi.org/10.1038/OBY.2009.282

ClinVar. (2023, febrero 18). *VCV000003520.77 - ClinVar - NCBI*. National Library of Medicine. https://www.ncbi.nlm.nih.gov/clinvar/variation/3520/?oq=((18559[AlleleID]))&m=NM_005957.5(MTHFR):c.665C%3ET%20(p.Ala222Val)

Constenla Umaña, A. (1991). *Las lenguas del área intermedia: introducción a su estudio areal* (1. Ed). Editorial de la Universidad de Costa Rica.

COSMIC. (s. f.). *Mutation overview page MTHFR_ENST00000376590 - p.A222V (Substitution - Missense)*. COSMIC. Recuperado 19 de febrero de 2023, de https://cancer.sanger.ac.uk/cosmic/mutation/overview?id=114234410

den Dunnen, J. T., Dalgleish, R., Maglott, D. R., Hart, R. K., Greenblatt, M. S., Mcgowan-Jordan, J., Roux, A. F., Smith, T., Antonarakis, S. E., & Taschner, P. E. M. (2016). HGVS Recommendations for the Description of Sequence Variants: 2016 Update. *Human Mutation*, *37*(6), 564–569. https://doi.org/10.1002/humu.22981

Donnan, G. A., Fisher, M., Macleod, M., & Davis, S. M. (2008). Stroke. *The Lancet*, *371*(9624), 1612–1623. https://doi.org/10.1016/S0140-6736(08)60694-7

Ensembl. (s. f.-a). *rs1801133 (SNP) - Genes and regulation - Homo_sapiens - Ensembl genome browser 109*. Ensembl. Recuperado 18 de febrero de 2023, de https://www.ensembl.org/Homo_sapiens/Variation/Mappings?r=1:11795821-11796821;v=rs1801133;vdb=variation;vf=1280266

Ensembl. (s. f.-b). *Transcript: ENST00000376590.9 (MTHFR-204) - Variant table - Homo_sapiens - Ensembl genome browser 109*. Ensembl. Recuperado 24 de febrero de 2023, de

https://www.ensembl.org/Homo_sapiens/Transcript/Variation_Transcript/Table?db=core;g=ENSG00000177000;r=1:11785723-11805964;source=dbSNP;t=ENST00000376590;v=rs1801133;vdb=variation;vf=1280266

Froese, D. S., Kopec, J., Rembeza, E., Bezerra, G. A., Oberholzer, A. E., Suormala, T., Lutz, S., Chalk, R., Borkowska, O., Baumgartner, M. R., & Yue, W. W. (2018). Structural basis for the regulation of human 5,10-methylenetetrahydrofolate reductase by phosphorylation and S-adenosylmethionine inhibition. *Nature Communications*, *9*(1). https://doi.org/10.1038/s41467-018-04735-2

Frosst, P., Blom, H. J., Milos, R., Goyette, P., Sheppard, C. A., Matthews, R. G., Boers, G. J. H., den Heijer, M., Kluijtmans, L. A. J., van den Heuve, L. P., & Rozen, R. (1995). A candidate genetic risk factor for vascular disease: a common mutation in methylenetetrahydrofolate reductase. *Nature Genetics*, *10*(1), 111–113. https://doi.org/10.1038/ng0595-111

Gaughan, D. J., Barbaux, S., Kluijtmans, L. A. J., & Whitehead, A. S. (2000). The human and mouse methylenetetrahydrofolate reductase (MTHFR) genes: Genomic organization, mRNA structure and linkage to the CLCN6 gene. *Gene*, *257*(2), 279–289. https://doi.org/10.1016/S0378-1119(00)00392-9

Gnecchi-Ruscone, G. A., Sarno, S., de Fanti, S., Gianvincenzo, L., Giuliani, C., Boattini, A., Bortolini, E., Corcia, T. di, Mellado, C. S., Jesus D Avila Francia, T., Gentilini, D., di Blasio, A. M., Cosimo, P. di, Cilli, E., Gonzalez-Martin, A., Franceschi, C., Franceschi, Z. A., Rickards, O., Sazzini, M., … Pettener, D. (2019). Dissecting the Pre-Columbian Genomic Ancestry of Native Americans along the Andes-Amazonia Divide. *Molecular Biology and Evolution*, *36*(6), 1254–1269. https://doi.org/10.1093/molbev/msz066

Gonçalves, R., Jesus, J., Fernandes, A. T., & Brehm, A. (2002). Genetic profile of a multi-ethnic population from Guiné-Bissau (west African coast) using the new PowerPlex® 16 System kit. *Forensic Science International*, *129*(1), 78–80. https://doi.org/10.1016/S0379-0738(02)00204-9

Goyette, P., Sumner, J. S., Milos, R., Duncan, A. M. V., Rosenblatt, D. S., Matthews, R. G., & Rozen, R. (1994). Human methylenetetrahydrofolate reductase: isolation of cDNA, mapping and mutation identification. *Nature Genetics*, *7*(2), 195–200. https://doi.org/10.1038/ng0694-195

Gravel, S., Zakharia, F., Moreno-Estrada, A., Byrnes, J. K., Muzzio, M., Rodriguez-Flores, J. L., Kenny, E. E., Gignoux, C. R., Maples, B. K., Guiblet, W., Dutil, J., Via, M., Sandoval, K., Bedoya, G., Oleksyk, T. K., Ruiz-Linares, A., Burchard, E. G., Martinez-Cruzado, J. C., & Bustamante, C. D. (2013). Reconstructing Native American Migrations from Whole-Genome and Whole-Exome Data. *PLoS Genetics*, *9*(12). https://doi.org/10.1371/JOURNAL.PGEN.1004023

Graydon, J. S., Claudio, K., Baker, S., Kocherla, M., Ferreira, M., Roche-Lima, A., Rodríguez-Maldonado, J., Duconge, J., & Ruaño, G. (2019). Ethnogeographic prevalence and implications of the 677C>T and 1298A>C MTHFR polymorphisms in US primary care

populations. *Biomarkers in Medicine*, *13*(8), 649–661. https://doi.org/10.2217/bmm-2018-0392

Gurdasani, D., Barroso, I., Zeggini, E., & Sandhu, M. S. (2019). Genomics of disease risk in globally diverse populations. *Nature Reviews Genetics*, *20*(9), 520–535. https://doi.org/10.1038/s41576-019-0144-0

Hankey, G. J. (1999). Smoking and risk of stroke. *Journal of Cardiovascular Risk*, *6*, 207–211. https://academic.oup.com/eurjpc/article/6/4/207/5933730

Harris, D. N., Song, W., Shetty, A. C., Levano, K. S., Cáceres, O., Padilla, C., Borda, V., Tarazona, D., Trujillo, O., Sanchez, C., Kessler, M. D., Galarza, M., Capristano, S., Montejo, H., Flores-Villanueva, P. O., Tarazona-Santos, E., O'Connor, T. D., & Guio, H. (2018). Evolutionary genomic dynamics of Peruvians before, during, and after the Inca Empire. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(28), E6526–E6535. https://doi.org/10.1073/PNAS.1720798115

HGMD® home page. (s. f.). Recuperado 19 de febrero de 2023, de https://www.hgmd.cf.ac.uk/ac/index.php

Human Genome Variation Sequence. (s. f.). Nomenclatura internacional para el informe de variaciones de secuencia del genoma humano. En *Grupo CAHT*. Grupo CAHT. Recuperado 8 de abril de 2023, de https://www.grupocaht.com/wp-content/uploads/2019/04/Nomenclatura-Internacional.pdf

Jaén Suárez, O. (1998). La población del istmo de Panamá: estudio de geohistoria. En E. de C. H. y A. de C. Internacional (Ed.), *La población del Istmo de Panamá. Estudio de Geohistoria*. https://dialnet.unirioja.es/servlet/libro?codigo=184747

Jaén Suárez, O. 1942-. (1978). *La población del Istmo de Panamá del Siglo XVI al Siglo XX: estudio sobre la población y los modos de organización de las economías, las sociedades y los espacios geográficos /*. Instituto Nacional de Cultura, Impresora de La Nación,.

Jin, H., Cheng, H., Chen, W., Sheng, X., Levy, M. A., Brown, M. J., & Tian, J. (2018). An evidence-based approach to globally assess the covariate-dependent effect of the MTHFR single nucleotide polymorphism rs1801133 on blood homocysteine: a systematic review and meta-analysis. *The American journal of clinical nutrition*, *107*(5), 817—825. https://doi.org/10.1093/ajcn/nqy035

Jopling, C. F. (1994). *Indios y negros en Panamá en los siglos XVI y XVII: selecciones de los documentos del Archivo General de Indias*. Centro de Investigaciones Regionales de Mesoamérica.

Jorge-Nebert, L. F., Eichelbaum, M., Griese, E.-U., Inaba, T., & Arias, T. D. (2002). Analysis of six SNPs of NAT2 in Ngawbe and Embera Amerindians of Panama and determination of the Embera acetylation phenotype using caffeine. En *Pharmacogenetics* (Vol. 12). http://www.PHYLIP

Kelly, P. M., Weinberg, A. D., Bernell, S., & Howard, S. W. (2016). Use Your Words Carefully: What Is a Chronic Disease? *Front. Public Health*, *4*(159). https://doi.org/10.3389/fpubh.2016.00159

Khalil, R., Al-Awaida, W. J., Al-Ameer, H. J., Jarrar, Y., Imraish, A., al Bawareed, O., Qawadri, R., al Madhoon, F., & Obeidat, L. (2021). Investigation of ACE rs4646994, MTHFR rs1801133 and VDR rs2228570 Genotypes in Jordanian Patients with Fibromyalgia Syndrome. *Endocrine, metabolic & immune disorders drug targets*, *21*(10), 1920–1928. https://doi.org/10.2174/1871530321666201223104622

Kim, J., Kim, H., Roh, H., & Kwon, Y. (2018). Causes of hyperhomocysteinemia and its pathological significance. En *Archives of Pharmacal Research* (Vol. 41, Número 4, pp. 372–383). Pharmaceutical Society of Korea. https://doi.org/10.1007/s12272-018-1016-4

Leclerc, D., Sibani, S., & Rozen, R. (2013). Molecular Biology of Methylenetetrahydrofolate Reductase (MTHFR) and Overview of Mutations/Polymorphisms. En *Madame Curie Bioscience Database [Internet]*. Landes Bioscience. https://www.ncbi.nlm.nih.gov/books/NBK6561/

Li, M. M., Datto, M., Duncavage, E. J., Kulkarni, S., Lindeman, N. I., Roy, S., Tsimberidou, A. M., Vnencak-Jones, C. L., Wolff, D. J., Younes, A., & Nikiforova, M. N. (2017). Standards and Guidelines for the Interpretation and Reporting of Sequence Variants in Cancer: A Joint Consensus Recommendation of the Association for Molecular Pathology, American Society of Clinical Oncology, and College of American Pathologists. En *Journal of Molecular Diagnostics* (Vol. 19, Número 1, pp. 4–23). Elsevier B.V. https://doi.org/10.1016/j.jmoldx.2016.10.002

Liew, S. C., & Gupta, E. das. (2015). Methylenetetrahydrofolate reductase (MTHFR) C677T polymorphism: Epidemiology, metabolism and the associated diseases. *European Journal of Medical Genetics*, *58*(1), 1–10. https://doi.org/10.1016/j.ejmg.2014.10.004

Loewen, J. A. (1963). Choco II: Phonological Problems. *https://doi.org/10.1086/464751*, *29*(4), 357–371. https://doi.org/10.1086/464751

Mariscal Davy, R. R. (2021). Características relacionadas al sobrepeso y obesidad en estudiantes de la facultad de medicina de la Universidad de Panamá en diciembre 2018. *Ciencia e Investigación Medico Estudiantil Latinoamericana*, *25*(1). https://doi.org/10.23961/cimel.v26i1.1276

Maróti, Z., Boldogkoi, Z., Tombácz, D., Snyder, M., & Kalmár, T. (2018). Evaluation of whole exome sequencing as an alternative to BeadChip and whole genome sequencing in human population genetic analysis. *BMC Genomics*, *19*(778), 1–13. https://doi.org/10.1186/S12864-018-5168-X/FIGURES/5

Ministerio de Salud. (2018). *Análisis de Situación de Salud. Macro Visión Nacional de Salud*.

Moreno-Estrada, A., Gignoux, C. R., Fernández-López, J. C., Zakharia, F., Sikora, M., Contreras, A. v, Acuña-Alonzo, V., Sandoval, K., Eng, C., Romero-Hidalgo, S., Ortiz-Tello, P., Robles,

V., Kenny, E. E., Nuño-Arana, I., Barquera-Lozano, R., Macín-Pérez, G., Granados-Arriola, J., Huntsman, S., Galanter, J. M., … Bustamante, C. D. (2014). The genetics of Mexico recapitulates Native American substructure and affects biomedical traits. *Science*, *344*(6189), 1280–1285. https://doi.org/10.1126/science.1251688

Neves, W. A., & Hubbe, M. (2005). Cranial morphology of early Americans from Lagoa Santa, Brazil: Implications for the settlement of the New World. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(51), 18309–18314. https://doi.org/10.1073/PNAS.0507185102/SUPPL_FILE/07185DATASET1.XLS

New England BioLabs Inc. (s. f.). *NEBNext Ultra Ii DNA Library Prep | NEB*. NEBNext. Recuperado 7 de abril de 2023, de https://international.neb.com/applications/ngs-sample-prep-and-target-enrichment/illumina-library-preparation/nebnext-ultra-ii-dna-library-prep/nebnext-ultra-ii-dna-library-prep

Novogene. (2022). *Davis-Panama-UdeP-10-hWGS-90Gb-WBI-STD-WOBI-NVUS2021050622 Project Standard Analysis Report*.

O'Connor, T. D. (2018). Native American Genomic Diversity through Ancient DNA. En *Cell* (Vol. 175, Número 5, pp. 1173–1174). Cell Press. https://doi.org/10.1016/j.cell.2018.10.058

OMIM. (s. f.). *Entry - *607093 - 5,10-Methylenetetrahydrofalate reductase; MTHFR - OMIM*. OMIM. Recuperado 19 de febrero de 2023, de https://www.omim.org/entry/607093#0003

Patiño Vásquez, A. (2014). *Revisión bibliográfica sobre el déficit de ácido fólico en la mujer embarazada y sus repercusiones sobre el feto.* [Tesis de Grado]. Universidade da Coruña.

Petersen, D., Kong, A., Jorge, L., Nebert, D., & Arias, T. (1991). Debrisoquine polymorphism: novel CYP2D6 gene Bam HI restriction fragment length polymorphism in the Ngawbé Guaymí Indian of Panama. *Pharmacogenetics*, *1*(3), 136–142. https://pubmed.ncbi.nlm.nih.gov/1688244/

Petrone, I., Bernardo, P. S., Santos, E. C. dos, & Abdelhay, E. (2021). MTHFR C677T and A1298C Polymorphisms in Breast Cancer, Gliomas and Gastric Cancer: A Review. *Genes*, *12*(4). https://doi.org/10.3390/GENES12040587

PLANAS, J., FUSTÉ, M., VIÑAS, J., & IRIZAR, J. L. (1966). Haptoglobin Types in the Iberian Peninsula. *Acta Genetica et Statistica Medica*, *16*(4), 371–376. https://www.jstor.org/stable/45104189

Popejoy, A. B., & Fullerton, S. M. (2016). Genomics is failing on diversity. *Nature*, *538*(7624), 161–164. https://doi.org/10.1038/538161a

Posth, C., Nakatsuka, N., Lazaridis, I., Skoglund, P., Mallick, S., Lamnidis, T. C., Rohland, N., Nägele, K., Adamski, N., Bertolini, E., Broomandkhoshbacht, N., Cooper, A., Culleton, B. J., Ferraz, T., Ferry, M., Furtwängler, A., Haak, W., Harkins, K., Harper, T. K., … Reich, D. (2018). Reconstructing the Deep Population History of Central and South America. *Cell*, *175*(5), 1185-1197.e22. https://doi.org/10.1016/j.cell.2018.10.027

Priya, S. S., Sankaran, R., Ramalingam, S., Sairam, T., & Somasundaram, L. S. (2016). Genotype phenotype correlation of genetic polymorphism of PPAR gamma gene and therapeutic response to pioglitazone in type 2 diabetes mellitus- a pilot study. *Journal of Clinical and Diagnostic Research*, *10*(2), 11–14. https://doi.org/10.7860/JCDR/2016/16494.7331

Ramos, C., Castro-Pérez, E., Molina-Jiron, C., & Trejos, D. (2018). Analysis of 30 INDEL Polymorphic Markers in the Panamanian Population: Gene Admixture Estimates, Population Structure and Forensic Parameters. *Journal of Forensic Research*, *09*(01). https://doi.org/10.4172/2157-7145.1000413

Rasmussen, M., Anzick, S. L., Waters, M. R., Skoglund, P., Degiorgio, M., Stafford, T. W., Rasmussen, S., Moltke, I., Albrechtsen, A., Doyle, S. M., Poznik, G. D., Gudmundsdottir, V., Yadav, R., Malaspinas, A. S., Samuel Stockton White, V., Allentoft, M. E., Cornejo, O. E., Tambets, K., Eriksson, A., … Willerslev, E. (2014). The genome of a Late Pleistocene human from a Clovis burial site in western Montana. *Nature 2014 506:7487*, *506*(7487), 225–229. https://doi.org/10.1038/nature13025

Reich, D., Patterson, N., Campbell, D., Tandon, A., Mazieres, S., Ray, N., Parra, M. v., Rojas, W., Duque, C., Mesa, N., García, L. F., Triana, O., Blair, S., Maestre, A., Dib, J. C., Bravi, C. M., Bailliet, G., Corach, D., Hünemeier, T., … Ruiz-Linares, A. (2012). Reconstructing Native American population history. En *Nature* (Vol. 488, Número 7411, pp. 370–374). Nature Publishing Group. https://doi.org/10.1038/nature11258

Reyes, G. (2022, abril 8). Enfermedades no transmisibles representan el 52% del total de las causas de muertes. *La Prensa*. https://www.prensa.com/sociedad/enfermedades-no-transmisibles-representan-el-52-del-total-de-las-causas-de-muertes/

Reyes, L., Godfrey, D., Ming, L. J., MacLean, C., Gonzalez, F. J., & Madrigal, L. (2021). The distribution in native populations from Mexico and Central America of the C677T variant in the MTHFR gene. *American Journal of Human Biology*, *33*(6). https://doi.org/10.1002/ajhb.23567

Ribeiro-dos-Santos, A. M., Vidal, A. F., Vinasco-Sandoval, T., Guerreiro, J., Santos, S., Ribeiro-dos-Santos, Â., & de Souza, S. J. (2020). Exome Sequencing of Native Populations From the Amazon Reveals Patterns on the Peopling of South America. *Frontiers in Genetics*, *11*. https://doi.org/10.3389/fgene.2020.548507

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), 405–424. https://doi.org/10.1038/gim.2015.30

Robbins, C., Torres, J. B., Hooker, S., Bonilla, C., Hernandez, W., Candreva, A., Ahaghotu, C., Kittles, R., & Carpten, J. (2007). Confirmation study of prostate cancer risk variants at 8q24 in African Americans identifies a novel risk locus. *Genome Research*, *17*(12), 1717. https://doi.org/10.1101/GR.6782707

Romoli, K., & Por, R. E. (1987). Los de la lengua de cueva los grupos indígenas del istmo oriental en la época de la conquista española. *Boletín Museo del Oro*, *19*, 141–142. https://publicaciones.banrepcultural.org/index.php/bmo/article/view/7206

Rosenberg, N., Murata, M., Ikeda, Y., Opare-Sem, O., Zivelin, A., Geffen, E., & Seligsohn, U. (2002). The Frequent 5,10-Methylenetetrahydrofolate Reductase C677T Polymorphism Is Associated with a Common Haplotype in Whites, Japanese, and Africans. En *Am. J. Hum. Genet* (Vol. 70).

Roychoudhury, A. K., & Nei, Masatoshi. (1988). *Human Polymorphic Genes: World Distribution*. Oxford University Press.

rs1801133 RefSNP Report - dbSNP - NCBI. (2022, septiembre 21). dbSNP Short Genetic Variations. https://www.ncbi.nlm.nih.gov/snp/rs1801133#clinical_significance

rs1801133 (SNP) - Population genetics - Homo_sapiens - Ensembl genome browser 109. (s. f.). Ensembl. Recuperado 18 de febrero de 2023, de https://www.ensembl.org/Homo_sapiens/Variation/Population?r=1:11795821-11796821;v=rs1801133;vdb=variation;vf=1280266

Rozen, R. (1997). Genetic predisposition to hyperhomocysteinemia: deficiency of methylenetetrahydrofolate reductase (MTHFR). Thrombosis and Haemostasis, 78(1), 523–536.

Rubio, S., Pacheco-Orozco, R. A., Gómez, A. M., Perdomo, S., & García-Robles, R. (2020). Secuenciación de nueva generación (NGS) de ADN: presente y futuro en la práctica clínica. *Universitas Médica*, *61*(2), 1–15. https://doi.org/10.11144/javeriana.umed61-2.sngs

Sant Joan de Déu Barceona Hospital. (2014, noviembre 11). *El exoma y su papel en el diagnóstico de errores congénitos del metabolismo*. https://metabolicas.sdjhospitalbarcelona.org/noticia/exoma-su-papel-diagnostico-errores-congenitos-metabolismo

Sasson, M., Lee, M., Jan, C., Fontes, F., & Motta, J. (2014). Prevalence and Associated Factors of Obesity among Panamanian Adults. 1982–2010. *PLOS ONE*, *9*(3), e91689-. https://doi.org/10.1371/journal.pone.0091689

Scheib, C. L., Li, H., Desai, T., Link, V., Kendall, C., Dewar, G., Griffith, P. W., Mörseburg, A., Johnson, J. R., Potter, A., Kerr, S. L., Endicott, P., Lindo, J., Haber, M., Xue, Y., Tyler-Smith, C., Sandhu, M. S., Lorenz, J. G., Randall, T. D., … Kivisild, T. (2018). Ancient human parallel lineages within North America contributed to a coastal expansion. *Science*, *360*(6392), 1024–1027. https://doi.org/10.1126/SCIENCE.AAR6851/SUPPL_FILE/AAR6851_SCHEIB_SM.PDF

Shan, X., Wang, L., Hoffmaster, R., & Kruger, W. D. (1999). Functional characterization of human methylenetetrahydrofolate reductase in Saccharomyces cerevisiae. *Journal of Biological Chemistry*, *274*(46), 32613–32618. https://doi.org/10.1074/jbc.274.46.32613

Sirugo, G., Williams, S. M., & Tishkoff, S. A. (2019). The Missing Diversity in Human Genetic Studies. En *Cell* (Vol. 177, Número 1, pp. 26–31). Cell Press. https://doi.org/10.1016/j.cell.2019.02.048

Skoglund, P., Mallick, S., Bortolini, M. C., Chennagiri, N., Hünemeier, T., Petzl-Erler, M. L., Salzano, F. M., Patterson, N., & Reich, D. (2015). Genetic evidence for two founding populations of the Americas. *Nature*, *525*(7567), 104–108. https://doi.org/10.1038/nature14895

Spina bifida: MedlinePlus Genetics. (s. f.). Recuperado 30 de enero de 2023, de https://medlineplus.gov/genetics/condition/spina-bifida/

Tarazona-Santos, E., Carvalho-Silva, D. R., Pettener, D., Luiselli, D., de Stefano, G. F., Labarga, C. M., Rickards, O., Tyler-Smith, C., Pena, S. D. J., & Santos, F. R. (2001). Genetic differentiation in South Amerindians is related to environmental and cultural diversity: Evidence from the Y chromosome. *American Journal of Human Genetics*, *68*(6), 1485–1496. https://doi.org/10.1086/320601

Tratamiento del cáncer de próstata (PDQ®)–Versión para pacientes - NCI. (s. f.). Recuperado 25 de enero de 2023, de https://www.cancer.gov/espanol/tipos/prostata/paciente/tratamiento-prostata-pdq

UniProt. (s. f.). *MTHFR - Methylenetetrahydrofolate reductase (NADPH) - Homo sapiens (Human) | UniProtKB | UniProt*. UniProtKB. Recuperado 30 de enero de 2023, de https://www.uniprot.org/uniprotkb/P42898/entry

Variant: rs1801133. (s. f.). GWAS Catalog. Recuperado 18 de febrero de 2023, de https://www.ebi.ac.uk/gwas/variants/rs1801133

Vidal, E. A., Moyano, T. C., Bustos, B. I., Pérez-Palma, E., Moraga, C., Riveras, E., Montecinos, A., Azócar, L., Soto, D. C., Vidal, M., di Genova, A., Puschel, K., Nürnberg, P., Buch, S., Hampe, J., Allende, M. L., Cambiazo, V., González, M., Hodar, C., … Gutiérrez, R. A. (2019). Whole Genome Sequence, Variant Discovery and Annotation in Mapuche-Huilliche Native South Americans. *Scientific Reports*, *9*(1). https://doi.org/10.1038/s41598-019-39391-z

Wang, K., Li, M., & Hakonarson, H. (2010). Annovar: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*, *38*(16), e164. https://doi.org/10.1093/NAR/GKQ603

Wang, S., Lewis, C. M., Jakobsson, M., Ramachandran, S., Ray, N., Bedoya, G., Rojas, W., Parra, M. v., Molina, J. A., Gallo, C., Mazzotti, G., Poletti, G., Hill, K., Hurtado, A. M., Labuda, D., Klitz, W., Barrantes, R., Bortolini, M. C., Salzano, F. M., … Ruiz-Linares, A. (2007). Genetic Variation and Population Structure in Native Americans. *PLOS Genetics*, *3*(11), e185. https://doi.org/10.1371/JOURNAL.PGEN.0030185

Wang, X., Fu, J., Li, Q., & Zeng, D. (2016). Geographical and ethnic distributions of the MTHFR C677T, A1298C and MTRR A66G gene polymorphisms in Chinese populations: A meta-analysis. *PLoS ONE*, *11*(4). https://doi.org/10.1371/journal.pone.0152414

Westenberger, S. (2020). Illumina Sequencing Overview: Library Prep to Data Analysis with Scott Westenberger, PhD - View Webinar | Ambry Genetics. En *Ambry Genetics*. Ambry Genetics. https://www.ambrygen.com/providers/webinar/134/illumina-sequencing-overview-library-prep-to-data-analysis-with-scott-westenberger-phd

Whole Exome Sequencing | Detect exonic variants. (s. f.). Illumina. Recuperado 20 de octubre de 2022, de https://www.illumina.com/techniques/sequencing/dna-sequencing/targeted-resequencing/exome-sequencing.html

Witt, M., & Erickson, R. P. (1989). A rapid method for detection of Y-chromosomal DNA from dried blood specimens by the polymerase chain reaction. En *Hum Genet* (Vol. 82).

Wolfe, J., Darling, S. M., Erickson, R. P., Craig, I. W., Buckle, V. J., Rigby, P. W. J., Willard, H. F., & Goodfellow, P. N. (1985). Isolation and characterization of an alphoid centromeric repeat family from the human Y chromosome. *Journal of Molecular Biology*, *182*(4), 477–485. https://doi.org/10.1016/0022-2836(85)90234-7

World Health Organization. (2005). *WHO steps surveillance manual: the WHO stepwise approach to chronic disease risk factor surveillance*. WHO.

World Health Organization. (2016). *Noncommunicable diseases*. https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases

Yamada, K., Chen, Z., Rozen, R., & Matthews, R. G. (2001). Effects of common polymorphisms on the properties of recombinant human methylenetetrahydrofolate reductase. *Proceedings of the National Academy of Sciences*, *98*(26), 14853–14858. https://doi.org/10.1073/pnas.261469998

Zeigler-Johnson, C. M., Spangler, E., Jalloh, M., Gueye, S. M., Rennert, H., & Rebbeck, T. R. (2008). Genetic susceptibility to prostate cancer in men of African descent: implications for global disparities in incidence and outcomes. *The Canadian journal of urology*, *15*(1), 3872–3882.

ANEXOS

- **ANEXO 1: Ejemplo de tabla de resultados de anotación de Illumina**

Tabla 27. Información de la anotación.

CHROM	POS	ID	REF	ALT	QUAL	FILTER	GeneName	Func	Gene	GeneDetail	ExonicFur	AAChang	Gencod	cytoBand	wgRna	genomicSuperDups	Repeat
chr1	16288	rs113141985	C	G	109,77	PASS	WASH7P	ncRNA_introni	NR_024540	.	.	.	.	1p36.33	.	Score=0.993729;Name=chr9:10843	.
chr1	16298	rs2747966	C	T	337,77	PASS	WASH7P	ncRNA_introni	NR_024540	.	.	.	.	1p36.33	.	Score=0.993729;Name=chr9:10843	.
chr1	17614	rs201057270	G	A	111,77	PASS	WASH7P	ncRNA_exonic	NR_024540	.	.	.	.	1p36.33	.	Score=0.993729;Name=chr9:10843	.
chr1	54844	rs372902470	G	A	73,28	PASS	.	intergenic	NR_026818,NM	dist=18763;dis	.	.	.	1p36.33	.	Score=0.994828;Name=chr19:6000	Score=680;Name="70:FLAM

avsnp150	cosmic	CLNAI	CLNDI	CLND	CLNREVST	CLNSIG	gwasCatal	1000g2015aug_	1000g2015aug_	1000g2015aug	1000g2015aug	1000g2015aug	1000g2015aug_	esp6500siv2_
rs113141985	.	.	.	.	.	.	.	.	.	.	.	.	.	.
rs2747966	.	.	.	.	.	.	.	.	.	.	.	.	.	.
rs201057270	.	.	.	.	.	.	.	.	.	.	.	.	.	.
rs372902470	.	.	.	.	.	.	.	.	.	.	.	.	.	.

ExAC_	ExAC_	ExAC_	ExAC_	ExAC_	ExAC_	ExAC_	ExAC_	gnomad_	gnomad_	gnomad_	gnomad_	gnomad_	gnomad_	gnomad_	gnomad_ex	gnomad_	gnomad_	gnomad_	gnomad_	gnomad_	gnomad_genom	gnomad_g	gnomad_g	gnomad_g	gnomad_g	gnomad_g
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	0,1944	0,2789	0,1072	.	0,2934	0,378	0,2169	0,2343	0,1714	0,2133
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	0,5438	0,5377	0,5131	.	0,5794	0,5041	0,5504	0,6067	0,5125	0,5267
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	0,0403	0,0749	0,0124	.	0,0617	0,2262	0,044	0,0242	0,0253	0,0405
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	0,3799	0,3887	0,2652	.	0,3536	0,0883	0,4871	0,4291	0,475	0,4582

SIFT	Polyphen2_HDI	Polyphen2_HV/	LRT	MutationTaster	MutationAsses	FATHMM	phyloP100way_	phyloP20way_n	SiPhy_29way_l	gerp++gt2	CADD
.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.

INFO	FORMAT	P715	Ori_REF	Ori_ALT	OMIM	GWAS_Pubmed	HGMD_ID_Dis	HGMD_mutatio	GO_BP	GO_CC	GO_MF	KEGG_PATHW	PID_PATHWA	BIOCARTA_I	REACTOME_PATI
AC=1;AF=0.5;A	GT:AD:DP:GQ:I	0/1:33,10:43:99:138,0,1449	C	G	.	.	.	.	.	.	.	.	.	.	.
AC=1;AF=0.5;A	GT:AD:DP:GQ:I	0/1:27,11:38:99:366,0,976	C	T	.	.	.	.	.	.	.	.	.	.	.
AC=1;AF=0.5;A	GT:AD:DP:GQ:I	0/1:32,8:40:99:140,0,1114	G	A	.	.	.	.	.	.	.	.	.	.	.
AC=2;AF=1;AN	GT:AD:DP:GQ:I	1/1:0,3:3:9:101,9,0	G	A	.	.	.	.	.	.	.	.	.	.	.

La información de anotación incluye cinco partes: Regiones del cromosoma y estructuras génicas relacionadas con esta variación (1-18), Anotación de la base de datos (19-61), Predicción funcional (62-73), Información básica sobre la variación (74-78), y Anotación de la función y la vía génica (79-89).

El primer punto es regiones cromosómicas y estructuras de genes relacionados con esta variación--El gen donde se localizan los sitios de la variación puede estar asociado con ciertas enfermedades. Novogene anota regiones cromosómicas y estructuras de genes relacionados con esta variación, para ayudar a los profesores a entender la estructura de genes y áreas de información.

Leyenda:

(1) CHROM: Código de identificación del cromosoma.

(2) POS: La posición de la variante en los cromosomas. El valor se refiere a la posición de la primera base en la REF Sting.

(3) ID: El número rs de la variante en dbSNP.

(4) REF: Base(s) de referencia.

(5) ALT: Base(s) alternativa(s). Lista separada por comas de alelos alternativos no de referencia detectados en al menos una de las muestras.

(6) QUAL: Valor de calidad de la variante. Puntuación de calidad en escala de Phred para la afirmación realizada en ALT. Por ejemplo, -10log10 prob (la llamada en ALT es incorrecta).

(7) FILTER: Estado del filtro, PASS si la posición ha pasado todos los filtros.

(8) GeneName: Nombres de los genes en los que se encuentra esta variante según las anotaciones refGene.

(9) Func: Este campo indica si la variante afecta a exones o afecta a regiones intergénicas, o afecta a intrones, o afecta a genes de ARN no codificante. El valor de este campo tiene la siguiente precedencia: exónico = corte y empalme > ARNnc > UTR5/UTR3 > intrónico > aguas arriba/aguas abajo > intergénico. Notas: 1. Cuando una variante afecta a diferentes genes o transcripciones, la variante puede encajar en múltiples categorías funcionales, y entonces se utiliza la precedencia mencionada anteriormente para decidir qué función se imprime; 2. "exónico" aquí se refiere sólo a la porción exónica codificante, pero no a la porción UTR, ya que hay dos palabras clave (UTR5, UTR3) que están específicamente reservadas para las anotaciones UTR; 3. Si una variante se localiza tanto en la región 5'UTR como en la 3'UTR (posiblemente para dos genes diferentes), entonces se imprimirá "UTR5,UTR3" como salida; 4. El término "corte y empalme" en ANNOVAR se define como la variante que se encuentra por defecto a 2 pb del límite exón/intrón; 5. El término "splicing" en ANNOVAR sólo se refiere a los 2 pb del intrón que están cerca de un exón; 6. Los términos " aguas arriba " y " aguas abajo " se definen a 1 kb del sitio de inicio de la transcripción o del sitio final de la transcripción, respectivamente, teniendo en cuenta la cadena del ARNm. Si una variante se localiza tanto en la región aguas abajo como aguas arriba (posiblemente para 2 genes diferentes), entonces se imprimirá como salida 'aguas arriba, aguas abajo'.

(10) Gen: Nombre(s) de la transcripción. Si una variante tiene 'intergenic' en el campo 'Func', este campo dará dos transcripciones vecinas. Si una variante afecta a varias transcripciones con diferentes categorías funcionales, sólo se mostrarán los nombres de las transcripciones de acuerdo con el valor del campo "Func". Por ejemplo, rs333970 afecta al exónico, corte y empalme, intrónico, exónico de los cuatro transcritos del gen CSF1, el valor 'Func' será 'exónico; corte y empalme' y el valor 'Gene' será 'NM_000757, NM_172210, NM_172212' (NM_172211 será ignorado).

(11) GeneDetail: Descripción del cambio de secuencia en UTR, corte y empalme, ARNnc_corte y empalme o región intergénica. Si "Func" es "exonic; splicing" o "splicing", este campo indica el cambio de secuencia en la(s) región(es) de splicing; por ejemplo, NM_172210:exon6:c.1090+5C>A, NM_172210 es el identificador del transcrito; exon6:c.1090+5C>A es el cambio de secuencia y significa que esta sustitución C>A se encuentra en la quinta base aguas abajo del sexto exón (1090 es la posición final del sexto exón del ADNc). Si 'Func' es 'intergénico', este campo da la distancia a los transcritos vecinos, como 'dist=1366; dist=22344'. Si 'Func' es 'UTR*', este campo da el cambio de secuencia en UTR; por ejemplo, NM_198576:c.*19C>T significa que esta sustitución C>T está en la 19ª base aguas abajo del codón de parada en NM_198576.

(12) ExonicFunc: Este campo indica las consecuencias funcionales de la variante (valores posibles: SNV sin sentido, SNV sinónimo, inserción con desplazamiento de marco, deleción con desplazamiento de marco, inserción sin desplazamiento de marco, deleción sin desplazamiento de marco, sustitución con desplazamiento de marco, sustitución sin desplazamiento de marco, ganancia de codón de parada, pérdida de codón de parada, desconocido).

(13) AAChange: Este campo indica los cambios de aminoácidos como resultado de la variante exónica. Sólo las variantes exónicas tienen información en este campo, es decir, cuando "Func" es "exónica" o "exónica; corte y empalme", este campo da el cambio de aminoácido en cada transcripción relacionada. Por ejemplo, AIM1L:NM_001039775:exon2:c.C2768T:p.P923L, AIM1L es el nombre del gen; NM_001039775 es el identificador del transcrito; exon2 significa que esta variante se encuentra en el segundo exón de NM_001039775; c. C2768T es el cambio de secuencia y significa que esta sustitución C>T está en la posición 2.768 del ADNc; p.P923L es el cambio de aminoácido y significa que el aminoácido 923 de la proteína cambia de Pro a Leu debido a esta variante. Otro ejemplo, NADK:NM_001198995:exon10:c.1240_1241insAGG:p.G414delinsEG, c.1240_1241insAGG es el cambio de secuencia y significa que hay una inserción de 3 pb entre la posición 1.240 y 1.241 en el ADNc; p.G414delinsEG es el cambio de aminoácido y significa que Gly en el aminoácido 414 de la proteína se cambia a Glu-Gly.

(14) Gencode: El(los) nombre(s) de la(s) transcripción(es) en la(s) que se localiza esta variante según las definiciones de genes Gencode.

(15) cytoband: Este campo indica las bandas cromosómicas teñidas con Giemsa. Cuando una variante abarca varias bandas, éstas se conectarán mediante un guión (por ejemplo, 1q21.1-q23.3).

(16) wgRna: Nombres de genes de ARN pequeños nucleolares y microARNs basados en la Liberación miRBase y snoRNABase.

(17) genomicSuperDups: Este campo indica si la variante coincide con duplicaciones segmentarias en el genoma de referencia. Las variantes que coinciden con duplicaciones segmentarias son probablemente errores de alineamiento de secuencias y deben tratarse con extrema precaución. El campo "Puntuación" en la salida es la identidad de secuencia que va de 0 a 1 entre dos segmentos genómicos. El campo 'Nombre' representa los otros segmentos "coincidentes" en el genoma. Por ejemplo, 'Score=0.994828; Name=chr19:60000' significa que el fragmento en la posición de chr19:60000 es homólogo al fragmento que contiene esta variante, y la identidad de secuencia es 0.994828. Tenga en cuenta que, para que una región se incluya en las duplicaciones segmentarias, al menos 1 Kb de la secuencia total (que contenga al menos 500 pb de secuencia no RepeatMasked) debe alinearse y se requiere una identidad de secuencia de al menos el 90%.

(18) Repetición: Este campo indica si la variante encuentra repeticiones intercaladas y secuencias de ADN de baja complejidad producidas por el programa RepeatMasker, como SINE, LINE y repeticiones simples. Por ejemplo, 'Score=180;Name=1385:(CACCC) n(Simple_repeat)', 180 es la puntuación de la repetición, (CACCC) n es el nombre de la repetición, 'Simple_repeat' es el tipo de repetición. Tenga en cuenta que las variantes asignadas a repeticiones son probablemente falsas y deben tratarse con extrema precaución.

El segundo punto es la anotación de la base de datos: existe un gran número de polimorfismos comunes en la población humana, mientras que muchas variantes deletéreas son raras o de baja frecuencia. Esta parte proporciona la frecuencia alélica y la información clínica de cada variante.

Leyenda:

(19) avsnp150: El número RS de la variante en la base de datos dbSNP (build 150).

(20) cosmic70: El Código de identificación Cosmic de la variante en la base de datos Catalogue Of Somatic Mutations In Cancer (COSMIC).

(21) CLNALLELEID: El Código de identificación del alelo en ClinVar.

(22) CLNDN: El nombre de enfermedad preferido de ClinVar para el concepto especificado por los identificadores de enfermedad en CLNDISDB.

(23) CLNDISDB: Pares etiqueta-valor del nombre e identificador de la base de datos de enfermedades, por ejemplo, OMIM: NNNNNN.

(24) CLNREVSTAT: Significancia clínica ClinVar para el Código de identificación de la variación.

(25) CLNSIG: Significancia clínica para esta variante única.

(26) gwasCatalog: Este campo indica si esta variante se ha asociado previamente a enfermedades o rasgos en estudios de asociación de genoma completo. Enumera los nombres de las enfermedades relacionadas con esta variación. "." significa que esta variación no ha sido reportada por un estudio GWAS publicado.

(27) 1000g2015aug_eas: Este campo proporciona la frecuencia alélica para el alelo en ALT en el 1000 Genomes Project (publicado en agosto de 2015) en la población de Asia Oriental.

(28) 1000g2015aug_sas: Este campo proporciona la frecuencia alélica para el alelo en ALT en el 1000 Genomes Project (publicado en agosto de 2015) en la población de Asia Meridional.

(29) 1000g2015aug_eur: Este campo proporciona la frecuencia alélica para el alelo de ALT en el 1000 Genomes Project (publicado en agosto de 2015) en la población europea.

(30) 1000g2015aug_afr: Este campo proporciona la frecuencia alélica para el alelo en ALT en el 1000 Genomes Project (publicado en agosto de 2015) en la población africana.

(31) 1000g2015aug_amr: Este campo da la frecuencia alélica para el alelo en ALT en 1000 Genomes Project (publicado en agosto de 2015) en población de Americanos Mestizos.

(32) 1000g2015aug_all: Este campo da la frecuencia alélica para el alelo en ALT en 1000 Genomes Project (publicado en agosto de 2015) en toda la población (ALL).

(33) esp6500siv2_all: El ESP es un proyecto de secuenciación del exoma financiado por el NHLBI que tiene como objetivo identificar variantes genéticas en regiones exónicas de más de 6000 individuos, incluyendo sanos, así como sujetos con diferentes enfermedades. Este campo da la frecuencia alélica alternativa para la variante en ESP.

(34) ExAC_ALL: ExAC es la abreviatura de Exome Aggregation Consortium. El conjunto de datos abarca 60.706 individuos no emparentados y debería servir como conjunto de referencia útil de frecuencias alélicas para estudios de enfermedades graves. Los grupos de población admitidos actualmente son ALL, AFR (africanos), AMR (americanos mixtos), EAS (asiáticos orientales), FIN (finlandeses), NFE (europeos no finlandeses), OTH (otros) y SAS (sudasiáticos). ExAC_ALL proporciona la frecuencia alélica alternativa para la variación en TODAS las muestras ExAC.

(35) ExAC_AFR: La frecuencia alélica alternativa para la variación en ExAC para población africana.

(36) ExAC_AMR: La frecuencia alélica alternativa para la variación en ExAC para la población de Americanos Mestizos.

(37) ExAC_EAS: La frecuencia alélica alternativa para la variación en ExAC para la población de Asia Oriental.

(38) ExAC_FIN: La frecuencia alélica alternativa para la variación en ExAC para la población finlandesa.

(39) ExAC_NFE: La frecuencia alélica alternativa para la variación en ExAC para la población europea no finlandesa.

(40) ExAC_OTH: La frecuencia alélica alternativa para la variación en ExAC para otra población.

(41) ExAC_SAS: La frecuencia alélica alternativa para la variación en ExAC para la población del Sur de Asia.

(42) gnomAD_exome_AF: La Base de Datos de Agregación Genómica (gnomAD), es una coalición de investigadores que buscan agregar y armonizar los datos de secuenciación genómica y exómica de una variedad de proyectos de secuenciación a gran escala, y hacer que los datos resumidos estén disponibles para la comunidad científica en general. Esta versión incluye dos conjuntos de llamadas: datos de secuenciación del exoma de 123.136 individuos y secuenciación del genoma completo de 15.496 individuos. Se utilizó GenomAD 2.1.1 para el análisis de anotación de los sitios de mutación. gnomAD_exome_AF proporciona la frecuencia alélica para la variación en todas las muestras de exoma.

(43) gnomAD_exome_AF_raw: La frecuencia alélica para la variación en gnomAD_exome sin filtrar.

(45) gnomAD_exome_AF_sas: La frecuencia alélica para la variación en gnomAD_exome para la población del sur de Asia.

(46) gnomAD_exome_AF_amr: La frecuencia alélica para la variación en gnomAD_exome para la población Latina/Americana Mestiza.

(47) gnomAD_exome_AF_eas: La frecuencia alélica para la variación en gnomAD_exome para la población de Asia Oriental.

(48) gnomAD_exome_AF_nfe: La frecuencia alélica para la variación en gnomAD_exome para la población europea no finlandesa.

(49) gnomAD_exome_AF_fin: La frecuencia alélica para la variación en gnomAD_exome para la población finlandesa.

(50) gnomAD_exome_AF_asj: La frecuencia alélica para la variación en gnomAD_exome para la población judía Ashkenazi.

(51) gnomAD_exome_AF_oth: La frecuencia alélica para la variación en gnomAD_exome para Otra (población no asignada) población.

(52) gnomAD_genome_AF: gnomAD_genome_AF da la frecuencia alélica para la variación en todas las muestras del genoma de gnomAD.

(53) gnomAD_genome_AF_raw: La frecuencia alélica para la variación en gnomAD_genoma sin filtrar.

(54) gnomAD_genoma_AF_afr: La frecuencia alélica para la variación en gnomAD_genome para la población africana.

(55) gnomAD_genome_AF_sas: La frecuencia alélica para la variación en gnomAD_genoma para la población del sur de Asia.

(56) gnomAD_genome_AF_amr: La frecuencia alélica para la variación en gnomAD_genome para población Latina/Americana Mestiza.

(57) gnomAD_genome_AF_eas: La frecuencia alélica para la variación en gnomAD_genome para la población de Asia Oriental.

(58) gnomAD_genome_AF_nfe: La frecuencia alélica para la variación en gnomAD_genoma para la población europea no finlandesa.

(59) gnomAD_genome_AF_fin: La frecuencia alélica para la variación en gnomAD_genoma para la población finlandesa.

(60) gnomAD_genome_AF_asj: La frecuencia alélica para la variación en gnomAD_genome para la población judía Ashkenazi.

(61) gnomAD_genoma_AF_oth: La frecuencia alélica para la variación en gnomAD_genome para población Otra (población no asignada).

El tercer punto es la predicción funcional: estas anotaciones pueden ayudar a evaluar la nocividad de una variación. Nota 1: SIFT, Polyphen2, MutationTaster, LRT, MutationAssessor y FATHMM son similares y todos predicen si una sustitución de aminoácidos afecta a la función de la proteína; sólo las variantes codificantes tienen estas anotaciones. Nota 2: phyloP, SiPhy, gerp++ y CADD son similares y predicen el nivel de conservación del sitio; estos tipos de "puntuaciones de conservación" sólo consideran el nivel de conservación en la base actual, y no se preocupan por la verdadera identidad del nucleótido, por lo que las variantes sinónimas y no sinónimas en el mismo sitio recibirán la misma puntuación; estas puntuaciones se utilizan para encontrar sitios funcionalmente importantes, por lo que las variantes que confieren una mayor susceptibilidad pueden recibir una buena puntuación.

Leyenda:

(62) SIFT: anotación SIFT (dbNSFP versión 3.3a). La anotación consta de puntuación y predicción categórica; las puntuaciones y las predicciones están separadas por comas. Hay dos predicciones posibles: D (Deletéreo, puntuación<=0,05)); T (Tolerado, puntuación>0,05).

(63) Polyphen2_HDIV: anotación PolyPhen 2 (dbNSFP versión 3.3a) basada en la base de datos HumanDiv. Esta anotación debe utilizarse al evaluar alelos raros en loci potencialmente implicados en fenotipos complejos, cartografía densa de regiones identificadas por estudios de asociación de todo el genoma y análisis de selección natural a partir de datos de secuencias. La anotación consiste en una puntuación y una predicción categórica. Hay tres predicciones posibles: D (probablemente perjudicial, puntuación>=0,957), P (posiblemente perjudicial, 0,453<=puntuación<=0,956), B (benigno, puntuación<=0,452).

(64) Polyphen2_HVAR: anotación PolyPhen 2 (dbNSFP versión 3.3a) basada en la base de datos HumanVar. Esta anotación debe utilizarse para el diagnóstico de enfermedades mendelianas. La anotación consiste en una puntuación y una predicción categórica. Hay tres predicciones posibles: D (probablemente dañino, puntuación>=0,909), P (posiblemente dañino, 0,447<=puntuación<=0,909), B (benigno, puntuación<=0,446).

(65) LRT: anotación LRT (dbNSFP versión 3.3a). La anotación consta de puntuación y predicción categórica. Hay tres predicciones posibles: D (Deletéreo), N (Neutro), U (Desconocido).

(66) MutationTaster: Anotación MutationTaster (dbNSFP versión 3.3a). La anotación consiste en una puntuación y una predicción categórica. Hay cuatro predicciones posibles: 'A' (Enfermedad_causante_automática), 'D (Enfermedad_causante), 'N' (Polimorfismo), 'P' (Polimorfismo_automático'). D y N se categorizan sólo por la puntuación, mientras que A y P se categorizan por la puntuación y otra información (si el SNV no sinónimo conduce a una ganancia de codón de parada, la variación se predecirá como 'A'; si los tres genotipos del SNV no sinónimo tienen información de frecuencia en HapMap, la variación se predecirá como 'P'). Por lo tanto, tanto A como D deben considerarse deletéreas.

(67) MutationAssessor: Anotación MutationAssessor (dbNSFP versión 3.3a). La anotación consiste en una puntuación y una predicción categórica. Hay cuatro predicciones posibles: H (alta), M (media), L (baja), N (neutra). H/M significa funcional y L/N significa no funcional.

(68) FATHMM: anotación FATHMM (dbNSFP versión 3.3a). La anotación consiste en una puntuación y una predicción categórica. Hay dos predicciones posibles: D (Deletéreo, puntuación<=-1,5)); T (Tolerado, puntuación>-1,5).

(69) phyloP100way_vertebrate: Puntuación PhyloP (dbNSFP versión 3.3a) basada en el alineamiento del genoma completo de 100 vertebrados. En general, cuanto mayor es la puntuación, más conservado está el sitio.

(70) phyloP20way_mammalian: Puntuación PhyloP (dbNSFP versión 3.3a) basada en la alineación del genoma completo de 20 mamíferos.

(71) SiPhy_29way_logOdds: Puntuación SiPhy (dbNSFP versión 3.3a) basada en la alineación del genoma completo de 29 genomas de mamíferos. Cuanto mayor es la puntuación, más conservado está el sitio.

(72) gerp++gt2: Puntuaciones GERP++ para todas las mutaciones con GERP++>2 en el genoma humano, ya que este umbral suele considerarse evolutivamente conservado y potencialmente funcional. Las variantes con '.' en este campo deben considerarse no conservadas. Cuanto mayor es la puntuación, más conservado está el sitio.

(73) CADD: Puntuación CADD (Combined Annotation Dependent Depletion). En la salida, los valores delimitados por comas son puntuaciones brutas y puntuaciones escaladas por phred. Para las puntuaciones en escala de grises, 10 significa el percentil 10% de las puntuaciones más altas, 20 significa el percentil 1% de las puntuaciones más altas y 30% significa el percentil 0,1% de las puntuaciones más altas. El sitio web oficial del CADD sugiere

15 como límite; en los estudios publicados, se utiliza 10 o 15 como límite. Nota: Todas las puntuaciones CADD son de la versión 3.3a de dbNSFP y Novogene sólo da puntuaciones no inferiores a 10; por lo tanto, los SNV anotados con "." pueden ser SNV no sinónimos o SNV de sitio de corte y empalme con una puntuación CADD inferior a 10 o ser SNV sinónimos sin puntuación CADD en dbNSFP.

El cuarto punto es la información básica sobre la variación: esta parte muestra la información detallada de los sitios de variación, incluidas las profundidades alélicas, los tipos de base de ADN antes y después de la mutación, y la información sobre el genotipo, etc. Esta información desempeña un papel fundamental en el análisis del pedigrí.

Leyenda:

(74) INFO: Información sobre esta variación procedente del software de llamada de variantes. Los campos INFO se codifican como una serie separada por punto y coma de claves cortas con valores opcionales en el formato: <clave>=<datos>[datos].

(75) FORMAT: El campo FORMAT especifica los tipos de datos y el orden (cadena alfanumérica separada por dos puntos). Le sigue un campo por muestra, con los datos separados por dos puntos correspondientes a los tipos especificados en el FORMAT.

GT: genotipo, codificado como valores alélicos separados por / o |. Los valores alélicos son 0 para el alelo de referencia (lo que hay en el campo Ori_REF), 1 para el primer alelo listado en Ori_ALT, 2 para el segundo alelo listado en Ori_ALT y así sucesivamente. 0/0 y 1/1 representan homocigotos. 0/1 representa heterocigoto. '.' significa que no se puede realizar una llamada para una muestra en un locus determinado.

AD: Profundidades alélicas para los alelos ref y alt en el orden listado (Allelic depths).

DP: Profundidad de lectura aproximada (se filtran las lecturas con MQ=255 o con malas parejas).

GQ: Calidad del genotipo.

PL: Probabilidades normalizadas, escaladas por Phred para genotipos como se define en la especificación VCF.

(76) SampleID: Los datos separados por dos puntos de esta muestra corresponden a los tipos especificados en el FORMATO.

(77) Ori_REF: El alelo de referencia (lo que aparece en el campo REF) en el archivo VCF. De acuerdo con el flujo de trabajo de anotación en Novogene (mencionado anteriormente), para InDel, el alelo en el campo "REF" en este archivo puede ser diferente (normalmente más corto) que el "REF" en el archivo VCF.

(78) Ori_ALT: El(los) alelo(s) alternativo(s) (lo que está en el campo ALT) en el archivo VCF. En este archivo, el alelo en el campo "ALT" corresponde a un alelo en el campo "Ori_ALT"; de acuerdo con el flujo de trabajo de anotación en Novogene (mencionado

anteriormente), para InDel, el alelo en el campo "ALT" puede ser diferente de (normalmente más corto que) el alelo correspondiente en el campo "Ori_ALT".

El quinto elemento es la función del gen y la anotación de la ruta: Estas anotaciones son para genes que contienen esta variación.

Leyenda:

(79) OMIM: Anotación de Online Mendelian Inheritance in Man (OMIM).

(80) GWAS_Pubmed_pValue: Anotación del catálogo GWAS del NHGRI-EBI. El valor es como 'pubmedID(p-value);pubmedID(p-value)'. pubmedID' es el Código de identificación de PubMed de la publicación del estudio que informó de la asociación entre la variación y la enfermedad. p-value'' es el valor p correspondiente en la publicación.

(81) HGMD_ID_Diseasename: anotación de la base de datos de mutaciones genéticas humanas (HGMD®). El valor es como ID_HGMD(Nombre_enfermedad); ID_HGMD(Nombre_enfermedad)'. ID_HGMD es el identificador interno de HGMD. Nombre_enfermedad es el nombre de la enfermedad o afección asociada a la mutación.

(82) HGMD_mutation: Anotación de la base de datos de mutaciones genéticas humanas (HGMD®). El valor es información sobre esta variante.

(83-85) GO_BP, GO_CC, GO_MF:Anotación de Gene Ontology. BP es Proceso Biológico; CC es componente celular; MF es función molecular.

(86) KEGG_PATHWAY: Anotación de la base de datos KEGG PATHWAY.

(87) PID_PATHWAY: Anotación de PID (Pathway Interaction Database).

(88) BIOCARTA_PATHWAY: Anotación de BioCarta.

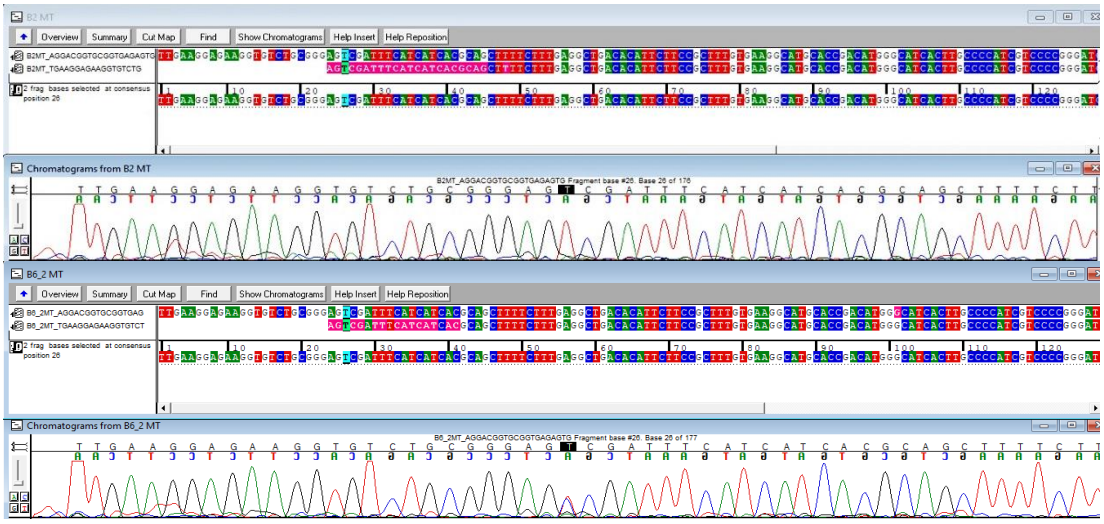
(89) REACTOME_PATHWAY: Anotación de Reactome Pathway Database.

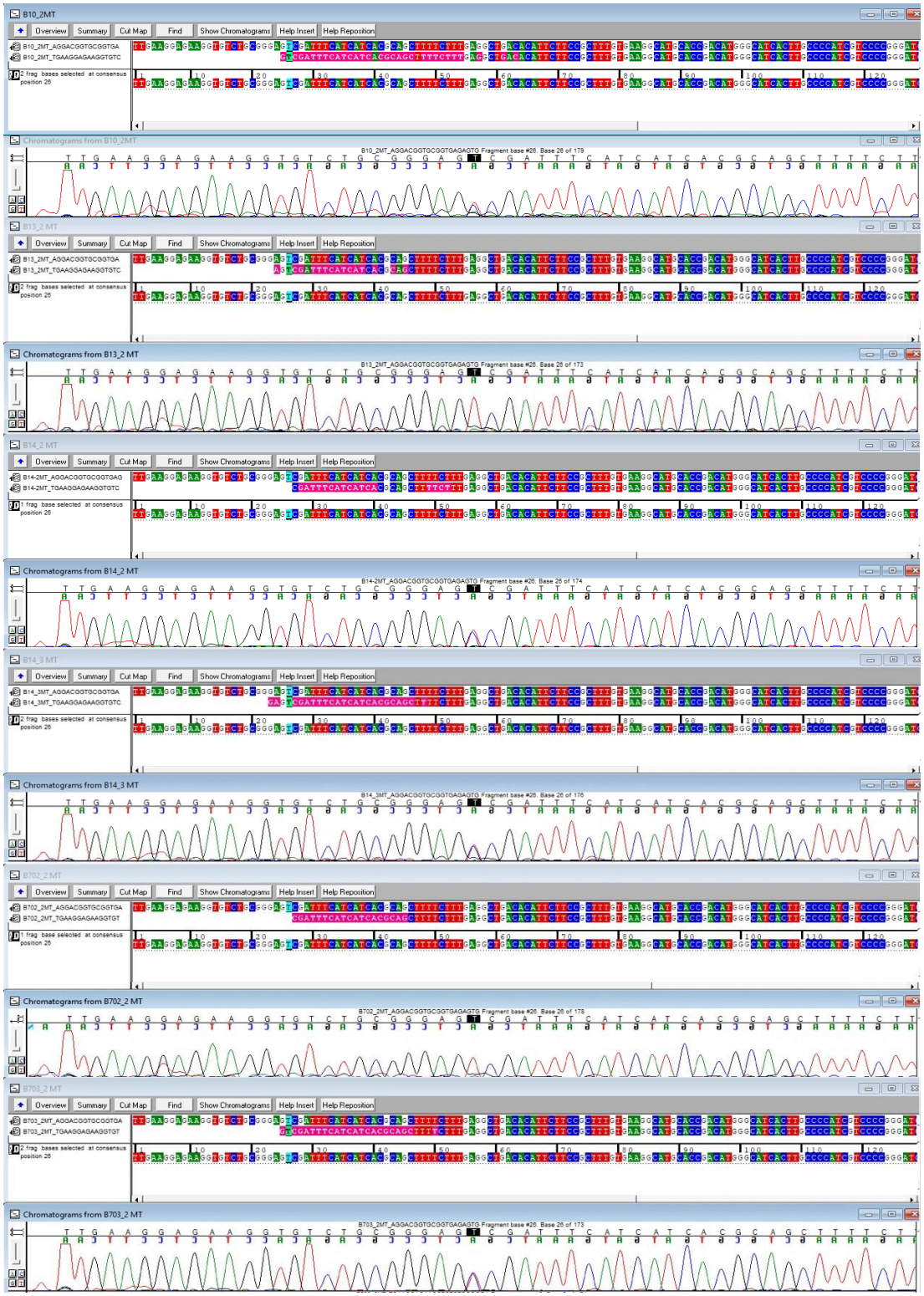
- **ANEXO 2: Programas usados en los análisis realizados por *Novogene***

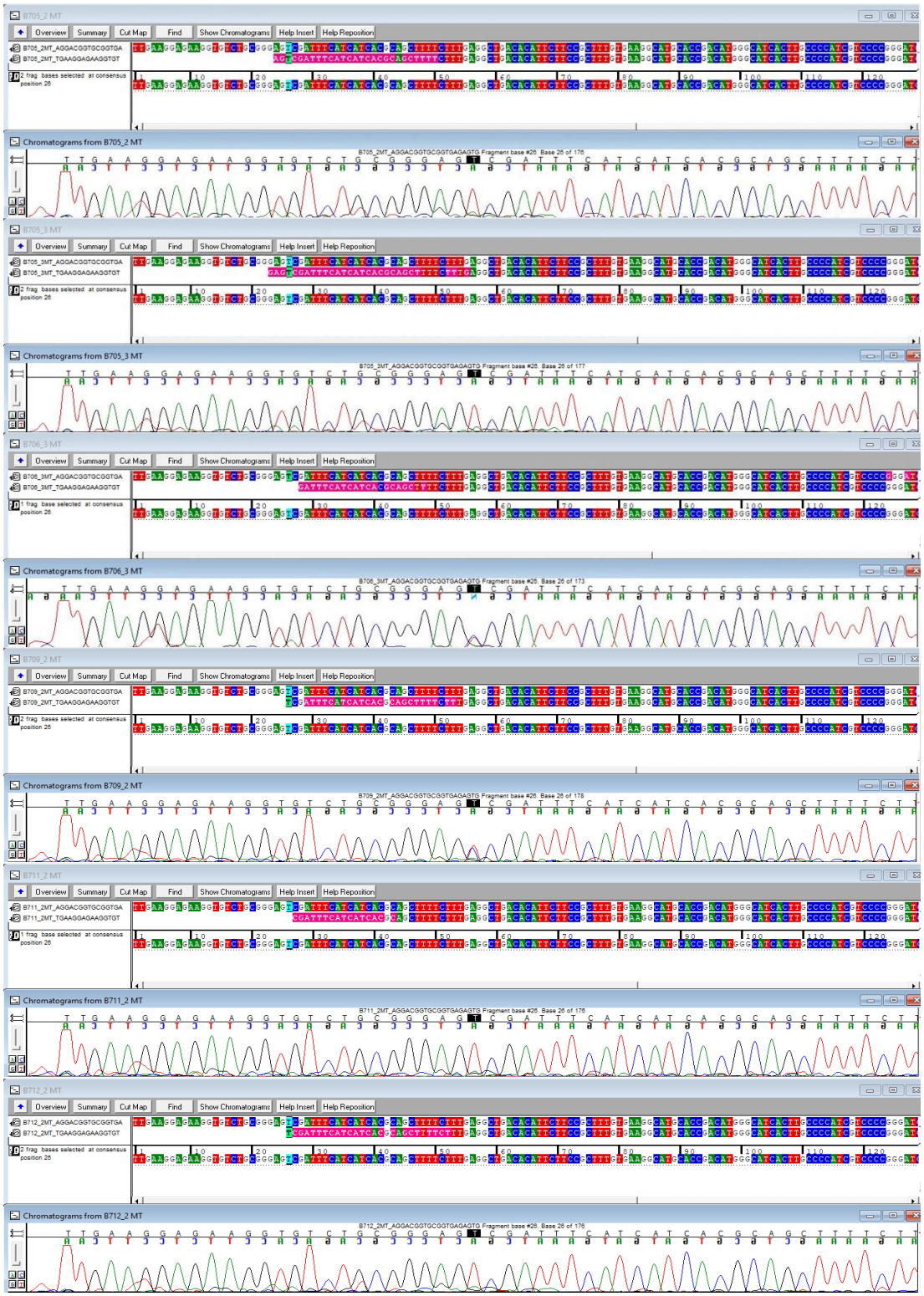
Tabla 28. Principales programas usados en el análisis de secuenciación del genoma completo.

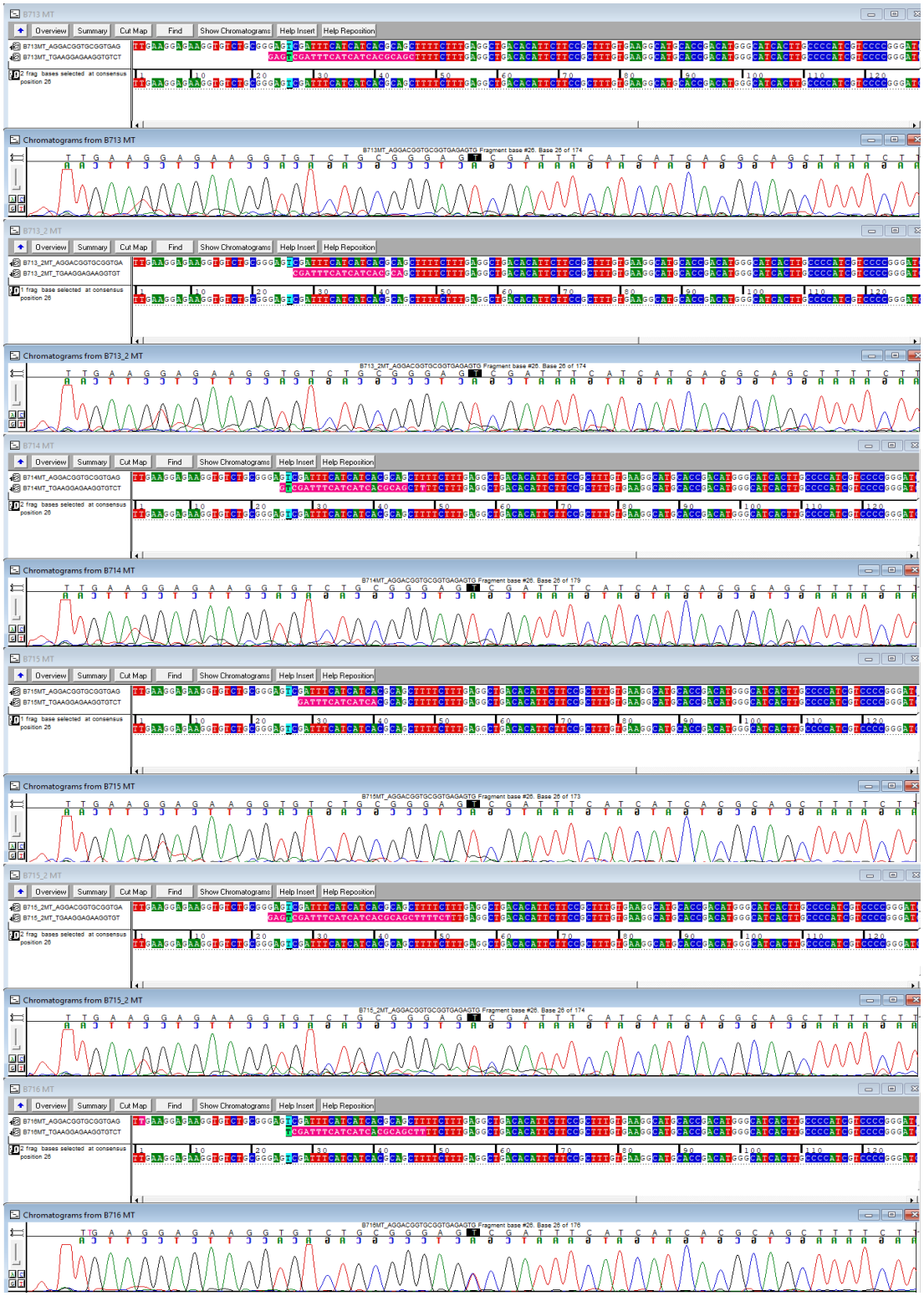
Contenido analítico	Programas	Comentarios	Versión
Alineamiento	BWA	Mapea las lecturas de secuenciación al genoma de referencia, y envía los archivos de los alineamientos en el formato bam	v0.7.17
Samtools	Ordena los archivos bam	v1.8	
Picard	Une todos los archivos bam desde las mismas muestras y marca las lecturas duplicadas	v2.18.9	
Detección de SNP/InDel	GATK	Detecta y filtra SNPs/InDels	v4.0
Detección de SV	DELLY	Detecta SVs	v0.8.7
Detección de CNV	control-FREEC	Detecta CNVs	v11.4
Anotación	ANNOVAR	Anota variantes	2015Dec14

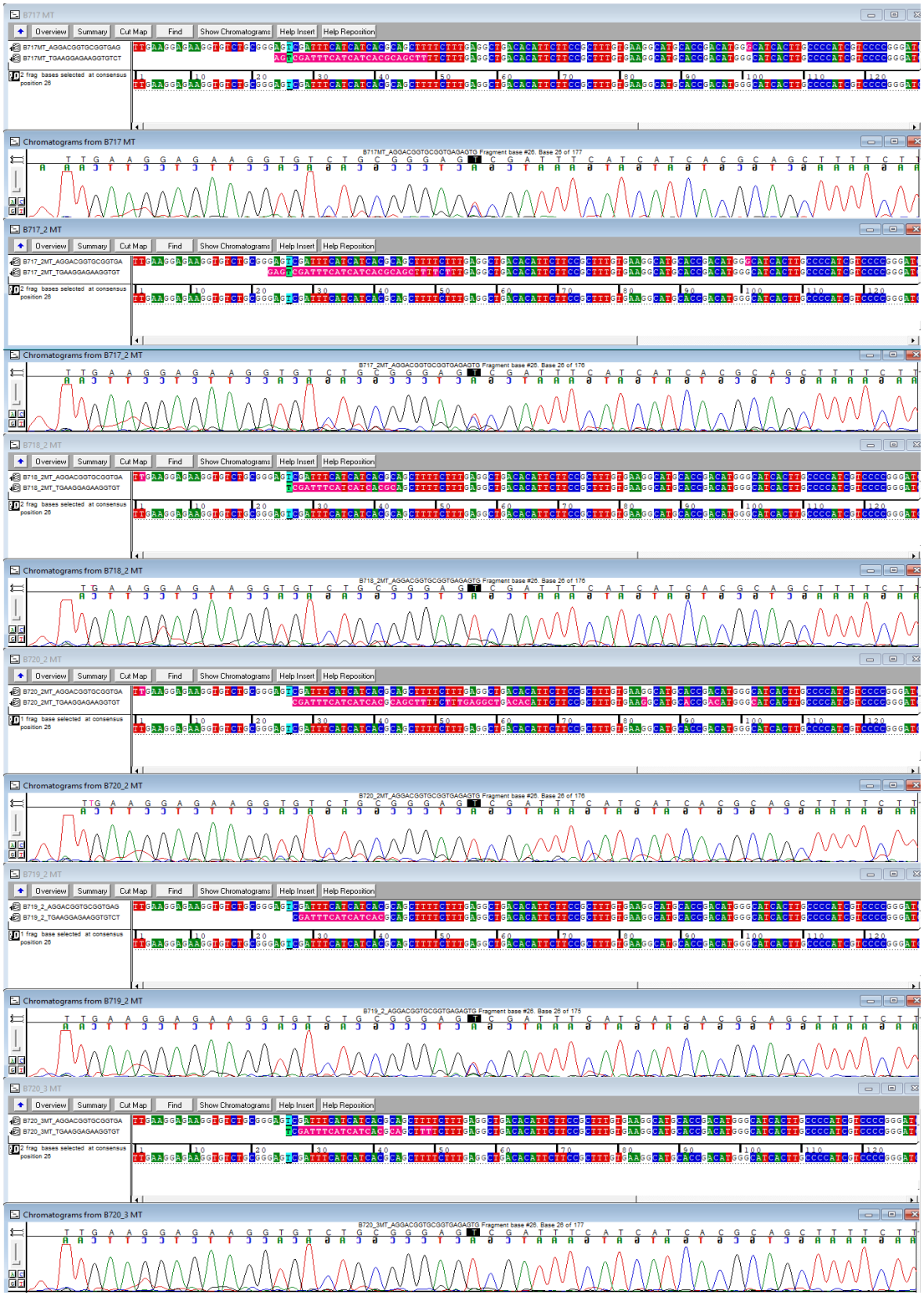
- **ANEXO 3: SNP encontrado en el cromatograma de Sequencher de las muestras**

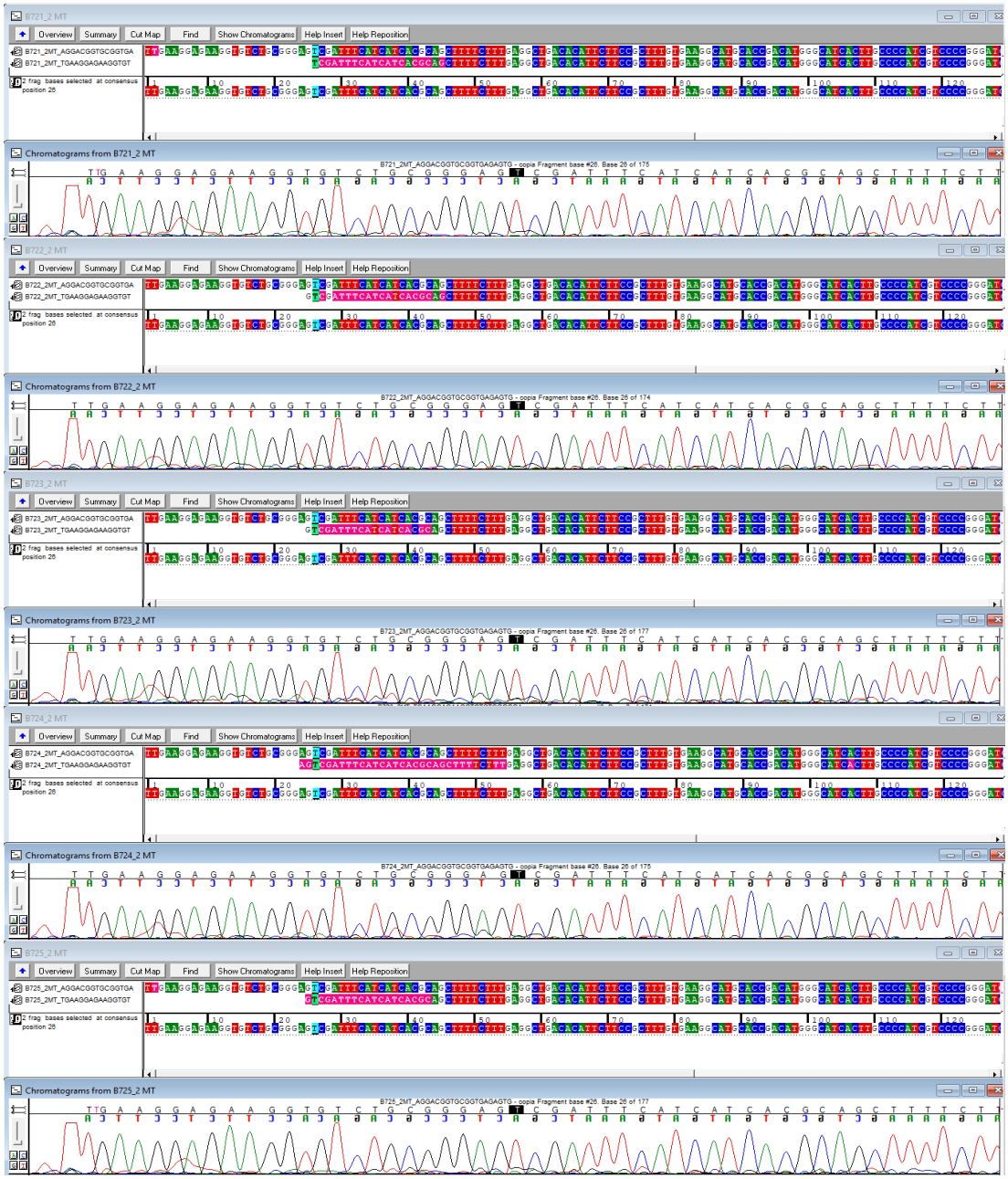


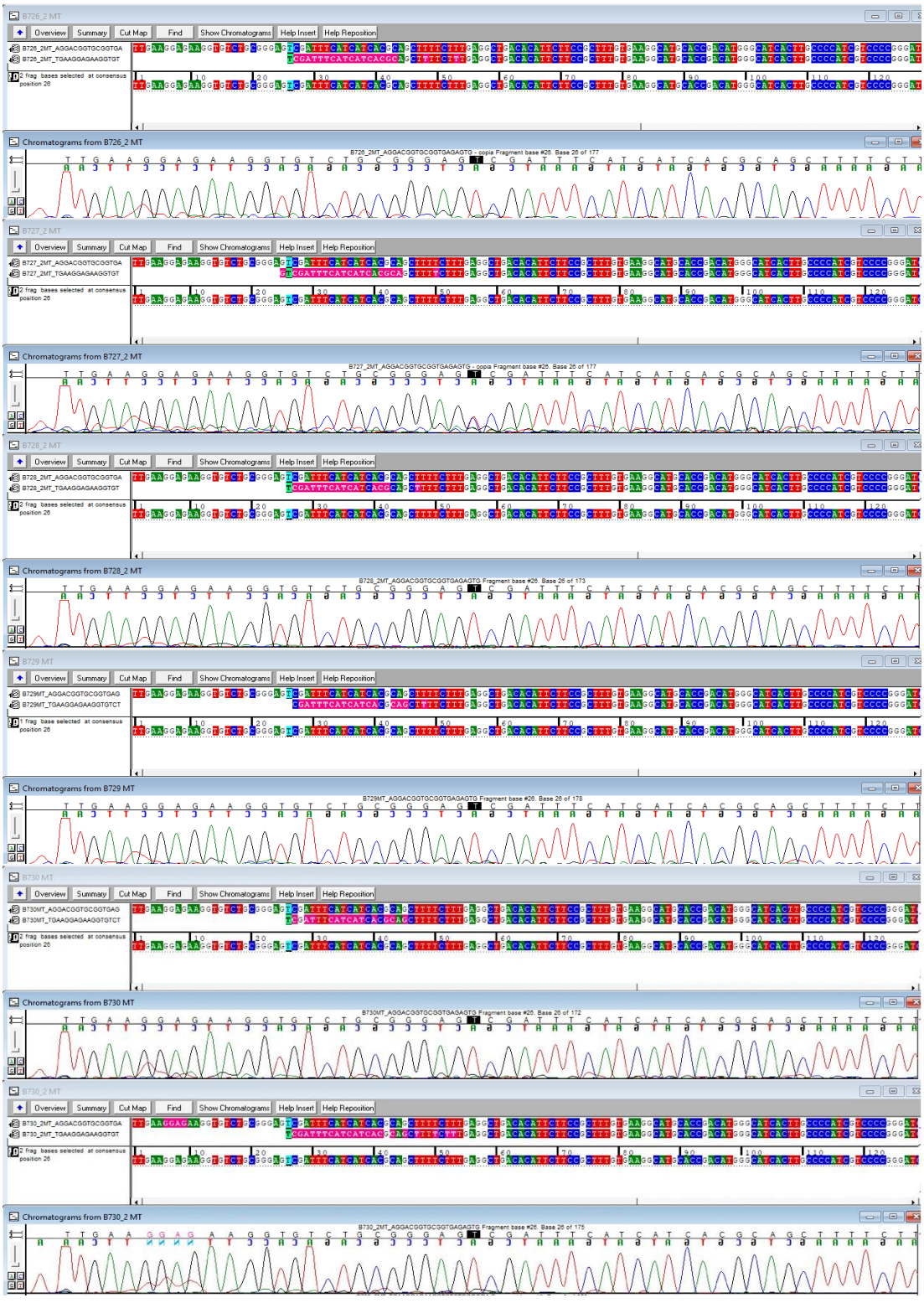

B10_2MT
Overview
Summary
Cut Map
Find
Show Chromatograms
Help Insert
Help Reposition
Chromatograms from B10_2MT
B13_2 MT
Chromatograms from B13_2 MT
B14_2 MT
Chromatograms from B14_2 MT
B14_3 MT
Chromatograms from B14_3 MT
B702_2 MT
Chromatograms from B702_2 MT
B703_2 MT
Chromatograms from B703_2 MT

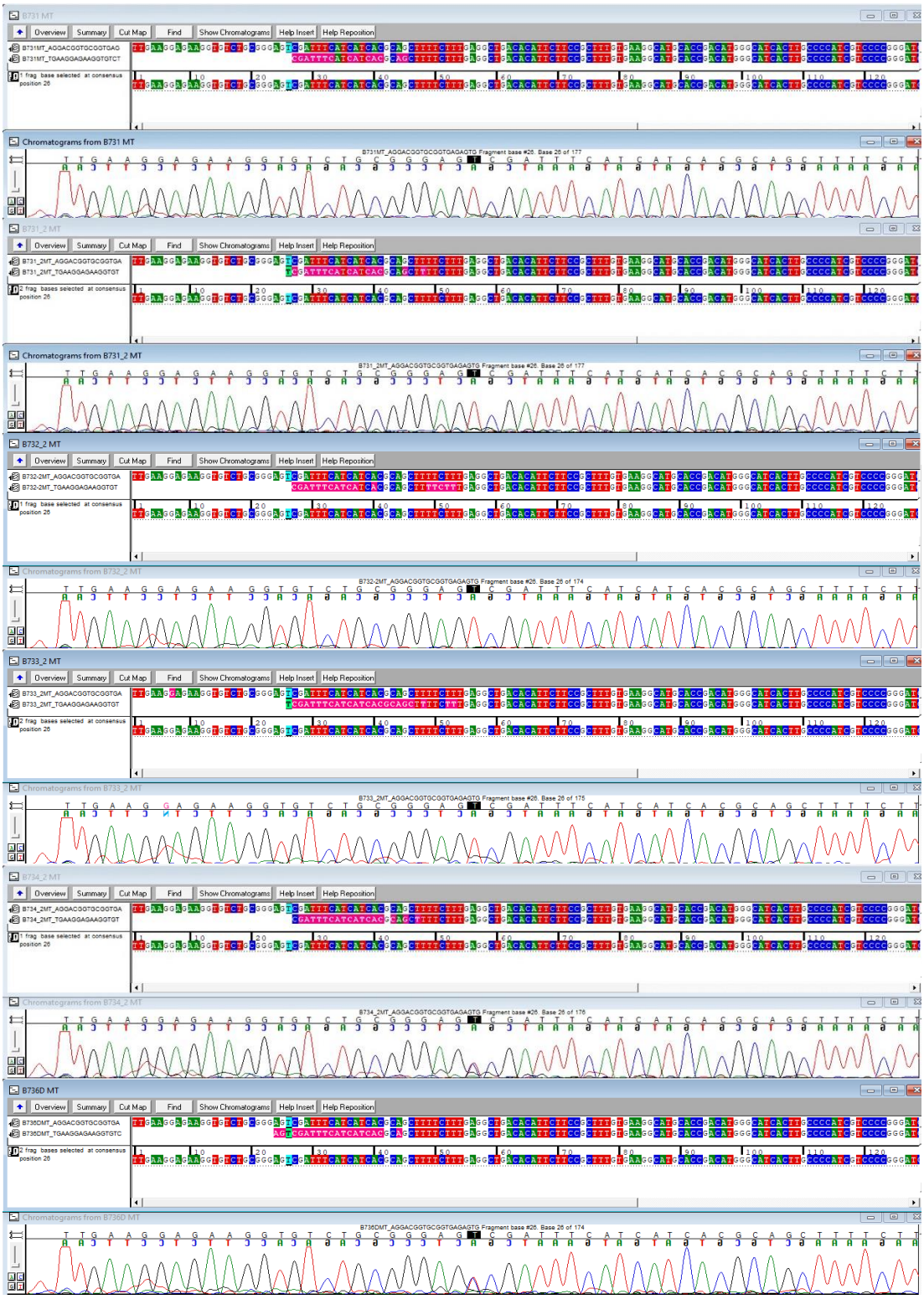

B705_2 MT
Overview
Summary
Cut Map
Find
Show Chromatograms
Help Insert
Help Reposition
B705_2MT_AGGACGGTGCGGTGA
B705_2MT_TGAAGGAGAAGGTGT
Chromatograms from B705_2 MT
B705_2MT_AGGACGGTGCGGTGAGAGTG Fragment base #26, Base 26 of 176
B705_3 MT
B705_3MT_AGGACGGTGCGGTGA
B705_3MT_TGAAGGAGAAGGTGT
Chromatograms from B705_3 MT
B705_3MT_AGGACGGTGCGGTGAGAGTG Fragment base #26, Base 26 of 177
B706_3 MT
B706_3MT_AGGACGGTGCGGTGA
B706_3MT_TGAAGGAGAAGGTGT
Chromatograms from B706_3 MT
B706_3MT_AGGACGGTGCGGTGAGAGTG Fragment base #26, Base 26 of 173
B709_2 MT
B709_2MT_AGGACGGTGCGGTGA
B709_2MT_TGAAGGAGAAGGTGT
Chromatograms from B709_2 MT
B709_2MT_AGGACGGTGCGGTGAGAGTG Fragment base #26, Base 26 of 178
B711_2 MT
B711_2MT_AGGACGGTGCGGTGA
B711_2MT_TGAAGGAGAAGGTGT
Chromatograms from B711_2 MT
B711_2MT_AGGACGGTGCGGTGAGAGTG Fragment base #26, Base 26 of 176
B712_2 MT
B712_2MT_AGGACGGTGCGGTGA
B712_2MT_TGAAGGAGAAGGTGT
Chromatograms from B712_2 MT
B712_2MT_AGGACGGTGCGGTGAGAGTG Fragment base #26, Base 26 of 176

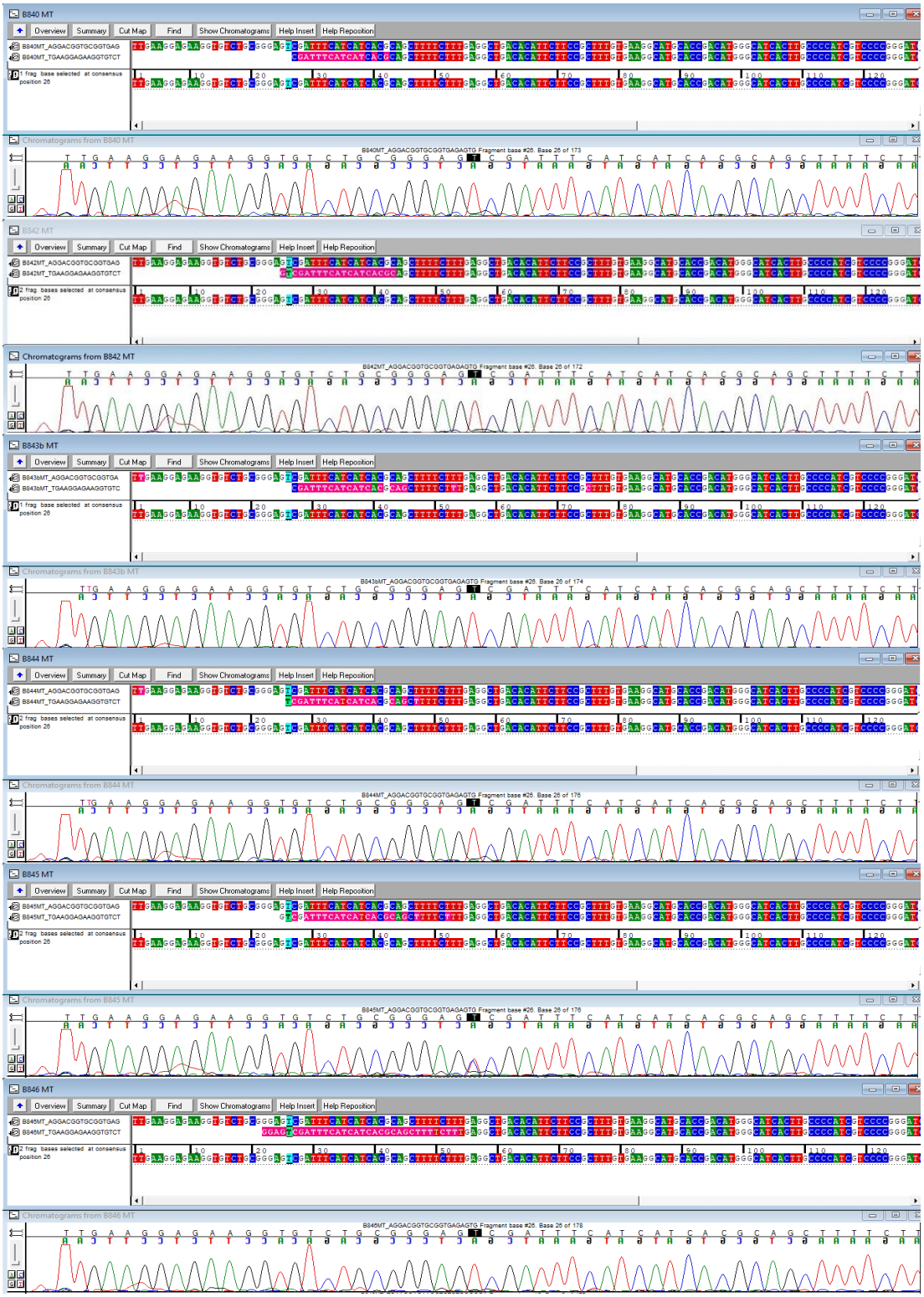

B713 MT
Overview
Summary
Cut Map
Find
Show Chromatograms
Help Insert
Help Reposition
B713MT_AGGACGGTGCGGTGAG
B713MT_TGAAGGAGAAGGTGTCT
2 frag. bases selected at consensus position 26
Chromatograms from B713 MT
B713MT_AGGACGGTGCGGTGAGAGTG Fragment base #26. Base 26 of 174
B713_2 MT
B713_2MT_AGGACGGTGCGGTGA
B713_2MT_TGAAGGAGAAGGTGT
1 frag. base selected at consensus position 26
Chromatograms from B713_2 MT
B713_2MT_AGGACGGTGCGGTGAGAGTG Fragment base #26. Base 26 of 174
B714 MT
B714MT_AGGACGGTGCGGTGAG
B714MT_TGAAGGAGAAGGTGTCT
Chromatograms from B714 MT
B714MT_AGGACGGTGCGGTGAGAGTG Fragment base #26. Base 26 of 179
B715 MT
B715MT_AGGACGGTGCGGTGAG
B715MT_TGAAGGAGAAGGTGTCT
Chromatograms from B715 MT
B715MT_AGGACGGTGCGGTGAGAGTG Fragment base #26. Base 26 of 173
B715_2 MT
B715_2MT_AGGACGGTGCGGTGA
B715_2MT_TGAAGGAGAAGGTGT
Chromatograms from B715_2 MT
B715_2MT_AGGACGGTGCGGTGAGAGTG Fragment base #26. Base 26 of 174
B716 MT
B716MT_AGGACGGTGCGGTGAG
B716MT_TGAAGGAGAAGGTGTCT
Chromatograms from B716 MT
B716MT_AGGACGGTGCGGTGAGAGTG Fragment base #26. Base 26 of 176


B717 MT
Overview
Summary
Cut Map
Find
Show Chromatograms
Help Insert
Help Reposition
Chromatograms from B717 MT
B717_2 MT
Chromatograms from B717_2 MT
B718_2 MT
Chromatograms from B718_2 MT
B720_2 MT
Chromatograms from B720_2 MT
B719_2 MT
Chromatograms from B719_2 MT
B720_3 MT
Chromatograms from B720_3 MT


B721_2 MT
Overview
Summary
Cut Map
Find
Show Chromatograms
Help Insert
Help Reposition
Chromatograms from B721_2 MT
B722_2 MT
Chromatograms from B722_2 MT
B723_2 MT
Chromatograms from B723_2 MT
B724_2 MT
Chromatograms from B724_2 MT
B725_2 MT
Chromatograms from B725_2 MT


B726_2 MT
Overview
Summary
Cut Map
Find
Show Chromatograms
Help Insert
Help Reposition
Chromatograms from B726_2 MT
B727_2 MT
Chromatograms from B727_2 MT
B728_2 MT
Chromatograms from B728_2 MT
B729 MT
Chromatograms from B729 MT
B730 MT
Chromatograms from B730 MT
B730_2 MT
Chromatograms from B730_2 MT


B731 MT
Overview
Summary
Cut Map
Find
Show Chromatograms
Help Insert
Help Reposition
Chromatograms from B731 MT
B731_2 MT
Chromatograms from B731_2 MT
B732_2 MT
Chromatograms from B732_2 MT
B733_2 MT
Chromatograms from B733_2 MT
B734_2 MT
Chromatograms from B734_2 MT
B736D MT
Chromatograms from B736D MT


B840 MT
Overview
Summary
Cut Map
Find
Show Chromatograms
Help Inset
Help Reposition
Chromatograms from B840 MT
B842 MT
Chromatograms from B842 MT
B843b MT
Chromatograms from B843b MT
B844 MT
Chromatograms from B844 MT
B845 MT
Chromatograms from B845 MT
B846 MT
Chromatograms from B846 MT

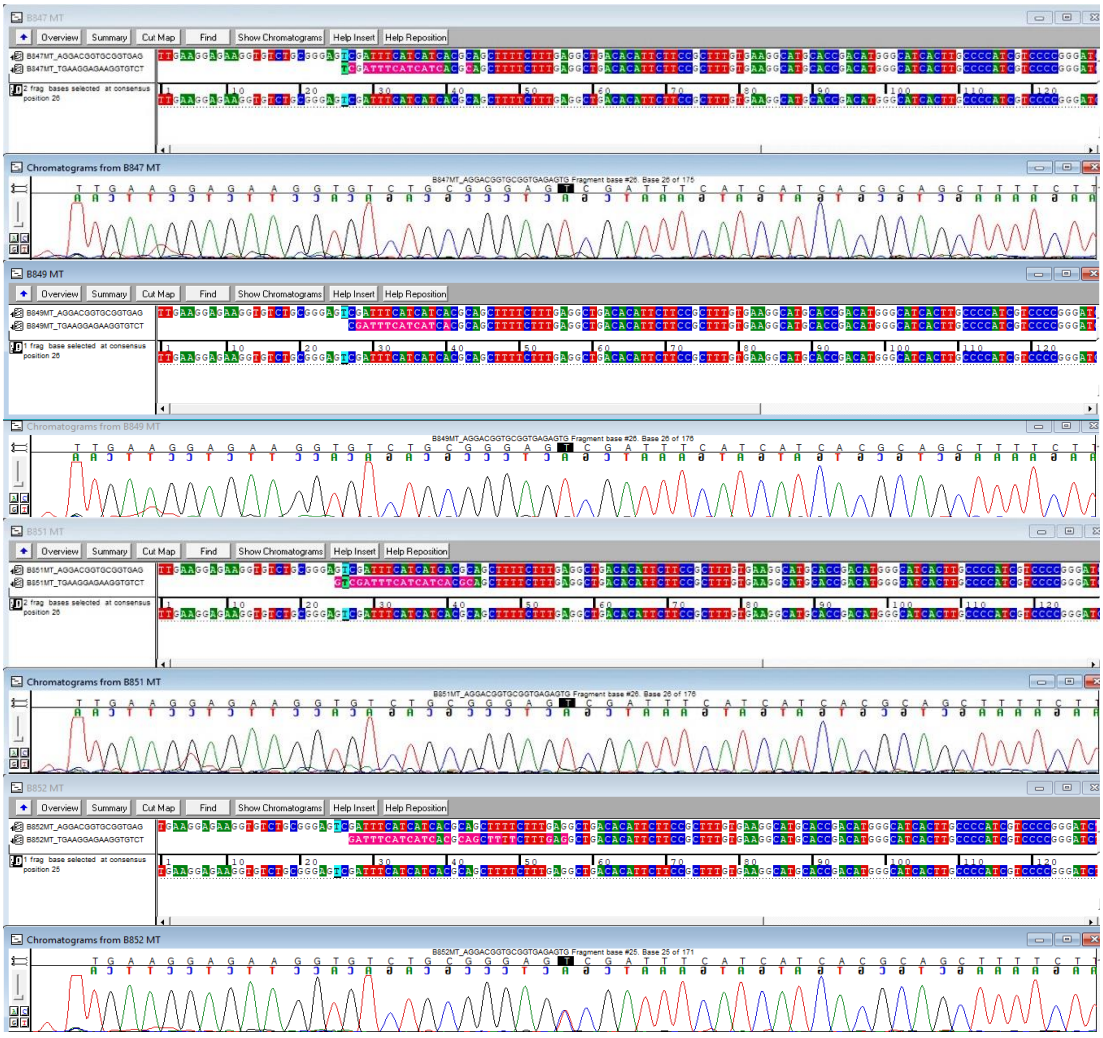


Figura 40. Posición del SNP en Sequencher en el resto de las muestras.

- **ANEXO 4: Corte de la enzima de restricción HinfI en NEBcutter dentro las muestras**

HinfI's cuts

B2

Save as text file

#	Cut position (blunt - 5' ext. - 3' ext.)	5'... Site with flanks ...3'
1	*23/26	13 GTGTCTGCGG G▾AGT▴C GATTTCATCA

HinfI's cuts

B6 2 T

Save as text file

#	Cut position (blunt - 5' ext. - 3' ext.)	5'... Site with flanks ...3'
1	*23/26	13 GTGTCTGCGG G▾AGT▴C GATTTCATCA

HinfI's cuts

B10 2MT T

Save as text file

#	Cut position (blunt - 5' ext. - 3' ext.)	5'... Site with flanks ...3'
1	*23/26	13 GTGTCTGCGG G▾AGT▴C GATTTCATCA

HinfI's cuts

B13 2

Save as text file

#	Cut position (blunt - 5' ext. - 3' ext.)	5'... Site with flanks ...3'
1	*23/26	13 GTGTCTGCGG G^AGT_C GATTTCATCA

HinfI's cuts

B14 2 T

Save as text file

#	Cut position (blunt - 5' ext. - 3' ext.)	5'... Site with flanks ...3'
1	*23/26	13 GTGTCTGCGG G^AGT_C GATTTCATCA

HinfI's cuts

B14 3 T

Save as text file

#	Cut position (blunt - 5' ext. - 3' ext.)	5'... Site with flanks ...3'
1	*23/26	13 GTGTCTGCGG G^AGT_C GATTTCATCA

HinfI's cuts

B702 2

Save as text file

#	Cut position (blunt - 5' ext. - 3' ext.)	5'... Site with flanks ...3'
1	*23/26	13 GTGTCTGCGG G^AGT_C GATTTCATCA

HinfI's cuts

B703 2 T

Save as text file

#	Cut position (blunt - 5' ext. - 3' ext.)	5'... Site with flanks ...3'
1	*23/26	13 GTGTCTGCGG G^AGT_C GATTTCATCA

HinfI's cuts

B705 2

Save as text file

#	Cut position (blunt - 5' ext. - 3' ext.)	5'... Site with flanks ...3'
1	*23/26	13 GTGTCTGCGG G▾AGT▴C GATTTCATCA

HinfI's cuts

B705 3

Save as text file

#	Cut position (blunt - 5' ext. - 3' ext.)	5'... Site with flanks ...3'
1	*23/26	13 GTGTCTGCGG G▾AGT▴C GATTTCATCA

HinfI's cuts

B706 3 T

Save as text file

#	Cut position (blunt - 5' ext. - 3' ext.)	5'... Site with flanks ...3'
1	*23/26	13 GTGTCTGCGG G▾AGT▴C GATTTCATCA

HinfI's cuts

B709 2 T

Save as text file

#	Cut position (blunt - 5' ext. - 3' ext.)	5'... Site with flanks ...3'
1	*23/26	13 GTGTCTGCGG G▾AGT▴C GATTTCATCA

HinfI's cuts

B711 2

Save as text file

#	Cut position (blunt - 5' ext. - 3' ext.)	5'... Site with flanks ...3'
1	*23/26	13 GTGTCTGCGG G▾AGT▴C GATTTCATCA

HinfI's cuts

B712 2

Save as text file

#	Cut position (blunt - 5' ext. - 3' ext.)	5'... Site with flanks ...3'
1	*23/26	13 GTGTCTGCGG G▾AGT▴C GATTTCATCA

HinfI's cuts

B713

Save as text file

#	Cut position (blunt - 5' ext. - 3' ext.)	5'... Site with flanks ...3'
1	*23/26	13 GTGTCTGCGG G▾AGT▴C GATTTCATCA

HinfI's cuts

B713 2

Save as text file

#	Cut position (blunt - 5' ext. - 3' ext.)	5'... Site with flanks ...3'
1	*23/26	13 GTGTCTGCGG G▾AGT▴C GATTTCATCA

HinfI's cuts

B714

Save as text file

#	Cut position (blunt - 5' ext. - 3' ext.)	5'... Site with flanks ...3'
1	*23/26	13 GTGTCTGCGG G▾AGT▴C GATTTCATCA

HinfI's cuts

B715

Save as text file

#	Cut position (blunt - 5' ext. - 3' ext.)	5'... Site with flanks ...3'
1	*23/26	13 GTGTCTGCGG G▾AGT▴C GATTTCATCA

HinfI's cuts

B715 2

Save as text file

#	Cut position (blunt - 5' ext. - 3' ext.)	5'... Site with flanks ...3'
1	*23/26	13 GTGTCTGCGG G▾AGT▴C GATTTCATCA

HinfI's cuts

B716 T

Save as text file

#	Cut position (blunt - 5' ext. - 3' ext.)	5'... Site with flanks ...3'
1	*23/26	13 GTGTCTGCGG G▾AGT▴C GATTTCATCA

HinfI's cuts

B717 T

Save as text file

#	Cut position (blunt - 5' ext. - 3' ext.)	5'... Site with flanks ...3'
1	*23/26	13 GTGTCTGCGG G▾AGT▴C GATTTCATCA

HinfI's cuts

B717 2 T

Save as text file

#	Cut position (blunt - 5' ext. - 3' ext.)	5'... Site with flanks ...3'
1	*23/26	13 GTGTCTGCGG G▾AGT▴C GATTTCATCA

HinfI's cuts

B718 2

Save as text file

#	Cut position (blunt - 5' ext. - 3' ext.)	5'... Site with flanks ...3'
1	*23/26	13 GTGTCTGCGG G▾AGT▴C GATTTCATCA

HinfI's cuts

B719 2 T

Save as text file

#	Cut position (blunt - 5' ext. - 3' ext.)	5'... Site with flanks ...3'
1	*23/26	13 GTGTCTGCGG G^AGT_C GATTTCATCA

HinfI's cuts

B720 2

Save as text file

#	Cut position (blunt - 5' ext. - 3' ext.)	5'... Site with flanks ...3'
1	*23/26	13 GTGTCTGCGG G^AGT_C GATTTCATCA

HinfI's cuts

B720 3

Save as text file

#	Cut position (blunt - 5' ext. - 3' ext.)	5'... Site with flanks ...3'
1	*23/26	13 GTGTCTGCGG G^AGT_C GATTTCATCA

HinfI's cuts

B721 2

Save as text file

#	Cut position (blunt - 5' ext. - 3' ext.)	5'... Site with flanks ...3'
1	*23/26	13 GTGTCTGCGG G^AGT_C GATTTCATCA

HinfI's cuts

B722 2

Save as text file

#	Cut position (blunt - 5' ext. - 3' ext.)	5'... Site with flanks ...3'
1	*23/26	13 GTGTCTGCGG G^AGT_C GATTTCATCA

HinfI's cuts

B723 2

Save as text file

#	Cut position (blunt - 5' ext. - 3' ext.)	5'... Site with flanks ...3'
1	*23/26	13 GTGTCTGCGG G▾AGT▴C GATTTCATCA

HinfI's cuts

B724 2

Save as text file

#	Cut position (blunt - 5' ext. - 3' ext.)	5'... Site with flanks ...3'
1	*23/26	13 GTGTCTGCGG G▾AGT▴C GATTTCATCA

HinfI's cuts

B725 2

Save as text file

#	Cut position (blunt - 5' ext. - 3' ext.)	5'... Site with flanks ...3'
1	*23/26	13 GTGTCTGCGG G▾AGT▴C GATTTCATCA

HinfI's cuts

B726 2

Save as text file

#	Cut position (blunt - 5' ext. - 3' ext.)	5'... Site with flanks ...3'
1	*23/26	13 GTGTCTGCGG G▾AGT▴C GATTTCATCA

HinfI's cuts

B727 2 T

Save as text file

#	Cut position (blunt - 5' ext. - 3' ext.)	5'... Site with flanks ...3'
1	*23/26	13 GTGTCTGCGG G▾AGT▴C GATTTCATCA

HinfI's cuts

B728 2

Save as text file

#	Cut position (blunt - 5' ext. - 3' ext.)	5'... Site with flanks ...3'
1	*23/26	13 GTGTCTGCGG G^AGT_C GATTTCATCA

HinfI's cuts

B729

Save as text file

#	Cut position (blunt - 5' ext. - 3' ext.)	5'... Site with flanks ...3'
1	*23/26	13 GTGTCTGCGG G^AGT_C GATTTCATCA

HinfI's cuts

B730

Save as text file

#	Cut position (blunt - 5' ext. - 3' ext.)	5'... Site with flanks ...3'
1	*23/26	13 GTGTCTGCGG G^AGT_C GATTTCATCA

HinfI's cuts

B730 2

Save as text file

#	Cut position (blunt - 5' ext. - 3' ext.)	5'... Site with flanks ...3'
1	*23/26	13 GTGTCTGCGG G^AGT_C GATTTCATCA

HinfI's cuts

B731

Save as text file

#	Cut position (blunt - 5' ext. - 3' ext.)	5'... Site with flanks ...3'
1	*23/26	13 GTGTCTGCGG G^AGT_C GATTTCATCA

HinfI's cuts

B731 2

Save as text file

#	Cut position (blunt - 5' ext. - 3' ext.)	5'... Site with flanks ...3'
1	*23/26	13 GTGTCTGCGG G▾AGT▴C GATTTCATCA

HinfI's cuts

B732 2

Save as text file

#	Cut position (blunt - 5' ext. - 3' ext.)	5'... Site with flanks ...3'
1	*23/26	13 GTGTCTGCGG G▾AGT▴C GATTTCATCA

HinfI's cuts

B733 2

Save as text file

#	Cut position (blunt - 5' ext. - 3' ext.)	5'... Site with flanks ...3'
1	*23/26	13 GTGTCTGCGG G▾AGT▴C GATTTCATCA

HinfI's cuts

B734 2 T

Save as text file

#	Cut position (blunt - 5' ext. - 3' ext.)	5'... Site with flanks ...3'
1	*23/26	13 GTGTCTGCGG G▾AGT▴C GATTTCATCA

HinfI's cuts

B736D T

Save as text file

#	Cut position (blunt - 5' ext. - 3' ext.)	5'... Site with flanks ...3'
1	*23/26	13 GTGTCTGCGG G▾AGT▴C GATTTCATCA

HinfI's cuts

B840

Save as text file

#	Cut position (blunt - 5' ext. - 3' ext.)	5'... Site with flanks ...3'
1	*23/26	13 GTGTCTGCGG G^AGT_C GATTTCATCA

HinfI's cuts

B842

Save as text file

#	Cut position (blunt - 5' ext. - 3' ext.)	5'... Site with flanks ...3'
1	*23/26	13 GTGTCTGCGG G^AGT_C GATTTCATCA

HinfI's cuts

B843b

Save as text file

#	Cut position (blunt - 5' ext. - 3' ext.)	5'... Site with flanks ...3'
1	*23/26	13 GTGTCTGCGG G^AGT_C GATTTCATCA

HinfI's cuts

B844

Save as text file

#	Cut position (blunt - 5' ext. - 3' ext.)	5'... Site with flanks ...3'
1	*23/26	13 GTGTCTGCGG G^AGT_C GATTTCATCA

HinfI's cuts

B845 T

Save as text file

#	Cut position (blunt - 5' ext. - 3' ext.)	5'... Site with flanks ...3'
1	*23/26	13 GTGTCTGCGG G^AGT_C GATTTCATCA

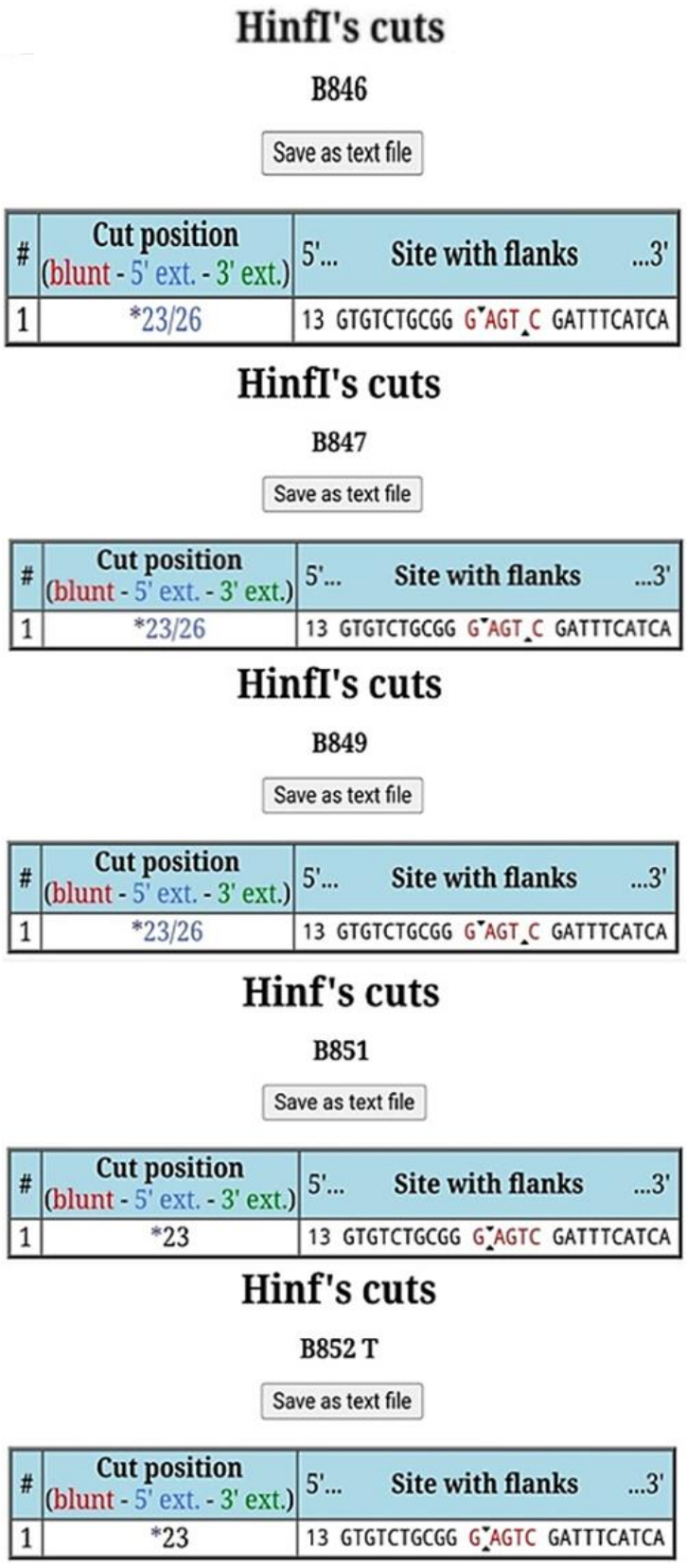


Figura 41. Sitio de corte de restricción de HinFI en el resto de las muestras.

- **ANEXO 5: sitios *webs* y *softwares* usados en el presente trabajo**

Sitios web utilizados en el presente trabajo

CADD: http://cadd.gs.washington.edu
ClinVar: http://www.ncbi.nlm.nih.gov/clinvar
COSMIC: https://cancer.sanger.ac.uk/cosmic
dbSNP: Https://www.ncbi.nlm.nih.gov/snp/
Ensembl: https://www.ensembl.org/index.html
ExPasy: https://web.expasy.org/
Fathmm: http://fathmm.biocompute.org.uk
Gencode: https://www.gencodegenes.org/
GWAS Catalog: https://www.ebi.ac.uk/gwas/
HGMD: https://www.hgmd.cf.ac.uk/ac/index.php
Illumina: https://www.illumina.com/
Mutationtaster: http://www.mutationtaster.org
NCBI: https://www.ncbi.nlm.nih.gov/
NEBCutter: https://nc2.neb.com/NEBcutter2/?noredir
Novogene: https://www.novogene.com
OMIM: http://www.omim.org
RefSeq: https://www.ncbi.nlm.nih.gov/refseq/
Uniprot: https://www.uniprot.org/
Varsome: Https://varsome.com/

Softwares utilizados en el presente trabajo
MEGA
PopGene
Sequencher