



UNIVERSIDAD
NACIONAL
DE COLOMBIA

Un modelo de cálculo para Índices (IP) y probabilidad de paternidad (W) por intervalos de confianza.

María Fernanda Mogollón Olivares

Universidad Nacional de Colombia

Instituto de Genética

Bogotá, Colombia

2023

Un modelo de cálculo para Índices (IP) y probabilidad de paternidad (W) por intervalos de confianza.

María Fernanda Mogollón Olivares

Trabajo de tesis presentado como requisito parcial para optar al título de:

Magíster en Genética Humana

Director:

William Usaquén Martínez, MSc., PhD.
UNIVERSIDAD NACIONAL DE COLOMBIA

Grupo de Investigación:

Genética de Poblaciones e Identificación (GPI)

Instituto de Genética.

Universidad Nacional de Colombia.

UNIVERSIDAD NACIONAL DE COLOMBIA
SEDE BOGOTÁ
FACULTAD DE MEDICINA
MAESTRÍA EN GENÉTICA HUMANA
BOGOTÁ, 2023

*A todas las mujeres en STEM que se atrevieron a romper patrones heteropatriarcales en la Ciencia y la Academia e hicieron que niñas como yo creyéramos con el anhelo de investigar y preguntar sin temor ni miedos.
Sin embargo, el miedo aún nos persigue.*

A todas las mujeres que se les exige ser tan excelente madre como si no trabajaran pero que trabajen como si no fueran madres.

En Colombia, el 37% de las personas que ocupan cargos de investigación son mujeres.

A nivel mundial, sólo el 30% de las matrículas de áreas de ciencias biológicas, matemáticas y estadística pertenecen a mujeres.

En 120 años de historia – sin contar el actual -, el premio Nobel ha sido otorgado 934 personas. De ellas, casi el 93.79% han sido hombres (876) y tan sólo el 6,42% (60) han sido mujeres.

*Cuando dicen que tenemos más protagonismo, más facilidades, más oportunidades y mayor visibilidad,
¿A qué se refieren?*

*To all the women in STEM who dared to break heteropatriarchal patterns in Science and Academia
and made girls like me grow up with the desire to investigate and ask questions without fear or
doubts.
However, fear still haunts us.*

*To all the women who are expected to be as excellent mothers as if they didn't work
but work as if they weren't mothers.*

In Colombia, 37% of people holding research positions are women.

*On a global level, only 30% of enrollments in biological sciences, mathematics, and statistics belong
to women.*

*In 120 years of history - not counting the current one - the Nobel Prize has been awarded to 934
individuals. Of them, almost 93.79% have been men (876), and only 6,42% (60) have been women.*

*When they say we have more visibility, more opportunities, and more prominence,
What do they mean?*

Agradecimientos

En este viaje de exploración y conocimiento, no puedo dejar de expresar mi gratitud a un compañero inseparable de la vida en la Tierra: el sexo y su fiel aliada, la reproducción sexual. La incesante danza de genes y cromosomas ha permitido que hoy podamos estudiar la variabilidad genética y los efectos de esta sobre las dinámicas poblacionales.

A las personas que han confiado en la Universidad Nacional de Colombia, como laboratorio acreditado y certificado para realizar sus pruebas de paternidad y dieron su consentimiento para investigaciones en Genética de Poblaciones.

Al grupo de Genética de Poblaciones e Identificación por permitirme trabajar con ellos en estos últimos cinco años. Infinitas gracias al Profesor William Usaquén por todas sus enseñanzas, su guía fue fundamental para llevar a cabo este trabajo. A Andrea, Dani, Ana quienes me acogieron como una familia. A Diana y Freddy por orientarme en todos los procesos administrativos y técnicos del laboratorio. A Julie y Dayana, por su valiosa amistad y enseñanzas académicas.

A mi familia, a quienes les agradezco y les pido excusa por ser la hija ausente. Doce años lejos de ustedes, no ha sido fácil. Por todos los desayunos y tardes de domingo que he perdido a su lado. A mi Mami y Papi que me apoyaron en el 2012 con la idea de irme a estudiar a una ciudad que ni ellos ni yo conocíamos. A Candelaria Naranjo que siempre está en mi corazón. A mis hermanos: Anghie, Fernando y sobrinos: Julián, Maryfer, Valentina, Samara. A quienes se han vuelto mi nueva familia y apoyo: Alejo y Negrito.

A las amistades que dejó la Maestría en Genética Humana y a los que ya estaban antes de este camino: Bran, Clau, Juan o llegaron en el proceso. Todos han sido mi red de apoyo en una ciudad que a veces siento mía pero que muchas más veces siento ajena.

And last but not least, me.

Título:

Un modelo de cálculo para Índices (IP) y probabilidad de paternidad (W) por intervalos de confianza.

Resumen

Durante las últimas tres décadas en Colombia se han llevado a cabo investigaciones que han permitido exponer la variabilidad de sus poblaciones desde una perspectiva genética; también se han realizado diversos reportes de frecuencias alélicas y estimadores forenses para poblaciones humanas específicas que ayudan a caracterizar la población colombiana teniendo en cuenta su complejidad y los diferentes procesos de mezcla. Dado a que el campo de la genética forense genera una cantidad basta de datos poblacionales de polimorfismos tipo STR en muestras distribuidas globalmente, se ha estudiado el poder de estos sets de datos para responder preguntas relacionadas a la evolución humana y su diversidad, teniendo en cuenta dos tipos de recursos: las frecuencias alélicas disponibles en las bases de datos y datos genotípicos que se pueden encontrar en pocas bases de datos o en artículos científicos. Esta información es publicada como reporte de frecuencias en artículos científicos y trabajos de tesis, sin embargo, no existe constancia en la publicación como tampoco un registro de base de datos estandarizado que sean de acceso público. Las frecuencias y estimadores son empleados tanto en estudios de genética poblaciones para corroborar hipótesis de estructura genética y de poblamiento, como también para formular parámetros en cálculos de índices de paternidad útiles en pruebas de paternidad y filiación al establecer el parentesco de individuos como en el caso del derecho de identidad según lo contemplado en el Artículo 25 de la Ley 1098/06. Sin embargo, estos reportes de probabilidades de paternidad son expresados como estimadores puntuales a pesar de que, en la práctica, las frecuencias con las que se obtienen pertenecen a una muestra de población y no a la población total. Teniendo en cuenta el número de casos de pruebas de filiación estandarizados y registrados en la base de datos del Grupo de Genética de Poblaciones de la Universidad Nacional de Colombia, se propone realizar un modelo de cálculo que matemáticamente exprese el resultado de una prueba en función de un intervalo de confianza, a partir de técnicas de remuestreos aleatorizados por métodos de Monte Carlo. Realizar un estudio de carácter teórico con fundamentos matemáticos y estadísticos permitirá establecer líneas de análisis robustas para la expresión de resultado en las pruebas y la interpretación de estos.

Palabras clave: Identificación, pruebas de paternidad, STRs autosómicos, índice de paternidad, bases de datos, probabilidad de exclusión, filiación genética, remuestreo, aleatorización, Monte Carlo.

Title:

A calculation model for Indices (IP) and probability of paternity (W) by confidence intervals.

Abstract:

During the last three decades in Colombia, research has been carried out that has allowed exposing the variability of its populations from a genetic perspective; Various reports of allelic frequencies and forensic estimators have also been made for specific human populations that help characterize the Colombian population, considering its complexity and the different admixture processes. Given that the field of forensic genetics generates a vast amount of population data on STR-type polymorphisms in globally distributed samples, the power of these data sets to answer questions related to human evolution and its diversity has been studied, considering two types of resources: allelic frequencies available in databases and genotypic data that can be found in few databases or in scientific articles. This information is published as a frequency report in scientific articles and thesis works, however, there is no record in the publication nor a standardized database record that is publicly accessible. Frequencies and estimators are used both in population genetics studies to corroborate hypotheses of genetic structure and population, as well as to formulate parameters in calculations of useful paternity indices in paternity and filiation tests when establishing the kinship of individuals, as in the case of the right of identity as contemplated in Article 25 of Law 1098/06. However, these reports of paternity probabilities are expressed as punctual estimators even though, in practice, the frequencies with which they are obtained belong to a population sample and not to the total population. Considering the number of cases of standardized filiation tests registered in the database of the Population Genetics Group of the National University of Colombia, it is proposed to make a calculation model that mathematically expresses the result of a test based on a confidence interval, based on randomized resampling techniques using Monte Carlo methods. Carrying out a theoretical study with mathematical and statistical foundations will allow establishing robust lines of analysis for the expression of results in the tests and their interpretation.

Keywords: Human identification, paternity tests, autosomal STRs, paternity index, databases, exclusion probability, genetic affiliation, resampling, randomization, Monte Carlo.

Tabla de contenido

Capítulo 1. Potencialidad de las bases de datos: Aplicaciones y nuevas metodologías análisis en genética de poblaciones.	14
1.1. Introducción.	14
1.1.1. Bases de datos en genética forense e identificación	16
1.1.2. Formulación y aplicación de modelos teóricos y métodos de análisis estadístico en genética a partir del uso de información reposada en bases de datos.....	18
1.1.3. Aleatorización y métodos de Monte Carlo.	20
1.1.3.1. Elementos de una simulación de Monte Carlo y consideraciones generales de la prueba: 22	
1.2. Planteamiento del problema	24
1.3. Objetivos.....	27
1.3.1. Objetivo general	27
1.3.2. Objetivos específicos	27
1.4. Referencias	28
Capítulo 2: Modelo de cálculo para Índices (IP) y probabilidad de paternidad (W) por intervalos de confianza.	39
2.1. Introducción.....	39
2.2. Materiales y métodos.	42
2.2.1. Poblaciones de referencia.	42
Región1: R1 San Andrés.....	43
Región 2: R2 Wayúu.....	43
Región 3: R3 Amazonas.	44
Región 4: R4 Bogotá.....	44
Región 5: R5 Caribe.....	45
2.2.2. Casos analizados.	46
2.2.3. Implementación de los métodos de aleatorización de Monte Carlo para la selección de Mn muestras, cálculos de frecuencias alélicas, IP y W por muestreo.	46
2.2.3.1. Selección de Mn muestras por región.	46
2.2.3. Representación de resultados.	50
2.3. Resultados.	50

2.3.1. Análisis de diversidad y genética descriptiva.	50
2.3.2. Índices de Paternidad por región, submuestra y replica.....	52
2.3.2.1. Región 1: San Andrés y Providencia.....	52
2.3.2.2. Región 2: Wayuu.....	54
2.3.2.3. Región 3: Amazonas.....	56
2.3.2.4. Región 4: Bogotá.....	58
2.3.2.4. Región 5: Caribe.....	60
2.3.3. Probabilidades de Paternidad por región, submuestra y replica.	61
2.3.3.1. Región 1: San Andrés.	61
2.3.3.2. Región 2: Wayuu.....	61
2.3.3.3. Región 3: Amazonas.....	64
2.3.3.4. Región 4: Bogotá.....	65
2.3.3.5. Región 5: Caribe.....	66
2.3.4. Casos atípicos o raros.	67
2.3.4.1. Según el índice de paternidad (IP).	67
2.3.4.2. Según la probabilidad de paternidad (W).	69
2.3.3. Un modelo de cálculo para Índices (IP) y probabilidad de paternidad (W) por intervalos de confianza.....	69
2.4. Discusión.....	71
2.5. Conclusiones.....	73
2.6. Referencias.	74

Capítulo 3: Efecto de la ausencia de la madre en pruebas de paternidad y el número de falsos presuntos padres no excluidos a partir de 15 marcadores STRs utilizando una base de datos genética de Bogotá, Colombia. 80

3.1 Introducción.	82
3.2 Materiales y métodos.....	83
3.2.1. Casos analizados.	83
3.2.2. Extracción de ADN, tipificación de STRs y análisis de fragmentos.	84
3.2.3. Análisis estadístico y comparación de perfiles genéticos de presuntos padres e hijos no emparentados.	84
3.3 Resultados.	86

3.3.1	Comparaciones de hijos y presuntos padres no relacionados teniendo en cuenta la presencia de la madre biológica.	88
3.3.1	Comparaciones de hijos y presuntos padres no relacionados sin la presencia de la madre biológica.	88
3.4	Discusiones.	90
3.5	Conclusiones.	91
3.6	Referencias.	92
Capítulo 4: Genética poblacional y forense: Herramientas para la reconstrucción histórica y social en la era del posconflicto colombiano.		96
4.1	Poder de los muestreos dirigidos y a conveniencia en genética de poblaciones y su aplicación en forense.	99
4.2	¿Por qué unir fuerzas e intenciones de investigación?	99
4.3	Limitaciones y conclusiones.	101
4.4	Referencias.	102

Índice de figuras

Figura 1 . Generalidad Métodos de Monte Carlo: A través de los métodos de Monte Carlo, puedo llegar a inferir un parámetro de una población a partir de un estadístico proveniente de una muestra de la población de interés.	21
Figura 2 <i>Las frecuencias alélicas de la población de referencia corresponden a un estadístico y no a un parámetro.</i>	25
Figura 3 Distribución de frecuencias alélicas del alelo a13 al a19 del marcador D3S1358 en la población de Bogotá. Tomado de Usaquén Martínez, 2012.....	26
Figura 4 Variables asociadas al tipo de muestreo y selección de muestras en estudios genético poblaciones.....	40
Figura 5 Distribución geográfica de las poblaciones de referencia empleadas en las pruebas de paternidad. Región R1 San Andrés (SAN) se representa en rojo, Región R2 Wayuu (WAY) en gris, Región R3 Amazonas (UIT: Uitoto, COC: Cocama, TIC: Ticuna) en azul claro, Región R4 Bogotá (BOG) en púrpura y Región R5 Caribe en azul oscuro (ZEN: Zenú, MOK: Mokana, ARH: Arhuaco, KAN: Kankuamo). Tomado de Mogollon Olivares et al., 2020).....	42
Figura 6 Diagrama del experimento realizado por cada una de las regiones desde R1 a R5. A. Generación de tres submuestras (S1, S2, S3) de la región R con diferentes tamaños muestrales n, n/2, n/4: A cada submuestra se le realizan cuatro replicas. B. Cálculo de distribución de frecuencias alélicas para cada replica de cada submuestra: Generación de tablas de distribución de frecuencia alélica incluyendo las frecuencias alélicas mínimas para cada replica de cada submuestra. C. Resolución del caso trío con cada una de las 12 poblaciones de referencia generadas en cada región: Un caso en particular es resultado con las frecuencias alélicas de las 12 poblaciones de referencia generadas. D. Cálculo IP por gen por subpoblación por cada replica: Cálculo de índices de paternidad para cada uno de los genes evaluados en el caso trío. E. Cálculo IP combinado por gen por subpoblación por cada réplica (IPC R1S1r1 a IPC R1S3r4) y generación del IPC por submuestra: Desde R1S1r1 a R1S2r4 se calculan los IPC. F. Cálculo W combinado por gen por subpoblación por cada réplica (WC R1S1r1 a WC R1S3r4) y generación del IWC por submuestra.	48
Figura 7 Diagrama del experimento realizado en los 64.824 casos trío de paternidad con diferentes 150 padres. A. Cálculo de distribución de frecuencias alélicas para la población de referencia: Bogotá fue la población de referencia empleada. B. Resolución del caso trío con madre (M) e hijo (H) constantes. Se tomaron 296 dúos madre e hijo sin exclusión materna que fueron analizados con 219 presuntos padres (PP), desde el padre 1 hasta el n; analizándose desde el caso 1 al n formado por el dúo madre e hijo n con el presunto padre n. C. Cálculo del índice y probabilidad de paternidad por caso con resultado de no exclusión de paternidad: La probabilidad de paternidad fue calculada como una probabilidad a posteriori Bayesiana. D. Conteo de casos trío con no exclusión de paternidad con el mismo presunto padre: Se realizó una consulta que permitiera determinar si se encontraba el mismo presunto padre en varios dúos (madre e hijo) en donde el resultado de ese trío no correspondiera a una exclusión de la paternidad.....	87

Índice de gráficas

- Gráfica 1** Medidas de diversidad genética para cada uno de los muestreos simulados realizados en la Región del Amazonas desde AMAS1r1 a AMAS1r4 ($n = n/4$), AMAS2r1 a AMAS2r4 ($n = n/2$) y AMAS3r1 a AMAS3r4 (n). Tamaño muestral, número de alelos reportados, número de alelos comunes, número de alelos raros, número de alelos efectivos y frecuencia alélica mínima. 51
- Gráfica 2** Frecuencias alélicas mínimas para cada uno de los muestreos simulados realizados en la Región del Amazonas desde AMAS1r1 a AMAS1r4 ($n = n/4$), AMAS2r1 a AMAS2r4 ($n = n/2$) y AMAS3r1 a AMAS3r4 (n)..... 51
- Gráfica 3** Gráfico de dispersión del PCA construido a partir de los índices de paternidad con los sistemas D16S539, D18S51, D19S433, D21S11, D2S1338, D3S1358, D8S1179, FGA, TH01, VWA calculados para 1797 casos tríos analizados en este estudio. Población de referencia: Región R1 San Andrés con diferentes tamaños muestrales: A. $n4.$, B. $n4.$, y C. n 53
- Gráfica 4** Gráfico de dispersión del PCA construido a partir de los índices de paternidad con los sistemas D16S539, D18S51, D19S433, D21S11, D2S1338, D3S1358, D8S1179, FGA, TH01, VWA calculados para 1797 casos tríos analizados en este estudio. Población de referencia: Región R2 Wayuu con diferentes tamaños muestrales: A. A. $n4.$, B. $n4.$, y C. n 55
- Gráfica 5** Gráfico de dispersión del PCA construido a partir de los índices de paternidad con los sistemas D16S539, D18S51, D19S433, D21S11, D2S1338, D3S1358, D8S1179, FGA, TH01, VWA calculados para 1797 casos tríos analizados en este estudio. Población de referencia: Región R3 Caribe con diferentes tamaños muestrales: A. $n4.$, B. $n2.$, y C. n 57
- Gráfica 6** Gráfico de dispersión del PCA construido a partir de los índices de paternidad con los sistemas D16S539, D18S51, D19S433, D21S11, D2S1338, D3S1358, D8S1179, FGA, TH01, VWA calculados para 1797 casos tríos analizados en este estudio. Población de referencia: Región R4 Bogotá con diferentes tamaños muestrales: A. $n4.$, B. $n2.$, y C. n 59
- Gráfica 7** Gráfico de dispersión del PCA construido a partir de los índices de paternidad con los sistemas D16S539, D18S51, D19S433, D21S11, D2S1338, D3S1358, D8S1179, FGA, TH01, VWA calculados para 1797 casos tríos analizados en este estudio. Población de referencia: Región R5 Caribe con diferentes tamaños muestrales: A. $n4.$, B. $n2.$, y C. n 60
- Gráfica 8** Gráfico de dispersión del PCA construido a partir de las probabilidades de paternidad con los sistemas D16S539, D18S51, D19S433, D21S11, D2S1338, D3S1358, D8S1179, FGA, TH01, VWA calculados para 1797 casos tríos analizados en este estudio. Población de referencia: Región R1 San Andrés con diferentes tamaños muestrales: A. $n4.$, B. $n2.$, y C. n 62

Gráfica 9 Gráfico de dispersión del PCA construido a partir de las probabilidades de paternidad con los sistemas D16S539, D18S51, D19S433, D21S11, D2S1338, D3S1358, D8S1179, FGA, TH01, VWA calculados para 1797 casos tríos analizados en este estudio. Población de referencia: Región R2 Wayuu con diferentes tamaños muestrales: A. $n4$., B. $n2$., y C. n 63

Gráfica 10 Gráfico de dispersión del PCA construido a partir de las probabilidades de paternidad con los sistemas D16S539, D18S51, D19S433, D21S11, D2S1338, D3S1358, D8S1179, FGA, TH01, VWA calculados para 1797 casos tríos analizados en este estudio. Población de referencia: Región R3 Amazonas con diferentes tamaños muestrales: A. $n4$., B. $n2$., y C. n 64

Gráfica 11 Gráfico de dispersión del PCA construido a partir de las probabilidades de paternidad con los sistemas D16S539, D18S51, D19S433, D21S11, D2S1338, D3S1358, D8S1179, FGA, TH01, VWA calculados para 1797 casos tríos analizados en este estudio. Población de referencia: Región R4 Bogotá con diferentes tamaños muestrales: A. $n4$., B. $n2$., y C. n 65

Gráfica 12 Gráfico de dispersión del PCA construido a partir de las probabilidades de paternidad con los sistemas D16S539, D18S51, D19S433, D21S11, D2S1338, D3S1358, D8S1179, FGA, TH01, VWA calculados para 1797 casos tríos analizados en este estudio. Población de referencia: Región R4 Bogotá con diferentes tamaños muestrales: A. $n4$., B. $n2$., y C. n 66

Gráfica 13 Índice de paternidad de los sistemas D16S539, D18S51, D19S433, D21S11, D2S1338, D3S1358, D8S1179, FGA, TH01, VWA calculados para un caso trío analizado en este estudio. Se emplearon 12 poblaciones de referencia que corresponden a muestreos de diferentes tamaños muestrales de $S1 = n4$, $S2 = n2$ y $S3 = n$ de las regiones R1: San Andrés, R2: Wayuu, R3: Amazonas, R4: Bogotá y R5: Caribe. 68

Gráfica 14 Número de "presuntos padres adicionales" (hombres con menos de tres exclusiones y al menos 12 no exclusiones en los 15 loci STR analizados) para los hijos con los que fueron analizados. 89

Índice de tablas

Tabla 1 Cálculo IP combinado por gen por población y generación del intervalo de confianza para un caso determinado..... 70

Tabla 2 Cálculo W combinado por gen por población y generación del intervalo de confianza para un caso determinado..... 71

Tabla 3 Número de "presuntos padres adicionales" (hombres con menos de tres exclusiones y al menos 12 coincidencias en los loci STR) para los hijos de un dúo (M-H)..... 88

Tabla 4 Distribución de probabilidades de paternidad en pares hijo/hombre emparentados y no emparentados con 3 o menos exclusiones..... 89

Tabla S1. Frecuencias alélicas para las 60 poblaciones de referencia desde R1S1r1 a R5S3r4 empleadas en el experimento para los marcadores D16S539, D18S51, D19S433, D21S11, D2S1338, D2S1358, D8S1179, FGA, TH01 y VWA. Las celdas resaltadas en azul corresponden a los alelos ausentes en dicha población cuya frecuencia reportada es la frecuencia alélica mínima según Budowle et al., 1996.

Tabla S2. Medidas de diversidad genética para los marcadores D16S539, D18S51, D19S433, D21S11, D2S1338, D2S1358, D8S1179, FGA, TH01 y VWA y cálculo de frecuencias alélicas mínimas por marcador para las 60 poblaciones de referencia desde R1S1r1 a R5S3r4 empleadas en el experimento.

Tabla S3. Índices de paternidad calculados para los marcadores D16S539, D18S51, D19S433, D21S11, D2S1338, D2S1358, D8S1179, FGA, TH01 y VWA a partir de las poblaciones de referencia de las regiones San Andrés, Wayuu, Amazonas, Bogotá y Caribe con tamaños muestrales de $n/4$, $n/2$ y n .

Capítulo 1. Potencialidad de las bases de datos: Aplicaciones y nuevas metodologías de análisis en genética de poblaciones.

1.1. *Introducción.*

La población colombiana actual es resultado de la unión de una gran variedad de componentes étnicos, históricos, culturales, geográficos y biogeográficos que la han convertido en un país con alta diversidad biológica y cultural (CINEP, 1998b, 1998c, 1998a). Al igual que la mayoría de los países latinoamericanos, en Colombia se encuentran representantes de poblaciones pertenecientes a grupos poblaciones afrodescendientes, comunidades indígenas, poblaciones humanas autoderminadas como blancos y de ancestralidad múltiple (DANEa, 2007; DANEb, 2007; Ministerio de Asuntos Exteriores y de Cooperación & Oficina de Información Diplomática del Departamento de Relaciones Exteriores, 2017).

Desde el punto de vista genético, la población de Colombia posee un amplio interés de estudio como consecuencia del flujo genético de grupos humanos provenientes de diferentes ancestralidades que poblaron su territorio (Alonso Morales et al., 2018; Homburger et al., 2015; H. Ossa et al., 2015; Rishishwar, Conley, Wigington, et al., 2015). La llegada de colonizadores europeos a tierras pobladas por comunidades indígenas y el subsecuente arribo de grupos procedentes de África y posteriores movimientos poblacionales en el siglo XX permitieron lograr el panorama de diversidad genética actual de la población colombiana. (Alonso Morales et al., 2018; Keyeux & Usaquén, 2006; Mincultura, 2011; Humberto Ossa et al., 2016; Urbano, Portilla, Builes, Gusmão, & Sierra-Torres, 2016).

Se han realizado investigaciones que no sólo han permitido exponer la gran riqueza de la población Colombiana desde una perspectiva genética sino que también han revelado diferentes patrones de composición, mezcla, genealogía e historia natural de las poblaciones valiéndose no sólo de distintos marcadores genéticos sino también de información demográfica, cultural, histórica y lingüística (Alonso & Usaquén, 2012; Alonso Morales et al., 2018; Bolnick, Bolnick, & Smith, 2006; Bolnick, Raff, Springs, Reynolds, & Miró-Herrans, 2016; Builes et al., 2013; Callegari-Jacques, Tarazona-Santos, Gilman, Herrera, Cabrera, dos Santos, et al., 2011; Casas-Vargas et al., 2017; Da Costa Francez, Rodrigues, De Velasco, & Dos Santos, 2012; Hunley & Healy, 2011; Kohlrausch et al., 2005; Mesa et

al., 2000; Moreno-Estrada et al., 2013; Humberto Ossa et al., 2016; M. Y. Rojas, Alonso Morales, Sarmiento, Eljach, & Usaquén Martínez, 2013; Wang et al., 2007).

Los análisis genéticos descriptivos y cuantitativos establecen las características de estructura genética de la población colombiana que puedan determinar si existe una heterogeneidad en la población evaluando la variabilidad de los polimorfismos y su distribución espacial que ayudan comprender las fuerzas del cambio genético en nuestro país y también sus transformaciones culturales (M. Y. Rojas et al., 2013; Usaquén Martínez, 2012).

En Colombia se han realizado estudios de la diversidad genética y reportes de frecuencias alélicas y estimadores forenses para poblaciones específicas humanas como las realizadas en la costa norte (Martinez et al., 2017; Martínez, Builes, & Caraballo, 2008; Martínez et al., 2006; Martínez, Caraballo, Gusmão, Amorim, & Carracedo, 2005; M. Y. Rojas et al., 2013), el noreste (Hincapié et al., 2009; Ossa Reyes, Torres Ramírez, & Nieto Romero, 2009; Vargas, Castillo, Gil, Pico, & García, 2003; Castillo et al., 2013), el occidente (Franco-Candela & Barreto, 2017; Gómez, Reyes, Cárdenas, & García, 2003; Julieta Avila, Briceño, & Gómez, 2009; Palacio et al., 2006; Porras et al., 2008; Rondón, César Osorio, Viviana Peña, Andrés Garcés, & Barreto, 2008; Rondón G., Oribio, Braga, Cárdenas, & Barreto, 2006), la región del pacífico (Bravo et al., 2001), la región central (Bravo et al., 2001; Burgos et al., 2015; Castillo et al., 2013; Gaviria et al., 2004; Ibarra et al., 2014; Rey et al., 2003) y la región amazónica (Braga, Arias B., & Barreto, 2012; Rivera Franco, Braga, Espitia Fajardo, & Barreto, 2020; Yin et al., 2018). Estos trabajos realizaron en cálculo de las frecuencias alélicas por regiones geográficas. Otras publicaciones han reportado las frecuencias alélicas con base en una aproximación a nivel nacional (Benítez-Páez & Reyes, 2003; Durán et al., 2003; Paredes et al., 2003; Sánchez-Diz et al., 2009; Juan J Yunis & Yunis, 2013; Juan José Yunis et al., 2005).

Estas frecuencias y los estimadores forenses reportados son empleados en el estudio de genética poblaciones para evaluar hipótesis de estructura genética y de poblamiento, pero también ayudan a determinar parámetros en cálculos de índices de paternidad útiles en pruebas de paternidad y filiación usadas para establecer el parentesco de individuos como el derecho de identidad establecido en el Artículo 25 de la Ley 1098/06.

1.1.1. Bases de datos en genética forense e identificación

La tecnología del ADN basada en el análisis de la variabilidad genética de regiones cortas repetidas en tándem (en inglés: short tandem repeat STR) se ha convertido en el método de elección y en el estándar internacional de análisis de ADN más utilizado en los miles de laboratorios de genética forense de todo el mundo (Alonso Alonso, 2019). La información genética en las regiones de STR carece de valor hasta que es cotejada con el perfil genético de una muestra de referencia o una de interés para lograr establecer la posible identidad o grado de parentesco genético entre las muestras comparadas (Tillmar, 2010; Wurmb-schwark et al., 2015).

Esta particularidad hace necesario realizar un análisis comparativo entre la muestra problema y la de referencia; lo que ha llevado como resultado el desarrollo y la creación de bases de datos nacionales que permiten realizar búsquedas sistemáticas de perfiles genéticos sirviendo como estrategia de análisis en la identificación de personas en investigaciones criminales como de personas desaparecidas (Alonso Alonso, 2019; Bieber, Brenner, & Lazer, 2006; Santos, Machado, & Silva, 2013).

En las bases de datos de ADN, principalmente de marcadores STR, reposan hoy día más de 100 millones de perfiles de ADN distribuidos en unos 60 países y son una herramienta clave en la investigación criminal nacional e internacional. Su desarrollo, así como su estandarización, son llevados a cabo por organismos científicos como la Red Europea de Institutos de Ciencias Forenses (ENFSI -DNA) en Europa y el Laboratorio del FBI con el desarrollo del sistema CODIS en los Estados Unidos (European Network of Forensic Science Institutes, 2020; Federal Bureau of Investigations, 2020a, 2020b). Estas grandes bases de datos forenses juegan un papel importante en procesos de identificación de desaparecidos, identificación de víctimas en conflictos bélicos o en grandes catástrofes que requieren la discriminación de un número grande de individuos en donde el estado de conservación o el poco material encontrado puede limitar o imposibilitar la identificación de los cuerpos por métodos forenses convencionales (Alonso Alonso, 2019; Bieber et al., 2006; Goodwin, Linacre, & Hadi, 2011; Houck, 2015; Santos et al., 2013).

A nivel continental europeo, el Ministerio del Interior de Reino Unido en 1995 puso en funcionamiento la primera base de datos nacional de ADN del mundo. En esta base de datos originalmente sólo se almacenaban seis *loci* STR (FGA, TH01, VWA, D8S1179, D18S51 y D21S11) y en 1999 se expandió a 10 (D3S1358, D16S539, D2S1338 y D19S43) (UK National DNA Database, 2022a).

Para octubre 6 del 2022 (actualización más reciente a la fecha), cuenta con un registro de más de 8.000.000 de perfiles genéticos (UK National DNA Database, 2022b). Por otro lado, en España se encuentra la base de datos policial de identificadores obtenidos a partir del ADN, creada bajo la Ley Orgánica 10/2007 (Jefatura del Estado Español, 2007) donde se mantienen los perfiles genéticos de uso forense, que han sido obtenidos de muestras de personas (sospechosos, detenidos, imputados-investigados) y aquellos perfiles genéticos obtenidos de los indicios biológicos recogidos con la ocasión de un hecho delictivo. Actualmente, en esta base de datos reposan más de 400.000 registros de perfiles genéticos (Ministerio del Interior de España, 2018).

En nuestro continente, particularmente en Estados Unidos, el CODIS (Combined DNA Index System) del Laboratorio del FBI comenzó como un proyecto piloto de software en 1990, sirviendo a 14 laboratorios estatales y locales. La Ley de Identificación de ADN de 1994 formalizó la autoridad del FBI para establecer un Sistema Nacional de Índice de ADN (NDIS) para fines policiales (Federal Bureau of Investigations, 2020b). En 1998 se puso en funcionamiento y hoy día más de 190 laboratorios públicos de aplicación de la ley participan en NDIS en los Estados Unidos (Federal Bureau of Investigations, 2021a) y reposan en la base de datos aproximadamente 14.000.000 de perfiles. A nivel internacional, más de 90 laboratorios de más de 50 países utilizan el software CODIS para sus propias iniciativas de bases de datos (Federal Bureau of Investigations, 2021a).

Con un enfoque más académico, es importante mencionar la creación de una base de datos de referencia estándar del Instituto Nacional de Estándares y Tecnología (National Institute of Standards and Technology) del Departamento de Comercio de Estados Unidos, llamada STRBase (SRD - 130) creada en 1997 y cuyo objetivo es beneficiar la investigación y la aplicación de marcadores de ADN de repetición en tándem cortos para las pruebas de identidad humana en genética forense e identificación humana (Butler & Reeder, 1997). En esta base reposan las distribuciones de frecuencias alélicas por marcadores y por poblaciones tanto de caracterizaciones genéticas estatales y gubernamentales, así como de proyectos de investigación en genética de poblaciones.

En Latinoamérica, siguiendo con las aplicaciones forenses, un país ejemplo ha sido Argentina que realizó esfuerzos por estados para generar un archivo sistemático público de material genético y muestras biológicas de familiares de personas secuestradas y desaparecidas durante la dictadura militar argentina (Ministerio de Ciencia, Tecnología e Innovación. Gobierno de Argentina., 2021)

mediante la Ley 23.511 de 1987 (Asociación Civil Abuelas de Plaza de Mayo, 1987). La unificación de las bases de datos regionales y por estado dio lugar en el 2009 al Banco Nacional de Datos Genéticos (BNDG) como un espacio de obtención, almacenamiento y análisis de las muestras genéticas necesarias para el esclarecimiento de delitos de lesa humanidad en Argentina, que garantizara así la conservación de los perfiles genéticos de cada uno de los miembros de las familias que sufrieron el secuestro y desaparición de algún integrante y que depositaran sus muestras en él (Banco Nacional de Datos Genéticos, 2021).

En el caso específico de Colombia, se presentó el Proyecto en Cámara 326 del 2020 y en Senado 442 del 2021 por medio del cual se propone crear el Banco nacional de datos genéticos vinculados a la comisión de delitos violentos de alto impacto. [Banco nacional de datos genéticos]. Fue aprobado en primer (24/11/2020) y segundo debate (23/3/2021); sin embargo, para junio del 2021 fue archivado por vencimiento de términos (Senado de la República de Colombia, 2021).

1.1.2. Formulación y aplicación de modelos teóricos y métodos de análisis estadístico en genética a partir del uso de información reposada en bases de datos.

La historia de los humanos puede ser contada a través de las diferencias genéticas entre poblaciones formulando hipótesis conjuntas al conocimiento antropológico (Pickrell & Reich, 2014). Con la creciente cantidad de datos genéticos que reposan en bases de datos, así como el avance de los modelos teóricos, los procesos históricos y prehistóricos que juegan un papel importante en la configuración de la diversidad genética observada pueden identificarse mejor (Serre et al, 2004; Rosenberg et al., 2005).

La información genética dispuesta en grandes bases de datos pertenece en su mayoría a estudios del ámbito forense en el cual se han tipificado gran cantidad de individuos distribuidos en todo el mundo con un set pequeño (~20) de marcadores tipo STRs. Estas bases de datos han sido desarrolladas para responder preguntas relacionadas con la identificación individual y con poder discriminar a un individuo de una población particular (Kanitz, Guillot, Antoniazza, Neuenschwander, & Goudet, 2018; Poloni, Currat, & Silva, 2012). Sin embargo, lo que las hace informativas es que contienen muestras dispersas globalmente y por cada población estudiada han tipificado un número alto de individuos (Rosenberg et al., 2005).

Esto es lo contrario a los estudios en genética de poblaciones, en los que, a razón del costo de las investigaciones, que también se remiten a grupos poblacionales muy específicos, el tamaño de muestra suele ser pequeño debido a la selección a priori de individuos, por lo que la representación general de una población puede ser deficiente. Debido a que el campo de la genética forense genera una cantidad basta de datos poblacionales de polimorfismos tipo STR en muestras distribuidas globalmente, se ha estudiado el poder de estos set de datos para responder preguntas relacionadas a la evolución humana y su diversidad, teniendo en cuenta dos tipos de recursos: las frecuencias alélicas disponibles en las bases de datos y datos genotípicos que se pueden encontrar en pocas bases de datos o en artículos científicos (Bentayebi, Abada, Izhmad, & Amzazi, 2014; Callegari-Jacques, Tarazona-Santos, Gilman, Herrera, Cabrera, Dos Santos, et al., 2011; Houck, 2015; Khubrani, Wetton, & Jobling, 2019; Poloni et al., 2012; Sun et al., 2013).

Lo que nos plantea este escenario en el que se tienen un conjunto grande de datos con amplia distribución global, es que estos pueden constituir una fuente importante de información sobre la diversidad genética humana, a pesar del número relativamente bajo de marcadores tipificados. Debido a que la geografía como motor de la diversidad genética ha sido relacionada con la estructuración de las poblaciones (Cavalli-Sforza, 1994), los patrones genéticos observados muestran agrupaciones poblacionales principales asociadas con la ubicación continental (europeos, asiáticos, melanesios, nativos americanos y africanos) (Rosenberg et al, 2004; Prugnolle, Manica, Balloux, 2005; Fujimura et al., 2014).

Sin embargo, es necesario resaltar que también existe una relación entre el número de variables y el nivel de agrupación, en la que a menor número de loci empleados en el análisis de estructura, se encuentra menor número de clústers. Y cuando se analiza un mayor número de loci y se incrementa el número posible de poblaciones ancestrales (o de referencia) k , se encontrará un mayor número de clústers o de núcleos poblacionales (Falush, Stephens, & Pritchard, 2007; Holsinger & Weir, 2009; Pritchard, Wen, & Falush, 2009).

Es aquí cuando la formulación y aplicación de modelos teóricos y métodos de análisis estadístico toman cabida y muestran que la informatividad de los datos genéticos reposados en bases de datos puede tener un poder e impacto mayor al que se esperaba inicialmente gracias a las nuevas interpretaciones y aproximaciones teóricas. Como ejemplo de esto, se tienen los distintos modelos

de inferencia de ancestralidad usando distintas bases conceptuales como: 1. La evidencia histórica combinada con métodos bayesianos, 2. La aplicación del análisis discriminante, 3. El uso del análisis de componentes principales, 4. El análisis espacial (Bradburd, Coop, & Ralph, 2018; Byun et al., 2017; Rishishwar, Conley, Vidakovic, & Jordan, 2015; Thornton & Bermejo, 2014) y 5. La simulación de muestreos poblacionales aleatorizados mediante métodos de Monte Carlo (Mode & Gallop, 2008; Al-Dalky, Taha, Homouz & Qasaimeh, 2016; Baladeh & Khakzad, 2018; de Pádua, Pitombeira-Neto & de Athay de Prata, 2018).

1.1.3. Aleatorización y métodos de Monte Carlo.

Como se ha mencionado, una de las aplicaciones de métodos de análisis estadístico en aplicación de modelos teóricos en genética de poblaciones es el uso de los métodos de Monte Carlo. Estos son una clase de algoritmos computacionales que se basan en muestreos aleatorios repetitivos para obtener resultados numéricos. Su base está en usar la aleatoriedad para resolver los problemas que podrían ser en principio deterministas (Manly, 1991a). El objetivo es comprobar la hipótesis de que un conjunto de datos constituye una muestra aleatoria de una población específica cuya distribución tiene parámetros conocidos y está perfectamente definida (Losilla Vidal, 1994; Manly, 1991a). El procedimiento base consiste en simular muestras a partir de una población dada, obtener el valor del estadístico de contraste para cada muestra simulada y ubicar posteriormente en la distribución muestral los valores simulados del valor del estadístico de contraste obtenido al aplicar la prueba en la muestra real (Losilla Vidal, 1994).

Durante la selección de muestras aleatorias mediante el método de Monte Carlo, la significancia de una prueba estadística observada se evalúa comparándola con una muestra de pruebas estadísticas obtenidas al generar muestras aleatorias de acuerdo con algún modelo supuesto (Manly, 1991a). Si el modelo supuesto implica que todos los ordenamientos de datos son igualmente probables, entonces esto equivale a que la prueba de aleatorización logró un muestreo aleatorio con una distribución de probabilidad aleatoria. Por lo tanto, las pruebas de aleatorización pueden considerarse casos especiales dentro de una categoría más amplia de las pruebas de Monte Carlo (Manly, 1991a).

La aplicación de los Métodos de Monte Carlo se centra fundamentalmente en la investigación teórica de los métodos estadísticos en técnicas de muestreo, ya que en el contexto aplicado el conocimiento de los parámetros poblacionales es inusual (Losilla Vidal, 1994). Generalmente, se dispone sólo de una muestra de datos extraídos de una población y se desea conocer, a partir de estos datos, algunas características de la población raíz u origen (Manly, 1991b). Se puede expresar lo dicho como (Figura 1):

ESTADÍSTICO → PARÁMETRO

Promedio muestral → Promedio poblacional

Figura 1 . *Generalidad Métodos de Monte Carlo: A través de los métodos de Monte Carlo, puedo llegar a inferir un parámetro de una población a partir de un estadístico proveniente de una muestra de la población de interés.*

Tomando como ejemplo la Figura 1., a partir del promedio muestral de x_n observaciones [13], cuando n tiende a infinito podría acercarme al parámetro o el valor estimado $E(X)$ [14].

$$x_1, x_2, x_3, \dots, x_n \quad x \sim U(0,1) \quad [1]$$

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad n \rightarrow \infty \quad E(X) \quad [2]$$

Aplicando una función que llamaremos promedio desde $i = 1$ a $i = n$, se tiene:

$$f(x_1), f(x_2), f(x_3), \dots, f(x_n) \quad x \sim U(0,1) \quad [3]$$

$$\frac{1}{n} \sum_{i=1}^n f(x_i) \quad n \rightarrow \infty \quad E(X) \quad [4]$$

Dado que n tiende a ∞ , se aplica la integral definida de 0 a 1 y se obtiene el estimado poblacional a partir de la muestra origen o raíz:

$$\int_0^1 \sum_x f(x) dx = E(f(x)) \quad [5]$$

Los métodos de aleatorización de Monte Carlo también pueden ser utilizados en los casos en los que desee generar nuevas muestras de datos mediante técnicas de muestreo (con o sin reposición, manteniendo constante el tamaño muestral o no), a partir de los datos de la muestra original. Entre estas técnicas resaltan por su aplicabilidad y poder de generación de pruebas de aleatorización o *randomization tests* (Efron, 1982; Losilla Vidal, 1994; Manly, 1991a, 1991b):

- *El Jackknife*: Las estimaciones son realizadas a partir de las muestras resultantes de la exclusión de una observación o varias del conjunto inicial de datos (Tukey, 1958).
- *La validación recíproca (cross-validation)*: En esta técnica los datos se dividen en dos grupos o dos mitades ($n - l$ y l o $\frac{n}{2}$ y $\frac{n}{2}$). Sobre el primer grupo se realizan las estimaciones de los parámetros de interés y, por último, se verifican dichas estimaciones aplicándolas (mediante ecuaciones predictivas) al segundo grupo de datos, obteniéndose un indicador fiable de las posibles estimaciones para los nuevos conjuntos de datos.
- *El bootstrap*: Se basa en la generación aleatoria de nuevos conjuntos de datos mediante un muestreo aleatorio con reposición de los datos de la muestra original (Efron, 1979).

Los métodos de Monte Carlo también se pueden utilizar para calcular los intervalos de confianza para los parámetros de la población. Esencialmente la idea es usar datos generados por computadora para determinar la cantidad de variación que se espera en muestras estadísticas (Manly, 1991b, 1991a).

1.1.3.1. Elementos de una simulación de Monte Carlo y consideraciones generales de la prueba:

a. Construcción del modelo a simular

El mayor desafío y el elemento más importante de la simulación es la construcción del modelo o algoritmo a simular. Son varios factores a tener en cuenta, pero el más crucial es determinar cuál es la distribución de probabilidad que se ajusta a la variable de interés. En la Figura 1, se conoce que el modelo a simular es un promedio y se sabe que la variable x tiene una distribución uniforme de 0 a 1.

b. Identificación de variables aleatorias de entrada (INPUT) y de salida (OUTPUT) dentro del modelo

El investigador es quien diseña su experimento y por consiguiente debe preguntarse y plantear qué variables aleatorias serán las que introduce en su modelo. En [3] y [4] se conoce la variable aleatoria x , que es el *input* de la función $f(x)$ de la que se conoce su dominio. Siendo $\frac{1}{n} \sum_{i=1}^n f(x_i)$ la variable salida.

c. Correr la simulación

Al tener el diseño de la simulación de interés, solo restaría ejecutar el comando de corrida, pero nuevamente se hace necesario responderse las preguntas: ¿Cuántas iteraciones son necesarias para lograr acercarse al estadístico de interés a un parámetro estimado? ¿Hay algún otro estadístico que quiera calcular para describir el comportamiento de mis M_n muestras? ¿Qué método de representación gráfica aplicaré para visualizar el comportamiento de las M_n muestras en dicha simulación.

d. Resultado de la simulación:

Cuando se obtiene el resultado de una simulación como en [5], es cuando se deben tomar decisiones y establecer, por ejemplo, si los resultados de la simulación se ajustan a los datos poblacionales reales.

Teniendo en cuenta que el estadístico a hallar es una medida cuantitativa, derivada de un conjunto de datos de una muestra, queremos que este valor se aleje lo menor posible del valor poblacional o parámetro desconocido. En este momento, debemos tener en cuenta las características y bondades del estimador estadístico empleado:

- *Insesgamiento*: Esta propiedad hace referencia a cuán central se encuentra mi estimador referente al parámetro. Esto se evalúa cuando a medida que se distribuye el estimador en el muestreo coincide con el parámetro a estimar.
- *Eficiencia*: Un estimador es eficiente a medida que su varianza es mínima. Así, un estimador es más eficiente que otro si su varianza es menor al contrastado (intervalos de confianza por aleatorización).

- *Suficiencia*: Cuando un estimador resume la información relevante contenida en la muestra y ningún otro estimador puede dar información adicional acerca del parámetro desconocido es suficiente.
- *Consistencia*: Esta propiedad tiene relación con el tamaño de la muestra y el concepto de límite; así cuando el tamaño muestral es muy grande puede estimar el parámetro poblacional sin error.
- *Invarianza*: La invarianza hace referencia al método de estimación en la que se calcula el estimador no permita que haya variaciones ente cambios de escala sino de origen.
- *Robustez*: Cuando a pesar de que la hipótesis de partida es incorrecta, los resultados del estimador serán cercanos a los parámetros.

1.2. Planteamiento del problema

Durante los últimos 25 años en Colombia se han llevado a cabo investigaciones que han permitido exponer la riqueza de sus poblaciones desde una perspectiva genética y se han realizado diversos reportes de frecuencias alélicas y estimadores forenses para poblaciones humanas específicas que ayudan a caracterizar la población. Estas frecuencias y estimadores son empleados tanto en el estudio de genética poblaciones para aclarar hipótesis de estructura genética y de poblamiento como también para generar parámetros en cálculos de índices de paternidad útiles en pruebas de paternidad y filiación al establecer el parentesco de individuos en el caso del derecho de identidad como lo contempla el Artículo 25 de la Ley 1098/06.

Los reportes de probabilidades de paternidad son expresados como estimadores puntuales a pesar de que en la práctica las frecuencias con las que se obtienen pertenecen a una muestra de población y no a la población total. Es por esto que se propone realizar un modelo de cálculo que matemáticamente exprese el resultado de una prueba como intervalo de confianza a partir de remuestreos aleatorizados por los métodos de Monte Carlo.

La generación de modelos teóricos formales (matemáticos y estadísticos) para cálculos de índices y probabilidades de paternidad es una línea de investigación activa en grupos de trabajo de genética

estadística en el mundo (Anderson, 2006; European Network of Forensic Science Institutes, 2020; GHEP-ISFG, 2020; Houck, 2015; Hu et al., 2016; Kanitz et al., 2018; Moroni, Gasbarra, Arjas, Lukka, & Ulmanen, 2011; Silva, Pereira, Poloni, & Currat, 2012; Wurmb-schwark et al., 2015).

Alternativamente los cálculos de índices y probabilidades de paternidad se pueden reportar como intervalos de confianza y no como estimadores puntuales ya que su cálculo depende de las frecuencias alélicas obtenidas de una muestra de población, que pueden cambiar con el tipo de muestreo empleado (Figura 1).

Teniendo en cuenta que la relación $\frac{x}{y}$ para el *IP* en un caso de paternidad es calculada a partir de la distribución de frecuencias de una población de interés llamada *población de referencia*. Es importante mostrar que este término, aunque se refiere a una *población* realmente no es un parámetro sino un estadístico, ya las frecuencias son obtenidas a partir del recuento de alelos observados en una muestra poblacional (Figura 2).

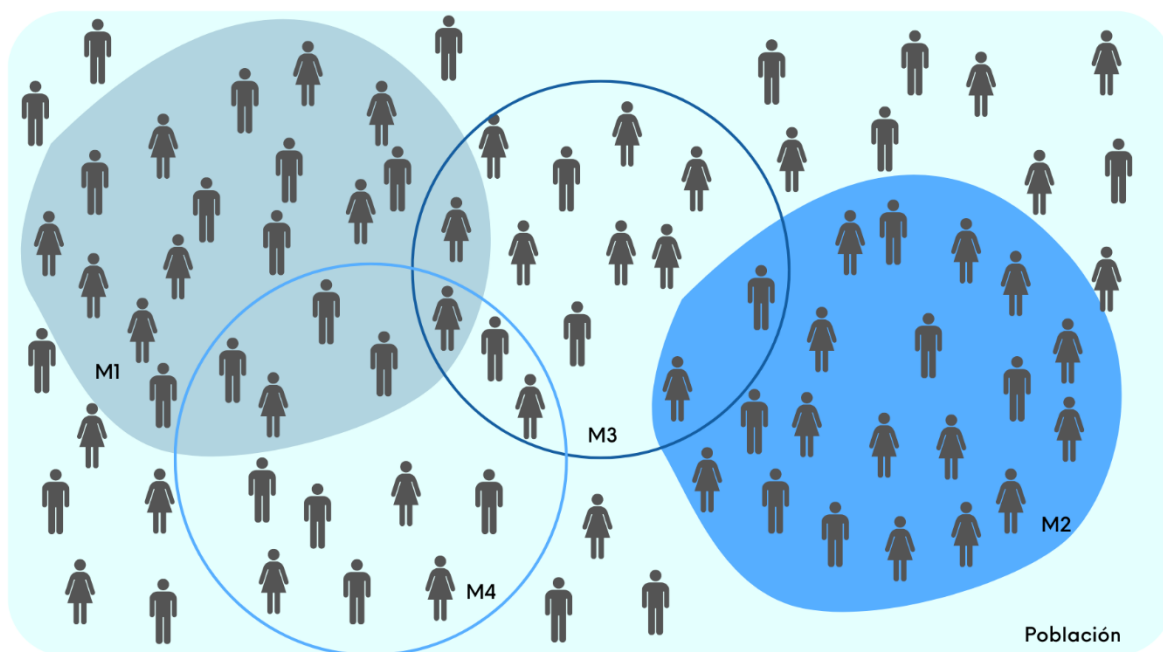


Figura 2 Las frecuencias alélicas de la población de referencia corresponden a un estadístico y no a un parámetro.

Supongamos que en un estudio de diversidad y caracterización genética de la Población (P) enmarcada en el cuadro azul de la Figura 2., sólo se pudo tomar información de M_1 porque eran los individuos disponibles en ese momento. Al poco tiempo, otro grupo de investigadores quiso realizar otro estudio de diversidad genética para la misma población P , pero sólo se encontraban disponibles los individuos de la muestra de población M_2 .

Ninguno de los dos grupos de investigadores puede asegurar realizar un censo porque sería muy costoso y además requeriría mucho tiempo de esfuerzo de muestreo; tampoco podría hacerlo un tercer o cuarto grupo de investigadores por lo que, en cada evento, se inferirán datos de una población a partir de M_3 y M_4 muestras. Y así, cuando un grupo de investigadores n se interese en la misma población, tomará una muestra poblacional M_n . Por consiguiente, se generarán n distribuciones de frecuencias alélicas que podrán ser empleadas en cálculos de índices y probabilidades de paternidad.

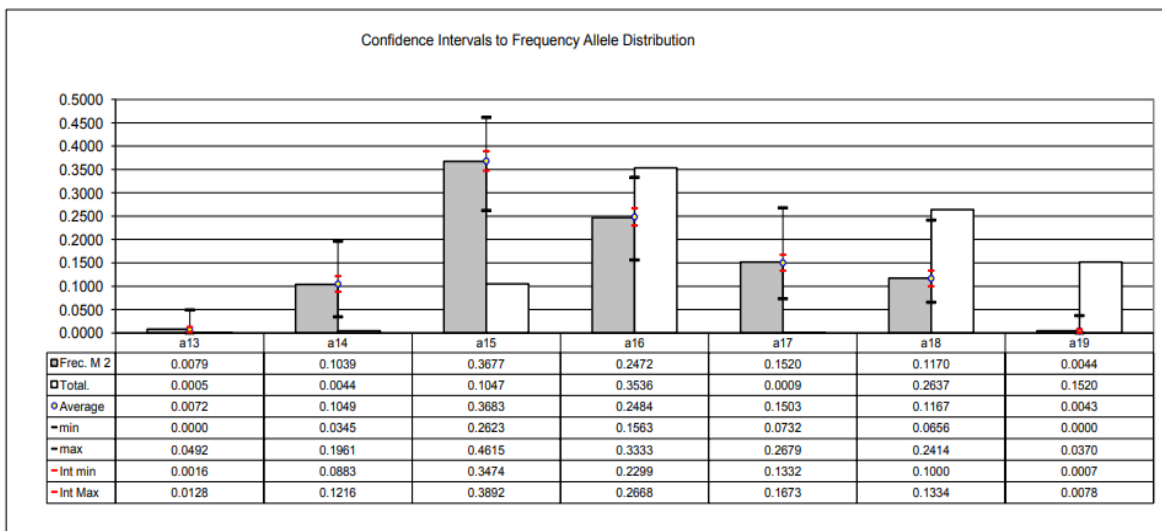


Figura 3 Distribución de frecuencias alélicas del alelo a13 al a19 del marcador D3S1358 en la población de Bogotá. Tomado de Usaquén Martínez, 2012

Si se toman las n distribuciones de frecuencias alélicas de las M_n muestras, se podría analizar su comportamiento y evidenciar en qué intervalos de mínima y máxima frecuencia se encuentra la distribución de cada alelo, como se muestra en la Figura 3. En este caso, en la población de Bogotá se realizan muestreos consecutivos aumentando el tamaño de muestra y a partir de estas se calcula

la distribución de frecuencias alélicas del alelo a13 al a19 del marcador D3S1358 para cada una de las muestras simuladas, logrando calcular el intervalo de confianza con el mínimo y máximo para esa población.

Tomando este ejercicio como ejemplo y dado que las distribuciones de frecuencias alélicas de esta población pueden expresarse en términos de intervalos de confianza y no como estimadores puntuales; también los índices de paternidad y probabilidad de paternidad pueden ser expresados como intervalos de confianza teniendo en cuenta un mismo caso a lo largo de diferentes poblaciones de referencia.

1.3. Objetivos.

Teniendo en cuenta esta inconsistencia teórica se han planteado los siguientes objetivos en esta investigación.

1.3.1. Objetivo general

- Proponer un modelo de cálculo por intervalos de confianza de los índices y probabilidades de paternidad empleando métodos de aleatorización de Monte Carlo que seleccionen individuos de una muestra de población creando n sub-muestras.

1.3.2. Objetivos específicos

- Reportar los resultados de los índices de paternidad en forma de intervalos de confianza en lugar de valores puntuales, mediante un análisis experimental que modifique la población de referencia utilizada en los cálculos.
- Plantear un modelo de cálculo de IP y W por uno que considere los intervalos de confianza.

- Determinar las coincidencias en perfiles genéticos que identifiquen que puede haber más de un presunto padre para un hijo en cuestión y una madre en la población total.

El realizar un estudio de carácter teórico con fundamentos matemáticos y estadísticos permite conocer el efecto que tiene la escogencia de una población de referencia en una prueba de identificación creando líneas de análisis robustas para la expresión del resultado y su interpretación, ya que al realizar un elevado número de repeticiones se disminuye la incertidumbre de la medición.

Se espera contribuir al fortalecimiento y robustecimiento de la línea teórica de los cálculos empleados en las pruebas de filiación y se espera incentivar el buen uso y reporte de estos índices al proponer una línea de análisis, soportada en la comprensión de los cálculos de índices de paternidad y probabilidad de paternidad para ser empleado por miembros de la comunidad académica y usuarios de pruebas de paternidad.

Todo este trabajo se realiza bajo sombra de la importancia de la genética de poblaciones en el quehacer de las pruebas de filiación, teniendo en cuenta que no sólo el factor genético moldea las dinámicas poblacionales sino también otros factores biológicos y culturales que caracterizan nuestro país. Para lograr caracterizar una población primero se debe conocer el contexto histórico y generar instrumentos de obtención de información complementaria (demografía, lengua, genealogía, factores migratorios y socioculturales) que permitan establecer qué población tipo es la que estamos estudiando y nos permitan generar clasificaciones *a priori* robustas, que en un futuro podrán ser empleadas para determinar las distribuciones de frecuencias de las poblaciones y emplearlas en el quehacer de las pruebas de paternidad. Esto con ayuda de la interacción entre disciplinas científicas como la matemática y ciencias biológicas.

1.4. Referencias

- Al-Dalky R., Taha K., Homouz D., Qasaimeh M. (2016). Applying Monte Carlo Simulation to Biomedical Literature to Approximate Genetic Network. IEEE/ACM Transactions on Computational Biology and Bioinformatics
- Alonso Alonso, A. (2019). Las bases de datos de ADN de interés forense. In M. C. Crespillo Márquez

-
- & P. A. Barrio Caballero (Eds.), *Genética Forense: Del laboratorio a los tribunales* (I, pp. 425–443). España: Díaz de Santos. Retrieved from <https://www.editdiazdesantos.com/libros/9788490522134/Crespillo-Marquez-Genetica-forense.html>
- Alonso, L. A., & Usaquén, W. (2012). Y-chromosome and surname analysis of the native islanders of San Andrés and Providencia (Colombia). *HOMO-Journal of Comparative Human Biology*, 1–14. <https://doi.org/10.1016/j.jchb.2012.11.006>
- Alonso Morales, L. A., Casas-Vargas, A., Castro, M. R., Resque, R., Ribeiro-dos-Santos, Â. K., Santos, S., ... Usaquén, W. (2018). Paternal portrait of populations of the middle Magdalena river region (Tolima and Huila, Colombia): New insights on the peopling of central America and northernmost South America. *PLoS ONE*, 13(11), 1–20. <https://doi.org/10.1371/journal.pone.0207130>
- Anderson, K. G. (2006). How well does paternity confidence match actual paternity? Evidence from worldwide nonpaternity rates. *Current Anthropology*, 47(3), 513–520. <https://doi.org/10.1086/504167>
- Asociación Civil Abuelas de Plaza de Mayo. (1987, May 13). Ley 23.511 – Banco Nacional de Datos Genéticos – Consejo de Derechos Humanos. Retrieved July 16, 2020, from <http://cdh.defensoria.org.ar/ley-23-511-banco-nacional-de-datos-geneticos/>
- Baladeh A. E. & Khakzad N. (2018). Integration of Genetic Algorithm and Monte Carlo Simulation for System Design and Cost Allocation Optimization in Complex Network. 2018 3rd International Conference on System Reliability and Safety (ICSRS)
- Banco Nacional de Datos Genéticos. Ministerio de Ciencia Tecnología e Innovación Productiva. (2020). Historia del BNDG. Retrieved July 16, 2020, from <https://www.argentina.gob.ar/ciencia/bndg/historia>
- Benítez-Páez, A., & Reyes, H. O. (2003). Allelic frequencies at 12 STR loci in Colombian population. *Forensic Science International*, 136(1–3), 86–88. [https://doi.org/10.1016/S0379-0738\(03\)00220-2](https://doi.org/10.1016/S0379-0738(03)00220-2)
- Bentayebi, K., Abada, F., Izhmad, H., & Amzazi, S. (2014). Genetic ancestry of a Moroccan population as inferred from autosomal STRs. *MGENE*, 2, 427–438. <https://doi.org/10.1016/j.mgene.2014.04.011>
- Bieber, F. R., Brenner, C. H., & Lazer, D. (2006, June 2). Finding criminals through DNA of their relatives. *Science*. American Association for the Advancement of Science.

<https://doi.org/10.1126/science.1122655>

- Bolnick, D. A., Bolnick, D. I., & Smith, D. G. (2006). Asymmetric Male and Female Genetic Histories among Native Americans from Eastern North America. *Molecular Biology and Evolution*, 23(11), 2161–2174. <https://doi.org/10.1093/molbev/msl088>
- Bolnick, D. A., Raff, J. A., Springs, L. C., Reynolds, A. W., & Miró-Herrans, A. T. (2016). Native American Genomics and Population Histories. *Annual Review of Anthropology*, 45(1), 319–340. <https://doi.org/10.1146/annurev-anthro-102215-100036>
- Bradburd, G. S., Coop, G. M., & Ralph, P. L. (2018). Inferring Continuous and Discrete Population Genetic. *Genetics*, 210(September), 33–52. <https://doi.org/10.1534/genetics.XXX.XXXXXX>
- Braga, Y., Arias B., L., & Barreto, G. (2012). Diversity and genetic structure analysis of three Amazonian Amerindian populations from Colombia. *Colombia Médica*, 43(2), 133–140. Retrieved from <http://www.scielo.org.co/pdf/cm/v43n2/v43n2a05.pdf>
- Bravo Aguilar, M. L. J. (2009a). Investigación de la Paternidad Biológica. In *La verdad genética de la paternidad* (I, pp. 45–80). Medellín, Antioquia: Universidad de Antioquia.
- Bravo Aguilar, M. L. J. (2009b). Microsatélites o secuencias cortas repetidas una a continuación de la otra en tándem y probabilidad de exclusión a priori de la paternidad. In *La verdad genética de la paternidad* (I, pp. 28–44). Medellín, Antioquia: Editorial Universidad de Antioquia.
- Bravo, M. L., Moreno, M. A., Builes, J. J., Salas, A., Lareu, M. V., & Carracedo, A. (2001). Autosomal STR genetic variation in negroid Chocó and Bogotá populations. *International Journal of Legal Medicine*, 115(2), 102–104. <https://doi.org/10.1007/s004140100223>
- Builes, J. J., Ospino, J. M., Manrique, A., Aguirre, D. P., Mendoza, L., Bravo, M. L. J., ... Gusmão, L. (2013). Genetic population data of 38 autosomal InDels for the Amerindian community Embera-Chami of Lapo, Antioquia-Colombia. *Forensic Science International: Genetics Supplement Series*, 4(1), 170–171. <https://doi.org/10.1016/j.fsigss.2013.10.088>
- Burgos, G., Restrepo, T., Ibarra, A., Gaviria, A., Machado, G., Mora, C., & Lizarazo, R. (2015). Allelic frequencies and forensic parameters for miniSTRs D10S1248, D14S1434 and D22S1045 (NC01) in a sample from Central Andean Colombian region. *Forensic Science International: Genetics Supplement Series*, 5, e81–e82. <https://doi.org/10.1016/j.fsigss.2015.09.033>
- Butler, J. M., & Reeder, D. J. (1997, February 10). STRBase: Short Tandem Repeat DNA Internet Data Base. Retrieved July 16, 2020, from <https://strbase.nist.gov/>
- Byun, J., Han, Y., Gorlov, I. P., Busam, J. A., Seldin, M. F., & Amos, C. I. (2017). Ancestry inference using

- principal component analysis and spatial analysis: A distance-based analysis to account for population substructure. *BMC Genomics*, 18(1), 1–12. <https://doi.org/10.1186/s12864-017-4166-8>
- Callegari-Jacques, S. M., Tarazona-Santos, E. M., Gilman, R. H., Herrera, P., Cabrera, L., dos Santos, S. E. B., ... Salzano, F. M. (2011). Autosomal STRs in native South America-Testing models of association with geography and language. *American Journal of Physical Anthropology*, 145(3), 371–381. <https://doi.org/10.1002/ajpa.21505>
- Casas-Vargas, A., Romero, L. M., Usaquén, W., Zea, S., Silva, M., Briceño, I., ... Rodríguez, J. V. (2017). Diversidad del ADN mitocondrial en restos óseos prehispánicos asociados al templo del sol en los andes orientales colombianos. *Biomedica*, 37(4), 1–41. <https://doi.org/10.7705/biomedica.v37i4.3377>
- Castillo, A., Gil, A., Pico, A., Vargas, C., Yurrebaso, I., & García, O. (2013). Genetic variation for 20 STR loci in a northeast Colombian population (Department of Santander). *Forensic Science International: Genetics Supplement Series*, 4(1). <https://doi.org/10.1016/j.fsigss.2013.10.152>
- CINEP. (1998a). *Colombia: País de Regiones. Región del Alto Magdalena - Región Suroccidental*. (F. Zambrano Pantoja, Ed.) (Tomo III). Santafé de Bogotá, Colombia: CINEP (Centro de Investigación y Educación popular), COLCIENCIAS.
- CINEP. (1998b). *Colombia: País de Regiones. Región Noroccidental - Región Cundiboyacense*. (F. Zambrano Pantoja, Ed.) (Tomo II). Santafé de Bogotá, Colombia: Investigación y Educación popular), COLCIENCIAS.
- CINEP. (1998c). *Colombia: País de Regiones. Región Occidental - Región Caribe*. (F. Zambrano Pantoja, Ed.) (Tomo I). Santafé de Bogotá, Colombia: CINEP (Centro de Investigación y Educación popular), COLCIENCIAS.
- Da Costa Francez, P. A., Rodrigues, E. M. R., De Velasco, A. M., & Dos Santos, S. E. B. (2012). Insertion-deletion polymorphisms-utilization on forensic analysis. *International Journal of Legal Medicine*, 126(4), 491–496. <https://doi.org/10.1007/s00414-011-0588-z>
- De Pádua Agripa Sales L., Pitombeira-Neto A., & de Athay de Prata. (2018). A genetic algorithm integrated with Monte Carlo simulation for the field layout design problem. *Oil & Gas Science and Technology - Rev. IFP Energies nouvelles*, 73, 24.
- Durán, R., Zarante, I., Acevedo, M. L., Villegas, M. R., Salazar, J., Bocanegra, B. Y., & Bernal, J. (2003). Allelic frequency of six STR loci in five Colombian cities. *Journal of Forensic Sciences*, 48(4), 887.

-
- Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/12877314>
- Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Annals of Statistics*, 7, 1–26.
- Efron, B. (1982). The jackknife, the bootstrap, and other resampling methods. *Society for Industrial and Applied Mathematics, CBMS-NSF(Monograph)*, 38.
- European Network of Forensic Science Institutes. (n.d.). DNA | ENFSI. Retrieved July 15, 2020, from <http://enfsi.eu/about-enfsi/structure/working-groups/dna/>
- Falush, D., Stephens, M., & Pritchard, J. K. (2007). Inference of population structure using multilocus genotype data: Dominant markers and null alleles. *Molecular Ecology Notes*, 7(4), 574–578. <https://doi.org/10.1111/j.1471-8286.2007.01758.x>
- Federal Bureau of Investigations. (2020a). CODIS - NDIS Statistics — FBI. Retrieved July 15, 2020, from <https://www.fbi.gov/services/laboratory/biometric-analysis/codis/ndis-statistics>
- Federal Bureau of Investigations. (2020b). Combined DNA Index System (CODIS) — FBI. Retrieved July 15, 2020, from <https://www.fbi.gov/services/laboratory/biometric-analysis/codis>
- Franco-Candela, F. A., & Barreto, G. (2017). Estructura genética de poblaciones indígenas del occidente colombiano mediante el uso de marcadores ligados al cromosoma Y. *Revista de La Academia de Ciencias Exactas, Físicas y Naturales*, 41(160), 281–289. Retrieved from <https://www.raccefyn.co/index.php/raccefyn/article/view/476/311>
- Gaviria, A., Ibarra, A. A., Jaramillo, N., Palacio, O. D., Acosta, M. A., Brion, M., & Carracedo, Á. (2004). Nineteen autosomal microsatellite data from Antioquia (Colombia). *Forensic Science International*, 143(1), 69–71. <https://doi.org/10.1016/j.forsciint.2004.01.007>
- GHEP-ISFG. (2020). Statistics Working group | GHEP-ISFG. Retrieved July 17, 2020, from <https://ghep-isfg.org/en/estadistica-genetica-forense/>
- Gómez, M. V., Reyes, M. E., Cárdenas, H., & García, O. (2003). Genetic variation for 12 STRs loci in a Colombian population (Department of Valle del Cauca). *Forensic Science International*, 137(2–3), 235–237. [https://doi.org/10.1016/s0379-0738\(03\)00297-4](https://doi.org/10.1016/s0379-0738(03)00297-4)
- Goodwin, W., Linacre, A., & Hadi, S. (2011). *An Introduction to Forensic Genetics. Journal of Chemical Information and Modeling* (Second Edi, Vol. 53). Wiley-Blackwell. A John Willy & Sons, Ltd., Publication. <https://doi.org/10.1017/CBO9781107415324.004>
- Holsinger, K. E., & Weir, B. S. (2009). Genetics in geographically structured populations : defining , estimating and interpreting F_{ST}. *Nature Reviews. Genetics*, 10, 639–650. <https://doi.org/10.1038/nrg2611>

-
- Homburger, J. R., Moreno-Estrada, A., Gignoux, C. R., Nelson, D., Sanchez, E., Ortiz-Tello, P., ... Bustamante, C. D. (2015). Genomic Insights into the Ancestry and Demographic History of South America. *PLoS Genetics*, *11*(12), 1–26. <https://doi.org/10.1371/journal.pgen.1005602>
- Houck, M. M. (2015). *Forensic Biology* (Advanced F). San Diego, CA, USA.: Elsevier Inc.
- Hu, P., Hsieh, M. H., Lei, M. J., Cui, B., Chiu, S. K., & Tzeng, C. M. (2016). A Simple Algorithm for Population Classification. *Scientific Reports*, *6*, 1–5. <https://doi.org/10.1038/srep23491>
- Hunley, K., & Healy, M. (2011). The impact of founder effects, gene flow, and European admixture on native American genetic diversity. *American Journal of Physical Anthropology*, *146*(4), 530–538. <https://doi.org/10.1002/ajpa.21506>
- Ibarra, A., Restrepo, T., Rojas, W., Castillo, A., Amorim, A., Martínez, B., ... Gusmão, L. (2014). Evaluating the X Chromosome-Specific Diversity of Colombian Populations Using Insertion / Deletion Polymorphisms. *PLoS ONE*, *9*(1), 1–10. <https://doi.org/10.1371/journal.pone.0087202>
- Jefatura del Estado Español. (2007). Ley Orgánica 10/2007, de 8 de octubre, reguladora de la base de datos policial sobre identificadores obtenidos a partir del ADN. Retrieved July 16, 2020, from http://noticias.juridicas.com/base_datos/Admin/lo10-2007.html
- Julieta Avila, S., Briceño, I., & Gómez, A. (2009). Genetic population analysis of 17 Y-chromosomal STRs in three states (Valle del Cauca, Cauca and Nariño) from Southwestern Colombia. *Journal of Forensic and Legal Medicine*, *16*(4), 204–211. <https://doi.org/10.1016/j.jflm.2008.12.002>
- Kanitz, R., Guillot, E. G., Antoniazza, S., Neuenschwander, S., & Goudet, J. (2018). Complex genetic patterns in human arise from a simple range-expansion model over continental landmasses. *PLoS ONE*, *13*(2), 1–16. <https://doi.org/10.1371/journal.pone.0192460>
- Keyeux, G., & Usaquén, W. (2006). Rutas migratorias hacia Sudamérica y poblamiento de las cuencas de los ríos Amazonas y Orinoco, deducidas a partir de estudios genéticos moleculares. In Gaspar Morcote, S. Mora, & C. Calvo (Eds.), *Pueblos y paisajes antiguos de la selva amazónica* (p. 415). Bogotá: Universidad Nacional de Colombia, Editorial Unibiblos.
- Khubrani, Y. M., Wetton, J. H., & Jobling, M. A. (2019). Forensic Science International : Genetics Analysis of 21 autosomal STRs in Saudi Arabia reveals population structure and the influence of consanguinity. *Forensic Science International: Genetics*, *39*(December 2018), 97–102. <https://doi.org/10.1016/j.fsigen.2018.12.006>
- Kohlrausch, F. B., Callegari-Jacques, S. M., Tsuneto, L. T., Petzl-Erler, M. L., Hill, K., Hurtado, A. M., ... Hutz, M. H. (2005). Geography influences microsatellite polymorphism diversity in Amerindians.

-
- American Journal of Physical Anthropology*, 126(4), 463–470.
<https://doi.org/10.1002/ajpa.20042>
- Losilla Vidal, J. M. (1994). *MonteCarlo Toolbox de Matlab: Herramientas para un laboratorio estadístico fundamentado en técnicas Monte Carlo*. Universitat Autònoma de Barcelona, Barcelona, España.
- Lucía Hincapié, M., Gil, A. M., Pico, A. L., Gusmão, L., Rondón, F., Vargas, C. I., & Castillo, A. (2009). Análisis de la estructura genética en una muestra poblacional de Bucaramanga, Departamento de Santander. *Colombia Médica*, 40(4), 1–12.
- Luque Gutiérrez, J. A. (2019). Estudio de las relaciones de parentesco. In M. C. Crespillo Márquez & P. A. Barrio Caballero (Eds.), *Genética Forense: Del laboratorio a los tribunales* (I, pp. 351–381). España: Díaz de Santos.
- Manly, B. F. J. (1991a). Monte Carlo and other computer-intensive methods. In *Randomization and Monte Carlo Methods in Biology* (I, pp. 21–30). London, Great Britain: Chapman and Hall.
- Manly, B. F. J. (1991b). Randomization test and confidence intervals. In *Randomization and Monte Carlo Methods in Biology* (I, pp. 2–20). London, Great Britain: Chapman and Hall.
- Martínez, B., Builes, J. J., Aguirre, D., Mendoza, L., Hernández, L., & Marrugo, J. (2017). Autosomic STR database for an afrodescendant population sample of San Basilio de Palenque, Colombia. *Forensic Science International: Genetics Supplement Series*, 6, e555–e557.
<https://doi.org/10.1016/j.fsigss.2017.09.217>
- Martínez, B., Builes, J. J., & Caraballo, L. (2008). Genetic data analysis of nine STRs in two Caribbean Colombian populations: César and Guajira. *Journal of Forensic Sciences*, 53(1), 254–255.
<https://doi.org/10.1111/j.1556-4029.2007.00631.x>
- Martínez, B., Caraballo, L., Barón, F., Gusmão, L., Amorim, A., & Carracedo, A. (2006). Analysis of STR loci in Cartagena, a Caribbean city of Colombia. *Forensic Science International*, 160(2–3), 223.
<https://doi.org/10.1016/j.forsciint.2005.05.035>
- Martínez, B., Caraballo, L., Gusmão, L., Amorim, A., & Carracedo, A. (2005). Autosomic STR population data in two Caribbean samples from Colombia. *Forensic Science International*, 152(1), 79–81.
<https://doi.org/10.1016/j.forsciint.2005.01.016>
- Mesa, N. R., Mondragon, M. C., Soto, I. D., Parra, M. V., Duque, C., Ortiz-Barrientos, D., ... Ruiz-Linares, A. (2000). Autosomal, mtDNA, and Y-chromosome diversity in Amerinds: Pre- and Post-Columbian patterns of gene flow in South America. *American Journal of Human Genetics*, 67(5),

1277–1286. [https://doi.org/10.1016/S0002-9297\(07\)62955-3](https://doi.org/10.1016/S0002-9297(07)62955-3)

- Ministerio de Asuntos Exteriores y de Cooperación, & Oficina de Información Diplomática del Departamento de Relaciones Exteriores. (2017). *Ficha País República de Colombia*. Retrieved from http://www.exteriores.gob.es/Documents/FichasPais/COLOMBIA_FICHA PAIS.pdf
- Ministerio de Ciencia Tecnología e Innovación. Gobierno de Argentina. (n.d.). Banco Nacional de Datos Genéticos: La ciencia y la tecnología al servicio de la reparación de graves violaciones a los derechos humanos. Retrieved July 16, 2020, from <https://www.argentina.gob.ar/ciencia/bndg>
- Ministerio del Interior. Gobierno de España. Centro Tecnológico de Seguridad. (2018). *Base de datos policial de identificadores obtenidos a partir de ADN: desde el inicio hasta diciembre 2018*.
- Mode C., Gallop R. (2008). A review on Monte Carlo simulation methods as they apply to mutation and selection as formulated in Wright–Fisher models of evolutionary genetics. *Mathematical Biosciences* 211. 205–225
- Moreno-Estrada, A., Gravel, S., Zakharia, F., McCauley, J. L., Byrnes, J. K., Gignoux, C. R., ... Bustamante, C. D. (2013). Reconstructing the Population Genetic History of the Caribbean. *PLoS Genetics*, 9(11). <https://doi.org/10.1371/journal.pgen.1003925>
- Moroni, R., Gasbarra, D., Arjas, E., Lukka, M., & Ulmanen, I. (2011). Effects of Reference Population and Number of STR Markers on positive evidence in Paternity Testing. *Journal of Forensic Research*, 02(02). <https://doi.org/10.4172/2157-7145.1000119>
- Ossa, H., Aquino, J., Sierra, S., Ramírez, A., Carvalho, E. F., & Gusmão, L. (2015). Analysis of admixture in Native American populations from Colombia. *Forensic Science International: Genetics Supplement Series*, 5, e332–e334. <https://doi.org/10.1016/j.fsigss.2015.09.132>
- Ossa, Humberto, Aquino, J., Pereira, R., Ibarra, A., Ossa, R. H., Pérez, L. A., ... Gusmão, L. (2016). Outlining the ancestry landscape of Colombian admixed populations. *PLoS ONE*, 11(10), 1–15. <https://doi.org/10.1371/journal.pone.0164414>
- Ossa Reyes, H., Torres Ramírez, L. J., & Nieto Romero, L. V. (2009). Frecuencias alélicas y haplotípicas del Sistema hla clase i (loci a*, b*) en una población de indígenas Motilón-Barí, Norte de Santander, Colombia. *Nova*, 7(12), 131. <https://doi.org/10.22490/24629448.426>
- Palacio, O. D., Triana, O., Gaviria, A., Ibarra, A. A., Ochoa, L. M., Posada, Y., ... Carracedo, A. (2006). Autosomal microsatellite data from Northwestern Colombia. *Forensic Science International*, 160(2–3), 217–220. <https://doi.org/10.1016/j.forsciint.2005.05.034>

-
- Paredes, M., Galindo, A., Bernal, M., Avila, S., Andrade, D., Vergara, C., ... Carracedo, Á. (2003). Analysis of the CODIS autosomal STR loci in four main Colombian regions. *Forensic Science International*, 137(1), 67–73. [https://doi.org/10.1016/S0379-0738\(03\)00271-8](https://doi.org/10.1016/S0379-0738(03)00271-8)
- Poloni, E. S., Currat, M., & Silva, N. M. (2012). Human Neutral Genetic Variation and Forensic STR Data, 7(11). <https://doi.org/10.1371/journal.pone.0049666>
- Porras, L., Beltrán, L., Ortiz, T., Sanchez-Diz, P., Carracedo, A., & Henao, J. (2008). Genetic polymorphism of 15 STR loci in central western Colombia. *Forensic Science International: Genetics*, 2(1), e7–e8. <https://doi.org/10.1016/j.fsigen.2007.08.004>
- Pritchard, J. K., Wen, X., & Falush, D. (2009). *Documentation for structure software: Version 2.3*. Retrieved from <https://web.stanford.edu/group/pritchardlab/structure.html>
- Rey, M., Gutiérrez, A., Schroeder, B., Usaquén, W., Carracedo, A., Bustos, I., & Giraldo, A. (2003). Allele frequencies for 13 STR's from two Colombian populations: Bogotá and Boyacá. *Forensic Science International*, 136(1–3), 83–85. [https://doi.org/10.1016/S0379-0738\(03\)00221-4](https://doi.org/10.1016/S0379-0738(03)00221-4)
- Rishishwar, L., Conley, A. B., Vidakovic, B., & Jordan, I. K. (2015). A combined evidence Bayesian method for human ancestry inference applied to Afro-Colombians. *Gene*, 574(2), 345–351. <https://doi.org/10.1016/j.gene.2015.08.015>
- Rishishwar, L., Conley, A. B., Wigington, C. H., Wang, L., Valderrama-Aguirre, A., & King Jordan, I. (2015). Ancestry, admixture and fitness in Colombian genomes. *Scientific Reports*, 5(12376), 1–16. <https://doi.org/10.1038/srep12376>
- Rivera Franco, N., Braga, Y., Espitia Fajardo, M., & Barreto, G. (2020). Identifying new lineages in the Y chromosome of Colombian Amazon indigenous populations. *American Journal of Physical Anthropology*, 172(2), 165–175. <https://doi.org/10.1002/ajpa.24039>
- Rojas, M. P. (1987). Regionalización de indígenas Choco Datos etnohistóricos, lingüísticos y asentamientos actuales. *Boletín Museo Del Oro*, 18, 46–63.
- Rojas, M. Y., Alonso Morales, L. A., Sarmiento, V. A., Eljach, L. Y., & Usaquén Martínez, W. (2013). Structure analysis of the la Guajira-Colombia population: A genetic, demographic and genealogical overview. *Annals of Human Biology*, 40(2), 119–131. <https://doi.org/10.3109/03014460.2012.748093>
- Rondón, F., César Osorio, J., Viviana Peña, Á., Andrés Garcés, H., & Barreto, G. (2008). Diversidad genética en poblaciones humanas de dos regiones colombianas, 39(2), 52–60.
- Rondón G., F., Oribio, R. F., Braga, Y. A., Cárdenas, H., & Barreto, G. (2006). Estudio de Diversidad

-
- Genética de Cuatro Poblaciones Aisladas del Centro y Suroccidente Colombiano. *Revista de La Universidad Industrial de Santander. Salud*, 38(1), 12–20. Retrieved from <https://www.redalyc.org/pdf/3438/343837061004.pdf>
- Sánchez-Diz, P., Acosta, M. A., Fonseca, D., Fernández, M., Gómez, Y., Jay, M., ... Restrepo, C. M. (2009). Population data on 15 autosomal STRs in a sample from Colombia. *Forensic Science International: Genetics*, 3(3). <https://doi.org/10.1016/j.fsigen.2008.08.002>
- Santos, F., Machado, H., & Silva, S. (2013). Forensic DNA databases in European countries: is size linked to performance? *Life Sciences, Society and Policy*, 9(1), 12. <https://doi.org/10.1186/2195-7819-9-12>
- Senado de la República de Colombia. (n.d.). Proyecto de Ley Estatuario 106 de 2018: Estado de los Proyectos de Ley y Actos Legislativos del Senado. Retrieved July 16, 2020, from <http://leyes.senado.gov.co/proyectos/index.php/textos-radicados-senado/p-ley-2018-2019/1242-proyecto-de-ley-106-de-2018>
- Silva, N. M., Pereira, L., Poloni, E. S., & Currat, M. (2012). Human Neutral Genetic Variation and Forensic STR Data. *PLoS ONE*, 7(11). <https://doi.org/10.1371/journal.pone.0049666>
- Sun, H., Zhou, C., Huang, X., Lin, K., Shi, L., Yu, L., ... Chu, J. (2013). Autosomal STRs Provide Genetic Evidence for the Hypothesis That Tai People Originate from Southern, 8(4). <https://doi.org/10.1371/journal.pone.0060822>
- Thornton, T. A., & Bermejo, J. L. (2014). Local and global ancestry inference and applications to genetic association analysis for admixed Populations. *Genetic Epidemiology*, 38(SUPPL.1). <https://doi.org/10.1002/gepi.21819>
- Tillmar, A. (2010). *Populations and Statistics in Forensic Genetics*.
- Tukey, J. W. (1958). Bias and confidence in not quite large samples. *Annals of Mathematical Statistics*, 29, 614.
- UK National DNA Database. (2020a). National DNA Database documents - GOV.UK. Retrieved July 16, 2020, from <https://www.gov.uk/government/collections/dna-database-documents>
- UK National DNA Database. (2020b). National DNA Database statistics - GOV.UK. Retrieved July 16, 2020, from <https://www.gov.uk/government/statistics/national-dna-database-statistics>
- Urbano, L., Portilla, E. ., Builes, J. J., Gusmão, L., & Sierra-Torres, C. H. (2016). Ancestral Genetic Composition of a human population from the Colombian Southwest using autosomal AIM-InDels. *Journal of Basic and Applied Genetics*, 27(2), 37–48. Retrieved from

http://www.sag.org.ar/sitio/wp-content/uploads/2019/05/V.XXVII_2016_Issue2_30122012.pdf

- Usaquén Martínez, W. (2012). *Validación y consistencia de información en estudios de diversidad genética humana a partir de marcadores microsatélites*. Universidad Nacional de Colombia. Retrieved from <http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:No+Title#0>
- Vargas, C. I., Castillo, A., Gil, A. M., Pico, A. L., & García, O. (2003). *Population genetic data for 13 STR loci in a northeast Colombian (department of Santander) population*. Retrieved from https://www.isfg.org/files/04ab68f72414cb3c6171bdb4a84e4c055c9eee46.02003370_957336538240.pdf
- Wang, S., Lewis, C. M., Jakobsson, M., Ramachandran, S., Ray, N., Bedoya, G., ... Ruiz-Linares, A. (2007). Genetic variation and population structure in Native Americans. *PLoS Genetics*, 3(11), 2049–2067. <https://doi.org/10.1371/journal.pgen.0030185>
- Wurmb-schwark, N. Von, Podruks, E., Schwark, T., Göpel, W., Fimmers, R., & Poetsch, M. (2015). About the power of biostatistics in sibling analysis — comparison of empirical and simulated data, 1201–1209. <https://doi.org/10.1007/s00414-015-1252-9>
- Y., O., I.M., S., & S., A. (1982). A Simple Method For Calculating The Probability of Excluding Paternity with Any Number of Codominant Alleles. *Forensic Science International*, 19, 93–98.
- Yin, C., Deng, C., Qian, X., Huang, H., Yu, Y., Hu, L., ... Chen, F. (2018). The genetic diversity and applicability assessment of autosomal STRs among Chinese populations by a novel Fixation Index and Nei ' s index, 31(January), 49–58. <https://doi.org/10.1016/j.legalmed.2017.12.012>
- Yunis, Juan J, & Yunis, E. J. (2013). Mitochondrial DNA (mtDNA) haplogroups in 1526 unrelated individuals from 11 Departments of Colombia. *Genetics and Molecular Biology*, 36(3), 329–335. <https://doi.org/10.1590/S1415-47572013000300005>
- Yunis, Juan José, Garcia, O., Cuervo, A. G., Guio, E., Pineda, C. R., & Yunis, E. J. (2005). Population data for PowerPlex 16 in thirteen departments and the capital city of Colombia - PubMed. *Journal of Forensic Sciences*, 50(3), 685–702. Retrieved from <https://pubmed.ncbi.nlm.nih.gov/15932109/>

Capítulo 2: Modelo de cálculo para Índices (IP) y probabilidad de paternidad (W) por intervalos de confianza.

2.1. Introducción.

Desde la creación del Instituto Colombiano de Bienestar Familiar en 1968 se han unido esfuerzos de diferentes organizaciones, grupos de estudio e investigadores nacionales por velar por la promoción y la garantía de los derechos de los niños, niñas y adolescentes (Correa Rubio & Sánchez Rodríguez, 2021). Siendo uno de los derechos de los niños, niñas y adolescentes reconocidos como fundamental, el derecho a la identidad en su componente filiación, haciendo parte de uno de los componentes básicos de su personalidad jurídica (Ortega Torres et al., 2015).

Los casos de filiación en Colombia realizados por el ICBF desde 1972 fueron llevados a cabo mediante exámenes antropoheredobiológicos, luego con grupos sanguíneos y hasta dos décadas después se implementaron los marcadores moleculares actuales. En 1995, la acogida de los marcadores STRs en pruebas de filiación; abrió las puertas de esta tecnología a nivel nacional, haciendo necesario establecer valores de referencia de frecuencias alélicas para los marcadores moleculares empleados en estas (Usaquén Martínez, 2012).

Los valores de referencia empleados eran otorgados por las casas comerciales que los calculaban a partir de población estadounidense y europea, desconociéndose las variantes alélicas propias de nuestras poblaciones. Entrado el nuevo milenio, proyectos independientes por parte de la Universidad Nacional de Colombia y del Instituto Nacional de Medicina legal y Ciencias Forenses (Paredes et al., 2003), realizaron los primeros reportes de frecuencias genéticas para la población colombiana con sistemas microsatélites en regiones del país (Usaquén Martínez, 2012).

Posteriormente, otras distribuciones de frecuencias alélicas han sido reportadas para poblaciones específicas Colombianas como las de la **Costa Caribe** (Martínez et al. 2006, 2008, 2017; Rojas et al., 2013), **el noreste** (Castillo et al., 2013; Lucía Hincapié et al., 2009; Ossa Reyes et al., 2009; Vargas et al., 2003), **el occidente** (Franco-Candela & Barreto, 2017; Gómez et al., 2003a, 2003b; Palacio et al., 2006; Porras et al., 2008; Rondón et al., 2008; Rondón G. et al., 2006), **la región del pacífico** (Bravo et al., 2001), **la región central** (Bravo et al., 2001; Burgos et al., 2015; Castillo et al., 2013; Gaviria et al.,

2004; Ibarra et al., 2014; Rey et al., 2003) y *la región amazónica* (Braga et al., 2012; Rivera Franco et al., 2020). Otras publicaciones han reportado las frecuencias alélicas con base en una aproximación a nivel nacional (Benítez-Páez & Reyes, 2003; Durán et al., 2003; Paredes et al., 2003; Sánchez-Diz et al., 2009; Yunis et al., 2013; Yunis & Yunis, 2013)

Tanto en los estudios regionales como en los que poseen una aproximación nacional, se pueden detectar diferencias en las frecuencias alélicas y estimadores genéticos reportados, dado que corresponden a muestreos completamente distintos sobre distintas muestras de población. El **tamaño de muestra** en estos trabajos depende mayoritariamente de los recursos que se cuentan para hacer la investigación o proyecto; cabiendo la duda de si el tamaño empleado es representativo de la población y por ende si los estimadores genéticos obtenidos son confiables o se aproximan al valor del parámetro poblacional (Figura 4A). Otra variable, es la **selección de la muestra**; ya que generalmente se utilizan técnicas de muestreos por norma o conveniencia en donde se emplean distintos mecanismos para facilitar la obtención: emplear casos relacionados con la casuística del laboratorio que realiza el reporte, realizar muestreos en hospitales o en ciudades principales sin contar con una herramienta de recolección de datos demográficos o genealógicos y usar como criterio de inclusión el lugar de nacimiento sin tener en cuenta migraciones de generaciones anteriores (Figura 4B). Estas dos variables traen como resultado que no sea posible la estimación del error muestral y no permiten evaluar si la variación de los estimadores poblacionales es debida a variaciones evolutivas y/o muestreos deficientes (Figura 4B). (Usaquén Martínez, 2012).

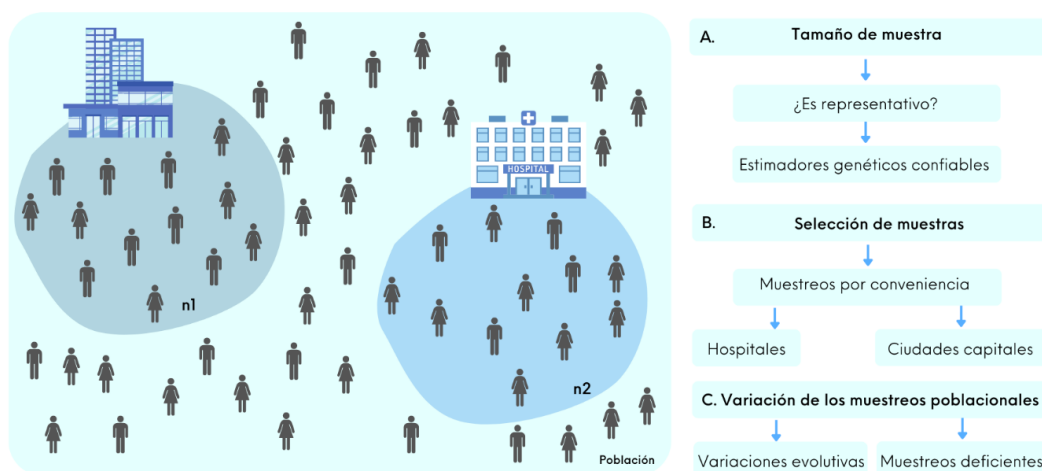


Figura 4 Variables asociadas al tipo de muestreo y selección de muestras en estudios genético poblacionales.

Actualmente, los laboratorios que realizan pruebas de filiación y paternidad en el país emplean la población de referencia que según criterio propio varía entre cada uno de estos. A Julio de 2022, son nueve los laboratorios acreditados y avalados por el Organismo Nacional de Acreditación de Colombia (ONAC) (Instituto Colombiano de Bienestar Familiar (ICBF), 2022) y según la gestión de estadísticas judiciales de la Corte Suprema de Justicia, en el año 2019 se presentaron 2.754 demandas de investigación de paternidad, 2.753 para el año 2020, y 2.589 para el 2021. Esta tipología de demanda se posiciona entre los quince tipos de demanda más usada en los últimos 5 años (Correa Rubio & Sánchez Rodríguez, 2021). Sin embargo, el número de demandas es mucho menor al número de solicitudes de realización de pruebas de paternidad al ICBF realizadas por el Laboratorio de Genética del Instituto Nacional de Medicina Legal y Ciencias Forenses; en donde se estima que al mes se realizan en promedio 545 casos mensuales, para un aproximado de 6540 casos anuales (Ortega Torres et al., 2015). Y esto es sólo estimando el número de casos de un laboratorio acreditado de los nueve que hay en el país, si se estima los casos realizados por este laboratorio en los últimos 10 años; se llegaría al orden de al menos 65.400 casos de pruebas de paternidades a nivel nacional.

En los últimos 10 años, este número creciente de paternidades en disputa en Colombia ha involucrado al menos a 65.400 padres de diferentes orígenes étnicos y ancestralidades. Se examinó qué tan sensibles son el Índice de Paternidad (PI) y la Probabilidad de Paternidad (W) a la selección de la base de datos STR de la población con distintas ancestralidades (*'Database Effect'*) (Moroni et al., 2011) y cómo podría expresarse el índice de paternidad y probabilidad de paternidad de un mismo caso como un intervalo de confianza; teniendo en cuenta la varianza generada en estos estimadores dada las distintas poblaciones de referencia empleadas. Para evaluar esto, realizamos pruebas de paternidad en 1797 casos tríos colombianos con un conjunto de marcadores STR de uso común y 5 poblaciones de referencia con ancestralidades distintas; en donde se varió el tamaño muestral por región (para cada población se evaluaron tres tamaños muestrales y por tamaño muestral se realizaron 4 réplicas; más detallado en materiales y métodos). Estas 5 poblaciones de referencia representan 60 diferentes distribuciones de frecuencias de distribución alélica y se estiman para cada población específica variando entre poblaciones. Como consecuencia, calcular probabilidades usando diferentes datos de población significa diferencias en los valores de las probabilidades y por ende permite poder expresar el resultado de una prueba de paternidad como un intervalo de confianza de IP y W.

2.2. Materiales y métodos.

2.2.1. Poblaciones de referencia.

Se evaluaron 5 poblaciones de referencias con distintas ancestralidades basadas en el estudio previo del Grupo de Genética e Identificación de Poblaciones (GPI) del Instituto de Genética de la Universidad Nacional de Colombia en el que se colectaron muestras en todo el país entre los años 2008 y 2017 (Mogollon Olivares et al., 2020). Cada muestra cuenta con su consentimiento informado y un análisis genealógico obtenido a través de una encuesta dirigida a evaluar variables de tiempo de residencia, genealogía, demografía e información sociocultural de todos los participantes. Un estudio histórico previo de las poblaciones muestreadas y un análisis exhaustivo de las variables permitió realizar una clasificación no basada en lugares de nacimiento, sino en genealogía y demografía, lo que la hizo más rigurosa y concordante con las categorías de ascendencia. Teniendo en cuenta los resultados de este estudio, en donde mediante marcadores informativos de ancestría InDels en adición a un análisis histórico y demográfico, se determinaron diferentes grupos poblacionales con ancestralidades contrastantes y se tomaron las tipificaciones de marcadores autosómicos STR almacenadas en la base de datos del GPI para recrear las cinco poblaciones de referencia. Su distribución geográfica es representada en la Figura 5.

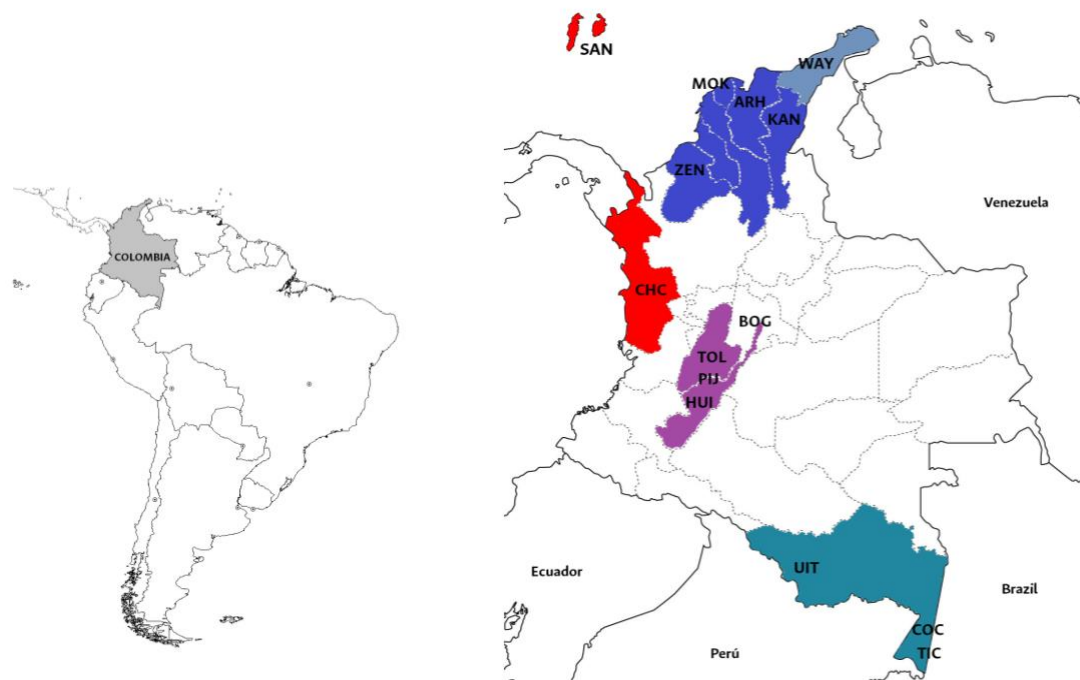


Figura 5 Distribución geográfica de las poblaciones de referencia empleadas en las pruebas de paternidad. Región R1 San Andrés (SAN) se representa en rojo, Región R2 Wayuu (WAY) en gris, Región R3 Amazonas (UIT: Uitoto, COC: Cocama, TIC: Ticuna) en azul claro, Región R4 Bogotá (BOG) en púrpura y Región R5 Caribe en azul oscuro (ZEN: Zenu, MOK: Mokana, ARH: Arhuaco, KAN: Kankuamo). Tomado de Mogollon Olivares et al., 2020).

Región 1: R1 San Andrés.

Respecto al aporte genético africano, se tomó a la población de San Andrés y Providencia, ya que históricamente constituyó un paso importante en la ruta esclavista de África, que luego se convirtió en un paraíso para el contrabando y los piratas (Parsons, 1985). Este territorio ha sido poblado por diferentes grupos culturales como los Miskitos una comunidad amerindia de Centro América, colonos europeos y esclavos africanos (Meisel, 2005; Vollmer, 1997). Debido a las disputas constantes entre los ingleses y españoles por el control de las islas y el poder comercial que representaban, en 1803 estas pasaron a manos del Nuevo Reino de Granada y en 1822, mediante la Constitución de Cúcuta las islas se adhirieron de manera oficial a la nación Colombiana (Parsons, 1985). Estas dinámicas dieron paso al surgimiento de la población raizal, que es el resultado del proceso mestizaje de estos grupos poblacionales (Alonso & Usaqué, 2012). Se denominó **Región 1 (R1)** a la población de San Andrés y Providencia a los 46 individuos que mediante genealogía y marcadores informativos de ancestría (Indels - AIMS) tuvieron un porcentaje de ancestralidad africana de al menos el 62.7% y que se autodeterminaron como *raizales*.

Región 2: R2 Wayúu.

Al extremo norte de Colombia y Suramérica, en la Llanura Caribe se encuentra la península de La Guajira, territorio donde se estableció la comunidad Wayuu perteneciente a la familia lingüística Arawak que migró desde la región del Río Amazonas – Río Negro donde se encuentra la actual ciudad de Manaos, hasta la costa occidental de Venezuela y la región más septentrional del Caribe Colombiano (Oliver, 1990). Fueron cazadores, recolectores y pescadores que comerciaron con perlas y con la sal de Manaure y poseían una organización sociopolítica de castas cuyos símbolos fueron animales, llamando la atención sus elaborados ritos funerarios (CINEP, 1998b). Según el Censo Nacional de Población DANE (DANE, 2007), la comunidad Wayuu representa el 44,9% de la población en esta región y son el mayor grupo amerindio en Colombia (20,5% de la población indígena nacional). A pesar de que esta población ha definido una identidad cultural, existen múltiples factores que han llevado a la convivencia entre los guajiros y migrantes que comparten el mismo territorio (Rojas et al., 2013). Se denominó **Región 2 (R2)** a la población Wayuu con 94 individuos que

mediante genealogía y marcadores informativos de ancestría (Indels - AIMS) tuvieron un porcentaje de ancestralidad nativoamericana de al menos el 93.2% y que se autodeterminaron como *wayuu*.

Región 3: R3 Amazonas.

Al extremo sur de Colombia, se encuentra la región biogeográfica de la Amazonía que posee altas proporciones ancestrales nativoamericanas y limita con Ecuador, Perú, Brasil. En el departamento del Amazonas, el 43,4% de la población residente se autodetermina como indígena (DANE, 2007) perteneciente a alguna de las comunidades de la región, siendo las más numerosas la Ticuna, Uitoto y Cocama. Estas comunidades tienen en común la cultura de selva tropical, la residencia en casas plurifamiliares llamadas malocas y mantienen una economía de roza y quema itinerante, además de la caza y la pesca (Morcote Ríos et al., 2006). Sin embargo, existen también diferencias en la forma de las malocas, los tipos de organización social interna y los rituales, por ejemplo, los Huitoto aún habitan en malocas de forma octogonal o circular, mientras que los Ticuna abandonaron la vida de maloca y adoptaron el patrón de asentamiento actual a raíz de la dispersión y redistribución de la población ocasionada por las sucesivas oleadas de explotación industrial de los recursos amazónicos y de la mano de obra indígena para la producción de quina, madera y caucho (Mincultura, 2005; Ministerio de Asuntos Exteriores y de Cooperación & Oficina de Información Diplomática del Departamento de Relaciones Exteriores, 2017). Cabe también mencionar que, en la Amazonía colombiana, habitan también poblaciones de ancestrías mixtas o mezcladas, colonos del interior del país y de los países vecinos, principalmente de Brasil, Perú, Venezuela y Ecuador (Moreno Bandeira, 2018). Se denominó ***Región 3 (R3)*** a la población indígena amazónica con 59 individuos que mediante genealogía y marcadores informativos de ancestría (Indels - AIMS) tuvieron un porcentaje de ancestralidad nativoamericana de al menos un 93.5% y que se autodeterminaron como *Cocama, Uitotos o Ticunas*.

Región 4: R4 Bogotá.

Como población de referencia de ancestría mezclada o múltiple, se consideró a Bogotá como una población en donde han confluído múltiples ancestrías, a pesar de que sus primeros pobladores fueron los Muisca, pertenecientes a la familia lingüística Chibcha y aun cuando al arribo de los

conquistadores se calcula que había medio millón de indígenas de este grupo (CINEP, 1998a). Aunque Bogotá careció de un flujo importante de inmigrantes extranjeros, según los censos llevados a cabo en el siglo XIX, la población tuvo un crecimiento regular; en 1832 tenía 36.465 habitantes; en 1881, 84.723 habitantes y hacia finales de siglo casi 100.000. Sin embargo, el establecimiento de Bogotá como la capital del país trajo consigo un crecimiento poblacional como consecuencia de un incremento en la oferta laboral en una gran variedad de campos e industrias, lo que dio como consecuencia una considerable ampliación física de la ciudad (Sitio Oficial Bogotá, 2019) y un alto flujo migratorio individuos de todas las regiones del país. Debido a que para esta población no se contaba con registros históricos o información demográfica detallada en tres generaciones, se tomaron como población de referencia **Región 4 (R4)** a 50 personas nacidas en Bogotá (BOG) de nuestra base de datos de casos de paternidad que presentaron los índices de paternidad más altos.

Región 5: R5 Caribe.

En la región de la Llanura del Caribe, también se asentaron comunidades nativoamericanas como los Arhuacos, Kankuamos, Mokane y Zenú. Los Arhuacos y Kankuamos, distribuidos en las laderas sur orientales de la Sierra Nevada de Santa Marta, se alimentaban de caracoles, conchas, pescados, maíz y yuca (CINEP, 1998b). Estos últimos, sufrieron un proceso de colonización y aculturación más agresivo que las otras comunidades de la región y debido a que los últimos hablantes de esta lengua fallecieron hacia 1960, se han realizado esfuerzos para recuperar la lengua y la identidad cultural (Talco Arias, 1994).

Los Mokane, ubicados en la zona costera del departamento del Atlántico se destacaban por sus habilidades para la navegación, la caza, la domesticación de la abeja para la producción de miel y su trabajo con la construcción de terrazas artificiales que les permitió evitar la erosión y conservar la humedad en los suelos (CINEP, 1998b). Actualmente, los Mokane están en proceso de recuperación de sus tradiciones y se realizan esfuerzos académicos para evitar su total extinción. Por último, el pueblo amerindio Zenú, distribuido en los departamentos de Córdoba y Sucre explotaron eficazmente los fértiles suelos de la Llanura Caribe, llegando a construir extensos camellones de cultivo en la zona del río San Jorge. Los Zenú también trabajaron el oro, la alfarería y los tejidos hasta

alcanzar un desarrollo técnico y artístico de gran calidad (CINEP, 1998b; Jaramillo & Turbay Ceballos, 2000).

Se denominó como **Región 5 (R5)** a la población indígena del caribe (n=57) Mokaná, Arhuacos, Kankuamos y Zenú así como a los individuos con padres, madres y abuelos de la región muestreada con 124 individuos que mediante genealogía y marcadores informativos de ancestría (Indels - AIMS) tuvieron un porcentaje de ancestralidad nativoamericana de al menos el 51.82% y que se autodeterminaron como *Mokaná, Arhuacos, Kankuamos y Zenú*.

2.2.2. Casos analizados.

Se utilizó la base de datos del laboratorio de Identificación Humana del Instituto de Genética de la Universidad Nacional de Colombia con 17766 personas que pertenecían a 6385 casos. Al hacer la depuración de los casos de paternidad en donde se contara con el consentimiento informado, la información de lugar de nacimiento, la información genética completa para al menos diez marcadores y que el caso que se analizara fuera un caso trío se obtuvo un total de 1797 casos de paternidad trío, con un total de 5391 personas incluidas en el experimento.

2.2.3. Implementación de los métodos de aleatorización de Monte Carlo para la selección de Mn muestras, cálculos de frecuencias alélicas, IP y W por muestreo.

2.2.3.1. Selección de Mn muestras por región.

Los métodos de aleatorización de Monte Carlo pueden ser utilizados en los casos en los que se desee generar nuevas muestras de datos mediante técnicas de muestreo (con o sin reposición, manteniendo constante el tamaño muestral o no), a partir de los datos de la muestra original (Efron, 1979, 1982; Losilla Vidal, 1994; Manly, 1991b, 1991a). Empleando el algoritmo secuencial, de la aplicación desarrollada en Access **Aplicación para cálculo de tamaños de muestra** (Usaquén Martínez, 2012), se utilizó el método de aleatorización de Monte Carlo *bootstrap* en el que la generación

aleatoria de nuevos conjuntos de datos mediante un muestreo aleatorio se hace con reposición de los datos de la muestra original, manteniendo el tamaño muestral (Efron, 1979, 1982).

A. Generación de tres submuestras ($S1, S2, S3$) de la región R con diferentes tamaños muestrales $n, n/2, n/4$. Para cada una de las regiones anteriormente descritas desde $R1$ a $R5$, se generaron tres submuestras $S1, S2$ y $S3$ (Figura 1.A) con una semilla $n = 5$. La submuestra $S1$ corresponde al tamaño muestral $\frac{n}{4}$, $S2$ a $\frac{n}{2}$ y finalmente $S3$ a n . Para cada submuestra $S1, S2$ y $S3$, se tomaron cuatro replicas. Teniendo en cuenta que son cinco regiones que se emplearon como poblaciones de referencias, y fueron tres submuestras para cada región con sus cuatro replicas, en total serían 60 poblaciones, desde $R1S1r1$ a $R5S3r4$.

B. Cálculo de distribución de frecuencias alélicas para cada replica de cada submuestra. Después de haber simulado poblaciones de referencia de tamaños $n, n/2$ y $n/4$ para cada submuestra desde $R1S1r1$ a $R5S3r4$ a partir de una semilla de las regiones $R1$ a $R5$ (San Andrés, Wayuu, Amazonas, Bogotá, Caribe), se realizó la distribución de las frecuencias alélicas para cada una de las nuevas submuestras y sus réplicas (Figura 1. B) (Tabla S1). Teniendo en cuenta que las distribuciones de frecuencias alélicas para poblaciones más cerradas o conservadas como Wayuu, Amazonas y San Andrés podían variar respecto a poblaciones más mezcladas como Bogotá, se realizó el cálculo de frecuencias alélicas mínimas para cada uno de los marcadores microsatélites empleados de acuerdo con (Budowle et al., 1996)

$$p_{min} = 1 - [1 - (1 - \alpha)^{\frac{1}{c}}]^{\frac{1}{2n}}$$

En donde p_{min} es la frecuencia alélica mínima, c es el número de alelos comunes que pueden ser estimados a partir del nivel de heterocigosidad o con una frecuencia mayor de 0.01 (Chakraborty R, 1981; Neel JV, 1973; Nei M, 1975), y n es el número de individuos del set de datos estudiado. Para este estudio siguiendo las recomendaciones del autor citado, α fue fijado como 0.05. Los cálculos de frecuencias alélicas mínimas y medidas de diversidad genética de las 60 submuestras se muestran en la tabla S2.



Región (R)

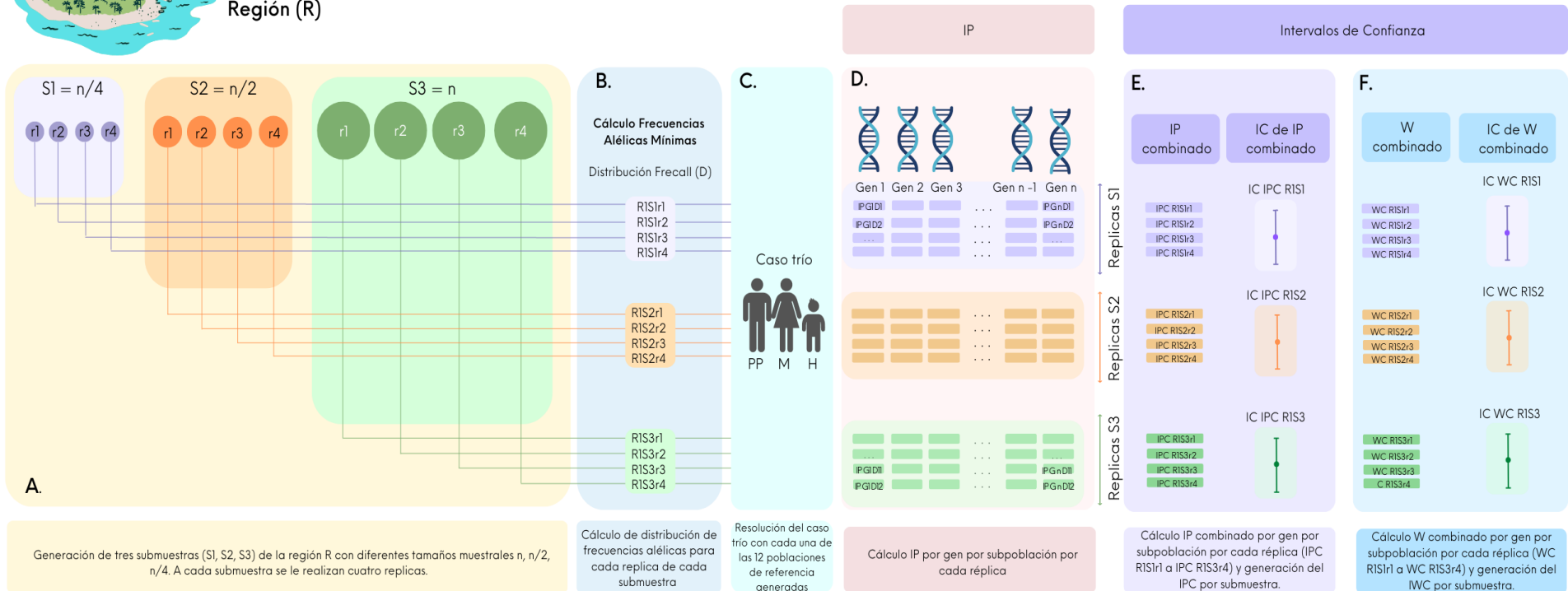


Figura 6 Diagrama del experimento realizado por cada una de las regiones desde R1 a R5. **A. Generación de tres submuestras (S_1 , S_2 , S_3) de la región R con diferentes tamaños muestrales n , $n/2$, $n/4$: A cada submuestra se le realizan cuatro replicas.** **B. Cálculo de distribución de frecuencias alélicas para cada réplica de cada submuestra:** Generación de tablas de distribución de frecuencia alélica incluyendo las frecuencias alélicas mínimas para cada réplica de cada submuestra. **C. Resolución del caso trío con cada una de las 12 poblaciones de referencia generadas en cada región:** Un caso en particular es resuelto con las frecuencias alélicas de las 12 poblaciones de referencia generadas. **D. Cálculo IP por gen por subpoblación por cada réplica:** Cálculo de índices de paternidad para cada uno de los genes evaluados en el caso trío. **E. Cálculo IP combinado por gen por subpoblación por cada réplica (IPC $R1S1r1$ a $IPC R1S3r4$) y generación del IPC por submuestra:** Desde $R1S1r1$ a $R1S2r4$ se calculan los IPC. **F. Cálculo W combinado por gen por subpoblación por cada réplica (WC $R1S1r1$ a $WC R1S3r4$) y generación del IWC por submuestra.**

-
- C. Resolución del caso trío con cada una de las 12 poblaciones de referencia generadas en cada región.** Como se muestra en la Figura 1, en la sección C, se toma un caso trío con un presunto padre, hijo y madre que es evaluado bajo las frecuencias alélicas desde la región 1 con las poblaciones de referencia R1S1r1, R1S1r2, R1S1r3, R1S1r4, R1S2r1, R1S2r2, R1S2r3, R1S2r4, R1S3r1, R1S2r2, R1S3r3 y R1S3r4 hasta la región 5 con las poblaciones de referencia R5S1r1, R5S1r2, R5S1r3, R5S1r4, R5S2r1, R5S2r2, R5S2r3, R5S2r4, R5S3r1, R5S2r2, R5S3r3 y R5S3r4.
- D. Cálculo IP por gen por subpoblación por cada réplica.** Para cada una de las réplicas de poblaciones de referencias que corresponden a submuestreos con distintos tamaños muestrales, se realizó el cálculo de índices de paternidad para cada uno de los genes evaluados en el caso trío (Figura 1. D).
- E. Cálculo IP combinado por gen por subpoblación por cada réplica (IPC R1S1r1 a IPC R1S3r4) y generación del IPC por submuestra.** Se calcularon los índices de paternidad combinados (IPC) para cada una de las regiones desde R1 a R5 (Figura 1.E). Posterior a estos se generaron 3 intervalos de confianza teniendo en cuenta el tamaño muestral de cada una de las submuestras S1, S2 y S3. De esta forma con las cuatro réplicas (r1, r2, r3, r4) de cada submuestra S1, S2 y S3 se calcularon los intervalos de confianza (IC1, IC2 y IC3) de los índices de paternidad combinados para un mismo caso.
- F. Cálculo W combinado por gen por subpoblación por cada réplica (WC R1S1r1 a WC R1S3r4) y generación del IWC por submuestra.** Se calcularon las probabilidades de paternidad combinadas (WC) para cada una de las regiones desde R1 a R5 (Figura 1.F) Posterior a estos se generaron 3 intervalos de confianza teniendo en cuenta el tamaño muestral de cada una de las submuestras S1, S2 y S3. De esta forma con las cuatro réplicas (r1, r2, r3, r4) de cada submuestra S1, S2 y S3 se calcularon los intervalos de confianza (IC1, IC2 y IC3) de los índices de paternidad combinados para un mismo caso.

2.2.3. Representación de resultados.

Para representar gráficamente la variabilidad de los resultados se realizó un análisis de componentes principales (PCA) utilizando datos de índices de paternidad y probabilidades de paternidad por cada caso y marcador mediante el software MultiVariate Statistical Package (MVSP), versión 3.22 (Kovach, 2007).

2.3. Resultados.

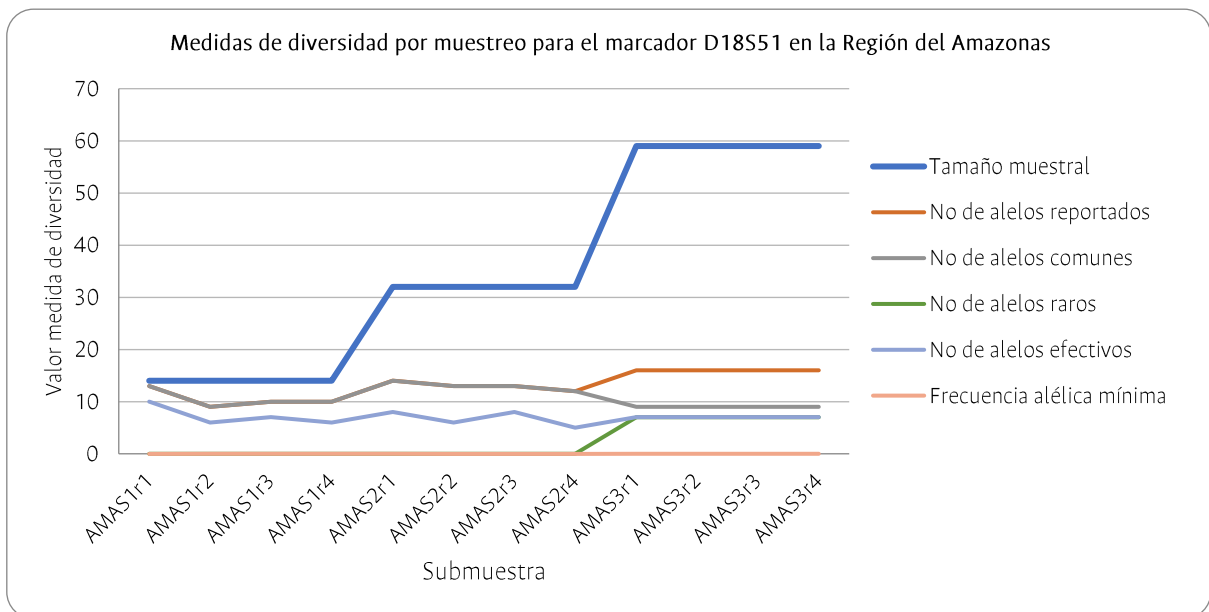
2.3.1. Análisis de diversidad y genética descriptiva.

Sesenta simulaciones independientes fueron realizadas para generar sesenta poblaciones de referencia correspondientes a cinco regiones colombianas, a partir de tres submuestros con tamaños $n/4$, $n/2$ y n . Para cada uno de los submuestros S1, S2, y S3 se tomaron cuatro replicas $r1$, $r2$, $r3$ y $r4$ (Figura 6). A partir de estas poblaciones de referencia simuladas teniendo como base las cinco regiones descritas en la metodología, se calcularon las 60 distribuciones de frecuencias alélicas desde la población R1S1r1 a la R5S3r4 (Tabla S1).

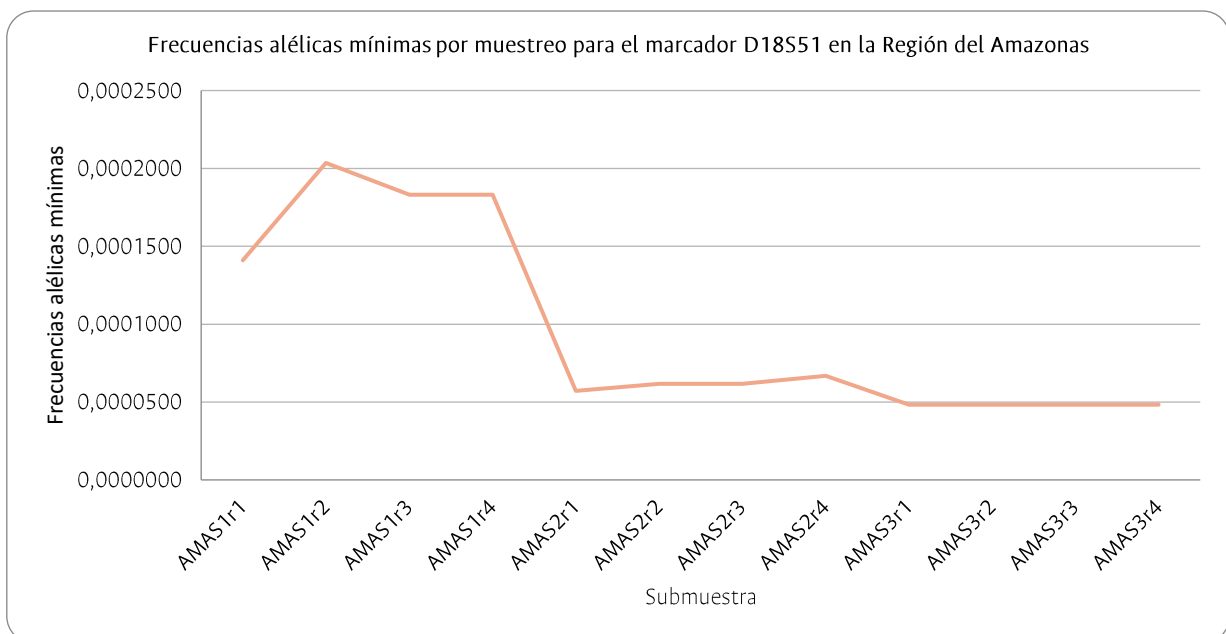
También, se realizó el cálculo de frecuencias alélicas mínimas para cada uno de los marcadores microsatélites empleados de acuerdo con Budowle et al., 1996. Los cálculos de frecuencias alélicas mínimas y medidas de diversidad genética de las 60 simulaciones se muestran en la Tabla S2. El número de alelos reportados, número de alelos comunes, número de alelos raros, número de alelos efectivos y frecuencia alélica mínima fueron calculados para cada una de las simulaciones de poblaciones de referencia.

A medida que se aumentó el tamaño muestral desde AMAS1r1 a AMAS1r4 ($n = n/4$), AMAS2r1 a AMAS2r4 ($n = n/2$) y AMAS3r1 a AMAS3r4 (n) (Gráfica 1), se observa que a mayor tamaño muestral los alelos encontrados en la población aumentan, así como el número de alelos raros. También el número de alelos efectivos aumentan y al llegar al mayor tamaño muestral este valor se mantiene constante; por otro lado, el número de alelos comunes disminuye y se vuelve constante. En la gráfica 2. se observa con mayor detalle, cómo se afectan las frecuencias alélicas mínimas al aumentar el tamaño muestral; a mayor tamaño muestral la frecuencia alélica mínima disminuye y se mantiene

constante. Esto es algo que no sólo se evidenció en la población de Amazonas sino en las cuatro restantes.



Gráfica 1 Medidas de diversidad genética para cada uno de los muestreos simulados realizados en la Región del Amazonas desde AMAS1r1 a AMAS1r4 ($n = n/4$), AMAS2r1 a AMAS2r4 ($n = n/2$) y AMAS3r1 a AMAS3r4 (n). Tamaño muestral, número de alelos reportados, número de alelos comunes, número de alelos raros, número de alelos efectivos y frecuencia alélica mínima.



Gráfica 2 Frecuencias alélicas mínimas para cada uno de los muestreos simulados realizados en la Región del Amazonas desde AMAS1r1 a AMAS1r4 ($n = n/4$), AMAS2r1 a AMAS2r4 ($n = n/2$) y AMAS3r1 a AMAS3r4 (n).

2.3.2. Índices de Paternidad por región, submuestra y replica.

Para cada una de las réplicas de poblaciones de referencias que corresponden a submuestreos con distintos tamaños muestrales, se realizó el cálculo de índices de paternidad para cada uno de los genes evaluados en los 1797 casos trío (Figura 3 - 6). Se representó gráficamente mediante análisis de componentes principales la variabilidad del resultado del IP por marcador de estos 1797 casos bajo las condiciones de diferentes tamaños n para las poblaciones de referencia empleadas.

2.3.2.1. Región 1: San Andrés y Providencia.

En la primera región se evidenció que muchos casos se agregaron en los mismos puntos (Gráfica 3), se esperaba que los datos fueran representados de forma dispersa en los cuatro cuadrantes, pero al contrario se observan pocas concentraciones donde los 1797 casos se encuentran representados casi que en la misma posición. Esto es explicado bajo la condición de la implementación de cálculos para frecuencias mínimas alélicas; ya que para un alelo que no se encontrara en la población de referencia pero sí en los perfiles de los casos, para un mismo marcador la frecuencia alélica mínima sería la misma para todos los alelos faltantes.

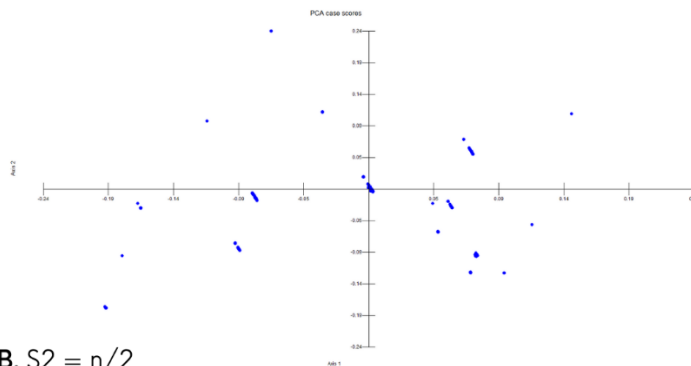
En la tabla S1, para R1S1r1 antes de agregar las frecuencias alélicas mínimas en el marcador D16S539 se habían encontrado 7 alelos. Para completar la distribución con el fin de correr los casos se agregaron 9 alelos con la frecuencia alélica mínima de 0,00006784, es decir, más de la mitad de las frecuencias corresponden a un valor constante empleado para hacer los cálculos de índices de paternidad en 1797 casos. Lo mismo es observado en marcador D2S1358, donde la distribución inicial reportaba 7 alelos y la final contando agregados con frecuencias alélicas mínimas dio una distribución total de 16 alelos, es decir más del doble de alelos que presentan frecuencias alélicas mínimas que alelos reportados en la distribución inicial. Ya que los alelos faltantes son agregados con esta frecuencia.

Además, cabe resaltar, que los cálculos de frecuencias alélicas mínimas dependen del tamaño muestral, que es constante para todos los marcadores en cada submuestra S1, S2 y S3. El valor en el cálculo que cambia por marcador es c o el número de alelos comunes, que ya depende de cada marcador, sin embargo, al observar la frecuencia alélica mínima en cada marcador, encontrados valores muy similares, por ejemplo 0,0000678 y 0,0000668 para los marcadores D16S539 y D2S1358, respectivamente. Para D18S51, D19S433 y D21S11 ($p_{min} = 0,0000338, 0,0000324$ y $0,0000372$;

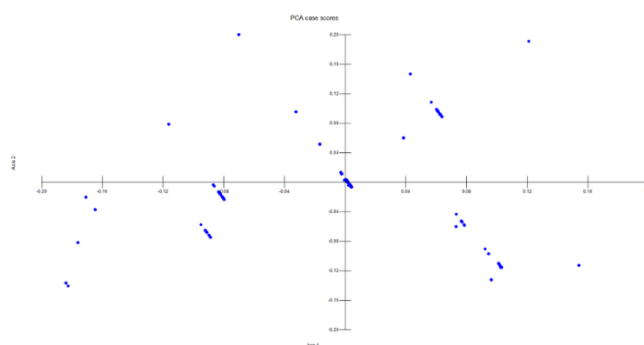
respectivamente); D2S1338 ($p_{min}=0,0000432$) y FGA ($p_{min}=0,0000413$); y finalmente D8S1179 ($p_{min}=0,0000516$) y VWA

Índice de Paternidad - R1: San Andrés

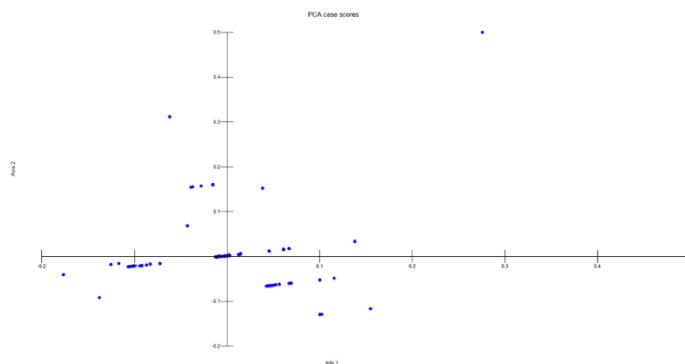
A. $S1 = n/4$



B. $S2 = n/2$



C. $S3 = n$



Gráfica 3 Gráfico de dispersión del PCA construido a partir de los índices de paternidad con los sistemas D16S539, D18S51, D19S433, D21S11, D2S1338, D3S1358, D8S1179, FGA, TH01, VWA calculados para 1797 casos tríos analizados en este estudio. Población de referencia: Región R1 San Andrés con diferentes tamaños muestrales: A. $n/4$, B. $n/4$, y C. n .

($p_{min}=0,0000572$). Esta variable, aunque es diferente para cada marcador, muestra que hay marcadores que tiene p_{min} muy parecidas, lo que nuevamente juega un papel importante en el agrupamiento de los casos en componentes principales en los mismos puntos, esto dado a que hay muchos IP para cada marcador que se van a repetir a lo largo de los casos ensayados.

Ahora, al observar qué ocurre con los IP, a lo largo que se aumenta el tamaño n de la población de referencia, evidenciamos que de S1 a S2 la gráfica de PCA disminuye un poco en dimensión (magnitud en la dispersión de los ejes x y y), para S2 a S3 vemos que los datos se encuentran más agregados.

2.3.2.2. Región 2: Wayuu.

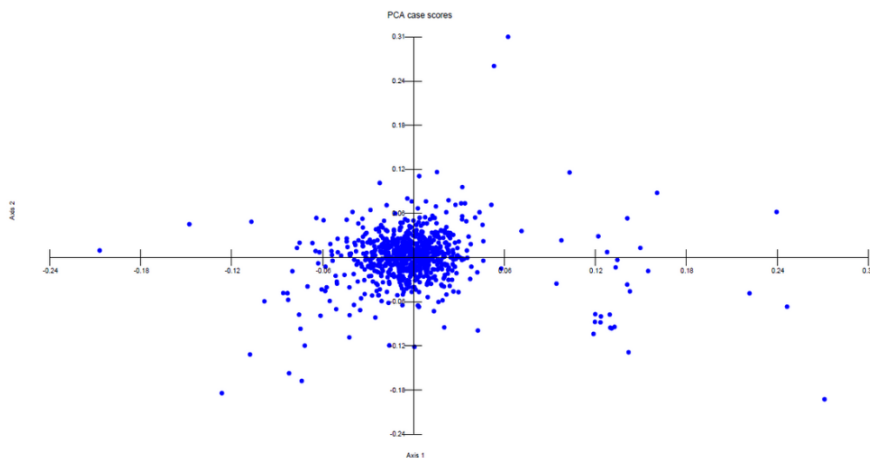
Para la población de referencia Wayuu, también fue necesario completar la distribución de frecuencias alélicas para los diez marcadores analizados mediante el cálculo de frecuencia alélica mínima descrita por Budowle et al., 1996 (Tabla S2). Cuando los índices de paternidad son calculados en poblaciones de referencia más conservadas o con distribuciones de frecuencia restringidas como esta, se evidencia que, debido al amplio uso de frecuencias alélicas mínimas, los 1797 casos se agrupan en puntos poco diferenciables o más bien una gran cantidad de casos, no logra diferenciarse en grupos de inercia distantes (Gráfica 4.B).

Este agrupamiento, nuevamente puede explicarse no solamente por el tamaño muestral constante en cada subpoblación, sino por las frecuencias alélicas mínimas similares entre marcadores, por ejemplo, para R2S1r1, D16S539 ($p_{min}= 0,0000737$) y VWA ($p_{min}= 0,0000777$) tienen frecuencias alélicas mínimas similares, así como D18S51 ($p_{min}= 0,0000452$), D19S433 ($p_{min}= 0,0000486$), D21S11 ($p_{min}= 0,0000401$), D2S1338 ($p_{min}= 0,0000401$).

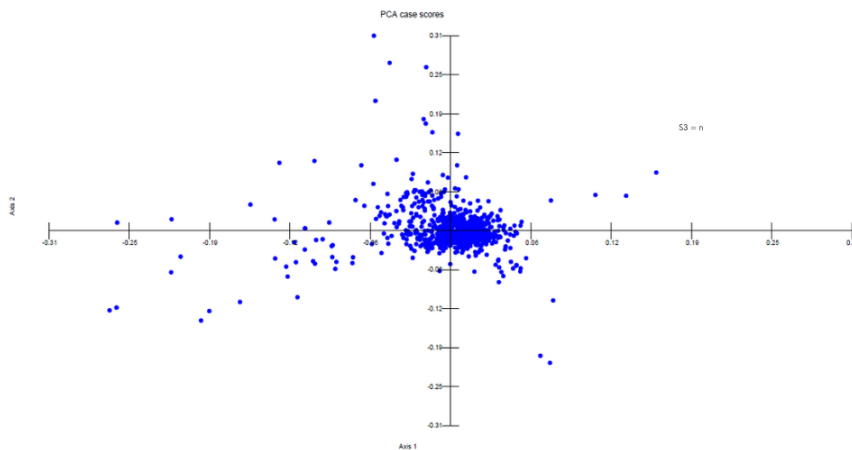
Al aumentar el tamaño de muestra de las llamadas poblaciones de referencia Wayuu desde S1 (Gráfica 4A) a S2 (Gráfica 4B), $n/4$ y $n/2$, respectivamente, observamos que a pesar de haber aumentado el tamaño muestral al doble no hay una agrupamiento drástico; sin embargo al aumentar el tamaño muestral a n desde S2 (Gráfica 4B) a S3 (Gráfica 4C) se observan núcleos de inercia más consolidados y densos. esto es evidente desde S3 a S2, y aún más marcado en S3 a S1.

Índice de Paternidad - R2: Wayuu

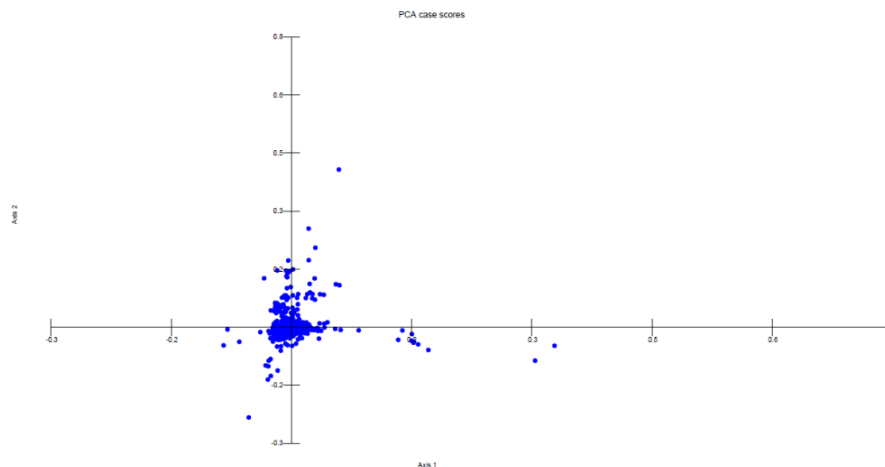
A. $S1 = n/4$



B. $S2 = n/2$



C. $S3 = n/3$



Gráfica 4 Gráfico de dispersión del PCA construido a partir de los índices de paternidad con los sistemas D16S539, D18S51, D19S433, D21S11, D2S1338, D3S1358, D8S1179, FGA, TH01, VWA calculados para 1797 casos tríos analizados en este estudio. Población de referencia: Región R2 Wayuu con diferentes tamaños muestrales: A. $n/4$, B. $n/4$, y C. n .

2.3.2.3. Región 3: Amazonas.

El resultado de análisis de componentes principales de la población del Amazonas (Gráfica 5) llamó la atención porque resultó distinto a lo esperado. Una población conservada con un componente ancestral nativoamericano mayor al 93.5% y cuyos integrantes se autodeterminaron como *Cocama*, *Uitotos* o *Ticunas* se espera que al ser empleada como población de referencia los índices de paternidad calculados con sus frecuencias resultasen con poca dispersión.

La población del Amazonas muestra una mayor dispersión de los casos en los cuadrantes (respecto a las otras cuatro regiones); resultado que no se esperaría en una población que consideramos más conservada o con distribuciones de frecuencias alélicas restringidas. Al observar, la gráfica de PCA con un tamaño muestral $n/4$ (Gráfica 5A); se evidencia una amplia dispersión de los resultados. Eso sí, se evidencian pocos puntos de los totales 1797 casos analizados.

También es de resaltar la presencia de mayor número de casos atípicos o casos no agrupados en ningún punto de inercia respecto a las demás poblaciones de referencia empleadas, específicamente en los cuadrantes I y IV (Gráfico 5A.). Al aumentar el tamaño muestral de $n/4$ a $n/2$, se observa una mayor dispersión de los datos en los cuadrantes III y IV, sin embargo, los datos se encuentran concentrados en más puntos por lo que se ve menos datos que en la población S1 con $n/4$. En el cuadrante IV se observan datos atípicos o muy alejados del resto de casos.

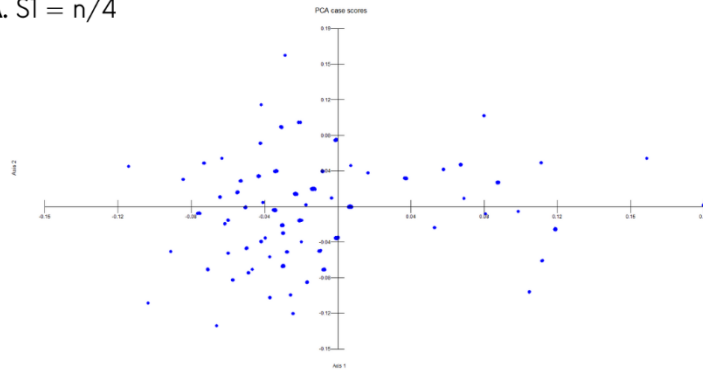
Aquí se hace necesario tener en cuenta que esos puntos apartados del resto corresponden a casos que para esta población presentaron índices de paternidad muy elevados; siendo estos IP de dos o tres órdenes de magnitud mayores para un marcador respecto a los demás. Lo que refuerza la idea que es una población de referencia con presencia de alelos en baja frecuencia que, empleadas en casos trío de paternidad pertenecientes a poblaciones mezcladas como Bogotá, aumentan los IP.

Al emplear el tamaño muestral n (Gráfica 5C), se evidencia que hay una concentración de puntos aún mayor que en tamaños muestrales menores ($n/4$ y $n/2$) por lo que se observan menos puntos en el plano, mostrando que los casos se agregan en un mismo punto dado que los IP para cada marcador debido al uso de las frecuencias alélicas mínimas no sólo correspondieron al mismo valor para el marcador, sino que también fueron muy similares entre marcadores. Las frecuencias alélicas mínimas fueron similares en los marcadores: D19S433 ($p_{min}= 0,0000221$), D21S11 ($p_{min}= 0,0000214$), FGA ($p_{min}= 0,0000267$) y TH01 ($p_{min}= 0,0000267$). Y, por último, D16S539 ($p_{min}= 0,0000777$) y VWA

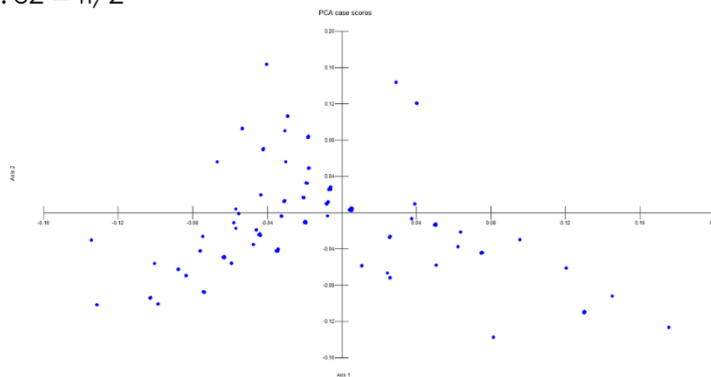
($p_{min} = 0,0000792$). En el caso de D21S11, se empleó la frecuencia mínima sólo en dos alelos de 17 alelos en total.

Índice de Paternidad - R3: Amazonas

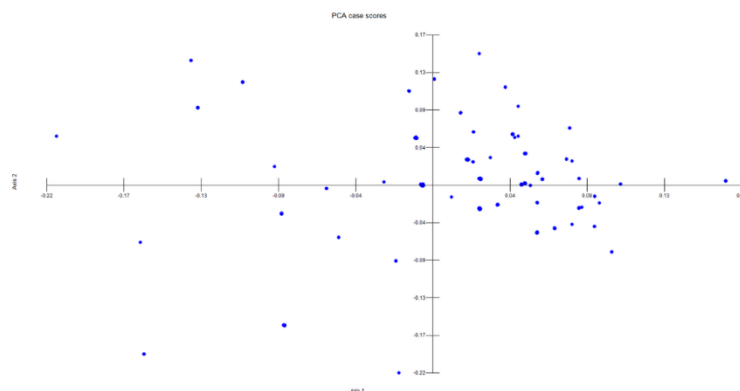
A. $S1 = n/4$



B. $S2 = n/2$



C. $S3 = n$



Gráfica 5 Gráfico de dispersión del PCA construido a partir de los índices de paternidad con los sistemas D16S539, D18S51, D19S433, D21S11, D2S1338, D3S1358, D8S1179, FGA, TH01, VWA calculados para 1797 casos tríos analizados en este estudio. Población de referencia: Región R3 Amazonas con diferentes tamaños muestrales: A. $n/4$, B. $n/2$, y C. n .

2.3.2.4. Región 4: Bogotá.

En una población con patrones de mezcla altos y elevada migración como Bogotá, al ser empleada como referencia en los cálculos de índices de paternidad, observamos que al aumentar el tamaño de la población de referencia hay una concentración en los valores de IP hacia puntos de inercia más agregados (Gráfica 6). Este agrupamiento es evidente al duplicar el tamaño muestral desde $n/4$ a $n/2$ (Gráfica 6A y 6B). Al igual que las anteriores poblaciones de referencia analizadas (R1 a R3) se evidenciaron dos puntos aislados correspondientes a casos con IP elevados (Gráfica 6A: cuadrantes I y II; Gráfica 6B: Cuadrante I y II; Gráfica 6C: Cuadrante I y II).

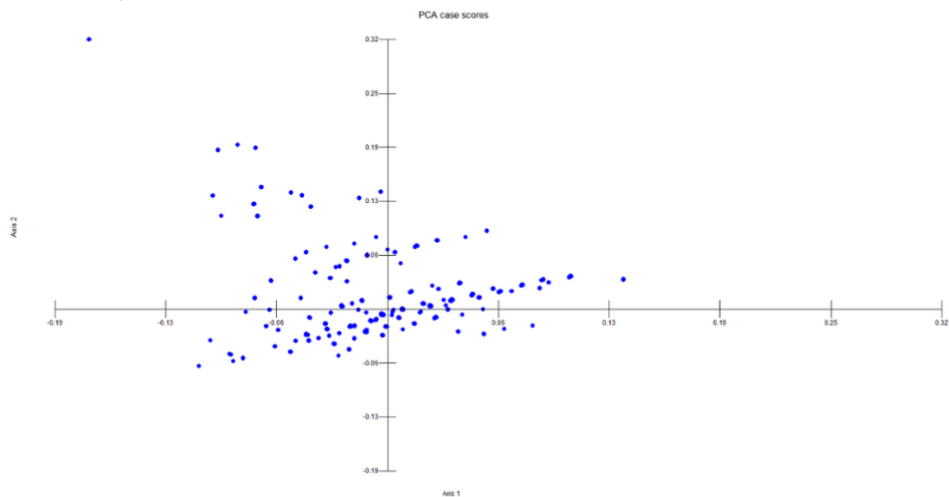
En $n/4$, se observa que los casos se encuentran dispersos a lo largo del eje x y y ; y aunque muchos casos se agrupan bajo la misma coordenada, hay una distribución no agregada en nubes de casos. Al doblar el tamaño muestral en $n/2$, se ve menor número de casos ya que muchos de estos se ubicaron en la misma coordenada (x, y) ; pero también aparecen los primeros clústers conformados por grupos de casos (Gráfico 6B). Con el tamaño muestral n , se hace aún más evidente cómo los IP se van concentrando en conglomerados mucho más densos (Gráfico 6C.)

Examinando las frecuencias alélicas mínimas por marcadores para la población de referencia Bogotá, encontramos que nuevamente hay marcadores con frecuencias mínimas calculadas similares: D16S539 ($p_{min}= 0,000512$), D21S11 ($p_{min}= 0,000474$); por otro lado D18S51 ($p_{min}= 0,000256$), D2S1358 ($p_{min}= 0,000320$), D19S433 ($p_{min}= 0,000366$), y TH01 ($p_{min}= 0,000284$).

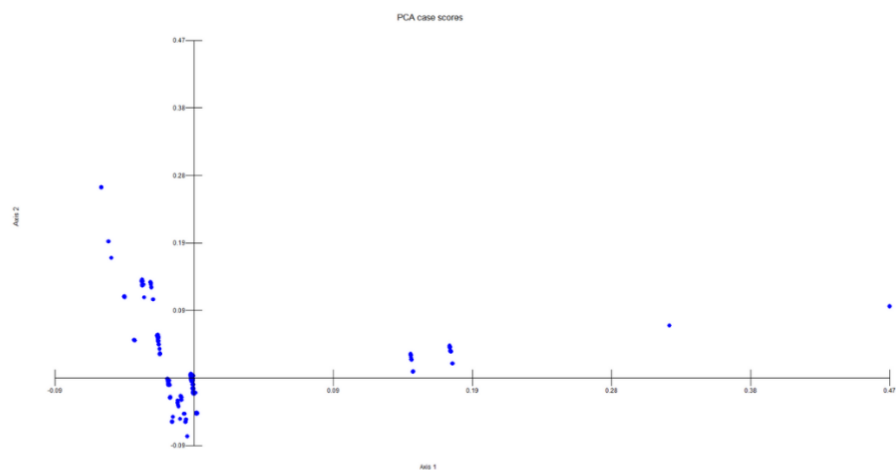
Al observar la distribución de frecuencias alélicas en cada uno de los diez marcadores, en D16S539 sólo 5 alelos se encontraban en la distribución inicial y los 14 restantes fueron calculados como frecuencia alélica mínima. Lo mismo ocurre en los demás marcadores, más de la mitad de la distribución es completada bajo el cálculo de Budowle et al., 1996; esto es lo que explica que muchos IP calculados por marcador resulten en el mismo valor y haya una sobreposición en la ubicación espacial del cartesiano de PCA para diferentes casos.

Índice de Paternidad - R4: Bogotá

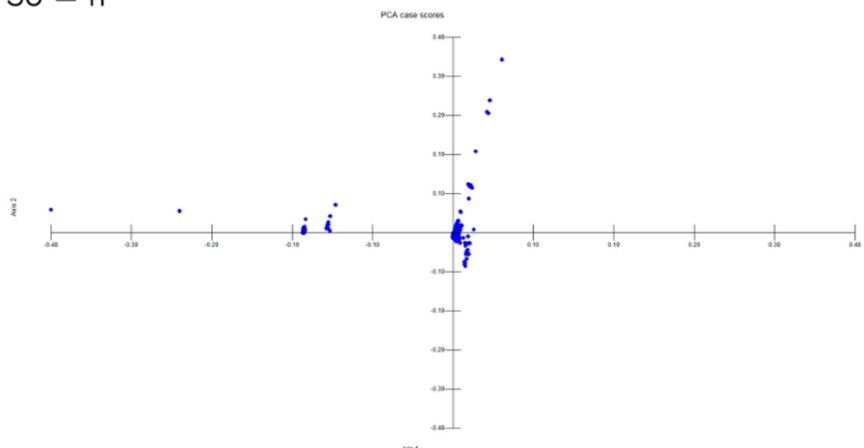
A. $S1 = n/4$



B. $S2 = n/2$



C. $S3 = n$



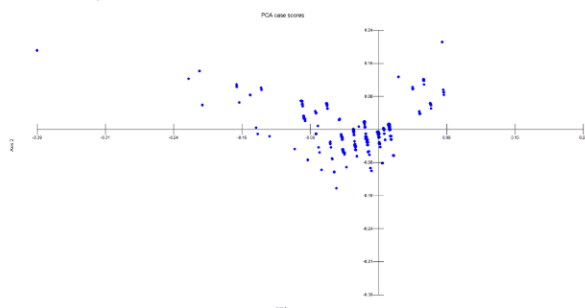
Gráfica 6 Gráfico de dispersión del PCA construido a partir de los índices de paternidad con los sistemas D16S539, D18S51, D19S433, D21S11, D2S1338, D3S1358, D8S1179, FGA, TH01, VWA calculados para 1797 casos tríos analizados en este estudio. Población de referencia: Región R4 Bogotá con diferentes tamaños muestrales: A. $n/4$, B. $n/2$, y C. n .

2.3.2.4. Región 5: Caribe.

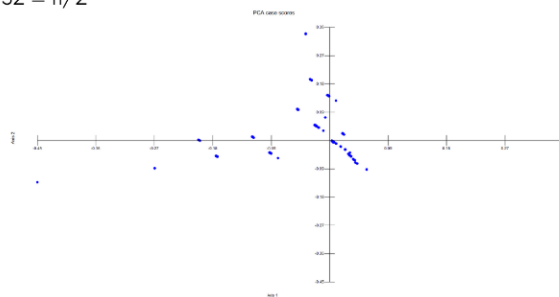
En la última región ensayada, al aumentar el tamaño de la muestra se observó el mismo comportamiento de agrupamiento hacia un centro de inercia de los valores de IP calculados (Gráfica 7). El cambio en la distribución de los casos se hace notorio al duplicar el tamaño muestral de $n/4$ a $n/2$ (Gráfica 7A. y Gráfica 7B.), de donde pasamos a una dispersión generalizada de los casos a la formación de conglomerados. Nuevamente observamos casos atípicos con IP elevados, esta vez en los cuadrantes I y II en $n/4$, II y III en $n/2$ y IV en n .

Índice de Paternidad - R5: Caribe

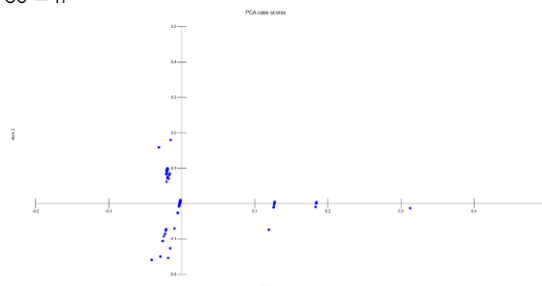
A. $S1 = n/4$



B. $S2 = n/2$



C. $S3 = n$



Gráfica 7 Gráfico de dispersión del PCA construido a partir de los índices de paternidad con los sistemas D16S539, D18S51, D19S433, D21S11, D2S1338, D3S1358, D8S1179, FGA, TH01, VWA calculados para 1797 casos tríos analizados en este estudio. Población de referencia: Región R5 Caribe con diferentes tamaños muestrales: A. $n/4$, B. $n/2$, y C. n .

Para esta región, por ejemplo, en $n/4$, encontramos que los marcadores D19S433 ($p_{min}=0,0000752$) y D2S1338 ($p_{min}=0,0000752$) tienen el mismo valor en frecuencia alélica mínima; que al igual que en las otras poblaciones de referencia (R1 a R4) explica que en diferentes casos para diferentes marcadores resulten IP muy parecidos.

2.3.3. Probabilidades de Paternidad por región, submuestra y replica.

Para cada una de las réplicas de poblaciones de referencias que corresponden a submuestreos con distintos tamaños muestrales, se realizó el cálculo de índices de paternidad (W) para cada uno de los genes evaluados en los 1797 casos trío (Figura 6F). Se representó gráficamente mediante análisis de componentes principales la variabilidad del resultado de W por marcador de estos 1797 casos bajo las condiciones de diferentes tamaños n para las poblaciones de referencia empleadas.

2.3.3.1. Región 1: San Andrés.

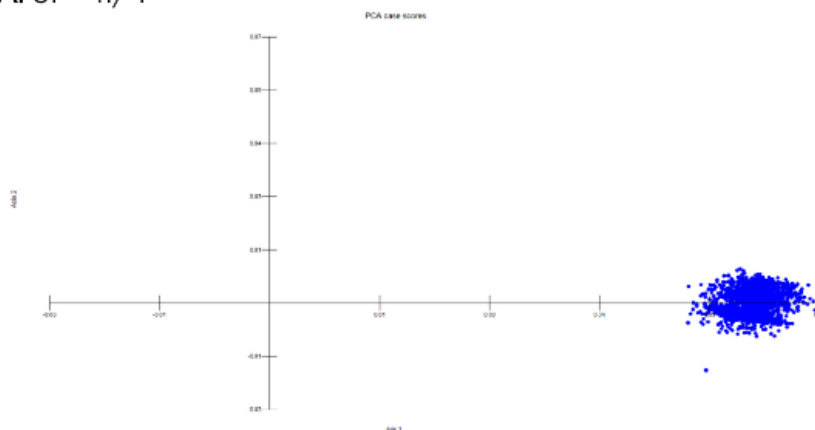
Al evaluar el comportamiento de las probabilidades de paternidad para la Región 1 (Gráfica 8), se evidenció que no hay una mayor diferenciación entre la nube de puntos del ensayo con un tamaño muestral de $n/4$ a n . Lo que resalta de este análisis son los casos externos en cada uno de los ensayos, sobre todo el caso que llamamos atípico que se observa en todas las regiones fuera de la nube de puntos agrupada en el cuadrante IV.

2.3.3.2. Región 2: Wayuu.

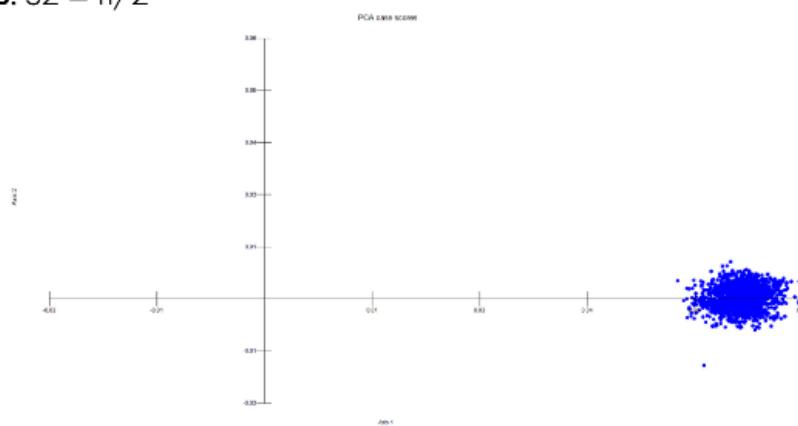
Al emplear como población de referencia a la Región 2 *Wayuu* (Gráfica 9), de igual forma a R1: San Andrés se evidenció que no hay una mayor diferenciación entre la nube de puntos del ensayo con un tamaño muestral de $n/4$ a n . Nuevamente en el cuadrante IV llamó la atención un caso aislado del resto de la densidad de puntos (Gráfica 9A y 9B).

Probabilidad de Paternidad - R1: San Andrés

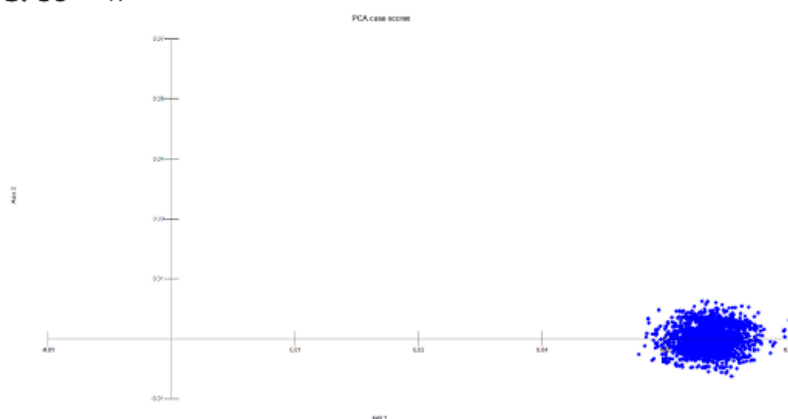
A. $S1 = n/4$



B. $S2 = n/2$



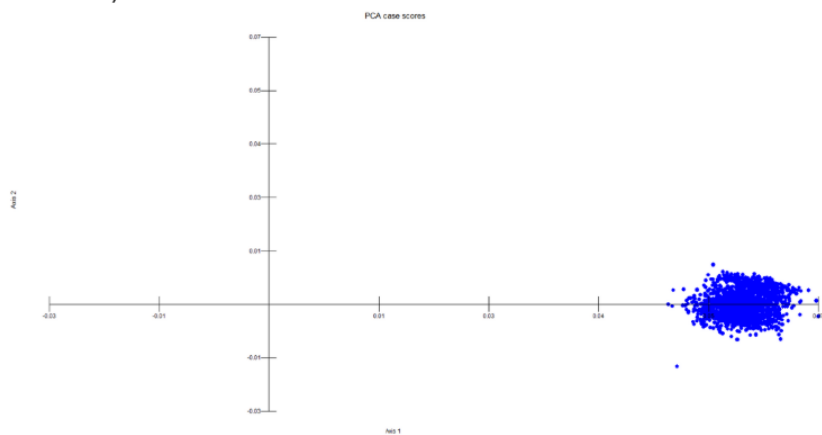
C. $S3 = n$



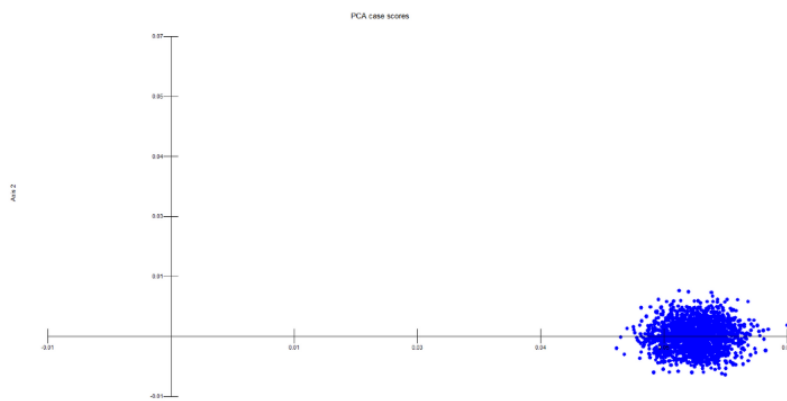
Gráfica 8 Gráfico de dispersión del PCA construido a partir de las probabilidades de paternidad con los sistemas D16S539, D18S51, D19S433, D21S11, D2S1338, D3S1358, D8S1179, FGA, TH01, VWA calculados para 1797 casos trios analizados en este estudio. Población de referencia: Región R1 San Andrés con diferentes tamaños muestrales: A. $n/4$, B. $n/2$, y C. n .

Probabilidad de Paternidad - R2: Wayuu

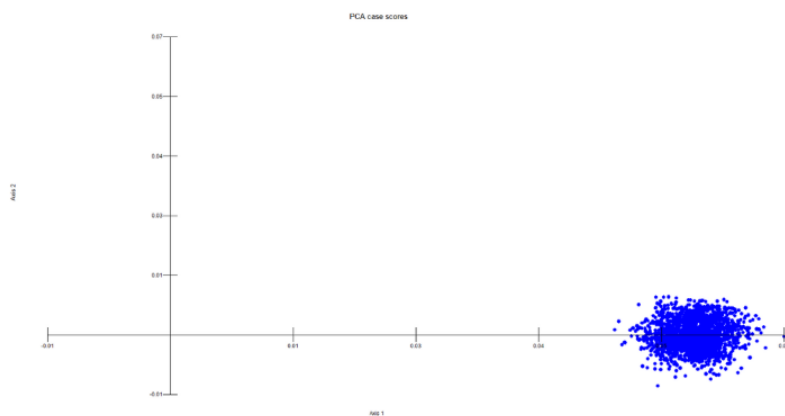
A. $S1 = n/4$



B. $S2 = n/2$



C. $S3 = n$



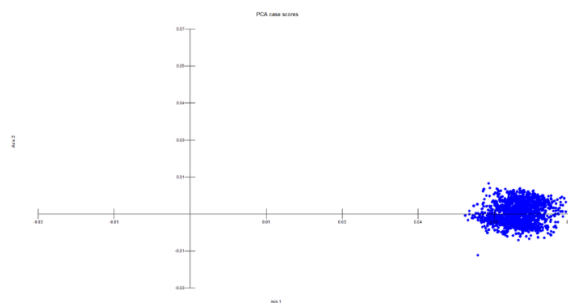
Gráfica 9 Gráfico de dispersión del PCA construido a partir de las probabilidades de paternidad con los sistemas D16S539, D18S51, D19S433, D21S11, D2S1338, D3S1358, D8S1179, FGA, TH01, VWA calculados para 1797 casos tríos analizados en este estudio. Población de referencia: Región R2 Wayuu con diferentes tamaños muestrales: A. $n/4$, B. $n/2$, y C. n .

2.3.3.3. Región 3: Amazonas.

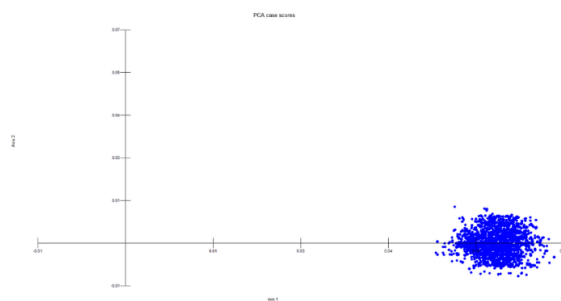
Siguiendo con la Región 3: Amazonas (Gráfico 10), una población conservada como Wayúu se evidenció que los resultados de probabilidad de paternidad no mostraron una mayor diferenciación entre la nube de puntos del ensayo con un tamaño muestral de $n/4$ a n . Sin embargo, a medida que se aumenta el tamaño muestral se evidencia como aumenta la aglomeración de los casos, y cómo uno de los casos se acerca a la densidad de puntos hasta hacer parte del clúster.

Probabilidad de Paternidad - R3: Amazonas

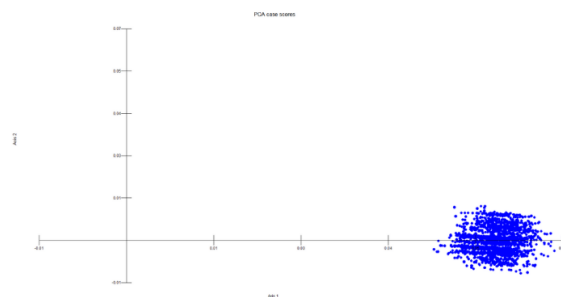
A. $S1 = n/4$



B. $S2 = n/2$



C. $S3 = n$



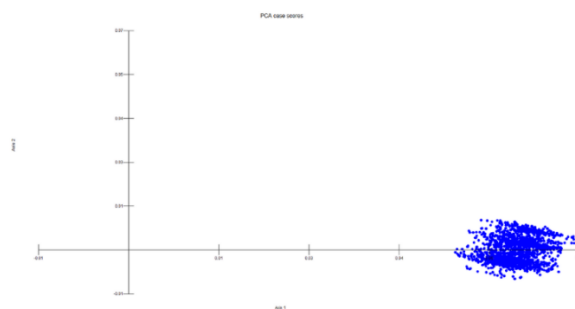
Gráfica 10 Gráfico de dispersión del PCA construido a partir de las probabilidades de paternidad con los sistemas D16S539, D18S51, D19S433, D21S11, D2S1338, D3S1358, D8S1179, FGA, TH01, VWA calculados para 1797 casos tríos analizados en este estudio. Población de referencia: Región R3 Amazonas con diferentes tamaños muestrales: A. $n/4$, B. $n/2$, y C. n .

2.3.3.4. Región 4: Bogotá.

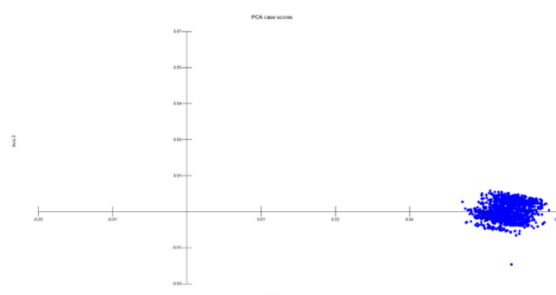
A diferencia de poblaciones de referencia con distribuciones de frecuencias alélicas más limitadas (San Andrés, Wayúu y Amazonas) y con mayor empleo de frecuencias alélicas mínimas para completar la distribución y poder correr los casos de paternidad; en la Región 4: Bogotá se encontró que medida que se aumenta el tamaño muestral (Gráfica 11) y los casos se agregan a un núcleo de inercia más denso, los casos atípicos se diferencian aún más del resto de casos agregados (Gráfica 11A y 11B).

Probabilidad de Paternidad - R4: Bogotá

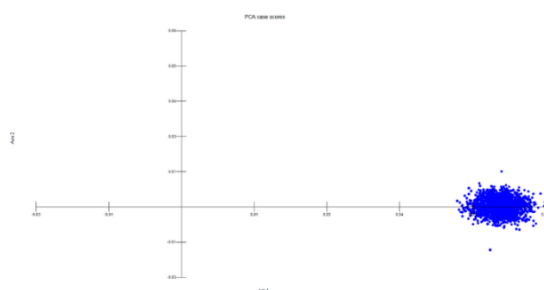
A. $S1 = n/4$



B. $S2 = n/2$



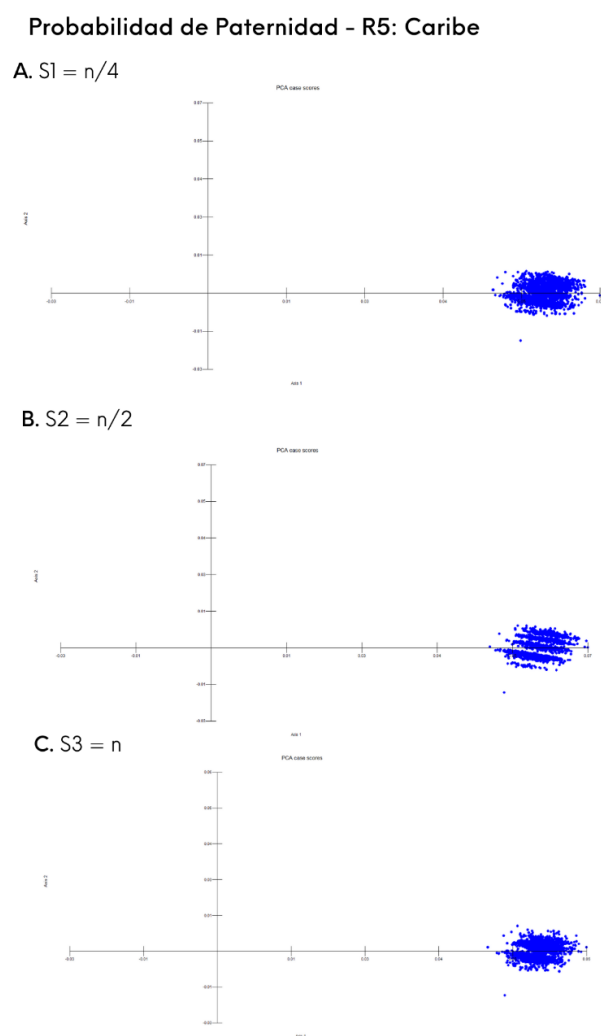
C. $S3 = n$



Gráfica 11 Gráfico de dispersión del PCA construido a partir de las probabilidades de paternidad con los sistemas D16S539, D18S51, D19S433, D21S11, D2S1338, D3S1358, D8S1179, FGA, TH01, VWA calculados para 1797 casos tríos analizados en este estudio. Población de referencia: Región R4 Bogotá con diferentes tamaños muestrales: A. $n/4$, B. $n/2$, y C. n .

2.3.3.5. Región 5: Caribe.

En la última población de referencia ensayada Región 5: Caribe, se encontró el mismo patrón que en las poblaciones más conservadas en los resultados de las probabilidades de paternidad individuales calculadas por marcador por caso (Gráficas 8, 9, 10 y 12). Desde $n/4$ a $n/2$ se observó como los casos aglomerados en una densidad de puntos formaron unos grupos de casos separados entre sí (Gráfica 12A y 12B); en n se observa cómo se agrupan aún más los datos sin embargo el dato atípico sigue sin agregarse al clúster formado (Gráfica 12C).



Gráfica 12 Gráfico de dispersión del PCA construido a partir de las probabilidades de paternidad con los sistemas D16S539, D18S51, D19S433, D21S11, D2S1338, D3S1358, D8S1179, FGA, TH01, VWA calculados para 1797 casos tríos analizados en este estudio. Población de referencia: Región R4 Bogotá con diferentes tamaños muestrales: A. $n/4$, B. $n/2$, y C. n .

2.3.4. Casos atípicos o raros.

Se definió como **caso raro o atípico** a aquel caso que se observó fuera de la nube de puntos mediante interpretación gráfica. En total fueron seleccionados 12 casos, los cuales fueron llamados: 17619, 16228, 18549, 16367, 20192, 11066, 13304, 16290, 19071, 17093, 13356. En todas las regiones a excepción de Bogotá, cuando el tamaño muestral aumenta de $n/4$ a $n/2$, estos casos aislados se agregan a la nube de puntos.

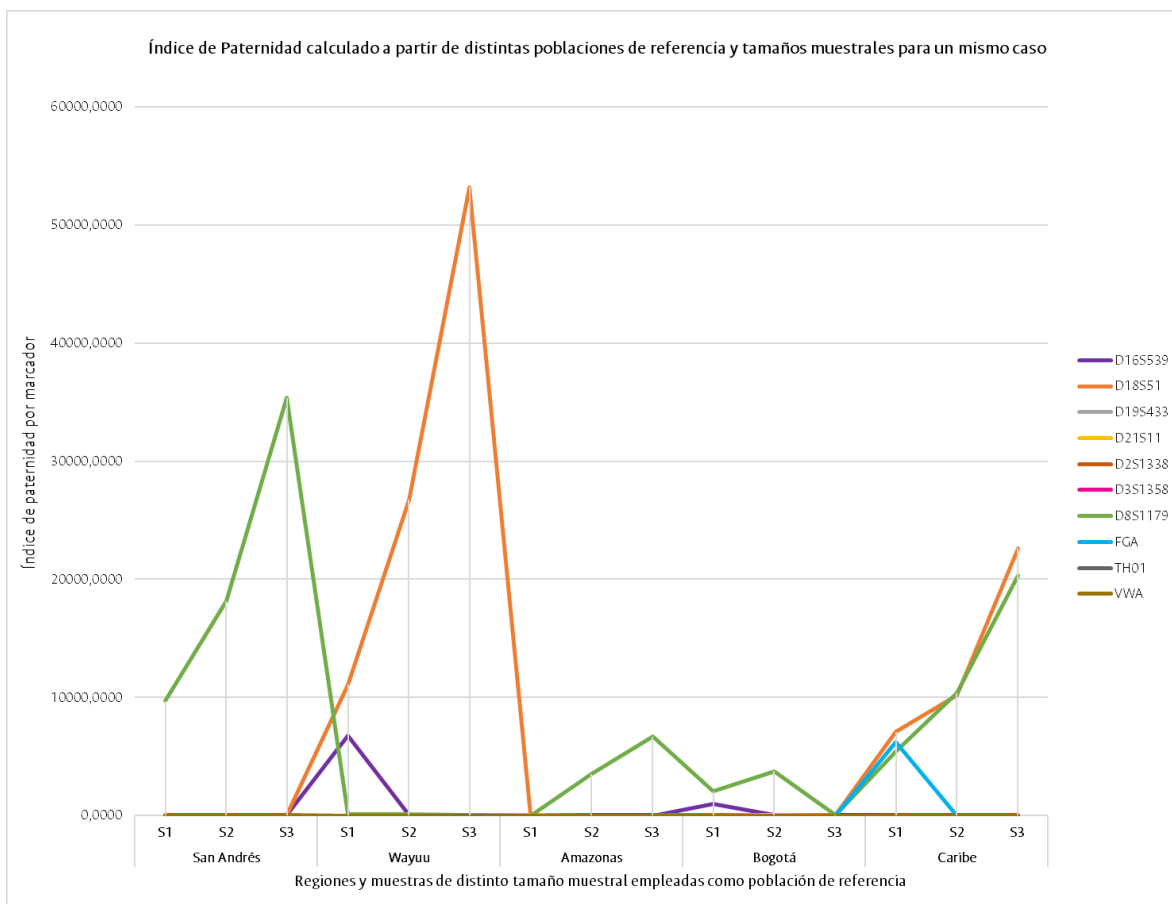
2.3.4.1. Según el índice de paternidad (IP).

Teniendo en cuenta las gráficas de componentes principales para los índices de paternidad, se establecieron como **casos atípicos o raros** aquellos que se encontraron en unitario o separados de las densidades de puntos agrupadas en diferentes clústers. Lo que se observa en estos casos, es que para un marcador particular el IP es mayor que el resto de los marcadores en uno o dos órdenes de magnitud o se evidencia una exclusión materna o paterna.

Tomando como ejemplo los resultados del caso mostrado en la Gráfica 13, fue de interés el comportamiento de los marcadores D18S51(en naranja) y el D8S1179 (en verde). En la Tabla S3, se observan los índices de paternidad calculados por marcador por región y tamaño muestral empleado para la población de referencia. Para el marcador D18S51, en todos los tres muestreos S1, S2 y S3 de las cinco regiones se evidenció que a medida que aumentó el tamaño muestral, también aumentó el índice de paternidad. Comportamiento esperado ya que como a medida que se aumenta el tamaño muestral la frecuencia alélica mínima disminuye (Sección 2.3.1: Gráfica 2), haciendo que cuando en algún caso se encontraran alelos raros o poco comunes el IP tuviera un valor alto. En D8S1179, se evidenció el mismo comportamiento, a excepción de las poblaciones Wayúu y Bogotá.

Las diferencias en IP a medida que se cambia una población de referencia con diferente tamaño muestral y ancestralidad son contrastantes hasta en tres órdenes de magnitud. En D18S51, con un tamaño muestral n , en San Andrés, Amazonas y Bogotá el IP osciló entre 15,6666 y 44,0833, contrastando con Caribe y Wayúu con IP tres órdenes de magnitud mayores (22615,2919 y 53191,4894; respectivamente) (Tabla S3). El alelo 11 del marcador D18S51, empleado en el cálculo de índice de paternidad, en las poblaciones de Wayuu y Caribe corresponde a una frecuencia mínima

calculada, y para el resto de las poblaciones corresponde a un alelo común; esto explica el por qué índices de paternidad tan distantes entre poblaciones.



Gráfica 13 Índice de paternidad de los sistemas D16S539, D18S51, D19S433, D21S11, D2S1338, D3S1358, D8S1179, FGA, TH01, VWA calculados para un caso trío analizado en este estudio. Se emplearon 12 poblaciones de referencia que corresponden a muestreos de diferentes tamaños muestrales de $S1 = n/4$, $S2 = n/2$ y $S3 = n$ de las regiones R1: San Andrés, R2: Wayuu, R3: Amazonas, R4: Bogotá y R5: Caribe.

Ahora, en D8S1179 para S3 se observó que las poblaciones de Wayuu y Bogotá ($IP_{WAY} : 65,6250$ e $IP_{BOG} : 45,5000$) presentan IP menores en dos órdenes de magnitud a Amazonas ($IP_{AMA} : 6667,7882$) y tres órdenes de magnitud a Caribe y San Andrés ($IP_{CAR} : 20353,7877$ e $IP_{SAI} : 35346,0007$) (Tabla S3). El alelo 8 del marcador D8S1179 empleado en el cálculo de IP para las poblaciones de San Andrés y Amazonas corresponde a un alelo de frecuencia mínima calculada y para Wayuu y Bogotá, corresponde a un alelo de frecuencia menor al 0.01. Este comportamiento en los índices de paternidad diferencial por población de referencia muestra la necesidad de evaluar la fluctuación de

estos resultados a la luz del muestreo realizado y de la diversidad genética de la población que queremos tomar como referente.

2.3.4.2. Según la probabilidad de paternidad (W).

Teniendo en cuenta los PCA construidos a partir de las probabilidades de paternidad con los sistemas D16S539, D18S51, D19S433, D21S11, D2S1338, D3S1358, D8S1179, FGA, TH01, VWA calculados para 1797 casos tríos analizados en este estudio (Gráfica 8 a 12), hubo un caso que siempre permaneció fuera del clúster del resto de casos. Al examinar el caso más alejado, se encontró que presenta una exclusión materna en el marcador D8S1179, ya que el genotipo del presunto padre corresponde a 13/14, el del hijo 13/16 y de la madre 14/15.

Revisando otros casos atípicos como el número 16228, también se encontró una exclusión materna en el marcador D21S11 (PP: 31/32,2; H: 31/32,2, M:29/30). En el caso 18549, no se encuentran exclusiones, pero el valor de W para el marcador D3S1358 es menor que 0,5. Para el 16367, tampoco se encuentran exclusiones, pero sí probabilidades de paternidad en tres marcadores inferiores a 0,6 (D16S539, D8S1179, TH01), al examinar las tipificaciones se encontraron que presentaban alelos de frecuencia común y mayores a 0,2, lo que puede incidir en un IP más bajo que el resto de los marcadores.

A pesar de que estos casos se aíslan del núcleo principal de puntos, no es fácil generar una generalización causal a su comportamiento, ya que los casos presentaron particularidades como las exclusiones maternas y paternas e IP y W bajos para ciertos marcadores.

2.3.3. Un modelo de cálculo para Índices (IP) y probabilidad de paternidad (W) por intervalos de confianza.

Teniendo en cuenta el diagrama del experimento (Figura 6) realizado por cada una de las regiones desde R1 a R5, se muestra el proceso en un caso atípico para ejemplificar el modelo de cálculo de índices de paternidad y probabilidad de paternidad mediante intervalos de confianza.

A. Población de referencia con tamaños muestrales n : Una de las variables que se deben tener claras son las poblaciones de referencia que van a hacer empleadas en el modelo y el tamaño de muestra a emplearse o que se empleó para realizar las distribuciones alélicas que han de emplearse. A lo largo de este capítulo, se ha mostrado que el tamaño muestral sí tiene efecto sobre el resultado de índice y probabilidad de paternidad por lo que el más adecuado a emplear será el mayor tamaño muestral que se tenga para cada población. Para este ejercicio y ser enfático en la variabilidad del IP y W según el muestreo realizado para tomar la población de referencia, se han empleado los tres tamaños muestrales.

B. Cálculo de distribución de frecuencias alélicas para cada población de referencia a emplear: Se generaron las tablas de distribución de frecuencia alélica y se realizó el cálculo de frecuencias alélicas mínimas para cada población de referencia a emplear (Tabla S1 y S2).

C. Resolución del caso trío con cada una de las poblaciones de referencia con tamaño muestral distinto generadas en cada región: Se resolvió el caso trío con las frecuencias alélicas de las 15 poblaciones de referencia generadas y se obtuvieron los resultados de la Tabla S3.

D. Cálculo IP combinado por gen por población y generación del intervalo de confianza: Los resultados de índices de paternidad combinados se muestran en la Tabla 1.

Tabla 1 Cálculo IP combinado por gen por población y generación del intervalo de confianza para un caso determinado.

Población	IPC para cada tamaño muestral				
	S1	S2	S3	n	
San Andrés	9,74E+02	1,82E+03	3,54E+03	<i>Media</i>	1,74E+03
Wayúu	1,80E+03	2,69E+03	5,33E+03	<i>Desviacion</i>	1631,2891
Amazonas	5,20E+00	3,57E+02	6,73E+02	<i>Intervalo</i>	8,26E+02
Bogotá	3,06E+02	3,79E+02	8,96E+00	<i>Int min</i>	915,1289
Caribe	1,88E+03	2,06E+03	4,30E+03	<i>Int max</i>	2655,7888

El índice de paternidad para este caso sería representado como: $IP = 1,74E+03 \pm 8,26E+02$.

E. Cálculo W combinado por gen por población y generación del intervalo de confianza: Los resultados de índices de paternidad combinados se muestran en la Tabla 2.

Tabla 2 Cálculo W combinado por gen por población y generación del intervalo de confianza para un caso determinado.

Población	WC para cada tamaño muestral				
	S1	S2	S3	<i>n</i>	
San Andrés	0,99897	0,99945	0,99972	<i>Media</i>	0,981601
Wayúu	0,99944	0,99963	0,99981	<i>Desviación</i>	0,047068
Amazonas	0,83881	0,99721	0,99852	<i>Intervalo</i>	0,023819
Bogotá	0,99674	0,99737	0,89959	<i>Int min</i>	0,957781
Caribe	0,99947	0,99951	0,99977	<i>Int max</i>	1,939382

La probabilidad de paternidad para este caso sería representada como: $W = 0,981601 \pm 0,047068$.

La interpretación final de este caso sería:

NO EXCLUSIÓN: En los resultados obtenidos se observa que es **1740,6599081984 ± 825,53** veces más probable que el **PP** sea el padre biológico de **H1**, hijo de **M1**, con una probabilidad de **0,981601 ± 0,047068**. Esta probabilidad se calcula por comparación con un hombre no relacionado biológicamente, analizado en las poblaciones de referencia de San Andrés, Wayúu, Amazonas, Bogotá y Caribe.

2.4. Discusión.

Actualmente, los laboratorios que realizan pruebas de filiación y paternidad en el país emplean la población de referencia que según criterio propio es escogida y varía entre cada uno de estos. Entre algunos de estos criterios está la ubicación del laboratorio en una región determinada del país, que reportes de frecuencias alélicas se han realizado recientemente por grupos de investigación de universidades regionales o por el mismo laboratorio y finalmente qué publicación científica tiene mayor soporte por el alcance y reconocimiento de los autores ante entidades competentes en los casos de disputa de paternidad.

Tanto en los estudios regionales como en los que poseen una aproximación nacional, se pueden detectar diferencias en las frecuencias alélicas y estimadores genéticos reportados, debido a que corresponden a muestreos completamente distintos sobre distintas muestras de población. Dos laboratorios que emiten el dictamen de un mismo caso, al emplear poblaciones de referencia con distribuciones alélicas diferenciales, presentarán diferencias no en el resultado final de exclusión y

no exclusión de una paternidad sino en el dato puntual de IP y W; que generan una duda a los implicados del caso y dificulta la interpretación los resultados de las autoridades.

Un acercamiento a cómo podemos expresar estas diferencias entre resultados puntuales de IP y W y englobarlas en un mismo dato, es teniendo en cuenta que la relación $\frac{x}{y}$ para el *IP* en un caso de paternidad es calculada a partir de la distribución de frecuencias de una población de interés llamada *población de referencia*, llamada de esta manera aunque formalmente no se esté calculando un parámetro sino un estadístico, ya que las frecuencias son obtenidas a partir del recuento de alelos observados en una muestra poblacional.

Para examinar los posibles cambios de resultados de Índice de Paternidad (PI) y la Probabilidad de Paternidad (W) fue importante seleccionar bases de datos STR de la población colombiana con distintas ancestralidades; pues permitió expresar índice y probabilidad de paternidad de un mismo caso como un intervalo de confianza; teniendo en cuenta la varianza generada en estos estimadores dada las distintas poblaciones de referencia empleadas (Tabla 2).

En el presente estudio llama la atención que la probabilidad y los índices de paternidad presenten diferencias causadas por el tamaño de muestra por muestreo (Tabla 2), donde se observan valores de W de 0,83881 para Wayuu ($n/2$) y para Bogotá de 0,89959 (n). La interpretación de estos resultados sería un *indicio de paternidad o paternidad no significativa* (Bravo Aguilar, 2009; Fábrega Ruíz, 1998), resultados que no interfieren en la no exclusión de paternidad, dado que para las demás 13 poblaciones la *paternidad es extremadamente probable*. Que tanto la frecuencia alélica empleada tanto en Wayuu como Bogotá sea frecuencia común y en adición el número de marcadores empleados en el estudio sea bajo (Pritchard et al., 2000) pueden explicar el por qué para estas dos poblaciones de los 15 totales se presentó un $W < 90$.

Esto ocurre cuando evaluamos cada probabilidad de paternidad como un resultado puntual por población de referencia empleada, sin embargo, al emplear el intervalo de confianza ($IP = 0,981601 \pm 0,047068$) se obtiene un resultado que no deja duda a la interpretación y que, además está soportada en que fue calculada mediante cinco poblaciones de referencia de ancestralidades diferentes reconociendo así la importancia de la genética de poblaciones en el qué hacer de las pruebas de filiación y permitiendo hacer un acercamiento al valor estimado de IP y W en relación con la variabilidad genética presente en la población y la presencia de variantes en baja frecuencia originadas por procesos de migración y mutación que traerán consigo IP más altos.

2.5. Conclusiones.

En los últimos 10 años, el creciente número de paternidades en disputa en Colombia ha involucrado al menos a 65.400 padres de diferentes orígenes étnicos y ancestralidades. Se examinó qué tan sensibles son el Índice de Paternidad (PI) y la Probabilidad de Paternidad (W) a la selección de la base de datos STR de la población con distintas ancestralidades, los resultados sugieren que, aunque no hay diferencias en el resultado de exclusión y no exclusión de la paternidad, los datos numéricos puntuales varían de una población a otra al menos para el IP en hasta tres órdenes de magnitud. También se hace evidente el efecto de la población de referencia sobre todo cuando se trabajan con poblaciones muy conservadas con altos porcentajes de pertenencia a un componente ancestral, ya que se hace necesario incluir a las distribuciones de frecuencias alélicas el cálculo de la frecuencia alélica mínima que eleva los IP.

Respecto al modelo de cálculo de índice de paternidad y probabilidad de paternidad de un mismo caso como un intervalo de confianza, se tuvo en cuenta la varianza generada en estos estimadores dada las distintas poblaciones de referencia empleadas. Se realizaron 1797 casos tríos colombianos con un conjunto de 10 marcadores STR de uso común y 5 poblaciones de referencia con ancestralidades distintas; en donde se varió el tamaño muestral por región, se encontró que a mayor tamaño muestral el número de alelos efectivos se mantenía constante lo que permitía IP más elevados.

Las diferencias encontradas en valores de las probabilidad e índice de paternidad empleando poblaciones de referencia de distinto origen étnico, ancestralidad y tamaño muestral permitieron poder expresar el resultado de una prueba de paternidad como un intervalo de confianza de IP y W. Sería interesante poder ahondar con un set de poblaciones de referencia con mayor tamaño muestral y que hayan sido tipificadas con los mismos marcadores para poder aumentar el número de estos en el experimento y evaluar cómo se ven afectados estos valores de IP y W, no solo por el efecto de la base de datos sino por el efecto de número de marcadores (9, 10, 12 y 15), ya que en otras investigaciones se ha observado que se disminuye la informatividad a menor número de marcadores empleados.

El enfoque estadístico propuesto tiene como ventaja principal la capacidad de consolidar los resultados obtenidos de una misma población por diferentes investigadores que han empleado

diferentes métodos de muestreo. Esto permite que los datos se combinen de manera que se acerquen más a una evaluación precisa de la población en cuestión.

En este sentido, se recomienda expresar los resultados en forma de intervalos de confianza. Esta práctica no solo facilita el uso apropiado de esta herramienta, sino que también contribuye a que los resultados reflejen de manera más precisa las características de la población de referencia. Este enfoque es de particular importancia en el campo de la genética forense, ya que unificar los datos permite a todos los laboratorios del país utilizar bases de datos estandarizadas.

La utilidad última de este modelo de cálculo radica en que el reporte de resultados para la Administración de Justicia será coherente en todos los laboratorios. Esto significa que se utilizará la misma aproximación matemática para evaluar cómo varían los índices y probabilidades de paternidad en la población colombiana o cualquiera de interés, lo que, en última instancia, conduce a una comprensión más sólida de los informes de resultados en el ámbito forense.

2.6. Referencias.

- Alonso, L. A., & Usaqué, W. (2012). Y-chromosome and surname analysis of the native islanders of San Andrés and Providencia (Colombia). *HOMO-Journal of Comparative Human Biology*, 1–14. <https://doi.org/10.1016/j.jchb.2012.11.006>
- Benítez-Páez, A., & Reyes, H. O. (2003). Allelic frequencies at 12 STR loci in Colombian population. *Forensic Science International*, 136(1–3), 86–88. [https://doi.org/10.1016/S0379-0738\(03\)00220-2](https://doi.org/10.1016/S0379-0738(03)00220-2)
- Alcaldía de Bogotá (2019). *Sitio oficial Portal Bogotá*. Historia. <https://bogota.gov.co/historia-de-bogota-recorrido-por-la-historia-de-la-ciudad-de-bogota>
- Braga, Y., Arias B., L., & Barreto, G. (2012). Diversity and genetic structure analysis of three Amazonian Amerindian populations from Colombia. *Colombia Médica*, 43(2), 133–140. <http://www.scielo.org.co/pdf/cm/v43n2/v43n2a05.pdf>
- Bravo Aguilar, M. L. J. (2009). Investigación de la Paternidad Biológica. In *La verdad genética de la paternidad* (I, pp. 45–80). Universidad de Antioquia.
- Bravo, M. L., Moreno, M. A., Builes, J. J., Salas, A., Lareu, M. v., & Carracedo, A. (2001). Autosomal STR genetic variation in negroid Chocó and Bogotá populations. *International Journal of Legal Medicine*, 115(2), 102–104. <https://doi.org/10.1007/s004140100223>

-
- Budowle, B., Monson, K. L., & Chakraborty, R. (1996). Estimating minimum allele frequencies for DNA profile frequency estimates for PCR-based loci. *International Journal of Legal Medicine*, *108*, 173–176.
- Burgos, G., Restrepo, T., Ibarra, A., Gaviria, A., Machado, G., Mora, C., & Lizarazo, R. (2015). Allelic frequencies and forensic parameters for miniSTRs D10S1248, D14S1434 and D22S1045 (NC01) in a sample from Central Andean Colombian region. *Forensic Science International: Genetics Supplement Series*, *5*, e81–e82. <https://doi.org/10.1016/j.fsigss.2015.09.033>
- Castillo, A., Gil, A., Pico, A., Vargas, C., Yurrebaso, I., & García, O. (2013). Genetic variation for 20 STR loci in a northeast Colombian population (Department of Santander). *Forensic Science International: Genetics Supplement Series*, *4*(1). <https://doi.org/10.1016/j.fsigss.2013.10.152>
- Chakraborty R. (1981). Expected number of alleles per locus in a sample and estimation of mutation rates. *American Journal of Human Genetics*, *33*, 481–484.
- CINEP. (1998a). *Colombia: País de Regiones. Región Noroccidental - Región Cundiboyacense* (F. Zambrano Pantoja, Ed.; Tomo II). Investigación y Educación popular), COLCIENCIAS.
- CINEP. (1998b). *Colombia: País de Regiones. Región Occidental - Región Caribe*. (F. Zambrano Pantoja, Ed.; Tomo IV). CINEP (Centro de Investigación y Educación popular), COLCIENCIAS.
- Correa Rubio, C. N., & Sanchez Rodriguez, P. S. (2021). *La paternidad evadida en Colombia: El derecho a la filiación de los menores versus el derecho a la intimidad y la autonomía de la voluntad del presunto padre*. Universidad Cooperativa de Colombia - Sede Ibagué, Espinal.
- Durán, R., Zarante, I., Acevedo, M. L., Villegas, M. R., Salazar, J., Bocanegra, B. Y., & Bernal, J. (2003). Allelic frequency of six STR loci in five Colombian cities. *Journal of Forensic Sciences*, *48*(4), 887. <http://www.ncbi.nlm.nih.gov/pubmed/12877314>
- Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Annals of Statistics*, *7*, 1–26.
- Efron, B. (1982). The jackknife, the bootstrap, and other resampling methods. *Society for Industrial and Applied Mathematics, CBMS-NSF(Monograph)*, 38.
- Departamento de Estadística (2007). *Colombia una nación multicultural: Su diversidad étnica*.
- Fábrega Ruíz, C. F. (1998). *Pruebas Biológicas de Paternidad. Aspectos científicos y jurídicos de las mismas*.
- Franco-Candela, F. A., & Barreto, G. (2017). Estructura genética de poblaciones indígenas del occidente colombiano mediante el uso de marcadores ligados al cromosoma Y. *Revista de La Academia de Ciencias Exactas, Físicas y Naturales*, *41*(160), 281–289. <https://www.raccefyn.co/index.php/raccefyn/article/view/476/311>
- Gaviria, A., Ibarra, A. A., Jaramillo, N., Palacio, O. D., Acosta, M. A., Brion, M., & Carracedo, Á. (2004). Nineteen autosomal microsatellite data from Antioquia (Colombia). *Forensic Science International*, *143*(1), 69–71. <https://doi.org/10.1016/j.forsciint.2004.01.007>

- Gómez, M. V., Reyes, M. E., Cárdenas, H., & García, O. (2003a). Genetic variation for 7 STR loci in a Colombian population (Department of Valle del Cauca). *Journal of Forensic Science*, 48(3), 687–688. <https://pubmed.ncbi.nlm.nih.gov/12762550/>
- Ibarra, A., Freire-Aradas, A., Martínez, M., Fondevila, M., Burgos, G., Camacho, M., Ostos, H., Suarez, Z., Carracedo, A., Santos, S., & Gusmão, L. (2014). Comparison of the genetic background of different Colombian populations using the SNPforID 52plex identification panel. *International Journal of Legal Medicine*, 128(1), 19–25. <https://doi.org/10.1007/s00414-013-0858-z>
- Instituto Colombiano de Bienestar Familiar (ICBF). (2022). *Filiación - Pruebas de ADN | Portal ICBF - Instituto Colombiano de Bienestar Familiar ICBF*. <https://www.icbf.gov.co/bienestar/proteccion/filiacion-pruebas-adn>
- Jaramillo, S., & Turbay Ceballos, S. (2000). Los indígenas Zenúes. In *Geografía Humana de Colombia, Región Andina Central* (Tomo IV, V). Instituto Colombiano de Cultura Hispánica.
- Kovach, W. L. (2007). *MVSP - A MultiVariate Statistical Package for Windows*. (Version 3.2.2.; pp. 1–135). Kovach Computing Services. <https://www.kovcomp.co.uk/mvsp/>
- Losilla Vidal, J. M. (1994). *MonteCarlo Toolbox de Matlab: Herramientas para un laboratorio estadístico fundamentado en técnicas Monte Carlo*. Universitat Autònoma de Barcelona.
- Lucía Hincapié, M., Gil, A. M., Pico, A. L., Gusmão, L., Rondón, F., Vargas, C. I., & Castillo, A. (2009). Análisis de la estructura genética en una muestra poblacional de Bucaramanga, Departamento de Santander. *Colombia Médica*, 40(4), 1–12.
- Manly, B. F. J. (1991a). Monte Carlo and other computer-intensive methods. In *Randomization and Monte Carlo Methods in Biology* (I, pp. 21–30). Chapman and Hall.
- Manly, B. F. J. (1991b). Randomization test and confidence intervals. In *Randomization and Monte Carlo Methods in Biology* (I, pp. 2–20). Chapman and Hall.
- Martínez, B., Builes, J. J., Aguirre, D., Mendoza, L., Hernández, L., & Marrugo, J. (2017). Autosomic STR database for an afrodescendant population sample of San Basilio de Palenque, Colombia. *Forensic Science International: Genetics Supplement Series*, 6, e555–e557. <https://doi.org/10.1016/j.fsigss.2017.09.217>
- Martínez, B., Builes, J. J., & Caraballo, L. (2008). Genetic data analysis of nine STRs in two Caribbean Colombian populations: César and Guajira. *Journal of Forensic Sciences*, 53(1), 254–255. <https://doi.org/10.1111/j.1556-4029.2007.00631.x>
- Martínez, B., Caraballo, L., Barón, F., Gusmão, L., Amorim, A., & Carracedo, A. (2006). Analysis of STR loci in Cartagena, a Caribbean city of Colombia. *Forensic Science International*, 160(2–3), 223. <https://doi.org/10.1016/j.forsciint.2005.05.035>
- Martínez, B., Pereira, R., Meza, K., Hernández, L., Amorim, A., Marrugo, J., & Gusmão, L. (2017). Forensic Science International: Genetics Supplement Series Ancestry estimates in afrodescendant population from San Basilio de Palenque, Colombia. *Forensic Science*

International: Genetics Supplement Series, 6, e224–e225.
<https://doi.org/10.1016/j.fsigss.2017.09.105>

- Meisel, A. (2005). La continentalización de la isla de San Andrés, Colombia: Panyas, raizales y turismo. In *Economías locales del Caribe colombiano: siete estudios de caso*. (Colección, pp. 12–43). Banco de la República.
- Mincultura, M. de C. (2005). *Caracterización de los pueblos Indígenas de Colombia. Dirección de Poblaciones. Tikuna, los hijos de Yoi e Ipi, y gente de tierra firme*. <http://www.mincultura.gov.co/areas/poblaciones/noticias/Documents/Caracterización del pueblo Tikuna.pdf>
- Ministerio de Asuntos Exteriores y de Cooperación, & Oficina de Información Diplomática del Departamento de Relaciones Exteriores. (2017). *Ficha País República de Colombia*. http://www.exteriores.gob.es/Documents/FichasPais/COLOMBIA_FICHA PAIS.pdf
- Mogollon Olivares, F., Moncada Madero, J., Casas-vargas, A., Zea Montoya, S., Suárez Medellín, D., & Usaquén, W. (2020). Contrasting the ancestry patterns of three distinct population groups from the northernmost region of South America. *American Journal of Physical Anthropology*, e24130. <https://doi.org/10.1002/ajpa.24130>
- Morcote Ríos, G., Mora Camargo, S., & Franky Calvo, C. (2006). *Pueblos y paisajes de la selva Amazónica*. Universidad Nacional de Colombia, Fundación Taraxacum.
- Moreno Bandeira, V. H. (2018). *Sitio oficial de la Gobernación de Amazonas*. Historia. <http://www.amazonas.gov.co/departamento/nuestro-departamento>
- Moroni, R., Gasbarra, D., Arjas, E., Lukka, M., & Ulmanen, I. (2011). Effects of Reference Population and Number of STR Markers on positive evidence in Paternity Testing. *Journal of Forensic Research*, 02(02). <https://doi.org/10.4172/2157-7145.1000119>
- Neel JV. (1973). “Private” genetic variants and the frequency of mutations among South American Indians. *Proc Natl Acad Sci USA*, 70, 3311–3315.
- Nei M. (1975). Molecular population genetics and evolution. . *North Holland/American Elsevier*, 118.
- Oliver, J. (1990). Reflexiones sobre los posibles orígenes del wayuu (guajiro). In *La Guajira: de la memoria al porvenir: Una visión antropológica*.
- Ortega Torres, J., Rueda, O. L., & Jaime, L. A. (2015). *DOCUMENTO GUÍA PRUEBAS DE ADN PARA INVESTIGACIÓN DE PATERNIDAD Y/O MATERNIDAD (Versión actualizada-Enero-2015)* .
- Ossa Reyes, H., Torres Ramírez, L. J., & Nieto Romero, L. V. (2009). Frecuencias alélicas y haplotípicas del Sistema hla clase i (loci a*, b*) en una población de indígenas Motilón-Barí, Norte de Santander, Colombia. *Nova*, 7(12), 131. <https://doi.org/10.22490/24629448.426>
- Palacio, O. D., Triana, O., Gaviria, A., Ibarra, A. A., Ochoa, L. M., Posada, Y., Maya, M. C., Lareu, M. V., Bríon, M., Acosta, M. A., & Carracedo, A. (2006). Autosomal microsatellite data from

- Northwestern Colombia. *Forensic Science International*, 160(2–3), 217–220. <https://doi.org/10.1016/j.forsciint.2005.05.034>
- Paredes, M., Galindo, A., Bernal, M., Avila, S., Andrade, D., Vergara, C., Rincón, M., Romero, R. E., Navarrete, M., Cárdenas, M., Ortega, J., Suarez, D., Cifuentes, A., Salas, A., & Carracedo, Á. (2003). Analysis of the CODIS autosomal STR loci in four main Colombian regions. *Forensic Science International*, 137(1), 67–73. [https://doi.org/10.1016/S0379-0738\(03\)00271-8](https://doi.org/10.1016/S0379-0738(03)00271-8)
- Parsons, J. J. (1985). *San Andrés y Providencia: Una geografía histórica de las islas colombianas del Caribe*. El Ancora Editores.
- Porras, L., Beltrán, L., Ortiz, T., Sanchez-Diz, P., Carracedo, A., & Henao, J. (2008). Genetic polymorphism of 15 STR loci in central western Colombia. *Forensic Science International: Genetics*, 2(1), e7–e8. <https://doi.org/10.1016/j.fsigen.2007.08.004>
- Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, 155, 945–959.
- Rey, M., Gutiérrez, A., Schroeder, B., Usaquén, W., Carracedo, A., Bustos, I., & Giraldo, A. (2003). Allele frequencies for 13 STR's from two Colombian populations: Bogotá and Boyacá. *Forensic Science International*, 136(1–3), 83–85. [https://doi.org/10.1016/S0379-0738\(03\)00221-4](https://doi.org/10.1016/S0379-0738(03)00221-4)
- Rivera Franco, N., Braga, Y., Espitia Fajardo, M., & Barreto, G. (2020). Identifying new lineages in the Y chromosome of Colombian Amazon indigenous populations. *American Journal of Physical Anthropology*, 172(2), 165–175. <https://doi.org/10.1002/ajpa.24039>
- Rojas, M. Y., Alonso, L. A., Sarmiento, V. A., Eljach, L. Y., & Usaque, W. (2013). Structure analysis of the La Guajira-Colombia population: A genetic, demographic and genealogical overview. *Annals of Human Biology*, 40(2), 119–131. <https://doi.org/10.3109/03014460.2012.748093>
- Rondón, F., César Osorio, J., Viviana Peña, Á., Andrés Garcés, H., & Barreto, G. (2008). *Diversidad genética en poblaciones humanas de dos regiones colombianas*. 39(2), 52–60.
- Rondón G., F., Oribio, R. F., Braga, Y. A., Cárdenas, H., & Barreto, G. (2006). Estudio de Diversidad Genética de Cuatro Poblaciones Aisladas del Centro y Suroccidente Colombiano. *Revista de La Universidad Industrial de Santander. Salud*, 38(1), 12–20. <https://www.redalyc.org/pdf/3438/343837061004.pdf>
- Sánchez-Diz, P., Acosta, M. A., Fonseca, D., Fernández, M., Gómez, Y., Jay, M., Alape, J., Lareu, M. V., Carracedo, A., & Restrepo, C. M. (2009). Population data on 15 autosomal STRs in a sample from Colombia. *Forensic Science International: Genetics*, 3(3). <https://doi.org/10.1016/j.fsigen.2008.08.002>
- Talco Arias, J. (1994). *Los kankuamos: Un pueblo indígena en reconstrucción* (Ediciones Turdakede, Ed.). Organización Nacional Indígena Kankuama.
- Usaquén Martínez, W. (2012). *Validación y consistencia de información en estudios de diversidad genética humana a partir de marcadores microsatélites* [Universidad Nacional de Colombia]. <http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:No+Title#0>

- Vargas, C. I., Castillo, A., Gil, A. M., Pico, A. L., & García, O. (2003). *Population genetic data for 13 STR loci in a northeast Colombian (department of Santander) population*. https://www.isfg.org/files/04ab68f72414cb3c6171bdb4a84e4c055c9eee46.02003370_957336538240.pdf
- Vollmer, L. (1997). *La historia del poblamiento del archipiélago de San Andrés, Vieja Providencia y Santa Catalina*. (Ediciones).
- Yunis, J. J., Acevedo, L. E., Campo, D. S., & Yunis, E. J. (2013). Geno-geographic origin of Y-specific STR haplotypes in a sample of Caucasian-Mestizo and African-descent male individuals from Colombia. *Biomédica*, 33(3), 459–467. <https://doi.org/10.7705/biomedica.v33i3.807>
- Yunis, J. J., & Yunis, E. J. (2013). Mitochondrial DNA (mtDNA) haplogroups in 1526 unrelated individuals from 11 Departments of Colombia. *Genetics and Molecular Biology*, 36(3), 329–335. <https://doi.org/10.1590/S1415-47572013000300005>

Capítulo 3: Efecto de la ausencia de la madre en pruebas de paternidad y el número de *falsos presuntos padres no excluidos* a partir de 15 marcadores STRs utilizando una base de datos genética de Bogotá, Colombia.

Resumen.

En los últimos 10 años, el creciente número de paternidades en disputa en Colombia ha involucrado al menos a 65.400 padres de diferentes orígenes étnicos y ancestralidades. En este estudio, se determinó cuántas inclusiones de paternidad erróneas podrían detectarse en Bogotá, la sexta ciudad más poblada de Latinoamérica con un área geográfica de 1.775 km²; al comparar los resultados empíricos de 296 dúos madre (M) e hijo (H) y 219 presuntos padres (PP) (con 15 marcadores microsátélites). Esta comparación entre cada dúo (M-H) y todos los hombres no emparentados (PP) (es decir, todos los presuntos padres de los otros casos) se realizó con la implementación de un módulo de un programa informático especialmente diseñado para realizar cálculos de pruebas de paternidad y maternidad que dio como resultado 64.824 tríos M/H/PP. Para los 64.824 tríos M/H/PP se encontraron menos de tres STR excluyentes en 156 pruebas, es decir, el 0.2412 % de las pruebas trío dieron como resultado una no exclusión de la paternidad. Para 1 dúo M-H (es decir, 0.3378 %) se detectó uno o más hombres no emparentados con cero o tres STR excluyentes. Se realizó el mismo experimento con casos dúo PP – H no emparentados, de 65.262 pruebas dúo generadas, se pudo confirmar el padre putativo propuesto en 375 pruebas, es decir, el 0,5746 % de las pruebas dúo dieron como resultado una no exclusión de la paternidad. Se encontraron 55 Hijos con 1 presunto padre adicional; 29 con 2 P-P adicionales; 11 con 3 P-P adicionales; 2 hijos con 4 y 5 P-P adicionales y 1 hijo con 6 y 7 P-P adicionales, respectivamente. Las probabilidades de paternidad oscilaron entre 0,9974 y 0,9999. Lo que muestra que si se excluye a la madre de la prueba podemos hallar coincidencias en perfiles genéticos del presunto padre e hijo no emparentado. Estos resultados resaltan los posibles sesgos que se pueden presentar en los casos de paternidad sin madre utilizando solo análisis STR y recomiendan mucha precaución al asignar predicados verbales como "paternidad probada" sino seguir empleando términos como "paternidad biológica no excluida".

Palabras clave: pruebas de paternidad, inclusión errónea de paternidad, probabilidad de no exclusión, casos de paternidad sin madre, paternidad probada, paternidad biológica.

Abstract

In the last ten years, Colombia's growing number of paternity disputes has involved at least 65,400 fathers of different ethnic origins and ancestry. In this study, we determined how many erroneous paternity inclusions could be detected in Bogotá, the sixth most populous city in Latin America with a geographic area of 1,775 km², by comparing the empirical results of 296 mothers (M) and son (S) duos and 219 alleged fathers (AF) (with 15 microsatellite markers). This comparison between each duo (M-S) and all unrelated males (AF) (i.e., all alleged fathers of the other cases) was performed with the implementation of a specially designed software module to perform test calculations of paternity and maternity, which resulted in 64,824 M/S/AF trios. For the 64,824 M/S/AF trios, fewer than three excluding STRs were found in 156 tests; that is, 0.2412% of the trio tests resulted in a non-exclusion of paternity. For 1 M-S duo (i.e., 0.3378%), one or more unrelated males with zero or three excluding STRs were detected. The same experiment was performed with unrelated AF-S duo cases; out of 65,262 duo tests generated, the proposed putative father could be confirmed in 375 tests; that is, 0.5746% of the duo tests resulted in a non-exclusion of the paternity. 55 Children were found with one additional presumed father, 29 with two additional AF, 11 with three additional AFs, two children with 4 and 5 additional AF, and one son with 6 and 7 additional P-P, respectively. Paternity probabilities ranged from 0.9974 to 99.999%, which shows that if the mother is excluded from the test, we can find matches in the genetic profiles of the alleged father and unrelated child. These results highlight the potential biases that can occur in motherless paternity cases using only STR analysis and recommend great caution in assigning verbal predicates such as "*proven paternity*" rather than continuing to use terms such as "*non-excluded biological paternity*"

Keywords: paternity disputes, erroneous paternity inclusions, non-exclusion of paternity, motherless paternity cases, proven paternity, biological paternity.

3.1 Introducción.

El análisis de ADN y cotejo de alelos son una actividad rutinaria en la casuística de los laboratorios especializados en pruebas de filiación y paternidad, tanto en los casos que incluyen presunto padre e hijo, como la madre (Bravo Aguilar, 2009; Brinkmann et al., 2001; Poetsch et al., 2006; Thomson et al., 1999). Las pruebas de paternidad actuales **no prueban la paternidad biológica** del presunto padre; sino que **prueban la no paternidad** al excluir a los hombres cuyo genotipo es incompatible con el del niño en cuestión (Anderson, 2006; Luque Gutiérrez, 2019; Pena & Chakraborty, 1994). Encontrar menos de dos exclusiones o no encontrar exclusión alguna en el cotejo de los perfiles genéticos del presunto padre e hijo puede tomarse como **prueba de la no exclusión de paternidad** si la probabilidad de excluir a personas que no son padres biológicos es extremadamente alta (Anderson, 2006; Houck, 2015; Pena & Chakraborty, 1994; Tillmar, 2010). Esta probabilidad se calcula empleando el Teorema de Bayes que considera las frecuencias de los alelos bajo consideración en la población general y tiene valores superiores a 0.9999, de modo que, de 10.000 pruebas de paternidad, los no padres biológicos serán 9.999 veces excluidos (Luque Gutiérrez, 2019; Mickey et al., 1986; Pena & Chakraborty, 1994; Tagliabracci, 2010).

En Colombia, en el año 2021 se presentaron 2.589 procesos jurídicos de impugnación relacionados a disputas de paternidad, posicionándose entre los primero quince tipos de demanda más usada en los últimos 5 años (Correa Rubio & Sanchez Rodriguez, 2021). Esta estadística es casi tres veces menor al número de solicitudes de realización de pruebas de paternidad al ICBF (Instituto Colombiano de Bienestar Familiar) ejecutadas por el Laboratorio de Genética del Instituto Nacional de Medicina Legal y Ciencias Forenses; en donde se estima que al año se realizan aproximadamente 6540 pruebas (Ortega Torres et al., 2015). Si este dato se extrapola a los últimos diez años; se llegaría al orden de al menos 65.400 casos de pruebas de paternidades realizadas a nivel nacional desde que se implementó esta tecnología en el país en un único laboratorio.

En los últimos 10 años, el creciente número de paternidades en disputa en Colombia ha involucrado al menos a 65.400 padres de diferentes orígenes étnicos y ancestralidades. Teniendo en cuenta que, en las pruebas de paternidad, cuando los individuos involucrados pertenecen a la misma subpoblación, es menos probable que los presuntos padres muestren discrepancias de alelos en un sistema determinado con el genotipo del hijo (el Andari et al., 2018; Jacewicz et al., 2004; Lee et al.,

2013; Poetsch et al., 2006; Sánchez et al., 2008); se quiso determinar cuántas inclusiones de paternidad erróneas podrían detectarse en la ciudad de Bogotá; al comparar los resultados empíricos de 296 dúos madre (M) e hijo (H) y 219 presuntos padres (PP) al emplear 15 marcadores microsatélites. Esta comparación entre cada dúo (M-H) y todos los hombres no emparentados (PP) (es decir, todos los presuntos padres de los otros casos) se llevó a cabo mediante la realización módulo de programa informático especialmente diseñado para realizar cálculos probabilidad de paternidad y maternidad que dio como resultado 64.824 tríos M/H/PP.

Se tomó en cuenta únicamente la población de Bogotá, dado a que varios estudios han evidenciado que la población de referencia de las frecuencias alélicas de STR empleadas no tiene un impacto significativo en los cálculos de las pruebas de filiación, donde el uso de diferentes bases de datos de frecuencias poblacionales podría conducir al mismo resultado tal como se evidenció en el capítulo anterior (Andari et al., 2018; Fernandes et al., 2004; Mickey et al., 1986; Moroni et al., 2011). Debido a que la probabilidad de una **falsa no exclusión de paternidad** es mayor cuando se analiza un presunto padre y un hijo en cuestión (caso dúo: PP - H) que cuando hay una madre confirmada adicional (caso trío: PP – H- H) (Aguilar et al., 2021; Lee et al., 2013), se evaluaron los mismos 296 hijos (H) y 219 presuntos padres (PP) de los 64.824 tríos M/H/PP analizados con 15 marcadores, para examinar si el número de **falsos padres no excluidos** aumentaba o disminuía al no tomar a la madre en el caso.

3.2 *Materiales y métodos*

3.2.1. *Casos analizados.*

Se tomaron 296 hijos (H) y 219 presuntos padres (PP) que participaron en casos de paternidad tríos (presunto padre, madre e hijo) durante 2013 al 2021 en el laboratorio de Identificación Humana del Instituto de Genética de la Universidad Nacional de Colombia. Se tuvo en cuenta que no se presentaran exclusiones en marcadores entre madre e hijo. Al realizar la depuración de los casos de paternidad se revisó que se contara con el consentimiento informado, la información de lugar de nacimiento, y la información genética completa para los 15 marcadores como se estableció en el aval del Comité de Ética Médica de la Universidad Nacional de Colombia; cuidando también del anonimato de las personas involucradas en el estudio.

3.2.2. Extracción de ADN, tipificación de STRs y análisis de fragmentos.

El ADN se extrajo utilizando dos punches por muestra y se lavó con tampón de purificación de reactivos FTA. Para definir los perfiles genéticos, el genotipado de los loci STR se realizó con el kit comercial AmpFISTR® Identifier (Applied Biosystems, Warrington, U.K.) (15 marcadores: CSF1PO, D13S317, D16S539, D18S51, D19S433, D21S11, D2S1338, D5S818, D7S820, D8S1179, FGA, TPOX, vWA). Estas amplificaciones se realizaron en un Applied Biosystems 2720 Thermal Cycler™. La electroforesis capilar y la detección de productos amplificados se realizaron con el analizador genético ABI PRISM 310™. Las muestras se analizaron con el software GeneMapper ID™, versión 3.2 (Applied Biosystem).

3.2.3. Análisis estadístico y comparación de perfiles genéticos de presuntos padres e hijos no emparentados.

- A. Cálculo de distribución de frecuencias alélicas para la población de referencia.** Bogotá fue empleada como población de referencia en este análisis por tener un comportamiento de ancestralidad múltiple. Bogotá es una población en donde han confluído múltiples ancestrías, a pesar de que sus primeros pobladores fueron los Muisca, pertenecientes a la familia lingüística Chibcha cuyo tamaño población previo a la conquista y procesos de colonización se estima en medio millón de personas. Bogotá careció de un flujo importante de inmigrantes extranjeros, según los censos llevados a cabo en el siglo XIX, sin embargo, la población tuvo un crecimiento regular; en 1832 tenía 36.465 habitantes; en 1881, 84.723 habitantes y hacia finales de siglo casi 100.000. El establecimiento de Bogotá como la capital del país trajo consigo un crecimiento poblacional como consecuencia de un incremento en la oferta laboral en una gran variedad de campos e industrias, lo que dio como consecuencia una considerable ampliación física de la ciudad (Sitio oficial Portal Bogotá, 2022) y un flujo migratorio de alto individuos de todas las regiones del país; actualmente Bogotá con una extensión de 1.775 km² alberga 7,9 millones de personas es la sexta ciudad más poblada de Latinoamérica albergando a personas de todos los orígenes geográficos del país.

Teniendo en cuenta lo anteriormente descrito, las frecuencias alélicas de Bogotá del Laboratorio de Genética de Poblaciones de la Universidad Nacional de Colombia fueron empleadas como población de referencia para calcular las *probabilidades de paternidad* como probabilidades bayesianas a posteriori en los casos a resolver. También fueron calculadas las frecuencias alélicas mínimas para cada uno de los marcadores microsatélites empleados de acuerdo con Budowle et al., 1996:

$$p_{min} = 1 - [1 - (1-\alpha)^{\frac{1}{c}}]^{\frac{1}{2n}}$$

En donde p_{min} es la frecuencia alélica mínima, c es el número de alelos comunes que pueden ser estimados a partir del nivel de heterocigosidad o con una frecuencia mayor de 0.01 (Nei 1975; Neel 1973; Chakraborty 1981), y n es el número de individuos del set de datos estudiado. Para este estudio siguiendo las recomendaciones del autor citado, α fue fijado como 0.05.

- B. Resolución del caso trío con madre (M) e hijo (H) constantes.** Se tomaron 296 dúos madre e hijo sin exclusión materna que fueron analizados con 219 presuntos padres (PP), desde el padre 1 hasta el n ; analizándose desde el caso 1 al n formado por el dúo madre e hijo n con el presunto padre n (Figura 7B). Esta comparación entre cada dúo (M-H) y todos los hombres no emparentados (PP) (es decir, todos los presuntos padres de los otros casos) fue posible gracias a la realización de un módulo de un programa informático especialmente diseñado (desarrollado en Access 2019) para realizar cálculos de pruebas de paternidad y maternidad que dio como resultado 64.824 tríos M/H/PP. (Figura 7B)
- C. Cálculo del índice y probabilidad de paternidad por caso con resultado de no exclusión de paternidad.** La probabilidad de paternidad fue calculada como una probabilidad a posteriori Bayesiana (Bravo Aguilar, 2009; Fábrega Ruíz, 1998; Pena & Chakraborty, 1994) empleando un programa informático comercialmente disponible (Aplicación A.P.S: Manejo y Análisis de Paternidad y Maternidad, Bioanalítica, Bogotá, Colombia). El programa es utilizado para el análisis de paternidad de rutina en el Laboratorio de Genética de Poblaciones e Identificación de la Universidad Nacional de Colombia – Sede Bogotá así como por al menos otros 3 laboratorios nacionales que se encuentran habilitados y acreditados por ONAC en la norma

ISO/IEC 17025:2017 y que participan anualmente en el Ejercicio de Intercomparación para investigación de polimorfismos de ADN del Grupo de Habla Española y Portuguesa de la *International Society for Forensic Genetics* (GHEP-ISFG). Es por esto por lo que el cálculo de la probabilidad de paternidad realizado es recomendado por las directrices colombianas y del Grupo de Habla Española y Portuguesa de la *International Society for Forensic Genetics* (GHEP-ISFG) (Figura 7C).

Este programa se empleó como herramienta base para realizar las múltiples combinaciones de madre e hijo como dúo constante y presunto padre como elemento variable; dado a que más allá de tener la capacidad de realizar los diferentes casos de paternidad como operaciones de manera simultánea es un sistema de gestión en base de datos, que almacena y organiza dentro de una sólida estructura entidad-relación la información que se produce en este experimento de análisis genético; evitando de esta forma que se generen duplicados y se generen errores en las iteraciones. Teniendo en cuenta los procedimientos anteriormente descritos y diagramados en la Fig. 7., se realizaron modificaciones en las relaciones de la base de datos que permitieran desarrollar simultáneos casos en una misma corrida, y también permitieran correr diferentes casos dejando como variable al presunto padre y no al dúo madre e hijo.

- D. ***Conteo de casos trío con no exclusión de paternidad con el mismo presunto padre:*** Después de analizar los 64.824 casos trío de paternidad con diferentes 150 padres, se realizó una consulta que permitiera determinar si se encontraba el mismo presunto padre en varios dúos (madre e hijo) en donde el resultado de ese trío no correspondiera a una exclusión de la paternidad (Figura 7D)

3.3 Resultados.

Se testeó el algoritmo modificado de la aplicación A.P.S. (Aplicación para manejo y análisis en casos de paternidad) de BioAnalítica, para corridas simultáneas de varios casos de paternidad, utilizando los datos poblacionales registrados por el grupo de Genética de Poblaciones e Identificación del Instituto de Genética de la Universidad Nacional de Colombia (GPI-IGUN).



Región (R)

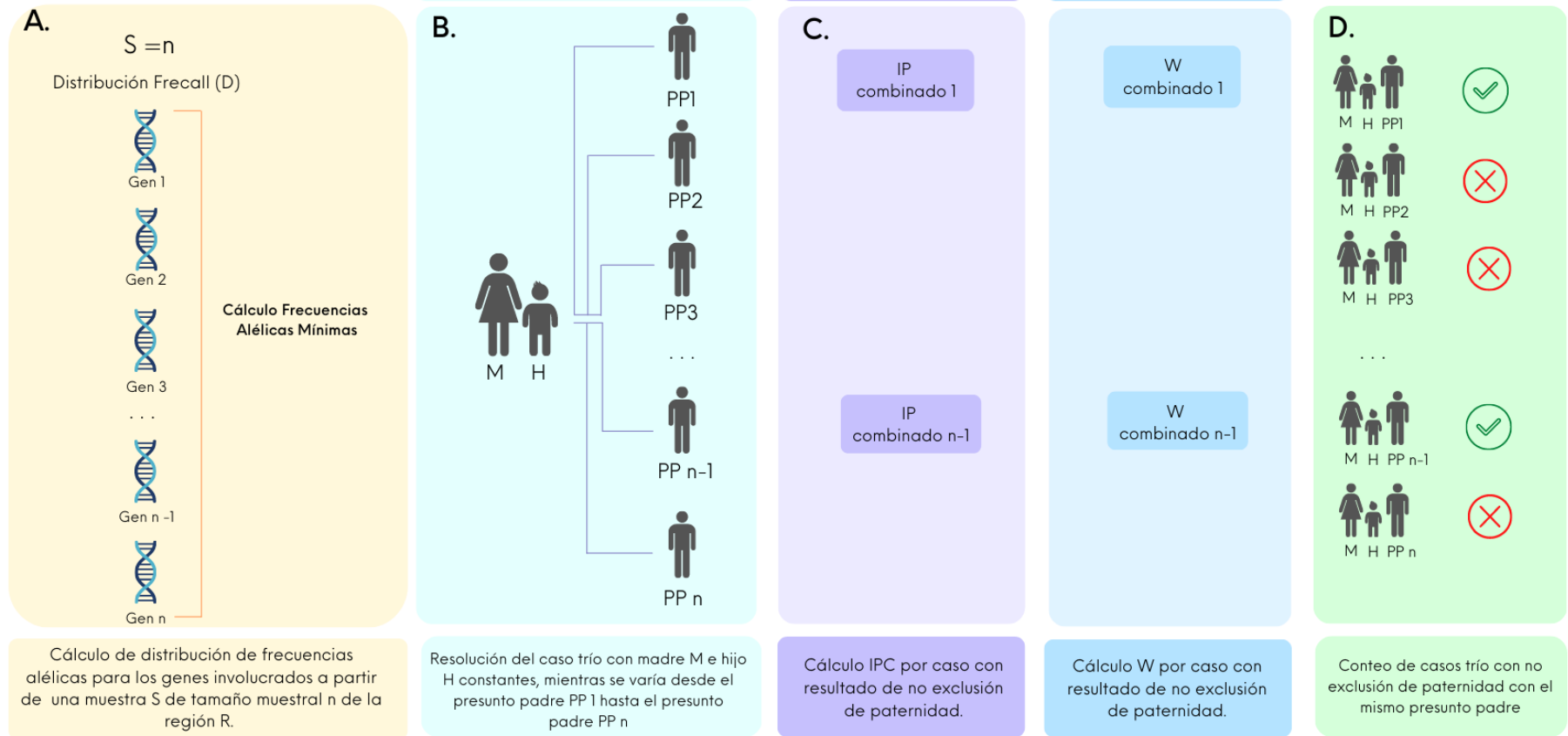


Figura 7 Diagrama del experimento realizado en los 64.824 casos trío de paternidad con diferentes 150 padres. **A. Cálculo de distribución de frecuencias alélicas para la población de referencia:** Bogotá fue la población de referencia empleada. **B. Resolución del caso trío con madre (M) e hijo (H) constantes.** Se tomaron 296 dúos madre e hijo sin exclusión materna que fueron analizados con 219 presuntos padres (PP), desde el padre 1 hasta el n ; analizándose desde el caso 1 al n formado por el dúo madre e hijo n con el presunto padre n . **C. Cálculo del índice y probabilidad de paternidad por caso con resultado de no exclusión de paternidad:** La probabilidad de paternidad fue calculada como una probabilidad a posteriori Bayesiana. **D. Conteo de casos trío con no exclusión de paternidad con el mismo presunto padre:** Se realizó una consulta que permitiera determinar si se encontraba el mismo presunto padre en varios dúos (madre e hijo) en donde el resultado de ese trío no correspondiera a una exclusión de la paternidad.

3.3.1 Comparaciones de hijos y presuntos padres no relacionados teniendo en cuenta la presencia de la madre biológica.

Se tuvieron en cuenta 296 dúo madre e hijo obtenidos a partir de la base datos del Genética de Poblaciones e Identificación del Instituto de Genética de la Universidad Nacional de Colombia (GPI-IGUN) que fueron analizados con 219 presuntos padres, para un total de 64.824 pruebas de paternidad tríó realizadas de forma simultánea.

Del total de 64.824 pruebas tríó generadas, se pudo confirmar el padre putativo propuesto en 156 pruebas, es decir, el 0.2412 % de las pruebas tríó dieron como resultado una no exclusión de la paternidad. Para 1 dúo M-H (es decir, 0.3378 %) se detectó uno o más hombres no emparentados con cero o tres STR coincidentes. El número de "presuntos padres" adicionales (hombres con al menos 12 coincidencias con niños no emparentados) para cada dúo M-H se muestra en la Tabla 3. Se encontraron 8 dúos M-H con 1 presunto padre adicional y 1 dúo M-H con dos padres adicionales.

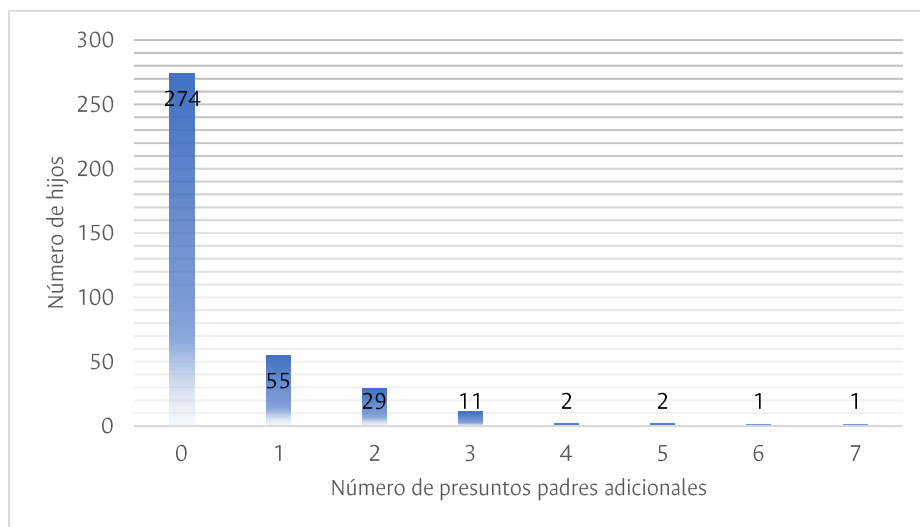
Tabla 3 Número de "presuntos padres adicionales" (hombres con menos de tres exclusiones y al menos 12 coincidencias en los loci STR) para los hijos de un dúo (M-H).

Número de Presuntos padres adicionales	Número de casos dúo (M-H)
1	8
2	1

3.3.1 Comparaciones de hijos y presuntos padres no relacionados sin la presencia de la madre biológica.

Para evaluar el efecto de la ausencia de la madre se realizó el mismo procedimiento anterior, realizando pruebas dúo Presunto Padre – Hijo en cuestión (PP - H). Se tuvieron en cuenta 298 hijos obtenidos a partir de la base datos del Genética de Poblaciones e Identificación del Instituto de Genética de la Universidad Nacional de Colombia (GPI-IGUN) que fueron analizados cada uno con 219 presuntos padres distintos, para un total de 65.262 pruebas dúo realizadas simultáneamente.

Del total de 65.262 pruebas dúo generadas, se pudo confirmar el padre putativo propuesto en 375 pruebas, es decir, el 0,5746 % de las pruebas trío dieron como resultado una no exclusión de la paternidad. El número de "presuntos padres" adicionales (hombres con al menos 12 coincidencias con niños no emparentados) para cada hijo se muestra en la Gráfica 14. Se encontraron 55 Hijos con 1 presunto padre adicional; 29 con 2 P-P adicionales; 11 con 3 P-P adicionales; 2 hijos con 4 y 5 P-P adicionales y 1 hijo con 6 y 7 P-P adicionales, respectivamente.



Gráfica 14 Número de "presuntos padres adicionales" (hombres con menos de tres exclusiones y al menos 12 no exclusiones en los 15 loci STR analizados) para los hijos con los que fueron analizados.

En la Tabla 4. se muestra la distribución de probabilidad de paternidad de las 375 prueba dúo analizadas de las cuales no hubo una exclusión biológica de paternidad; aun sabiendo que no eran dúos emparentados.

Tabla 4 Distribución de probabilidades de paternidad en pares hijo/hombre emparentados y no emparentados con 3 o menos exclusiones.

Probabilidad de Paternidad	Número de dúo presunto padre / hijo (n=375)
0,999 – 0,9999	287
0,99 – 0,9989	88
0,95 – 0,99	0
0,90 – 0,95	0
<0,90	0

3.4 Discusiones.

La población de Bogotá, ha sido empleada como modelo para el análisis de tamaño de muestra (Usaquén Martínez, 2012) y variabilidad de polimorfismos STRs en previas investigaciones (Benítez-Páez & Reyes, 2003; Bravo et al., 2001; Durán et al., 2003; Mogollón Olivares, 2017; Paredes et al., 2003; Pereira et al., 2012; Rey et al., 2003; K. M. Rojas et al., 2011; W. Rojas et al., 2010) ya que cuenta con una población de aproximadamente 7,9 millones de habitantes en la que se da un activo proceso de migración positiva desde todo el país (Usaquén Martínez, 2012).

En este estudio, se determinó cuántas inclusiones de paternidad erróneas podrían detectarse en Bogotá, la sexta ciudad más poblada de Latinoamérica con un área geográfica de 1.775 km². Teniendo en cuenta que, en los últimos 10 años, el creciente número de paternidades en disputa en Colombia ha involucrado al menos a 65.400 padres de diferentes orígenes étnicos y ancestralidades, se analizaron 64.824 pruebas de paternidad trío de forma simultánea en donde se encontraron 8 padres que podían ser los presuntos padres de 1 o dos niños adicionales; para un total de 19 tríos M-PP e hijo con probabilidades de paternidad mayores a 0,9993.

Otras investigaciones se han basado en analizar la implicación de la ausencia de madre en los cálculos de probabilidad de paternidad (J.A. Thomson, 2001; Poetsch et al., 2006; von Wurmb-Schwark et al., 2006; Wurmb-schwark et al., 2015). Para el presente estudio, el eliminar de la prueba a la madre aumentó 5,3158 veces el número de presuntos padres adicionales no emparentados, resultado que no difiere con otros estudios realizados, dado que la probabilidad de una **falsa no exclusión de paternidad** es mayor cuando se analiza un presunto padre y un hijo en cuestión (caso dúo: PP - H) que cuando hay una madre confirmada adicional (caso trío: PP – H- H) (Aguiar et al., 2021; Lee et al., 2013).

De Ungria y colaboradores, encontraron de 5253 dúos PP -H de la población Filipina, 195 dúos PP -H con tres exclusiones o menos al analizar siete loci STR (de Ungria et al., 2002); esta alta tasa de coincidencias en perfiles genéticos de PP e hijos no emparentados debe considerarse al tamaño poblacional de Filipina y al grado de consanguinidad que puede presentarse en su población (Poetsch et al., 2006); a diferencia de la población Bogotana donde se da un activo proceso de migración positiva desde todo el país.

Sin embargo, el tomar una población con alta variabilidad genética como Bogotá, nos hace pensar en otras poblaciones Colombianas más conservadas como las previamente estudiadas como Wayuu, San Andrés y Amazonas; donde se debían realizar cálculos de frecuencias alélicas mínimas para poder realizar los cálculos de índice y probabilidad de paternidad. Las probabilidades de paternidad para los presuntos padres biológicos adicionales fueron en todos los casos mayores a 0,99 y para el 76,5333% fueron mayores a 0,9990 (Tabla 4), en 122 de estos casos presentaron probabilidades mayores a 0,9999 indicando que se puede encontrar un presunto padre adicional en al menos 10.000 hombres.

Del total de 65.262 pruebas dúo generadas, se pudo confirmar el padre putativo propuesto en 375 pruebas, es decir, el 0,5746 % de las pruebas trío dieron como resultado una no exclusión de la paternidad en casos no emparentados. A pesar de este bajo porcentaje, este resultado cuestiona el significado de la probabilidad de paternidad en los casos donde se excluye a la madre, cuando no se conocen pruebas no biológicas complementarias, es decir que haya habido algún tipo de relación entre el padre y madre del hijo en cuestión. Especialmente declaraciones verbales como “paternidad biológica no excluida” deben seguir usándose con mucho cuidado en las investigaciones de padres solteros/hijos.

3.5 Conclusiones.

Las pruebas de paternidad en donde participan presunto padre (PP) e hijo (H) son solicitadas en los laboratorios que realizan pruebas de paternidades si el PP ha registrado ya a el niño en cuestión; y más si implica que no sea necesaria la presencia de la madre en dicha prueba. En este estudio, se determinó cuántas inclusiones de paternidad erróneas podrían detectarse en Bogotá, la sexta ciudad más poblada de Latinoamérica con un área geográfica de 1.775 km²; al comparar los resultados empíricos de 296 dúos madre (M) e hijo (H) y 219 presuntos padres (PP) (con 15 marcadores microsatélites); y para las pruebas dúo con 298 hijos y y 219 presuntos padres (PP) (con 15 marcadores microsatélites).

Para los 64.824 tríos M/H/PP se encontraron menos de tres STR excluyentes en 156 pruebas, es decir, el 0,2412 % de las pruebas trío dieron como resultado una no exclusión de la paternidad. Para los 65.262 casos dúo PP – H no emparentados, se pudo confirmar el padre putativo propuesto en 375 pruebas, es decir, el 0,5746 % de las pruebas dúo dieron como resultado una no exclusión de la paternidad. Se encontraron 55 Hijos con 1 presunto padre adicional; 29 con 2 P-P adicionales; 11 con

3 P-P adicionales; 2 hijos con 4 y 5 P-P adicionales y 1 hijo con 6 y 7 P-P adicionales, respectivamente. Las probabilidades de paternidad oscilaron entre 0,9974 y 99,999%, lo que muestra que ***una alta probabilidad de paternidad (> 0,99) no necesariamente indica una paternidad biológica probada o no excluida*** en casos sin madre.

Si se excluye a la madre de la prueba podemos hallar coincidencias en perfiles genéticos del presunto padre e hijo no emparentado; y estas coincidencias serán mayores en poblaciones con menor diversidad genética o muy conservadas, por lo que se sugiere que para evitar los posibles sesgos que se pueden presentar en los casos de paternidad sin madre utilizando solo análisis STR que el laboratorio pueda contar con la capacidad de aumentar el número de loci STR autosómicos en la prueba al tener resultados poco concluyentes, soporte sus resultados con tipificación de STR X y Y, y haga un estudio previo de su población donde pueda conocer las medidas de diversidad genética de la población de referencia que emplea así como los alelos más comunes.

3.6 Referencias.

- Aguilar, V. R. C., de Castro, A. M., Pinto, L. M., Ferreira, A. C. S., dos Santos, E. V. W., & Louro, I. D. (2021). Assessing false paternity risk in simulated motherless cases from more than 20 000 real exclusion trios. *Transfusion*, 61(3), 678–681. <https://doi.org/10.1111/trf.16153>
- Anderson, K. G. (2006). How well does paternity confidence match actual paternity? Evidence from worldwide nonpaternity rates. *Current Anthropology*, 47(3), 513–520. <https://doi.org/10.1086/504167>
- Benítez-Páez, A., & Reyes, H. O. (2003). Allelic frequencies at 12 STR loci in Colombian population. *Forensic Science International*, 136(1–3), 86–88. [https://doi.org/10.1016/S0379-0738\(03\)00220-2](https://doi.org/10.1016/S0379-0738(03)00220-2)
- Bravo Aguilar, M. L. J. (2009). Investigación de la Paternidad Biológica. In *La verdad genética de la paternidad* (I, pp. 45–80). Universidad de Antioquia.
- Bravo, M. L., Moreno, M. A., Builes, J. J., Salas, A., Lareu, M. v., & Carracedo, A. (2001). Autosomal STR genetic variation in negroid Chocó and Bogotá populations. *International Journal of Legal Medicine*, 115(2), 102–104. <https://doi.org/10.1007/s004140100223>
- Brinkmann, B., Pfeiffer, H., Schürenkamp, M., & Hohoff, C. (2001). The evidential value of STRs: An analysis of exclusion cases. *International Journal of Legal Medicine*, 114(3), 173–177. <https://doi.org/10.1007/s004140000174>
- Correa Rubio, C. N., & Sanchez Rodriguez, P. S. (2021). *La paternidad evadida en Colombia: El derecho a la filiación de los menores versus el derecho a la intimidad y la autonomía de la voluntad del presunto padre*. Universidad Cooperativa de Colombia - Sede Ibagué, Espinal.

- de Ungria, M. C. A., Frani, A. M., Magno, M. M. F., Tabbada, K. A., Calacal, G. C., Delfin, F. C., & Halos, S. C. (2002). Parentage Testing Evaluating DNA tests of motherless cases using a Philippine genetic database. *TRANSFUSION*, 954–957.
- Durán, R., Zarante, I., Acevedo, M. L., Villegas, M. R., Salazar, J., Bocanegra, B. Y., & Bernal, J. (2003). Allelic frequency of six STR loci in five Colombian cities. *Journal of Forensic Sciences*, 48(4), 887. <http://www.ncbi.nlm.nih.gov/pubmed/12877314>
- el Andari, A., Daouk, A., & Mansour, I. (2018). Effect of DNA Profile Size, Reference Population Database, and Parents Availability on Parentage Testing in Consanguineous and Endogamous Populations: The Lebanese Case. *Journal of Forensic Research*, 09(04). <https://doi.org/10.4172/2157-7145.1000425>
- Fábrega Ruíz, C. F. (1998). *Pruebas Biológicas de Paternidad. Aspectos científicos y jurídicos de las mismas*.
- Fernandes, A. T., Gonçalves, R., & Brehm, A. (2004). Databases: The real importance in paternity testing. *International Congress Series*, 1261(C), 463–464. [https://doi.org/10.1016/S0531-5131\(03\)01766-7](https://doi.org/10.1016/S0531-5131(03)01766-7)
- Houck, M. M. (2015). *Forensic Biology (Advanced F)*. Elsevier Inc.
- J.A. Thomson, K. L. A. V. P. M. N. B. J. I. H. W. P. G. D. (2001). Analysis of disputed single-parent/child and sibling relationships using 16 STR loci. *Int. J. Legal Med*, 115, 128–134.
- Jacewicz, R., Berent, J., Prośniak, A., Dobosz, T., Kowalczyk, E., & Szram, S. (2004). Non-exclusion paternity case with a triple genetic incompatibility. *International Congress Series*, 1261(C), 511–513. [https://doi.org/10.1016/S0531-5131\(03\)01649-2](https://doi.org/10.1016/S0531-5131(03)01649-2)
- Lee, J. C. I., Tsai, L. C., Chu, P. C., Lin, Y. Y., Lin, C. Y., Huang, T. Y., Yu, Y. J., Linacre, A., & Hsieh, H. M. (2013). The risk of false inclusion of a relative in parentage testing - an in silico population study. *Croatian Medical Journal*, 54(3), 257–262. <https://doi.org/10.3325/cmj.2013.54.257>
- Lique Gutiérrez, J. A. (2019). Estudio de las relaciones de parentesco. In M. C. Crespillo Márquez & P. A. Barrio Caballero (Eds.), *Genética Forense: Del laboratorio a los tribunales* (I, pp. 351–381). Díaz de Santos.
- Mickey, M. R., Gjertson, D. W., & Terasaki, P. I. (1986). Empirical Validation of the Essen-Moller Probability of Paternity. In *Am J Hum Genet* (Vol. 39).
- Mogollón Olivares, F. (2017). *Variabilidad y diversidad genética de la población humana Colombiana en cuatro regiones biogeográficas mediante marcadores autosómicos STR*.
- Moroni, R., Gasbarra, D., Arjas, E., Lukka, M., & Ulmanen, I. (2011). Effects of Reference Population and Number of STR Markers on positive evidence in Paternity Testing. *Journal of Forensic Research*, 02(02). <https://doi.org/10.4172/2157-7145.1000119>
- Ortega Torres, J., Rueda, O. L., & Jaime, L. A. (2015). *DOCUMENTO GUÍA PRUEBAS DE ADN PARA INVESTIGACIÓN DE PATERNIDAD Y/O MATERNIDAD (Versión actualizada-Enero-2015)*.
- Paredes, M., Galindo, A., Bernal, M., Avila, S., Andrade, D., Vergara, C., Rincón, M., Romero, R. E., Navarrete, M., Cárdenas, M., Ortega, J., Suarez, D., Cifuentes, A., Salas, A., & Carracedo, Á. (2003). Analysis of

- the CODIS autosomal STR loci in four main Colombian regions. *Forensic Science International*, 137(1), 67–73. [https://doi.org/10.1016/S0379-0738\(03\)00271-8](https://doi.org/10.1016/S0379-0738(03)00271-8)
- Pena, S. D. J., & Chakraborty, R. (1994). Paternity testing in the DNA era. *Trends in Genetics*, 10(6), 204–209.
- Pereira, R., Phillips, C., Pinto, N., Santos, C., dos Santos, S. E. B., Amorim, A., Carracedo, Á., & Gusmão, L. (2012). Straightforward inference of ancestry and admixture proportions through ancestry-informative insertion deletion multiplexing. *PLoS ONE*, 7(1). <https://doi.org/10.1371/journal.pone.0029684>
- Poetsch, M., Lüdcke, C., Repenning, A., Fischer, L., Mályusz, V., Simeoni, E., Lignitz, E., Oehmichen, M., & von Wurmb-Schwark, N. (2006). The problem of single parent/child paternity analysis-Practical results involving 336 children and 348 unrelated men. *Forensic Science International*, 159(2–3), 98–103. <https://doi.org/10.1016/j.forsciint.2005.07.001>
- Rey, M., Gutiérrez, A., Schroeder, B., Usaquén, W., Carracedo, A., Bustos, I., & Giraldo, A. (2003). Allele frequencies for 13 STR's from two Colombian populations: Bogotá and Boyacá. *Forensic Science International*, 136(1–3), 83–85. [https://doi.org/10.1016/S0379-0738\(03\)00221-4](https://doi.org/10.1016/S0379-0738(03)00221-4)
- Rojas, K. M., Roa, M., Briceño, I., Guaneme, C., & Gómez, A. (2011). Polimorfismos de 17 marcadores STR del cromosoma-Y en una muestra poblacional del altiplano cundiboyacense. *Colombia Medica*, 42(1), 88–97. <https://www.scopus.com/inward/record.uri?eid=2-s2.0-79953011667&partnerID=40&md5=fe2c2fa3b361573147d85f9be5f33bf3>
- Rojas, W., Parra, M. V., Campo, O., Caro, M. A., Lopera, J. G., Arias, W., Duque, C., Naranjo, A., García, J., Vergara, C., Lopera, J., Hernandez, E., Valencia, A., Caicedo, Y., Cuartas, M., Gutiérrez, J., López, S., Ruiz-Linares, A., & Bedoya, G. (2010). Genetic make up and structure of Colombian populations by means of uniparental and biparental DNA markers. *American Journal of Physical Anthropology*, 143(1), 13–20. <https://doi.org/10.1002/ajpa.21270>
- Sánchez, D., González-Andrade, F., Bolea, M., & Jarreta, B. M. (2008). False inclusion in a deficient paternity case with two alleged fathers. *Forensic Science International: Genetics Supplement Series*, 1(1), 525–527. <https://doi.org/10.1016/j.fsigss.2007.10.105>
- Tagliabracci, A. (2010). *Introduzione Alla Genetica Forense*. <https://doi.org/10.1007/978-88-470-1512-8>
- Thomson, J. A., Pilotti, V., Stevens, P., Ayres, K. L., & Debenham, P. G. (1999). Validation of short tandem repeat analysis for the investigation of cases of disputed paternity. In *Forensic Science International* (Vol. 100).
- Tillmar, A. (2010). *Populations and Statistics in Forensic Genetics* (Issue 1175).
- Usaquén Martínez, W. (2012). *Validación y consistencia de información en estudios de diversidad genética humana a partir de marcadores microsatélites* [Universidad Nacional de Colombia]. <http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:No+Title#0>

von Wurmb-Schwark, N., Mályusz, V., Simeoni, E., Lignitz, E., & Poetsch, M. (2006). Possible pitfalls in motherless paternity analysis with related putative fathers. *Forensic Science International*, *159*(2–3), 92–97. <https://doi.org/10.1016/j.forsciint.2005.07.015>

Wurmb-schwark, N. von, Podruks, E., Schwark, T., Göpel, W., Fimmers, R., & Poetsch, M. (2015). *About the power of biostatistics in sibling analysis — comparison of empirical and simulated data*. 1201–1209. <https://doi>.

Capítulo 4: Genética poblacional y forense: Herramientas para la reconstrucción histórica y social en la era del posconflicto colombiano.

La información genética dispuesta en grandes bases de datos pertenece en su mayoría a estudios del ámbito forense en el cual se han tipificado gran cantidad de individuos distribuidos en todo el mundo con un set pequeño (~20) de marcadores tipo STRs. Estas bases de datos han sido desarrolladas para responder preguntas relacionadas con la identificación individual y con poder discriminar a un individuo de una población particular (Kanitz, Guillot, Antoniazza, Neuenschwander, & Goudet, 2018; Poloni, Currat, & Silva, 2012); de esta forma se analiza y estudia la variación dentro de la población, más no la diferencia entre las poblaciones (Jobling, 2022). A pesar de esto, pueden llegar a ser informativas porque contienen muestras dispersas globalmente y por cada población estudiada han tipificado un número alto de individuos (Rosenberg et al., 2005).

Debido a que se estudia la variación intrapoblacional, para evitar el sesgo gracias a las coincidencias entre cotejo de perfiles genéticos, se calculan las frecuencias alélicas empleadas en poblaciones de referencia que se encuentren en estado de panmixia (Weir, 1992). Este supuesto plantea dos interrogantes ampliamente debatidos en la década de 1990 por Lewontin y Daniel Hartl en las llamadas “*DNA fingerprinting wars*”: ¿Cuál población de referencia debo emplear para un caso particular? ¿Puede la subestructura de la población de referencia invalidar las suposiciones hechas en los cálculos de probabilidad de coincidencia aleatoria (*RMP: random mating probability*)? (Jobling, 2022; Lander & Budowle, 1994; Lewontin, 1994; Lewontin & Hartl, 1991).

Tal como observamos en el capítulo anterior en donde se evidenció más de un posible padre para un hijo en cuestión en la misma población, o en el caso de un delito donde la única fuente de información disponible sobre el sospechoso es su perfil de ADN; la elección de la población de referencia para calcular el RMP se vuelve particularmente relevante. Es aquí donde cabe la importancia del cálculo de alelos raros y frecuencias mínimas; aún más si se evidencia que existen grandes diferencias en las frecuencias de alelos entre poblaciones aisladas o muy conservadas (Budowle et al., 1996; Jobling, 2022; Lewontin & Hartl, 1991).

Estas recomendaciones y aclaraciones se acotaron en el informe del Consejo Nacional de Investigación de los Estados Unidos (National Research Council Committee on DNA Technology in Forensic Science., 1996) que sugiere que para las poblaciones de EE. UU., se emplee un valor conservador de $\vartheta=0,01$ (al menos un orden de magnitud más alto que los valores medidos empíricamente) y para "algunas poblaciones pequeñas y aisladas" un $\vartheta=0,03$. A pesar de los esfuerzos realizados para responder las mayores preocupaciones de Lewontin y Harlt; siguen sin resolverse los problemas subyacentes a la escogencia adecuada de poblaciones de referencia y la subestructura de la población que vuelven a salir a luz con cada nuevo desarrollo en tecnología forense (Brinkmann et al., 2001; J.A. Thomson, 2001; Jobling, 2022; Moroni et al., 2011). Por otro lado, en genética forense a diferencia de la poblacional, los perfiles genéticos suelen ser parciales (faltan un conjunto completo de loci o alelos) dado a la degradación de la muestra o la dificultad de obtención de la misma, lo que puede aumentar los *RMP* y dificultar la interpretación de resultados (Fernandes et al., 2004; Lewontin, 1994; Pena & Chakraborty, 1994; Poetsch et al., 2006; von Wurmb-Schwark et al., 2006).

Otro problema que se presenta respecto a la escogencia de poblaciones de referencia es la forma arbitraria en la que se catalogan y seleccionan individuos dentro de estas. Es probable que el genetista de poblaciones actual use un sistema de clasificación basado en la afiliación etnolingüística, la geografía y la autodefinición (Ansari-Pour et al., 2016; Bedoya et al., 2006; Builes et al., 2013; Garavito et al., 2015; Homburger et al., 2015; Jobling, 2022; Mogollon Olivares et al., 2020; Urbano et al., 2016). En la práctica forense, por el contrario, el análisis se lleva a cabo dentro de los marcos sociopolíticos de los sistemas nacionales de justicia penal que están basados en sus propios censos de población y, a menudo, se remontan al pasado al emplear términos coloniales como blancos – mestizos, negroides o los clasificadores raciales del siglo XVIII de Blumenbach, al emplear categorías como caucásicos, personas de color, hispanos o mongoloides (Jobling, 2022; Mogollon Olivares et al., 2020).

Adentrándonos en los estudios en genética de poblaciones, en los que, a razón del costo de las investigaciones, que también se remiten a grupos poblacionales muy específicos, el tamaño de muestra suele ser pequeño debido a la selección a priori de individuos, por lo que la representación general de una población puede ser deficiente. Dado a que el campo de la genética forense genera una cantidad basta de datos poblacionales de polimorfismos tipo STR en muestras distribuidas

globalmente, se ha estudiado el poder de estos set de datos para responder preguntas relacionadas a la evolución humana y su diversidad, teniendo en cuenta dos tipos de recursos: las frecuencias alélicas disponibles en las bases de datos y datos genotípicos que se pueden encontrar en pocas bases de datos o en artículos científicos (Bentayebi, Abada, Izhmad, & Amzazi, 2014; Callegari-Jacques, Tarazona-Santos, Gilman, Herrera, Cabrera, Dos Santos, et al., 2011; Houck, 2015; Khubrani, Wetton, & Jobling, 2019; Poloni et al., 2012; Sun et al., 2013). Sin embargo, se sigue generando la misma duda, ¿estás categorías se han formado de forma arbitraria o fueron pensadas en un marco etnolingüístico, geográfico y cultural?

En Colombia, la caracterización de poblaciones de nuestro territorio ha sido realizada gracias al reporte de las frecuencias alélicas y estadísticos forenses indispensables en el qué hacer de los laboratorios de filiación y de pruebas de paternidad. En adición, estos reportes han sido realizados por los laboratorios empleando las muestras de participantes en las pruebas de identificación que realizan; y por los grupos de investigación de los laboratorios de universidades que han podido caracterizar a nivel regional a lo largo de los años sus comunidades de interés cercanas. Los estudios que son dirigidos a grupos poblaciones muy específicos han sido llevados a cabo con mucho esfuerzo por proyectos que han podido abarcar tamaños de muestra pequeños debido a la dificultad para solicitar permisos, el poco tiempo de muestreo o la entrada a las comunidades por rutas de difícil acceso que encarecen el proceso de muestreo. Además de esto, los investigadores deben realizar una la selección a priori de individuos y determinar de acuerdo con su pregunta de investigación e hipótesis qué participantes pueden incluir en los ensayos moleculares dado al limitado presupuesto para realizar un número elevado de tipificaciones.

Lo que nos plantea este escenario en el que se tienen un conjunto de datos con una distribución nacional conglomerada por regiones o departamentos, que además en su mayoría es obtenida a través de casos de pruebas de paternidad o proyectos enfocados en un marco genético poblacional es que se hace indispensable unir esfuerzos entre las dos líneas de análisis. Y de esta forma poder constituir una fuente importante de información sobre la diversidad genética humana de nuestro país, a pesar del número relativamente bajo de marcadores tipificados.

4.1 *Poder de los muestreos dirigidos y a conveniencia en genética de poblaciones y su aplicación en forense.*

Al analizar las poblaciones desde un contexto histórico, demográfico y cultural previamente antes de adentrarse al muestreo en campo permite conocer mejores formas de acercamiento con la población y determinar a priori individuos o comunidades claves para esclarecer preguntas de la formación de la comunidad, primeros pobladores y cómo se conforma la localidad de interés.

Este acercamiento previo permite formular preguntas adecuadas a realizar en la población a forma de encuesta o etnografías que dirijan a conocer patrones en la población, sus características sociales y demográficas (Mogollon Olivares et al., 2020; Usaquén Martínez, 2012). Esto da pie a que no sólo el investigador tome a la población como contexto de estudio, sino que lo involucre activamente en la investigación y el también participe como generador de conocimiento y sea quien cuente su historia a través de su vivencia de la mano con las hipótesis planteadas por metodologías genéticas.

Toda información adicional a la que se obtendrá después del campo; tendrá más informatividad que si no se tuviera junto con un contexto histórico y social. Además, permite que poblaciones conservadas sean estudiadas y caracterizadas desde múltiples aristas; y así a un futuro poder tener información valiosa en el marco forense que pueda ser difícil de obtener por el acceso a la población o la urgencia con lo que se necesite resolver un caso. Estas urgencias o preguntas particulares a un caso en comunidades muy conservadas o de difícil acceso pueden ser resueltas a futuro por el apoyo que la genética de poblaciones pudo ofrecer gracias a investigaciones previas.

4.2 *¿Por qué unir fuerzas e intenciones de investigación?*

Aún pasados 65 años de conflicto armado interno, Colombia sigue enfrentando los desafíos de un proceso de paz por medio de la planeación de estrategias desde frentes estatales, organizaciones no gubernamentales e instituciones académicas (Centro Nacional de Memoria Histórica, 2023). A la fecha de corte del 31 de marzo del 2023, el *Observatorio de Memoria y Conflicto (OMC)* ha registrado 269.306 muertes categorizadas en 11 modalidades de violencia en el marco del conflicto armado entre 1958 y 2022. A partir de la integración de 653 fuentes y 33.497 bases de datos y documentos,

el OMC ha contribuido al esclarecimiento histórico y al reconocimiento de la pluralidad de memorias; documentando las circunstancias de modo, tiempo y lugar de los hechos, los responsables y las víctimas del conflicto armado (Centro Nacional de Memoria Histórica, 2023).

En este contexto, todavía es buen momento para resaltar la importancia de contar con una base de datos completa y sólida de las frecuencias alélicas de marcadores genéticos utilizados en el área de identificación humana y filiación genética, que sean propios y representativos de las poblaciones diversas que habitan el territorio colombiano, en aras de apoyar los procesos de esclarecimiento histórico y reparación de víctimas. Con esta finalidad, en el 2016 se desarrollaron los Estándares forenses mínimos para la búsqueda de personas desaparecidas y la recuperación e identificación de cadáveres (Martínez Neira et al., 2017); documento que sirve como referente para el mejoramiento de los procesos técnicos al interior de las instituciones vinculadas con estas actividades. Dentro del marco establecido, se hace constante énfasis sobre la necesidad de que los *esfuerzos investigativos* se enfoquen en la *recuperación de información* suficiente sobre las víctimas y sobre la *rigurosidad* en la obtención de datos durante los diferentes estadios de las investigaciones.

Siguiendo estos estándares, se establece la necesidad de *confirmación y reporte de alelos nuevos, pérdidas alélicas, exclusiones aisladas, micro variantes, patrones trialélicos y mutaciones*. Además, se plantea que estas novedades deben incluirse en los cálculos estadísticos. Adicionalmente, para los cálculos de pruebas de identificación que arrojen resultados no excluyentes, deben ser validados estadísticamente empleando datos poblacionales de referencia para Colombia reposados en bases de datos avaladas por el Comité Interinstitucional de Genética del Instituto Nacional de Medicina Legal y Ciencias Forenses (Martínez Neira et al., 2017).

En este orden de ideas, no sólo es suficiente que durante los procesos de identificación y reparación de víctimas los laboratorios encargados de estos procesos se basen en recomendaciones científicas de buenas prácticas, que estén acreditados bajo normas internacionales de calidad o realicen ejercicios de Intercomparación; sino que resulta de suma importancia y premura la centralización de datos y la coordinación, depuración y operacionalización de los mismos en bases de datos relacionales para mantener su integridad e informatividad (Alonso Alonso, 2019). Esta tarea es un reto no sólo por el hecho de centralizar la información sino porque la construcción de la base de

datos debe ser una colaboración interinstitucional generada por la discusión interdisciplinar de antropólogos, científicos forenses, sociólogos, genetistas poblacionales, etc. (Revista Semana, 2019) que permita generar una estructura sólida que se base en el curado y limpieza de datos, estructura de la base de datos, escogencia de módulos, tablas y variables necesarias para la correcta interpretación y tratamiento estadístico (Alonso Alonso, 2019; García, 2019).

4.3. Limitaciones y conclusiones.

Es necesario mencionar que a pesar de todo este esfuerzo y trabajo conjunto; la reparación a víctimas y el esclarecer la verdad no podrá ser lograda totalmente por una perspectiva académica científica. Los procesos de identificación humana de víctimas en contextos de post conflicto de otros países nos enseñan no sólo a enriquecer las metodologías que se utilizan en la práctica forense, sino también nos invitan a considerar que se llegará a un punto en el que *no será posible identificar* todos los restos.

El Comité internacional de la Cruz Roja sobre los programas forenses humanitarios en Chipre y Kosovo, informa que al pasar el tiempo se presenta una disminución en el porcentaje de identificaciones genéticas exitosas, dado que en general la información más confiable o de fácil acceso se investiga primero y la menos confiable se deja para el final, los entierros clandestinos se vuelven más difíciles de encontrar y los recuerdos de los testigos se desvanecen (Mikellide, 2017). En el caso de Chipre, por ejemplo, tras 40 años de conflicto y tras 10 años de operaciones sistemáticas para la identificación, sólo el 31% de las personas desaparecidas fue identificada. En contraste, en Kosovo 13 años después del conflicto y tras 10 años de operaciones, fue posible resolver más de un 60% de los casos de personas desaparecidas. Es por esto, que estos procesos no deben ser cobijados sólo bajo la luz de una mirada académica y científica, sino que se debe plantear en conjunto de las familias y víctimas de conflicto otras metodologías que amparen un reconocimiento de quienes perpetraron los crímenes de lesa humanidad, ya sea el Estado u otros agentes mediante justicia especial y realización de actos de reparación social que no borren la historia sino que se pueda conocer como un hecho que no hay que repetir.

Lo anteriormente mencionado, sumado a un marco académico que fomente nuevas generaciones de genetistas forenses y poblacionales con herramientas sociales, antropológicas y científicas; permitirá

establecer equipos interdisciplinarios necesarios capaces de asumir los retos que depara la reconstrucción histórica y social en la era del posconflicto colombiano para que asuman una mayor y mejor participación en este proceso.

4.4. Referencias.

- Alonso Alonso, A. (2019). Las bases de datos de ADN de interés forense. In M. C. Crespillo Márquez & P. A. Barrio Caballero (Eds.), *Genética Forense: Del laboratorio a los tribunales* (I, pp. 425–443). Díaz de Santos. <https://www.editdiazdesantos.com/libros/9788490522134/Crespillo-Marquez-Genetica-forense.html>
- Ansari-Pour, N., Moñino, Y., Duque, C., Gallego, N., Bedoya, G., & Thomas, M. (2016). Palenque de San Basilio in Colombia: genetic data support an oral history of a paternal ancestry in Congo. *Proceedings of the Royal Society of London B: Biological Sciences*, 283(1827), 1–9. <https://doi.org/https://doi.org/10.1098/rspb.2015.2980>
- Bedoya, G., Montoya, P., Garcia, J., Soto, I., Bourgeois, S., Carvajal, L., Labuda, D., Alvarez, V., Ospina, J., Hedrick, P. W., & Ruiz-Linares, A. (2006). Admixture dynamics in Hispanics: A shift in the nuclear genetic ancestry of a South American population isolate. *Proceedings of the National Academy of Sciences*, 103(19), 7234–7239. <https://doi.org/10.1073/pnas.0508716103>
- Brinkmann, B., Pfeiffer, H., Schürenkamp, M., & Hohoff, C. (2001). The evidential value of STRs: An analysis of exclusion cases. *International Journal of Legal Medicine*, 114(3), 173–177. <https://doi.org/10.1007/s004140000174>
- Budowle, B., Monson, K. L., & Chakraborty, R. (1996). Estimating minimum allele frequencies for DNA profile frequency estimates for PCR-based loci. *International Journal of Legal Medicine*, 108, 173–176.
- Builes, J. J., Ospino, J. M., Manrique, A., Aguirre, D. P., Mendoza, L., Bravo, M. L. J., Pereira, R., & Gusmão, L. (2013). Genetic population data of 38 autosomal InDels for the Amerindian community Embera-Chami of Lapo, Antioquia-Colombia. *Forensic Science International: Genetics Supplement Series*, 4(1), 170–171. <https://doi.org/10.1016/j.fsigss.2013.10.088>
- Centro Nacional de Memoria Histórica. (2023). *Observatorio de Memoria y Conflicto*. <https://micrositios.centrodememoriahistorica.gov.co/observatorio/sievcac/fuentes/>
- Fernandes, A. T., Gonçalves, R., & Brehm, A. (2004). Databases: The real importance in paternity testing. *International Congress Series*, 1261(C), 463–464. [https://doi.org/10.1016/S0531-5131\(03\)01766-7](https://doi.org/10.1016/S0531-5131(03)01766-7)
- Garavito, G., Martínez, B., Builes, J. J., Aguirre, D., Mendoza, L., & Afanador, C. H. (2015). *Forensic Science International: Genetics Supplement Series Indels markers set and ancestry estimates in a population sample from Atlantic Department of Colombia*. 5, 177–178.

- García, O. (2019). Interpretación y valoración estadística de perfiles genéticos mezcla: Problemática asociada, repercusión, estrategias de mejora y evaluación de resultados. In M. C. Crespillo Márquez & P. A. Barrio Caballero (Eds.), *Genética Forense: Del laboratorio a los tribunales* (I, pp. 383–404). Díaz de Santos.
- Homburger, J. R., Moreno-Estrada, A., Gignoux, C. R., Nelson, D., Sanchez, E., Ortiz-Tello, P., Pons-Estel, B. A., Acevedo-Vasquez, E., Miranda, P., Langefeld, C. D., Gravel, S., Alarcón-Riquelme, M. E., & Bustamante, C. D. (2015). Genomic Insights into the Ancestry and Demographic History of South America. *PLoS Genetics*, *11*(12), 1–26. <https://doi.org/10.1371/journal.pgen.1005602>
- J.A. Thomson, K. L. A. V. P. M. N. B. J. I. H. W. P. G. D. (2001). Analysis of disputed single-parent/child and sibling relationships using 16 STR loci. *Int. J. Legal Med*, *115*, 128–134.
- Jobling, M. A. (2022). Forensic genetics through the lens of Lewontin: Population structure, ancestry and race. In *Philosophical Transactions of the Royal Society B: Biological Sciences* (Vol. 377, Issue 1852). Royal Society Publishing. <https://doi.org/10.1098/rstb.2020.0422>
- Lander, E. S., & Budowle, B. (1994). DNA fingerprinting dispute laid to rest. *Nature*, *371*, 735–738.
- Lewontin, R. C. (1994). Forensic DNA typing dispute. *Nature*, *372*, 398.
- Lewontin, R. C., & Hartl, D. L. (1991). Population Genetics in Forensic DNA Typing. *Science*, *254*, 1745–1750. www.sciencemag.org
- Martínez Neira, N. H., Riveros Dueñas, M. P., Valdés Moreno, C. E., Niño Izquierdo, C. I. V., García-Flno, C. A. del P., & Cuestas Gómez, Y. (2017). *Estándares forenses mínimos para la búsqueda de personas desaparecidas, y la recuperación e identificación de cadáveres*.
- Mikellide, M. (2017). Recovery and identification of human remains in post-conflict environments: A comparative study of the humanitarian forensic programs in Cyprus and Kosovo. *Forensic Science International*, *279*, 33–40. <https://doi.org/10.1016/j.forsciint.2017.07.040>
- Mogollon Olivares, F., Moncada Madero, J., Casas-vargas, A., Zea Montoya, S., Suárez Medellín, D., & Usaquén, W. (2020). Contrasting the ancestry patterns of three distinct population groups from the northernmost region of South America. *American Journal of Physical Anthropology*, e24130. <https://doi.org/10.1002/ajpa.24130>
- Moroni, R., Gasbarra, D., Arjas, E., Lukka, M., & Ulmanen, I. (2011). Effects of Reference Population and Number of STR Markers on positive evidence in Paternity Testing. *Journal of Forensic Research*, *02*(02). <https://doi.org/10.4172/2157-7145.1000119>
- National Research Council Committee on DNA Technology in Forensic Science. (1996). *The evaluation of forensic DNA evidence*.
- Pena, S. D. J., & Chakraborty, R. (1994). Paternity testing in the DNA era. *Trends in Genetics*, *10*(6), 204–209.
- Poetsch, M., Lüdcke, C., Repenning, A., Fischer, L., Mályusz, V., Simeoni, E., Lignitz, E., Oehmichen, M., & von Wurmb-Schwark, N. (2006). The problem of single parent/child paternity analysis-Practical

results involving 336 children and 348 unrelated men. *Forensic Science International*, 159(2–3), 98–103. <https://doi.org/10.1016/j.forsciint.2005.07.001>

Revista Semana. (2019). *Más de un tercio de los desaparecidos nunca serán encontrados*. Entrevista. <https://www.semana.com/nacion/articulo/francisco-etxeberria-hablo-con-semana-sobre-la-desaparicion-forzada-en-colombia/630531/>

Urbano, L., Portilla, E. C., Builes, J. J., Gusmão, L., & Sierra-Torres, C. H. (2016). Ancestral Genetic Composition of a human population from the Colombian Southwest using autosomal AIM-InDels. *Journal of Basic and Applied Genetics*, 27(2), 37–48. http://www.sag.org.ar/sitio/wp-content/uploads/2019/05/V.XXVII_2016_Issue2_30122012.pdf

Usaquén Martínez, W. (2012). *Validación y consistencia de información en estudios de diversidad genética humana a partir de marcadores microsatélites* [Universidad Nacional de Colombia]. <http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:No+Title#0>

von Wurmb-Schwark, N., Mályusz, V., Simeoni, E., Lignitz, E., & Poetsch, M. (2006). Possible pitfalls in motherless paternity analysis with related putative fathers. *Forensic Science International*, 159(2–3), 92–97. <https://doi.org/10.1016/j.forsciint.2005.07.015>

Weir, B. S. (1992). Review Population genetics in the forensic DNA debate. In *Proc. Natl. Acad. Sci. USA* (Vol. 89).