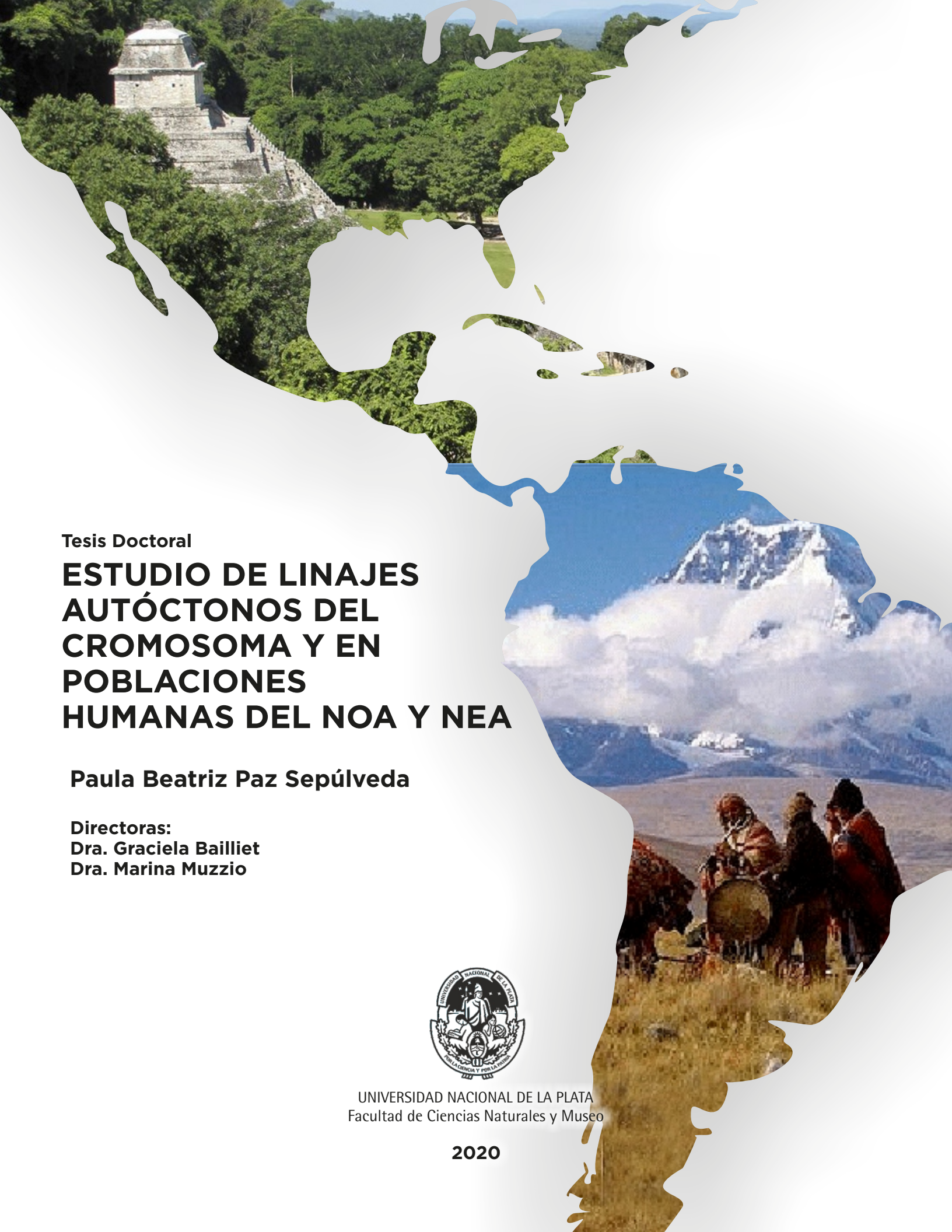


Tesis Doctoral

ESTUDIO DE LINAJES AUTÓCTONOS DEL CROMOSOMA Y EN POBLACIONES HUMANAS DEL NOA Y NEA

Paula Beatriz Paz Sepúlveda

Directoras:
Dra. Graciela Bailliet
Dra. Marina Muzzio

UNIVERSIDAD NACIONAL DE LA PLATA
Facultad de Ciencias Naturales y Museo

2020

ESTUDIO DE LINAJES AUTÓCTONOS DEL CROMOSOMA Y EN POBLACIONES HUMANAS DEL NOA Y NEA

Paula Beatriz Paz Sepúlveda

Directoras: Dra. Graciela Bailliet y Dra. Marina Muzzio

Universidad Nacional de La Plata
Facultad de Ciencias Naturales y Museo

Tesis doctoral
2020

A los pueblos originarios de América que lucharon y luchan por la libertad. Pongo a su disposición este trabajo como elemento para afirmar sus reclamos como autoridad que controla la disposición de las tierras que ocuparon sus antepasados.

AGRADECIMIENTOS

A la Universidad Nacional de La Plata por mantener una educación pública y gratuita. A la Facultad de Ciencias Naturales y Museo por posibilitarme acceder a una formación de posgrado. A los directivos del IMBICE, por ofrecerme un lugar para desarrollar este trabajo.

A mis directoras Graciela y Marina, por elegirme y darme la posibilidad de formar parte del Proyecto, por la calidez humana que me dieron, por estar siempre dispuestas y facilitarme todo lo que necesité para poder avanzar y terminar. También, por poner a mi disposición a Loki y Thor, sin estas fieles computadoras este trabajo no hubiera sido posible.

A todas las personas que participaron de este trabajo, quienes donaron su tiempo y sus muestras.

A Ezequiel y Jonathan de la Plataforma Bioinformática, por ayudarme siempre que necesité.

A Pille Hallast y Thomaz Pinotti, por ayudarme con los métodos por e-mail, desde grandes distancias.

A Leo, el diseñador, por su tiempo y paciencia en la creación de imágenes.

A las compañeras de GMP: Marisol, Daniela, Eliana, Belén, Camila, Mariela y Rita, con ustedes todo fue más lindo. A Jose Motti por sus sugerencias en este escrito. A becarixs y la gente del IMBICE.

A toda mi gran familia platense: Verónica, Cecilia, Guadalupe, Camila, Lara, Santiago, Lola, Martín, Esteban, Juan, Charo, Javier, Andrea, Benjamín, Manuel, tía Ana, tío Mario; por tantos momentos compartidos, por el cariño y lo divertido que es siempre estar con ustedes.

A todas las amistades dispersas por el mundo, a mi gran familia sanjuanina y la parte que lleva el linaje nativo americano: tía Ana, Hugo, Guillermo, Eduardo, abuela; por su cariño y compañía.

A mis amigas Iriel y Analía, el cariño de ustedes siempre me acompaña, más allá de la distancia.

A Sofía, por ser parte de mi familia, por tantos momentos compartidos, por el cuidado y amor dado a Simón mientras escribía este trabajo.

A mi amiga Coni, por ser parte de mi familia, por la compañía de estos años, por ser incondicional y por sus aportes inmensos en la realización de este trabajo.

A mis padres, por tanto amor, por ser incondicionales, por cuidar a Simón mientras escribía este trabajo. A mi hermana y hermano por estar, por alentarme siempre a avanzar y terminar.

A Simón por su amor, compañía y presencia.

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

ÍNDICE DE TABLAS ADJUNTAS

RESUMEN

El poblamiento temprano de América ha sido foco de debate incesante durante más de 100 años. Varios sitios arqueológicos han encontrado evidencias de ocupación humana temprana en Mesoamérica y Sudamérica que datan de 18000 años y antes. El cromosoma Y humano, posee el tramo más largo de ADN no recombinante de todo el genoma humano y es transmitido por completo de padres a hijos, contiene un registro de la historia del linaje paterno siendo utilizado como una herramienta altamente informativa para investigar la historia de las poblaciones humanas. Para realizar un estudio sobre las relaciones filogenéticas de linajes nativos americanos e inferir sobre el poblamiento de América, realizamos un análisis del haplogrupo Q del cromosoma Y, el cual es el único haplogrupo panamericano y representa prácticamente todos los linajes nativos americanos en Mesoamérica y Sudamérica.

Se construyó un árbol filogenético calibrado para el haplogrupo Q en base a 102 secuencias completas de cromosomas Y, de las cuales, 13 secuencias son nuevas presentadas en este trabajo. Se definieron 17 sub-haplogrupos Q. De estos, 13 sub-haplogrupos son específicos de nativos americanos y pertenecen a Q-Z780 y Q-M3 (incluye a Q-M848). Una de las secuencias realizadas en este trabajo se identificó dentro del sub-linaje Q-M346* (el sufijo "*" está indicando en este caso, que es derivado para Q-M346 pero ancestral para Q-L54), para un individuo de San Juan, Argentina, no identificado antes en esta región. Q-M346* podría ser un tercer sub-linaje autóctono de América, pero se necesitan más estudios para su confirmación. Otras dos secuencias obtenidas en este trabajo aportan nueva información a Q-Z780; cinco secuencias aportan nueva información a sub-linajes definidos dentro de Q-M848; y cinco secuencias forman parte junto a otras 12 secuencias de las bases de datos, de ramas dentro de Q-M3 donde sus relaciones filogenéticas no pueden resolverse y representan la gran variabilidad presente en linajes nativos americanos que todavía no se explican con los datos disponibles de secuencias. Se presentan 72 SNPs validados que aportan nueva información a Q-M346*, Q-Z780 y Q-M848, denominados como Q-GMP1 a Q-GMP72.

Los tiempos de divergencia y la estructura poblacional encontrada dentro de Q-Z780 aportan soporte genético a evidencias arqueológicas de ocupación humana temprana anteriores a 18000 años en Mesoamérica y Sudamérica. Estudios más exhaustivos en linajes nativos americanos más antiguos, como Q-Z780 (y quizás Q-M346*), permitirían acceder a la historia humana más ancestral en estas regiones.

ABSTRACT

The early settlement of America has been the focus of incessant debate for more than 100 years. Several archaeological sites have shownevidence of early human occupation in Mesoamerica and South America dating back 18000 years and before. The human Y chromosome has the longest stretch of non-recombinant DNA in the entire human genome and is completely transmitted from parent to child, thus contains a register of the paternal lineage history and is used as a highly informative tool to investigate the history of human populations. To carry out a study on the phylogenetic relationships of Native American lineages and infer the history of the American settlement, we analyzed the Y-chromosome Q haplogroup, which is the only Pan-American haplogroup and represents practically all Native American lineages in Mesoamerica and South America.

A calibrated phylogenetic tree for Haplogroup Q was constructed based on 102 whole Y-chromosome sequences, of which 13 sequences are new presented in this work. We defined 17 Q-subhaplogroups. Of these, 13 subhaplogroups are Native American specific and belong to Q-Z780 and Q-M3 (includes Q-M848). One of the sequences presented in this work was identified within the Q-M346* sub-lineage (the suffix "*" is indicating in this case that it is derived for Q-M346 but ancestral for Q-L54), for an individual of San Juan, Argentina, not previously identified in America. Q-M346* could be a third Native American sub-lineage, however, more studies are needed for confirmation. Two other sequences obtained in this work provide new information to Q-Z780; five sequences contribute new information to defined sub-lineages within Q-M848; and five sequences are part, along with 12 other sequences in the databases, of branches within Q-M3 where their phylogenetic relationships cannot be resolved and represent the great variability present in Native American lineages that are not yet explained with the available data from sequences. We present 72 validated SNPs that provide new information to Q-M346*, Q-Z780 and Q-M848, referred to as Q-GMP1 to Q-GMP72.

The divergence times and the population structure found within Q-Z780 provide genetic support for archaeological evidence of early human occupation in Mesoamerica and South America before 18000 years. More exhaustive studies in older Native American lineages, such as Q-Z780 (and perhaps Q-M346*) would allow access to the most ancient human history in these regions.

1 INTRODUCCIÓN

1.1 El Cromosoma Y

El material genético nuclear de las células somáticas de individuos humanos normales se divide en 46 moléculas separadas: los cromosomas. Estos a su vez se pueden dividir en 23 pares, donde uno de cada par se hereda de cada progenitor. Los cromosomas del par 1 al 22 se conocen como autosomas. Los dos cromosomas restantes se conocen como los cromosomas sexuales, porque difieren entre los sexos [1]. En este trabajo simplificaremos la definición de los géneros teniendo en cuenta los cromosomas sexuales, de esta manera, las mujeres poseen dos copias del cromosoma X y los hombres presentan un cromosoma X y un cromosoma Y; debido a que las monosomías y polisomías de los mismos, no se abarcan en este estudio [2, 3]. El cromosoma Y es determinante del sexo a través de la expresión del gen *SRY* (del inglés, sex-determining region, Y), en el desarrollo temprano, su expresión causa que el precursor de la gónada se convierta en un testículo en lugar de en un ovario [1].

El cromosoma Y posee propiedades únicas que lo diferencian del resto del genoma humano: está presente solamente en hombres y no es esencial para la vida de un individuo (las mujeres no lo tienen), la mitad consiste en repeticiones en tándem (ADN satélite), presenta pocos genes, es haploide (cada célula masculina presenta una copia cromosómica-Y), contiene el tramo más largo de ADN no recombinante de todo el genoma humano y se transmite por completo de padres a hijos [4, 5]. Por lo que, el cromosoma Y contiene un registro de la historia del linaje paterno, y es una herramienta excelente para investigar la evolución humana desde la perspectiva masculina, presentando roles importantes en los estudios de historia de las poblaciones humanas, genealogía, estudios en genética forense y genética médica masculina [6].

1.2 Estructura genética del cromosoma Y

El cromosoma Y es uno de cromosomas más pequeños del genoma humano, presenta una longitud de secuencia aproximada de 60 Mb. En la Figura 1.1 se esquematiza la estructura del Cromosoma Y. Existen dos segmentos de homología de secuencia llamadas Regiones Pseudoautosómicas, en las puntas de los brazos cortos y brazos largos del cromosoma Y, donde se produce el entrecruzamiento meiótico entre los cromosomas X e Y. Entre estas regiones se encuentra la región masculina específica del cromosoma Y (MSY, del inglés, male-specific region of the Y chromosome), también conocida como región no recombinante del cromosoma Y (NRY, del inglés, non-recombining region of the Y chromosome) que escapa del entrecruzamiento [6].

La región MSY, presenta en aproximadamente la mitad de su longitud un bloque de heterocromatina constitutiva en el brazo largo de longitud variable que incluye al centrómero y en la longitud restante presenta un bloque de eucromatina de aproximadamente 23 Mb compuesta por tres clases de secuencia que por su complejidad y grado de homología con el cromosoma X, son definidas como: la clase X-degenerada (XDG) que presenta una longitud de secuencia de 8,6

Mb; la clase X-transpuesta (XTR) con una longitud de 3,4 Mb; y la clase amplicónica con una longitud total de 10,2 Mb [7].

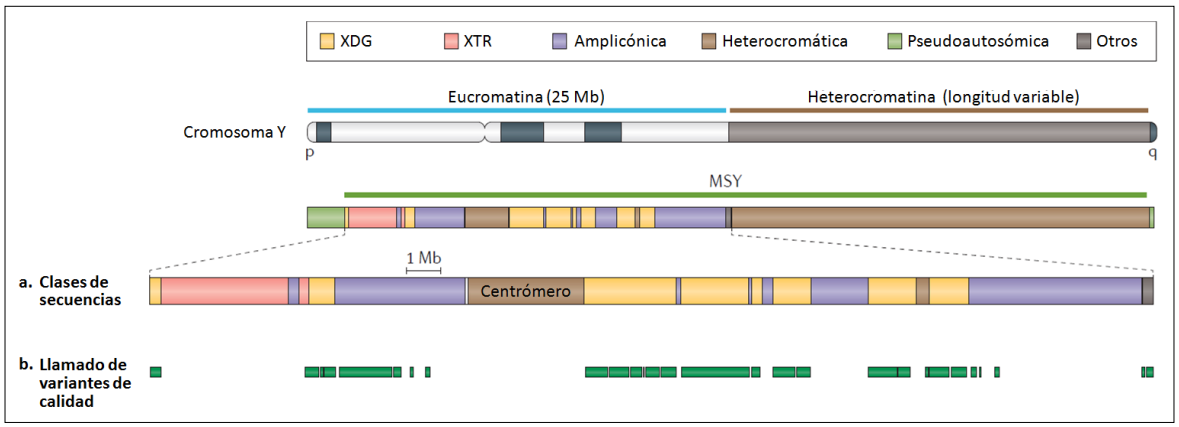


Figura 1.1: Representación esquemática del Cromosoma Y. El brazo corto del cromosoma Y es nombrado como p y el brazo largo como q. Imagen extraída y modificada de la referencia [6].

En los estudios de secuenciación completa de cromosoma Y, los cuales serán abordados en este trabajo, es necesario determinar las diferencias (variantes) entre la muestra secuenciada y el genoma de referencia, a este proceso se lo denomina llamado de variantes (en inglés, variant calling). Se ha encontrado una alta similitud intercromosómica e intracromosómica entre la región amplicónica y la región XTR, que dificulta la interpretación de los datos de secuenciación y sólo en una pequeña porción las variantes pueden ser llamadas y leídas sin ambigüedad (ver figura 1.1.b) [6]. Por lo tanto, existen sitios dentro del MSY en los cuales el llamado de variantes se puede realizar de manera confiable, siendo la clase XDG en su mayoría la región que produce datos de llamado de secuencias de calidad. Mientras que las regiones amplicónicas y XTR tienden a fallar en la calidad y son regiones deficientes en llamado de variantes sin ambigüedad. En síntesis la longitud aproximada de llamado de variantes de calidad de la región MSY da un total de ~10 Mb [8].

1.3 Polimorfismos y filogenias del cromosoma Y

Entre los polimorfismos presentes en la región MSY, se conocen las repeticiones cortas de ADN en tándem (STR, del inglés, Short Tandem Repeats), también llamados microsatélites, y además, los polimorfismos de un solo nucleótido (SNP, del inglés, Single Nucleotide Polymorphism), que se caracterizan por el cambio de una única base, y son también llamados marcadores bialélicos o marcadores binarios, debido a que en las poblaciones humanas presentan la posibilidad de segregarse en una de dos clases alélicas [9].

Los SNPs bialélicos surgen como eventos mutacionales con baja frecuencia y como consecuencia es muy pequeña la probabilidad de que dos eventos mutacionales consecutivos ocurran exactamente en el mismo locus, lo que los hace muy confiables para la reconstrucción filogenética

del cromosoma Y [10]. Por otro lado, las mutaciones STRs se producen por la adición o sustracción de un número determinado de repeticiones (generalmente de a una), debido a esto, se caracterizan por presentar alelos con diferente longitud de repeticiones. Ambos tipos de marcadores se han utilizado juntos en la determinación de linajes. Mientras los SNPs definen grupos monofiléticos de linajes llamados haplogrupos, los STRs definen sub-linajes de los haplogrupos llamados haplotipos [11].

Los primeros polimorfismos identificados en el NRY se identificaron en 1985 con el surgimiento de la técnica de polimorfismos en la longitud de los fragmentos de restricción (RFLP, del inglés Restriction Fragment Length Polymorphism), en donde, secuencias específicas de nucleótidos son reconocidas y cortadas por enzimas de restricción [12, 13]. Para esa época, se habían distinguido un número muy bajo de polimorfismos, lo cual dificultaba la construcción de árboles filogenéticos [14].

Durante las décadas siguientes se sumaron una gran cantidad de estudios sobre la variación del cromosoma Y, contribuyendo a la creación de árboles filogenéticos que combinaban SNPs (en números limitados) más STRs [15]. El despliegue de múltiples STRs, que presentan tasas de mutación más altas y que son variables en todas las poblaciones, pueden revelar el nivel de variación dentro de los haplogrupos descriptos por los SNPs y también pueden proporcionar alguna información sobre sus profundidades de tiempo (es decir, el tiempo transcurrido desde que ocurrió la mutación definitoria de haplogrupo), así los haplogrupos más antiguos albergarán una mayor diversidad de haplotipos STRs que los más jóvenes [16]. En 2002 los miembros el Consorcio del Cromosoma Y (YCC, del inglés, *Y Chromosome Consortium*) constituyeron un sistema de nomenclatura normalizada para clasificar los polimorfismos de la región no recombinante del cromosoma Y, y publicaron un único árbol de parsimonia de alta resolución para 153 haplogrupos [17]. En 2008 se presentó un árbol actualizado y revisado de 311 haplogrupos [15].

Aunque los estudios combinados de Y-SNPs más Y-STRs prosperaron, generando la sub-disciplina de la filogeografía masculina, han sido superados por los avances en la técnica de secuenciación completa de cromosoma Y, proporcionadas por las plataformas de secuenciación NGS (del inglés, next-generation sequencing, traducida como secuenciación de próxima generación). En la actualidad la mejor manera de identificar la variación en el cromosoma Y es secuenciándolo [18]. Este enfoque comenzó a utilizarse más ampliamente desde el 2010 cuando las plataformas de secuenciación NGS comenzaron a disminuir sus costos [19].

Los árboles de cromosoma Y construidos con secuenciación NGS permiten crear árboles filogenéticos robustos *de novo* en base a un gran conjunto de datos de SNPs, en donde las longitudes de ramificación son proporcionales al número de mutaciones (SNP) [20]. A partir del 2010 varios estudios demostraron el valor de secuencias completas de cromosoma Y para caracterizar y calibrar la filogenia del cromosoma Y humano, y también sobre la forma en la que esta filogenia puede proporcionar información sobre la evolución humana reciente, generando nuevas ideas sobre la historia humana masculina [8, 21-23].

En el año 2016 el Proyecto 1000 Genomas publicó un árbol filogenético calibrado basado en 60555 SNPs obtenidos con secuenciación NGS de cromosoma Y de un conjunto de 1244 individuos seleccionados aleatoriamente de 26 poblaciones del mundo [22]. Este trabajo recapituló y refinó la estructura filogenética del cromosoma Y para la mayoría de los haplogrupos y brindó nueva información base para muchos estudios posteriores.

1.3.1 Tasa mutacional del cromosoma Y desde datos NGS

Como cualquier parte del genoma, la región MSY acumula SNPs a través de eventos mutacionales, pero lo hace a una tasa promedio más alta que los otros cromosomas [6]. Se cree que el motivo de la tasa más alta de mutación del NRY se debe a la gran cantidad de divisiones celulares y, por lo tanto, a las replicaciones de ADN que se producen en la línea germinal masculina [24].

La medición directa de la tasa de mutación de SNPs del MSY comenzó en el 2008 en un estudio donde los cromosomas Y de dos individuos separados por 13 generaciones de un pedigrí chino, se secuenciaron con la profundidad (depth o depth of coverage, representa el número promedio de veces que cada base en el genoma es secuenciada en los fragmentos de ADN) de 11-20x y las mutaciones candidatas se examinaron para la medición de la tasa mutacional [25].
También se han utilizado puntos de calibración arqueológicos para estimar la tasa de mutación de SNPs del MSY, en un estudio de la expansión del cromosoma Y en América a 15 kya (del inglés, kilo years ago, miles de años atrás) [8]. En el presente estudio en general, se expresarán en kya las dataciones que provienen de estudios moleculares.
La estimación de las tasas de mutación del ADN antiguo (ADNa) requiere una secuencia antigua que esté fechada con precisión y sea lo suficientemente antigua como para que le falten muchas mutaciones [26], que es una combinación aparentemente poco probable. Sin embargo, afortunadamente se ha informado una de esas secuencias a partir de un fémur de 45,000 años de Ust'-Ishim encontrado en el oeste de Siberia [27].

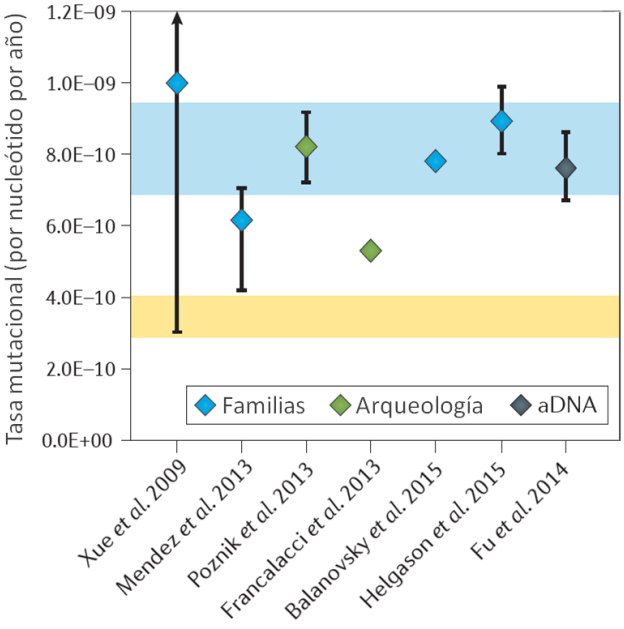


Figura 1.2: Representación gráfica de las tasas mutacionales de cromosoma Y obtenida desde diferentes estudios. Los diamantes representan la tasa de mutación estimada y su color indica si el valor se deriva de estudios familiares, evidencias arqueológicas o de ADN antiguo. Las barras verticales, representadas para algunos estudios, muestran los intervalos de confianza del 95%. Las estimaciones de las tasas mutacionales de estos estudios son razonablemente consistentes y sus intervalos de confianza se superponen. La tasa de mutación consenso SNP del MSY (sombreado azul) se contrasta con la tasa de mutación SNP autosómica (sombreado amarillo). Imagen e información tomada de [6].

1.3.2 Calibración de filogenias producidas con NGS

Además de producir una filogenia con una estructura robusta, los datos NGS dan como resultado ramas para las cuales las longitudes se basan en el número de mutaciones. Por lo que, si conocemos la tasa de mutación y ésta ha sido constante, esta información se puede convertir en tiempo para generar una filogenia calibrada [22].

Sin embargo, evaluar la constancia de la tasa de mutación en el tiempo y en diferentes lugares es difícil. El número de mutaciones de la línea masculina aumenta con la edad paterna [28] y, por lo tanto, la variación en el tiempo de generación masculina podría conducir a una variación de la tasa mutacional [29], que en el MSY podría conducir a diferentes longitudes de ramas desde la raíz a la punta para diferentes linajes [30].

Como se mencionó anteriormente, se han utilizado diferentes enfoques para estimar la tasa mutacional de los Y-SNPs, y sigue en estudio qué tasa mutacional es más confiable. Los tres enfoques mencionados dan estimaciones razonablemente consistentes: sin embargo, existe una diferencia del 15% entre la estimación basada en genealogía de $8.9x10^{-10}$ mutaciones por base por año [28] y la estimación basada en ADNa de $7.6x10^{-10}$ mutaciones por base por año [27].

Dado a que en este estudio realizaremos cálculos temporales de la filogenia de cromosoma Y del Haplogrupo Q característico de América para hacer inferencias históricas sobre poblaciones nativas americanas y compararemos nuestros cálculos temporales con otros trabajos que realizan inferencias similares, utilizaremos la tasa mutacional más utilizada, basada en ADNa [22, 31].

1.3.3 Manejo de datos obtenidos por secuenciación NGS

En la práctica, la secuenciación del NRY no está exenta de dificultades. Incluso con la disponibilidad de una secuencia de referencia de alta calidad [7], la compleja estructura altamente repetitiva del NRY y las lecturas cortas (<200 pares de bases) producidas por la mayoría de las tecnologías de secuenciación actuales hacen que el mapeo inequívoco (es decir, la alineación con el genoma de referencia) sea posible solo en regiones únicas del cromosoma. Estas regiones son discontinuas y se dispersan a lo largo del NRY sumando una longitud total aproximada de 10 Megabases (Mb, equivale un millón de bases) (ver figura 1.1.b) [6, 22]. Algunos estudios se han enriquecido específicamente para los subconjuntos de 0.5 - 3.7 Mb de estas regiones [30, 32, 33], mientras que otros, han secuenciado el genoma completo y posteriormente extrajeron las lecturas relevantes con técnicas bioinformáticas [8, 21, 23]. La profundidad de secuenciación también es importante. La baja profundidad utilizada en varios estudios iniciales probablemente hayan resultado en el descubrimiento menos eficiente de variantes raras que están presentes en solo uno o unos pocos individuos [34].

Varios otros factores técnicos también influyen en el conjunto final de variantes y, por lo tanto, en el árbol filogenético como la plataforma de secuenciación, el algoritmo utilizado para la llamada de variantes y las estrategias de filtrado y validación. En consecuencia, los resultados presentados

en diferentes estudios filogenéticos generados con técnicas NGS no se pueden combinar o comparar de manera simple. Se requiere un nuevo análisis a partir de las lecturas de secuencia [6, 20, 35]. A pesar de estas complejidades, los conjuntos de variantes llamadas y las filogenias basadas en SNPs de estudios NGS independientes son altamente congruentes [8, 36, 37].

1.3.4 Filogenia del Haplogrupo Q-M242 basada en datos NGS

El haplogrupo Q es el único haplogrupo panamericano y representa prácticamente todos los cromosomas Y de los nativos americanos en Mesoamérica y Sudamérica. Como mencionamos, en el año 2016 el Proyecto 1000 Genomas publicó un árbol filogenético calibrado basado en secuenciación NGS de cromosoma Y. En este trabajo se incluyeron 57 secuencias de individuos pertenecientes al haplogrupo Q [22], el cual fue base para estudios recientes de reconstrucción filogenética del haplogrupo Q [31, 38], que serán analizados posteriormente en este estudio. En la figura 1.3 se muestra el árbol filogenético para el haplogrupo Q presentado por el Proyecto 1000 Genomas.

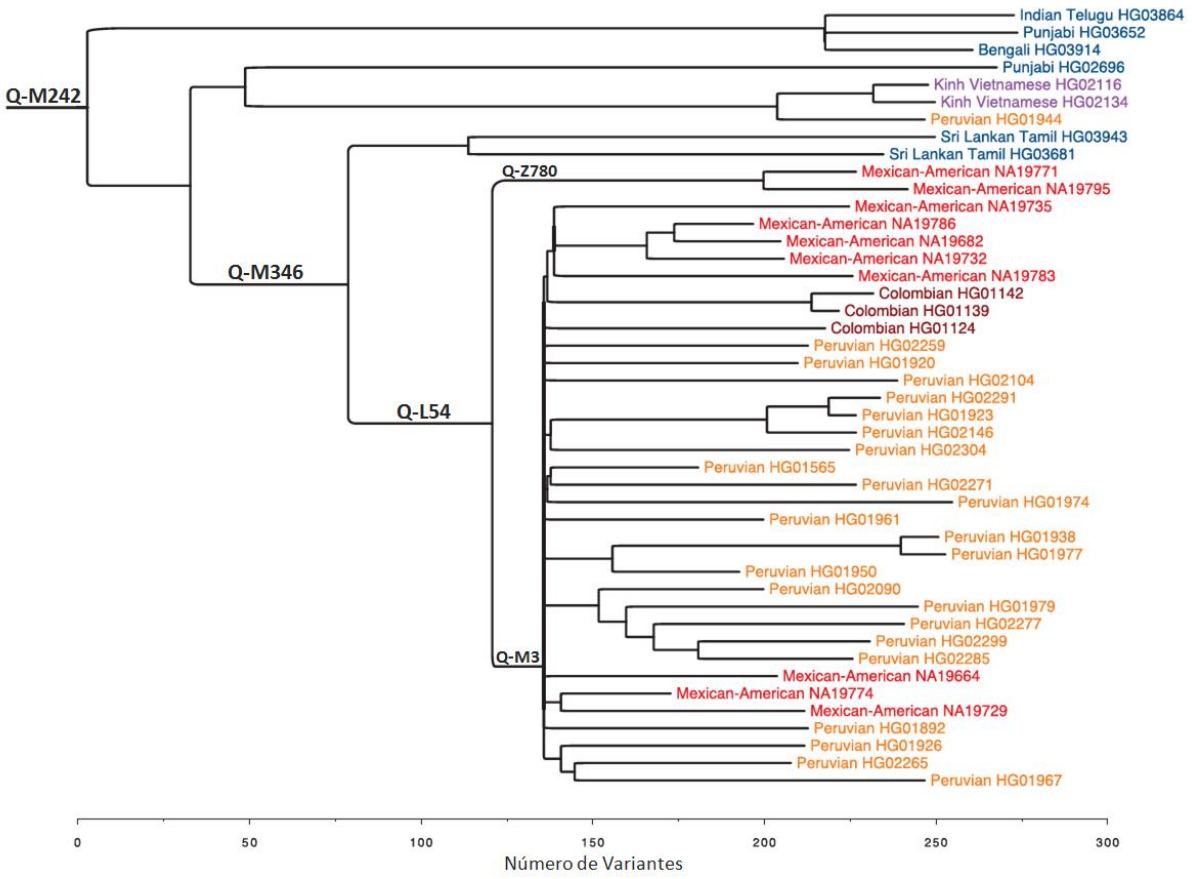


Figura 1.3. Árbol filogenético para el haplogrupo Q basado en datos NGS presentado por el Proyecto 1000 Genomas. Las longitudes de las ramas están en función del número de variantes encontradas. Los individuos representados en tono azul indican sudasiáticos, el violeta indica vietnamitas, el naranja peruanos, el bordó colombianos y el rojo individuos de Los Ángeles de origen mexicano. Imagen tomada y modificada de [22].

De la filogenia completa del cromosoma Y, el haplogrupo Q-M242 es una de las dos ramas de P-M45, la otra rama es R-M207 [22]. El haplogrupo Q se define por la mutación Q-M242 [39], ha sido descripto en frecuencias bajas en Europa, Asia oriental y Medio Oriente y en frecuencias altas en algunos pueblos siberianos y en América [40]. El sub-haplogrupo Q-M346 se describió en Pakistán [41], en el sur de Siberia y el este de Europa [42], en Afganistán, Pakistán, Irán, Kirguistán y Mongolia [43, 44]. Q-L54, se encontró presente en la República de Tuva, Noreste de Siberia y México [45], Canadá [46], América [47] y en grupos kalmyks, siendo estos una rama de los mongoles [41]. Q-Z780 se ha descripto como un linaje nativo americano de baja frecuencia [41]. Q-M3 es el principal linaje hallado entre los nativos americanos, y es considerado como un linaje fundador de América [39]. Q-M3 ha sido descripto en alta frecuencia en América y en muy baja frecuencia en Siberia [48, 49].

1.4 Usos de datos genéticos en investigaciones poblacionales

Cuando se usan datos genéticos para realizar inferencias relacionadas al poblamiento de América, uno de los problemas surgen del mestizaje que tuvo lugar entre nativos americanos con otras poblaciones, que incluyen contingentes europeos y contingentes africanos que ingresaron a diferentes lugares de América en los últimos cientos de años [50-53]. Este mestizaje tiene consecuencias importantes: varía la diversidad genética e introduce linajes foráneos que podrían conducir a conclusiones incorrectas si no se identifican sus orígenes. Esto ha podido resolverse en gran parte gracias a los nuevos avances en la caracterización de diferentes linajes del cromosoma Y de diferentes poblaciones del mundo y a la incorporación de técnicas moleculares desarrolladas para su identificación [44, 54].

Además, la Sociedad Internacional de Genealogía Genética (ISOGG, en inglés, International Society of Genetic Genealogy), una organización independiente sin fines de lucro de genealogistas genéticos, dirigida por voluntarios de más de 70 países [55], presenta la base de datos pública más completa para cromosoma Y, donde se puede acceder al árbol filogenético del haplogrupo Q actualizado 2019-2020 [56] y una lista con todas las posiciones para SNPs validados según su ubicación filogenética [57].

Cuando se estudia al Cromosoma Y como fuente de información genética de los linajes nativos americanos, se destaca de manera contrastante la reducción drástica de los linajes masculinos autóctonos de América después de la colonización europea. Esto, sumado a la alta tasa histórica de mestizaje masculino en las comunidades nativas americanas, ha dificultado la identificación de los linajes autóctonos de América [58].

El conocimiento genético se ha desarrollado arraigado a la historia del colonialismo y a la explotación en la investigación antropológica y tiene el potencial de reproducir injusticias y desequilibrios de poder [59]. Ésto ha llevado a la creación de organizaciones como el Consejo de Pueblos Indígenas sobre Biocolonialismo (Indigenous Peoples Council on Biocolonialism, IPCB), con sede en Nevada, Estados Unidos, el cual tiene como objetivo la protección de los recursos genéticos, conocimientos indígenas, derechos culturales y humanos de los efectos negativos de la

biotecnología [60]. Se entiende por biocolonialismo a la apropiación indebida de recursos biológicos de países en desarrollo por grandes centros de investigación, laboratorios e industrias de ingeniería genética o bioingeniería [61]. Por lo que representa un desafío actual para las políticas que financian el área, así como para centros de investigación e investigadores, el romper con ciclos de desequilibrios de poder y prácticas biocoloniales en la antropología molecular. Es por esto que en el presente trabajo se procura contribuir desde una perspectiva genética al conocimiento de linajes nativos americanos, sus vínculos y su profundidad temporal en territorios americanos. No se pretende deslegitimar la narrativa de los pueblos originarios sobre su historia y sus orígenes sino poner a disposición este estudio como elemento de reclamo territorial.

1.5 Estudio del poblamiento de América

El origen de los humanos en América ha atraído enorme atención y controversia, así como años de investigaciones multidisciplinarias que plantean diferentes Modelos e Hipótesis de Poblamiento con la intención de dar respuestas a interrogantes como: ¿Cuándo entraron los humanos por primera vez en América? ¿De dónde vinieron? Estos interrogantes siguen sin tener respuestas certeras y existe una continua controversia entre diferentes disciplinas sobre el poblamiento de América.

La genética de poblaciones realiza contribuciones para conocer y comprender cómo fue el poblamiento americano, sin embargo es necesario relacionar estos datos con evidencias fósiles, sitios arqueológicos, relaciones lingüísticas, así como cambios ambientales que ocurrieron para interpretar los datos genéticos desde una visión integrada.

1.5.1 Cambios ambientales, extinción de la megafauna y cambios culturales

El término "Último Máximo Glacial" (abreviado como LGM del inglés, Last Glacial Maximum) se usa ampliamente para referirse al episodio en el que el volumen de hielo global alcanzó por última vez su máximo y los niveles del mar asociados estaban en su nivel más bajo, abarcando un intervalo de tiempo de 23.000 y 14.000 cal AP (años calibrados antes del presente) [62]. Los niveles del mar más bajos existieron durante el LGM, se ha explicado que deshielos repentinos de las capas de hielo y rotura de represas que retienen lagos liberaron enormes volúmenes de agua, que dieron como resultado un rápido aumento del nivel del mar conocido como pulsos de agua de deshielo (meltwater pulses). Tales cambios rápidos en el nivel del mar son parte de un patrón complejo de interacciones entre la atmósfera, los océanos, las capas de hielo y la tierra sólida, y cada uno de los cuales tienen diferentes escalas de tiempo de respuesta [63]. Se ha propuesto que dos pulsos de agua de deshielo marcaron un período frío, conocido como Younger Dryas (YD), que ocurrió entre 12.900-11.600 cal AP [64]. Existen registros que indican que el aumento del nivel del mar en todo el mundo fue de aproximadamente (~) 120 m hacen ~14.000 cal AP [1].

Se ha propuesto al cambio climático como una de las posibles causas de extinción de unos 35 géneros de grandes mamíferos en América hacen ~14.000 cal AP, incluidos mamuts, mastodontes, camélidos, caballos, perezosos terrestres gigantes, osos y felinos dientes de sable. Otra explicación

ha sido la excesiva caza por humanos y/o enfermedades traídas por humanos [1]. Pero no se han encontrado evidencias de enfermedad pandémica en ese periodo y dado a que las extinciones fueron demasiado amplias y ecológicamente profundas es poco probable que hayan sido el resultado solo del enfriamiento climático y la destrucción excesiva de humanos [65].

La hipótesis del impacto de Younger Dryas (YD) postula que los fragmentos de un gran asteroide / cometa en desintegración golpearon América del Norte, América del Sur, Europa y Asia occidental hacen ~12.800 cal AP [65-67]. Múltiples explosiones de aire / impacto produjeron la capa límite YD (YDB, Younger Dryas boundary), depositando concentraciones máximas de una gran diversidad de marcadores de impacto [65]. El evento de impacto propuesto provocó una gran quema de biomasa, cambio climático YD, cambios abruptos anómalos en la distribución de plantas y animales, extinción de la megafauna, así como, cambios culturales y disminución de la población humana [66]. La diversidad del conjunto de marcadores relacionados con el impacto cósmico se ubican en más de 50 sitios, que se encuentran principalmente dentro del hemisferio norte [65], pero también en tres sitios en el hemisferio sur en Venezuela [68], Antártida [69] y en la Patagonia de Chile [66].

1.5.2 Evidencia ósea humana

Los restos óseos proporcionan evidencia inequívoca de la presencia de humanos en América y sus dataciones brindan información valiosa sobre el momento del poblamiento humano en estas regiones. Su estudio también plantea cuestiones éticas y morales, donde se han realizado importantes avances en distintos países en cuestiones legales de restituciones de restos por parte de las comunidades originarias que los reclaman [70, 71].

En 1940 en las Cuevas llamadas Spirit Cave, en Nevada, Estados Unidos, se encontraron los restos óseos en una representación de entierro de un hombre bajo de entre 40 y 45 años con abscesos dentales, varias lesiones y parcialmente momificado por las condiciones secas. Los estudios de datación proporcionaron una antigüedad de 10.600 cal AP [72].

En 1968 se encontró en Montana, Estados Unidos los restos de un niño varón encontrado con más de 115 herramientas líticas, fue enterrado espolvoreado con ocre rojo, lo que sugiere un funeral honorario. Fue nombrado como el niño Anzick, su datación dio aproximadamente 12.700-12.500 cal AP. Se lo asoció a la Cultura Clovis y la secuenciación genómica lo mostró derivado para Q-Z780 [38, 73, 74].

En 1975 se encontró en Lapa Vermelha, Minas Gerais, Brasil los restos óseos de una mujer nombrada Luzia. Las dataciones realizadas varían en valores de 12.700 cal AP a 16.000 cal AP, se considera como uno de los esqueletos paleoindios más antiguos encontrados en América [75].

En 1980 se encontró una bola de pelo congelada en Groenlandia perteneciente a la cultura paleo-esquimal llamada Saqqaq. Su genoma fue secuenciado recientemente, su datación dio 4 kya, se

encontró desde estudios del cromosoma Y que es derivado para el haplogrupo Q y su variación genómica mostró afinidad biológica con tres pueblos árticos del Lejano Oriente siberiano: los nganasanos, koryaks y chukchis [38, 76].

En 1989 se encontró en una cantera cerca de Buhl, Idaho, Estados Unidos los restos óseos de una mujer con artefactos que incluyen, una biface de obsidiana y una aguja de hueso, lo que sugiere un entierro. Fue nombrada Buhl Woman o mujer de Buhl, las dataciones dieron ~12.9 cal AP [77]. En 1991 se realizó un nuevo entierro, de acuerdo con la Ley de Protección y Repatriación de Tumbas de los Nativos Norteamericanos [78].

En 1996 se encontró en Kennewick, Estados Unidos el esqueleto casi completo de un hombre de 40-55 años que fue nombrado como el hombre Kennewick, los análisis de datación estimaron una antigüedad 8.300-9.200 cal AP y su asociación al linaje Q-M3 [79].

En 1996 se encontró en el sitio arqueológico On Your Knees Cave en Alaska, en la Isla Príncipe de Gales, los restos óseos de un hombre que fue nombrado el "Hombre de la Isla Príncipe de Gales" o "Prince of Wales Island Man". Las dataciones dieron ~10.300 cal AP y los análisis de ADN antiguo realizados en estos restos revelaron su pertenencia al haplogrupo Q-M3 [80].

1.5.3 Restos arqueológicos

En la actualidad existen grandes controversias entre diferentes arqueólogos que realizan dataciones en sitios vinculados a evidencias de los primeros pobladores del continente americano, existen más de 400 sitios arqueológicos dispersos por Sudamérica que han sido fechados con dataciones radiocarbónicas calibradas y muestran una mayor densidad de evidencias para la ocupación humana posterior a 14.500 cal AP [81]. A continuación mencionaremos algunos sitios arqueológicos relacionados a estas fechas, así como también sitios con dataciones anteriores.

Cultura Clovis y Folsom: Los primeros restos arqueológicos de la Cultura Clovis fueron encontrados en New Mexico, Estados Unidos. Se caracterizan por presentar las llamadas "puntas Clovis", puntas de proyectil estriadas (ranuradas) construidas a partir de diferentes tipos de piedra, las dataciones de los primeros restos arqueológicos Clovis datan de alrededor de 13.500 cal AP [82]. La cultura Clovis no se encuentra fuera de Norteamérica, y se conoce que eran cazadores ya que sus puntas de piedra se han encontrado junto a material óseo de animales como mamuts, mastodontes y animales más pequeños [83]. Los restos arqueológicos de la Cultura Folsom han sido datados en 12.900-12000 cal AP, las puntas de piedra de estos cazadores se distinguen por un tamaño más pequeño y más largo [82].

Caverna da Pedra Pintada: El sitio arqueológico de Cueva de la Roca Pintada se sitúa cerca de Monte Alegre, en el Estado de Pará, Brasil. Este sitio contiene rocas con pinturas rupestres y restos biológicos que indican el uso de nueces de Brasil, peces, tortugas y mejillones, así como puntas triangulares hechas de cuarzo y calcedonia que podrían haber sido usadas para cazar animales

más grandes. Las dataciones realizadas sobre algunos de estos restos indican una ocupación humana aproximada (~) de 13.500 -10.000 cal AP [84].

Huaca Prieta: ubicado cerca de la costa del Pacífico del norte de Perú. Las dataciones de radiocarbono indican una presencia humana intermitente fechada entre ~15.000 cal AP y 8.000 cal AP [85].

Arroyo Seco 2: ubicado al sur de la región pampeana argentina. Contiene un rico registro arqueológico que evidencia la ocupación humana y su interacción con los mamíferos extintos del Pleistoceno con dataciones de radiocarbono de ~12.170 cal AP (13.814-14.147 cal AP) [86].

Monte Verde II: ubicado al sur de Chile, este sitio arqueológico arrojado evidencias de presencia humana que han sido datadas en ~14.500 cal AP [87].

Pilauco: ubicado en la ciudad de Osorno, Chile. Se encuentra a 100 km del sitio arqueológico Monte Verde y se ha establecido como contemporáneo a éste. Proporciona una gran variedad de evidencia sobre la coexistencia humana, floral y faunística del Pleistoceno tardío en el noroeste de la Patagonia [88].

Pedra Furada: yacimiento arqueológico y de pinturas rupestres localizado en São Raimundo Nonato, al este de Piauí (Brasil), registra evidencias de ocupación humana de ~ 32.000 cal AP [89]. Cercano a este sitio se encuentra otro yacimiento arqueológico importante, Boqueirão da Pedra Furada, donde se han encontrado pruebas de ocupación humana que se remontan a más de 20.000 cal AP [90].

Toca do Sítio do Meio: ubicado en Parque Nacional Serra da Capivara (Piauí, Brasil), donde se han encontrado evidencias de ocupación humana de aproximadamente 25.000 cal AP [91]. Otro sitio arqueológico en este parque es el llamado, Toca da Tira Peia, donde se han hallado evidencias de ocupación humana de 20.000 cal AP [92].

Santa Elina: ubicado en el lado sureste de la cordillera que une la Serra das Araras y la Serra da Água Limpa en Matogrosso, Brasil. Se han encontrado evidencias de ocupación humana con fechas de ~23.000 cal AP [93].

Arroyo del Vizcaíno: este yacimiento arqueológico situado en Uruguay ha sido fechado en ~30.000 cal AP y se encontraron más de mil restos de megafauna, algunos de ellos muestran marcas que fueron interpretadas como productos de la acción humana [94].

Cueva de Chiquihuite: ubicada en el centro-norte de México, esta cueva presenta evidencias que sugieren la presencia humana de ~26.000 cal AP y antes [95].

1.5.4 Lenguas nativas americanas

En la actualidad no hay un consenso respecto al número de lenguas nativas americanas que existen y las estimaciones de académicos lingüistas respetados han oscilado entre 400 y más de 2.500. Si bien, tampoco hay un acuerdo sobre cómo se deben clasificar los idiomas nativos americanos, la mayoría de investigadores lingüistas estiman que hay aproximadamente 150 familias de idiomas diferentes en América que actualmente no se puede demostrar que estén relacionadas entre sí [96].

La lingüística histórica busca rastrear la ascendencia de los diferentes idiomas que se hablan actualmente. Muchos de estos idiomas se remontan a varios idiomas ancestrales conocidos como proto-idiomas. Dado a que en este trabajo se intenta reconstruir la historia ancestral nativa americana, se hará uso del conocimiento disponible sobre la distribución geográfica de algunas familias lingüísticas nativas americanas y en algunos casos de la lingüística histórica para realizar inferencias sobre las relaciones filogenéticas de algunos sub-linajes del haplogrupo Q-M848 del presente estudio. Para un mejor entendimiento del uso de la lingüística en este estudio, los antecedentes sobre estos conocimientos se encuentran desarrollados en el capítulo 4 para sub-linajes Q-M848 que incluyen etnias para las cuales existe conocimiento sobre su lengua, sobre la distribución geográfica de su familia lingüística y sobre lingüística histórica.

1.5.5 Hipótesis de poblamiento americano

El poblamiento americano involucra uno de los debates multidisciplinarios más polémicos de la actualidad. Este tema involucra un gran cuerpo de investigaciones y números teorías sobre cómo y cuándo comenzó este evento. A continuación se resumen tres principales modelos que mantienen defensores vigentes en la actualidad.

La teoría conocida como Los Primeros Clovis (Clovis First), fue la hipótesis predominante entre los arqueólogos en la segunda mitad del siglo XX. Este modelo sostiene que la presencia humana en América comenzó hacen unos 13.000 cal AP con la cultura Clovis de Norteamérica. El principal apoyo para esto fue que no se había encontrado evidencia sólida de habitación humana anterior a Clovis. De acuerdo con la teoría más aceptada, la cultura Clovis cruzó el puente terrestre de Beringia sobre el estrecho de Bering desde Siberia a Alaska durante el período de descenso del nivel del mar durante la edad de hielo, y luego se dirigió hacia el sur a través de un corredor libre de hielo al oeste de Canadá a medida que los glaciares se retiraron [97].
Algunos de los argumentos que sostienen en la actualidad la teoría de Los Primeros Clovis se basa en los resultados de la secuenciación completa de Anzinck y su pertenencia al haplogrupo Q-Z780, el cual es un linaje masculino ancestral todavía extendido en nativos americanos, y el linaje Q-M3 es el descendiente más común, que según Dulik y col. 2012 surgió hacen unos 13.400. Por otro lado, las puntas estriadas de Clovis han sido reconocidas como las herramientas líticas construidas por la Cultura Clovis. En arqueología existe una tendencia en comparar las adaptaciones tecnológicas utilizadas por paleoindios en diferentes regiones. Las puntas estriadas encontradas en América del Sur han sido interpretadas por algunos investigadores norteamericanos como una

difusión de las puntas estriadas de Clovis de América del Norte [98]. Las puntas estriadas de América del Sur incluyen las puntas "cola de pescado" representadas en Cueva Fell (cueva natural y sitio arqueológico en el sur de la Patagonia) [99] y muchas otras regiones, que datan de la época de Clovis y llegan hasta Tierra del Fuego [100]. Estos son algunos de los fundamentos que han llevado a Stuart Fiedel y col. 2017 a afirmar que: "Los parientes cercanos que enterraron al niño Anzick 12.800 cal AP, fabricaron herramientas clásicas de Clovis y fueron inequívocamente los ancestros genéticos lineales de todos los pueblos nativos vivos del sur de Norteamérica, América Central y América del Sur. Las puntas de cola de pescado Fell I derivadas de Clovis rastrean la rápida migración hacia el sur de esta población ancestral hasta Tierra del Fuego. Cualquier población hipotetizada anterior, si alguna vez existió (improbablemente), debe haber sido reemplazada o genéticamente inundada por estos descendientes de Clovis" [98].

El segundo modelo, afirma una antigüedad intermedia y propone un poblamiento del continente americano pre-Clovis pero post-LGM, entre 18.500 - 13.000 cal AP, con una posterior entrada humana a América del Sur poco tiempo después [101].
Recientemente, se ha realizado una contribución importante a este modelo en un trabajo reciente llevado a cabo por Prates y col. 2020. En este estudio se recopila el conjunto de datos de 1661 fechas tempranas de radiocarbono (1543 realizadas en materiales culturales o restos relacionados y 118 en huesos/dientes humanos) de 454 sitios arqueológicos de América del Sur. En este análisis no se incluyen los sitios arqueológicos anteriores a 15.000 cal AP mencionados en la sección 1.5.3 porque se considera que no cumplen los requisitos estándares de validación de estos autores. En base a sus análisis estadísticos definieron un umbral cronológico más temprano para el poblamiento de América del Sur entre ~15.500 cal AP (16.600-15.100 cal AP). También realizaron una exploración para el cambio demográfico temprano usando probabilidades sumadas y encontraron que durante el primer período con evidencia humana en Sudamérica (15.100-13.500 cal AP) la intensidad de la señal arqueológica es extremadamente baja. La señal de intensidad arqueológica aumenta lentamente y alcanza un pico entre 12.500 cal AP, que coincide con el período del Younger Dryas y el momento de las principales extinciones de la megafauna. Hacia 11.000 cal AP el crecimiento de la población se estabiliza con un aumento gradual a largo plazo. Además en este trabajo se afirma que: "si realmente hubiera ocurrido una primera llegada antes de 15.500 cal AP, esta población temprana probablemente se habría extinguido, ya que la supuesta evidencia cultural antes de esta fecha muestra una discontinuidad sustancial (e inesperada) en la curva de densidad de probabilidad calibrada sumada, y los restos humanos están completamente ausentes antes de 12.600 cal. AP" [81].

En base a estudios genómicos de muestras modernas y antiguas se ha considerado que los nativos americanos derivan de un subconjunto del acervo genético euroasiático llevado a América por una población ancestral de Beringia en una cronología acorde al segundo modelo [102]. Estudios recientes basados en secuenciación del cromosoma Y pertenecientes al haplogrupo Q han presentado nuevos hallazgos a la historia genética de poblaciones nativas americanas aportando nuevas evidencias al segundo modelo de poblamiento [22, 31, 38]. Pinotti y col. 2019 realiza una reconstrucción filogenética calibrada de cromosoma Y en base a 222 secuencias completas del

mundo pertenecientes al Haplogrupo Q y C (con 20 nuevas secuencias de nativos de América del Sur, 19 del haplogrupo Q y uno del haplogrupo C3). En este trabajo en base a los resultados filogenéticos y a las dataciones estimadas se establece que la ocupación inicial de América del Norte se dio después de 19.5 kya. Para el linaje Q-M848 (dentro de Q-M3) se volvió a definir lo que ya había sido descripto por otros autores como una rápida expansión de ~15 kya asociada a la colonización inicial de América del Sur [22]. Por primera vez se estableció que los sub-linajes de Q-M848 presentaban una estructura espacial en América del Sur que surgió tan pronto como ~12.3 kya. Para Q-Z780 este trabajo encontró una profundidad temporal de 17 kya (15.0-19.3) que se interpretó con un escenario de asentamiento más complejo [31]. También en el año 2019 Grugni y colaboradores presentan un árbol filogenético calibrado del haplogrupo Q basado en una longitud de 1,5 Mb de 152 cromosomas Y, de ellos 34 nuevos re-secuenciados. En base a las dataciones estimadas para Q-Z780 proponen una entrada a América del Sur antes de 15 kya. Por otro lado, sobre los análisis en Q-M848 definen una fase importante de crecimiento de la población masculina después de 15 kya, seguida de un periodo de tamaño poblacional constante de 8 a 3 kya, después del cual se registra otro signo de crecimiento. Estos eventos de expansión comenzaron durante el Holoceno con la mejora de las condiciones climáticas [38]. Estos últimos dos trabajos aportan grandes avances en el conocimiento filogenético del haplogrupo Q y son importantes en las inferencias realizadas en el capítulo 4 de este estudio.

El tercer modelo, de Cronología Larga, propone una entrada al continente americano pre-LGM y defiende un poblamiento de América del Sur antes de 18.000 cal AP [90, 93, 95, 103].
Los defensores de este modelo son los académicos que han validado sitios anteriores a 18.000 cal AP que se encuentran descriptos en la sección 1.5.3. En la actualidad este modelo es muy cuestionado por la gran mayoría de académicos en desacuerdo con los criterios de validación de estos sitios.

Objetivo General

Ampliar el conocimiento de la diversidad genética de los linajes nativos americanos pertenecientes al Haplogrupo Q y contribuir con nuevas perspectivas a la historia del poblamiento americano.

Objetivos específicos

1. Realizar una reconstrucción filogenética del haplogrupo Q-M242 desde datos de secuenciación NGS.
2. Analizar la variabilidad e incrementar la resolución del haplogrupo Q.
3. Analizar la distribución de los linajes propios de nativos americanos.
4. Realizar una descripción de las relaciones filogenéticas de linajes nativos americanos y contrastar con la información arqueológica, histórica y lingüística disponible.
5. Proponer un modelo de poblamiento americano en base a las relaciones filogenéticas y a las dataciones de los sub-linajes nativos americanos encontrados.

2 MATERIALES Y MÉTODOS

2.1 Muestras

La colección de muestras del Laboratorio de Genética Molecular Poblacional consta en la actualidad de 1085 muestras masculinas ya tipificadas (con ADN extraído de sangre periférica) obtenidas de campañas realizadas en distintas poblaciones, llevadas a cabo en diferentes momentos por diferentes proyectos.

Las diversas poblaciones urbanas y suburbanas muestreadas son procedentes de las provincias argentinas de Mendoza, San Juan, La Rioja, Salta, Jujuy, Catamarca, Tucumán, Santa Fe, Entre Ríos y Corrientes (PICT 2005 32450). Otra parte de las bases de datos incluyen 145 muestras específicamente de comunidades indígenas de Gran Chaco: Mocoví, Toba, Wichi, Chorote, Lengua y Ayoreo (ADN extraído de saliva) (PICT 2003 Nº 01 14328 y Proyecto "De las Historias Étnicas a la Prehistoria en el Gran Chaco argentino") y de las comunidades Huiliche y Mapuche (ADN extraído de sangre periférica) (PICT 1998 1-4493).

Dos evaluaciones del Comité de Ética fueron aprobadas, el Comité de Bioética Provincial de la Provincia de Jujuy (Anexo I) y el Comité de Ética IMBICE (Instituto Multidisciplinario de Biología Celular) (Anexo II). Las muestras se obtuvieron a través del consentimiento informado de los donantes (Anexo III). Para mantener la confidencialidad, las muestras llegan al laboratorio codificadas por lugar de proveniencia y sexo, de manera de ser resguardada la identidad del voluntario. Se han devuelto a los participantes sus resultados individuales de ancestría continental y a las comunidades las proporciones poblacionales de las mismas.

Todas las muestras biológicas se acompañan de una encuesta genealógica, en que el voluntario revela su lugar de nacimiento, el lugar de nacimiento de su linaje materno y de su linaje paterno (Origen Paterno), así como su procedencia socio-étnica (si tienen conocimiento de ancestría nativa americana o foránea) y el conocimiento o no, de alguna lengua diferente al español (Anexo IV). Con estos datos se descartan personas emparentadas de modo tal que sus linajes no fueran analizados en forma duplicada.

En el caso de las muestras de sangre periférica, a partir de 15 ml de sangre proporcionada por el voluntario, se extrajo el ADN con el método de "salting out" [104]. En el caso de las muestras de saliva, el ADN se extrajo a través de un preparado comercial (Nucleo Spin Tissue, Macherey-Nagel). La determinación de los haplogrupos se realizó mediante la técnica PCR-APLP (reacción en cadena de la polimerasa-polimorfismos de longitud de los productos amplificados), diseñada por Umetsu y colaboradores [105, 106] y adaptada según Jurado-Medina y col. 2014 [107]. Los individuos del haplogrupo Q se identifican usando cebadores o primers específicos para los marcadores Q-M242, Q-M3 y Q-M346 siguiendo la técnica de PCR-APLP multiplex [107] y a través de secuenciación los marcadores Q-CTS2730 (equivalente a Q-Z780) y Q-Z19231 (equivalente a Q-F4674) [49, 108].

2.1.1 Selección de las muestras

Las muestras pertenecientes al haplogrupo Q-M3 fueron seleccionadas de acuerdo con su comportamiento en una red construída con 17 STR (YFiler) y son presentadas en el Anexo V, como una figura inédita de este trabajo. Las dos muestras de Tartagal (87FK8/TG71 y SQVCW/TG33) se seleccionaron por presentar linajes haplotípicos propios de Gran Chaco [109]. Las muestras de Jujuy se seleccionaron por formar parte de linajes haplotípicos con una distribución regional característica de la provincia de Jujuy (LD4PC/LQ60 y EKEFB/LQ6). Las muestras T4WQV/LQ41 de Jujuy, 6QHWE/SMA111 de Catamarca, TYEQC/SMA138 de Catamarca y M39DJ/LV59 de Mendoza, se seleccionaron por presentar haplotipos que se agrupaban con muestras de distinto origen geográfico [108]. Las muestras de Rio Negro (8A2QN) y Chubut (UCNEN) se seleccionaron por mostrarse aislados de otras muestras por más de 8 mutaciones y por representar linajes propios de Patagonia (datos inéditos presentados en esta tesis).

Las dos muestras Q-Z780 (Z8ZMY/Be89/229 y S8BAL/MLG132/222) también se seleccionaron en base a una red de STRs, esta red se presenta en el Anexo VI como una figura inédita de este trabajo. Esta red de haplotipos (ver Anexo VI) se ordena en tres ramas, partiendo de un ancestro que no fue encontrado en el muestreo, el cual está representado por un vector intermedio (un vector es una construcción teórica, donde la combinación de marcadores de STR que determinan el haplotipo son los que deberían estar presente para explicar la aparición de los haplotipos derivados). Las dos muestras se seleccionaron por formar parte de linajes haplotípicos ubicados en ramas diferentes. En una red completa publicada recientemente ambas muestras forman parte de uno de los 17 haplotipos encontrados en Argentina y Paraguay para este haplogrupo [49].

Para la muestra perteneciente a Q-M346* de San Juan (RUTBE), no pudieron identificarse marcadores derivados que definieran acabadamente el sub-haplogrupo en el trabajo de Jurado Medina y col. 2020 [49]. En dicho trabajo RUTBE se relacionaba con linajes haplotípicos de Medio Oriente. Fue seleccionada para la secuenciación en este trabajo para obtener más información sobre este haplogrupo.

Se seleccionaron sólo 13 muestras para la secuenciación por restricciones económicas. En la tabla 2.1 se presentan los detalles de las muestras seleccionadas.

Código-ID secuenciación	Código interno	SNP	Haplogrupo	Localización en Argentina	Fuente
8A2QN	B203	M3	Q1b1a1a	Bariloche (Rio Negro)	Este trabajo
LD4PC	LQ60	M3	Q1b1a1a	La Quiaca (Jujuy)	Jurado Medina y col. 2015
87FK8	TG71	M3	Q1b1a1a	Tartagal (Salta)	Jurado Medina y col. 2014
Z8ZMY	Be89	Z780	Q1b1a2	Belén (Catamarca)	Jurado Medina y col. 2020
SQVCW	TG33	M3	Q1b1a1a	Tartagal (Salta)	Jurado Medina y col. 2014
TYEQC	SMA138	M3	Q1b1a1a	Santa María (Catamarca)	Jurado Medina y col. 2015
T4WQV	LQ41	M3	Q1b1a1a	La Quiaca (Jujuy)	Jurado Medina y col. 2015
EKEFB	LQ6	M3	Q1b1a1a	La Quiaca (Jujuy)	Jurado Medina y col. 2015
RUTBE	SJN49	M346	Q1b	San Juan (San Juan)	Jurado Medina y col. 2020
M39DJ	LV59	M3	Q1b1a1a	Lavalle (Mendoza)	Jurado Medina y col. 2015
6QHWE	SMA111	M3	Q1b1a1a	Santa María (Catamarca)	Jurado Medina y col. 2015
UCNEN	TEH26	M3	Q1b1a1a	El Chalía (Chubut)	Este trabajo
S8BAL	MLG132	Z780	Q1b1a2	Malargüe (Mendoza)	Jurado Medina y col. 2020

Tabla 2.1. Detalle de las trece muestras seleccionadas para la secuenciación completa NGS (Next-Generation) de cromosoma Y.

2.2 Secuenciación completa de cromosoma Y

El ADN de las trece muestras seleccionadas fueron enviados a secuenciar a la compañía Full Genomes Corporation (FGC) [110], ubicada en Estados Unidos, para la secuenciación completa del cromosoma Y. La secuenciación fue realizada con el servicio "Y Elite 2.1" que utiliza el equipo Illumina HiSeq 4000 [111]. Las características del servicio Y Elite 2.1 son:

- Ofrece una longitud de lecturas finales de 150 pb y una secuenciación paired-end (del inglés, extremos emparejados). La secuenciación paired-end, secuencia cada uno de los fragmentos (de 150 pb) desde ambos extremos, produciendo el doble de lecturas en el mismo tiempo que la secuenciación single-read (simple lectura). La secuenciación paired-end permite una alineación más precisa y la capacidad de detectar con más precisión variantes de inserción-eliminación (indel) y elementos de secuencia repetitivos [112], útil para los pasos posteriores de limpieza y filtros de secuencias.
- Ofrece una cobertura (en inglés, coverage o depth) de secuenciación de 30x. La cobertura es el número medio de lecturas de cada nucleótido en la secuencia reconstruida, en nuestro caso, una media de 30 lecturas por nucleótido.

2.2.1 Principios de la técnica de Secuenciación NGS

La secuenciación de próxima generación (NGS) es una técnica que permite "leer" el contenido del material genómico. Sus características principales son un alto rendimiento a un precio relativamente bajo por base secuenciada. Como inconveniente, hay un alto costo inicial para la máquina, una tasa de error no despreciable, la necesidad de una infraestructura especializada y conocimientos de bioinformática para analizar los datos generados. Explicaremos el principio

básico de la tecnología de secuenciación Illumina [113], que fue la plataforma utilizada en nuestro estudio.

Después de extraer el ADN, este se corta y se fragmenta por medio de sonicación. Según el protocolo de preparación de muestras Illumina, los fragmentos se unen a adaptadores (~ 130 pb de longitud) que se unirán a una celda de flujo, el medio que contiene las moléculas de ADN para la secuenciación (Figura 2.1.a). El ADN fragmentado se selecciona por tamaño con electroforesis: el ADN se hace correr en un gel de agarosa bajo un campo eléctrico que moverá las moléculas a diferentes velocidades de acuerdo con su masa. Mediante marcadores adecuados, el gel se corta para seleccionar fragmentos de ADN de una longitud alrededor de 370 pb de longitud. Luego se realiza un paso de enriquecimiento por PCR. Los grupos de fragmentos son luego cargados en la celda de flujo propia, con un instrumento llamado cBot. Cada celda de flujo es un soporte de vidrio que contiene 8 canales llamados "carriles" que están llenos con una matriz de adaptadores para ligar los fragmentos. Los carriles son legibles de manera independiente, por lo tanto, se pueden verter hasta 8 muestras diferentes (o mezclas) para la secuenciación.

Una vez que los fragmentos se unen a los adaptadores en la celda de flujo, se unen, se extienden y se duplica el fragmento de ADN original, que en el otro lado contiene otro adaptador que es reconocido por otros adaptadores en la celda de flujo. Este proceso continúa varias veces y se crean localmente múltiples copias alrededor del fragmento original. Esto toma el nombre de PCR puente. Esto permite crear puntos de copias idénticas de los fragmentos de ADN que queremos secuenciar. Antes de comenzar la secuenciación, el ADN se desnaturaliza obteniendo grupos de fragmentos de ADN monocatenarios. Todo este proceso se denomina generación del clúster y se ilustra en la Figura 2.1.b.

Los fragmentos están listos para ser secuenciados. El proceso químico se llama secuenciación por síntesis química (SBS, del inglés sequencing-by-synthesis), ver figura 2.1.c. El proceso se realiza en ciclos y en el enésimo ciclo, cada fragmento de cada grupo se unirá a un nucleótido marcado fluorescentemente complementario a la enésima base a partir del adaptador. El nucleótido fluorescente tiene un terminador reversible que bloquea una mayor extensión del fragmento de ADN. La celda de flujo se ilumina luego con láseres con diferentes longitudes de onda correspondientes a las energías de excitación de los marcadores de fluorescencia transportadas por las bases incorporadas. Cada uno de los cuatro nucleótidos tiene su propia longitud de onda de fluorescencia diferente. Para cada una de las cuatro longitudes de onda, se toma una foto y se procesa el software. Para cada ciclo, el software calcula las coordenadas de los clústers y su contenido de base.

La fluorescencia luego se corta y también el terminador se desbloquea. Comienza el ciclo con una nueva capa de nucleótidos fluorescentes que llega a leer las bases en las secuencias contenidas en los grupos. Después de 100 ciclos se recopila una pila de lecturas de base para cada clústers, que se linealiza, dando las secuencias 3'-5 'de los primeros 100 pb de fragmentos.

Después de ese paso, se realiza el llamado "pair-end", en donde se realiza otra PCR puente y los fragmentos son invertidos y se realizan otros 100 ciclos o secuencias.

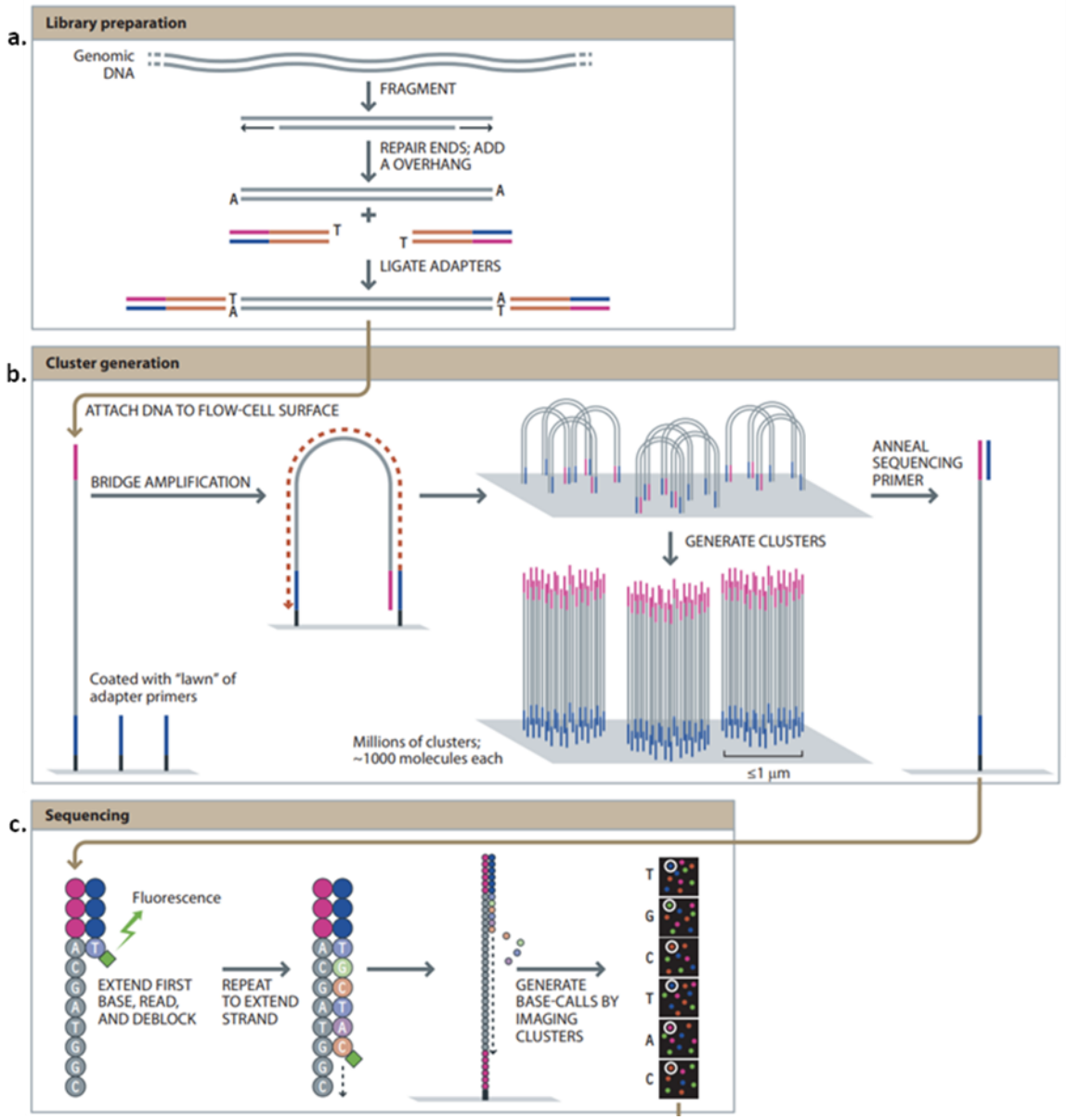


Figura 2.1. Pasos de secuenciación de Illumina: a) Preparación de la biblioteca, b) Generación Clúster, c) Secuenciación.

2.3 Resultados de la Secuenciación NGS

Durante la secuenciación por síntesis química, para cada uno de los clusters, las llamadas de bases son realizadas y almacenadas para cada ciclo de secuenciación por un software de análisis en tiempo real (Real-Time Analysis, RTA). Este software almacena los datos de llamadas de bases en forma de archivos BCL individuales (base call file). Cuando toda la secuenciación se completa, las

llamadas de bases en los archivos BCL deben convertirse en datos de secuencia. Este proceso se llama conversión de BCL a FASTQ [114] y fue realizado por la compañía FullGenomes Corp.

En el presente trabajo empezamos el manejo de secuencias a partir de este punto ya que recibimos todos los resultados de la secuenciación NGS en formato FASTQ. Para cada una de las trece muestras recibimos dos archivos en formato ".fastq" debido a que, en la secuenciación paired-end se crean un archivo "R1.fastq" y otro "R2.fastq" por cada muestra, con lecturas correspondientes a ambos extremos pareados. Además, los archivos .fastq fueron recibidos sin adaptadores de secuencia en los extremos y con un peso de aproximadamente 4.5 GB por archivo. Para reducir este peso los archivos .fastq son enviados comprimidos con el formato GNU zip, quedando con la extensión ".fastq.gz" y con un peso de 968,2 MB.
El formato .fastq es un archivo de texto con una forma común de almacenar lecturas de secuencia, cada lectura está representada por cuatro líneas:

- La línea 1 comienza con un carácter '@' seguida por un identificador de secuencia y una descripción opcional (que puede ser información de posicionamiento para saber el lugar en el carril de la celda de flujo).
- La línea 2 contiene la información de la secuencia de letras sin procesar, en el alfabeto de cuatro letras habitual para identificar a los nucleótidos (ATCG).
- La línea 3 comienza con un carácter "+" opcionalmente, le sigue el mismo identificador de secuencia y cualquier descripción nuevamente.
- La línea 4 representa las puntuaciones de calidad (Quality Score). Nos informa respecto a la probabilidad de que una llamada de base sea incorrecta. Cada símbolo representado con el alfabeto ASCII corresponde a un nivel de calidad en escala Phred.

```
@sequence_id
TACACCGAGACATTCCATTGCCCAGGGACGAGCCGGAGACAGATGCCTTCCTCTTATCTCAACTGCA
+
KKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKAFKKKKKKKKKKKKKFKKKKKFAFKKAFFK7<F7AFKKKKKKKKKK
```

Figura 2.2: Ejemplo de un fragmento de archivo ".fastq".

2.4 Procesamiento de datos NGS

El Procesamiento de datos de secuenciación NGS se realizó utilizando las recomendaciones de buenas prácticas (o Best Practices) de GATK (Genome Analysis Toolkik) versión 3.8.1. GATK [115, 116]. La plataforma GATK es una colección de herramientas de línea de comandos para analizar datos de secuenciación de alto rendimiento con el enfoque principal de descubrimiento de variantes.

Hemos agregado en Anexo VII, el pipeline creado basado en las buenas prácticas de GATK para el procesamiento de datos NGS de las trece muestras secuenciadas en este trabajo. El pipeline es un archivo en formato ".sh" ejecutable, que contiene múltiples comandos, que incluyen el

procesamiento de las secuencias desde el archivo en bruto ".fastq" hasta la generación de un GVCF "intermedio". El pipeline se ejecutó para cada archivo obtenido de la secuenciación (en total son 13 x 2 archivos .fastq, debido a que son 13 secuencias con resultados paired end).

En las siguientes secciones se explicarán los pasos de este flujo de trabajo, utilizando como ejemplo la muestra nombrada como LD4PC. La descripción del pipeline se da hasta la 2.4.6 titulada, "Llamado de variantes". Donde, con los resultados finales del pipeline, los archivos generados para cada una de las trece muestras se ensamblan con 32 secuencias de alta cobertura de secuenciación descargadas de las bases de datos, en pasos externos al pipeline.

También hemos obtenido de las bases de datos 56 secuencias con baja cobertura de secuenciación, que reciben otro procesamiento de datos de secuencias, explicados en la sección 2.4.8 Procesamiento de secuencias de baja cobertura de secuenciación.

El detalle del nombre de cada una de las secuencias utilizadas en este trabajo, así como su lugar de origen y referencia, se presenta como archivo adjunto de esta tesis. Para consultar mayor información sobre las mismas y debido a la extensión del contenido, se presenta un link para el acceso a la tabla nombrada como:

tabla adjunta I - información sobre las muestras.
https://docs.google.com/spreadsheets/d/1i-cnkj863o32zPFUoKfbfehI61EqjhULd4K6P-miXR8/edit?usp=sharing

Para una mejor comprensión y seguimiento de esta sección del trabajo a continuación presentamos la figura 2.3 que resume en un diagrama de flujo de trabajo la metodología utilizada.

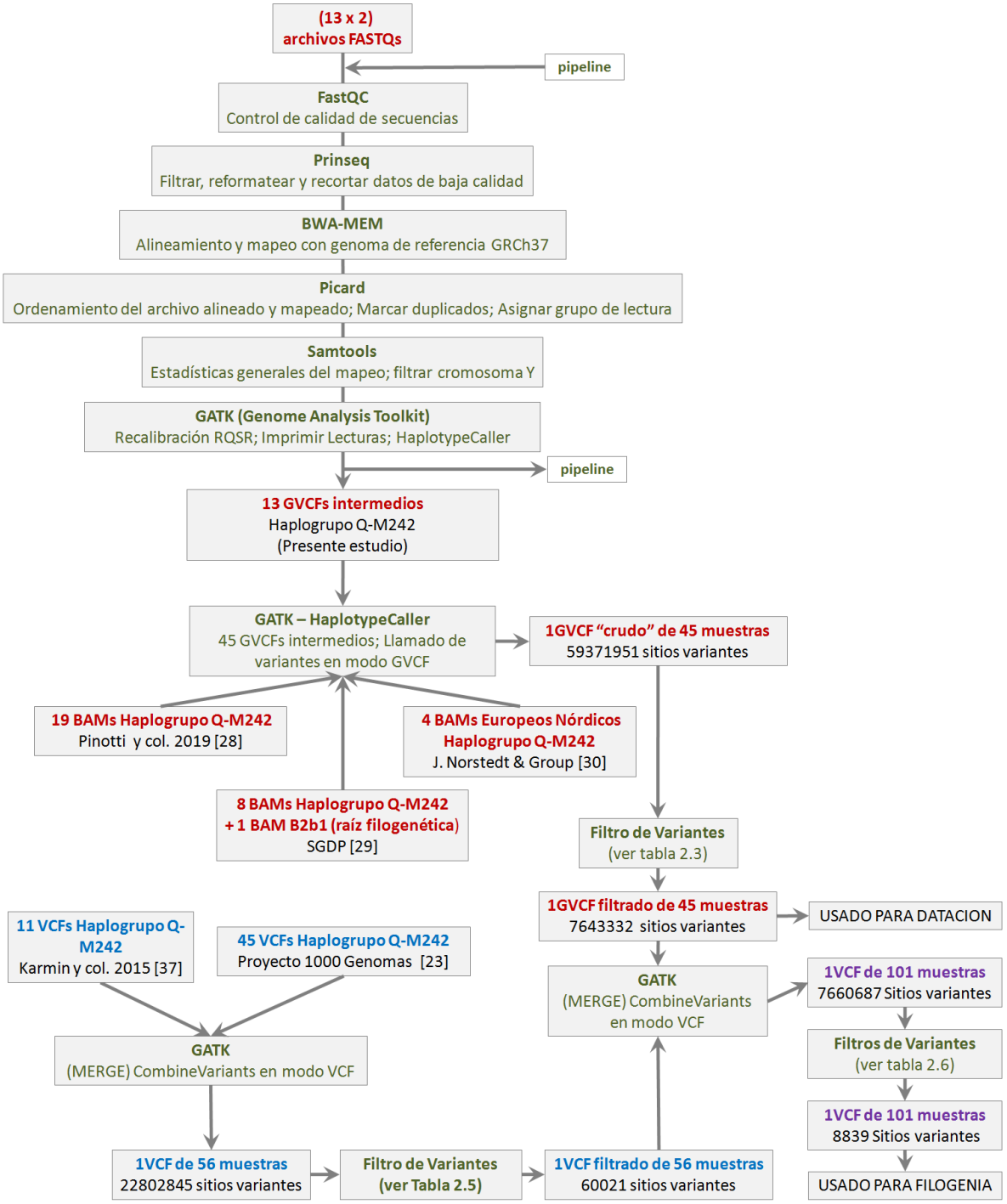


Figura 2.3. Diagrama de flujo de trabajo. En color rojo se representan las secuencias de alta cobertura de secuenciación, tanto del presente estudio, como las obtenidas de las bases de datos. En azul se representan las secuencias de baja cobertura descargadas de las bases de datos. En violeta se representan los archivos generados al unir ambas calidades de secuencia. En verde oscuro se representan los programas utilizados así como el tipo de modificación realizado en cada paso.

2.4.1 Control de la calidad de la secuenciación

El proceso de control de calidad de la secuenciación se realizó con el programa FastQC [117]. Los datos crudos en formato .fastq fueron sometidos a un control de calidad de datos de secuencia sin procesar. Los resultados proporcionan un conjunto de análisis estadísticos de las secuencias crudas, siendo necesarios para saber si los datos de la secuenciación tienen algún problema que se deba tener en cuenta antes de realizar cualquier análisis posterior.

El programa FastQC (Versión 0.11.8) fue descargado desde la plataforma de Babraham Bioinformatics [118], requiere un entorno Java adecuado y la instalación de la librería Picard, que está incluida en la descarga. Fue instalado en linux y el comando que se utilizó para cada archivo .fastq fue:

fastqc LD4PC_HNHYNCCXX_L6_1.clean.fq LD4PC_HNHYNCCXX_L6_2.clean.fq --extract

El anterior es un ejemplo para los dos archivos .fastq de la muestra LD4PC. Este comando se ejecutó para las trece muestras secuenciadas. Los resultados de los informes estadísticos se detallan en el capítulo III.

2.4.2 Filtros de datos de baja calidad de secuenciación

El software PRINSEQ [119] se utiliza como herramienta para filtrar, reformatear y recortar datos de secuenciación NGS que puedan tener baja calidad de secuenciación. Se descargó el programa prinseq-lite 0.20.4 desde [120]. El comando utilizado fue:

perl prinseq-lite.pl -fastq LD4PC_HNHYNCCXX_L6_1.clean.fq -fastq2 LD4PC_HNHYNCCXX_L6_2.clean.fq -out_good LD4PC_ QC -min_qual_mean 28 -trim_left 10 -trim_right 10 -log LD4PC_prinseq.log

Donde:

- -out_good: significa que se llamarán los resultados con buena calidad con el prefijo LD4PC_QC y se crearán dos archivos LD4PC_QC_1.fastq y LD4PC_QC_2.fastq
- -min_qual_mean 28: significa que las lecturas que tienen una calidad de base promedio inferior a 28 se filtran. Este parámetro es altamente utilizado en procesamiento de datos NGS.
- -trim_left 10: significa que se recortan las secuencias 5' terminal en un número de diez posiciones. -trim_right 10: significa que se recortan las secuencias 3' terminal en un número de diez posiciones. Siendo parámetros comúnmente utilizados en procesamiento de datos NGS [121, 122].
- -log: se refiere a que en caso de que algo haya afectado al proceso, se creará un archivo LD4PC_prinseq.log donde habrá un registro secuencial de todos los acontecimientos que pudiesen haber afectado al proceso en particular.

2.4.3 Alineamiento y mapeo contra el genoma humano de referencia

Una vez que las lecturas de secuencia fueron filtradas y obtuvimos aquellas con un umbral mínimo de calidad, se prosiguió al paso de alineamiento y mapeo en donde se ubican las secuencias en orden para reconstruir la secuencia completa del ADN amplificado. El mapeo se realiza utilizando

el genoma humano de referencia, y permite obtener un archivo que nos indica la posición que cada secuencia debería tener en el genoma humano. De esta forma se obtiene un "mapeo" (se llama así comúnmente a los alineamientos de miles de lecturas sobre una misma secuencia) con todas las lecturas superpuestas en su lugar correspondiente en el genoma de referencia.

El genoma humano de referencia utilizado fue la versión 37 (Genome Reference Consortium Human Build 37, GRCh37), obtenido en formato ".fasta" desde [123]. El formato ".fasta" es un archivo basado en texto, utilizado para representar secuencias de ADN, usando los cuatro códigos de una letra de las bases nucleicas (ACTG). El formato también permite incluir nombres de secuencias y comentarios luego del símbolo '>'.

```
>1 dna:chromosome  chromosome:GRCh37:1:1:249250621:1
ACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCA
ACCCTAACCCTAACCCTAACCTCTAACCCTAACCCTAACCCCTAACCCTAACCCTAACCCTAACCCTAACCTAACCCTAACCCTAACCC
TAACCCTAACCCTAACCCTAACCCTAACCCTAACCCCTAACCCTAACCCTAAACCCTAAACCCTAACCCTAACCCTAACCCTAACCCTA
```

Figura 2.4: Ejemplo de un fragmento de archivo ".fasta".

Para el proceso de alineamiento y mapeo utilizamos el software de Burrows-Wheeler Aligner (BWA) [124, 125]. Utilizamos el algoritmo BWA-MEM diseñado para secuencias cortas, entre 35 y 300 pb. Así, el comando utilizado para el mapeo y alineamiento fue:

bwa mem -M -t 9 human_g1k_v37_decoy.fasta LD4PC_QC_1.fastq LD4PC_QC_2.fastq > LD4PC _bwa_mem.sam > LD4PC _bwa_mem.log

Donde:

- -M: es un parámetro para que haya compatibilidad con la herramienta markDuplicates de Picard (utilizado en pasos posteriores).
- -t: es el número de procesadores del sistema operativo que se definen para correr este comando. En nuestro caso, utilizamos nueve procesadores.
- Inputs: son los dos archivos ".fastq" obtenidos del paso anterior y el archivo ".fasta" del genoma humano de referencia.
- Output: un archivo ".sam"

El resultado del alineamiento y mapeo se da en un archivo SAM [126], que significa mapa de alineamiento de secuencia (del inglés, Sequence Alignment Map). El archivo ".sam" consiste en un archivo de texto tabulado con un encabezado que contiene información general del secuenciamiento y alineamiento, así como, líneas correspondientes a cada lectura alineada. En la Figura 2.5 se muestra un ejemplo de un archivo SAM.

```
@HD VN:1.5 SO:coordinate
@SQ SN:ref LN:51304566
r001   99 ref v37 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002    0 ref v37 30 3S6M1P1I4M *  0 0 AAAAGATAAGGATA *
r003    0 ref v37 30 5S6M       * 0 0 GCCTAAGCTAA     * SA:Z:ref,29,-,6H5M,17,0;
r004    0 ref v37 30 6M14N5M    * 0 0 ATAGCTTCAGC     *
r003 2064 ref v37 17 6H5M       * 0 0 TAGGC * SA:Z:ref,9,+,5S6M,30,1;
r001  147 ref 37 30 9M        = 7 -39 CAGCGGCAT   * NM:i:1
```

Figura 2.5: Ejemplo de un fragmento de archivo ".sam". Comienza con un encabezado indicando con "@" que informa la versión del formato (VN); el tipo de ordenamiento de los alineamientos (SO, por coordenadas); y a continuación el listado de alineamientos. Por ejemplo, la primer línea especifica el nombre del fragmento "r0001"; un flag de "99" (suma de números que indican la condición del alineamiento); el nombre de la secuencia de referencia "ref"; su posición de mapeo "7" (en relación a la referencia); su calidad de mapeo de "30"; su código CIGAR "8M2I4M1D3M" (8 match seguidos, 2 inserciones, 4 match, 1 deleción y 3 match); la referencia del siguiente read en el fragmento es la misma (=) y su posición es "37"; la longitud observada del template es "39"; la secuencia del fragmento es "TTAGATAAAGGATACTG"; y por último la calidad de secuenciación (representada igual que en el formato ".fastq") en este caso no se almacenó "*".

2.4.4 Ordenamiento del archivo alineado y mapeado

Los archivos SAM tienen la característica de ser desordenados y en texto plano (el alineador realiza una línea por cada secuencia analizada según vaya apareciendo), y para que pueda ser correcta y rápidamente "leído" por futuros programas que trabajan de forma secuencial es necesario un proceso de "ordenado" por coordenadas. Además, para que el archivo ocupe menos espacio de disco (el SAM generado para la muestra LD4PC "LD4PC_bwa_mem.sam" pesa 24.5 GB) se comprime a su versión binaria e indexada BAM (".bam"), la cual ocupa menor espacio, pero tiene la desventaja de que no es posible acceder a su información por un procesador de texto (no es "human-readable" o legible por humanos).

Para realizar el ordenamiento utilizamos el programa Picard [127], el cual es un kit de herramientas de Java que presenta la función SortSam, la cual permite ordenar el archivo SAM y convertirlo a su versión binaria en BAM. El comando utilizado para esto fue:

java -jar picard.jar SortSam I=LD4PC_bwa_mem.sam O=LD4PC.sorted.bam SO=coordinate CREATE_INDEX=true > LD4PC.sortsam.log

Donde:

- SortSam: es el parámetro que ordena el archivo SAM
- I: es el input, en este ejemplo LD4PC_bwa_mem.sam
- O: es el output, en este ejemplo LD4PC_bwa_mem.sam
- SO: Orden de clasificación de salida, en este caso de coordenadas
- CREATE_INDEX=true: crea un índice para BAM generado

2.4.5 Procesamientos de archivos BAM

Los siguientes pasos serán una serie de procesamientos realizados sobre los archivos BAMs. Primero se verifican las estadísticas generales del mapeo, como la cantidad de lecturas que alinearon correctamente contra la referencia, con la herramienta "flagstat" de Samtools [128, 129]. Para esto se corre el comando:

samtools flagstat LD4PC.sorted.bam > LD4PC.flagstat_before.stats

En el siguiente paso, utilizando el parámetro "view" de Samtools junto a "REGIONS", se imprimieron todas las alineaciones asignadas a la secuencia de referencia solamente del cromosoma Y. El comando que se utilizó fue:

samtools view -bh -F 256 -f 2 LD4PC.sorted.bam $REGIONS Y > LD4PC.chrY.bam

Después de imprimir solamente las lecturas del cromosoma Y, vemos las estadísticas de mapeo solo para el cromosoma Y, que es lo que finalmente se utilizará en el análisis:

samtools flagstat LD4PC.chrY.bam > LD4PC.flagstat_after.stats

2.4.5.1 Marcar duplicados

Los múltiples procesos de PCR realizados, tanto en la generación de la librería como en la secuenciación en sí, pueden generar duplicados, y estos pueden contribuir a una profundidad de lectura ficticia. Las lecturas duplicadas pueden generar sesgos en las variantes de llamadas, corriendo el riesgo de tener una sobrerrepresentación de las secuencias en algunas áreas. Los duplicados por lo general no son eliminados, pero son identificados y marcados con una bandera, “FLAG” en los archivos SAM o BAM para que los programas llamadores de variantes (pasos posteriores) no los tomen en cuenta a la hora de establecer los genotipos para cada variante. El programa Picard [127] presenta herramientas que permiten marcar los duplicados de PCR del archivo “.bam”. El comando que se utilizó fue:

java -jar picard.jar MarkDuplicates I=LD4PC.sorted.bam O=LD4PC.dedupped.bam M=LD4PC.dedupped.metrics.txt > LD4PC.markduplicates.log

Donde:

- MarkDuplicates: es la herramienta que localiza y etiqueta las lecturas duplicadas en un archivo BAM o SAM.
- I= es el input generado en el paso anterior, en este ejemplo LD4PC.sorted.bam
- O= es el output que en este caso se llamará LD4PC.dedupped.bam
- M= Archivo de salida que proporciona métricas útiles para validación de las lecturas.

2.4.5.2 Asignar grupo de lectura

Para los pasos siguientes, el programa GATK requiere que se le indique las características del grupo de lectura, e indica qué tecnología se utilizó para generarlos (muestras secuenciadas en la misma corrida, el nombre de las muestras, etc.). El comando que se utilizó fue:

java -jar picard.jar AddOrReplaceReadGroups I=LD4PC.dedupped.bam O=LD4PC.RG.bam LB=library PL=Illumina SM=LD4PC CREATE_INDEX=true > LDP4PC.addorreplacereadgroups.log

Donde:

- AddOrReplaceReadGroups es el parámetro de Picard que permite agregar o reemplazar grupos de lectura en un único grupo de lectura en un nuevo archivo BAM.
- I es el input, en este caso LD4PC.dedupped.bam
- O es el output, en este caso el archivo de salida se llamará LD4PC.RG.bam
- LB Este parámetro es Library o Biblioteca. Permite identificar la biblioteca de preparación de ADN. Es necesario para determinar qué grupos de lectura podrían contener duplicados moleculares.
- PL es la tecnología utilizada para generar las lecturas, en este caso, Illumina.
- SM es el nombre de la muestra, en este ejemplo, LD4PC. Se debe especificar correctamente el nombre de la muestra secuenciada en este grupo de lectura. Las herramientas GATK tratan a todos los grupos de lectura con el mismo SM que contienen datos de secuencia para la muestra, SM. Este paso es fundamental, especialmente cuando se usan herramientas de muestras múltiples como el Genotipo Unificado (GVCF, que se verá más adelante).
- CREATE_INDEX=true Este parámetro crea un índice para el BAM generado. Necesario para utilizar ese BAM generado para los pasos siguientes.

2.4.5.3 Recalibración del nivel de calidad de base

La recalibración del nivel de calidad de base (Base Quality Score Recalibration, BQSR), es un paso de pre-procesamiento de datos que detecta errores sistemáticos cometidos por el equipo de secuenciación cuando estima la precisión de cada llamada de base. Los puntajes de calidad de base son estimaciones de error por base emitidas por los equipos de secuenciación y expresan cuán seguro estaba el equipo de llamar a la base correcta cada vez. Los puntajes de calidad son expresados por el equipo en escala de Phred, y son importantes debido a que nuestros algoritmos de llamadas de variante (utilizados en pasos posteriores) dependen en gran medida del puntaje de calidad asignado a las llamadas de base individuales en cada secuencia de lectura.

Para realizar este paso utilizamos las herramientas de análisis del genoma de la plataforma GATK (Genome Analysis Toolkit) [115, 116]. La función “BaseRecalibrator" de GATK genera un modelo de la covariación de los datos provistos con un set de variantes conocidas. El set de variantes conocidas fué descargado de las bases de datos del Proyecto 1000 Genomas fase I, tanto indels como SNPs [130]. El comando utilizado fue:

java -jar GenomeAnalysisTK.jar -nct 4 -T BaseRecalibrator -I LD4PC.RG.bam -R human_g1k_v37_decoy.fasta -knownSites 1000G_phase1.indels.b37.vcf -knownSites Mills_and_1000G_gold_standard.indels.b37.vcf -knownSites dbsnp_137.b37.vcf -o LD4PC.recal.table > LD4PC.baserecalibrator.log

Donde:

- -nct: es el número de procesadores de la CPU que se definen para correr este comando. En nuestro caso, 4.
- -T BaseRecalibrator: este parámetro genera una tabla de recalibración que será utilizada en un paso posterior.
- -I: input, en este ejemplo, LD4PC.RG.bam.
- -knownSites: archivo con sitios conocidos descargado del Proyecto 1000 genomas fase I, 1000G_phase1.indels.b37.vcf
- -knownSites: archivo con sitios conocidos descargado del Proyecto 1000 genomas fase I, Mills_and_1000G_gold_standard.indels.b37.vcf
- -knownSites: archivo con sitios SNPs conocidos descargado del Proyecto 1000 genomas fase I, dbsnp_137.b37.vcf
- -o: output, en este ejemplo llamado, LD4PC.recal.table

2.4.5.4 Imprimir Lecturas

En este paso se ajusta la calidad de las bases en la muestra basándose en el modelo creado del paso anterior, BQSR (Base Quality Score Recalibration). Se aplica un filtro llamado, filtro de lecturas bien llamadas (WellformedReadFilter) el cual comprueba si una lectura está "bien formada", es decir, está libre de grandes inconsistencias internas y problemas que podrían conducir a errores posteriores. Solamente se "imprimen las lecturas" (PrintReads) que pasan el filtro, creando un nuevo archivo BAM. Para este paso también usamos la plataforma GATK, el comando que se utilizó fue:

java -jar GenomeAnalysisTK.jar -nct 4 -T PrintReads -R human_g1k_v37_decoy.fasta -I LD4PC.RG.bam -BQSR LD4PC.recal.table -o LD4PC.recal.bam > LD4PC.printreads.log

Donde:

- -nct: es el número de procesadores de la CPU que se definen para correr este comando. En nuestro caso, 4.
- -T PrintReads: parámetro que imprime las lecturas que pasan por el filtro "WellformedReadFilter" utilizando el archivo LD4PC.recal.table generado en el paso anterior de BQSR.
- -R: es la referencia del genoma humano, en este ejemplo human_g1k_v37_decoy.fasta.
- -I: input, es el último archivo bam generado en procesos anteriores, LD4PC.RG.bam.
- -BQSR: parámetro recalibración del nivel de calidad de base, se usa el archivo creado en el paso anterior, LD4PC.recal.table.
- -o: Otput, en este caso fué nombrado, LD4PC.recal.bam

2.4.6 Llamado de variantes

El llamado de variantes es el proceso por el cual se analizan las regiones donde se presentan variaciones con respecto al genoma de referencia y se seleccionan aquellas que cumplan con ciertos criterios que las hagan elegibles como “variantes verdaderas” (es decir, distinguir las

variantes y los errores de secuenciación), tales como la calidad de la base secuenciada, calidad de mapeo y el número de lecturas independientes que den evidencia a favor de su presencia. Esta información se vuelca en un tipo de archivo llamado VCF (por sus siglas en inglés, Variant Call Format) [131], y fue introducido por el consorcio responsable del Proyecto 1000 Genomas [130].
Los archivos VCFs consisten de un archivo de texto tabulado en el cual, al eliminarse toda la información redundante con el genoma de referencia, sólo se obtiene información sobre las particularidades del ADN analizado, lo que lo hace más compacto y fácil de leer (ver figura 2.6).

Para el proceso de llamado de variantes, utilizamos la herramienta “HaplotypeCaller” de GATK siguiendo las prácticas recomendadas del Broad Institute [132, 133]. La herramienta "HaplotypeCaller" es capaz de llamar a los SNPs e indels simultáneamente a través del ensamblaje local *de novo* de haplotipos en una región activa. En otras palabras, cada vez que el programa encuentra una región que muestra signos de variación, descarta la información de mapeo existente y vuelve a ensamblar completamente las lecturas en esa región. Esto permite que HaplotypeCaller sea más preciso al llamar a regiones que tradicionalmente son difíciles de llamar, por ejemplo, cuando contienen diferentes tipos de variantes cercanas entre sí.

Debido a que en este trabajo se pretende construir un árbol filogenético, es necesario realizar el llamado de variantes para todas las muestras en conjunto, por un lado, porque GATK usa la información de todas las muestras para determinar las variantes (por ejemplo, variaciones compartidas entre muestras probablemente sean consideradas variantes verdaderas) y, por otro lado, para poder comparar variantes entre muestras. Para realizar esto, se trabaja con un tipo de VCF que contiene información adicional, llamado VCF genómico o GVCF (Genomic Variant Call Format), el cual es un VCF multi-muestra (multisample), que contiene información de todas las variantes de todas las muestras.

La construcción del GVCF, se realiza ejecutando primero la herramienta HaplotypeCaller, muestra por muestra para generar un GVCF intermedio (que no se utilizará en el análisis final), pero que luego se utiliza la herramienta "GenotypeGVCF" para generar un genotipado conjunto de múltiples muestras en una forma muy eficiente.

```
##fileformat=VCFv4.2
##ALT=<ID=NON_REF,Description="Represents any possible alternative allele at this location">
##FILTER=<ID=LowQual,Description="Low quality">
##FILTER=<ID=NOCALL,Description="Some or all of this record had no sequence call by Complete Genomics">
##FILTER=<ID=NONVARIANT,Description="GSNONVARSCORE != NA && GSNONVARSCORE >= 13.0">
##FILTER=<ID=PASS,Description="All filters passed">
##FORMAT=<ID=AD,Number=R,Type=Integer,Description="Allelic depths for the ref and alt alleles in the order listed">
##FORMAT=<ID=CN,Number=1,Type=Integer,Description="Copy number genotype for imprecise events">
##FORMAT=<ID=CNL,Number=.,Type=Float,Description="Copy number likelihoods with no frequency prior">
##FORMAT=<ID=CNP,Number=.,Type=Float,Description="Copy number likelihoods">
##FORMAT=<ID=CNQ,Number=1,Type=Float,Description="Copy number genotype quality for imprecise events">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Approximate read depth (reads with MQ=255 or with bad
mates are filtered)">
##FORMAT=<ID=FT,Number=.,Type=String,Description="Genotype-level filter">
##FORMAT=<ID=GP,Number=G,Type=Float,Description="Genotype likelihoods">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=PL,Number=G,Type=Integer,Description="Normalized, Phred-scaled likelihoods for genotypes as defined
in the VCF specification">
##reference=file:///home/marina/Paula/Reference_seq/HGREF/human_g1k_v37_decoy.fasta
#CHROM POS       ID          REF       ALT       QUAL     FILTER   INFO      FORMAT LD4PC
Y      2668456  rs2058276,dbsnp.94:rs2058276      T        C        51788.90 PASS
       AC=1;AF=1.00;AN=1;DP=32  GT:AD:DP:GQ:PL   1:0,32:32:99:934,0
Y      2694240  .           T         G         735.36   .        AC=1;AF=1.00;AN=1;DP=26   GT:AD:DP:GQ:PL
       1:0,26:26:99:776,0
```

Figura 2.6. Fragmento de un archivo VCF con dos posiciones variantes. Consiste de un encabezado, en el cual cada línea está precedida por "##", que puede contener información de la versión de VCF utilizada, la fecha de creación, la referencia utilizada, etc., pero además contiene información relevante a los campos encontrados en secciones como INFO, FILTER o FORMAT de una forma legible. Posteriormente hay una línea indicando que se encuentra ordenado en la región de las variantes, y está precedido por "#". Utilizando la primer línea de ejemplo, se encuentra una variante en el cromosoma (CHROM) "Y"; en la posición (POS) "2668456"; con un identificador conocido (ID), en caso que presente, en este caso "rs2058276"; existe un cambio de "T" en la referencia (REF) por una "C" en la variante (ALT), pasó los filtros (FILTER) aplicados ("PASS"). En el campo "INFO" se encuentra información separada por ";" y para la primera variante del ejemplo, AC (recuento de alelos) es uno; AF (frecuencia alélica) es 1.00; AN (número total de alelos) es uno; DP (profundidad de cobertura filtrada) es 32. El campo "FORMAT" indica el orden para el cual cada muestra tendrá información adicional separado por ":" Por lo tanto, en el ejemplo, la primera variante de la muestra LD4PC presenta en GT (Genotipo) valor de uno. En caso de llamadas haploides, como es el cromosoma Y, el GT solo tiene dos valores, 1 si presenta el alelo ALT y 0 si no presenta el alelo ALT; AD (profundidad del alelo sin filtrar) es 0.32; DP (profundidad de cobertura filtrada) en este caso 32; GQ (calidad del genotipo) confianza en escala Phred de que la asignación del genotipo (GT) es correcta (GT=1 -ProbError), en este caso el valor 99 es la máxima confianza que asigna GATK, si el valor fuese 60 habría más del 99,9999% de probabilidad que el alelo ALT sea correcto [134]; PL (likelihoods "normalizadas" según GATK) en escala Phred de los posibles genotipos, donde se le asigna 0 al genotipo más probable y se ajusta al resto en función, en este caso 943,0 , la variante ALT es la más probable y la REF es menos probable (likelihood en este caso es que tan probable es que el genotipo NO sea correcto.)

El comando que se utilizó fue:

java -jar GenomeAnalysisTK.jar -R human_g1k_v37_decoy.fasta -T HaplotypeCaller -L Y -I LD4PC.recal.bam -o LD4PC.haplotypecallerGVCF.g.vcf --emitRefConfidence BP_RESOLUTION -ploidy 1 > LD4PC.haplotypecallerGVCF.log

Donde:

- -R: es la referencia del genoma humano, en este ejemplo human_g1k_v37_decoy.fasta.
- -T HaplotypeCaller: especifica a GATK que utilice esa herramienta para generar el GVCF "intermedio".
- -I: input en este caso el input es el último .bam generado en el paso anterior.
- -o: el output será el GVCF "intermedio" que se utilizará en el siguiente paso para llamar las variantes en conjunto, llamado en este ejemplo LD4PC.haplotypecallerGVCF.g.vcf
- --emitRefConfidence BP_RESOLUTION: parámetro que emite puntajes de confianza de referencia, dando una estimación de confianza resumida por pb para un sitio.
- -ploidy 1: especifíca el número de cromosomas por muestra, en nuestro caso 1 al ser cromosoma Y.
- -L: este parámetro representa los intervalos genómicos sobre los cuales operar. En nuestro caso especificamos solamente los intervalos del cromosoma Y.

2.4.6.1 Procesamiento de secuencias de alta cobertura obtenidas de las bases de datos y llamado de variantes

La definición actual que se tiene de la filogenia del haplogrupo Q-M242 es baja, y esto se atribuye al bajo número de muestras secuenciadas que se tienen para este haplogrupo. En este trabajo se realizó una búsqueda en las bases de datos de secuencias del cromosoma Y pertenecientes al haplogrupo Q-M242 y así construir un árbol filogenético que incluya la mayor cantidad posible de secuencias de este haplogrupo. El Proyecto de Diversidad Genómica Simons (SGDP, del inglés Simons Genome Diversity Project) ha realizado la secuenciación genómica completa de más de cien poblaciones humanas diversas, incluimos algunas secuencias del haplogrupo Q de este proyecto en este estudio, junto a otras de otros autores que se detallan en la tabla 2.2.

Una parte de las muestras descargadas de la bibliografía están disponibles con una cobertura de secuenciación alta y en formato ".bam" (tabla 2.2) y otras están disponibles en una baja cobertura de secuenciación, descriptas en la tabla 2.4. La diferencia en la cobertura de secuenciación requiere procesamientos de datos NGS diferentes, por lo que en esta sección será detallado el procesamiento para el caso de las secuencias de alta cobertura de secuenciación (~30x).

Secuencias de alta cobertura del Cromosoma Y	Fuente
19 secuencias haplogrupo Q-M242	Pinotti y col. 2020 [31]
8 secuencias haplogrupo Q-M242	SGDP [135]
4 secuencias europeas nórdicas haplogrupo Q-M242	J. Norstedt & Group [136]
1 secuencia haplogrupo B1b1 para raíz filogenética	SGDP [135]

Tabla 2.2: Secuencias de alta cobertura de secuenciación descargadas de las bases de datos.

Para poder ensamblar las 13 muestras secuenciadas en el presente trabajo con las 32 muestras de alta cobertura descargadas de las bases de datos (en formato ".bam") y realizar el llamado de variantes en conjunto, primero generamos los "index" (necesarios para el procesamiento de los BAMs) de cada una, utilizando la plataforma Samtools [129, 137]. Para cada muestra se corrió el comando:

samtools index Nmuestra.sorted.bam

Luego se generó el GVCF intermedio para cada una de las 32 muestras. Por lo tanto, se corrió el comando de la sección 2.4.6 anterior, generando 32 archivos "NNN.haplotypecallerGVCF.g.vcf".

Luego se realizó el llamado de variantes en conjunto. Para esto, se corrió el comando:
java -jar GenomeAnalysisTK.jar -R human_g1k_v37_decoy.fasta -T GenotypeGVCFs -G StandardAnnotation -o 44samples.genotypeGVCF.vcf -allSites --variant nombredelamuestra.haplotypecallerGVCF.g.vcf > nombredelamuestra.haplotypecallerGVCF.log

Donde:
- -R human_g1k_v37_decoy.fasta: es el genoma humano de referencia.
- -T GenotypeGVCF: herramienta que permite realizar el genotipado en conjunto en una o más muestras desde el GVCF "intermedio" generado anteriormente con HaplotypeCaller.
- -G StandardAnnotation: proporciona el set standar de anotaciones para cada variante.
- -o: es el output que fue nombrado como multisample44.genotypeGVCF.vcf
- -allSites: proporciona información de todos los sitios, no solo de las variantes. Importante para cuando se juntan todas las variantes en un archivo.
- --variant: por simplificación para que el comando no quedara demasiado largo, se resumió. Pero aquí iría "--variant nombredelamuestra.haplotypecallerGVCF.g.vcf" para cada uno de los 45 GVCFs "intermedios" generados.

El resultado es un único archivo VCF nombrado como ejemplo, 44samples.genotypeGVCF.vcf, en el que todas las muestras se han genotipado conjuntamente. Este VCF suele ser llamado "crudo" y presenta un total de 59371951 sitios variantes.

2.4.7 Filtros de variantes de secuencias de alta cobertura de secuenciación

Para construir un árbol filogenético robusto y calibrado es necesario seguir una serie de pasos adicionales de filtrado para "limpiar" el VCF "crudo" generado. Los parámetros de filtros establecidos en este trabajo fueron fijados siguiendo tanto las recomendaciones de buenas prácticas genómicas de GATK (Best Practices) [116] como a las referentes a la manipulación de secuencias genómicas del cromosoma Y [8, 22, 30].

2.4.7.1 Eliminar regiones altamente repetitivas del Cromosoma Y

Como se explicó en la sección 1.2, el cromosoma Y presenta una longitud total aproximada de 60 Mb, pero solamente en un fragmento de una longitud de aproximadamente 10 Mb pueden ser

describirse sitios únicos, sin ambigüedades. Por lo tanto, aquí seleccionamos solamente estas regiones únicas para la búsqueda de variantes filogenéticas informativas, siguiendo las recomendaciones de [8]. El comando que se utilizó para esto fue:

java -jar GenomeAnalysisTK.jar -T SelectVariants -R human_g1k_v37_decoy.fasta --variant 44samples.genotypeGVCF.vcf -o REG.genotypeGVCF.vcf -L REGY.bed > REG.genotypeGVCF.log

Donde:
- -T SelectVariants esta herramienta de GATK permite seleccionar un subconjunto de variantes en función de diversos criterios.
- -R human_g1k_v37_decoy.fasta es la referencia del genoma humano.
- --variant es el input, en este caso es el VCF multimuestra generado 44samples.genotypeGVCF.vcf
- -o REG.genotypeGVCF.vcf es el nombre del output
- -L REGY.bed es un archivo que presenta las coordenadas únicas del cromosoma Y, obtenido de la referencia [29].

El archivo REG.genotypeGVCF.vcf presenta un total de 10445027 sitios variantes.

2.4.7.2 Eliminar indels

El árbol filogenético que se construyó se basó en polimorfismos bialélicos (SNPs) por lo que excluimos todos los indels en este paso. Para esto se corrió el comando:

java -jar GenomeAnalysisTK.jar -T SelectVariants -R human_g1k_v37_decoy.fasta --variant REG.genotypeGVCF.vcf --selectTypeToExclude INDEL -o noindels.genotypeGVCF.vcf > noindels.genotypeGVCF.log

Donde:
- -T SelectVariants: esta herramienta de GATK permite seleccionar un subconjunto de variantes en función de diversos criterios.
- -R human_g1k_v37_decoy.fasta: es el genoma humano de referencia.
- --variant REG.genotypeGVCF.vcf: es el VCF generado en el paso anterior, contiene las regiones únicas de interés del cromosoma Y.
- --selectTypeToExclude: este argumento permite excluir tipos particulares de variantes, en este caso indels.
- -o: el output, nombrado como noindels.genotypeGVCF.vcf

El archivo "noindels.genotypeGVCF.vcf" presenta un total de 10442285 sitios variantes.
En este paso es necesario chequear rigurosamente la ausencia de todos los indels del archivo VCF. Como algunos no fueron eliminados con el paso anterior, fueron encontrados manualmente con las herramientas "nano", "awk" y "sed" de linux y luego fueron anotados en un archivo ".txt" el cual presenta una única columna de posiciones. Finalmente se eliminaron utilizando el software VCFtools versión 0.1.15 [138], el cual presenta un conjunto de herramientas para manipular archivos VCF.

vcftools --vcf noindels.genotypeGVCF.vcf --exclude-positions indels_extras.txt --out noindels2.genotypeGVCF --recode

- Donde:
- --vcf: es el input, en este caso el archivo generado en paso anterior noindels.genotypeGVCF.vcf
- --exclude-positions: parámetro que elimina las posiciones especificadas en un archivo de texto, en este caso indels_extras.txt.
- --out el nombre del output, en este caso noindels2.genotypeGVCF
- --recode parámetro para generar de nuevo un archivo VCF que será llamado noindels2.genotypeGVCF.recode.vcf

El archivo noindels2.genotypeGVCF.recode.vcf un total de sitios variantes de 10441434.

2.4.7.3 Eliminar variantes "perdidas"

Las llamadas variantes "perdidas" o "missingness" surgen en la secuenciación debido a la falta de cobertura de secuenciación en algunas regiones cromosómicas. En la práctica, se eliminan las variantes que presentan una frecuencia de datos faltantes mayor a 0.1 [30].
Para realizar esto primero se genera un reporte de todos los sitios que presentan variantes perdidas con sus respectivas frecuencias, utilizando el parámetro "missing-site" de VCFtools.

vcftools --vcf noindels2.genotypeGVCF.recode.vcf -- VCFTools -missing-site

Donde:
- --vcf: es el input, en este caso es el archivo generado en el paso anterior noindels2.genotypeGVCF.recode.vcf.
- --missing-site: parámetro de VCFtools que genera un archivo llamado "out.lmiss" que informa las variantes faltantes por sitio.

El archivo "out.lmiss" generado es un ".txt" y presenta información por posición sobre la frecuencia de "variante perdida". (El archivo out.lmiss presenta un total de 10441435 posiciones, de las cuales 2798102 posiciones presenta valores de frecuencia de datos faltantes mayores a 0.1, esto representa la eliminación de un 30% del total de datos faltantes. Se conserva un 70% de datos faltantes con frecuencia mayor a 0.1).
Este archivo "out.lmiss" es manualmente ordenado de mayor a menor según la columna de la frecuencias y se genera un nuevo archivo de texto que presenta una única columna con las posiciones con frecuencia mayor a 0.1, nombrado "miss_0.1.txt", que presenta las posiciones que queremos eliminar. Luego con VCFtools corremos el comando:

vcftools --vcf noindels2.genotypeGVCF.recode.vcf --exclude-positions miss_0.1.txt --out alta_cobertura_limpio --recode

Donde:
- --vcf es el input, en este caso es el último VCF generado, noindels2.genotypeGVCF.recode.vcf

- --exclude-positions es un parámetro de VCFtools que permite eliminar posiciones ingresando un ".txt" que presenta únicamente una columna de posiciones, en este caso el archivo creado anteriormente miss_0.1.txt
- --out es el output, debemos especificar que nombre queremos que tenga el archivo de salida, en este caso, alta_cobertura_limpio
- --recode parámetro de VCFtools que genera un archivo VCF y lo nombrará como, alta_cobertura_limpio.recode.vcf

El archivo "alta_cobertura_limpio.recode.vcf" presenta 7643332 sitios variantes.

	Número sitios de variantes
44samples.genotypeGVCF.vcf "crudo" inicial	59371951
Eliminar regiones repetitivas (GATK SelectVariants, regiones únicas ChrY)	10445027
Eliminar indels (GATK -selectTypeToExclude INDEL)	10442285
Eliminar indels manual (VCFTools exclude-positions)	10441434
Eliminar por "missigness" > a 0.1 (VCFTools -missing-site)	7643332

Tabla 2.3. Resumen de filtros aplicados en secuencias de alta cobertura de secuenciación y números de variantes conservadas por filtro aplicado.

2.4.8 Procesamiento de secuencias de baja cobertura de secuenciación

Vista la dificultad en el procesamiento de conjunto de datos de alta y baja cobertura de secuenciación se eligió procesarlos de manera separada. En esta sección se ensamblan todas las secuencias Q-M242 de baja cobertura (~5x) que pudieron ser descargadas de la bibliografía. La tabla 2.4 detalla el número de secuencias descargadas y la fuente, y en la tabla adjunta I - información sobre las muestras, se amplía la información sobre las muestras.

Secuencias del haplogrupo Q-M242	**Fuente**
45 secuencias de cromosoma Y	Proyecto 1000 Genomas [130]
11 secuencias de cromosoma Y	M. Karmin y col. [23]

Tabla 2.4. Secuencias de baja cobertura de secuenciación descargadas de la referencia.

Las secuencias de baja cobertura descargadas están disponibles únicamente en formato VCF y lo primero que se hizo fue unirlas en un único archivo para poder compararlas de manera más fácil. Para esto se utilizó la plataforma GATK y se corrió el siguiente comando:

java -jar GenomeAnalysisTK.jar -T CombineVariants -R human_g1k_v37_decoy.fasta --variant cada_muestra_baja_cobertura.vcf -o 1000g_karmin.vcf -genotypeMergeOptions UNIQUIFY

Donde:

- -T CombineVariants: herramienta de GATK que permite generar un único archivo VCF desde varias muestras procedentes de flujos de trabajo diferentes. Esta es una herramienta

"inteligente" que es capaz de fusionar tratando las mismas muestras por separado o no, combinando anotaciones según corresponda.
- -R human_g1k_v37_decoy.fasta: genoma humano de referencia.
- --variant: son los inputs, en este caso por simplificación de comando se resumió, pero es un "--variant cada_muestra_baja_cobertura.vcf" por cada una de las 56 muestras de baja cobertura.
- -o: el nombre del output que en este caso es 1000g_karmin.vcf
- -genotypeMergeOptions UNIQUIFY parámetro que permite unificar genotipos.

El archivo generado 1000g_karmin.vcf presenta un total de 56 muestras y un total de 22802845 sitios variantes.

2.4.8.1 Filtros de variantes en secuencias de baja cobertura de secuenciación

Se realizaron una serie de filtros de limpieza de variantes sobre el archivo anteriormente generado. Como estos pasos son los mismos descriptos en la sección 2.4.7, acá simplificaremos la parte explicativa y solamente dejaremos constancia de los comandos y el registro del número de variantes que se fueron eliminando en cada paso.

2.4.8.1.1 Eliminar regiones altamente repetitivas del Cromosoma Y

java -jar GenomeAnalysisTK.jar -T SelectVariants -R human_g1k_v37_decoy.fasta --variant 1000g_karmin.vcf -o 1000g_karmin_sinposnik.genotypeGVCF.vcf -L REGY.bed > 1000g_karmin_sinposnik.genotypeGVCF.log

El archivo generado 1000g_karmin_sinposnik.genotypeGVCF.vcf presenta 10445971 sitios variantes.

2.4.8.1.2 Eliminar indels

java -jar GenomeAnalysisTK.jar -T SelectVariants -R human_g1k_v37_decoy.fasta --variant 1000g_karmin_sinposnik.genotypeGVCF.vcf --selectTypeToExclude INDEL -o 1000g_karmin_sinposnik_indell.genotypeGVCF.vcf > 1000g_karmin_sinposnik_indell.genotypeGVCF.log

El archivo generado 1000g_karmin_sinposnik_indell.genotypeGVCF.vcf presenta 10444021 sitios variantes. Sobre este VCF obtenido se hizo una búsqueda manual de indels no eliminados con las herramientas "nano", "awk" y "sed" de linux, anotando todos los indels en un ".txt" llamado: eliminar_1000g_karmin6.txt, para luego correr:

vcftools --vcf 1000g_karmin_sinposnik_indell.genotypeGVCF.vcf --exclude-positions eliminar_1000g_karmin6.txt --out 1000g_karmin_sinposnik_indell_raros.genotypeGVCF --recode

El archivo generado llamado 1000g_karmin_sinposnik_indell_raros.genotypeGVCF.recode.vcf presenta 10443805 presenta sitios variantes.

2.4.8.1.3 Eliminar variantes "perdidas" o "missingness"

vcftools --vcf 1000g_karmin_sinposnik_indell_raros.genotypeGVCF.recode.vcf --missing-site

El archivo "out.lmiss" generado con el comando anterior es procesado manualmente con herramientas "nano", "awk" y "sort" de linux para generar un archivo ".txt" que presenta una columna única de posiciones con frecuencias alélicas de datos faltantes mayores a 0.1. Las posiciones del ".txt" fueron eliminadas con el comando:

vcftools --vcf 1000g_karmin_sinposnik_indell_raros.genotypeGVCF.recode.vcf --exclude-positions recorte_1000g_karmin_missing5_mayor0.1_Y_POS.txt --out 1000g_karmin_0.1miss --recode

El archivo generado 1000g_karmin_0.1miss.recode.vcf presenta 60021 posiciones variantes.

	Número de sitios de variantes
Unión del VCF inicial sin filtros "1000g_karmin.vcf"	22802845
Eliminar regiones repetitivas (GATK SelectVariants, regiones unicas ChrY)	10445971
Eliminar indels (GATK -selectTypeToExclude INDEL)	10444021
Eliminar indels manual (VCFTools exclude-positions)	10443805
Eliminar por "missigness" > a 0.1 (VCFTools -missing-site)	60021

Tabla 2.5. Resumen de filtros aplicados en secuencias de baja cobertura de secuenciación y números de variantes conservadas por filtro aplicado.

2.4.9 Procesamiento del conjunto de todas las secuencias

2.4.9.1 Unión de todas las muestras

Una vez que los archivos de alta y baja cobertura de secuencias quedaron "limpios" de gran parte de variantes espurias, se prosiguió a unir los dos archivos. Para esto se utilizó la plataforma GATK, con el uso de los parámetros "CombineVariants" y "genotypeMergeOptions UNIQUIFY" que, como ya fueron explicados anteriormente, solo daremos especificaciones del comando y del archivo generado.

java -jar GenomeAnalysisTK.jar -R human_g1k_v37_decoy.fasta -T CombineVariants -genotypeMergeOptions UNIQUIFY --variant alta_cobertura_limpio.recode.vcf --variant 1000g_karmin_0.1miss.recode.vcf -o todas.vcf

El archivo "todas.vcf" presenta un total de 7660687 posiciones variantes.

2.4.9.2 Filtro de alelos monomórficos

Debido a que filogenéticamente los alelos monomórficos no son informativos, se prosiguió a filtrar esas posiciones del archivo. Debido a que la frecuencia de alelos monomórficos varía con la

ausencia o presencia de diferentes individuos, este filtro debe ser aplicado para todas las muestras en conjunto. Para esto se utilizó VCFtools y se corrió el comando:

vcftools --vcf todas.vcf --non-ref-ac-any 1 --out todas_listas --recode

Donde:
- --vcf: es el input, en este caso el VCF generado en el paso anterior "todas.vcf".
- --non-ref-ac-any 1: este parámetro de VCFtools permite eliminar todos los sitios donde no hay un alelo alternativo.
- --out: es el output, y fué nombrado como todas_casi_listas
- --recode: es el parámetro de VCFtools generar nuevamente un archivo VCF.

El archivo que se obtuvo, "todas_casi_listas.recode.vcf" presenta 10095 posiciones variantes.

2.4.9.3 Filtros de profundidad

En este último paso en la aplicación de filtros de variantes, nos aseguramos que todas las posiciones variantes presenten una profundidad mínima de lectura de dos, siguiendo las recomendaciones de [30, 31, 38]. Para esto utilizamos VCFtools y corrimos el siguiente comando:

vcftools --vcf todas_casi_listas.recode.vcf --out todas_listas --min-meanDP 2 --recode

Donde:
- --vcf: es el input, en este caso el archivo antes generado como "todas_casi_listas.recode.vcf"
- --out: es el output, en este caso fue nombrado como "todas_listas"
- --min-meanDP 2: este parámetro de VCFtools permite incluir solamente los sitios con valores medios de profundidad (DP) (sobre todos los individuos incluidos) mayores o iguales que el valor 2. Que están informadas en el campo "FORMAT" "DP" de todas las posiciones del VCF.
- --recode: genera un nuevo VCF nombrado como "todas_listas.recode.vcf".

El archivo "todas_listas.recode.vcf" presenta un total de 8839 sitios variantes.

	Número sitios de variantes
Unión (Merge) todas las secuencias "todas.vcf"	7660687
Filtro de alelos monomórficos (VCFtools non-ref-ac-any 1)	10095
Filtros de profundidad (vcftools --min-meanDP 2)	8839

Tabla 2.6. Resumen de filtros aplicados al archivo de unión de todas las secuencias con los números de variantes conservadas por filtro aplicado.

2.5 Construcción del árbol filogenético

La construcción del árbol filogenético se hizo en base al archivo "todas_listas.recode.vcf", el cual presenta todas las secuencias utilizadas en este trabajo, un total de 103 secuencias que incluyen, 102 pertenecientes al haplogrupo Q más una secuencia del haplogrupo B2b1 (del Congo de África,

utilizada como raíz filogenética). En la tabla 2.7 se encuentra el número de secuencias por país. En esta tabla y en el resto de este trabajo todos los individuos de Los Ángeles, Estados Unidos, que presentan un origen mexicano se consideran mexicanos, para más detalle consultar el detalle en la tabla adjunta I - información sobre las muestras.

	Número de secuencias por países
Perú	30
Argentina	25
México	16
Brasil	11
Colombia	4
Pakistán	2
Vietnam	2
Sri Lanka	2
Reino Unido	2
Bolivia	2
Ecuador	1
Irán	1
Bangladesh	1
India	1
Suecia	1
Noruega	1
República Democrática del Congo	1
Número total de secuencias	**103**

Tabla 2.7. Resumen del número de secuencias utilizadas por país para la construcción del árbol filogenético.

2.5.1 Alineación múltiple de secuencias

Para la construcción del árbol filogenético primero se realizó una alineación múltiple de secuencias (multiple sequence alignment, MSA). Para esto se utilizó el software VCF-kit [139, 140], el cual presenta un conjunto de herramientas para la manipulación de archivos VCFs. La opción "phylo fasta" de VCF-kit permite generar un archivo fasta de variantes concatenadas entre sí, que es equivalente a una alineación múltiple de secuencia, desde un VCF. En el FASTA generado, cada línea representa una muestra. En la figura 2.7 se detalla la estructura de un archivo ".fasta". Este tipo de archivo representa efectivamente una alineación de secuencia múltiple que solo incorpora los sitios variables de las muestras del VCF. El comando que se ejecutó para esto fue:

vk phylo fasta todas_listas.recode.vcf > todas_listas.fasta

Donde:

- phylo fasta es el parámetro de VCF-kit que genera un archivo fasta alineado. Los datos faltantes son representados con una N.

- todas_listas.recode.vcf es el input, en este caso, el último archivo VCF generado que se encuentra filtrado de variantes espurias.
- todas_listas.fasta es el output, nombre que tendrá el archivo FASTA alineado.
  Cada línea del archivo fasta alineado generado representa una muestra y presenta 8839 variantes.

```
>LD4PC.variant
CCCGGTGCACNCTCCGCGACCGTTGAGCGNNTACAAGTCAGACCACCACGAAGCGTCCATTCNACCGTCTTTTCG
>LP6005441-DNA_A01.variant
CCTGGTGCACACTCCGCGACCGTTTAGCGCCTACAAGTCAGACAACCTCGAAGCGTCCATTCCACCGTCTTTTCGG
>LP6005441-DNA_A08.variant
CCCGGTGCACTCTCAGCGGCCGTTTAACCCCTAGAAGTCAGGTAGCCTCGACGCATGCAGTCCACCGACCGTTCG
```

Figura 2.7. Ejemplo de un fragmento de archivo ".fasta" alineado. El formato FASTA es un ".txt" utilizado para representar secuencias de ácidos nucleicos. Los pares de bases se representan usando códigos de una única letra. El formato también permite incluir nombres de secuencias y comentarios precedidos del símbolo ">". Cada línea es una secuencia alineada con las múltiples muestras que contiene el archivo.

2.5.2 Construcción de árbol filogenético de máxima verosimilitud

Se utilizó el método de máxima verosimilitud siguiendo a [31, 141], este método utiliza técnicas estadísticas estándar para inferir distribuciones de probabilidad para asignar probabilidades a árboles filogenéticos posibles particulares. Para esto, utilizamos el programa RAxML v.8.2.12 (Randomized Axelerated Maximum Likelihood) [142, 143] que realiza inferencias secuenciales y paralelas basada en la máxima verosimilitud de grandes árboles filogenéticos.

El procedimiento seguido para la construcción del árbol filogenético fue hecho en tres pasos, primero se generó un "mejor árbol" o "best tree", el siguiente paso fue realizar un análisis de "bootstrap". Este último análisis consiste en un re-muestreo y re-construcción del árbol "N" veces, donde se prueba cuántas veces se recuperan los mismos nodos. Se hicieron cien inferencias rápidas de bootstrap para dar apoyo estadístico a los clados del árbol. Por último, se generó un árbol "consensus" entre el mejor árbol o "best tree" y los resultados de Bootstrap. Los parámetros utilizados se eligieron siguiendo lo sugerido en [31].

Generación del mejor árbol:

raxmlHPC -c 1 -m GTRGAMMA --asc-corr=lewis -s todas_listas.fasta -p 12345 -n ASC.test

El archivo resultante "ASC.test", puede visualizarse en un programa visualizador de árboles, como FigTree [144].

Generación de análisis de bootstrap:

raxmlHPC -c 1 -m GTRGAMMA --asc-corr=lewis -s todas_listas.fasta -n bootstrap.TEST -p 12345 -f a -x 12345 -# 100

El resultado "bootstrap.Test" también puede visualizarse en con el programa FigTree, y además, se pueden seleccionar funciones para visualizar los resultados del bootstrap para cada nodo del árbol filogenético.

Generación de un árbol consensus:
raxmlHPC -f b -t ASC.test -z bootstrap.TEST -m GTRGAMMA --asc-corr=lewis -n consensus

El árbol "consensus" es el árbol filogenético resultante que se presenta en el anexo VI. El árbol consensus fue enraizado manualmente desde FigTree, utilizando la opción "Midpoint root" dentro de las herramientas "Tree" que dispone el programa. De esta manera, la muestra africana del haplogrupo B2b1 queda como raíz filogenética. En el anexo VII se presenta el detalle los resultados del bootstrap para el árbol consensus obtenido.

2.5.3 Datación de los nodos filogenéticos

Los nodos del árbol filogenético fueron datados usando el estadístico ρ (rho) [145], el cual se basa en la topología de un árbol enraizado y mide el número promedio de mutaciones de una raíz de un clado, para cada una de las secuencias muestreadas de dicho clado. Cuando se divide por la tasa de mutación para toda la secuencia por unidad de tiempo, proporciona una estimación de la edad de un clado determinado en unidad de tiempo.

Se utilizó la tasa de mutación del cromosoma Y reportada en [27], la cual fue calculada utilizando la secuencia del genoma del "hombre Ust'-Ishim" (nombre dado al fósil humano hallado cerca del asentamiento Ust'-Ishim, en Rusia), datado en 45,000 años, que es hoy el genoma humano más antiguo secuenciado hasta la fecha. La tasa mutacional reportada es de $0.76x10^{-9}$ mutaciones por sitio por año, con un intervalo de confianza del 95% de $0.67x10^{-9}$ a $0.86x10^{-9}$.

La variación en la cobertura de secuenciación afecta el cálculo del número promedio de genotipos variantes entre secuencias, generando desequilibrios en los cálculos entre muestras y sesgos en la datación. Evitamos estos errores utilizando únicamente las muestras con alta cobertura de secuencia, que además fueron filtradas bajo los mismos criterios y suponemos un equilibrio entre el número de sitios variantes e invariantes. Por lo tanto, hemos calculado únicamente la datación de nodos entre muestras de alta cobertura de secuenciación, trabajando así, con el archivo VCF "alta_cobertura_limpio.recode.vcf" obtenido en la sección 2.4.7.3.

Para el cálculo del estadístico ρ (rho), seguimos el mismo método empleado por Pinotti y col. 2019 [31], cuyo autor contribuyó con el script utilizado. De esta manera, se contabilizaron las diferencias de genotipo por posición entre dos muestras que fueron convertidas en años utilizando la tasa mutacional de cromosoma Y conveniente.

Así por ejemplo, el resultado del cálculo ρ (rho) entre las muestras EKEFB y LD4PC es de $2.53x10^{-6}$ mutaciones por sitio y debido a que este valor contabiliza del número de variantes por sitio para ambas ramas, el número se duplica y debe dividirse por dos.

$$\frac{2.53\text{x}10^{-6}\ \text{mutaciones por sitio}}{2} = 1.26\text{x}10^{-6}\ \text{mutaciones por sitio}$$

Utilizando la tasa mutacional $0.76\text{x}10^{-9}$ obtenemos:

$$\frac{1.26\text{x}10^{-6}\ \text{mutaciones por sitio}}{0.76\text{x}10^{-9}\ \text{mutaciones por sitio por año}} \cong 1664\ \text{años}$$

Con el mismo procedimiento se calculan los años de divergencia con un intervalo de confianza del 95% utilizando las tasas mutacionales $0.67\text{x}10^{-9}$ y $0.86\text{x}10^{-9}$, obteniendo un límite inferior y superior de 1471 y 1888 años, respectivamente. Estos resultados son representados en el capítulo de resultados con una notación de 1,66 kya (1.47-1.89) entre las muestras LD4PC y EKEFB.

Este procedimiento fue aplicado para todos los pares de muestras que pueden formarse desde el archivo "alta_cobertura_limpio.recode.vcf". La tabla resultante construida presenta un total de 1146 filas, por lo que no es posible adjuntarla en anexos, pero a continuación se deja un link para su visualización y se le asigna el nombre:

tabla adjunta II - sección 2.5.3 - datación por pares de muestras.
https://docs.google.com/spreadsheets/d/1YBhbjRXogCbPnizRnLiEFAMzsgV6HYce8UsmEYuoN_g/edit?usp=sharing

Los resultados de datación entre pares de muestra se promedian cuando el clado está conformado por más de una muestra. En el siguiente capítulo de resultados se detalla para cada nodo datado, los valores en años entre las muestras que comparten clados.

2.5.3.1 Datación de los nodos filogenéticos

En base a la "tabla adjunta II - sección 2.5.3 - datación por pares de muestras" se construyó una tabla de cálculos de datación por nodo para el árbol filogenético construido. Cada pestaña de esta tabla presenta el resultado del cálculo de datación para el nodo representado por un SNP. Este cálculo se realiza promediando solamente las muestras que comparten el nodo filogenético que se quiere datar. Debido a la extensión de esta tabla, se presenta un link para su acceso y se asigna su nombre como:

tabla adjunta III - sección 2.5.3.1 - datación de los nodos filogenéticos.
https://docs.google.com/spreadsheets/d/18PR5sG7KXTnVv7b5_HpmG42VE3YW5RK25cwPIZay4tM/edit?usp=sharing

2.7 Búsqueda de SNPs de importancia filogenética

Teniendo en cuenta los clados del árbol filogenético consensus generado en la sección 2.5.2, utilizamos la herramienta "SelectVariants" de GATK, junto con el parámetro "isec" de BCFtools, para realizar la búsqueda de variantes filogenéticas de interés. A continuación, se ejemplifica el procedimiento para el clado de la sección 3.1.2.2.2.2.2.1.11 (Clado XI Haplogrupo Q1b1a1a1h: Q-Z5906), conformado por once muestras, el cual incluye dos muestras secuenciadas en este trabajo, LD4PC y EKEFB.

Por lo tanto, con el fin de encontrar todos los sitios variantes presentes en el clado soportado por Q-Z5906, se crea un archivo VCF que incluye solamente las once muestras del clado a analizar en este ejemplo.

Creación de un VCF que incluye solamente las once muestras del clado a analizar:

java -jar GenomeAnalysisTK.jar -T SelectVariants -R human_g1k_v37_decoy.fasta -V final_filtDP.recode.vcf -sn EKEFB -sn GRC14443115_S20_L00 -sn GS000020273-ASM -sn GS000020274-ASM -sn GS000016951-ASM -sn HG02291 -sn HG01923 -sn LD4PC -sn GS000016942-ASM -sn HG02146 -sn HG02304 -env -trimAlternates --selectTypeToExclude INDEL -o nodo138_con_clado_EKEFB_LD4PC.vcf

Donde:

- -T SelectVariants parámetro que permite seleccionar un subconjunto y extraer una o más muestras de un VCF en función del nombre de la muestra.
- -R human_g1k_v37_decoy.fasta genoma humano de referencia.
- -V es el input, en este caso "todas_listas.recode.vcf"
- -sn "samplename" nombre de la muestra que se quiere incluir, debe ser el nombre exacto que presenta el VCF para la muestra, en este caso son tantos "-sn" como muestras de ese clado.
- -env -trimAlternates elimina las posiciones que no tienen ninguna variante
- --selectTypeToExclude INDEL excluye cualquier INDEL que no haya sido filtrado.
- -o nombre del output, en este caso "nodo138_con_clado_EKEFB_LD4PC.vcf"

A continuación, con el mismo fin de determinar los sitios variantes presentes únicamente en el clado soportado por Q-Z5906, se crea otro archivo VCF que presenta todas las muestras del árbol filogenético menos las once muestras de este ejemplo:

java -jar GenomeAnalysisTK.jar -T SelectVariants -R human_g1k_v37_decoy.fasta -V final_filtDP.recode.vcf -xl_sn EKEFB -xl_sn GRC14443115_S20_L00 -xl_sn GS000020273-ASM -xl_sn GS000020274-ASM -xl_sn GS000016951-ASM -xl_sn HG02291 -xl_sn HG01923 -xl_sn LD4PC -xl_sn GS000016942-ASM -xl_sn HG02146 -xl_sn HG02304 -env -trimAlternates --selectTypeToExclude INDEL -o nodo138_sin_clado_EKEFB_LD4PC.vcf

Donde, los parámetros son los mismos del comando anterior con la excepción de:

- -xl_sn este parámetro excluye las muestras con los nombres mencionados, en este caso son diez "-xl_sn" con los nombres exactos de las muestras que figuran en el VCF, que se quieren excluir.

La intersección entre estos dos archivos VCFs generados (**nodo138_con_clado_EKEFB_LD4PC.vcf y nodo138_sin_clado_EKEFB_LD4PC.vcf**) contiene todas las variantes del clado XI (Q-Z5906). Por lo que, una vez generados ambos VCFs de deben convertir en ".gz" para con BCFtools generar un archivo de intersección entre ambos:

bcftools isec -p dir_nodo138_con_clado_EKEFB_LD4PC -c all nodo138_con_clado_EKEFB_LD4PC.vcf.gz nodo138_sin_clado_EKEFB_LD4PC.vcf.gz

Donde:

- -p es el directorio donde se creará el resultado
- -c all es el parámetro de "isec" que controla cómo serán tratados los registros. En este caso "all" realiza intersecciones de todas las líneas entre ambos VCFs y crea un VCF con todos los registros de líneas exclusivas, que es el de interés.

Para este ejemplo, el archivo VCF de interés, presenta 196 posiciones, las cuales representan tanto posiciones variantes únicas para las muestras, como posiciones variantes compartidas entre las muestras de este clado analizado.

Este procedimiento se aplicó para cada uno de clados del árbol filogenético construido. Se construyó una tabla que une todos los clados analizados. Cada pestaña de esta tabla presenta los datos del VCF resultante creado para conocer los SNPs de cada clado del árbol filogenético. Por tanto, cada pestaña contiene las posiciones variantes únicas o compartidas, únicamente entre las muestras del clado en estudio. El nombre de cada pestaña está dado según la importancia del SNP encontrado para el clado (en este ejemplo, Q-Z5906), seguido de alguna de las muestras que contiene el clado). En el caso de los clados que presenten sub-clados, dentro de la pestaña se nombran los SNPs de relevancia interna y en algunos casos se crean pestañas con sub-clados dentro de un nodo de interés. Debido a la extensión de esta tabla, se presenta un link para su acceso y se le asigna el nombre:

tabla adjunta IV - sección 2.7 - búsqueda de SNPs de importancia filogenética.
https://docs.google.com/spreadsheets/d/10-GPFNN6eVbF4aPWAprCoadw-So8DUAwj10ueqoyi_g/edit?usp=sharing

En esta última tabla, los colores han sido asignados únicamente para facilitar el conteo de los SNPs de los sub-linajes.

De manera manual se realizó una búsqueda en las bases de datos de SNPs de ISOGG [56, 57], así como en los últimos trabajos publicados en el tema [22, 23, 31, 33, 38], para conocer si los SNPs encontrados estaban descriptos y/o validados. Algunas de las variantes de interés filogenético presentes en nuestras muestras y en algunos casos compartidas con otras muestras, que no se encontraron validadas y/o descriptas en ISOGG, se eligieron para la validación y se describen en el siguiente capítulo. Como criterio se consideraron como más importantes las variantes compartidas

entre varias muestras de un sub-linaje que no se encontraron validadas, luego también se eligieron variantes únicas de las muestras que no se encontraban validadas.

En base a la información del archivo "final_filtDP.recode.vcf" más los resultados de búsqueda de SNPs de importancia filogenética, se construyó una tabla que contiene únicamente los SNPs de relevancia para la definición de todas las muestras en el árbol filogenético construido. Esta tabla contiene el ID de los SNPs, la posición en el cromosoma Y, los alelos de referencia y alternativos, los genotipos para cada una de las muestras, asi como los nombres de los haplogrupos según ISOGG para las variantes que ya se encuentran validadas, y las que aún no se han incorporado a ISOGG. . Esta tabla también se presenta con un link debido a su extensión y se nombró como:

tabla adjunta V - sección 2.7 - SNPs relevantes por nodo.
https://docs.google.com/spreadsheets/d/1VA1QXj0TQsTnfnlkqPiXdpJ9_oYd3gUDth4PTtbLTYk/edit?usp=sharing

2.8 Validación de SNPs de importancia filogenética

Los errores propios de la metodología NGS llevan a que los SNPs nuevos que se detectan en una investigación deban ser validados mediante una metodología independiente [148]. El estándar de oro para realizar esto es la secuenciación Sanger de los segmentos donde se encontró la variante [149], aunque se ha discutido su utilidad [150] y el trabajo más reciente disponible sobre este debate afirma que si la calidad es muy alta no es necesario [151], pero es fundamental notar que los rangos de calidad analizados por estos investigadores comenzaban en 173x, muy superior al 30x de la presente tesis.

- Muestras: se utilizaron las trece muestras secuenciadas en este trabajo, detalladas en la tabla 2.1.
- Cuantificación de ADN: las concentraciones de ADN en solución se obtuvieron a partir de la determinación de la absorbancia a 260 nm mediante espectrofotometría. Se utilizó el espectrofotómetro UV visible de espectro completo utilizado para cuantificar y evaluar la pureza de ADN, ARN, proteínas NanoDrop 2000 (Thermo Scientific™). Como blanco de muestra se utilizó agua Mili-Q. Posteriormente se realizaron las diluciones necesarias para obtener una concentración de 10 ng/ul de cada muestra.

- Reacción en cadena de la polimerasa (PCR): Los productos de amplificación fueron obtenidos mediante amplificación por PCR utilizando el termociclador Eppendorf Mastercycler Nexus (Eppendorf, Alemania) y el termociclador Biometra T3000 (Biometra, Alemania).
  Los cebadores fueron diseñados con las herramientas Primer3 [152] y Oligoanalyzer (IDT) [153].

Los ensayos de PCR fueron optimizados variando la temperatura de annealing de los cebadores o cambiando la polimerasa, con un volumen final de reacción de 25 µl.
Las concentraciones de los reactivos utilizados en la Mezcla de reacción se detallan a continuación:

Reactivo	1X (ul)	3.5X
Buffer GO 5X	5	
dNTPs 2mM	0.4	0.03 mM
Primer Fw 2.5 µM	0.25	0.25 µM
Primer Rv 2.5 µM	0.25	0.25 µM
Taq	0.125	
DNA	1	10ng/µl
Agua	17.975	
Volumen final	25	

Reactivo	1X (ul)	3.5X
Buffer 10X	2.5	8.75
MgCl2 (50mM)	1	2 mM
dNTPs 2mM	0.4	0.03 mM
Primer Fw 2.5 uM	0.25	0.25 µM
Primer Rv 2.5 uM	0.25	0.25 µM
Taq Platinum	0.125	
DNA	1	10ng/µl
Agua	19.5	
Volumen final	25	

Tablas 2.8. Condiciones de PCR utilizadas

Esquema del programa la PCR utilizado:

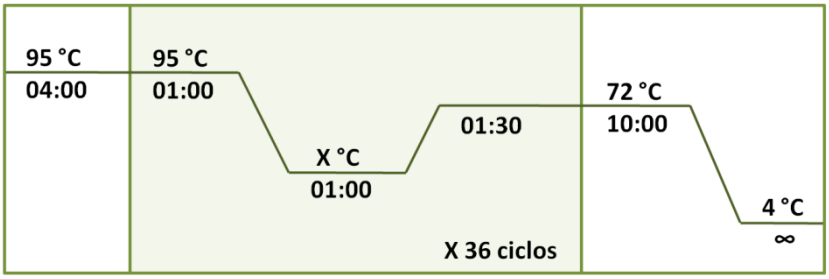


Figura 2.8: Esquema del programa de PCR utilizado. Para ambas Polimerasas se utilizó el mismo programa donde solamente se variaron las temperaturas de annealing (X °C).

La siguiente tabla resume la secuencia de los cebadores utilizados y las condiciones de cada par:

Clado	GRCh37	Cebador FW	Secuencia FW	Secuencia RV	Cebador RV	T° annealing	Polimerasa
RUTBE	6931449	Rutbe_1_FW	GGTGACCCCTTACAGATGGA	GACAGCCATCAGCTCACAAG	Rutbe_1_RV	60	Taq GO
	9023670	Rutbe_2_FW	AGCTGGCTGTTCCAGTCTGT	TCTTTGTGGGACTGATGTGG	Rutbe_2_RV	60	Taq GO
	14169926	Rutbe_3_FW	ATCCCCTTGACCTTCTTCGT	AACAAATTCCCAGGTGATGC	Rutbe_3_RV	60	Taq GO
	15479160	Rutbe_4_FW	TCTGTCTCAGCCCCTCAAGT	CAAGTCCCCTCCCCAGTAAT	Rutbe_4_RV	60	Taq GO
	7540846	Rutbe_5_FW	GTCTCAAGGTGGTGGAGCAT	ACAGGTTTTCACCCTGTTGG	Rutbe_5_RV	NA	NA
	8385051	Rutbe_6_FW	GTGTGGGGTGCTGTAGTGTG	CAGACTCCCGCAGATAGGAG	Rutbe_6_RV	58	Taq GO
	8702379	Rutbe_7_FW	GAAAGGGACCTGGGAATCAT	GGGGTGAGGAACTGACAGAA	Rutbe_7_RV	60	Taq GO
	8869476	Rutbe_8_FW	TTCACTGACACCTGCCTCTG	TCCCATAAGCCCTCTGATTG	Rutbe_8_RV	60	Taq GO
	14877756	Rutbe_9_FW	TGCCTCCTGTCAGATCAGTG	CTCAGGCTGGAAACAACACA	Rutbe_9_RV	64	Taq GO
	19060346	Rutbe_10_FW	CCTGGAATCTGGATCAGGAA	AGTTCAGATGCCTGGGAATG	Rutbe_10_RV	60	Taq GO
Z8ZMY/S8BAL	7948308	Z8/S8_ FW	CTTCATTCCATTTGGGCTGT	TATGACTCCCCTGCCTTGTC	Z8/S8_ RV	60	Taq GO
	2875332	S8BAL_1_FW	TCGAGAAAGTTGTGCTGGTG	ACCTAGCTTGTGGTGGGTTG	S8BAL_1_RV	60	Taq GO
	7828787	S8BAL_2_FW	CCAAACCACAAGAAGCCCTA	CAAAGAAAGAGCCCCTGTTG	S8BAL_2_RV	60	Taq GO
	21084351	S8BAL_3_FW	CTTTGAAACCAAGCCTGAGC	AATGGAGTTGGGCAGTTGTC	S8BAL_3_RV	60	TAQ PLATINUM
	7868875	Z8ZMY_1_FW	AAAACACCAACATGGGGAAA	TTCTCATCCTGCCATTTTCA	Z8ZMY_1_RV	60	Taq GO
	7905270	Z8ZMY_2_FW	TCTGCCTGCTTCACAACAAC	TGGCAAACGAATACCCTTTC	Z8ZMY_2_RV	60	Taq GO
N8A2QN	2749149	N8A2QN_1_FW	CCTTCCTTTGTGCCATCCTA	GCCATCTCATCAACCGTCTT	N8A2QN_1_RV	60	Taq GO
	2804456	N8A2QN_2_FW	TCGGCTACGCTTTAGGTGAC	TAGCACTGTCCAACGCTCAC	N8A2QN_2_RV	60	Taq GO
	7134535	N8A2QN_3_FW	GGGCTCAAGCAATACTTCCA	CAATCTTGGCTCACTGCAAA	N8A2QN_3_RV	62	Taq GO/2 min de extensión
	7267390	N8A2QN_4_FW	CACTCCAGGACCAACACCTT	TGGCATTGGGACCTAGAAAC	N8A2QN_4_RV	60	Taq GO
	8131788	N8A2QN_5_FW	CAGCTTGGCAAATATGGTGA	TGATGCAAAATGCCTGATGT	N8A2QN_5_RV	60	Taq GO
	14886685	N8A2QN_6_FW	ATGCTGACATGAACGATGGA	TCTGAACAAAGCCGTTTGC	N8A2QN_6_RV	60	Taq GO
	15954669	N8A2QN_7_FW	GTTTCCCCACATAAGCTCCA	CTTGGCACTTCTCCTTCCTG	N8A2QN_7_RV	62	Taq GO/2 min de extensión
	17351850	N8A2QN_8_FW	CCACCCACACTTGTCGTATG	CTCAATCCAGTTCCCCAAGA	N8A2QN_8_RV	60	Taq GO
	18830601	N8A2QN_9_FW	CACTGGGAAACAAATCCACA	CTTGGCAAACATGGAACAAA	N8A2QN_9_RV	64	TAQ PLATINUM
	19087030	N8A2QN_10_FW	TGGAAATGGGAAAAGTCTGC	CTCATTGGTGCCATCTTCCT	N8A2QN_10_RV	60	Taq GO
N87FK8	14198967	N87FK8_1_fw	TCAAAGACGGGACCAGAATC	GCCGTAACGAGTTGAGAACC	N87FK8_1_rv	60	TAQ PLATINUM
	23567702	N87FK8_2_fw	TAACCCTGCTGGGTATCTGG	TCTCCCTTGTGAGTGCTGTG	N87FK8_2_rv	58	Taq GO
	23592805	N87FK8_3_fw	ACCCTATCCAGAGAAGGTGTAT	GGTAATGGTCCAGGATGGATTT	N87FK8_3_rv	56	Taq GO
	23962077	N87FK8_4_fw	GACCCTGTCTCTTAACCCAATAC	GAAGACTGCCAGCCTCATAAA	N87FK8_4_rv	62	Taq GO/2 min de extensión
	23984584	N87FK8_5_fw	AGCCACAACCTGGAACATAG	ACAGAGTGTGAGCAGCAATAA	N87FK8_5_rv	62	Taq GO/2 min de extensión
	9143566	N87FK8_6_fw	CTATACCAGCCGTGTCATGTT	CCATGAGGACTCTTTGCTTCT	N87FK8_6_rv	60	Taq GO
	17263815	N87FK8_7_fw	GGAGGCTGAGGAGAGAGAATTA	TGCTGGGATGAGAGGTATGA	N87FK8_7_rv	66	TAQ PLATINUM
	23247806	N87FK8_8_fw	TCTGGGATCTTTTCCACAGG	CACTTTTGACAGCCTCCACA	N87FK8_8_rv	56	Taq GO
	23765443	N87FK8_9_fw	ACTTGAAGGAGGAGGCATTTC	CTCATTGTGGAGTGGGAGTTT	N87FK8_9_rv	60	Taq GO
	23575633	N87FK8_10_fw	CCTGCACACCTGCTTTAACA	GCCCAATGTGGTTTTGATTT	N87FK8_10_rv	60	Taq GO
T4WQV	14587968	T4WQV_1_fw	AGTCAGGGCAGAGCAGGTAG	TGCAGCTCACCAAGAAAATG	T4WQV_1_rv	66	Taq GO
	6678425	T4WQV_2_fw	GACTGTCCCCTCTGATCTGC	AGGGTCCTCCACTCTGGTTT	T4WQV_2_rv	62	Taq GO
	7353313	T4WQV_3_fw	ACTCTGCCATCTCCAACACC	TCATATCCACTGGGAGCACA	T4WQV_3_rv	64	Taq GO
	7566319	T4WQV_4_fw	GAACCCACAAGCTGCTAACTA	GCTCACATCCTCCAAAGACTAA	T4WQV_4_rv	64	Taq GO
	7673168	T4WQV_5_fw	GTTCCTGGCAGAAAAGTTGC	GGCATCCTGGTGATTTGACT	T4WQV_5_rv	54	TAQ PLATINUM/1,5 mM MgCl2
	7848322	T4WQV_6_fw	CCCTTAGGACAGGACACATTTAG	TGACAGAACCCAGCAGAAAG	T4WQV_6_rv	60	Taq GO
	7887814	T4WQV_7_fw	CCATTGCCAAAGAAGGTGTT	TACAGCCGTGGGTAAAGTCC	T4WQV_7_rv	62	Taq GO
	8251637	T4WQV_8_fw	TTTGTGCAGGGTCCATTACA	ACAACCACAGAGGGAAGTGG	T4WQV_8_rv	62	Taq GO
	8446496	T4WQV_9_fw	TTGCTGGATGGGACTACCTC	ATCTCTGTGGTTGGGACCTG	T4WQV_9_rv	62	Taq GO
M39DJ/UCNEN	6631920	M39DJ_1_fw	CATCCTCCTCTTTCATCATCCC	CCTGCTTGCATAGTGCCTATAA	M39DJ_1_rv	58	Taq GO
	7571644	M39DJ_2_fw	CAGGATGAATCCAGGGTCTAAC	GCTCCAGATGGTGGTGAAATA	M39DJ_2_rv	62	TAQ PLATINUM
	7765120	M39DJ_3_fw	TGCACAGATGCTCTCAAATACA	GACTTGGGTAATCTCCTGCTATG	M39DJ_3_rv	60	TAQ PLATINUM
	8359844	M39DJ_4_fw	CACCAGCAACCAAGTGTTATG	CCAAGGCAGGAGAGAAGAAA	M39DJ_4_rv	60	TAQ PLATINUM
	8560447	M39DJ_5_fw	CTTCTGTCTGGTGTGGATATGG	TGTGGGTTCAGGTGGTAGTA	M39DJ_5_rv	60	Taq GO
	6965772	UCNEN_1_fw	TCAGAGGGCACACAGACAAG	GGAGTCACAGGGTTGCAGAT	UCNEN_1_rv	66	Taq GO
	7419588	UCNEN_2_fw	GGTGTATGTGTGCATGGATTTC	CTCTGGGCAGAGAGTGTATTT	UCNEN_2_rv	60	Taq GO
	8133490	UCNEN_3_fw	GCCTCACCATAGCCATCATAAA	GGGTTGTTTCCACCACAGTAT	UCNEN_3_rv	60	Taq GO
	8440075	UCNEN_4_fw	CTGTGGGACTCTGTGTTCTTT	GGATTCTGATGAGGGTTGTTCT	UCNEN_4_rv	60	Taq GO
	14044033	UCNEN_5_fw	TGCAATGAAATGGTTCTCCA	TGAGGAAGTCAGCAGGGAGT	UCNEN_5_rv	62	Taq GO
6QHWE	2747337	**6QHWE_1_FW**	CACCATCCCAACTTCAGCTT	CTCCTCCAGCACAGACATCA	6QHWE_1_RV	64	Taq GO
	6904459	**6QHWE_2_FW**	ATGGGAGTGTGTGCATTCAA	TCACCATGAGCTGCCTACTG	6QHWE_2_RV	60	Taq GO
	7644074	**6QHWE_3_FW**	GGAGCTGCTCTGCATTCTCT	CCTGCAATGAAAAGCAAACA	6QHWE_3_RV	64	Taq GO
	7893507	**6QHWE_4_FW**	GGGAGCAAAGCACAGGTAAA	ACCTGGAGGAAGTTCCCAGT	6QHWE_4_RV	60	Taq GO
	8051637	**6QHWE_5_FW**	AGCAATGAAACCCGAGGATAG	GACCAGCCAATACTTAGGGAAA	6QHWE_5_RV	64	Taq GO
	8539196	**6QHWE_6_FW**	ATTGGGCTGTTGCAGGATAG	CACTGGAGGTGGAGGTGTTT	6QHWE_6_RV	60	Taq GO
	8833006	**6QHWE_7_FW**	ACTTTGGCACAGGTGTTTGG	CCCCTAAGAGGAAGCAAAGG	6QHWE_7_RV	64	Taq GO

Clado	GRCh37	Cebador FW	Secuencia FW	Secuencia RV	Cebador RV	T° annealing	Polimerasa
SQVCW	6656300	SQVCW_1_FW	GCAGCCACATCTTTCTGTCA	CAAAAGCAGCCCTCATTCTC	SQVCW_1_RV	60	Taq GO
	6921840	SQVCW_2_FW	TTTCCCTCCAGAGCCTACTAT	CAGCTAGTCCTGTCTCTCAAATC	SQVCW_2_RV	60	Taq GO
	7779334	SQVCW_3_FW	CCTTGCTTCCCTCAGGTTAAT	GGTTCAGAGTGTGGGAGAATAC	SQVCW_3_RV	64	Taq GO
	8023779	SQVCW_4_FW	TTGCCCAGAGGTCCAATATC	CACAAAAAGGAAACCCCAGA	SQVCW_4_RV	60	Taq GO
	8036434	SQVCW_5_FW	GGCTGAAGCAGGAGAATGAC	TGGACTGACCCTCTGGTTTC	SQVCW_5_RV	64	Taq GO
	8594266	SQVCW_6_FW	AGGGATGCAGTTGGAAACAC	GTGGCTTGGCAGAGAAAAAG	SQVCW_6_RV	66	Taq GO
	15261236	SQVCW_7_FW	GAGATGGAGTCTTGGTCTGTTG	GTGAACACTGGTAGAAAGGAAGTA	SQVCW_7_RV	64	Taq GO
TYEQC	14246232	TYEQC_1_FW	CCCTCCAGATAGCACACATTTC	TGTTCATATCCTTCGCCCAATAC	TYEQC_1_RV	60	Taq GO
	23583455	TYEQC_2_FW	AAACCAAGATGTGGGCAAAG	AAACTACCTTCCCGCTCCAT	TYEQC_2_RV	60	Taq GO
	8601728	TYEQC_3_FW	CCAGCAACAGCTCAATGAAA	GCACAAGGGCTAGTCTCAGG	TYEQC_3_RV	64	Taq GO
	8030403	TYEQC_4_FW	TCTGCTGACAGTGTGCTTCC	TCCCCTTTAGGCAATCATCA	TYEQC_4_RV	64	Taq GO
	7748881	TYEQC_5_FW	CTGTTAAGCCAGGAGAGTCAAA	GCCATGACCACAGATGAAATTG	TYEQC_5_RV	60	Taq GO
	7588274	TYEQC_6_FW	CTGCCTTGGTGGTTAGGATAAG	GTAGCAGCAGGCGTTGTATAG	TYEQC_6_RV	64	Taq GO
	7383562	TYEQC_7_FW	TCCAGGACATAACACCAGCA	AAAAAGGGGAGAAACCCTCA	TYEQC_7_RV	64	Taq GO
LD4PC/EKEFB	6793301	LD_EK_1_FW	CTGTCCCAATTTCAGCCACT	AGATCTCAGCCAGGCACAGT	LD_EK_1_RV	64	Taq GO
	7218975	LD_EK_2_FW	GCAAGACGAGCAACTGAAATG	CAATGGAGTGGGTGGAATGA	LD_EK_2_RV	64	Taq GO
	14754418	LD_EK_3_FW	CATCTCTTGACCTGGTGATCTG	CTGAGGGAATGGCATGACTATAA	LD_EK_3_RV	64	TAQ PLATINUM
	16835476	LD_EK_4_FW	CATCAGTGAGTGAGCCAGTAAG	ACATCTAAACAAGGAGGAGCTAAA	LD_EK_4_RV	60	Taq GO
	23748402	LD_EK_5_FW	AGAAACCGCCACTGCTATAC	CTCACATCTGCCTTCTCATAGG	LD_EK_5_RV	60	Taq GO
	23785274	LD_EK_6_FW	TGATTGGCAGTTGGTTGAAA	TTCCTTTTGCATCTGCACTG	LD_EK_6_RV	60	Taq GO
	8806607	LD_EK_7_FW	CAGAGCTGCATGGTAGTAGTG	GCTAGACAGAGAGTGCTGATTG	LD_EK_7_RV	64	Taq GO
	14196672	LD_EK_8_FW	GCACTCTATGCTGGGAAACA	GAGGAGCTTGCACTAACTAAGG	LD_EK_8_RV	64	Taq GO
	23987283	LD_EK_9_FW	GGCTGTCTGTGGTGATAGTAAG	TAGAGTGTGAGCAGCAATAAGG	LD_EK_9_RV	64	Taq GO

Tabla 2.9. Secuencia de los cebadores utilizados y las condiciones de cada par.

- Electroforesis en gel de agarosa:

  Los fragmentos de ADN obtenidos se verificaron mediante electroforesis en gel de agarosa (Genbiotech). El porcentaje de agarosa usado fue de 1.5%. Los geles se prepararon a partir de la fundición de una solución de agarosa en buffer TBE 1X (Tris Base, ácido bórico, EDTA, pH 8). Las muestras de ADN, antes de ser sembradas, en el gel de agarosa, fueron diluidas en buffer de siembra 6X, el cual contiene GelRed. Para estimar el tamaño del fragmento de ADN sometido a la electroforesis, se utilizó el marcador de 100 pb (Genbiotech).

  Las condiciones de electroforesis empleadas dependieron del tamaño del fragmento de ácido nucleico a resolver. En este sentido, en todos los casos se utilizó una solución buffer TBE 1X como buffer de corrida, mientras que el voltaje de las distintas electroforesis fue constante de 90 V. El tiempo de separación también fue variable entre 45-60 minutos. La posterior visualización de los ácidos nucleicos resueltos en el gel de agarosa se realizó en un transiluminador Gel Doc XR (Bio-Rad), aprovechando las propiedades del GelRed, el cual emite fluorescencia al ser irradiado con luz UV, luego de intercalarse entre las bases del ácido nucleico.

- Purificación de productos de PCR por precipitación con PE (polietilenglicol):

1. A 20 µl de producto de PCR se le agregan 20 µl de una solución de 20% PEG-2,5 M NaCl y luego la mezcla se incuba durante 15 minutos a 37ºC.
2. Se centrifuga a 13.000 rpm durante 20 minutos a temperatura ambiente.
3. El sobrenadante se descarta con pipeta. El pellet es incoloro y queda adherido a la pared del tubo.
4. Se agregan 50 µl de etanol 70% suavemente por la pared del tubo y se deja reposar 1 minuto. Luego, con pipeta, se descarta el sobrenadante para eliminar la mayor cantidad posible de etanol. Se centrifuga 20 min a 13.000 rpm.
5. El pellet se seca a 37ºC durante 10-15 minutos; asegurándose que el pellet estuviera seco sin restos de etanol.

6. Se resuspende el pellet en 25 µl de agua bidestilada estéril con pipeta o vortex a temperatura ambiente o 37ºC y por último se mide la concentración del producto purificado.

- Secuenciación de ADN

  Los productos de PCR que luego de la purificación mostraron bandas nítidas, fueron enviados a la empresa Macrogen, Inc (Seúl, Corea de Sur), siguiendo las especificaciones de la misma.

  Los resultados obtenidos fueron analizados utilizando varias herramientas disponibles en la web, un visualizador de secuencias software Chromas [154]), el analizador de SNV e Indel software Indigo [155] y para realizar alineamientos el software BLAST [156].

3 RESULTADOS

En el presente capítulo se analiza clado por clado el árbol filogenético obtenido (como se describió en la sección 2.5.2), representado en la figura 3.1. El total de muestras que incluye este árbol es de 103, incluyendo la raíz filogenética perteneciente al haplogrupo B2b1 (LP6005441-DNA_A08). La descripción del mismo comienza con el haplogrupo más antiguo, siguiendo en orden descendente por cada clado. Además, se tuvo en cuenta para este análisis los resultados de las relaciones filogenéticas encontradas, los valores de datación de los nodos que pudieron ser determinados, los SNPs relevantes y el análisis del soporte estadístico evaluado con los bootstrap. Se realizó un análisis más detallado de los SNPs de clados que contienen muestras secuenciadas en el presente trabajo.

Se ha respetado el nombre utilizado para nombrar los haplogrupos de Q-M242 asignados con la letra Q seguido de un código de letras y números utilizado por la plataforma ISOGG [56, 57].

Los análisis realizados sobre los datos genéticos siguiendo diferentes criterios, que se explicará a lo largo de este capítulo, se encuentran resumidos en tablas Excel. Estas tablas son extensas, por lo que no pueden ser presentadas en este formato y se adjuntan los links de acceso desde Google Docs.

Los SNPs obtenidos como se describe en la sección 2.7 se pueden consultar en las siguientes tablas:

- Tabla adjunta IV - sección 2.7 - búsqueda de SNPs de importancia filogenética.
  https://docs.google.com/spreadsheets/d/10-GPFNN6eVbF4aPWAprCoadw-So8DUAwj10ueqoyi_g/edit?usp=sharing
- Tabla adjunta V - sección 2.7 - SNPs relevantes por nodo.
  https://docs.google.com/spreadsheets/d/1VA1QXj0TQsTnfnlkqPiXdpJ9_oYd3gUDth4PTtbLTYk/edit?usp=sharing

Las dataciones obtenidas como se describe en la sección 2.5.3.1 por nodos pueden ser consultadas en la siguiente tabla:

- Tabla adjunta III - sección 2.5.3.1 - datación de los nodos filogenéticos.
  https://docs.google.com/spreadsheets/d/18PR5sG7KXTnVv7b5_HpmG42VE3YW5RK25cwPIZay4tM/edit?usp=sharing

Los datos de las muestras utilizadas para construir el árbol filogenético se encuentran resumidos en la siguiente tabla:

- tabla adjunta I - información sobre las muestras.
  https://docs.google.com/spreadsheets/d/1i-cnkj863o32zPFUoKfbfehI61EqjhULd4K6P-miXR8/edit#gid=1670853527

Se recomienda para una mejor lectura y entendimiento de este capítulo tener presente el árbol filogenético presentado en la figura 3.1, el cual resume los SNPs relevantes y la datación de los nodos que pudieron ser calculados.

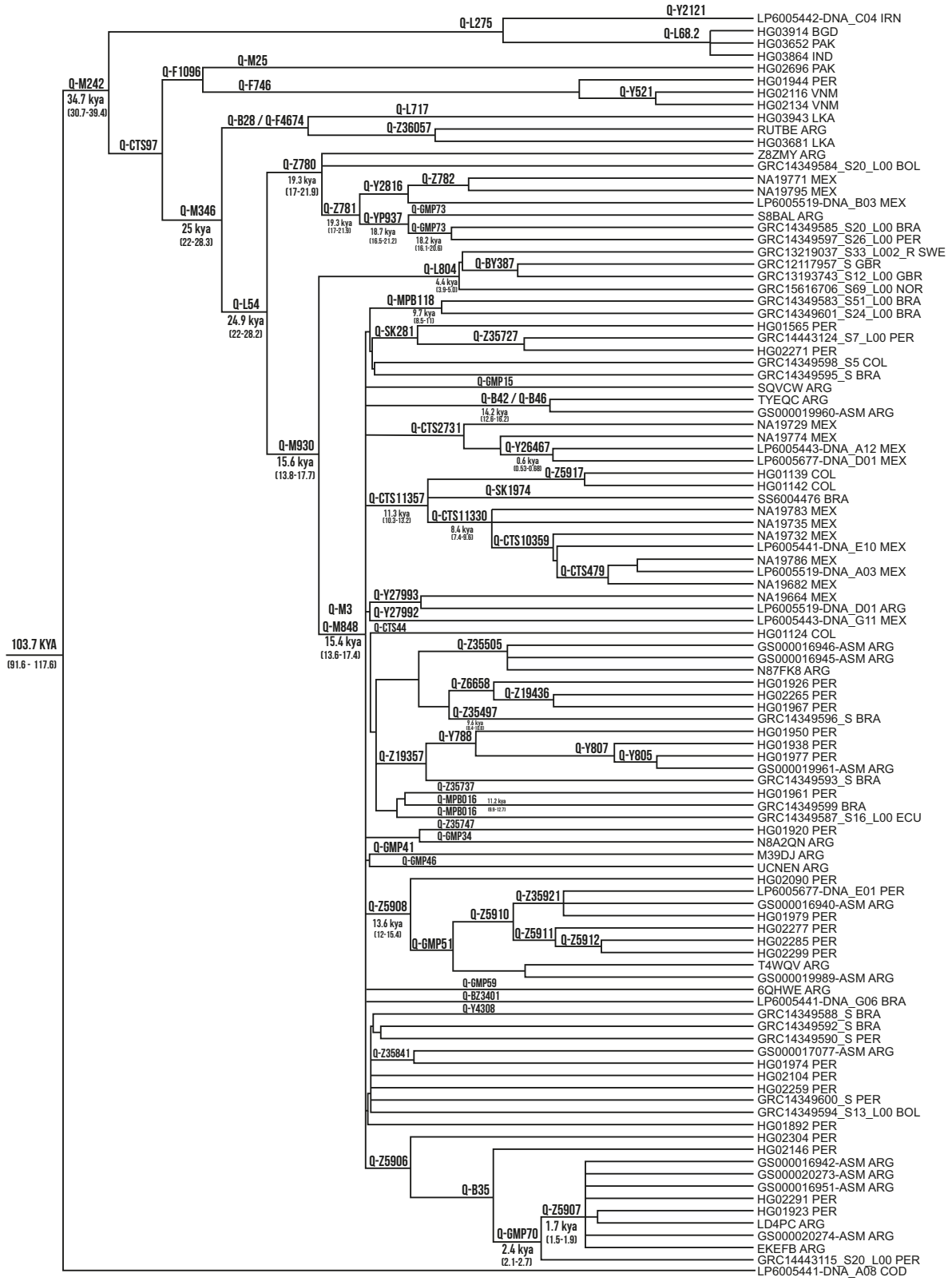


Figura 3.1. Representación esquemática del árbol filogenético del haplogrupo Q-M242. Se incluyen algunos SNPs relevantes y la datación calculada para algunos nodos. La longitud de las ramas no está en función de ninguna variable. En anexos VIII se presentan los resultados filogenéticos obtenidos con el programa RAxML sin modificar, el cual fue utilizado como base para la construcción de esta imagen.

3.1.1 Raíz del árbol filogenético

Para enraizar el árbol filogenético del haplogrupo Q se utilizó como outgroup la muestra LP6005441-DNA_A08 del pueblo Mbuti del Congo de África, perteneciente al haplogrupo B2b1. Se conoce que los haplogrupos africanos A y B, presentan la mayor diversidad dentro de la filogenia del cromosoma Y. La datación encontrada para el nodo que forma esta muestra con el resto del árbol es de 103.7 kya (91.6 - 117.6). Se realizó este cálculo de datación únicamente para contrastar respecto a la estimación temporal de Q-M242, ya que la inclusión de esta muestra en este trabajo tiene como único objetivo utilizar una muestra con gran divergencia con respecto a nuestros datos para un correcto enraizamiento y construcción del árbol filogenético del haplogrupo Q-M242.

3.1.2 Filogenia de Q-M242

El marcador Q-M242 es el que define nuestro conjunto de datos en estudio. Todas las secuencias seleccionadas presentan este marcador. El nodo con la mayor antigüedad en el árbol filogenético presentado se da para este marcador, con una datación promedio de 34.7 kya (30.7-39.4). En nuestro conjunto de datos Q-M242 se divide en Q-L275 y Q-CTS97.

3.1.2.1 Haplogrupo Q2: Q-L275

La primera bifurcación dentro de Q-M242 se da para el marcador Q-L275, presentada para cuatro muestras del sur de Asia: LP6005442-DNA_C04 de Irán, HG03914 de Bangladesh, HG03652 un individuo punyabí de Pakistán y HG03864 un indio Telugu del Reino Unido.
Los resultados muestran un alto soporte estadístico para esta rama con valores de 100 de Bootstrap. Sin embargo, la única muestra de alta cobertura dentro de este clado es LP6005442-DNA_C04 por lo que no se puede establecer una datación interna para este marcador. No se han encontrado nuevos marcadores para este sub-linaje, todos los SNPs analizados para este clado ya han sido descriptos en la bibliografía (ver Tabla adjunta IV - sección 2.7 - búsqueda de SNPs de importancia filogenética). Dado a que en general, la mayoría de los SNPs ya han sido descriptos por otros trabajos, solamente mencionaremos cuando se hayan encontrado nuevos dentro de los clados.

Q-L275 se bifurca en otras dos ramas, la primera contiene a LP6005442-DNA_C04 y presenta de forma privada a Q-Y2121 (marcador equivalente a Q-L245), y la segunda rama contiene a las otras tres HG03914, HG03652 y HG03864 que comparten Q-L68.2.

3.1.2.2 Haplogrupo Q1: Q-CTS97

Q-CTS97 se bifurca en Q-F1096 y Q-M346. La datación de Q-CTS97 no fue posible dado a la baja cobertura de secuenciación de las muestras que componen su sub-linaje Q-F1096.

3.1.2.2.1 Haplogrupo Q1a: Q-F1096

Q-F1096 es compartido por cuatro muestras: HG02696 punyabí de Pakistán al Sur de Asia, HG01944 de Lima, Perú y por HG02116 y HG02134 ambas de Vietnam.

Este es un clado que tiene alto soporte estadístico con 100 de Bootstrap, de igual manera sus clados internos también presentan alto soporte estadístico. Este nodo no pudo ser datado ya que todas las muestras que lo conforman tienen baja cobertura de secuenciación.
La muestra HG02696 presenta de forma privada a Q-M25. Por otro lado, HG01944, HG02116 y HG02134 comparten a Q-F746 y Q-M120. Este último linaje se divide en un sub-linaje soportado por Q-Y521 compartido por HG02116 y HG02134.

3.1.2.2.2 Haplogrupo Q1b: Q-M346

El marcador Q-M346 es uno de los que se encuentran mayormente representados en las muestras de este estudio. Todas las muestras secuenciadas en este estudio lo presentan. Se encontró un alto soporte estadístico para su nodo, con valores de 100 de Bootstrap. La datación encontrada para Q-M346 fue de 25 kya (22-28.3). Se bifurca en dos grandes ramas Q-F4674 y Q-L54.

3.1.2.2.2.1 Haplogrupo Q1b2a: Q-F4674

Q-F4674 es un sub-linaje de Q-M346 y presenta una ubicación precisa en la plataforma ISOGG, por lo cual se basó en este trabajo la clasificación de este clado utilizando este marcador. Este marcador se encuentra presente en las muestras: HG03943 de Sri Lanka, HG03681 de Pakistán y RUTBE de San Juan, Argentina. Esta última representa una de las muestras enviadas a secuenciar en este trabajo.

Todas las ramas de este clado presentan un soporte estadístico elevado de 100 de Bootstrap. Debido a que la muestra RUTBE de San Juan es la única de este clado que presenta alta cobertura de secuenciación, no fue posible establecer una datación para este nodo.

Del análisis de marcadores detallados en la tabla anexa 2.9 encontramos un total de 251 SNPs exclusivos para este clado. Q-F4674 se encontró compartido por las tres muestras, junto a otros 22 paralelos al SNP anterior, los cuales se encuentran descriptos en ISOGG. También paralelo a Q-F4674, se encontró un SNP compartido por las tres muestras, ausente en ISOGG, descripto en la bibliografía sin una clara ubicación filogenética, el cual se validó en este trabajo y se le asignó el nombre Q-GMP2, tabla anexa VIII. Dado a que todos los SNPs validado se encuentran descriptos en esta última tabla mencionada, remitirse a la misma cuando sean mencionados.

La primera bifurcación de este clado se da para la muestra HG03943 que presenta Q-L717 de manera exclusiva. La otra bifurcación se da para las muestras RUTBE y HG03681 que comparten el marcador Q-Z36057, junto a otros 63 SNPs paralelos al mismo. De los cuales, 44 se encuentran descriptos y 20 fueron encontrados nuevos para ambas muestras en este trabajo, de los cuales 3 fueron validados (Q-GMP1, Q-GMP3 y Q-GMP4).
Por otro lado, la muestra RUTBE presenta un SNP privado descripto en la bibliografía, junto a 39 SNPs privados ausentes en ISOGG, nuevos de este trabajo, de los cuales 5 fueron validados (Q-GMP5 al Q-GMP9). Debido a la baja cobertura de secuenciación las muestras HG03681 y HG03943 no tienen datos para estas posiciones.

3.1.2.2.2.2 Haplogrupo Q1b1a: Q-L54

Q-L54 es un sub-linaje de Q-M346, está presente en 92 muestras estudiadas en este trabajo. Obtuvimos un alto soporte estadístico para este nodo, con valores de 100 de bootstrap y la datación encontrada fue de 24.9 kya (22-28.2). Se divide en dos grandes ramas, Q-Z780 y Q-M930.

3.1.2.2.2.2.1 Haplogrupo Q1b1a2: Q-Z780

En nuestro conjunto de datos tenemos ocho muestras que representan este linaje, siendo dos de ellas secuenciadas en el presente trabajo, Z8ZMY de Belén, provincia de Catamarca y S8BAL de Malargüe, provincia de Mendoza, las demás muestras son GRC14349584_S20_L00 individuo Tsimané de Asunción de Quiquibey en Bolivia, NA19771 de México, NA19795 de México, LP6005519-DNA_B03 Zitlala Nahua de México, GRC14349585_S20_L00 individuo Maxacali de Brasil, GRC14349597_S26_L00 individuo Aymara de Perú.

Este linaje presenta una amplia distribución en Sudamérica, con representantes en varios países, por lo que no presenta una estructura espacial restringida a una única región.

El nodo representado por Q-Z780, presenta seis muestras de alta cobertura de secuenciación por lo que la datación que encontramos para este nodo es 19.3 kya (17-21.9), con un alto soporte estadístico de 100 de bootstrap.

Del análisis de SNPs para este clado encontramos un total 631 SNPs. De los cuales, 16 SNPs se comparten por todas las muestras y se encuentran descriptas en la bibliografía. Una de ellas fue validada en este trabajo, ya que si bien se encontraba descripta, no se encontraba en ISOGG validada, se le asignó el nombre Q-GMP10.

Se encontró que las muestras Z8ZMY y GRC14349584_S20_L00 no comparten ningún marcador, por lo que ambas muestras quedan sin una sub-clasificación definida dentro de Q-Z780. La muestra Z8ZMY presentó 60 SNPs privados, ausentes en ISOGG, (8 descriptos en bibliografía como equivalentes a Q-FGC47478, sin ubicación en ISOGG). Se validaron 2 SNPs privados para la muestra Z8ZMY ausentes en ISOGG, nombrados como Q-GMP13 y Q-GMP14.

Dos SNPs se encontraron compartidos por seis muestras de este clado, GRC14349585_S20_L00, GRC14349597_S26_L00, LP6005519-DNA_B03, NA19771, NA19795 y S8BAL. Ubicando a estas seis muestras dentro de un sub-linaje de Q-Z780, definido por el marcador Q-Z781, datándose en este trabajo con valores de 19.3 kya (17-21.9) y soportado por valores de 86 de bootstrap.

El marcador Q-Y2816 define un sub-linaje compartido por NA19771, NA19795 y LP6005519-DNA_B03. Y el marcador Q-Z782 define un sub-linaje dentro de Q-Y2816 que es compartido por NA19771 y NA19795. Estos nodos no pueden ser datados debido a la baja cobertura de secuenciación de las muestras. El sub-linaje Q-Y2816 presenta una estructura espacial definida dentro de México.

Otro sub-linaje de Q-Z781 está conformado por las muestras GRC14349585_S20_L00, GRC14349597_S26_L00 y S8BAL que comparten Q-YP937 y 4 SNPs más. Este nodo pudo ser datado con valores de 18.7 kya (16.5-21.2) y presenta un soporte estadístico alto, con valores de bootstrap de 85.
Q-YP937 se encontró conformado por un sub-linaje definido en este trabajo por dos SNPs compartidos por las muestras GRC14349597_S26_L00 y S8BAL, no definidos en la plataforma ISOGG. Uno de los cuales, fue reportado por Pinotti y col. 2019 [31] como privado para la muestra GRC14349597_S26_L00, y el otro SNP es nuevo, encontrado en este trabajo. Si bien este marcador nuevo no pudo ser validado por Sanger, se le asigna el nombre Q-GMP73 dado a que será retomado en los siguientes capítulos. La datación encontrada para Q-GMP73 es de 18.2 kya (16.1-20.6).

Las muestras GRC14349585_S20_L00 y GRC14349597_S26_L00 se encuentran compartiendo un nodo con 95 de bootstrap, pero no se encontró un marcador compartido entre ellas.

Por otro lado, la muestra S8BAL presentó 83 SNPs privados (4 descriptos en la bibliografía), de los cuales se validaron 2 SNPs ausentes en ISOGG, asignados como GMP11 y GMP12.

3.1.2.2.2.2.2 Haplogrupo Q1b1a1: Q-M930

Como un sub-linaje de Q-L54 se encuentra el marcador Q-M930, este último se encuentra presente en 84 muestras analizadas en este estudio. Este nodo presenta un alto soporte estadístico con valores de 100 de bootstrap. La datación para este nodo es de 15.6 kya (13.8-17.7). Q-M930 se subdivide en Q-L804 y Q-M3.

3.1.2.2.2.2.2.1 Haplogrupo Q1b1a1b: Q-L804

Este marcador ocurre únicamente entre individuos europeos. Filogenéticamente Q-L804 es un linaje hermano de los linajes más frecuentes nativos americanos, que son Q-Z780 y Q-M3. En el presente trabajo, cuatro muestras comparten este marcador, GRC13219037_S33_L002_R de Suecia, GRC12117957_S y GRC13193743_S12_L00 de Inglaterra y GRC15616706_S69_L00 de Noruega. El soporte estadístico para este clado es alto con valores de 100 de bootstrap. El cálculo de datación es de 4.4 kya (3.9-5.0).

Se encontró un sub-linaje derivado de Q-L804 compartido por las muestras GRC12117957_S y GRC13193743_S12_L00, soportado por el marcador Q-BY387 con un bootstrap 94.

3.1.2.2.2.2.2.2 Haplogrupo Q1b1a1a: Q-M3

Q-M3 es el haplogrupo más frecuente en nativos americanos y se encuentra en un total de 79 muestras de este estudio. Se encontró para el nodo Q-M3 un alto soporte estadístico, con valores de 100 de bootstrap. La datación presentó valores de 15.4 kya (13.6-17.4).
Este haplogrupo se divide en Q-M848 y en Q-Y4308. Del total de secuencias Q-M3, 78 son Q-M848 y solamente una es Q-Y4308.

3.1.2.2.2.2.2.1 Haplogrupo Q1b1a1a1 Q-M848

En nuestro conjunto de datos, la muestra GRC14349588_S Tupi Guaraní de Brasil, que es Q-Y4308 y no presenta Q-M848, está representada en la figura 3.1 dentro de Q-M848, debido a la baja representatividad en número de muestras Q-M3 y Q-Y4308. Debido a que Q-M3 y Q-M848 están representados en la figura 3.1 en el mismo nodo, tanto el bootstrap como la datación dan los mismos valores.

3.1.2.2.2.2.2.1.1.1 Clado I

El primer clado dentro Q-M3 del árbol filogenético encontrado es el formado por siete muestras GRC14349583_S51_L00 de la comunidad Aranã del Sureste de Brasil, GRC14349601_S24_L00 de la comunidad Xavante del Oeste de Brasil, HG01565 de Lima, Perú, GRC14443124_S7_L00 de Cuzco, Perú, HG02271 de Lima, Perú, GRC14349598_S5 de Pasto, Ecuador, GRC14349595_S de la comunidad Nambikwaran del Oeste de Brasil.

Los resultados del análisis de SNPs muestran que no hay un marcador compartido para todas las muestras de este clado. El bajo soporte estadístico encontrado para este clado, con valores de bootstrap menores a 10, hacen que datar este nodo no tenga un sentido biológico.
Estos resultados nos llevan a analizar este clado de manera separada, encontrando para un conjunto de muestras una mejor definición, que será explicado a continuación. Las muestras GRC14349598_S5 y GRC14349595_S además de presentar valores de bootstrap bajos no comparten marcadores por lo que quedan sin una ubicación definida dentro de Q-M848.

3.1.2.2.2.2.2.1.1.a Q-MPB118

Las muestras GRC14349583_S51_L00 y GRC14349601_S24_L00 presentan un alto soporte estadístico con valores de bootstrap de 93. Del análisis de marcadores encontramos 27 SNPs compartidos entre ellas, no descriptos en ISOGG dentro del haplogrupo Q. De los cuales, 21 SNPs fueron validados (MPB116 al MPB137) por Pinotti y col. 2019 [31] y los 6 restantes son encontrados en este trabajo como nuevos marcadores informativos de esa relación filogenética. La datación encontrada para ambas muestras fue de 9.7 kya (8.5-11). Este sub-linaje se encuentra restringido a individuos de la población brasilera.

3.1.2.2.2.2.2.1.1.b Haplogrupo Q1b1a1a1l: Q-SK281/Q-Z6659

El nodo conformado por las muestras HG01565, GRC14443124_S7_L00 y HG02271 presenta un soporte estadístico de 61 de bootstrap. Del análisis de marcadores se encontró un único marcador Q-SK281 compartido por las tres muestras, dando a estas tres muestras una ubicación definida en la filogenia de ISOGG.

Un sub-linaje dentro del mismo está conformado por las muestras GRC14443124_S7_L00 y HG02271, presentando alto soporte estadístico con valores de 100 de bootstrap. Del análisis de marcadores encontramos que estas dos últimas muestras comparten 63 SNPs entre ellos Q-Z35727. Este sub-linaje no pudo ser datado debido a que una sola de las muestras presenta alta

cobertura de secuenciación. El sub-linaje Q-SK281 muestra una estructura espacial característica de Perú.

3.1.2.2.2.2.2.1.2 Clado II

El siguiente clado está representado por las muestras SQVCW de Tartagal en Salta, TYEQC de Santa María en Catamarca, ambas secuenciadas en este trabajo, y GS00019960-ASM individuo perteneciente a la comunidad Colla de la provincia de Salta.

En conjunto este clado tiene un soporte estadístico bajo, con valores de bootstrap de 45. Del análisis de SNPs encontramos un total de 200 sitios variantes para este conjunto de muestras, pero ninguna de ellas es compartida entre las tres muestras. Si bien en este clado las muestras SQVCW y TYEQC presentan alta cobertura de secuenciación, dado el bajo bootstrap y la ausencia de SNPs compartidos, calcular la datación de este nodo no tiene significado biológico.

Se encontró para la muestra SQVCW 90 SNPs privados, ninguno se encontró descripto en las bases de datos, se validaron 6 SNPs de los mismos (GMP15 al GMP20).

3.1.2.2.2.2.2.1.2.a Haplogrupo Q1b1a1a1k2~: Q-B46_eq /B42

Las muestras TYEQC y GS00019960-ASM, forman un clado con alto soporte estadístico con valores de bootstrap de 94. Del análisis de polimorfismos para ambas muestras encontramos que ambas muestras comparten 2 SNPs, uno de los cuales es nuevo encontrado en este trabajo y el otro es descripto en la bibliografía como equivalente a Q-B46 [23]. Por lo que la ubicación de ambas muestras queda definida en ISOGG con la ubicación filogenética del mencionado marcador. Cuando la plataforma ISOGG representa el nombre del haplogrupo seguido con el símbolo "**~**", significa que no se tiene un claro conocimiento de la ubicación de este linaje en el árbol y solo tiene una asignación aproximada de su ubicación. Debido a que solamente una muestra de este nodo tiene alta cobertura de secuenciación no fue posible su datación. El marcador Q-B46 se encuentra restringido a la población argentina.

Es importante mencionar que la muestra de baja cobertura de secuenciación GS00019960-ASM presenta 195 posiciones sin información, de los 200 sitios variantes de este clado.
La muestra TYEQC presenta 108 SNPs privados, de los cuales 12 SNPs son nuevos, no descriptos por la bibliografía, de los cuales 6 SNPs fueron validados (Q-GMP15 al Q-GMP20).

3.1.2.2.2.2.2.1.3 Clado III Haplogrupo Q1b1a1a1m: Q-CTS2731

El siguiente clado está formado por cuatro muestras de las bases de datos, NA19729 de México, NA19774 de México, LP6005443-DNA_A12 y LP6005677-DNA_D01 ambos identificados como Zapotecas de México.

El nodo que sostiene a estas cuatro muestras presenta un alto soporte estadístico, con valores de 100 de bootstrap. Del análisis de polimorfismos encontramos que las cuatro muestras comparten

5 SNPs que las definen dentro de Q-CTS2731, teniendo por tanto una ubicación clara en árbol filogenético de la plataforma ISOGG. Este linaje presenta una diferenciación regional característica de México. Si bien en dos casos la localización de las muestras es de Los Ángeles, Estados Unidos, se conoce que el origen de esos individuos es mexicano.

Un sub-linaje dentro del anterior está formado por las muestras LP6005443-DNA_A12 y LP6005677-DNA_D01, las cuales comparten 89 SNPs, siendo Q-Y26467 el marcador más representativo de este linaje. En la filogenia de ISOGG, Q-Y26467 aún no presenta una ubicación clara. Los valores de bootstrap encontrados para este nodo son de 70.

Las dos muestras de alta cobertura de secuenciación de este clado pertenecen al sub-linaje Q-Y26467, y la datación calculada para este nodo presentó valores de 0.60 kya (0.53-0.68).

3.1.2.2.2.2.2.1.4 Clado IV Haplogrupo Q1b1a1a1e: Q-CTS11357/Q-M925

El siguiente es un clado númeroso formado por un total de 10 muestras de las bases de datos, las cuales se presentan como: HG01139 de Colombia, HG01142 de Colombia, SS6004476 Karitiana de Brasil, NA19783 de México, NA19735 de México, NA19732 de México, LP6005441-DNA_E10 Pima de México, NA19786 de México, LP6005519-DNA_A03 de Zitlala en México y NA19682 México.

Si bien en conjunto este nodo presenta un soporte estadístico bajo, con valores de bootstrap de 43. Del análisis de polimorfismos encontramos que las diez muestras comparten el SNP Q-CTS11357, el cual presenta una ubicación definida dentro de la filogenia de ISOGG. Las muestras de alta cobertura de secuenciación de este nodo son LP6005519-DNA_A03, LP6005441-DNA_E10 y SS6004476 y el cálculo de datación entre estas muestras es de 11.3 kya (10.3-13.2).

Q-CTS11357 muestra una amplia distribución presente tanto en Mesoamérica, como en Colombia y Brasil.

3.1.2.2.2.2.2.1.4.a Haplogrupo Q1b1a1a1e2: Q-Z5917

Las muestras HG01139 y HG01142 forman un sub-linaje dentro de Q-CTS11357 con alto soporte estadístico con valores de 100 de bootstrap. Del análisis de marcadores encontramos el SNP Q-Z5917 junto a otros 44 marcadores compartidos por ambas muestras con una ubicación clara en la filogenia de la plataforma ISOGG.

3.1.2.2.2.2.2.1.4.b Haplogrupo Q1b1a1a1e3~: SK1974

La muestra Karitiana de Brasil no se encontró compartiendo sub-linaje con otra muestra dentro de Q-CTS11357, pero se encontró que presenta de manera exclusiva el marcador Q-SK1974 sin una clara ubicación en ISOGG, pero clasificado como un sub-linaje de Q-CTS11357.

3.1.2.2.2.2.2.1.4.c Haplogrupo Q1b1a1a1e1: Q-CTS11330

Como otro sub-linaje de Q-CTS11357 las muestras NA19783, NA19735, NA19732, LP6005441-DNA_E10, NA19786, LP6005519-DNA_A03 y NA19682 comparten un nodo con un soporte estadístico alto, con valores de 88 de bootstrap. Del análisis de marcadores encontramos 2 SNPs compartidos por estas siete muestras, entre ellas Q-CTS11330. Siendo este un linaje bien definido en ISOGG. Este sub-linaje presenta una diferenciación regional restringida para la población mexicana.

Como un sub-linaje dentro de Q-CTS11330, las muestras NA19732, LP6005441-DNA_E10, NA19786, LP6005519-DNA_A03 y NA19682 comparten un nodo con valores de 100 de bootstrap y 15 SNPs. De los cuales, Q-CTS10359 es el representativo de este sub-linaje y presenta una ubicación bien definida en ISOGG.
Como un sub-linaje dentro de Q-CTS10359, las muestras NA19786, LP6005519-DNA_A03 y NA19682 comparten el nodo con valores de bootstrap de 100 y 3 SNPs compartidos, siendo Q-CTS479 representativo de este sub-linaje con clara ubicación en la plataforma ISOGG.
Las únicas muestras de alta cobertura de secuenciación para el nodo Q-CTS11330 son LP6005519-DNA_A03 y LP6005441-DNA_E10 y el cálculo de datación calculado entre ambas es de 8.4 kya (7.4-9.6).

3.1.2.2.2.2.2.1.5 Clado V Haplogrupo Q1b1a1a1n~ - Q-Y27993/Q-Y27992

El siguiente clado lo forman tres muestras, NA19664 individuo de Los Ángeles (con ancestría mexicana), LP6005519-DNA_D01 perteneciente a la comunidad Chané de Tartagal en Salta, Argentina y LP6005443-DNA_G11 de San Andrés Nuxiño en México.
Los resultados dan bajo soporte estadístico para este nodo con valores de bootstrap de 26. No se encontró ningún SNP compartido entre las tres muestras.

Las muestras LP6005519-DNA_D01 y NA19664 comparten un nodo con soporte estadístico bajo, con valores de bootstrap de 55. Del análisis de marcadores encontramos solamente el SNP Q-Y27993 compartido por ambas muestras, el cual pertenece al haplogrupo Q1b1a1a1n~ que no tiene una ubicación definida en la plataforma ISOGG.
Si bien en nuestros análisis no se encontró ningún marcador compartido para las tres muestras de este clado, la muestra LP6005443-DNA_G11 presenta de manera privada el marcador Q-Y27992 que también pertenece a Q1b1a1a1n~ según la plataforma ISOGG. Los restantes 88 SNPs que presenta esta última muestra son exclusivos y no se encuentran descriptos en la plataforma ISOGG.

La datación encontrada entre las muestras de alta cobertura de secuenciación del haplogrupo Q1b1a1a1n~ es de 16.1 kya (14.2-18.2). La baja definición de este haplogrupo en ISOGG, con solamente dos marcadores, y el alto valor de datación encontrada podrían indicar que son dos linajes diferentes. Este sub-linaje presenta una distribución en México y Argentina.

3.1.2.2.2.2.2.1.6 Clado VI

El siguiente es un clado números conformado por dieciséis muestras, HG01124 de Colombia, GS000016946-ASM individuo Wichi de Embarcación-Salta, GS000016945-ASM individuo Wichi de Embarcación-Salta, N87FK8 de Tartagal - Salta (secuenciada en este trabajo), HG01926 de Perú, HG02265 de Perú, HG01967 de Perú, GRC14349596_S individuo Parecy de Brasil, HG01950 de Perú, HG01938 de Perú, HG01977 de Perú, GS000019961-ASM individuo Colla de San Antonio de los Cobres - Salta, GRC14349593_S individuo Maxacali de Brasil, HG01961 de Perú, GRC14349599 individuo Hupda de la amazonia de Brasil, GRC14349587_S16_L00 individuo Cañari de Ecuador.
Este clado en conjunto presenta un soporte estadístico muy bajo, con valores de dos de bootstrap. El análisis de polimorfismos no reveló ningún SNP compartido. Estos resultados nos llevaron a analizar este clado de manera separada.

3.1.2.2.2.2.2.1.6.a Haplogrupo Q1b1a1a1p: Q-Z35505

Las muestras GS000016946-ASM, GS000016945-ASM y N87FK8 comparten un nodo con un soporte estadístico moderado, con valores de boostrap de 73. El análisis de SNPs de este nodo reveló un total de 51 SNPs, de los cuales solamente Q-Z35505 se encontró compartido por las tres muestras. Los 50 SNPs restantes, son únicos de la muestra N87FK8, 9 de los cuales son nuevos, no descriptos por la bibliografía [22, 23, 31, 38] ni presentes en las bases de datos [56].
Este linaje presenta una diferenciación regional restringido a la Argentina.

Es importante mencionar que GS000016946-ASM y GS000016945-ASM tienen baja cobertura de secuenciación y 50 sitios variantes de este clado no tienen datos para ambas muestras, este es el motivo por el cual no se observan diferencias en la longitud entre las ramas de las tres muestras en el gráfico filogenético.

Se seleccionaron 4 SNPs únicos para N87FK8 (sin datos para GS000016946-ASM y GS000016945-ASM), no descriptos en ISOGG para la validación (GMP30 al GMP33).
Este nodo no pudo ser datado ya que presenta una única muestra de alta cobertura de secuenciación.

La muestra HG01124 no se pudo definir compartiendo polimorfismos con otra muestra, presentando de manera privada a Q-CTS44, el cual la ubica dentro del haplogrupo Q1b1a1a1r~, sin una clara definición filogenética en ISOGG.

3.1.2.2.2.2.2.1.6.b Haplogrupo Q1b1a1a1k1 – Q-Z6658/Q-Z5915

El siguiente sub-clado está formado por las muestras HG01926, HG02265 y HG01967. Presenta un soporte estadístico de 78 de bootstrap. Del análisis de marcadores se encontró que las tres muestras comparten el SNP Q-Z6658, junto a otros 3 SNPs descriptos por ISOGG dentro del mismo haplogrupo. Este nodo no puede ser datado debido a la baja cobertura de secuenciación de las muestras que lo componen. Este linaje presenta una diferenciación regional característica de Perú.

Las muestras HG02265 y HG01967 a su vez comparten un nodo sostenido por Q-Z19436.

La muestra GRC14349596_S ha sido ubicada en el árbol filogenético compartiendo un nodo junto a las tres muestras citadas arriba. Este nodo presenta un soporte estadístico bajo, con valores de bootstrap de 37. Del análisis de SNPs no se encontró ningún marcador compartido entre esta muestra de Brasil y las tres muestras Q-Z6658. Se encontraron 26 SNPs compartidos por las muestras GRC14349596_S y N87FK8 (sin datos para las muestras GS000016946-ASM y GS000016945-ASM), entre los cuales Q-Z35497. Cinco de estos SNPs no se encontraron descriptos en ISOGG, de los cuales, pudieron validarse 4 (GMP26 al GMP29).
Las únicas dos muestras de alta cobertura de secuenciación entre estas siete muestras son la N87FK8 y GRC14349596_S, la datación entre ellas nos dio 9.6 kya (8.4-10.8).

3.1.2.2.2.2.2.1.6.c Haplogrupo Q1b1a1a1j - Q-Z19357

El sub-clado formado por HG01950, HG01938, HG01977, GS000019961-ASM y GRC14349593_S en conjunto forman un nodo con soporte estadístico con valores de 60 de bootstrap. Del análisis de polimorfismos se encontró un SNP compartido por las cinco muestras, el Q-Z19357, que les da una ubicación definida dentro de la filogenia de ISOGG. Este linaje presenta una distribución en Perú, Chaco y Brasil.

Este nodo no puede ser datado ya que solamente la muestra GRC14349593_S presenta alta cobertura de secuenciación.

El clado Q-Z19357 presenta un sub-clado formado por las muestras HG01950, HG01938, HG01977 y GS000019961-ASM, el cual presenta un soporte estadístico alto con valores de bootstrap de 100. Se encontraron un total de 12 SNPs compartidos, entre ellos Q-Z19354, los cuales forman un sub-linaje bien definido en ISOGG.

3.1.2.2.2.2.2.1.6.d Sin ubicación clara según ISOGG

El sub-clado conformado por las muestras HG01961, GRC14349599 y GRC14349587_S16_L00 presenta un soporte estadístico bajo, con valores de bootstrap de 30. No se encontró ningún polimorfismo compartido para las tres muestras.

La muestra HG01961 presenta SNPs privados que la ubican en el haplogrupo Q1b1a1a1s~ sin clara ubicación filogenética en ISOGG, entre ellos Q-Z35737.

Las muestras GRC14349599 y GRC14349587_S16_L00 comparten 15 SNPs ausentes en ISOGG, 7 de los cuales fueron validados por Pinotti y col. 2019 (Q-MPB016 al Q-MPB023) [31]. Debido a que ambas muestras presentan alta cobertura de secuenciación, la datación encontrada entre ambas es de 11.2 kya (9.9-12.7).

3.1.2.2.2.2.2.1.7 Clado VII

El siguiente clado está conformado por las muestras HG01920 individuo peruano de Lima, N8A2QN de Bariloche en la provincia de Rio Negro, M39DJ de Lavalle en provincia de Mendoza y UCNEN individuo Tehuelche de El Chalía en la provincia de Chubut. Las tres últimas muestras fueron secuenciadas en este trabajo.

El nodo que representa estas muestras tiene un soporte estadístico muy bajo, con valores de bootstrap de 3. El análisis de polimorfismos no reveló ningún SNP compartido. Estos resultados nos llevaron a analizar este clado de manera separada..

3.1.2.2.2.2.2.1.7.a Sin definición en ISOGG

Las muestras N8A2QN y HG01920 forman un nodo que presenta un bootstrap bajo de 29. El análisis de polimorfismos no reveló ningún SNP compartido por ambas muestras. Se encontró un total de 122 SNPs en este nodo pero, debido a la baja cobertura de secuenciación de la muestra HG01920, las posiciones encontradas como únicas para la muestra N8A2QN, no tiene datos para HG01920.

Se encontraron 65 SNPs privados para N8A2QN, ninguno de los cuales se encontró descripto en ISOGG, por lo que esta muestra representa una nueva rama filogenética. Validamos 6 SNPs, los cuales fueron asignados como GMP34 al GMP 40.

La muestra HG01920 de Perú presenta SNPs privados, entre ellos Q-Z35747 que la ubican en de Q1b1a1a1u~, según ISOGG.

3.1.2.2.2.2.2.1.7.b Sin definición en ISOGG

El siguiente clado se encuentra formado por dos muestras secuenciadas en este trabajo, M39DJ de Lavalle en la provincia de Mendoza y UCNEN individuo Tehuelche de El Chalía en la provincia de Chubut.

Este nodo presenta un soporte estadístico bajo, con valores de bootstrap de 40. Del análisis de polimorfismos encontramos que no hay SNPs compartidos por ambas muestras.

La muestra M39DJ presentaron 81 SNPs privados, no descriptos en ISOGG. De los cuales, se eligieron para validar 5, identificados como Q-GMP41 al Q-GMP45.

La muestra UCNEN presentó 103 SNPs privados, ausentes en ISOGG. Se eligieron 5 SNPs para validar, los cuales fueron nombrados Q-GMP46 al Q-GMP50.

3.1.2.2.2.2.2.1.8 Clado VIII Haplogrupo Q1b1a1a1i: Q-Z5908/Q-B48

El siguiente clado está formado por nueve muestras, HG02090 de Perú, LP6005677-DNA_E01 individuo Quechua de Perú, GS000016940-ASM individuo identificado como Cachi de Salta,

HG01979 de Perú, HG02277 de Perú, HG02285 de Perú, HG02299 de Perú, T4WQV de La Quiaca en la provincia de Jujuy (secuenciada en este trabajo) y GS000019989-ASM Colla de San Antonio de los Cobres en la provincia de Salta.

Este nodo presenta un elevado soporte estadístico, con valores de 100 de bootstrap. Del análisis de polimorfismos encontramos que las nueve muestras comparten 8 SNPs, de los cuales Q-Z5908 es el más representativo. Este linaje presenta solamente dos muestras de alta cobertura de secuenciación siendo estas T4WQV y LP6005677-DNA_E01 danto valores de datación de 13.6 kya (12.0-15.4). Este linaje presenta una ubicación definida dentro de ISOGG. Este linaje tiene una diferenciación regional característica para Perú y la región del Noroeste Argentino.

3.1.2.2.2.2.2.1.8.a Haplogrupo definido por Q-GMP51

Como un sub-linaje de Q-Z5908, se encontró un nuevo SNP compartido por LP6005677-DNA_E01, GS000016940-ASM, HG01979, HG02277, HG02285, HG02299, T4WQV y GS000019989-ASM, el cual se validó en este trabajo como Q-GMP51.

3.1.2.2.2.2.2.1.8.b Haplogrupo Q1b1a1a1i1a: Q-Z5910

Formando un sub-linaje dentro del anterior se encontraron las muestras LP6005677-DNA_E01, GS000016940-ASM, HG01979, HG02277, HG02285 y HG02299 que comparten el marcador Q-Z5910, descripto en ISOGG.

3.1.2.2.2.2.2.1.8.c Haplogrupo Q1b1a1a1i1a2: Q-Z35921

Otro sub-linaje lo forman las muestras LP6005677-DNA_E01, GS000016940-ASM y HG01979 comparten Q-Z35921, descripto en ISOGG.

3.1.2.2.2.2.2.1.8.d Haplogrupo Q1b1a1a1i1a1: Q-Z5911

Las muestras HG02277, HG02285 y HG02299 se encontraron compartiendo el marcador Q-Z5911, descripto en ISOGG.

3.1.2.2.2.2.2.1.8.e Haplogrupo Q1b1a1a1i1a1a: Q-Z5912

HG02285 y HG02299 comparten Q-Z5912, descripto en ISOGG.

3.1.2.2.2.2.2.1.8.f Haplogrupo no definido en ISOGG

Las muestras T4WQV y GS000016940-ASM comparten un nodo soportado con valores de 84 de bootstrap. No se encontraron SNPs compartidos entre ambas muestras, T4WQV presentó 67 SNPs exclusivos, sin datos para GS000016940-ASM. Dos de estos SNPs se encontraron validados por Pinotti y col. 2019 como MPB171 y MPB172 [31]. Los restantes marcadores encontrados son nuevos para esta muestra, no descriptos en ISOGG, de los cuales se validaron 7 SNPs, asignados como Q-GMP52 al Q-GMP58.

3.1.2.2.2.2.2.1.9 Clado IX:

El siguiente nodo está formado por la muestra 6QHWE, secuenciada en este trabajo, procedente de Santa María, provincia de Catamarca y LP6005441-DNA_G06 individuo Karitiana de la amazonia brasilera.
El soporte estadístico de este nodo es muy bajo con valores de bootstrap de 8. Los resultados del análisis de polimorfismos muestran que ningún SNP se comparte entre ambas muestras, por lo que se analiza este nodo de manera separada.

3.1.2.2.2.2.2.1.9.a Haplogrupo Q1b1a1a1v~: Q-BZ3401

La muestra LP6005441-DNA_G06 se identifica de manera privada presentando a Q-BZ3401, el cual no es un linaje bien definido en ISOGG.

3.1.2.2.2.2.2.1.9.b Haplogrupo no definido en ISOGG

La muestra 6QHWE, presentó 96 SNPs nuevos privados, no descriptos en ISOGG. De los cuales se validaron 6, designados como Q-GMP59 al Q-GMP64.

3.1.2.2.2.2.2.1.10 Clado X

El siguiente clado está conformado por GRC14349588_S Tupi Guaraní de Brasil, GRC14349592_S Kayapó de Brasil, GRC14349590_S Jíbaro de Perú, GS000017077-ASM Cachi de Salta, HG01974 de Perú, HG02104 de Perú, HG02259 de Perú, GRC14349600_S Uro de Puno en Perú, GRC14349594_S13_L00 de Asunción de Quiquibey de La Paz en Bolivia y HG01892 de Lima, Perú.
Este es un clado sin soporte estadístico. En conjunto presenta un total de 648 SNPs, de los cuales solo un marcador se encontró compartido únicamente entre dos muestras y se explica a continuación.

3.1.2.2.2.2.2.1.10.a Haplogrupo Q1b1a1a1f: Q-Z35841

Las muestras HG01974 y GS000017077-ASM presentan un soporte estadístico moderado, con valores de bootstrap de 64. Del análisis de marcadores se encontró únicamente a Q-Z35841 compartido para ambas muestras. Este SNP sostiene un linaje bien definido en ISOGG.

3.1.2.2.2.2.2.1.10.b Haplogrupo Q1b1a1a2: Q-Y4308

La muestra GRC14349588_S Tupi Guaraní de Brasil es la única muestra que en la figura 3.1 se encuentra aguas abajo de Q-M848 pero no presenta este último SNP, y presenta de manera exclusiva Q-Y4308. Todas las demás muestras de este trabajo que presentan el marcador Q-M3 también presentan Q-M848.

Las restantes muestras del clado X, GRC14349592_S, HG02259, GRC14349600_S y GRC14349594_13_L00, no pudieron asignarse compartiendo un nodo con otras muestras, y todos los marcadores se encontraron para estas muestras son exclusivos para las mismas.

3.1.2.2.2.2.2.1.11 Clado XI Haplogrupo Q1b1a1a1h: Q-Z5906

El último nodo que presenta la filogenia del haplogrupo Q en este trabajo, está representado por las muestras HG02304 de Perú, HG02146 de Perú, GS000016942-ASM individuo Cachi de Salta, GS000020273-ASM individuo Colla de San Antonio de los Cobres en Salta, GS000016951-ASM individuo Colla de San Antonio de los Cobres, HG02291 de Perú, HG01923 de Perú, LD4PC de La Quica en la provincia de Jujuy (secuenciada en este trabajo), GS000020274-ASM individuo colla de San Antonio de los Cobres, EKEFB de La Quiaca (secuenciada en este trabajo) y GRC14443115_S20_L00 de Cusco, Perú.

En conjunto las once muestras comparten un nodo con 60 de Bootstrap y soportado por el marcador Q-Z5906, defino en la plataforma ISOGG. Este nodo numeroso está conformado en su mayoría por muestras de baja cobertura por lo que estimar su datación no fue posible. Este linaje presenta una diferenciación regional característica de Perú y Argentina.

3.1.2.2.2.2.2.1.11.a Haplogrupo Q1b1a1a1h1: Q-B35

Formando un sub-clado dentro del anterior, las muestras HG02146, GS000016942-ASM, GS000020273-ASM, GS000016951-ASM, HG02291, HG01923, LD4PC, GS000020274-ASM, EKEFB y GRC14443115_S20_L00 comparten un nodo con 100 de bootstrap, soportando por Q-B35 junto a otros 48 SNPs equivalentes encontrados descriptos. De los cuales, se validaron 4 (Q-GMP65 al Q-GMP69) que no se encontraban validados en ISOGG. Se aportan 2 SNPs (Q-GMP76 y Q-GMP77) encontrados nuevos en este trabajo, no validados por Sanger, que podrían ser equivalentes a Q-B35, pero la falta de datos para las muestras de baja cobertura dificulta la confirmación.

3.1.2.2.2.2.2.1.11.b Haplogrupo Q-GMP70

Las nueve muestras GS000016942-ASM, GS000020273-ASM, GS000016951-ASM, HG02291, HG01923, LD4PC, GS000020274-ASM, EKEFB y GRC14443115_S20_L00 formando un nodo con 100 de boostrap, compartiendo marcadores no descriptos en ISOGG que se validaron asignando los nombres Q-GMP70 al Q-GMP72.

Tres muestras de este nodo presentan alta cobertura de secuenciación y los resultados de datación de 2.4 kya (2.1-2.7).

3.1.2.2.2.2.2.1.11.c Haplogrupo Q1b1a1a1h1a: Q-Z5907

Como un sub-clado del anterior, las muestras GS000016942-ASM, GS000020273-ASM, GS000016951-ASM, HG02291, HG01923, LD4PC, GS000020274-ASM, EKEFB comparten un nodo con valores de bootstrap de 86 y compartido por Q-Z5907. Se encontró un SNP nuevo no descripto que podría ser equivalente a Q-Z5907 pero la faltan datos para las muestras de baja cobertura para esta posición para su confirmación, se le asigna el nombre Q-GMP78.

Dos muestras de este nodo presentan alta cobertura de secuenciación y los resultados de datación entre ambas (LD4PC y EKEFB) es de 1.7 kya (1.5-1.9).

3.1.2.2.2.2.2.1.11.d Haplogrupo Q1b1a1a1h1a3~: Q-Z35471

Aguas abajo de Q-Z5907, se encuentra otro sub-clado soportado por Q-Z35471 presente en LD4PC, ausente para EKEFB y sin datos para las restantes muestras. Este último linaje no se encuentra bien definido dentro de la filogenia de ISOGG.

3.1.2.2.2.2.2.1.11.e Haplogrupo Q1b1a1a1h2~: Q-Z35929

De manera exclusiva la muestra HG02304 presenta Q-Z35929, el cual aun no presenta una ubicación bien establecida en ISOGG.

3.1.2.2.2.2.2.1.11.f Haplogrupo Q1b1a1a1h1b~: Q-Z35465

La muestra HG02146 presenta de manera exclusiva el SNP Q-Z35465, el cual no tiene una clara posición en la filogenia de ISOGG.

3.1.2.2.2.2.2.1.11 Haplogrupo Q1b1a1a1h1b~: No descripto en ISOGG

Se encontró un SNP nuevo no descripto antes entre las muestras EKEFB y GRC14443115_S20_L00, negativo para LD4PC y sin datos para las demás muestras, por lo que todavía no se puede precisar su ubicación. Se le asigna en nombre Q-GMP75 pero no fue validado por Sanger.

Se encontró también para la muestra EKEFB 9 SNPs exclusivos nuevos, no descriptos en ISOGG. Por otro lado, para la muestra LD4PC se encontraron 4 SNPs exclusivos nuevos, no descriptos en ISOGG.

4 DISCUSIÓN

Las secuencias completas de cromosoma Y de este estudio obtenidas desde la bibliografía han sido analizadas y reportadas en otros estudios filogenéticos del cromosoma Y [22, 23, 31, 38]. Sin embargo, en estudios de esta índole, como el presente trabajo, agregar nuevas secuencias con diferente cobertura de secuenciación y origen, permite redefinir algunos clados ya reportados o definir nuevas asociaciones filogenéticas no reportadas antes. Es decir, en algunos casos se encontraron los mismos clados ya reportados por otros autores y en otros se definen o describen clados y asociaciones filogenéticas nuevas de este estudio. Por lo tanto, todas las asociaciones entre las muestras analizadas, como los SNPs que permiten definir o redefinir un clado específico, serán discutidas en este apartado. Además, las relaciones filogenéticas y las dataciones estimadas para los sub-linajes definidos dentro de Q-M848 se contrastan con datos arqueológicos, históricos y lingüísticos en el intento de reconstruir la historia antigua de estos sub-linajes.

Las dataciones estimadas y descriptas en el capítulo 3 serán comparadas con las dataciones de trabajos recientes que reconstruyen el árbol filogenético del haplogrupo Q desde datos NGS [31, 38]. En general las dataciones que se obtienen en este trabajo son similares a las obtenidas en ambos trabajos citados, pero son más próximas a las encontradas por Pinotti y col. 2019, dado a que en este punto hemos seguido su misma metodología. Los tiempos de divergencia obtenidos por Grugni y col. 2019 han sido estimados mediante el software BEAST 1.8.3 [157], el cual es un paquete de análisis flexible para la estimación de parámetros evolutivos y permite calibrar los tiempos de divergencia especificando una tasa mutacional. Son varios los factores que generan variaciones en las estimaciones del tiempo de divergencia de los nodos, entre ellos, el número de muestras que contiene cada nodo, diferencias en la cobertura de secuenciación de las muestras, la variabilidad propia que presentan las muestras de cada estudio, el desequilibrio entre el número de sitios variantes e invariantes resultantes del método de procesamiento de secuencias. A pesar de estas diferencias, en el intento de reconstruir la historia ancestral americana, consideramos importante las estimaciones temporales de estos trabajos y serán utilizadas sus dataciones para los nodos que no pudieron datarse en este trabajo.

Se debe tener en cuenta que en el presente estudio los patrones de distribución espacial y la diferenciación regional encontrada para ciertos sub-linajes, son basados en los datos de secuencias actuales disponibles, con el sesgo de la baja cantidad de datos disponibles actualmente para cromosoma Y. Todavía existe un gran vacío informativo de secuencias de cromosoma Y de regiones como Chile, Paraguay, Uruguay, Venezuela, Las Guayanas y países centroamericanos. Países que representan vastos territorios en América cuentan con muy pocos datos disponibles: Brasil presenta solamente 11 secuencias, Ecuador sólo 2, Bolivia y Colombia sólo 1 secuencia cada uno. Por este motivo los datos que se discuten en el presente estudio, así como las hipótesis de poblamiento americano, deben tomarse como parte de los primeros avances en el uso de secuencias de cromosoma Y para el conocimiento de la distribución regional de linajes así como de uso para inferencias históricas.

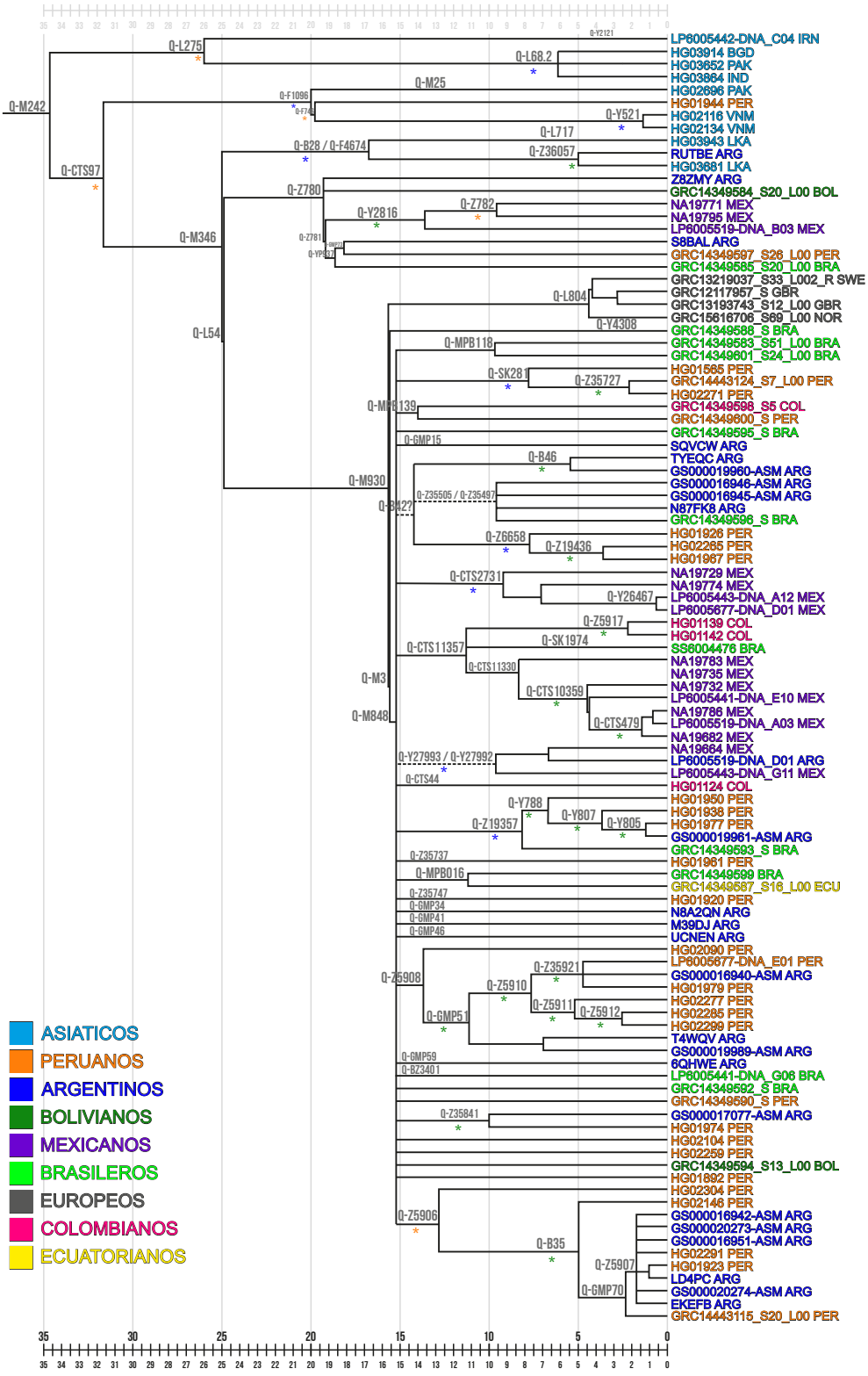


Figura 4.1: Filogenia calibrada del haplogrupo Q-M242. Las líneas discontinuas se utilizan para representar ramas que requieren un mayor estudio para una mejor definición. Los asteriscos verdes son los nodos que no pudieron datarse y la longitud de sus ramas no representa una profundidad temporal definida. Los nodos que no presentan asteriscos son los nodos datados en este estudio, los asteriscos azules representan las dataciones extraídas de [38] y los asteriscos naranjas de [31]. Para una lista completa de las dataciones utilizadas en esta figura, ver tabla anexa XI.

Haplogrupo Q-M242

Todas las secuencias analizadas en este trabajo pertenecen al haplogrupo Q-M242, siendo uno de los objetivos de esta tesis incrementar la resolución filogenética de este linaje. Q-M242 es una de las dos ramas de P1-M45, siendo la otra rama R-M207 [158]. Q-M242 se ha encontrado predominante en nativos americanos, tanto en Norteamérica, en poblaciones de habla canadiense [46], como en nativos sudamericanos [44], además tiene amplia dispersión en Asia [40, 159-161] y en Europa del Este [42].

La datación encontrada en este trabajo para este marcador es de 34.7 kya (30.7-39.4), similar a lo encontrado en la bibliografía de 35.71 kya (31.56-40.41) [31]. Si bien en la actualidad sigue sin estar claro el origen de Q-M242, se ha planteado la hipótesis de que se originó alrededor del área de las montañas de Altai (entre Asia Central y Oriental) [58]. En el capítulo V se presenta una Hipótesis de poblamiento americano basada en el haplogrupo Q-M242.

Haplogrupo Q2: Q-L275

La primera bifurcación que encontramos en el árbol filogenético (ver figura 4.1) se da para cuatro muestras del sur de Asia obtenidas de las bases de datos. La inclusión de muestras de las bases de datos, que se encuentran descriptas en la bibliografía [22, 38, 162] y que se redefinen en este estudio, son necesarias ya que permiten mejorar la definición de las nuevas secuencias que se aportan en este trabajo, así como mejorar la profundidad filogenética del haplogrupo Q.

Q-L275 no ha sido identificado en grupos nativos americanos y se ha propuesto que los orígenes de este marcador en poblaciones actuales americanas provienen de migraciones pos-colombinas desde Eurasia [47]. Se conoce que Q-L275 ocurre en poblaciones de Europa, Asia Central y Sur de Asia y se ha sugerido un origen probable en Asia occidental o Asia central [162]. Q-L275 no pudo ser datado, pero se ha encontrado reportado con una antigüedad de 26 kya (27.8 - 32.5) [31].

Haplogrupo Q1a: Q-F1096

Al igual que el clado anterior, este es un linaje que se redefine en este estudio ya que se encontraron a las mismas muestras descriptas en la bibliografía dentro de los mismos sub-linajes [22, 38]. No se pudo establecer una datación para Q-F1096 pero se encontró reportado en la bibliografía con 19.3 kya (16.7-21.9) [38]. Se conoce que Q-F1096 se divide en Q-M25 y Q-F746 [38].

Q-F746 se ha encontrado en Vietnam, Rusia, Indonesia, en nativos athabaskanos de Norteamérica [46] y presenta alta frecuencia en Groenlandia [163]. En nuestro conjunto de datos tres muestras presentan Q-F746, dos de ellas del sudeste asiático y un individuo peruano. La presencia de un individuo peruano en este sub-linaje ha sido interpretado como el resultado de una mezcla asiática post-colombina [22]. No hemos podido datar este marcador pero ha sido encontrada su datación en la bibliografía con valores de 20.7 kya (18.3-23.5) [31].

Q-M25 se observa en Europa del Este, Asia Central y en el Sur de Asia [42, 43, 47]. En nuestro conjunto de datos este linaje tiene un representante del Sur de Asia. Q-M25 no se ha observado

en los nativos americanos actuales, pero se ha informado recientemente en antiguos isleños aleutianos y antiguos athabaskanos de Norteamérica [164]. La datación reportada en la bibliografía para Q-M25 fue de 12.4 kya (10.2-14.6) [38].

Haplogrupo Q1b: Q-M346

El marcador Q-M346 ha sido descripto en Afganistán, Pakistán, Irán, Kirguistán, Mongolia, en el sur de Siberia y en Europa del Este [41-44, 49]. Los linajes nativos americanos más frecuentes Q-Z780 y Q-M3, son derivados de Q-M346 [44]. La antigüedad encontrada para este nodo en este trabajo fue de 25 kya (22-28.3), similar a lo encontrado por uno de los autores de 25.8 kya (22.8-29.2) [31] y más antigua a 17.5 kya (16.4-19.6) encontrado por [38].

Haplogrupo Q1b2a: Q-F4674

Q-F4674 es un sub-linaje de Q-B28, este último actualmente no tiene una ubicación clara en ISOOG, pero es el marcador más descripto en bibliografía. Q-B28 ha sido descripto en el sur y este de Asia y Europa [38, 56]. Si bien este marcador no pudo ser datado en el presente trabajo, fue encontrado con una antigüedad de 16.8 kya (14.7-18.9) [38]. Q-B28 no ha sido descripto antes en América, por lo que su presencia en la muestra RUTBE de San Juan es un hallazgo nuevo para la distribución de este marcador.

La muestra RUTBE de San Juan ha sido estudiada con microsatélites STRs en un trabajo reciente de nuestro grupo (haplotipo 58), junto a otras muestras de nuestra colección [49]. Se ha encontrado que dos individuos de San Juan, incluido la muestra RUTBE, presentaron el haplogrupo Q-M346* (derivado para Q-M346 y ancestral para Q-L54). RUTBE presento un haplotipo STR cercano a los linajes de Oriente Medio y Asia, en concordancia con lo encontrado en este trabajo, dado a que comparte rama filogenética con dos individuos del Sur de Asia (ver figura 4.1). Esto podría interpretarse de dos maneras distintas. La primera, como resultado de un evento reciente de migración post colombina desde Medio Oriente, ya que la migración interna podría haber llevado a este individuo desde provincias argentinas, como Tucumán, Santiago del Estero, Jujuy y La Rioja, donde la migración del Medio Oriente fue más significativa (Censo-INDEC 2001), a San Juan, Centro-Oeste de Argentina. La segunda seria que debido a la amplia distribución de Q-M346 en América, este marcador puede ser parte del acervo genético de los linajes paternos nativos americanos fundadores y Q-M346* y sus sub-linajes no hayan tenido tanto éxito como los de Q-M3 y Q-Z780 (con mayor frecuencia actualmente [44, 165]). Para confirmar esta hipótesis debería realizarse un estudio de ancestría para este individuo mediante marcadores autosómicos, que permitirían ver si tiene antepasados en esa región.

Sin embargo, el otro individuo de San Juan analizado con STRs (haplotipo 57 de [49]) también Q-M346*, presentó un haplotipo que se encuentra cerca de haplotipos siberianos y de haplotipos nativos americanos de Perú y Bolivia. En futuros análisis se buscarán los marcadores Q-B28 y Q-Z36057 presentes en RUTBE, para esta otra muestra sanjuanina.

Estos análisis han llevado a proponer que Q-M346* podría ser un tercer sub-haplogrupo autóctono de América, junto a Q-M3 y Q-Z780 [49], pero se necesitan mayores estudios.

Creemos que deberían encontrarse más individuos derivados para Q-M346 y ancestrales para Q-L54 en América, pero estudios en linajes nativos americanos que han estudiado Q-M346 no han analizado Q-L54 [166], lo cual dificulta el registro de sub-linajes Q-M346* en América.

Haplogrupo Q-L54

Se conoce que el linaje Q-L54 está presente en el pueblo Kalmyk de Asia Central y disperso en Asia Central, Norte de Asia, Canadá y América del Sur [31, 41, 45-47]. Q-L54 se presenta agua arriba de los linajes Q-Z780 y Q-M3 [44]. En este trabajo se encontró una datación de 24.9 kya (22-28.2), la cual es más antigua que la reportada en otros estudios de 18.9 kya (16.7-21.4) [31] y 15.6 kya (14.8- 17.4) [38], nuestros resultados son esperables dado que al incorporar más secuencias de alta cobertura a la filogenia completa del haplogrupo Q se incorpora más variabilidad y al profundizar su definición se puede correr la estimación del tiempo de divergencia de algunos nodos.

Haplogrupo Q-Z780

El haplogrupo Q-Z780 ha sido descripto como un linaje fundador del cromosoma Y en América y es reconocido en baja frecuencia [108]. Este linaje se encuentra ampliamente distribuido en América, con representantes de México, Colombia, Perú, Bolivia, Brasil, Argentina y Paraguay [38, 41, 49, 167]. Dada a la baja disponibilidad de secuencias que presenta actualmente este haplogrupo, se conoce poco de sus sub-linajes y se encuentran en estudio. De acuerdo a los marcadores mayormente conocidos, puede clasificarse en Q-Z781 y Q-FGC47539 [56], quedando otras ramificaciones por resolverse.

Las dos muestras incorporadas en este estudio al sub-haplogrupo Q-Z780 (Z8ZMY y S8BAL) permitieron ampliar su profundidad temporal, dando valores de 19.3 kya (17-21.9), de mayor antigüedad a lo reportado en la bibliografía de 17 kya (15.0-19.3) [31] y 14.3 kya (12.7-15.9) [38]. Estas dataciones además, son más antiguas que la encontrada para el cromosoma Y antiguo Anzick-1, perteneciente a Q-FGC47539 de 12.6 kya [168].

Dos muestras dentro de Q-Z780 no pudieron definirse dentro de un sub-clado. Una de las cuales, la muestra de nuestra colección Z8ZMY, presenta 60 SNPs exclusivos ausentes en ISOGG que aportan nueva información a este clado. La otra muestra es GRC14349584_S20_L00 que presenta 87 SNPs exclusivos ausentes en ISOGG. Por lo tanto, investigaciones futuras que incluyan un mayor número de muestras Q-Z780 serán necesarias para una mejor resolución de las muestras antes mencionadas.

El marcador Q-Z781 es el sub-linaje más representado de Q-Z780 en la actualidad y no presenta una clara ubicación en la filogenia de ISOGG. Dado el bajo número de secuencias actuales de Q-

Z780 las dataciones encontradas para Q-Z781 con valores 19.3 kya (17-21.9) se superponen a las de Q-Z780, siendo también más antiguas que las dataciones en otros trabajos con valores de 16 kya (14.1-18.1) [31] y 12.5 kya (11.0-14.0) [38] para Q-Z781.

El linaje Q-Z781 incluye un sub-linaje definido por Q-Y2816, el cual en la actualidad se encuentra presente únicamente en individuos mexicanos [169], este linaje con tres representantes presenta actualmente una diferenciación regional restringida a México.

Otro sub-linaje de Q-Z781 incluye a Q-YP937, soportados por 4 SNPs compartidos entre tres muestras, siendo una de ellas de nuestra colección. Dentro de los cuales, un SNP es nuevo reportado en este trabajo, no validado por secuenciación Sanger. Tres SNPs no se encuentran descriptos en ISOGG pero fueron reportados por Pinotti y col. 2019 [31] sin una ubicación clara en la filogenia. Aquí se pudo definir la ubicación de estos marcadores como paralelos a Q-YP937. Este nodo pudo ser datado en 18.7 kya (16.5-21.2) con mayor antigüedad a lo reportado en la bibliografía de 12.5 kya (11-14) [38].

Hemos encontrado un sub-linaje derivado de Q-YP937, no descrito en ISOGG, soportado por 2 SNPs compartidos por las muestras GRC14349597_S26_L00 y S8BAL, uno de estos marcadores fue nombrado en este trabajo como Q-GMP73 y presentó una datación de 18.2 kya (16.1-20.6). Si bien la datación de este sub-linaje podría ajustarse cuando se incorporen más muestras al mismo, esta datación permite estimar la profundidad temporal para la diferenciación regional de Q-GMP73. La asociación filogenética encontrada con Q-GMP73 evidencia vínculos entre individuos andinos y del centro oeste argentino con una antigüedad ~18.2 kya de la cual no se tienen registros arqueológicos con esa profundidad temporal para estos grupos humanos.

Por otro lado, la muestra de nuestra colección, S8BAL presenta 83 SNPs privados no descriptos en ISOGG que podrían ayudar a mejorar la profundidad del sub-linaje Q-YP937 y además expone la gran diversidad presente en este linaje todavía sin resolver.

Haplogrupo Q1b1a1: Q-M930

Q-L54 presenta como sub-linaje a Q-M930 el cual se subdivide en Q-L804 y Q-M3. La datación encontrada para Q-M930 en este trabajo es de 15.6 kya (13.8-17.7), dentro del rango encontrado en la bibliografía 17.3 kya (15.2-19.5) [31].

Haplogrupo Q1b1a1b: Q-L804

Q-L804 es un clado que se redefine en este trabajo ya que las cuatro muestras que conforman este linaje han sido descriptas en bibliografía [31] para individuos del norte de Europa. Se encontró en este trabajo una datación de 4.4 kya (3.9-5.0) coincidentes con la datación encontrada reportada de 4.4 kya (3.8-4.9) [31] para este linaje. El proyecto nórdico Q de FTDNA ha analizado este clado con un mayor número de muestras y ha encontrado que es un linaje reciente de aproximadamente 3.1 kya, específico de Europa con ocurrencia en Inglaterra, Noruega, Francia,

Escocia, Suecia y Alemania [170].

Haplogrupo Q1b1a1a: Q-M3

El haplogrupo Q-M3 ha sido descripto anteriormente como un linaje fundador del cromosoma Y en América [39, 171, 172] y es el sub-linaje más frecuente entre nativos americanos actuales [44, 47, 173, 174]. Además de ser ampliamente descripto en América, su presencia ha sido descripta también en algunas poblaciones de Siberia, pero no se sabe si estos son restos del linaje fundador o si es evidencia de migraciones regresivas desde Beringia al este de Asia [175].

La datación encontrada para este marcador en este trabajo fue de 15.4 kya (13.6-17.4), dentro del rango presentado en la bibliografía de 15 kya (13.2-16.8) [31] y 12.9 kya (11.3-14.5) [38].

En las últimas décadas, se ha ampliado la resolución interna de Q-M3 y actualmente se conoce que este linaje se subdivide en dos ramas, Q-M848 y Q-Y4308 [31, 38]. Publicaciones recientes que describen el haplogrupo Q por NGS han proporcionado un escenario más completo de la diferenciación Q en América [31, 38]. Q-M848 es un haplogrupo muy diverso con muchos linajes derivados que describiremos con más detalle.

Haplogrupo Q1b1a1a2: Q-Y4308

Como se ha mencionado anteriormente en este capítulo, Q-Y4308 es una de las dos ramas principales de Q-M3. En este trabajo no se pudo datar este nodo, pero se encontró en la bibliografía una antigüedad de 12.2 kya (10.8-13.9) [31]. Hasta ahora este marcador ha sido encontrado en poblaciones de Estados Unidos y México [176]. Se ha relacionado este linaje a individuos que hablan el idioma Algonquian [38], siendo este uno de los grupos de idiomas nativos de Norteamérica más poblados y extendidos. Además, Q-Y4308 ha sido encontrado en pueblos esquimal del extremo noreste de Asia [38, 177] y en un individuo Tupi Guaraní del Sur de Brasil como sugiere la muestra GRC14349588_S (ver figura 4.1), reportada antes en la bibliografía dentro de este sub-linaje [31].

La muestra GRC14349588_S es la única muestra de nuestro conjunto de datos que presenta Q-Y4308, Q-M3 y no presenta Q-M848. En la figura 4.1 se le asigna a la muestra GRC14349588_S su ubicación correcta de acuerdo al análisis detallado de SNPs.

Haplogrupo Q1b1a1a1: Q-M848

Q-M848 es la rama más representada del haplogrupo Q en América, tanto en tiempos modernos como en antiguos, se conoce que es más frecuente en América del Sur que en América del Norte [38, 177]. Se ha encontrado previamente a Q-M848 con una topología en estrella, donde todas las ramas que presentan este linaje se unen en el nodo central Q-M848 [22, 31, 38], ver figura 4.1.

Dada la gran representatividad de muestras Q-M848 y baja de Q-Y4308 en este trabajo, la datación de Q-M848 se superpone a la de Q-M3 con valores de 15.4 kya (13.6-17.4), quedando

dentro de los rangos estimados en la bibliografía 14.9 (13.1–16.9) kya [31] y de 12.5 kya (10.9-14.1) [38]. Los restos fósiles del hombre de Kennewick [178], encontrado a las orillas del río Columbia en Estados Unidos, pertenece al haplogrupo Q-M848 y ha sido datado en 8.3-9.2 kya [31], siendo una evidencia arqueológica que brinda margen temporal de la presencia de este linaje en Norteamérica.

Haplogrupo no clasificado por ISOGG: MPB118

La relación filogenética encontrada para las muestras GRC14349601_S24_L00 y GRC14349583_S51_L00 se ha encontrado soportada por 27 SNPs. De los cuales, 21 SNPs fueron encontrados validados en la bibliografía [31]. Los 6 SNPs restantes se aportan en este trabajo como información nueva a este linaje.

La datación que se encontró para este nodo fue de 9.7 kya (8.5-11), similar a lo reportado de 10.5 kya (9.3-10.5) [31]. Este sub-linaje dentro de Q-M848 todavía no ha sido descripto en la plataforma ISOGG. La plataforma privada YFull [179] ha incorporado ambas muestras al árbol filogenético, así como la definición de los marcadores, además incorporan a un tercer individuo de Brasil a este linaje [180], por lo que por el momento este linaje se encuentra restringido en la población brasilera.

La muestra GRC14349583_S51_L00 pertenece a la comunidad indígena Aranã ubicada en Araçuaí, Minas Gerais, Brasil. Los Aranã en su mayoría viven en la mesorregión de Jequitinhonha situada al norte de Minas Gerais. Se tienen registros de que esta etnia experimentó migraciones recientes a São Paulo, Belo Horizonte y Porto Alegre, pero todos consideran las ciudades de Coronel Murta e Araçuaí como su tierra natal [181]. Durante el siglo XVI el fuerte proceso de colonización que tuvo lugar en Brasil causó un gran impacto en los pueblos originarios, generando varios problemas, entre ellos, la pérdida de identidad étnica de varias comunidades [181]. Según el informe del Centro de Documentación Eloy Ferreira da Silva (CEDEFES) [182], la historiografía oficial señala que los Aranãs se extinguieron en el siglo XIX. Los antiguos Aranãs pertenecían a un subgrupo de los Botocudos, los cuales se dispersaban en la región del valle del río Doce en Minas Gerais [183]. A finales de la década de 1990 el grupo hoy denominado Aranã se insertó en el movimiento indígena dando inicio a la búsqueda de su identidad [183]. Con el asesoramiento de instituciones e investigadores que recurrieron a la historiografía, contribuyeron a la legitimación del uso del etnónimo Aranã. Según Carvalho, la historiografía fue fundamental para la construcción de la identidad Aranã, ya que existían lagunas e inconsistencias en el discurso de su historia y el grupo no conocía la tribu indígena a la cual pertenecían. Se utilizaron varios registros históricos para remodelar su narrativa y cimentar su cohesión social Aranã [181]. El reconocimiento étnico Aranã contemporáneo por el gobierno de Brasil data del año 2003 [183].
Por otro lado, la muestra GRC14349601_S24_L00 pertenece a un individuo del pueblo Xavante. Este grupo originario se autodenomina como A'we, y forman junto al pueblo Xerente (autodenominados Akwe) un conjunto etnolingüístico conocido en la literatura antropológica como Acuen, perteneciente a la familia lingüística Jê. Este grupo lingüístico vivía disperso en la

región hoy conocida como Centro-Oeste de Brasil, antes de la llegada de los Colonos. El nombre "Xavante" fue asignado por los no-indios con el objetivo de diferenciarlos de los otros Acuen, particularmente de los Xerente [184].

A comienzos del siglo XVIII, después del descubrimiento de oro en la región Centro-Oeste de Brasil, la llegada de mineros, conquistadores, colonos y misioneros presionó a las poblaciones indígenas locales, provocando conflictos entre éstas y los nuevos habitantes. Las poblaciones nativas reaccionaron de diferentes modos, algunas recurrieron a la práctica de ataques repentinos y a la guerra, otras, al establecimiento en el área o a la migración. A finales del siglo XVIII los antepasados de los Xavante cruzaron el río Araguaia, lo que separó definitivamente a los Xavante de los Xerente. Después de cruzar el río Araguaia, los Xavante se establecieron en la región de la Serra do Roncador, en lo que ahora es el estado de Mato Grosso. Actualmente los Xavante habitan diversos territorios geográficamente discontinuos ubicadas en el Estado de Mato Grosso, algunos de los cuales son designados hoy como territorios indígenas [185, 186].

Se tienen registros arqueológicos de ocupaciones humanas de cazadores-recolectores en el Centro Oeste y Sureste de Brasil datadas en el inicio del Holoceno (11000 a 8500 AP) [187]. Existiendo también fechas más antiguas, como por ejemplo las encontradas en Mato Grosso en los sitios Abrigo do Sol (19400 ± 1100 AP y 14470 ± 140 AP) [188] y Santa Elina (23320 ± 1000 AP y 22500 ± 500 AP) [189].

Estudios arqueológicos indican que la forma de vida de los cazadores-recolectores persistió en muchos lugares de Brasil, mucho después del advenimiento de los horticultores que producían cerámica. El Centro-Oeste brasileño parece haber sido una región de confluencia donde varias sociedades indígenas, sobre todo las agricultoras y ceramistas, se movilizaban por motivos variados. Así, en épocas pre-cabralinas (etapa histórica de Brasil antes de la llegada de los conquistadores portugueses en 1500, dirigidos por el navegante Pedro Alvares Cabral) gran parte del Centro-Oeste brasileño presentaba un extraordinario mosaico cultural [190].

Q-MPB118 da soporte a una ancestralidad de linaje compartida de grupos indígenas que hoy se reconocen como Aranã y Xavante, de los cuales no se tienen evidencias históricas de interacción. Desde su diferenciación hace aproximadamente 9.7 kya, y si bien aún es necesario estudiar más la distribución de Q-MPB118, este marcador ha estado presente entre grupos humanos del Centro-Oeste y Sureste de Brasil lo que evidencia los grandes movimientos e interacciones entre grupos humanos de dichas regiones.

Haplogrupo Q1b1a1a1l: Q-SK281

En la actualidad este es un linaje restringido a individuos peruanos [38]. En este trabajo encontramos que las muestras de las bases de datos de Perú, GRC14443124_S7_L00, HG01565 y HG02271 comparten el marcador Q-SK281. Estas últimas dos muestras han sido descriptas dentro del mismo linaje [38]. Las muestras GRC14443124_S7_L00 y HG02271 forman un sub-linaje dentro

del anterior definido por Q-Z35727, siendo este un nodo soportado por otros 63 SNPs. Sin embargo el trabajo que describe la muestra GRC14443124_S7_L00, no la define dentro de este sub-haplogrupo [31]. Por lo tanto esta última muestra pudo ser definida dentro de Q-SK281 en el presente estudio.

Si bien no se pudo datar este clado, se ha encontrado en la bibliografía valores de 7.8 kya (6.3-9.3) para Q-Z6659 [38], el cual es un nombre sinónimo de Q-SK281.

Se han encontrado en el valle de Chicama (La Libertad) los restos óseos del humano más antiguo hallado hasta el momento en Perú, nombrado como el hombre de Paiján y fechado en aproximadamente 8000 A.P. [191]. En la región Andina se han encontrado restos arqueológicos con dataciones de aproximadamente 8500 A.P [192] que evidencian la fabricación de cerámica y la domesticación de plantas y animales de horticultores seminómades [193].
Q-SK281 puede haberse diferenciado en Perú hacen aproximadamente 7.8 kya y se mantiene hasta el momento, como un linaje exclusivo de dicha región.

Haplogrupo sin definición en ISOGG: Q-MPB139

Las muestras GRC14349600_S y GRC14349598_S5 no fueron definidas juntas en el estudio filogenético presentado en la figura 3.1. Sin embargo, ambas muestras se encontraron descriptas juntas en la bibliografía soportadas por Q-MPB139 [31], el cual pudo encontrarse en nuestros datos para las mismas muestras (ver tabla anexa sección 2.7 - SNPs relevantes por nodo). Por lo que corroboramos la asociación filogenética presentada por este autor [31] y redefinimos para nuestro trabajo la ubicación de ambas muestras en el árbol filogenético de la figura 4.1. La datación encontrada en el presente estudio fue de 14 kya (12.4-15.9), similar a lo reportado por el mismo autor para este marcador con valores de 13.59 kya (12.01-15.38) [31].

La muestra GRC14349598_S5 pertenece a la comunidad indígena Pasto de la Provincia de Carchi en Ecuador. Los Pastos son una etnia indígena que habitaban junto con los Quillacingas el Area Septentrional Andina Norte, lo que actualmente es el departamento de Nariño, al sur de Colombia, y en la provincia de Carchi, al norte de Ecuador [194]. Se conoce que los Pastos estuvieron bajo el dominio del Imperio Inca poco antes de la conquista de los españoles en el siglo XVI [194]. Al tiempo que la conquista española llegó a lo que hoy es el departamento de Nariño, este territorio era habitado por un gran número de etnias diferentes. No existe, que se sepa, un informe sobre las tradiciones, usos, creencias e idiomas de los distintos grupos indígenas de esa región antes y después de la invasión española [195]. La región Andina de Nariño presenta vacíos arqueológicos con respecto a la definición de procesos culturales prehispánicos. Los estudios de la etapa paleo-indígena son prácticamente inexistentes, y por este motivo no se conoce el periodo en el cual comenzó a poblarse este territorio. Las evidencias sobre el Paleo-indio geográficamente más cercanas a los altiplanos nariñenses, provienen de los sitios del río Calima (en el departamento del Valle), y no corresponden al Área Septentrional Andina Norte [196]. Las fechas allí obtenidas son del 9000 al 7000 A.P [197]. Por lo que, en la actualidad los Pasto son un grupo

étnico en búsqueda de identidad arqueológica [196].
Por otro lado, la muestra GRC14349595_S pertenece a la etnia Uro de Puno, Perú. En tiempos pre-colombinos, esta etnia se distribuía en extensos territorios del Altiplano andino o meseta del lago Titicaca, el cual abarca territorios de Bolivia, Perú y zonas vecinas de Chile [166] y los valles interandinos de la cuenca del Pacífico [198]. La lengua de la etnia Uru en tiempos de pre-conquista se conoce como uruquilla, este idioma fue desapareciendo gradualmente luego de la conquista por el Imperio Inca [199], ocurrida entre los siglos XIII y XVI [200], y luego por la colonización española en el siglo XVI, cuando se impuso el quechua y aymara para facilitar las actividades administrativas y de evangelización [201, 202]. Actualmente, el idioma uruquilla es un lenguaje que se considera extinto [203], la mayoría de los residentes del Altiplano hablan aymara, quechua (consideradas ambas como lenguas "hermanas" de la familia andina) y castellano. La población Uro ha disminuido drásticamente como resultado del dominio inca y español [204], actualmente se distribuyen en cuatro asentamientos diferentes dispersos a lo largo de las áreas acuáticas del Altiplano, siendo conocidos en Bolivia como Uru-Chipaya, Uru-Poopo y Uru-Irohito, y en Perú como los Uros de Puno [205, 206]. La comunidad Uro de Perú, está compuesta por islas flotantes en la Bahía Puno del Lago Titicaca [207].

Los Uros han sido considerados por antropólogos y lingüistas diferentes de las etnias vecinas andinas desde todo punto de vista: tenían su propia lengua no relacionada, sus propias costumbres y creencias, su propia manera de practicar la caza, pesca, recolección y también agricultura pero, en particular, se llamaban a sí mismos como "Qhas Qut suñi", que significa "hombres del agua" en el idioma uruquilla [200, 207, 208]. Según algunos investigadores, los Uros fueron los primeros pobladores del Altiplano Andino, sin embargo, su origen es desconocido y actualmente es un tema de debate académico [206, 207, 209-211]. Se ha encontrado evidencia arqueológica que estima que los antepasados del pueblo Uro ocuparon el área de la Meseta del lago Titicaca en el año 1200 A.P. [212]. También se los considera como el posible sustrato cultural que formó parte de la antigua civilización Tiwanaku [200], la cual se encuentra en discusión la cronología de su periodo formativo. Existe hoy un abanico temporal en las estimaciones del inicio formativo de Tiwanaku que van desde 1500 A.P. según Ponce Sangines [213], y según Posnansky tuvo lugar hacen aproximadamente 14000 A.P. [214].

El marcador Q-MPB139 aporta evidencia genética a lo considerado por antropólogos y lingüistas ya el individuo Uro analizado se separan filogenéticamente de etnias vecinas andinas como Quechua y Aymara. Este sub-linaje también corrobora a estudios anteriores realizados con microsatélites de cromosoma Y que encuentran que los Uros poseen linajes exclusivos diferente a los haplotipos Aymara, Quechua y Arawak [166], a los cuales fueron asociados en otros estudios [215]. Además Q-MPB139 evidencia una ancestralidad de linaje compartida entre los Uros de Perú y los Pasto del altiplano ecuatoriano, con una diferenciación producida hacen aproximadamente 14 kya (12.4-15.9), lo que muestra además grandes movimientos de grupos humanos entre el Área Septentrional Andina Norte con el Área Andina Central en tiempos remotos de los cuales no se tiene información.

Haplogrupo Q-M848, sin mayor definición filogenética

GRC14349595_S queda sin una sub-clasificación dentro de Q-M848, en concordancia a lo presentado en la bibliografía para la misma muestra [31].

Haplogrupo no clasificado por ISOGG: GMP15

La muestra de Tartagal, SQVCW de nuestra colección no pudo ser descripta dentro de una ubicación conocida en la filogenia del haplogrupo Q. Los 90 SNPs encontrados para esta muestra son nuevos, 6 de los cuales fueron validados (GMP15 al GMP20). Los marcadores encontrados aportan nuevas posiciones informativas dentro de la filogenia de Q-M848, los cuales serian útiles en futuros estudios para incrementar la definición de este linaje desconocido.

El gran número de SNPs nuevos encontrados para esta muestra representa la existencia de una gran diversidad dentro de Q-M848, todavía sin explicar y de la cual, todavía se necesitan más datos de secuencias para intentar reconstruir la historia de las dinámicas poblacionales que existieron en la región chaqueña.

Haplogrupo Q1b1a1a1k2~: Q-B46

Las muestras TYEQC (de Catamarca) y GS00019960-ASM (Colla de Salta), forman un clado soportado por marcadores ubicados en Q-B46 dentro de la filogenia del haplogrupo Q [56]. No se pudo establecer una datación para este nodo, y otros trabajos que incluyen la muestra GS00019960-ASM tampoco han podido establecer una datación para el mismo [23, 38]. En la actualidad este linaje todavía no tiene una ubicación definida dentro de esta plataforma y en el presente trabajo aportamos 13 SNPs nuevos no descriptos en las bases de datos, siendo importantes para mejorar el conocimiento que se tiene de este linaje.

El marcador Q-B46 ha sido descripto como característico de individuos Colla [23, 177]. A mediados del siglo XVI los colonizadores españoles denominaron como Colla a todos los pueblos autóctonos que existían en el noroeste argentino a su llegada. Los pueblos del noroeste argentino de manera progresiva, sufrieron una aculturación que inició con la conquista Inca y se profundizó con la española, por lo que su cultura fue desplazada y muchos simplemente se identifican como Collas, desconociendo su verdadero origen [216].

Si bien no hemos podido establecer una datación para Q-B46, creemos que este linaje puede remontarse hasta hace 12000 años ya que se cree que los miembros del pueblo Colla son herederos de los primeros pobladores que habitaron la región andina, más tarde formaron parte del Tawantinsuyu, el gran estado inca que alcanzó su apogeo en el siglo XV. El sudeste del Incario fue llamado Kollasuyo, término que proviene del nombre de un gran grupo étnico del Lago Titicaca. Sus habitantes eran mayoritaria pero no exclusivamente Aymaras. Parte de la zona andina de esa región correspondiente al actual noroeste argentino incluía una variedad de grupos étnicos: chichas, atacamas, casabindos, cochinocas, lípez, atapamas y uros en la Puna; omaguacas, uquías, tilcaras o fiscaras o tiscaras, purmamarcas y tilianes en la Quebrada de Humahuaca;

ocloyas, paypayas, churumatas, gaypetes, osas, yalas, azamatas, tomatas, omanatas y yapanatas en el pedemonte andino [216].

La afinidad filogenética entre las muestras TYEQC de Catamarca y GS00019960-ASM (Colla) de Salta es esperable ya que los pueblos autóctonos de la región del noroeste argentino se mantuvieron desde tiempos ancestrales permanentemente relacionados entre sí, a través del intercambio, comercio, la migración, y la difusión de estilos artísticos y artesanales [217]. Agregamos a esta dinámica de interacción entre grupos humanos del noroeste argentino, el flujo génico que existió evidenciado a través de Q-B46.

Haplogrupo Q1b1a1a1p: Q-Z35505 / Q-Z35497 / Q-B43

La relación filogenética encontrada en este trabajo para las muestras GS000016946-ASM, GS000016945-ASM (ambas de Embarcación, Salta) y N87FK8 (de Tartagal de nuestra colección) se encontró soportada por Q-Z35505, el cual según ISOGG es paralelo a Q-B43 [56]. Se encontró además que N87FK8 comparte con la muestra GRC14349596_S veintiséis SNPs, entre ellos Q-Z35497, el cual también es paralelo a los dos marcadores arriba mencionados [56]. Se aportan 8 SNPs nuevos a este sub-linaje y se reconstruye en la figura 4.1 su estructura en base al análisis de SNPs.

En la actualidad el sub-linaje Q-B43 se ha descripto en individuos Wichis de Salta, Argentina [23], en individuos de Paraguay y Brasil [38] y se agrega en el presente estudio un individuo de la comunidad Pareci, de Mato Grosso, Brasil obtenido de las bases de datos [31], así como otro individuo de Salta de nuestra colección.

La datación encontrada en este trabajo para Q-Z35497, calculado únicamente entre N87FK8 y GRC14349596_S presentó valores de 9.6 kya (8.4-10.8). Este linaje se encontró datado en la bibliografía con valores 1.5 kya (0.9-2.1) [38]. Dado que en dicho trabajo la datación de Q-Z35505 fue realizada únicamente entre dos muestras (GS000016946-ASM y GS000016945-ASM) de baja cobertura de secuenciación y misma ubicación geográfica, el valor puede estar sujeto a sesgo informativo. Creemos que nuestros cálculos de datación son más ajustados para este sub-linaje dado a que son realizados con muestras de alta cobertura de secuenciación y de mayor distancia geográfica. Pero también consideramos que la datación obtenida de 1.5 kya (0.9-2.1) podría estar representando un tiempo de divergencia regional más localizado en Salta.

Tres muestras del sub-linaje Q-Z35505 son oriundas del norte de Salta, dos de las cuales pertenecen a la comunidad Wichí. Los Wichis son uno de los pueblos nativos numéricamente mayoritarios que habitan el territorio argentino. Este pueblo acondicionaba su economía al medio en que habitaban: el monte chaqueño, siendo sus principales actividades la recolección de frutos silvestres, la caza y la pesca, con alguna práctica agrícola [218]. El idioma Wichi pertenece a uno de los cuatro miembros de la familia lingüística Mataco-mataguayo [219]. En un intento de calcular la profundidad temporal del protoidioma Mataguayo se realizaron cálculos glotocronológicos que

arrojaron una fecha de 17 siglos mínimos de divergencia interna [220]. Antes de la llegada de los Europeos en el siglo XVI, la familia lingüística Mataguayo abarcaba extensiones dentro del Gran Chaco Central y Austral, en las regiones que incluyen, parte sur de Bolivia a la altura del río Pilcomayo, Noreste Argentino incluyendo noreste de Salta y Jujuy, parte noroeste de Chaco siguiendo el curso del río Bermejo limítrofe a Formosa, noroeste de Formosa y noroeste de Paraguay incluyendo la triple Frontera. Al momento de los primeros contactos con los europeos, los pueblos mataguayo parecen haber vivido dentro de los límites arriba mencionados, quedando hoy fuera de estos límites el grupo mataguayo-maká, que fue trasladado frente a la ciudad de Asunción después de la guerra del Chaco en el siglo [219].

Otro grupo étnico perteneciente a este sub-linaje son los Pareci de Mato Grosso, Brasil. El pueblo nativo Pareci se autodenomina Halití (personas del pueblo), hablan un idioma que es clasificado por los lingüistas como perteneciente a la familia Arawak [221]. El idioma Pareci se relaciona con la rama Maipure de la familia Arawak, y se estima una profundidad cronológica de unos 3.000 años [222]. Los Parecis se dividen en 4 subgrupos distintos: Kazíiniti, Waimaré, Warére y Káwali, habitaron territorios con límites bien definidos dentro de una extensa meseta que va desde las cabeceras de los ríos Arinos y Paraguay hasta las cabeceras de los ríos Guaporé y Juruena, en el Medio-Oeste del estado de Mato Grosso. Los registros escritos de la presencia de Pareci en el lugar se remontan al siglo XVIII [223, 224]. Hasta el comienzo de la colonización europea, los Parecis ocuparon este extenso territorio, actualmente están contenidos en áreas delimitadas políticamente como tierras indígenas, con diferente y reducido tamaño y diversidad ambiental. Incluso hoy, ocupan partes de su territorio tradicional, abarcando la ubicación de su mito de origen: el Ponte de Pedra, ubicado a 70 km de la ciudad de Campo Novo dos Parecis. Según la cosmología de Pareci, "en Ponte de Pedra la humanidad, el mundo, habría comenzado", siendo una región sagrada para el pueblo de Pareci [225].

El Gran Chaco se encuentra en el centro de Sudamérica, es el segundo bioma más grande de América del Sur, después del Amazonas, es compartido por cuatro países, Paraguay, Bolivia, Brasil y Argentina. En la actualidad, en todos estos países el Gran Chaco representa focos de vida silvestre bajo el creciente estrés de la expansión demográfica y la explotación económica [226]. Su área de extensión limita al norte con la Amazonia, la zona de transición la constituyen los llanos de Chiquitos, al este limita con la meseta brasileña y sus extensiones en la paraneña paraguaya y la meseta misionera, el límite con esta zona la constituyen los ríos Paraguay y Paraná. Limita al sur con la Pampa y está separada de ésta por el río Salado y la laguna Mar Chiquita, y al oeste con la región andina y sub-andina. Poco se sabe sobre el origen de las diferentes etnias del Gran Chaco, debido a la geografía de esta región caracterizada por sabanas y praderas inundables por los ríos, con ciclos de incendios [227, 228], los datos arqueológicos son escasos. Se cree que la región ha estado habitada durante al menos los últimos 4000 o 5000 años, y antes de esto toda el área era un pantano enorme [229]. La arqueología chaqueña se ha referenciado en estrecha vinculación con el Área Andina y/o Amazónica [230]. Los pueblos del Gran Chaco han sido cazadores-recolectores móviles [231] con practica de explotación animal y las plantas en los microambientes regados por los ríos [232], y con algo de practica horticultura [228]. Estas poblaciones se han

agrupado en varios grupos etnolingüísticos, llamados Mataguayos, Guaycurú, Tupí-Guaraní, Maskoy, Zamuco y Lule-Vilela [233]. A lo largo de los siglos, esta región sirvió como una especie de santuario para las poblaciones locales cuando los europeos desembarcaron en América, así como para otros que se establecieron allí como consecuencia de la presión colonial. Fue solo a principios del siglo XX que los países que afirmaban tener derecho a la región obtuvieron el control territorial y una presencia masiva que obligó a los nativos a la sedentarización [234].

En el presente estudio presentamos evidencias genéticas que relacionan dentro de un mismo sub-linaje, individuos de habla Mataguaya de Gran Chaco e individuos de habla Arawak de la región matogrossense, lindante al Gran Chaco. Respecto a esto se ha planteado anteriormente, que la población de habla Mataguaya puede haberse movido hacia el sureste debido a la presión de los grupos amazónicos, hablantes de lenguas Arawak [235]. Antes de llegar a la zona, debe haberse llevado a cabo algún tipo de intercambio entre Mataguayos y agricultores locales Arawak, debido a que algunos sitios arqueológicos del Gran Chaco revelan una cerámica decorada similar pero más rudimentaria [234]. Presentamos soporte genético a estas teorías agregando evidencias de una ancestralidad compartida entre grupos de habla Mataguaya y Arawak con una profundidad temporal de aproximadamente 9.6 kya. No podemos definir si ambos grupos presentaron un origen común o si ambos grupos presentan orígenes diferentes y al vincularse se mezclaron dejando rasgos genéticos compartidos. Las dataciones encontradas brindan la posibilidad de que Gran Chaco haya sido habitado antes de lo estimado [229].

Haplogrupo Q1b1a1a1k1 – Q-Z6658/Q-Z5915

Este es un clado que se redefine en este trabajo ya que es un sub-linaje formado únicamente para muestras de las bases de datos. Actualmente este linaje se encuentra restringido para muestras peruanas corroborando lo descripto en otros trabajos para las mismas muestras [31, 38, 236]. Si bien no hemos podido presentar una datación para Q-Z5915 se encontró en la bibliografía valores de 7.7 kya (9.2-6.2) [38].

Los evidencias líticas y óseas de los primeros cazadores-recolectores del Perú Arcaico, han sido hallados en Cueva Guitarrero con dataciones de 10610 A.P [237], las de Pikimachay, el Complejo lítico Paiján entre 11000 a 7500 A.P. [238, 239]. Se ha encontrado evidencia de cultivos humanos de calabaza con dataciones de aproximadamente 9240 - 7660 A.P., de maní de 7840 A.P., de quinua 8000 - 7500 A.P. y algodón 5490 A.P. en el valle de Ñanchoc, al norte de Perú [240]. En Cueva de Guitarrero, yacimiento arqueológico de Perú, se ha demostrado la práctica del cultivo de frijoles comunes y pallares hacen aproximadamente 7680 A.P. [241]. En Waynuna, al sur de Perú, se atestigua el uso de maíz con dataciones de 4000 A.P. [242] y en la costa árida peruana se han encontrado sitios con presencia de papa que datan de entre 4000 y 3000 A.P. [243]. La domesticación de camélidos en las excavaciones en Telarmachay, en la puna peruana, ha puesto sus orígenes en alrededor de 6000 A.P. [244, 245].

La alta cultura en Perú se da con el surgimiento de la civilización Caral, ubicada en el valle de Supe al norte de Lima, es uno de los yacimiento arqueológico más grandes de América. Algunos arqueólogos han considerado la cultura Caral como la más antigua del continente americano, con dataciones de 4090 A.P. [246]. Caral presenta elementos de arquitectura monumentales, el pueblo basó su economía en agricultura y pesca en el litoral del océano Pacífico, establecieron intercambio entre culturas y comercio con poblaciones vecinas en Perú, así como también con culturas muy alejadas geográficamente, ya que se conoce que tuvieron acceso al molusco Spondylus, característico de las aguas tropicales de Ecuador, además obtuvieron sodalita, un mineral que proviene de Bolivia, e incluso replicaron, en el entierro de un niño, el tratamiento que se daba a los muertos en la cultura Chinchorro de Chile [247]. Se conoce también que contemporáneas a Caral existieron otros centros culturales tempranos como Galgada, Kotosh en las tierras altas [248, 249] y aldeas costeras como El Paraíso, Bandurria, Huaca Prieta, Río Seco, Alto Salaverry, Culebras, Huaynuna y Tortugas [250]. La cultura Chavín de Huántar, con un origen estimado en 1000 - 600 a.C, hoy se conoce que fue la continuación de una tradición cultural mucho más antigua que fue Caral [247]. No hay evidencias de un fin catastrófico de estas civilizaciones, más bien sirvieron como base para la formación de culturas posteriores [251].

Si bien hemos mencionado antes una falta de consenso entre arqueólogos respecto la cronología temporal de Tiwanaku, algunos arqueólogos ubican esta civilización posterior a Caral y contemporáneo Wari, se los considera como dos grandes imperios en las tierras altas de los Andes entre 600 - 1000 A.D [252]. Le siguió a estos el Imperio Inca, quienes asimilaron y desarrollaron las influencias culturales de todo el territorio que dominaron. El imperio incaico, se fundó aproximadamente en el año 1200 d. C. y alcanzó su apogeo en el siglo XV, abarcó los actuales territorios correspondientes al extremo suroccidental de Colombia en la frontera, pasando por el oeste de Ecuador, Perú, el oeste de Bolivia, la mitad norte de Chile y el norte, noroeste y oeste de Argentina. La incorporación de sociedades étnicas al dominio inca, que muchas veces fue muy violenta, se dio en el marco de una misma gran tradición ecológico-cultural que permitió que las comunidades gozaran de cierta continuidad en el acceso a los recursos productivos de los diversos ambientes, en las formas de organización social y política, en una religiosidad focalizada en las potencias de la naturaleza, y en su cosmovisión. Los Incas tuvieron la intención de crear un Estado imperial cuyo fin era la unificación del mundo andino, pero cuya consolidación se vio truncada por la invasión española desde inicios del siglo XVI. La conquista española alteró profundamente la organización de este amplio espacio cultural, así como también lo hizo el surgimiento de las repúblicas a comienzos del siglo XIX [216]. Para fines de ese siglo, el desarrollo de los Estados y el capitalismo aumentaron drásticamente la fragmentación de estos pueblos, en una lucha constante por intereses territoriales y de recursos que sigue vigente en nuestros tiempos.

El marcador Q-Z6658 se habría diferenciado regionalmente en la región que hoy es Perú hacen aproximadamente 7.7 kya, estuvo presente entre cazadores recolectores de la región y posiblemente los individuos masculinos de culturas preincaicas portaron este linaje. Que luego, atravesó el proceso de conquista Inca y española, siendo un marcador que se mantiene presente hasta nuestros días entre individuos de esa región.

Haplogrupo Q-B42, sin mayor definición filogenética

Karmin y col. 2015 describe al marcador Q-B42 como ancestral para Q-B43 (paralelo a Q-Z35505 y Q-Z35497) y Q-B46 [23]. Se conoce que Q-B42 es una mutación recurrente que se usa para describir otro sub-linaje perteneciente al haplogrupo europeo R (R1b1a1b1a1a2c1a3a2a1d3). Dada esta característica de Q-B42, la plataforma ISOGG no incluye a este marcador dentro del haplogrupo Q, pero sigue siendo utilizado en trabajos actuales de reconstrucción filogenética del haplogrupo Q [38]. En el presente estudio Q-B42 ocurre entre individuos pertenecientes a los sub-linajes Q-Z35505, Q-B46 y Q-Z6658 (discutidos anteriormente) pero de manera inconstante para algunos individuos dentro de estos últimos dos sub-linajes. (ver tabla anexa sección 2.2 - información sobre las muestras).

La ocurrencia de Q-B42 se da entre individuos de: Catamarca, Colla de Salta, Lima (Perú), Wichi de Salta y Tartagal. En la figura 4.1 se reconstruye la posición de estos clados teniendo en cuenta la presencia de Q-B42 en algunas muestras. La datación encontrada en el presente estudio para Q-B42 es de 14.2 kya (12.6-16.2), más antigua que lo encontrado en la bibliografía de 10.1 kya (8.4-11.8) [38].

La región andina que ocupa hoy el territorio de Perú se considera como cuna de culturas matriz o madre, que iniciaron el proceso civilizatorio cultural originario andino [253, 254]. Se conoce la influencia que tuvieron las primeras civilizaciones del altiplano andino en el noroeste argentino, como es la cultura Tiwanaku, cuyo legado cultural ha sido encontrado en Perú, Chile y en el noroeste argentino [255]. También hemos mencionado que el Collasuyu formó parte del imperio Inca, Tawantinsuyu, y que se expandió hasta el noroeste argentino [216]. Q-B42 se diferenció hacen aproximadamente 14.2 kya (12.6-16.2) quizás entre los primeros habitantes del altiplano andino y luego posiblemente formó parte del acervo genético de las culturas que se establecieron en la región. Este sub-linaje evidencia el vinculo y flujo génico que existió entre grupos humanos de lo que actualmente es el territorio de Perú, Bolivia y el noroeste argentino desde tiempos ancestrales, y cuyo legado se conserva entre individuos autóctonos dispersos en estas regiones.

Se suma además como sub-linaje derivado de Q-B42, el Q-Z35497, que como ya mencionamos presenta una datación de 9.6 kya (8.4-10.8), ambos marcadores estarían evidenciando vínculos y flujo génico entre grupos andinos, chaqueños y amazónicos. Desde la arqueología se han encontrado evidencias culturales que reflejan que los grupos humanos chaqueños han recibido influencias periféricas tanto andinas como amazónicas [256]. Las relaciones filogenéticas encontradas en el presente estudio para Q-B42 y Q-Z35497 dan soporte genético a estos hallazgos. Es posible que las características del territorio chaqueño, con estacionalidad fluctuante en relación con los niveles de inundación del terreno, no hayan sido un obstáculo para una constante interrelación entre los grupos humanos andinos, chaqueños y amazónicos.

Haplogrupo Q1b1a1a1m: Q-CTS2731

El marcador Q-CTS2731 se encontró compartido por cuatro muestras mexicanas obtenidas de las

bases de datos. Este es un linaje que se redefine en este trabajo, siendo la relación filogenética encontrada ya reportada en otros estudios para las mismas muestras [31, 38]. Si bien no pudimos datar el nodo Q-CTS2731, se ha encontrado en la bibliografía una datación de 9.2 kya (7.5-10.9) [38]. Las bases de datos definen a este marcador restringido a poblaciones nativas de Estados Unidos y México [177].

El sub-linaje dentro del anterior sostenido por Q-Y26467 compartido por dos individuos Zapotecos de Oaxaca (LP6005443-DNA_A12 y LP6005677-DNA_D01), no tiene una ubicación clara en la filogenia de ISOGG [56]. Este sub-linaje también ha sido reportado en otros trabajos para las mismas muestras [31, 38]. La datación encontrada en este trabajo para este nodo fue de 0.6 kya (0.53-0.68), cercano a lo encontrado en la bibliografía de 0.5 kya (0.2-0.8) [38]. Se ha descripto a Q-Y26467 como característico de la población masculina zapoteca de México [177].

Actualmente existe mucha discusión sobre las dataciones encontradas para la primera ocupación humana en México. Estudios arqueológicos consideran que México cuenta con buena evidencia de la ocupación humana al menos desde 40 kya [257-261], pero las controversias respecto a las dataciones de dicho periodo todavía existen. Hay menos controversia relacionada con los sitios entre 20-30 kya y la ocupación humana en la cuenca de México y así como en diferentes sitios dispersos por México comunes en el período 10.5-13 kya [260, 261].

Se han hallado en diversas partes de México las tradiciones líticas conocidas como puntas Clovis, los usuarios de estas herramientas desarrollaron hacia 10000 A.P. una rica cultura en la que la cacería de grandes animales era frecuente, y se calcula que duró por lo menos 700 años [262]. No obstante, cabe advertir que la recolección no dejó de practicarse y probablemente fue la actividad principal de los primeros grupos humanos. En Chihuahua y San Luis Potosí tuvo lugar otra tradición más reciente de puntas denominadas Folsom (7500 A.P.), la cual se cree que se desarrolló por 1200 años y se asociaba con la cacería de bisontes hoy extintos. La diversidad de tradiciones líticas encontradas en sitios mexicanos muestran una gran cantidad de subtipologías, indicio de una considerable especialización tecnológica entre diferentes grupos humanos [261].

El territorio que hoy ocupa México ha sido considerado como cuna de grandes civilizaciones que irradiaron desde ahí influencia expansiva hacia otras regiones. Los olmecas, oriundos del trópico húmedo de la costa sur del Golfo de México, son considerados como la primera sociedad compleja que se desarrolló en Mesoamérica entre 1800 - 600 cal A.C. Su desarrollo incluyó relaciones comerciales de larga distancia con áreas adyacentes de Guatemala y las tierras altas de México [263, 264]. La cultura Maya, que se extendió por el sur de México y Centroamérica, incluía la Península de Yucatán al norte, así como los países actuales de Honduras, Belice, El Salvador y sur de Guatemala. Los mayas fueron contemporáneos a los Olmecas, mantuvieron contacto con esta cultura [265, 266]. Los Mayas tuvieron su florecimiento hacia 300 D.C a 900 D.C, construyeron grandes ciudades, la mayoría de las cuales tenían majestuosos templos piramidales. Desarrollaron métodos de agricultura, la matemática y astronomía avanzadas y también un sistema de escritura. Los Toltecas fueron dominantes de aproximadamente 900 - 1200 D.C. La civilización Azteca estaba

en su apogeo desde alrededor de 1200 D.C. hasta el momento de la llegada de los españoles, alrededor de 1500 D.C [267]. Los aztecas se establecieron en México Tenochtitlan en el centro del Valle de México, expandiendo su control hacia ciudades-estado ubicadas en los actuales estados de México, Morelos, Veracruz, Guerrero, Puebla, Oaxaca; la costa de Chiapas, Hidalgo, y parte de Guatemala [268, 269].

Otra cultura que ocurrió en México es la Zapoteca, se tienen registros de escrituras Zapotecas de 600 a.C. en los Valles de Oaxaca. El sistema de escritura zapoteco es el menos estudiado de las civilizaciones mexicanas, como la maya, aztecas y mixtecos. La escritura zapoteca temprana se encuentra principalmente en forma de inscripciones en monumentos de piedra y pinturas en las paredes de las tumbas en el Valle de Oaxaca. Estos incluyen descripciones del calendario zapoteca, la organización política, religión, gramática y el vocabulario del idioma zapoteca [270]. Los mitos zapotecas explican cómo se fundaron ciertos pueblos, pero no explican su origen. Los mitos sobre los orígenes afirman que los zapotecos nacieron de las rocas, cuevas y árboles de las regiones. Tales mitos son comunes en el sur de México [271].

Q-CTS2731 puede haberse diferenciado en México hacen aproximadamente 9.2 kya (7.5-10.9) (dataciones de [38]), estuvo presente en los primeros grupos humanos que habitaron Mesoamérica y sur de Norteamérica. Este marcador puede haber sido parte del acervo genético de las primeras civilizaciones mesoamericanas. La cultura Zapoteca presenta un marcador regional característico de Oaxaca definido como Q-Y26467 con dataciones de 0.6 kya, pero que deriva de un linaje más antiguo como es Q-CTS2731.

Haplogrupo Q1b1a1a1e: Q-CTS11357/Q-M925

Q-CTS11357 es un linaje que cuenta con diez individuos, se encontró ampliamente distribuido en América, con representantes en Colombia, Brasil y México en el presente estudio. Este es un linaje que se redefine ya que se encuentra conformado por muestras de las bases de datos que se encontraron reportadas en la bibliografía dentro del mismo linaje [31, 38]. La datación encontrada en este trabajo para Q-CTS11357 fue de 11.3 kya (10.3-13.2), cercano a lo encontrado en la bibliografía de 12.3 kya (10.9-13.9) [31] y 9.8 kya (8.4-11.2) [38]. Se conoce que la distribución de Q-M925 (equivalente a Q-CTS11357) comienza en el suroeste de Estados Unidos y se extiende a través de México en América Central [272].

Q-CTS11357 se clasifica en los siguientes sublinajes:

- El sub-linaje Q-Z5917 además de en individuos de Colombia, se ha encontrado en las bases de datos presente en individuos de Panamá y Nicaragua [38, 273].
- Q-SK1974 (descripto también con el marcador Q-Y26547), es un sub-linaje que no presenta una clara ubicación filogenética en ISOGG [56]. Se ha encontrado en este trabajo restringido a un individuo brasilero perteneciente a la comunidad Karitiana. En la actualidad solo se han encontrado representantes de este marcador en individuos de Brasil [31, 38, 274].
- El sub-linaje Q-CTS11330 se ha encontrado descripto en este y en otros trabajos representado

por individuos mexicanos [31, 38]. La plataforma privada Yfull incorpora a este sub-linaje un individuo de San Salvador. La datación encontrada en este estudio para Q-CTS11330 es de 8.4 kya (7.4-9.6), cercano a lo presentado en la bibliografía de 8.5 kya (7.1-9.9) [38]. Se aclarará aquí, que si bien en el árbol filogenético algunos individuos de este linaje se localizan geográficamente en Los Ángeles, sus orígenes son mexicanos.

- Se ha encontrado en la plataforma YFull un cuarto sub-linaje soportado por el marcador Q-BZ4012, el cual se encuentra ausente en ISOGG y del cual no se tienen datos en este estudio. Por el momento este sub-linaje se encuentra restringido a individuos de Norteamérica [38, 273].

Se conoce que entre el periodo aproximado entre 15 kya - 10 kya, tanto en Norteamérica como en Sudamérica la tipología de las puntas de proyectil halladas por la arqueología indican una gran variabilidad en las formas creadas por cazadores-recolectores de la época [275-277]. Se conoce también que las poblaciones humanas americanas de ese periodo de tiempo, al mismo tiempo y a escala continental poseían una manera de tratar las bases de las puntas de proyectil, conocido como "acanalado" que consiste en el adelgazamiento basal del material lítico [278, 279]. En Norteamérica, la distribución de diseños acanalados se extiende desde Alaska [280, 281] hasta el Norte de México [282]. Por otra parte, en Centroamérica y Sudamérica también hubo variabilidad de diseños acanalados y se los encuentra desde Guatemala [283, 284] hasta el estrecho de Magallanes [285]. El empleo del análisis de las "afinidades" tecnológicas existentes entre puntas de proyectil de Norte y Sudamérica han sido discutidas ampliamente en el intento de reconstrucción histórica y dilucidación de migraciones culturales [286-288].

Se conoce también que en el transcurso de los asentamientos humanos en América, varias especies se domesticaron o semi-domesticaron en un proceso que comenzó hace unos 10000 años [289]. Estas especies fueron la base de muchos de los cultivos que comemos hoy. A través de la influencia de los humanos, especies como el maíz, tomate, zapallo, frijol, maní, mandioca, entre muchos otros, comenzaron a domesticarse a partir de sus ancestros salvajes [290-292]. Estas plantas domesticadas fueron extendidas por los humanos desde sus respectivos centros de origen y, a la llegada de los europeos en el siglo XV, las principales especies domesticadas se extendían por toda América. Por ejemplo, se estaba plantando maíz en las tres Américas, su rango alcanzaba altas latitudes en ambos hemisferios y desde el nivel del mar hasta las grandes altitudes. Este rango refleja la diversidad genética y la plasticidad del maíz que le permite ocupar una variedad de ambientes diversos. Toda esta diversidad se explica en parte por la importancia, pasada y presente, del maíz como alimento básico para la mayoría de los habitantes de América [293].

La presencia del marcador Q-CTS11357 entre individuos de las comunidades indígenas Pima, Nahua y Karitiana exige la presentación de una revisión historia de estas culturas.
Pima: Los Akimel O'odham han sido conocidos por los occidentales como los Pima. El nombre Pima, se deriva de la frase nativa pi-nyi-match, que significa "no sé". Se aplicó a la tribu cuando los indios la usaron en respuesta a preguntas de los primeros exploradores españoles. Su nombre nativo, significa "gente del río", para distinguirse del Tohono O'odham o "gente del desierto", también conocido como Papago. Los dialectos relacionados de los dos pueblos son de la familia

lingüística uto-azteca, a veces agrupados como Los Pimas. Los Akimel O'odham ocuparon tierras ancestrales ahora mapeadas como parte del sur de Arizona, Estados Unidos y el norte de Sonora, México. Se dividieron en dos grupos principales: históricamente llamados Pima Alto y Pima Bajo. Los Pima alto vivían a lo largo de los ríos Gila y Salt. El Bajo Pima, o Nevone, como se les conoce en México, vivió a lo largo de los ríos Yaqui y Sonora mucho más al sur. El Tohono O'odham vivía al oeste inmediato del Pima Alto. No se conoce el origen de estos pueblos pero se cree que los ancestros antiguos de ambos pueblos eran los indios Hohokam, que significa "desaparecidos" en Akimel O'odham [267].

Nahua: los Aztecas, son también los llamados como Nahua. El pueblo Nahua, al igual que los Pima, pertenecen a la familia lingüística Uto-Azteca. Esta familia de idiomas exhibe hoy una extensión geográfica norte-sur excepcionalmente grande, desde el sur de Idaho, Estados Unidos, hasta El Salvador y Nicaragua en Centroamérica. La familia lingüística Uto-Azteca también es notable debido a la variedad de adaptaciones culturales entre las comunidades de hablantes, como los Shoshone, recolectores de Estados Unidos, hasta los constructores de estados urbanos como los Nahua en Mesoamérica [294]. El relato estándar sobre los orígenes de esta familia lingüística es que los Uto-Aztecas comenzaron su carrera como recolectores en el suroeste de Estados Unidos y en el noroeste de México [295-297].

Se han encontrado evidencias de maíz cultivado en el centro de México ubicadas en aproximadamente 5600 A.C. En Estados Unidos, el maíz más antiguo data del 3740 A.C. encontrado en Bat Cave en el oeste de Nuevo México [298]. Se ha sugerido que la brecha de tiempo, relativamente corta, entre el primer maíz en las tierras altas de México y el primer maíz en el suroeste de Estados Unidos significa que la migración sea el proceso cultural principal que condujo al establecimiento de la agricultura de maíz en el suroeste de los Estados Unidos [298].
Existen argumentos que atribuyen como muy probable que la comunidad de habla Proto-Uto-Azteca se haya formado en el centro de México y hayan participado en la domesticación primaria del maíz. Su expansión hacia el norte se vio impulsada por la presión demográfica resultante de un creciente compromiso con el cultivo de esta gramínea [299, 300].

Karitiana: La comunidad Karitiana de Brasil se encuentra ubicada actualmente en el norte del estado brasilero de Rondônia, más precisamente las Áreas Indígenas Karipuna e Karitiana, en las márgenes del rio igarapé Sapoti. No se conoce el origen ni la etimología de la palabra Karitiana, los propios indios afirman que les fue atribuida por caucheros que penetraron su territorio a fines del siglo XIX e inicios del siglo XX. Se autodenominan como Yjxa (traducido como, nosotros o gente), la lengua karitiana es la única remanente de la familia lingüística Arikem. Es muy poco lo que se sabe de la historia de los Karitiana, los primeros contactos con los blancos habrían ocurrido a fines del siglo XVIII, intensificándose con la llegada en masa de los caucheros al final del siglo XIX. Los Karitiana, no obstante, permanecieron ariscos al contacto sistemático hasta los años cincuenta, y la presencia de los blancos se tornó permanente sólo a partir de mediados de esta década, con la intervención del SPI (Servicio de Protección al Indio) y de misioneros salesianos [301].

El linaje Q-CTS11357 evidencia que existió hacen aproximadamente 11.3 kya (10.3-13.2) un foco poblacional en México que se extendió hasta el sudoeste de Estados Unidos, Centroamérica, llegando hasta Colombia y la Amazonia brasileña. El linaje Q-CTS11357 es parte del acervo genético de la familia lingüística Uto-Azteca. La mayor diversificación de este linaje se da entre individuos mexicanos por lo que soportamos la teoría planteada por [299, 300] de que la comunidad de habla Proto-Uto-Azteca se haya formado en el centro de México, siendo este grupo los impulsores de la domesticación primaria del maíz. Su expansión hacia el norte se vio impulsada por la presión demográfica resultante de un creciente compromiso con el cultivo de esta gramínea. La interacción ancestral de grupos humanos de México y todo el corredor centroamericano hasta la Amazonia brasilera, queda demostrada con este linaje y se ha encontrado reportada también en un estudio sobre la diversidad genética del maíz, donde se demuestra que el maíz utilizado por las poblaciones indígenas brasileñas, incluidas las de la Amazonía, están genéticamente más cerca de las muestras de maíz de México que el de otras regiones como de los Andes, esto se aplica tanto a las muestras maíz contemporáneas como a las arqueológicas [293]. Quizás los grupos humanos de estas regiones presentaban orígenes independientes pero en su crecimiento mantuvieron interacciones génicas y compartieron su tecnología y cultura. Estos resultados dan más soporte a esos resultados, que indicando con el linaje Q-CTS11357 que los habitantes de la Amazonía brasilera tenían una fuerte relación con las poblaciones de América Central y el norte de América del Sur.

Haplogrupo Q1b1a1a1n~: Q-Y27993/Q-Y27992

Este haplogrupo representado por LP6005443-DNA_G11 Mixteco de Oaxaca (México), LP6005519-DNA_D01 Chané de Salta (Argentina) y NA19664 de origen mexicano, es redefinido en este estudio ya que está conformado por muestras de las bases de datos. En la actualidad este linaje ha sido encontrado para individuos de México y Argentina [38, 302]. Este es un linaje que no se encuentra bien definido en la plataforma ISOGG, actualmente el haplogrupo Q1b1a1a1n~ se encuentra soportado por Y27992 y Y27993, considerándose paralelos entre sí. Basados en dicha definición y tomando las dos muestras de alta cobertura de este sub-linaje, la datación encontrada es de 16.1 kya (14.2-18.2), calculado solamente entre LP6005443-DNA_G11 y LP6005519-DNA_D01. Siendo estos valores más antiguos que lo encontrado en la bibliografía de 9.6 kya (8-11.2) para Q-Y27992 y calculados solamente entre LP6005519-DNA_D01 y NA19664 [38]. Consideramos que ambos cálculos están sujetos a bajo número de muestras y a baja resolución del presente linaje. Si nuestros resultados fueran correctos harían revisar la antigüedad de Q-M3 y Q-M848, lo cual podría ser esperable dado a que al incorporar más diversidad a la filogenia de Q-M3 los cálculos de su datación podrían correrse. Pero consideraremos para este nodo lo calculado en la bibliografía de 9.6 kya (8-11.2) dado a que aún es un sub-linaje que debe ser más estudiado.

Las muestras LP6005519-DNA_D01 y NA19664 fueron encontradas en nuestro análisis compartiendo el marcador Q-Y27993. Ambas muestras se han encontrado en la bibliografía descriptas dentro del mismo haplogrupo pero definidas por el marcador Q-Y27992 [38], negativo en nuestros datos para ambas muestras. Este último marcador se encuentra descripto en ISOGG como paralelo a Q-Y27993 [56]. La muestra LP6005443-DNA_G11 fue definida dentro del mismo

haplogrupo ya que presenta el marcador Q-Y27992 (de manera privada), además la inclusión de esta muestra dentro de este linaje también fue encontrada en las bases de datos [302]. Otros trabajos no han podido encontrar una ubicación clara para esta última muestra [31, 38], siendo la novedad de este trabajo poder definir a estas tres muestras dentro de un mismo sub-haplogrupo.

La pertenencia de un individuo Chané a este linaje nos lleva a remontarnos a la historia de esta comunidad. En el río Itiyuro, al norte de la provincia de Salta, Argentina, existen hoy cuatro comunidades de indígenas Chané. Estas comunidades fueron tempranamente dominadas y esclavizadas por grupos guaraní-hablantes (ubicados en Paraguay, noreste y noroeste de Argentina, sur y suroeste de Brasil y sureste de Bolivia) al menos a partir del siglo XV y probablemente antes. El origen de los Chané es desconocido y existen controversias ya que los mismos Chané afirman ser autóctonos de esa región y por otra parte, investigadores historiadores suponen que los Chané llegaron al Itiyuro huyendo del dominio Guaraní [303]. El idioma hablado por los Chané pertenece al tronco lingüístico Arawak, una de las familias lingüísticas más dispersas de América. Geográficamente este idioma se distribuye desde las Antillas y Bahamas hasta las tierras bajas de América del Sur en Argentina y desde la desembocadura del río Amazonas hasta las estribaciones de los Andes. No se conoce el origen de los Arawak, se ha sugerido que el noroeste de la Amazonía es la patria Arawak basada en la diversidad léxica de esta familia lingüística de ese centro geográfico, así como también se ha planteado que podrían tener una tierra natal en la costa atlántica o en la Amazonía occidental [304]. Se conoce a los Arawak como pacíficos, eran cazadores-recolectores y agricultores. Sus cultivos más importantes fueron mandioca, maíz, papas, batatas, frijoles, maní, pimientos, algodón y tabaco. Eran buenos canoeros y llegaron a tener embarcaciones que podían albergar hasta 100 personas, estas embarcaciones también fueron utilizadas para comercializar con pueblos Caribe, y diferentes comunidades de Sudamérica y América Central. Los Arawak también comerciaban con los pueblos norteamericanos de la costa de Florida, como Timucua y Calusa [267].

La presencia de un individuo Mixtec a este linaje, requiere de un marco histórico para este grupo étnico. Los mixtecos se han establecido en lo que hoy en día es parte de los estados mexicanos de Guerrero, Oaxaca y Puebla. Se incluyen dentro de los grupos nativos mexicanos más grandes, después de los Nahuas [305], Mayas [306] y Zapotecas [307]. Las huellas culturales Mixtecas se encuentran en los sitios del siglo XV A.C. y se mezclan con los Zapotecas. El sitio arqueológico Monte Albán, que se cree se construyó hacen aproximadamente 600 años A.C. cerca de la ciudad de Oaxaca, se considera de construcción Zapoteca, pero también puede haber incorporado rasgos de la cultura Mixteca [308]. El idioma Mixtec pertenece a la rama lingüística Otomangue [309], la cual es una de las más grandes de Mesoamérica. Se conoce que este idioma se extendió ampliamente en el centro de México, en Honduras, Nicaragua y en el norte de Costa Rica, aunque hoy sólo sobreviven las lenguas Otomangues en territorio mexicano [310]. El análisis del vocabulario asociado a la agricultura indica que los pueblos que hablaron el "proto-otomangue" tuvieron una participación relevante en la domesticación del maíz y otros cultivos, junto con otras culturas de la región [311].

La relación genética encontrada entre Mixtecos y Arawakos en este trabajo ha sido evidenciada también en un estudio genético basado en haplotipos HLA de 15681 haplotipos HLA mundiales, encontrando distancias genéticas cercanas entre Mixtecos mexicanos y Arawakos de Brasil [312].
Por otro lado, estudios lingüísticos realizados para establecer fechas de proto-idiomas, estiman para el idioma proto-Arawak fechas de 4461-4085 A.P. y para el proto-Otomanguean 6591 A.P [313]. Se conoce también que los grandes movimientos de grupos humanos se han visto impulsados por los alimentos y que grupos cazadores-recolectores encontraron ventajas en el estilo de vida agricultor [314]. Se han relacionado a los grupos Arawak y Otomangue con la domesticación inicial de mandioca, existen dos centros de diversidad de mandioca, uno está en Mesoamérica, que según la evidencia actual puede ser el lugar de origen del género, con aproximadamente 17 especies, y el otro está en Brasil, con aproximadamente 80 especies [315]. La evidencia arqueobotánica de mandioca cultivada presenta fechas de 6500 A.P. para México y 5000 A.P para las tierras bajas de la Amazonía colombiana [316].

De esta manera, la dispersión del cultivo de mandioca entre Mesoamérica y la Amazonia puede evidenciar los vínculos poblacionales que existieron en torno al desarrollo de esta práctica agrícola entre los individuos de estas regiones. En este estudio presentamos evidencias de una ancestralidad compartida entre grupos Mixtecos de habla Otomangue y Chané de habla Arawak con una profundidad temporal de aproximadamente 9.6 kya (datación de [38]). La diferenciación de Q-Y27992 podría haber ocurrido Antillas, Mesoamérica o la Amazonia. No podemos definir si ambos grupos presentaron un origen común o si ambos grupos presentan orígenes diferentes y al vincularse se mezclaron dejando rasgos genéticos compartidos.

Como ha sido propuesto por otros autores, los Chané pueden haber llegado al territorio Argentino en tiempos más recientes, probablemente huyendo de los Guaraníes hacen aproximadamente 2000 A.P. [303]. Debido a que, según datos arqueológicos el proceso de ocupación Guaraní puede haber tenido un principio hacen 2000 A.P., y se conoce que dicho proceso no respetó a las poblaciones de las regiones conquistadas [317]. Los Guaraníes eran culturas esencialmente amazónicas, por lo que grupos Arawak pueden haber llegado a Salta por migraciones desde Bolivia, huyendo de los Chiriguanos (rama Guaraní) que ocupaban territorio boliviano [303], o desde Paraguay oriental donde los sitios arqueológicos muestran secuencias regionales que muestran la continua y densa presencia Guaraní [317].

Haplogrupo Q1b1a1a1r~: Q-CTS44

En este trabajo no se pudo encontrar un nodo compartido para la muestra HG01124, se encuentra representada de manera privada por Q-CTS44, el cual no tiene una definición clara en la filogenia de ISOGG. En la bibliografía no se encontró a esta muestra definida dentro de un sub-haplogrupo, sino formando una rama politómica que sale de la raíz de Q-M848 [31, 38].

Haplogrupo Q1b1a1a1j - Q-Z19357

Lo encontrado en el presente estudio sobre la relación filogenética soportada por Q-Z19357 entre individuos de Perú y un individuo Colla de Salta de las bases de datos se encontró también reportado en la bibliografía [38]. Se aporta en este trabajo la inclusión a este linaje de la muestra GRC14349593_S Maxakalí de Minas Gerais (Brasil), no definida antes a este sub-linaje. Si bien no pudimos datar este nodo, Q-Z19357 presenta en la bibliografía dataciones de 8.1 kya (9.5-6.7) [38].

La inclusión de un individuo Maxakalí a este sub-linaje requiere de un marco histórico para este grupo étnico. Los Maxakalí (palabra en lenguaje desconocido, aplicado por primera vez en el área del río Jequitinhonha), también llamados Naknenuk, no pueden ser identificados como un grupo único, sino como un conjunto de varios. La denominación de estos grupos proviene por una articulación política como aliados y se establecieron juntos, especialmente después de 1808, cuando hubo una invasión sistemática de sus territorios y aumentaron los conflictos con otros grupos. En épocas pre-cabralinas los diversos grupos de Maxakalí ocuparon un área entre los ríos Pardo y Doce, correspondiente al sureste de Bahía, el noreste de Minas Gerais y el norte de Espírito Santo. Los remanentes de estos grupos, conocidos actualmente como Maxakalí, viven en dos áreas indígenas: Água Boa y Pradinho, hoy unificadas en la Tierra Indígena Maxakalí, en la cabecera del río Umburanas, en el noreste de Minas Gerais. Los Maxakalí pertenecen al tronco lingüístico Macro-Jê. En la actualidad la comunicación entre grupos Maxakalí es completamente en su idioma nativo. Se conoce que desde la ocupación de la región de Umburanas por los ganaderos, ha habido una disminución en la población de Maxakalí debido a la reducción de la calidad de vida y los conflictos con los agricultores de la región [318].

Se ha propuesto a la familia lingüística Macro-Jê como un grupo lingüístico que incluye a la familia Jê y a varias otras, entre ellas Maxakalí [319-321]. Aunque muchas lenguas Macro-Jê se hablan en la Amazonia brasileña, la distribución geográfica es más circunamazónica. Abarcando así, la ecorregión de Cerrado brasileño, que limita al norte con la región de la Amazonia y al oeste y suroeste con el Gran Chaco. La mayoría de lenguas Macro-Jê se concentran en el este y noreste de Brasil, aunque algunos pocos grupos habitan el centro y el suroeste de Brasil. La única lengua Macro-Jê conocida hablada fuera del actual Brasil es el Otúke que se hablaba al este del río Paraguay en Bolivia [321].

El linaje Q-Z19357 aporta evidencias una ancestralidad compartida entre grupos humanos andinos de Perú y del noroeste argentino con la etnia Maxakalí de Brasil con una profundidad temporal de 8.1 kya (9.5-6.7) (datación de [38]). No podemos definir si estos grupos presentaron un origen común o si presentaban orígenes diferentes y al vincularse se mezclaron dejando rasgos genéticos compartidos. Pero es probable que este sub-linaje se haya diferenciado en Perú, región conocida por ser cuna de grandes civilizaciones de Sudamérica, y haya extendido sus vínculos hacia civilizaciones de habla Macro-Je de Brasil. Pensamos que estos vínculos también deberían estar asociados con grupos humanos chaqueños debido a que es la región que relaciona la región

andina con la Ecorregión del Cerrado brasilero, que fue extensamente habitado por hablantes Macro-Jê en épocas pre-cabralinas. Respecto a esto, estudios lingüistas han mostrado que en las lenguas de la familia Guaicurú (habladas por mocovíes, toba, pilagás y caduveos), propias de la región chaqueña y Mato Grosso do Sul, ocurren algunos morfemas gramaticales que guardan similitud con elementos de lenguas pertenecientes al tronco lingüístico Macro-Jê, ampliamente extendido por las regiones central y oriental del Brasil [322, 323]. Sería necesario la incorporación de más individuos chaqueños a estudios genómicos de cromosoma Y para comprender más los vínculos entre estas regiones.

Haplogrupo Q1b1a1a1s~: Q-Z35737

La muestra HG01961 de Perú se encontró sin formar parte de un sub-haplogrupo aguas abajo de Q-M848, lo mismo se reportó en la bibliografía a la muestra HG01961 [31, 38]. Se encontró Q-Z35737 de manera privada para esta muestra, pero se necesitan más datos de secuencia para poder ampliar la resolución filogenética de este linaje. Q-Z35737 no se encuentra bien definido en la plataforma ISOGG y ha sido reportado presente en peruanos [272].

Haplogrupo no clasificado por ISOGG: Q-MPB016

La relación filogenética encontrada en este trabajo entre las muestras GRC14349599, Hupda de Brasil y GRC14349587_S16_L00, Cañari de Ecuador, se encontró descripta en la bibliografía [31]. En este trabajo redefinidos este linaje y aportamos 6 SNPs nuevos compartidos entre ambas muestras, no descriptos en la bibliografía y ausentes en ISOGG. La datación encontrada en este estudio para este nodo es de 11.2 kya (9.9-12.7), similar a lo encontrado en la bibliografía de 9,62 kya (8.5-10.89) [31].

Los Hupda son un pueblo nativo que habita en la cuenca alta del río Negro, en el noroeste amazónico. Su localización se da en territorios fronterizos entre Brasil y Colombia, entre los ríos Papuri y Tiquiê. Los Hupda son uno de los seis grupos distintos de habla Makú, que habitan el noroeste amazónico [324]. En la literatura etnográfica sobre el Noroeste Amazónico, los nativos de habla Makú son reconocidos como "Indios do mato" (índios de la mata) ya que rondan entre las divisorias de agua y se establecen temporalmente donde encuentran condiciones ecológicas favorables para la caza y su modo de vida [325]. No hay una autodenominación común adoptada por el propio conjunto de los Makú [326]. El término Makú, de origen Arawak, significa “siervo”, “salvaje”, y tiene una connotación peyorativa, por lo que es rechazado por el conjunto de los Makú. Debido a la influencia del movimiento indígena en la región del Río Negro, desde mediados de los años ochenta, los nombres peyorativos como el Makú, están cayendo en desuso, pero hasta ahora no ha surgido un nombre alternativo genérico y neutro, por lo que sigue siendo mantenido en la literatura etnográfica [327, 328].
Existen otras seis lenguas Makú emparentadas entre sí, que forman lo que se ha llamado la familia lingüística Makú. Hasta donde se sabe, esa familia no tiene nada que ver con las familias Tukano o Arawak, que también habitan el noroeste amazónico, exceptuando algunos evidentes y pocos prestamos lingüísticos [328, 329]. Los Makú se distribuyen en un área que limita al noroeste con el

río Guaviare (uno de los afluentes colombianos del Orinoco), al norte con el Río Negro, al sur con el río Japurá, y al sudeste con el río Uneiuxi (uno de los afluentes brasileros del río Negro). Dentro de esos territorios los Makú ocupan las "manchas" de bosque de tierra firme, donde la caza es más abundante y la vegetación más rica en especies utilizables en la alimentación o en la confección de artefactos [326, 330].

Poco se conoce acerca de la historia de los Makú en tiempos de pre-conquista. Se cree que la ocupación humana en el área del noroeste amazónico se dio probablemente en dos oleadas: primero, los Makú se establecieron en las zonas interfluviales, en las "manchas" de tierra firme, seguidamente vinieron los Arawak y los Tukano, quienes se establecieron en los altos barrancos de los ríos, en medio al igapó (terreno ribereño bajo, inundable periódicamente durante las lluvias entre abril y septiembre). El contacto ya bastante antiguo entre esos pueblos de origen y lenguas diversas, donde cada cual ocupaba porciones de tierra ecológicamente distintas, resultó en un complejo sistema de intercambios comerciales y simbólicos [328].

El otro individuo de este sub-linaje pertenece a la etnia Cañari, el término "cañari" aparece en las primeras crónicas españolas en referencia a los habitantes que ocupaban una gran extensión dentro del actual territorio de Ecuador. Existen varias hipótesis sobre el significado de la palabra "cañari" y sobre la época exacta en que apareció este término, desde la arqueología, se ha preferido designar a este grupo étnico como "proto cañaris" [331]. La arqueología ha clasificado tradicionalmente que la cronología proto cañari comprende dos fases sucesivas: Tacalshapa (300 - 1200 antes de nuestra era) y Cashaloma (1000 a 1500 de nuestra era), las cuales fueron definidas a partir de dos tipos de cerámica distintos encontrados en sitios arqueológicos del Austro ecuatoriano [332]. En las crónicas españolas se encuentra que aunque compartían un idioma y tradiciones comunes, los Cañaris se habrían dividido en varios núcleos políticos, cada cual manejado por su propio dirigente y asentados en sus propios valles. Estos núcleos habrían asimismo estado en contacto permanente entre ellos a través de intercambios, pero también de conflictos [333].

Los incas protagonizaron una conquista masiva en los Andes, la cual llegó al territorio cañari bajo el comando del soberano Túpac Yupanqui hacia aproximadamente 1463 [334-336]. Con el propósito de evitar cualquier intento de levantamiento y de contar con mayor mano de obra, esta zona será el objeto de desplazamientos poblacionales desde y hacia el territorio Cañari. A estos desplazados se los conoce bajo el nombre de mitmakuna o mitimaes [334, 337]. A pesar de todo, los incas no lograron sino difícilmente tomar el control del territorio Cañari, tarea que fue además interrumpida por la llegada de los españoles a esta zona en 1533 [332, 335]. Sobre su devenir, para algunos autores, las guerras de conquista inca y española habrían acabado con los "Cañaris" precolombinos, lo cual explicaría la "incaización" y la "castellanización" casi total de quienes se reivindican actualmente como Cañaris [336]. Otros autores, si bien reconocen esta incaización y castellanización, plantean más bien que los Cañaris no desaparecieron, sino que optaron voluntariamente por adoptar costumbres incas y españolas con el propósito de "sobrevivir" culturalmente [338].

El linaje Q-MPB016 aporta evidencias de una ancestralidad compartida entre grupos humanos de la etnia Cañari de Ecuador y Makús del noroeste amazónico de Brasil con una profundidad temporal de 11.2 kya (9.9-12.7). No podemos definir si estos grupos presentaron un origen común o si presentaban orígenes diferentes y al vincularse se mezclaron dejando rasgos genéticos compartidos. Pero sí podemos notar su separación de linajes peruanos a los que han sido asociados los Cañaris por la incaización y de linajes amazónicos, como Arawak, a los que han sido asociados los Makú. Estos resultados son importantes para la reconstrucción histórica de ambos grupos étnicos ya que no se conocen registros históricos de estos vínculos y además plantean su presencia en dichos territorios desde tiempos Arcaicos.

Haplogrupo Q1b1a1a1u~: Q-Z35747

La muestra HG01920 no pudo ser definida compartiendo marcadores con otra muestra, si bien se encontró Q-Z35747 de manera privada para esta muestra, se necesitan más datos de secuencias para poder ampliar este linaje filogenético. En la bibliografía se ha encontrado a esta muestra sin compartir nodo con otra muestra dentro de Q-M848 [31]. En la actualidad este linaje presenta poca información y carece de conocimiento sobre su distribución geográfica [272].

Haplogrupo nuevo: Q-GMP34

La muestra N8A2QN secuenciada en este estudio, se encuentra formando una nueva rama, exclusiva para la misma, derivada de Q-M848. Se encontraron 65 nuevos SNPs únicos para esta muestra de Bariloche (Río Negro), de los cuales pudieron validarse los Q-GMP34 al Q-GMP40.

Estos resultados muestran que en la filogenia del haplogrupo Q todavía existe una gran diversidad que no puede ser explicada con los datos de secuencia que se disponen. Si bien los nuevos polimorfismos encontrados para esta muestra podrán ayudar en el futuro a conocer más sobre las relaciones filogenéticas de linajes patagónicos, por el momento no podemos intentar reconstruir la historia local ni asociar a nada su gran diversidad y acervo genético.

Haplogrupo nuevo: Q-GMP41

La muestra M39DJ de nuestra colección se encuentra formando una nueva rama, exclusiva para la misma, derivada de Q-M848. Este sub-linaje identificado para un individuo de Mendoza, presentó 81 SNPs nuevos, no descriptos en ISOGG, los cuales muestran la gran diversidad no estudiada presente en la región Centro-Oeste argentina. Los marcadores validados podrán ayudar en el futuro a comprender mejor este linaje.

Haplogrupo nuevo: Q-GMP46

La muestra UCNEN de nuestra colección se encuentra formando una nueva rama, exclusiva para la misma, derivada de Q-M848. Este sub-linaje identificado en un individuo Tehuelche de El Chalía (Chubut) presentó 103 SNPs nuevos, no descriptos en ISOGG, que muestran al igual que las anteriores, la gran diversidad contenida en Q-M848 aún no estudiada. Los SNPs validados, podrán

ayudar a conocer más este linaje patagónico cuando se incorporen más secuencias filogenéticas similares.

Haplogrupo Q1b1a1a1i – Q-Z5908/Q-B48

Q-Z5908, también descripto con el marcador paralelo Q-B48 ha sido encontrado en las bases de datos y en la bibliografía con representantes de Perú [31, 177, 339] y además ha sido descripto presente en individuos Colla de Salta y en individuo de Cachi, Salta [23, 38]. En el presente trabajo corroboramos lo presentado por los mencionados autores y adicionamos a este sub-linaje una muestra de nuestra colección originaria de La Quiaca (Jujuy).

La datación encontrada en el presente trabajo para Q-Z5908 es de 13.6 kya (12.0-15.4), de mayor antigüedad a lo encontrado en la bibliografía datado con valores de 9.4 kya (8.1-10.7) [38]. Las diferencias en estos valores pueden estar sesgado por la baja cantidad de muestras utilizadas en nuestros cálculos. Hemos descripto un nuevo sub-linaje dentro de Q-Z5908 definido por Q-GMP51 (ver figura 4.1).

Como se ha explicado antes en este capítulo, el modo de vida de los pueblos de los Andes Centrales se ha caracterizado por la interrelación entre todos los grupos humanos de dicha región [217]. Como ya explicamos para el linaje Q-B42, se conoce la influencia que tuvieron las primeras civilizaciones del altiplano andino en el noroeste argentino, como es la cultura Tiahuanaco, cuyo legado cultural ha sido encontrado en Perú, Chile y en el noroeste argentino [255]. También hemos mencionado que el Collasuyu formó parte del imperio Inca, Tawantinsuyu, y que se expandió hasta el noroeste argentino [216]. Las relaciones filogenéticas encontradas para Q-Z5908 se asemejan a los vínculos entre grupos andinos discutidos para Q-B42. Las dataciones encontradas de 13.6 kya (12.0-15.4) evidencian la presencia de este linaje en territorios andinos y sus vínculos con el noroeste argentino desde tiempos Arcaicos.

Haplogrupo nuevo: Q-GMP46

La muestra 6QHWE de Catamarca perteneciente a nuestra colección, no pudo ser definida dentro de la filogenia conocida de ISOGG. Por lo que se describe una nueva rama filogenética de Q-M848 sin compartir nodo con otra muestra. Este es otro ejemplo de la gran diversidad presente en Q-M848 que no se alcanza a explicar con los datos que se disponen actualmente de las secuencias de cromosoma Y.

Haplogrupo Q1b1a1a1v~: Q-BZ3401

Q-BZ3401 es un marcador que actualmente no presenta una ubicación bien definida en ISOGG. Se ha encontrado en las bases de datos y en la bibliografía a Q-BZ3401 describiendo únicamente a la muestra LP6005441-DNA_G06 Karitiana de Brasil [38, 340], corroboramos esos hallazgos con lo encontrado en este trabajo.

Haplogrupo Q-M848, sin mayor definición filogenética

La muestra GRC14349592_S en nuestro análisis no pudo resolverse dentro de un sub-linaje dentro de Q-M848. Se ha encontrado en la bibliografía a esta muestra dentro de Q-Z19357 [31], ausente en nuestros datos para esta muestra.

Haplogrupo Q-M848, sin mayor definición filogenética

La muestra HG02259 no pudo definirse dentro de un sub-linaje de Q-M848. Este resultado se encuentra también reportado en la bibliografía y en las bases de datos [31, 38, 341].

Haplogrupo Q-M848, sin mayor definición filogenética

GRC14349594_S13_L00 queda definida como derivada para Q-M848, sin una sub clasificación dentro del mismo, estos resultados están en concordancia a lo presentado en la bibliografía [31].

Haplogrupo Q1b1a1a1f: Q-Z35841

La relación filogenética encontrada para las muestras de las bases de datos HG01974 y GS000017077-ASM soportada por Q-Z35841, no ha sido encontrada reportada en otros trabajos. Pero la muestra GS000017077-ASM ha sido definida con Q-B47 [23], el cual se conoce que es que es un sub-linaje de Q-Z35841. En el presente estudio se ha encontrado un marcador equivalente a Q-B47 presente en GS000017077-ASM. No se ha podido datar el nodo Q-Z35841 en este trabajo, y no ha sido encontrada su datación en otros estudios.

La relación filogenética encontrada para este nodo entre un individuo peruano y un Colla de Salta suma a la evidencia de flujo génico encontrado entre los grupos humanos andinos. Por lo que este linaje refuerza lo discutido para Q-Z5908 y Q-B42.

Haplogrupo Q1b1a1a1h: Q-Z5906

El linaje Q-Z5906 ha sido encontrado descripto en las bases de datos y en la bibliografía con miembros de Perú, Bolivia, comunidades Calchaquíes y Collas de Argentina [23, 38, 177, 342]. Se aportan dos muestras de nuestra colección a este sub-linaje: LD4PC y EKEFB, ambas de La Quiaca, Jujuy. Si bien en nuestro estudio no se ha podido datar el nodo Q-Z5906, se ha encontrado reportado en la bibliografía con valores de 12,88 kya (11,38-14,57) [31].

La ancestralidad de linaje compartida entre grupos humanos andinos de Perú y del noroeste argentino se refleja una vez más con el sub-linaje Q-Z5906 y sus sub-linajes Q-B35, Q-GMP70 y Q-Z5907. Siendo de esta manera, sub-linajes que refuerzan lo discutido para Q-B42, Q-Z5908 y Q-Z35841. El sub-linaje Q-Z5906, al igual que Q-B42, Q-Z5908 evidencia su presencia en territorios andinos y sus vínculos con el noroeste argentino desde tiempos Arcaicos.

Un sub-linaje derivado de Q-Z5906, presentó dataciones recientes con valores de 1.7 kya (1.5-1.9), encontrando valores bibliográficos para el mismo sub-linaje de 3.6 kya (2.9-4.3) [38]. Indicando una diferenciación regional más reciente pero abarcando la gran extensión entre Perú y el noroeste argentino, por lo que por miles de años estos grupos han estado en constante interacción y flujo génico.

5 HIPOTESIS DE POBLAMIENTO AMERICANO

Como se menciona en la sección 1.5.3 cada vez existen más evidencias arqueológicas que dan prueba de la presencia humana temprana en el continente Americano. Las excavaciones arqueológicas en la cueva de Chiquihuite en el norte de México proporcionan evidencias de ocupación humana que dataría desde hace unos 26.500 años [95]. Si bien este estudio no encuentra evidencia de ADN humano antiguo en las muestras, esto no niega la presencia humana en la cueva de Chiquihuite, ya que la probabilidad de detectar ADN humano antiguo a partir de sedimentos de cuevas se ha demostrado previamente que es baja [343]. Este sitio mexicano ahora se une a varios otros sitios arqueológicos documentados en el noreste y centro de Brasil que han arrojado evidencia que sugiere fechas para la ocupación humana entre hace 20.000 y 30.000 años [89-93, 344, 345]. Todos estos sitios analizan artefactos que evidencian presencia cultural, siendo un desafío actual para arqueólogos y genetistas encontrar ADN humano antiguo con esa profundidad temporal.

Además, el sitio arqueológico Cerutti Mastodon, al sur de California, Estados Unidos, ha encontrado in situ martillos y yunques de piedra que ocurren en asociación espacio-temporal con restos fragmentarios de un mastodonte (Mammut americanum). El análisis de múltiples muestras de hueso han indicado fechas de 130.7 ± 9.4 mil años. Estos hallazgos podrían evidenciar la presencia de una especie no identificada de *Homo* en América a principios del Pleistoceno tardío y revisa sustancialmente el momento de llegada del *Homo* a América [103].

En esta sección, se analizarán los sub-linajes masculinos nativos americanos actuales definidos en el presente estudio y se interpretará su profundidad temporal para realizar inferencias sobre la historia antigua del poblamiento americano. En la Figura 5.1 se representa en un mapa geográfico los sub-linajes derivados de Q-Z780 compartidos entre diferentes individuos del presente trabajo y su profundidad temporal. En esta imagen los individuos que viven en Los Ángeles pero que tienen origen mexicano, se han representado de manera arbitraria en la ciudad de México dado a la ausencia de información respecto a la localidad de origen de esos individuos.

Se ha estimado para Q-Z780 y Q-Z781 una profundidad temporal de 19.3 kya (17-21.9). Si bien, como se explicó al inicio del capítulo anterior, estos valores pueden estar sujetos a sesgos por bajo número de muestras, las dataciones encontradas para este linaje y su presencia en México y Brasil son consistentes con las dataciones estimadas para la presencia humana en los sitios arqueológicos mencionados en México [95] y en el Nordeste y Centro-oeste de Brasil [89-93, 344, 345]. Como se observa en la figura 5.1, hace ~19 mil años Q-Z781 podría haber presentado una amplia dispersión en Mesoamérica y Sudamérica. Además, puede haber presentado una diferenciación regional de Q-Y2816 y Q-Z782 en México (dataciones sin resolver), Q-YP937 entre Perú, Brasil y Argentina hace ~18.7 kya (16.5-21.2) y Q-GMP73 entre individuos andinos de Perú y el Centro Oeste de Argentina hace ~18.2 kya (16.1-20.6). Si bien esta especificidad regional podría ajustarse al estudiar en mayor profundidad estos marcadores en la población americana,

podríamos estimar que hacen ~19 mil años habría existido un foco poblacional donde hoy es México y un foco poblacional en Sudamérica entre lo que hoy es Perú, Argentina y Brasil. Al estar relacionados por un linaje compartido, es probable que estas poblaciones hayan mantenido vínculos y realizado intercambios culturales. Sobre la base del presente estudio se propone un poblamiento de Cronología Larga para Mesoamérica y Sudamérica, anterior a 18000 años.

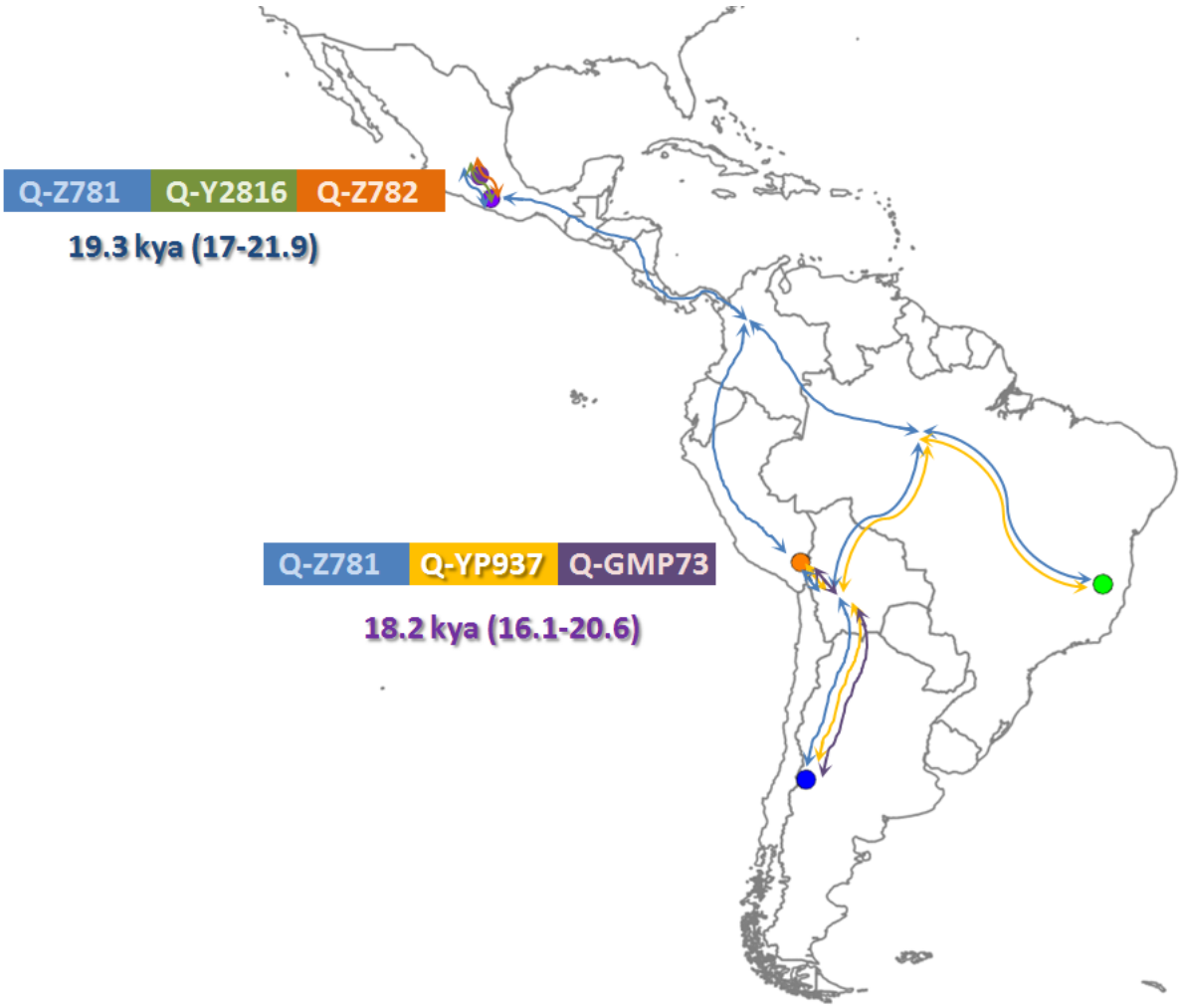


Figura 5.1. Profundidad temporal del sub-linaje Q-Z781, dispersión geográfica y diferenciación regional. El color de las rutas migratorias coincide con color del SNP representado. Las rutas migratorias representadas en diferentes colores son tentativas. La datación expresada en azul corresponde a Q-Z781 y la expresada en violeta corresponde a Q-GMP73.

Como se explicó en la sección 1.5.1, el periodo de tiempo que abarca el Younger Dryas (YD) 12.900-11.600 cal AP, está marcado por un evento singular que afectó el ambiente y el clima de todo el mundo. Si bien los orígenes de dicho evento siguen siendo estudiados y no existe un consenso, el estudio de cómo el evento YD afectó a las poblaciones humanas de la época viene siendo considerado por arqueólogos de diferentes partes del mundo [64].

Un estudio llevado a cabo por arqueólogos para probar si las poblaciones humanas en Norteamérica se vieron afectadas por un cambio climático abrupto u otros factores ambientales asociados con el YD, indican que una importante disminución de la población humana (cuello de botella), o alternativamente, reorganización de las poblaciones (es decir, cambios drásticos en los patrones de asentamiento), ocurrió en amplias áreas de América del Norte hacen aproximadamente 12.900 cal AP. También, evidencian que se produjeron disminuciones o cambios similares en gran parte del resto del hemisferio norte, Europa, partes de Asia y África [346]. En contraste, Medio Oriente no evidencia un declive significativo de la población humana al inicio del YD, en su lugar se indica un crecimiento seguido de una meseta prolongada, lo que sugiere que Medio Oriente puede haber servido como refugio para los humanos [64, 346, 347].

Lamentablemente los estudios arqueológicos abocados al poblamiento inicial América no se han enfocado en cómo afectó el evento YD en poblaciones de Mesoamérica y Sudamérica, por lo que prácticamente no se considera en la literatura arqueológica [348]. Esto podría deberse a una desestimación al considerar los eventos ambientales ocurridos entre 12.900-11.600 cal AP únicamente como evento de enfriamiento y como tal, no comparable al enfriamiento ocurrido en el hemisferio norte en dicho periodo. Por lo que, algunos arqueólogos sudamericanos que consideran que el clima no fue un factor importante en la determinación de la mayoría de las distribuciones humanas, no analizan los efectos de los episodios ambientales ocurridos en el periodo YD [349-351].

Estudios arqueológicos que intentan relacionar cómo influyó el periodo YD en las poblaciones humanas en Sudamérica, han realizado investigaciones en el sitio arqueológico patagónico de Pilauco, en la ciudad de Osorno al sur de Chile, el cual es reconocido como un importante recurso paleontológico y arqueológico debido a su rico y abundante conjunto de mamíferos extintos del Pleistoceno sudamericano y restos culturales. Se ha encontrado en este sitio una huella humana asociada con los restos de la megafauna y una semilla perforada manualmente que probablemente se usó como adorno [352]. Los artefactos líticos producidos por los primeros pobladores humanos de Pilauco solo se encuentran en estrecha asociación con restos de megafauna. Esta conexión observada sugiere que los humanos estaban explotando la megafauna extinta a través de la búsqueda y/o la caza. No se encontraron artefactos humanos después de la extinción de la megafauna en la capa YDB. Esta ausencia de restos sugiere que los humanos abandonaron el área después de las extinciones de la megafauna y/o experimentaron una disminución y/o reorganización de la población regional [66], similar a la propuesta para América del Norte [346]. Además, los estudios realizados en Pilauco para inferir la historia vegetativa local antes y después de YD encuentran una disminución importante en la abundancia y diversidad de materiales vegetales y un cambio pronunciado en la composición taxonómica de los conjuntos de plantas [66].

Estudios genéticos enfocados en el poblamiento inicial de América que aborden la profundidad temporal que abarca el YD son escasos, por lo que actualmente es un campo vacío el análisis de los efectos que causaron a nivel del genoma humano los cambios ambientales del YD. Cada vez

existen más evidencias de quema de biomasa a gran escala durante el YD, que además de ser registradas en Europa, Asia, Norteamérica, se incluyen evidencias en Sudamérica, donde el registro arqueológico es consistente con la actividad regional de incendios forestales, lo que indica que la quema de biomasa anómala alcanzó altas latitudes del hemisferio sur [66, 353]. Estas alteraciones medioambientales podrían haber afectado potencialmente a los genomas de poblaciones humanas arcaicas de múltiples formas: la materia particulada de la quema de biomasa se ha relacionado recientemente con diversas formas de daño del ADN tanto in vivo como in vitro [354, 355], y la extinción de las fuentes primarias de alimentos podrían haber alterado drásticamente las dietas de las poblaciones humanas, exponiéndolas así a nuevos mutágenos que los seres humanos aún no habían evolucionado para evitar o metabolizar [356, 357].

En este trabajo se propone que los focos poblacionales que existían en lo que hoy es México, Perú y otras partes de Sudamérica podrían haberse visto afectados de manera drástica por el episodio ambiental que ocurrió durante el YD. Si bien sería necesario la inclusión de modelos estadísticos que estimen el tamaño poblacional de Q-Z780 hacen ~19000 años, este linaje puede haber representado a una población mucho más grande antes del YD y luego de este evento, sufrió un declive abrupto y no tuvo un resurgimiento que iguale en número su amplia distribución antes del YD. Por lo que, los eventos ambientales ocurridos durante el YD podrían representar uno de los motivos fundamentales de la baja frecuencia actual de este linaje.

El linaje nativo americano más frecuente Q-M848, con una profundidad temporal de 15.4 kya (13.6-17.4), coexistió antes del evento YD junto a Q-Z780, pero a diferencia de Q-Z780, para Q-M848 proponemos que el evento YD podría haber actuado como impulsor de su expansión y diversificación, generando la topología en forma de estrella de este haplogrupo (ver Figura 4.3). Es sorprendente notar que el tiempo estimado para la capa YDB (~12.800 cal AP) cae dentro del límite inferior y superior de todos los tiempos de divergencia de los sub-linajes de Q-M848 encontrados, o posterior, ver figuras anexas X y XI.

Sub-linajes de Q-M848 que abarcan ~12.800 cal AP, muestran una gran expansión en esos tiempos: Q-CTS11357 muestra un foco poblacional en México y una mayor diferenciación regional en México y Centroamérica, con una dispersión hacia Norteamérica, Centroamérica y hasta la Amazonia; Q-MPB139 evidencia grandes distancias recorridas desde los Andes Centrales hacia los Andes Septentrionales (o viceversa), probablemente en búsqueda de mejores condiciones de vida para asentarse; Q-B42, Q-Z5908 y Q-Z5906 evidencian un foco poblacional en los Andes Centrales, en lo que hoy es Perú con dispersión hacia tierras bajas en lo que hoy es Bolivia y Argentina, en el caso de Q-B42 además, se encuentran vínculos entre grupos humanos andinos, chaqueños y amazónicos. Se evidencia como la Cordillera de los Andes, probablemente por sus características en extensión y altitud, actuó como refugio en esos tiempos, permitiendo la preservación de ciertos linajes y siendo un punto desde el cual pudieron dispersarse y diferenciarse diferentes grupos humanos.

Sub-linajes de Q-M848 que se diferenciaron en un periodo temporal posterior a ~12.800 cal AP: Q-MPB118 relaciona grupos humanos que se diferenciaron entre la región Centro-Oeste y Sureste de Brasil; Q-MPB016 evidencia los vínculos que existían entre grupos humanos andinos septentrionales y del noroeste amazónico, probablemente asociado a intereses en el intercambio de recursos y cultura. Q-SK281 se diferenció en los Andes Centrales y en la actualidad no se tienen registros de dispersiones externas a Perú. Q-Z35841 muestra una diferenciación en Perú y expansión hacia el Noroeste de Argentina donde se diferenció regionalmente en Q-B47. Q-CTS2731 con una diferenciación en México, expansión dentro de México y hacia Norteamérica. Q-Y27993/Q-Y27992 muestra vínculos entre linajes amazónicos Arawaks y mexicanos. Q-Z19357 vincula grupos humanos de los andes centrales con tierras bajas de Bolivia y Argentina y con comunidades nativas del centro-oeste de Brasil.

La mayor parte de registros genéticos en cuanto a sub-linajes nativos americanos masculinos que se disponen son aquellos derivados de Q-M848 que abarcan a ~12.800 cal AP y posteriores, por lo que gran parte de la diversidad genética actual presente en linajes nativos americanos masculinos estaría representada por los linajes que lograron subsistir y expandir su diversidad luego del YD. Lo que además podría sugerir una extinción y pérdida del acervo de linajes anteriores a ~12.800 cal AP.

Respecto a esto como se explicó en la sección 1.5.5 en el reciente trabajo de Prates y col. 2020, se ha observado un pico de intensidad arqueológica hace ~12.500 cal AP, este pico cae dentro de los limites superiores e inferiores de la mayoría de las dataciones los sub-linajes de Q-M848 encontrados en esta tesis. Esto podría representar mayormente a los grupos humanos que lograron subsistir a las adversidades ambientales ocurridas en el periodo YD. En el mismo trabajo se plantea que dada a la discontinuidad sustancial de evidencia cultural antes de 15.500 cal AP, si hubieran existido grupos humanos, se habrían extinguido. El marcador Q-Z780 podría formar parte de la extremadamente baja señal arqueológica antes de 15.500 cal AP, que logró subsistir pero casi se extinguió.

Considerar un declive poblacional abrupto en las poblaciones nativas americanas hace ~12.800 cal AP, trae más complejidades para comprender los procesos demográficos que subyacen a la distribución actual del haplogrupo Q-M242 en Euro-Asia y sus vínculos con los linajes americanos. Q-M242 podría haber presentado una amplia dispersión y flujo génico tanto América como en Asia desde hacen ~35.000-20.000 años, y luego del evento YD ciertos linajes no tuvieron éxito en América y si se encuentran, son muy poco frecuentes en la actualidad. Si en Medio Oriente no hubo un declive poblacional durante el YD sino que sirvió como refugio para los humanos [13-15], esto podría llevar a proponer, que los linajes del haplogrupo Q más antiguos (como Q-L275) podrían haberse preservado en la región de Medio Oriente y luego del YD se dispersó hasta obtener la diversidad existente hoy para el haplogrupo Q en Euro-Asia. Linajes más antiguos como Q-Z780 (y quizás Q-M346*) sufrieron un declive drástico en América luego del YD. Por lo que, pensar en un origen asiático para Q-M242 podría ser un error teniendo en cuenta la falta de datos

genéticos que existe actualmente para sub-linajes antiguos de Q-M242 (anteriores a 12.800 cal AP) en América.

Otro aspecto importante al estudiar el poblamiento inicial de América es que las costas terrestres en el Pleistoceno se inundaron en todo el mundo y los niveles globales del mar aumentaron de manera abrupta en ~120 m en el periodo aproximado de 14000 - 11500 años AP [1]. En el noroeste de América del Norte, existe un creciente interés por la arqueología prehistórica sumergida como una sub-disciplina incipiente fundamental para comprender las primeras ocupaciones costeras y las rutas migratorias [358]. Como es común, las revisiones de investigación sobre arqueología prehistórica marina sumergida, han abordado mayormente proyectos realizados en América del Norte [358-360]. En América del Sur, los sitios prehistóricos sumergidos tempranos son prácticamente desconocidos. Investigaciones realizadas recientemente en la costa del Pacífico de Chile central, han encontrado evidencias de hacen ~16,000 AP donde una parte significativa de territorios estaban expuestos debido a los niveles del mar más bajos y disponibles para la fauna terrestre. En el mismo sitio, se encontró una gran diversidad de megafauna terrestre extinta sumergida [361]. Estas evidencias sugieren condiciones paleoambientales que durante el Pleistoceno Terminal favorecieron la congregación de diversos recursos de mamíferos cazados por grupos humanos paleoindios alrededor de áreas productivas de tierras bajas como arroyos, lagunas, estuarios, llanuras fértiles y humedales, como se ha sugerido para el norte de Chile central [362].

Consideramos de gran importancia para el estudio del poblamiento americano la inclusión de los cambios ambientales que ocurrieron durante el YD, tanto en estudios arqueológicos como genéticos, así como la inclusión de la arqueología prehistórica sumergida. Las preguntas respecto al poblamiento inicial de América aún superan en número a las respuestas y se requiere de mayor estudio interdisciplinario, así como de avances tecnológicos que puedan revelarnos una mayor comprensión de nuestra historia antigua. Sin embargo, creemos que la secuenciación NGS de cromosoma Y es una herramienta importante en esta área. Futuros estudios que incluyan un mayor número de secuencias genómicas de cromosoma Y en regiones donde actualmente existen vacíos, así como un mayor número de sub-linajes nativos americanos más antiguos, ayudarían en la reconstrucción de la historia del poblamiento ancestral de América

6 CONCLUSIONES

Se obtuvo un árbol filogenético del haplogrupo Q-M242 a partir de 102 secuencias genómicas del cromosoma Y.

Se definieron 17 sub-haplogrupos dentro del árbol filogenético reconstruido para el haplogrupo Q. De estos, 13 sub-haplogrupos son específicos de nativos americanos y pertenecen a Q-Z780 y Q-M3. Dentro de Q-M3, para 17 ramas las relaciones filogenéticas no pueden resolverse y representan la gran variabilidad presente en linajes nativos americanos que todavía no pueden explicarse con los datos disponibles de secuencias.

Se aportaron 13 secuencias nuevas al haplogrupo Q, incrementando la variabilidad de 3 sub-haplogrupos Q-M346*, Q-Z780 y Q-M848 y validamos 72 nuevos SNPs para estos sub-haplogrupos. Siete de las nuevas secuencias ampliaron la resolución de 5 sub-haplogrupos: Q-M346*, Q-Z780, Q-B42, Q-Z5908, Q-Z5906. Las 6 secuencias nuevas restantes forman parte de las ramas no se pueden resolver dentro de Q-M3.

Se identificó un patrón de distribución de los linajes propios nativos americanos y una diferenciación regional característica para ciertos linajes:

- En México: Q-Y2816, Q-Z782, Q-Y26467, Q-CTS11330 (con expansión a Mesoamérica), Q-CTS2731 (con expansión a Norteamérica).
- En los Andes Septentrionales: Q-MPB139 (con expansión a los Andes Centrales), Q-MPB016 (con expansión alnoroeste amazónico).
- En los Andes Centrales: Q-SK281, Q-Z6658.
- En los Andes Centrales y el Noroeste Argentino: Q-Z5908, Q-Z5906, Q-B35, Q-GMP70, Q-Z5907, Q-Z35841, Q-Z19357 (con expansión alsudeste de Brasil), Q-B42, Q-Z35497 (con expansión a Gran Chaco y a la región matogrossense de Brasil).
- En el Noroeste argentino:Q-B46, Q-B47
- En el Centro-oeste y Sudeste de Brasil: Q-MPB118.
- En los Andes Centrales y el Centro Oeste Argentino: Q-GMP73

Se identificó una distribución amplia en Mesoamérica y Sudamérica de Q-Z780, Q-Z781, Q-CTS11357, Q-Y27992/Y27993 y para el caso de Q-YP937 una distribución amplia en Sudamérica.

Se estableció una relación de los sub-linajes definidos dentro de Q-M848 (Q-MPB118, Q-SK281, Q-MPB139, Q-B46, Q-Z35505 / Q-Z35497, Q-Z6658, Q-B42, Q-CTS2731, Q-CTS11357, Q-Y27993 / Q-Y27992, Q-Z19357, Q-MPB016, Q-Z5908, Q-Z35841 y Q-Z5906) y su datación con la información arqueológica, histórica y lingüística disponible.

Se propuso un modelo de poblamiento americano en base a las relaciones filogenéticas y a las dataciones de los sub-linajes nativos americanos encontrados. El sub-linaje Q-Z781 autóctono de

América, presenta una amplia distribución en Mesoamérica y Sudamérica y un tiempo de divergencia de ~19300 años, una diferenciación regional entre Los Andes Centrales y el Centro Oeste argentino para Q-GMP73 de ~18200 mil años, aportando soporte genético para un modelo de poblamiento de América del Sur anterior a 18000 años.

Se propone que la expansión en su diversidad y la topología en forma de estrella para Q-M848 se vio impulsada luego de ~12800 años por el evento ambiental del Younger Dryas. Lo contrario sucedió para Q-Z780, que con este evento sufrió una discontinuidad y baja sustancial. Q-M346* podría ser otro sub-linaje que sufrió una drástica reducción luego del Younger Dryas y por eso su baja frecuencia en América, pero se necesitan más estudios de este sub-linaje en América.

Este trabajo ha permitido integrar la información, ambiental, arqueológica, histórica y lingüística con los datos de secuencias NGS de cromosoma Y. Lo cual aporta un nuevo conocimiento sobre el modelo de poblamiento de América y constituye una herramienta importante para el estudio del mismo.

7 REFERENCIAS BIBLIOGRÁFICAS

[1] E. H. Mark A. Jobling, Matthew Hurles, Toomas Kivisild and Chris Tyler-Smith, *Human evolutionary genetics*, Second edition ed.: New York; London: Garland Science, ©2014., 2014.

[2] D. M. Snell and J. M. Turner, "Sex Chromosome Effects on Male–Female Differences in Mammals," *Current Biology,* vol. 28, pp. R1313-R1324, 2018.

[3] A. P. Arnold, L. A. Cassis, M. Eghbali, K. Reue, and K. Sandberg, "Sex hormones and sex chromosomes cause sex differences in the development of cardiovascular diseases," *Arteriosclerosis, thrombosis, and vascular biology,* vol. 37, pp. 746-756, 2017.

[4] M. A. Jobling and C. Tyler-Smith, "The human Y chromosome: an evolutionary marker comes of age," *Nature Reviews Genetics,* vol. 4, p. 598, 2003.

[5] F. R. Santos and C. Tyler-Smith, "Reading the human Y chromosome: the emerging DNA markers and human genetic history," *Braz. J. Genet,* pp. 665-670, 1996.

[6] M. A. Jobling and C. Tyler-Smith, "Human Y-chromosome variation in the genome-sequencing era," *Nature Reviews Genetics,* vol. 18, p. 485, 2017.

[7] H. Skaletsky, T. Kuroda-Kawaguchi, P. J. Minx, H. S. Cordum, L. Hillier, L. G. Brown*, et al.*, "The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes," *Nature,* vol. 423, p. 825, 2003.

[8] G. D. Poznik, B. M. Henn, M.-C. Yee, E. Sliwerska, G. M. Euskirchen, A. A. Lin*, et al.*, "Sequencing Y chromosomes resolves discrepancy in time to common ancestor of males versus females," *Science,* vol. 341, pp. 562-565, 2013.

[9] M. F. Hammer and S. L. Zegura, "The role of the Y chromosome in human evolutionary studies," *Evolutionary Anthropology: Issues, News, and Reviews: Issues, News, and Reviews,* vol. 5, pp. 116-134, 1996.

[10] A. Novelletto, "Y chromosome variation in Europe: Continental and local processes in the formation of the extant gene pool," *Annals of Human Biology,* vol. 34, pp. 139-172, 2007.

[11] P. de Knijff, "Messages through bottlenecks: on the combined use of slow and fast evolving polymorphic markers on the human Y chromosome," *The American Journal of Human Genetics,* vol. 67, pp. 1055-1061, 2000.

[12] M. Casanova, P. Leroy, C. Boucekkine, J. Weissenbach, C. Bishop, M. Fellous*, et al.*, "A human Y-linked DNA polymorphism and its potential for estimating genetic and evolutionary distance," *Science,* vol. 230, pp. 1403-1406, 1985.

[13] G. Lucotte and N. Ngo, "p49f, A highly polymorphic probe, that detects Taq1 RFLPs on the human Y chromosome," *Nucleic acids research,* vol. 13, p. 8285, 1985.

[14] M. A. Jobling, "A survey of long-range DNA polymorphisms on the human Y chromosome," *Human molecular genetics,* vol. 3, pp. 107-114, 1994.

[15] T. M. Karafet, F. L. Mendez, M. B. Meilerman, P. A. Underhill, S. L. Zegura, and M. F. Hammer, "New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree," *Genome research,* vol. 18, pp. 830-838, 2008.

[16] T. Zerjal, B. Dashnyam, A. Pandya, M. Kayser, L. Roewer, F. R. Santos*, et al.*, "Genetic relationships of Asians and Northern Europeans, revealed by Y-chromosomal DNA analysis," *American journal of human genetics,* vol. 60, p. 1174, 1997.

[17] Y. C. Consortium, "A nomenclature system for the tree of human Y-chromosomal binary haplogroups," *Genome research,* vol. 12, pp. 339-348, 2002.

[18] R. A. Rocca, G. Magoon, D. F. Reynolds, T. Krahn, V. O. Tilroe, P. M. O. den Velde Boots*, et al.*, "Discovery of Western European R1b1a2 Y chromosome variants in 1000 genomes project data: an online community approach," *PLoS One,* vol. 7, p. e41634, 2012.

[19] G. P. Consortium, "A map of human genome variation from population-scale sequencing," *Nature,* vol. 467, p. 1061, 2010.

[20] M. H. Larmuseau, A. Van Geystelen, M. Kayser, M. van Oven, and R. Decorte, "Towards a consensus Y-chromosomal phylogeny and Y-SNP set in forensics in the next-generation sequencing era," *Forensic Science International: Genetics,* vol. 15, pp. 39-42, 2015.

[21] W. Wei, Q. Ayub, Y. Chen, S. McCarthy, Y. Hou, I. Carbone*, et al.*, "A calibrated human Y-chromosomal phylogeny based on resequencing," *Genome research,* vol. 23, pp. 388-395, 2013.

[22] G. D. Poznik, Y. Xue, F. L. Mendez, T. F. Willems, A. Massaia, M. A. W. Sayres*, et al.*, "Punctuated bursts in human male demography inferred from 1,244 worldwide Y-chromosome sequences," *Nature genetics,* vol. 48, p. 593, 2016.

[23] M. Karmin, L. Saag, M. Vicente, M. A. W. Sayres, M. Järve, U. G. Talas*, et al.*, "A recent bottleneck of Y chromosome diversity coincides with a global change in culture," *Genome research,* vol. 25, pp. 459-466, 2015.

[24] J. F. Crow, "The origins, patterns and implications of human spontaneous mutation," *Nature Reviews Genetics,* vol. 1, p. 40, 2000.

[25] Y. Xue, Q. Wang, Q. Long, B. L. Ng, H. Swerdlow, J. Burton*, et al.*, "Human Y chromosome base-substitution mutation rate measured by direct sequencing in a deep-rooting pedigree," *Current Biology,* vol. 19, pp. 1453-1457, 2009.

[26] T. Kivisild, "The study of human Y chromosome variation through ancient DNA," *Human genetics,* vol. 136, pp. 529-546, 2017.

[27] Q. Fu, H. Li, P. Moorjani, F. Jay, S. M. Slepchenko, A. A. Bondarev*, et al.*, "Genome sequence of a 45,000-year-old modern human from western Siberia," *Nature,* vol. 514, p. 445, 2014.

[28] A. Helgason, A. W. Einarsson, V. B. Guðmundsdóttir, Á. Sigurðsson, E. D. Gunnarsdóttir, A. Jagadeesan*, et al.*, "The Y-chromosome point mutation rate in humans," *Nature genetics,* vol. 47, p. 453, 2015.

[29] P. Moorjani, Z. Gao, and M. Przeworski, "Human germline mutation and the erratic evolutionary clock," *PLoS biology,* vol. 14, p. e2000744, 2016.

[30] P. Hallast, C. Batini, D. Zadik, P. Maisano Delser, J. H. Wetton, E. Arroyo-Pardo*, et al.*, "The Y-chromosome tree bursts into leaf: 13,000 high-confidence SNPs covering the majority of known clades," *Mol Biol Evol,* vol. 32, pp. 661-73, Mar 2015.

[31] T. Pinotti, A. Bergström, M. Geppert, M. Bawn, D. Ohasi, W. Shi*, et al.*, "Y chromosome sequences reveal a short Beringian standstill, rapid expansion, and early population structure of Native American Founders," *Current Biology,* vol. 29, pp. 149-157. e3, 2019.

[32] S. Yan, C.-C. Wang, H.-X. Zheng, W. Wang, Z.-D. Qin, L.-H. Wei*, et al.*, "Y chromosomes of 40% Chinese descend from three Neolithic super-grandfathers," *PloS one,* vol. 9, p. e105691, 2014.

[33] S. Lippold, H. Xu, A. Ko, M. Li, G. Renaud, A. Butthof*, et al.*, "Human paternal and maternal demographic histories: insights from high-resolution Y chromosome and mtDNA sequences," *Investigative genetics,* vol. 5, p. 13, 2014.

[34] P. Francalacci, L. Morelli, A. Angius, R. Berutti, F. Reinier, R. Atzeni*, et al.*, "Low-pass DNA sequencing of 1200 Sardinians reconstructs European Y-chromosome phylogeny," *Science,* vol. 341, pp. 565-569, 2013.

[35] G. R. Magoon, R. H. Banks, C. Rottensteiner, B. E. Schrack, V. O. Tilroe, T. Robb*, et al.*, "Generation of high-resolution a priori Y-chromosome phylogenies using “next-generation” sequencing data," *bioRxiv,* p. 000802, 2013.

[36] P. Hallast, C. Batini, D. Zadik, P. Maisano Delser, J. H. Wetton, E. Arroyo-Pardo*, et al.*, "The Y-chromosome tree bursts into leaf: 13,000 high-confidence SNPs covering the majority of known clades," *Molecular biology and evolution,* vol. 32, pp. 661-673, 2014.

[37] R. Scozzari, A. Massaia, B. Trombetta, G. Bellusci, N. M. Myres, A. Novelletto*, et al.*, "An unbiased resource of novel SNP markers provides a new chronology for the human Y chromosome and reveals a deep phylogenetic structure in Africa," *Genome research,* vol. 24, pp. 535-544, 2014.

[38] V. Grugni, A. Raveane, L. Ongaro, V. Battaglia, B. Trombetta, G. Colombo*, et al.*, "Analysis of the human Y-chromosome haplogroup Q characterizes ancient population movements in Eurasia and the Americas," *BMC biology,* vol. 17, p. 3, 2019.

[39] M.-C. Bortolini, F. M. Salzano, M. G. Thomas, S. Stuart, S. P. Nasanen, C. H. Bau*, et al.*, "Y-chromosome evidence for differing ancient demographic histories in the Americas," *The American Journal of Human Genetics,* vol. 73, pp. 524-539, 2003.

[40] T. M. Karafet, L. P. Osipova, M. A. Gubina, O. L. Posukh, S. L. Zegura, and M. F. Hammer, "High levels of Y-chromosome differentiation among native Siberian populations and the genetic signature of a boreal hunter-gatherer way of life," *Human biology,* pp. 761-789, 2002.

[41] M. Geppert, Q. Ayub, Y. Xue, S. Santos, Â. Ribeiro-dos-Santos, M. Baeta*, et al.*, "Identification of new SNPs in native South American populations by resequencing the Y chromosome," *Forensic Science International: Genetics,* vol. 15, pp. 111-114, 2015.

[42] B. Malyarchuk, M. Derenko, G. Denisova, A. Maksimov, M. Wozniak, T. Grzybowski*, et al.*, "Ancient links between Siberians and Native Americans revealed by subtyping the Y chromosome haplogroup Q1a," *Journal of human genetics,* vol. 56, p. 583, 2011.

[43] J. Di Cristofaro, E. Pennarun, S. Mazières, N. M. Myres, A. A. Lin, S. A. Temori*, et al.*, "Afghan Hindu Kush: where Eurasian sub-continent gene flows converge," *PloS one,* vol. 8, p. e76748, 2013.

[44] G. Bailliet, V. Ramallo, M. Muzzio, A. García, M. R. Santos, E. L. Alfaro*, et al.*, "Brief communication: Restricted geographic distribution for Y-Q* paragroup in South America," *American Journal of Physical Anthropology: The Official Publication of the American Association of Physical Anthropologists,* vol. 140, pp. 578-582, 2009.

[45] M. Regueiro, J. Alvarez, D. Rowold, and R. J. Herrera, "On the origins, rapid expansion and genetic diversity of native Americans from hunting-gatherers to agriculturalists," *American journal of physical anthropology,* vol. 150, pp. 333-348, 2013.

[46] M. C. Dulik, A. C. Owings, J. B. Gaieski, M. G. Vilar, A. Andre, C. Lennie*, et al.*, "Y-chromosome analysis reveals genetic divergence and new founding native lineages in Athapaskan-and Eskimoan-speaking populations," *Proceedings of the National Academy of Sciences,* vol. 109, pp. 8471-8476, 2012.

[47] V. Battaglia, V. Grugni, U. A. Perego, N. Angerhofer, J. E. Gomez-Palmieri, S. R. Woodward*, et al.*, "The first peopling of South America: new evidence from Y-chromosome haplogroup Q," *PLoS One,* vol. 8, p. e71390, 2013.

[48] L. Roewer, M. Nothnagel, L. Gusmão, V. Gomes, M. González, D. Corach*, et al.*, "Continent-wide decoupling of Y-chromosomal genetic variation from language and geography in native South Americans," *PLoS Genet,* vol. 9, p. e1003460, 2013.

[49] L. S. Jurado Medina, P. B. Paz Sepulveda, V. Ramallo, C. Sala, J. Beltramo, M. Schwab*, et al.*, "Continental origin for Q Haplogroup Patrilineages in Argentina and Paraguay," *manuscrito en prensa en Human Biology,* 2020.

[50] D. R. Carvalho-Silva, F. R. Santos, J. Rocha, and S. D. Pena, "The phylogeography of Brazilian Y-chromosome lineages," *The American Journal of Human Genetics,* vol. 68, pp. 281-286, 2001.

[51] V. Ramallo, J. Mucci, A. García, M. Muzzio, J. Motti, M. Santos*, et al.*, "Comparison of Y-chromosome haplogroup frequencies in eight Provinces of Argentina," *Forensic Science International: Genetics Supplement Series,* vol. 2, pp. 431-432, 2009.

[52] E. J. Parra, A. Marcini, J. Akey, J. Martinson, M. A. Batzer, R. Cooper*, et al.*, "Estimating African American admixture proportions by use of population-specific alleles," *The American Journal of Human Genetics,* vol. 63, pp. 1839-1851, 1998.

[53] S. Wang, N. Ray, W. Rojas, M. V. Parra, G. Bedoya, C. Gallo*, et al.*, "Geographic patterns of genome admixture in Latin American Mestizos," *PLoS genetics,* vol. 4, p. e1000037, 2008.

[54] L. S. J. Medina, V. Ramallo, H. Calandra, G. Lamenza, J. Braunstein, S. Salceda*, et al.*, "Gran Chaco paternal lineages, a DNA approach."

[55] ISOGG. Available: https://isogg.org/

[56] ISOGG. *2019-2020 Haplogroup Q Tree*. Available: https://docs.google.com/spreadsheets/d/1bcVNnQ5y4tkY5NL4SuxSTO4ofR1ymh_1Joc9DgCYnoY/edit#gid=1268900795

[57] (2019-2020). *ISOGG index SNP*. Available: https://docs.google.com/spreadsheets/d/1UY26FvLE3UmEmYFiXgOy0uezJi_wOut-V5TD0a_6-bE/edit#gid=1934392066

[58] S. L. Zegura, T. M. Karafet, L. A. Zhivotovsky, and M. F. Hammer, "High-resolution SNPs and microsatellite haplotypes point to a single, recent entry of Native American Y chromosomes into the Americas," *Molecular biology and evolution,* vol. 21, pp. 164-175, 2004.

[59] S. Abel and H. Schroeder, "From country marks to DNA markers: the genomic turn in the reconstruction of African identities," *Current Anthropology,* vol. 61, pp. S000-S000, 2020.

[60] IPCB. (1999). *Indigenous Peoples Council on Biocolonialism*. Available: http://www.ipcb.org/

[61] J.-F. Morin, "Une réplique du Sud à l'extension du droit des brevets: la biodiversité dans le régime international de la propriété intellectuelle," *Droit et société,* pp. 633-653, 2004.

[62] P. D. Hughes and P. L. Gibbard, "A stratigraphical basis for the Last Glacial Maximum (LGM)," *Quaternary International,* vol. 383, pp. 174-185, 2015.

[63] K. Lambeck, T. M. Esat, and E.-K. Potter, "Links between climate and sea levels for the past three million years," *Nature,* vol. 419, p. 199, 2002.

[64] S. Blockley and R. Pinhasi, "A revised chronology for the adoption of agriculture in the Southern Levant and the role of Lateglacial climatic change," *Quaternary Science Reviews,* vol. 30, pp. 98-108, 2011.

[65] R. B. Firestone, A. West, J. P. Kennett, L. Becker, T. E. Bunch, Z. S. Revay*, et al.*, "Evidence for an extraterrestrial impact 12,900 years ago that contributed to the megafaunal extinctions and the Younger Dryas cooling," *Proceedings of the National Academy of Sciences,* vol. 104, pp. 16016-16021, 2007.

[66] M. Pino, A. M. Abarzúa, G. Astorga, A. Martel-Cea, N. Cossio-Montecinos, R. X. Navarro*, et al.*, "Sedimentary record from Patagonia, southern Chile supports cosmic-impact triggering of biomass burning, climate change, and megafaunal extinctions at 12.8 ka," *Scientific reports,* vol. 9, pp. 1-27, 2019.

[67] C. R. Kinzie, S. S. Que Hee, A. Stich, K. A. Tague, C. Mercer, J. J. Razink*, et al.*, "Nanodiamond-rich layer across three continents consistent with major cosmic impact at 12,800 cal BP," *The Journal of Geology,* vol. 122, pp. 475-506, 2014.

[68] W. C. Mahaney, D. Krinsley, and V. Kalm, "Evidence for a cosmogenic origin of fired glaciofluvial beds in the northwestern Andes: Correlation with experimentally heated quartz and feldspar," *Sedimentary Geology,* vol. 231, pp. 31-40, 2010.

[69] W. Mahaney, D. H. Krinsley, M. W. Milner, R. Fischer, and K. Langworthy, "Did the Black-Mat Impact/Airburst Reach the Antarctic? Evidence from New Mountain Near the Taylor Glacier in the Dry Valley Mountains," *The Journal of Geology,* vol. 126, pp. 285-305, 2018.

[70] K. Squires, D. Errickson, and N. Márquez-Grant, *Ethical Approaches to Human Remains: A Global Challenge in Bioarchaeology and Forensic Anthropology*: Springer Nature, 2019.

[71] R. García-Mancuso, M. Plischuk, B. Desántolo, G. Garizoain, and M. L. Sardi, "Ethical Considerations in Human Remains Based Research in Argentina," in *Ethical Approaches to Human Remains*, ed: Springer, 2019, pp. 447-463.

[72] R. L. Jantz and D. W. Owsley, "Variation among early North American crania," *American Journal of Physical Anthropology: The Official Publication of the American Association of Physical Anthropologists,* vol. 114, pp. 146-155, 2001.

[73] M. Rasmussen, S. L. Anzick, M. R. Waters, P. Skoglund, M. DeGiorgio, T. W. Stafford Jr*, et al.*, "The genome of a Late Pleistocene human from a Clovis burial site in western Montana," *Nature,* vol. 506, p. 225, 2014.

[74] J. A. Raff and D. A. Bolnick, "Palaeogenomics: genetic roots of the first Americans," *Nature,* vol. 506, pp. 162-163, 2014.

[75] J. Feathers, R. Kipnis, L. Piló, M. Arroyo-Kalin, and D. Coblentz, "How old is Luzia? Luminescence dating and stratigraphic integrity at Lapa Vermelha, Lagoa Santa, Brazil," *Geoarchaeology,* vol. 25, pp. 395-436, 2010.

[76] M. Rasmussen, Y. Li, S. Lindgreen, J. S. Pedersen, A. Albrechtsen, I. Moltke*, et al.*, "Ancient human genome sequence of an extinct Palaeo-Eskimo," *Nature,* vol. 463, p. 757, 2010.

[77] T. J. Green, B. Cochran, T. W. Fenton, J. C. Woods, G. L. Titmus, L. Tieszen*, et al.*, "The Buhl burial: a Paleoindian woman from southern Idaho," *American Antiquity,* vol. 63, pp. 437-456, 1998.

[78] U. S. C. Senate., "Native American Graves Protection and Repatriation Act," in *Committee on Indian Affairs (1993-)*, ed, 2005.

[79] M. Rasmussen, M. Sikora, A. Albrechtsen, T. S. Korneliussen, J. V. Moreno-Mayar, G. D. Poznik*, et al.*, "The ancestry and affiliations of Kennewick Man," *Nature,* vol. 523, p. 455, 2015.

[80] B. M. Kemp, R. S. Malhi, J. McDonough, D. A. Bolnick, J. A. Eshleman, O. Rickards*, et al.*, "Genetic analysis of early holocene skeletal remains from Alaska and its implications for the settlement of the Americas," *American Journal of Physical Anthropology: The Official Publication of the American Association of Physical Anthropologists,* vol. 132, pp. 605-621, 2007.

[81] L. Prates, G. G. Politis, and S. I. Perez, "Rapid radiation of humans in South America after the last glacial maximum: A radiocarbon-based study," *PloS one,* vol. 15, p. e0236023, 2020.

[82] R. E. Taylor, C. V. Haynes, and M. Stuiver, "Clovis and Folsom age estimates: stratigraphic context and radiocarbon calibration," *Antiquity,* vol. 70, pp. 515-525, 1996.

[83] T. Goebel, M. R. Waters, and D. H. O'Rourke, "The late Pleistocene dispersal of modern humans in the Americas," *science,* vol. 319, pp. 1497-1502, 2008.

[84] M. Michab, J. K. Feathers, J.-L. Joron, N. Mercier, M. Selo, H. Valladas*, et al.*, "Luminescence dates for the Paleoindian site of Pedra Pintada, Brazil," *Quaternary science reviews,* vol. 17, pp. 1041-1046, 1998.

[85] T. D. Dillehay, S. Goodbred, M. Pino, V. F. V. Sánchez, T. R. Tham, J. Adovasio*, et al.*, "Simple technologies and diverse food strategies of the Late Pleistocene and Early Holocene at Huaca Prieta, Coastal Peru," *Science Advances,* vol. 3, p. e1602778, 2017.

[86] G. G. Politis, M. A. Gutiérrez, D. J. Rafuse, and A. Blasi, "The arrival of Homo sapiens into the Southern Cone at 14,000 years ago," *PLoS One,* vol. 11, p. e0162870, 2016.

[87] T. Dillehay, "Monte Verde: A Late Pleistocene Settlement in Chile, Volume II: The Archaeological Context," *Smithsonian Institution, Washington, DC,* 1997.

[88] K. Moreno, J. E. Bostelmann, C. Macías, X. Navarro-Harris, R. De Pol-Holz, and M. Pino, "A late Pleistocene human footprint from the Pilauco archaeological site, northern Patagonia, Chile," *PloS one,* vol. 14, p. e0213572, 2019.

[89] N. Guidon and G. Delibrias, "Carbon-14 dates point to man in the Americas 32,000 years ago," *Nature,* vol. 321, pp. 769-771, 1986.

[90] E. Boëda, I. Clemente-Conte, M. Fontugne, C. Lahaye, M. Pino, G. D. Felice*, et al.*, "A new late Pleistocene archaeological sequence in South America: The Vale da Pedra Furada (Piauí, Brazil)," *Antiquity,* vol. 88, pp. 927-955, 2014.

[91] E. Boëda, R. Rocca, A. Da Costa, M. Fontugne, C. Hatté, I. Clemente-Conte*, et al.*, "New data on a pleistocene archaeological sequence in South America: Toca do Sítio do Meio, Piauí, Brazil," *PaleoAmerica,* vol. 2, pp. 286-302, 2016.

[92] C. Lahaye, M. Hernandez, E. Boëda, G. D. Felice, N. Guidon, S. Hoeltz*, et al.*, "Human occupation in South America by 20,000 BC: the Toca da Tira Peia site, Piauí, Brazil," *Journal of Archaeological Science,* vol. 40, pp. 2840-47, 2013.

[93] D. Vialou, M. Benabdelhadi, J. Feathers, M. Fontugne, and A. V. Vialou, "Peopling South America's centre: the late Pleistocene site of Santa Elina," *antiquity,* vol. 91, pp. 865-884, 2017.

[94] R. A. Fariña, "Bone surface modifications, reasonable certainty, and human antiquity in the Americas: The case of the Arroyo del Vizcaíno Site," *American Antiquity,* pp. 193-200, 2015.

[95] C. F. Ardelean, L. Becerra-Valdivia, M. W. Pedersen, J.-L. Schwenninger, C. G. Oviatt, J. I. Macías-Quintero*, et al.*, "Evidence of human occupation in Mexico around the Last Glacial Maximum," *Nature,* vol. 584, pp. 87-92, 2020.

[96] L. Campbell, *American Indian Languages. The Historical Linguistics of Native America*. 198 Madison Avenue, New York, New York 10016: First published in 1997 by Oxford University Press, Inc., 1997.

[97] T. Flannery, *The eternal frontier: an ecological history of North America and its peoples*: Grove Press, 2002.

[98] S. J. Fiedel, "The Anzick genome proves Clovis is first, after all," *Quaternary International,* vol. 444, pp. 4-9, 2017.

[99] J. Bird, "Antiquity and migrations of the early inhabitants of Patagonia," *Geographical Review,* vol. 28, pp. 250-275, 1938.

[100] M. Massone, "Los cazadores paleoindios de Tres Arroyos (Tierra del Fuego)," in *Anales del Instituto de la Patagonia*, 1987.

[101] G. Politis, L. Prates, and S. I. Perez, "Early Asiatic migration to the Americas: a view from South America," in *Mobility and Ancient Society in Asia and the Americas*, ed: Springer, 2015, pp. 89-102.

[102] R. Nielsen, J. M. Akey, M. Jakobsson, J. K. Pritchard, S. Tishkoff, and E. Willerslev, "Tracing the peopling of the world through genomics," *Nature,* vol. 541, pp. 302-310, 2017.
[103] S. R. Holen, T. A. Deméré, D. C. Fisher, R. Fullagar, J. B. Paces, G. T. Jefferson*, et al.*, "A 130,000-year-old archaeological site in southern California, USA," *Nature,* vol. 544, pp. 479-483, 2017.
[104] S. Miller, D. Dykes, and H. Polesky, "A simple salting out procedure for extracting DNA from human nucleated cells," *Nucleic acids research,* vol. 16, p. 1215, 1988.
[105] K. Umetsu, M. Tanaka, I. Yuasa, N. Adachi, A. Miyoshi, S. Kashimura*, et al.*, "Multiplex amplified product-length polymorphism analysis of 36 mitochondrial single-nucleotide polymorphisms for haplogrouping of East Asian populations," *Electrophoresis,* vol. 26, pp. 91-98, 2005.
[106] K. Umetsu, M. Tanaka, I. Yuasa, N. Saitou, T. Takeyasu, N. Fuku*, et al.*, "Multiplex amplified product-length polymorphism analysis for rapid detection of human mitochondrial DNA variations," *Electrophoresis,* vol. 22, pp. 3533-3538, 2001.
[107] L. S. J. Medina, M. Muzzio, M. Schwab, M. L. B. Costantino, G. Barreto, and G. Bailliet, "Human Y-chromosome SNP characterization by multiplex amplified product-length polymorphism analysis," *Electrophoresis,* vol. 35, pp. 2524-2527, 2014.
[108] L. S. Jurado Medina, "Tipificación de marcadores uniparentales en poblaciones mestizas de Argentina," tesis doctoral, Universidad Nacional de La Plata

Facultad de Ciencias Naturales y Museo, 2015.
[109] L. S. Jurado Medina, V. Ramallo, H. Calandra, G. Lamenza, J. Braunstein, S. Salceda*, et al.*, "Linajes paternos del Gran Chaco, un abordaje desde el ADN," *Folia Histórica del Nordeste,* 2014.
[110] *Full Genomes Corporation*. Available: www.fullgenomes.com
[111] I. hiseq-4000. Available: https://www.illumina.com/systems/sequencing-platforms/hiseq-3000-4000.html
[112] T. Nakazato, T. Ohta, and H. Bono, "Experimental design-based functional mining and characterization of high-throughput sequencing data in the sequence read archive," *PLoS One,* vol. 8, p. e77910, 2013.
[113] Illumina. Available: http://www.illumina.com
[114] P. J. Cock, C. J. Fields, N. Goto, M. L. Heuer, and P. M. Rice, "The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants," *Nucleic acids research,* vol. 38, pp. 1767-1771, 2010.
[115] *GATK (Genome Analysis Toolkit)*. Available: https://gatk.broadinstitute.org/hc/en-us
[116] A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky*, et al.*, "The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data," *Genome research,* vol. 20, pp. 1297-1303, 2010.
[117] S. Andrews, "FastQC-A Quality Control application for FastQ files," ed, 2010.
[118] S. A. y. col. *FASTQC software*. Available: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
[119] R. Schmieder and R. Edwards, "Quality control and preprocessing of metagenomic datasets," *Bioinformatics,* vol. 27, pp. 863-864, 2011.
[120] *PRINSEQ software*. Available: https://sourceforge.net/projects/prinseq/
[121] M. G. Di Giglio, M. Muttenthaler, K. Harpsøe, Z. Liutkeviciute, P. Keov, T. Eder*, et al.*, "Development of a human vasopressin V 1a-receptor antagonist from an evolutionary-related insect neuropeptide," *Scientific reports,* vol. 7, p. 41002, 2017.

[122] T. Eri, K. Kawahata, T. Kanzaki, M. Imamura, K. Michishita, L. Akahira*, et al.*, "Intestinal microbiota link lymphopenia to murine autoimmunity via PD-1+ CXCR5−/dim B-helper T cell induction," *Scientific reports,* vol. 7, p. 46037, 2017.
[123] *Genome Reference Consortium Human Build 37 (GRCh37)*. Available: https://www.ncbi.nlm.nih.gov/genome/guide/human/
[124] H. Li and R. Durbin, "Fast and accurate short read alignment with Burrows–Wheeler transform," *bioinformatics,* vol. 25, pp. 1754-1760, 2009.
[125] H. Li and R. Durbin, "Fast and accurate long-read alignment with Burrows–Wheeler transform," *Bioinformatics,* vol. 26, pp. 589-595, 2010.
[126] T. S. B. F. S. W. Group. (31 Dec 2019). *Sequence Alignment/Map Format Specification*. Available: https://samtools.github.io/hts-specs/SAMv1.pdf
[127] *Picard*. Available: https://broadinstitute.github.io/picard/
[128] *Samtools*. Available: https://samtools.github.io/hts-specs/SAMv1.pdf
[129] *Samtools Manual*. Available: http://www.htslib.org/doc/1.6/samtools.html
[130] G. P. Consortium, "A global reference for human genetic variation," *Nature,* vol. 526, p. 68, 2015.
[131] F. F. T. Team, "The Variant Call Format (VCF) Version 4.2 Specification. Available at ht tps," *github. com/samtools/hts-specs,* 2015.
[132] M. A. DePristo, E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire, C. Hartl*, et al.*, "A framework for variation discovery and genotyping using next-generation DNA sequencing data," *Nature genetics,* vol. 43, p. 491, 2011.
[133] G. A. Van der Auwera, M. O. Carneiro, C. Hartl, R. Poplin, G. Del Angel, A. Levy-Moonshine*, et al.*, "From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline," *Current protocols in bioinformatics,* vol. 43, pp. 11.10. 1-11.10. 33, 2013.
[134] *Phred-scaled-quality-scores*. Available: https://gatk.broadinstitute.org/hc/en-us/articles/360035531872-Phred-scaled-quality-scores
[135] S. Mallick, H. Li, M. Lipson, I. Mathieson, M. Gymrek, F. Racimo*, et al.*, "The Simons genome diversity project: 300 genomes from 142 diverse populations," *Nature,* vol. 538, p. 201, 2016.
[136] J. N. A. S. K. D. K. R. Q. N. F. T. D. group. Available: http://www.familytreedna.com/groups/qnordic/about
[137] N. T. Weeks and G. R. Luecke, "Optimization of SAMtools sorting using OpenMP tasks," *Cluster Computing,* vol. 20, pp. 1869-1880, 2017.
[138] *VCFtools manual*. Available: http://vcftools.sourceforge.net/man_latest.html
[139] D. E. Cook and E. C. Andersen, "VCF-kit: assorted utilities for the variant call format," *Bioinformatics,* vol. 33, pp. 1581-1582, 2017.
[140] *VCK-kit*. Available: https://vcf-kit.readthedocs.io/en/latest/
[141] V. Grugni, V. Battaglia, U. A. Perego, A. Raveane, H. Lancioni, A. Olivieri*, et al.*, "Exploring the Y chromosomal ancestry of modern panamanians," *PloS one,* vol. 10, p. e0144223, 2015.
[142] A. Stamatakis, "RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies," *Bioinformatics,* vol. 30, pp. 1312-1313, 2014.
[143] *RaxML 8.2 Manual*. Available: https://cme.h-its.org/exelixis/resource/download/NewManual.pdf
[144] A. Rambaut, "FigTree, a graphical viewer of phylogenetic trees," 2007.

[145] P. Forster, R. Harding, A. Torroni, and H.-J. Bandelt, "Origin and evolution of Native American mtDNA variation: a reappraisal," *American journal of human genetics,* vol. 59, p. 935, 1996.

[146] D. H. Parks, T. Mankowski, S. Zangooei, M. S. Porter, D. G. Armanini, D. J. Baird*, et al.*, "GenGIS 2: Geospatial analysis of traditional and genetic biodiversity, with new gradient algorithms and an extensible plugin framework," *PloS one,* vol. 8, 2013.

[147] D. H. Parks, M. Porter, S. Churcher, S. Wang, C. Blouin, J. Whalley*, et al.*, "GenGIS: A geospatial information system for genomic data," *Genome research,* vol. 19, pp. 1896-1904, 2009.

[148] G. Matthijs, E. Souche, M. Alders, A. Corveleyn, S. Eck, I. Feenstra*, et al.*, "Guidelines for diagnostic next-generation sequencing," *European Journal of Human Genetics,* vol. 24, pp. 2-5, 2016.

[149] W. Mu, H.-M. Lu, J. Chen, S. Li, and A. M. Elliott, "Sanger confirmation is required to achieve optimal sensitivity and specificity in next-generation sequencing panel testing," *The Journal of molecular diagnostics,* vol. 18, pp. 923-932, 2016.

[150] T. F. Beck, J. C. Mullikin, and N. C. S. P. B. L. G. l. m. n. gov, "Systematic evaluation of Sanger validation of next-generation sequencing variants," *Clinical chemistry,* vol. 62, pp. 647-654, 2016.

[151] R. De Cario, A. Kura, S. Suraci, A. Magi, A. Volta, R. Marcucci*, et al.*, "SANGER VALIDATION OF HIGH-THROUGHPUT SEQUENCING IN GENETIC DIAGNOSIS: STILL THE BEST PRACTICE?," *Frontiers in Genetics,* vol. 11, p. 1496, 2020.

[152] *Primer3*. Available: http://bioinfo.ut.ee/primer3-0.4.0/

[153] IDT. *Oligoanalyzer*. Available: https://www.idtdna.com/pages/tools/oligoanalyzer?utm_source=google&utm_medium=cpc&utm_campaign=ga_oligoanalyzer&utm_content=ad_group_oligo_analyzer&gclid=Cj0KCQjwwuD7BRDBARIsAK_5YhUCUENqH-anzBNEZS_InuWLUvfUGPeDEBbGaf4wAfTzXuCDk_l6LicaAuLvEALw_wcB

[154] Chromas. Available: https://technelysium.com.au/wp/chromas/

[155] I. s. Indigo. Available: https://www.gear-genomics.com/indigo/

[156] BLAST. Available: https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome

[157] A. J. Drummond and A. Rambaut, "BEAST: Bayesian evolutionary analysis by sampling trees," *BMC evolutionary biology,* vol. 7, pp. 1-8, 2007.

[158] W.-H. Shou, E.-F. Qiao, C.-Y. Wei, Y.-L. Dong, S.-J. Tan, H. Shi*, et al.*, "Y-chromosome distributions among populations in Northwest China identify significant contribution from Central Asian pastoralists and lesser influence of western Eurasians," *Journal of human genetics,* vol. 55, pp. 314-322, 2010.

[159] C.-C. Wang and H. Li, "Inferring human history in East Asia from Y chromosomes," *Investigative genetics,* vol. 4, p. 11, 2013.

[160] J. A. Trejaut, E. S. Poloni, J.-C. Yen, Y.-H. Lai, J.-H. Loo, C.-L. Lee*, et al.*, "Taiwan Y-chromosomal DNA variation and its relationship with Island Southeast Asia," *BMC genetics,* vol. 15, p. 77, 2014.

[161] T. M. Karafet, B. Hallmark, M. P. Cox, H. Sudoyo, S. Downey, J. S. Lansing*, et al.*, "Major east–west division underlies Y chromosome stratification across Indonesia," *Molecular biology and evolution,* vol. 27, pp. 1833-1844, 2010.

[162] O. Balanovsky, V. Gurianov, V. Zaporozhchenko, O. Balaganskaya, V. Urasin, M. Zhabagin*, et al.*, "Phylogeography of human Y-chromosome haplogroup Q3-L275 from an academic/citizen science collaboration," *BMC evolutionary biology,* vol. 17, p. 18, 2017.
[163] J. K. Olofsson, V. Pereira, C. Børsting, and N. Morling, "Peopling of the North Circumpolar Region–insights from Y chromosome STR and SNP typing of Greenlanders," *PloS one,* vol. 10, p. e0116573, 2015.
[164] P. Flegontov, N. E. Altinisik, P. Changmai, E. J. Vajda, J. Krause, and S. Schiffels, "Na-Dene populations descend from the Paleo-Eskimo migration into America," *bioRxiv,* p. 074476, 2016.
[165] A. García, M. Pauro, G. Bailliet, C. M. Bravi, and D. A. Demarchi, "Genetic variation in populations from central Argentina based on mitochondrial and Y chromosome DNA evidence," *Journal of Human Genetics,* vol. 63, pp. 493-507, 2018.
[166] J. R. Sandoval, D. R. Lacerda, M. S. Jota, A. Salazar-Granara, P. P. R. Vieira, O. Acosta*, et al.*, "The genetic history of indigenous populations of the Peruvian and Bolivian Altiplano: the legacy of the Uros," *PLoS One,* vol. 8, 2013.
[167] M. S. Jota, D. R. Lacerda, J. R. Sandoval, P. P. R. Vieira, D. Ohasi, J. E. Santos-Júnior*, et al.*, "New native South American Y chromosome lineages," *Journal of human genetics,* vol. 61, p. 593, 2016.
[168] S. J. Fiedel, "Origins of the first Americans: Before and after the Anzick genome," *Reviews in Anthropology,* vol. 46, pp. 164-179, 2017.
[169] YFull. *Q-Y2816*. Available: https://www.yfull.com/tree/Q-Y2816/
[170] J. Norstedt, "Q Nordic project," 2016.
[171] S. A. Avena, M. L. Parolin, C. B. Dejean, M. C. R. Part, G. Fabrikant, A. S. Goicoechea*, et al.*, "Mezcla génica y linajes uniparentales en Comodoro Rivadavia (provincia de Chubut, Argentina)," *Revista Argentina de antropología biológica,* vol. 11, pp. 25-41, 2009.
[172] R. Bisso-Machado, M. S. Jota, V. Ramallo, V. R. Paixão-Côrtes, D. R. Lacerda, F. M. Salzano*, et al.*, "Distribution of Y-chromosome q lineages in native americans," *American Journal of Human Biology,* vol. 23, pp. 563-566, 2011.
[173] V. Ramallo, M. Muzzio, M. R. Santos, J. M. B. Motti, L. S. Jurado Medina, C. M. Bravi*, et al.*, "Native male founder lineages of America," 2011.
[174] P. A. Underhill, L. Jin, R. Zemans, P. J. Oefner, and L. L. Cavalli-Sforza, "A pre-Columbian Y chromosome-specific transition and its implications for human evolutionary history," *Proceedings of the national Academy of sciences,* vol. 93, pp. 196-200, 1996.
[175] B. Malyarchuk, M. Derenko, G. Denisova, A. Maksimov, M. Wozniak, T. Grzybowski*, et al.*, "Ancient links between Siberians and Native Americans revealed by subtyping the Y chromosome haplogroup Q1a," *Journal of human genetics,* vol. 56, pp. 583-588, 2011.
[176] *YFull Q-Y4308/Q-Y4276*. Available: https://www.yfull.com/tree/Q-Y4276/
[177] Available: https://haplogroup.org/native-american-q-m3-tree-p2-q-m242-news-6-nov-2016/
[178] R. M. Seidemann, "Time for a Change-The Kennewick Man Case and Its Implications for the Future of the Native American Graves Protection and Repatriation Act," *W. Va. L. Rev.,* vol. 106, p. 149, 2003.
[179] YFull. Available: https://www.yfull.com/
[180] YFull, "Q-MPB122," 2018.
[181] L. C. Carvalho, "Os Arana e sua indianidade: disputas internas por legitimidade e o reconhecimento oficial como grupo indígena," tesis, Belo Horizonte 2008.
[182] "Centro de Documentação Eloy Ferreira da Silva (CEDEFES). Aranã: a luta de um povo no Vale do Jequitinhonha. Contagem: CEDEFES; 2003.."

[183] L. S. d. M. Cardoso, A. C. L. Queiroz, J. L. Pena, G. L. L. Machado-Coelho, and L. Heller, "Aranãs do médio Jequitinhonha: aspectos socioeconômicos, demográficos e sanitários de uma população indígena desaldeada," *Ciência & Saúde Coletiva,* vol. 21, pp. 3859-3870, 2016.

[184] J. V. Neel, F. M. Salzano, P. C. Junqueira, F. Keiter, and D. Maybury-Lewis, "Studies on the Xavante indians of the Brazilian Mato Grosso," *American journal of human genetics,* vol. 16, p. 52, 1964.

[185] B. Ricardo and F. Ricardo, *Povos indígenas no Brasil: 2001/2005*: Instituto Socioambiental, 2006.

[186] L. Graham. (2008). *Povos Indigenas no Brasil (Equipo de Edición de la Enciclopédia Pueblos Inginenas en Brasil ed.)*. Available: https://pib.socioambiental.org/es/Povo:Xavante

[187] P. I. Schmitz, "Prehistoric hunters and gatherers of Brazil," *Journal of World Prehistory,* vol. 1, pp. 53-126, 1987.

[188] E. T. Miller, "Pesquisas arqueológicas paleoindígenas no Brasil Ocidental," *Estudios Atacameños,* pp. 37-61, 1987.

[189] A. Vilhena-Vialou and D. Vialou, "Les premiers peuplements préhistoriques du Mato Grosso," *Bulletin de la Société préhistorique française,* vol. 91, pp. 257-263, 1994.

[190] J. E. de Oliveira and S. A. Viana, "O centro-oeste antes de Cabral," *Revista Usp,* pp. 142-189, 1999.

[191] C. Chauchat and J. P. Lacombe, "El hombre de Paiján:¿ el más antiguo peruano?," *Gaceta Arqueológica Andina,* vol. 11, pp. 4-6, 1984.

[192] B. Bauer, "Cuzco Antiguo. Tierra Natal de los Incas. Traducción de Javier Flores. Centro Bartolomé de las Casas. Cuzco.," 2008.

[193] P. Kaulicke and T. D. Dillehay, "Introducción:¿ por qué estudiar el Periodo Arcaico en el Perú?," 1999.

[194] H. Santacruz, "Origen de los pueblos Pastos," *Resumen de investigación. Disponible desde Internet en: http://www. rupestreweb. info/Pastos. html (con acceso 12/12/2013),* 2009.

[195] K. Romoli, "Las tribus de la antigua jurisdicción de Pasto en el siglo XVI," *Revista Colombiana de Antropología,* vol. 21, pp. 11-55, 1977.

[196] F. C. ARROYO, "PASTOS Y QUILLACINGAS: DOS GRUPOS ETNICOS EN BUSCA DE IDENTIDAD ARQUEOLOGICA," *Revista Colombiana de Antropología,* vol. 29, 1992.

[197] L. Herrera, M. C. De Schrimpff, W. Bray, and P. Botero, "Nuevas fechas de radiocarbono para el Precerámico en la cordillera occidental de Colombia. En: O. Ortiz-Troncoso y Thomas van der Hammen (ed.). Archaeology and Environment in Latin America," in *Proceedings of a symposium held at the 46th International Congress of Americanists, Amsterdam*, 1988, pp. 145-163.

[198] L. Chero, "Potencialidades de una integración real en la cuenca sudamericana del pacífico: Los intentos en Ecuador, Perú, y Chile; y sus retos a comienzos del siglo XX," *CA Bello, Integración Regional,* vol. 4, 2008.

[199] J. Tamayo Herrera, "Nuevo compendio de historia del Perú," *Editorial Osiris, Lima, Peru,* vol. 372, 1987.

[200] V. Sáenz, "VISIONES SOBRE GENTE URU EN BOLIVIA," *Revista Textos Antropológicos,* vol. 14, p. 55, 2003.

[201] J. A. Canahuire Ccama, "Evolución histórica y social de las naciones Collas," *Editor LACG,* p. 84, 1999.

[202] R. Cerrón Palomino, "Examen de la teoría aimarista de Uhle," *Max Uhle y el Perú antiguo,* 1998.

[203] K. Hannß, *Uchumataqu: The Lost Language of the Urus of Bolivia; a Grammatical Description of the Language as Documented Between 1894 and 1952*: CNWS publications, 2008.
[204] C. Peruano. *http: // www. inei.gob.pe*.
[205] A. Torero, "Acerca de la familia lingüística uruquilla (Uru-Chipaya)," *Revista Andina,* vol. 10, pp. 171-189, 1992.
[206] N. Wachtel, "Men of the water: The Uru problem (XVI and XVII centuries)," ed: Cambridge University Press & Editions de la Maison des Sciences de L'Homme, 1986.
[207] C. P. R, "Reconstrucción del proto-Uro:fonología. Lexis, v. XXXI/1–2, Lima, 47–104.," 2007.
[208] R. Cerrón-Palomino, "Reconstrucción del proto-uro: fonología," *Lexis,* vol. 31, pp. 47-104, 2007.
[209] T. Bouysse-Cassagne, "Poblaciones humanas antiguas y actuales," 1991.
[210] P. Muysken, "Uchumataqu: Research in progress on the Bolivian Altiplano," *Inter J Multicult Soc,* vol. 4, pp. 235-247, 2002.
[211] H. E. Manelis de Klein, "Los Urus: El extraño pueblo del altiplano," *Estudios Andinos,* vol. 3, pp. 129-150, 1973.
[212] D. V. Delgadillo, *La nación de los urus: Chipaya 1984* vol. 4: Centro Diocesano de Pastoral Social, 1998.
[213] C. Ponce Sanginés, "La cerámica de la época I de Tiwanaku," *Publicación del Instituto Nacional de Arqueologıa,* vol. 18, 1976.
[214] A. Posnansky, "Tiahuanaco, la cuna del hombre Americano: 4 Tomos en 2 volúmenes," *Tomo I y II, New York,* 1945.
[215] A. Arnaiz-Villena, V. Gonzalez-Alcos, J. I. Serrano-Vela, R. Reguera, L. Barbolla, C. Parga-Lozano*, et al.*, "HLA genes in Uros from Titikaka Lake, Peru: origin and relationship with other Amerindians and worldwide populations," *International Journal of Immunogenetics,* vol. 36, pp. 159-167, 2009.
[216] G. KARASIK and R. MACHACA, "Kollas de Jujuy," *Un pueblo, muchos pueblos. Pueblos indígenas en la Argentina cap,* vol. 6, 2016.
[217] G. Ibáñez, *Los collas*: Ediciones del Sol, 2008.
[218] M. Juárez, "Los Wichís Matacos, una cultura aborigen del Gran Chaco argentino: fotografías en blanco y negro de una cultura condenada," 2006.
[219] A. Fabre, "Los pueblos del Gran Chaco y sus lenguas, segunda parte: Los mataguayo," *Suplemento Antropológico,* pp. 313-435, 2005.
[220] M. Swadesh, *Mapas de clasificación lingüística de México y las Américas*: Universidad Nacional Atónoma de México, 1959.
[221] R. M. F. Montserrat, "Línguas indígenas no Brasil contemporâneo," *Índios no Brasil. Brasília: MEC,* pp. 93-104, 1994.
[222] G. Urban, "A história da cultura brasileira segundo as línguas nativas," *História dos índios no Brasil,* vol. 2, pp. 87-102, 1992.
[223] N. Badariotti, *Exploração no norte de Matto Grosso, região do alto Paraguay e planalto dos Parecis: apontamentos de historia natural, etnographia, geographia e impressões*: Escola typ. salesiana, 1898.
[224] J. V. d. Silva, "A Capitania de Mato Grosso: política de povoamento e população–século XVIII," *São Paulo: Tese de Doutorado: DH/FFLCH/USP,* 1994.
[225] F. P. Moi and W. F. Morales, "Arqueologia e gestão de recursos culturais entre os Paresi da Chapada dos Pareci, MT (Brasil)," *Especiaria: Cadernos de Ciências Humanas,* vol. 11, 2015.
[226] E. Krebs and J. Braunstein, "The renewal of Gran Chaco studies," *History of Anthropology Newsletter,* vol. 28, pp. 9-19, 2011.

[227] J.-A. Alvarsson, "The mataco of the Gran Chaco," *Uppsala: University of Uppsala,* 1988.
[228] J. Braunstein and E. Miller, "Ethnohistorical introduction," *Peoples of the Gran Chaco,* vol. 15, p. 22, 1999.
[229] D. A. Demarchi and A. G. Ministro, "Genetic structure of native populations from the Gran Chaco region, South America," *International Journal of Human Genetics,* vol. 8, pp. 131-141, 2008.
[230] J. A. Braunstein, S. Salceda, H. A. Calandra, M. Méndez, and S. Ferrarini, "Historia de los chaqueños–Buscando en la "papelera de reciclaje" de la antropología sudamericana"," *Acta Americana. Journal of the Swedish Americanist Society,* vol. 10, pp. 63-93, 2002.
[231] J. S. Saeger, *The Chaco Mission Frontier: The Guaycuruan Experience*: University of Arizona Press, 2000.
[232] M. Mendoza, "Range area and seasonal campsites of Toba bands in western Chaco, Argentina," *Before Farming,* vol. 2003, pp. 1-12, 2003.
[233] Č. Loukotka and J. Wilbert, *Classification of South American indian languages* vol. 7: Latin American Center, University of California, Los Angeles, 1968.
[234] J. Braunstein and A. Vidal, "En prensa. The Gran Chaco: convergence of languages and peoples," ed: The Languages of Hunter-Gatherers. Historical and global perspectives ….
[235] B. Sušnik, *Dimensiones migratorias y pautas culturales de los pueblos del Gran Chaco y de su periferia: enfoque etnológico*: Instituto de Historia, Facultad de Humanidades, Universidad Nacional del …, 1972.
[236] Yfull-Q-Z5915. Available: https://www.yfull.com/tree/Q-Z5915/
[237] T. F. Lynch and K. A. Kennedy, "Early human cultural and skeletal remains from Guitarrero Cave, Northern Peru," *Science,* vol. 169, pp. 1307-1309, 1970.
[238] C. Chauchat, "Additional observations on the Paijan complex," *Nawpa Pacha,* vol. 16, pp. 51-64, 1978.
[239] P. Ossa, "Paijan in early Andean prehistory: the Moche Valley evidence," *New evidence for the Pleistocene peopling of the Americas. University of Alberta, Edmonton,* pp. 290-295, 1978.
[240] T. D. Dillehay, J. Rossen, T. C. Andres, and D. E. Williams, "Preceramic adoption of peanut, squash, and cotton in northern Peru," *Science,* vol. 316, pp. 1890-1893, 2007.
[241] L. Kaplan, T. F. Lynch, and C. Smith, "Early cultivated beans (Phaseolus vulgaris) from an intermontane Peruvian valley," *Science,* vol. 179, pp. 76-77, 1973.
[242] L. Perry, D. H. Sandweiss, D. R. Piperno, K. Rademaker, M. A. Malpass, A. Umire*, et al.*, "Early maize agriculture and interzonal interaction in southern Peru," *nature,* vol. 440, pp. 76-79, 2006.
[243] D. Ugent, S. Pozorski, and T. Pozorski, "Archaeological potato tuber remains from the Casma Valley of Peru," *Economic Botany,* vol. 36, pp. 182-192, 1982.
[244] J. C. Wheeler, "On the origin and early development of camelid pastoralism in the Andes," *Animals and archaeology,* vol. 4, pp. 1-13, 1984.
[245] J. C. Wheeler, "Patrones prehistóricos de utilización de los camélidos sudamericanos," *Boletín de arqueología pucp,* pp. 297-305, 1999.
[246] R. S. Solis, J. Haas, and W. Creamer, "Dating Caral, a preceramic site in the Supe Valley on the central coast of Peru," *Science,* vol. 292, pp. 723-726, 2001.
[247] R. Shady Solís, "La civilización de Caral-Supe: 5000 años de identidad cultural en el Perú," ed: Lima: Instituto Nacional de Cultura, 2005.
[248] T. Grieder, A. B. Mendoza, C. Smith Jr, R. Malina, and L. Galgada, "Peru: A Preceramic Culture in Transition," ed: Univ. of Texas Press, Austin, 1988.

[249] S. Izumi and T. Sono, *Andes I-: Excavations at Kotosh, Peru, 1960* vol. 2: Kadokawa Publishing Company, 1963.
[250] T. D. Dillehay, P. J. Netherly, and J. Rossen, "Middle Preceramic public and residential sites on the forested slope of the western Andes, northern Peru," *American Antiquity,* pp. 733-759, 1989.
[251] P. Kaulicke, "Perspectivas regionales del periodo formativo en el Perú: una introducción," *Boletín de arqueología PUCP,* pp. 9-13, 1998.
[252] N. Sharratt, M. Golitko, P. Ryan Williams, and L. Dussubieux, "Ceramic production during the Middle Horizon: Wari and Tiwanaku clay procurement in the Moquegua Valley, Peru," *Geoarchaeology: An International Journal,* vol. 24, pp. 792-820, 2009.
[253] J. C. Tello, *Chavín, cultura matriz de la civilización andina* vol. 2: la Universidad de San Marcos, 1960.
[254] J. Ramírez Beltrán, "Civilización Caral (Supe)-Lima-Perú," *Jornada de Técnicas de Reparación y Conservación del Patrimonio,* vol. 4, 2015.
[255] C. Pereyra, *Breve historia de América*: Aguilar México, 1958.
[256] H. A. Calandra and S. A. Salceda, "Registro arqueológico regional chaqueño," 2006.
[257] P. R. Renne, J. M. Feinberg, M. R. Waters, J. Arroyo-Cabrales, P. Ochoa-Castillo, M. Perez-Campa*, et al.*, "Age of Mexican ash with alleged 'footprints'," *Nature,* vol. 438, pp. E7-E8, 2005.
[258] S. González, D. Huddart, M. R. Bennett, and A. González-Huesca, "Human footprints in Central Mexico older than 40,000 years," *Quaternary Science Reviews,* vol. 25, pp. 201-222, 2006.
[259] R. Adler, "The first Americans," *New Scientist, 190, No 2546, 42-46,* 2006.
[260] S. Gonzalez and D. Huddart, "The late pleistocene human occupation of Mexico," in *Memoria de Simposio Internacional de FUMDHAM*, 2008.
[261] E. Velásquez García, "Los habitantes más antiguos del actual territorio mexicano," *Nueva historia general de México,* pp. 17-70, 2010.
[262] M. G. Sanchez, "A synopsis of Paleo-Indian archaeology in Mexico," *Kiva,* vol. 67, pp. 119-136, 2001.
[263] A. Cyphers, *Las capitales olmecas de San Lorenzo y La Venta*: Fondo de Cultura Económica el Colegio de México, Fideicomiso Historia de las …, 2018.
[264] K. Hirth, A. Cyphers, R. Cobean, J. De León, and M. D. Glascock, "Early Olmec obsidian trade and economic organization at San Lorenzo," *Journal of Archaeological Science,* vol. 40, pp. 2784-2798, 2013.
[265] M. E. Pohl, K. O. Pope, and C. von Nagy, "Olmec origins of Mesoamerican writing," *Science,* vol. 298, pp. 1984-1987, 2002.
[266] M. Kramme, *Mayan, Incan, and Aztec Civilizations, Grades 5-8*: Mark Twain Media, 2012.
[267] C. Waldman, *Encyclopedia of native American tribes*: Infobase Publishing, 2006.
[268] P. Carrasco, "Estructura político-territorial del imperio technoca: la Triple Alianza de Tenochtitlan," *Tetzcoco y Tlacopan,* p. 267, 1996.
[269] H. D. Chávez and R. A. C. Aguilar, "Los pueblos de alta cultura de Mesoamérica," 2008.
[270] J. Marcus, "Zapotec writing," *Scientific American,* vol. 242, pp. 50-67, 1980.
[271] E. Z. Vogt, *Handbook of Middle American Indians, Volumes 7 and 8: Ethnology*, 2015.
[272] "Haplogroup https://haplogroup.org/native-american-q-m3-tree-p2-q-m242-news-6-nov-2016/."
[273] "YFull BZ4012."
[274] *YFull SK1974/Q-Y26547*. Available: https://www.yfull.com/tree/Q-Y26547/

[275] *Variaciones climáticas y ambientales en Patagonia durante los últimos cinco millones de años*, 1996.
[276] R. Bonnichsen, D. Stanford, and J. L. Fastook, "Environmental change and developmental history of human adaptive patterns: the Paleoindian case," *The geology of North America and adjacent oceans during the last deglaciation,* vol. 3, pp. 403-424, 1987.
[277] D. S. Amick, "Patterns of technological variation among Folsom and Midland projectile points in the American Southwest," *Plains Anthropologist,* vol. 40, pp. 23-38, 1995.
[278] D. E. Crabtree, "A STONEWORKER'SAPPROACH TO ANALYZING AND REPLICATING THE LINDENMEIER FOLSOM," *Ariel,* vol. 129, p. 92.22, 1966.
[279] E. Callahan, *The basics of biface knapping in the eastern fluted point tradition: a manual for flintknappers and lithic analysts*: Eastern States Archeological Federation, 1979.
[280] D. W. Clark and A. MCFADYEN CLARK, "Batza Téna. Trail to obsidian: archaeology at an Alaskan obsidian source," *Paper-Archaeological Survey of Canada,* pp. i-xvi, 1993.
[281] E. J. Dixon, *Quest for the Origins of the First Americans*: University of New Mexico Press, 1993.
[282] H. G. Nami, "Investigaciones actualísticas para discutir aspectos técnicos de los cazadores-recolectores del tardiglacial: El problema Clovis-Cueva Fell," in *Anales del Instituto de la Patagonia*, 1997, pp. 151-186.
[283] A. L. Bryan, *Paleoamerican prehistory as seen from South America*: Center for the Study of Early Man, University of Maine, 1986.
[284] A. J. Ranere and R. G. Cooke, "Paleoindian occupation in the Central American tropics," *Clovis: Origins and adaptations,* pp. 237-253, 1991.
[285] J. Bird, "A comparison of south Chilean and Ecuadorian “fishtail” projectile points," *Kroeber Anthropological Society Papers,* vol. 40, pp. 52-71, 1969.
[286] L. R. Binford, "Contemporary model building: paradigms and the current state of Palaeolithic research," *Models in archaeology,* vol. 10, 1972.
[287] L. S. Klejn, "Marxism, the systemic approach, and archaeology," *The explanation of culture change: models in prehistory. London, Duckworth,* pp. 691-710, 1973.
[288] C. Gnecco, "Fluting technology in South America," *Lithic Technology,* vol. 19, pp. 35-42, 1994.
[289] B. D. Smith, "The initial domestication of Cucurbita pepo in the Americas 10,000 years ago," *Science,* vol. 276, pp. 932-934, 1997.
[290] J. R. Harlan, "Agricultural origins: centers and noncenters," *Science,* vol. 174, pp. 468-474, 1971.
[291] J. R. Harlan, "Crops and Man. American Society of Agronomy," *Crop Science Society of America, Madison, Wisconsin,* vol. 16, pp. 63-262, 1992.
[292] P. Gepts and D. Debouck, "Origin, domestication, and evolution of the common bean (Phaseolus vulgaris L.)," *Common beans: research for crop improvement,* vol. 7, p. 53, 1991.
[293] F. O. Freitas and P. G. Bustamante, "Amazonian maize: diversity, spatial distribution and historical-cultural diffusion," *Tipití: Journal of the Society for the Anthropology of Lowland South America,* vol. 11, pp. 60-65, 2013.
[294] J. H. Hill, "Proto-Uto-Aztecan: a community of cultivators in Central Mexico?," *American Anthropologist,* vol. 103, pp. 913-934, 2001.
[295] L. Campbell, *American Indian languages: the historical linguistics of Native America* vol. 4: Oxford University Press on Demand, 2000.
[296] C. S. Fowler, "Some lexical clues to Uto-Aztecan prehistory," *International Journal of American Linguistics,* vol. 49, pp. 224-257, 1983.

[297] K. Hale and D. Harris, "Historical linguistics and archaeology," *Handbook of North American Indians,* vol. 9, pp. 170-177, 1979.
[298] R. G. Matson, "The spread of maize to the Colorado Plateau," *Archaeology Southwest,* vol. 13, pp. 10-11, 1999.
[299] P. Bellwood, "Prehistoric cultural explanations for the existence of widespread language families," *Archaeology and linguistics: Aboriginal Australia in global perspective,* pp. 123-34, 1997.
[300] P. Bellwood, "Austronesian prehistory and Uto-Aztecan prehistory: similar trajectories," *University of Arizona Department of Anthropology Lecture series. January,* vol. 27, 1999.
[301] S. Luciana and V. V. F. Ferreira. (2018). *Karitiana, Povos Indigenas no Brasil*. Available: https://pib.socioambiental.org/es/Povo:Karitiana
[302] YFull/Q-Y27992. *YFull*. Available: https://www.yfull.com/tree/Q-Y27992/
[303] I. Combès, "De Sanandita al Itiyuro: los chanés, los chiriguanos (¿ y los tapietes?) al sur del Pilcomayo," *Indiana,* vol. 24, pp. 259-289, 2007.
[304] R. S. Walker and L. A. Ribeiro, "Bayesian phylogeography of the Arawak expansion in lowland South America," *Proceedings of the Royal Society B: Biological Sciences,* vol. 278, pp. 2562-2567, 2011.
[305] G. Vargas-Alarcon, J. Moscoso, J. Martinez-Laso, J. M. Rodriguez-Perez, C. Flores-Dominguez, J. I. Serrano-Vela*, et al.*, "Origin of Mexican Nahuas (Aztecs) according to HLA genes and their relationships with worldwide populations," *Molecular immunology,* vol. 44, pp. 747-755, 2007.
[306] E. Gómez-Casado, J. Martínez-Laso, J. Moscoso, J. Zamora, M. Martin-Villa, M. Perez-Blas*, et al.*, "Origin of Mayans according to HLA genes and the uniqueness of Amerindians," *Tissue Antigens,* vol. 61, pp. 425-436, 2003.
[307] M. Petzl-Erler, C. Gorodezky, Z. Layrisse, W. Klitz, L. Fainboim, and C. Vullo, "Anthropology report for the Latin-American region: Amerindian and admixed populations," *Genetic diversity of HLA. Functional and Medical Implication,* vol. 1, 1997.
[308] M. R. Oudijk, "Mixtecos y zapotecos en la época prehispánica," *Arqueología mexicana,* vol. 15, pp. 58-62, 2008.
[309] M. Ruhlen, *A guide to the world's languages: Classification* vol. 1: Stanford University Press, 1991.
[310] J. K. Josserand, M. Winter, and N. A. Hopkins, *Essays in Otomanguean culture history*: Anthropology Section, Department of Sociology and Anthropology, Vanderbilt …, 1984.
[311] P. S. Bellwood, "First farmers: the origins of agricultural societies," 2005.
[312] A. Arnaiz-Villena, G. Vargas-Alarcón, J. Granados, E. Gómez-Casado, J. Longás, M. Gonzalez-Hevilla*, et al.*, "HLA genes in Mexican Mazatecans, the peopling of the Americas and the uniqueness of Amerindians," *Tissue Antigens,* vol. 56, pp. 405-416, 2000.
[313] C. H. Brown, C. R. Clement, P. Epps, E. Luedeling, and S. Wichmann, "The Paleobiolinguistics of Domesticated Manioc (Manihot esculenta)," *Ethnobiology Letters,* vol. Vol. 4 (2013), pp. pp. 61-70 (10 pages).
[314] J. Diamond and P. Bellwood, "Farmers and their languages: the first expansions," *Science,* vol. 300, pp. 597-603, 2003.
[315] A. Duputié, J. Salick, and D. McKey, "Evolutionary biogeography of Manihot (Euphorbiaceae), a rapidly radiating Neotropical genus restricted to dry environments," *Journal of Biogeography,* vol. 38, pp. 1033-1043, 2011.
[316] C. Isendahl, "The domestication and early spread of manioc (Manihot esculenta Crantz): a brief synthesis," *Latin American Antiquity,* vol. 22, pp. 452-468, 2011.

[317] F. S. Noelli, "La distribución geográfica de las evidencias arqueológicas guaraní," *Revista de Indias,* vol. 64, pp. 17-34, 2004.
[318] M. H. B. P. P. U. F. d. Bahia). (20 de agosto de 2018). *Maxakalí - Povos indígenas em Minas Gerais*. Available: https://pib.socioambiental.org/pt/Povo:Maxakali
[319] J. Mason, "Alden: The languages of South American Indians. Handbook of South American Indians, Bd. 6," *Bull. Bur. Amer. Ethnol,* 1950.
[320] A. D. I. Rodrigues, *Línguas brasileiras: para o conhecimento das línguas indígenas* vol. 11: Edições Loyola, 1994.
[321] A. D. Rodrigues, "Macro-jê," *The amazonian languages,* pp. 165-206, 1999.
[322] J. P. V. BARROS, "MACRO-GUAICURÚ-MACRO-JE."
[323] A. D. Rodrigues, "Typological parallelism due to social contact: Guató and Kadiwéu," in *ANNUAL MEETING OF THE BERKELEY LINGUISTICS SOCIETY*, 1983, pp. 218-222.
[324] J. Pozzobon, *Sociedade e improviso: estudo sobre a (des) estrutura social dos índios Maku*: Museu do Índio, FUNAI, 2011.
[325] A. da Silva Campos, *Conhecendo as raízes do Brasil: História e cultura dos povos indígenas*: Cultural Brasil, 2017.
[326] G. C. Becerra, *Viviendo en el bosque. Un siglo de investigaciones sobre los makú del Noroeste amazónico*: Universidad Nacional de Colombia, 2015.
[327] P. A. Lolli, "A plasticidade Maku," *Ilha Revista de Antropologia,* vol. 18, pp. 177-198, 2016.
[328] J. Pozzobon, "maku (1999)," *Instituto Socioambiental-ISA. Recuperado el,* vol. 26, p. 07.
[329] V. Martins, *Reconstrução fonológica do Protomaku oriental*: Netherlands Graduate School of Linguistics, 2005.
[330] J. A. Pozzobon, "Parenté et demographie chez les Indiens Makú," Paris 7, 1991.
[331] J. Idrovo and D. Gomis, "Historia de una región formada en el Austro del Ecuador y sus conexiones con el norte del Perú," *Imprenta América Latina, Cuenca,* pp. 41-44, 2009.
[332] C. A. L. Illescas, "TRADICIÓN CERÁMICA Y OCUPACIÓN PRECOLOMBINA DEL PIEDEMONTE ORIENTAL DE LOS ANDES: EL CASO DEL VALLE DEL RÍO CUYES (MORONA SANTIAGO, ECUADOR)," *Anales de la Universidad Central del Ecuador,* vol. 1, 2016.
[333] M. d. Gauiria, "Relación que envió a mandar su majestad se hiciese de esta ciudad de Cuenca y de toda su provincia (Cañaribamba). In Relaciones Geográficas de Indias – Perú," *Atlas, pp. 281–87,* vol. Madrid: Jiménez de la Espada., 1965.
[334] C. Itier, *Les Incas*: Les Belles lettres, 2008.
[335] H. Pablos, "Relación que envió a mandar Su Magestad se hiziese desta ciudad de Cuenca y de toda su provincia," *Relaciones Geográficas de Indias: Perú,* vol. 2, pp. 265-270, 1965.
[336] L. Hirschkind, "Historia de la población indígena del Cañar," *Revista de Antropología,* vol. 20, pp. 41-78, 2013.
[337] P. Cieza de León, "La crónica del Perú [1553]," *Ed. M. Ballesteros,* 1962.
[338] F. Salomon, "Ancestros, huaqueros y los posibles antecedentes del "Incaísmo" cañari," *Revista de Antropología,* pp. 7–40, 2013.
[339] *YFull Q-Z5908/Q-B48*. Available: https://www.yfull.com/tree/Q-Z5908/
[340] "YFull Q-BZ3401."
[341] *YFull Q-M848**. Available: https://www.yfull.com/tree/Q-M848*/
[342] *YFull Q-Z5906*. Available: https://www.yfull.com/tree/Q-Z5906/
[343] V. Slon, C. Hopfe, C. L. Weiß, F. Mafessoni, M. De La Rasilla, C. Lalueza-Fox*, et al.*, "Neandertal and Denisovan DNA from Pleistocene sediments," *Science,* vol. 356, pp. 605-608, 2017.
[344] E. Boëda, A. Lourdeau, C. Lahaye, G. Felice, S. Viana, I. Clemente-Conte*, et al.*, "Paleoamerican odyssey," 2013.

[345] C. Lahaye, G. Guérin, M. Gluchy, C. Hatté, M. Fontugne, I. Clemente-Conte*, et al.*, "Another site, same old song: the Pleistocene-Holocene archaeological sequence of Toca da Janela da Barra do Antonião-North, Piauí, Brazil," *Quaternary Geochronology,* vol. 49, pp. 223-229, 2019.

[346] D. G. Anderson, A. C. Goodyear, J. Kennett, and A. West, "Multiple lines of evidence for possible human population decline/settlement reorganization during the early Younger Dryas," *Quaternary International,* vol. 242, pp. 570-583, 2011.

[347] O. Bar-Yosef and A. Belfer-Cohen, "The Dawn of Farming in the Near East, Studies in Early Near Eastern Production, Subsitence, and Environment," 2002.

[348] M. I. Eren, *Hunter-gatherer behavior: human response during the Younger Dryas*: Left Coast Press, 2012.

[349] L. A. Borrero, "The prehistoric exploration and colonization of Fuego-Patagonia," *Journal of World Prehistory,* vol. 13, pp. 321-355, 1999.

[350] T. D. Dillehay, "The settlement of the Americas: a new prehistory," 2000.

[351] E. León, "Orígenes humanos en los Andes del Perú," *Editorial Universidad San Martín de Porres. Lima, Perú,* 2007.

[352] M. Pino, M. Chávez-Hoffmeister, X. Navarro-Harris, and R. Labarca, "The late pleistocene Pilauco site, Osorno, south-central Chile," *Quaternary International,* vol. 299, pp. 3-12, 2013.

[353] W. Mahaney, V. Kalm, D. Krinsley, P. Tricart, S. Schwartz, J. Dohm*, et al.*, "Evidence from the northwestern Venezuelan Andes for extraterrestrial impact: The black mat enigma," *Geomorphology,* vol. 116, pp. 48-57, 2010.

[354] C. G. Park, H. K. Cho, H. J. Shin, K. H. Park, and H. B. Lim, "Comparison of mutagenic activities of various ultra-fine particles," *Toxicological research,* vol. 34, pp. 163-172, 2018.

[355] N. de Oliveira Alves, A. T. Vessoni, A. Quinet, R. S. Fortunato, G. S. Kajitani, M. S. Peixoto*, et al.*, "Biomass burning in the Amazon region causes DNA damage and cell death in human lung cells," *Scientific reports,* vol. 7, pp. 1-13, 2017.

[356] K. Ye and Z. Gu, "Recent advances in understanding the role of nutrition in human genome evolution," *Advances in Nutrition,* vol. 2, pp. 486-496, 2011.

[357] J. Carlson, W. S. DeWitt, and K. Harris, "Inferring evolutionary dynamics of mutation rates through the lens of mutation spectrum variation," *Current Opinion in Genetics & Development,* vol. 62, pp. 50-57, 2020.

[358] A. E. Gusick and M. K. Faught, "Prehistoric archaeology underwater: A nascent subdiscipline critical to understanding early coastal occupations and migration routes," in *Trekking the Shore*, ed: Springer, 2011, pp. 27-50.

[359] L. L. Johnson, *Paleoshorelines and prehistory: an investigation of method*: CRC Press, 1991.

[360] P. M. Masters, "Detection and assessment of prehistoric artifact sites off the coast of southern California," *Quaternary coastlines and marine archaeology,* pp. 189-213, 1983.

[361] I. Cartajena, R. Simonetti, P. López, C. Morales, and C. Ortega, "Submerged paleolandscapes: Site GNL Quintero 1 (GNLQ1) and the first evidences from the Pacific Coast of South America," in *Prehistoric Archaeology on the Continental Shelf*, ed: Springer, 2014, pp. 131-149.

[362] L. Núñez, J. Varela, and R. Casamiquela, "Ocupación paleoindio en el Centro-Norte de Chile: adaptación circunlacustre en las tierras bajas," *Estudios Atacameños,* pp. 142-185, 1987.

ANEXO I - Aprobación del proyecto por los comités de ética nacionales

COMITÉ PROVINCIAL DE BIOETICA
PROVINCIA DE JUJUY

Ref. Evaluación de Protocolo

San Salvador de Jujuy, 22 de diciembre del 2008.

Sr. Ministro de Salud
Dr. Victor Urbani

PRESENTE

Por la presente me dirijo a Ud., a fin de elevar para su conocimiento, el informe de APROBACION realizado por el Comité Provincial de Bioética de Jujuy, el día 22 de Diciembre del 2008, sobre el protocolo de investigación, presentado por el Dr. Dipierri, sobre "Aportes continentales diferenciales en la conformación de las poblaciones humanas de América Latina"

Sin más, saluda a Ud. muy atte.

Dra. Maria de La Paz Bessio
Vice Pres

Lic. Maria Giardina
Presidente

Ministerio de Salud
Provincia de Jujuy

ANEXO II - Aprobación del proyecto por los comités de ética nacionales

CONICET

IMBICE

Comisión de Investigaciones Científicas
Gobierno de la Provincia de Buenos Aires

La Plata, 13 de Agosto de 2009.-

Dra. Graciela Bailliet
Dr. Claudio M. Bravi
IMBICE
S/D

Estimados Dres. Bailliet / Bravi:

Tenemos el agrado de dirigirnos a uds. con el objeto de informarles que el Comité de Etica en Investigaciones Biomédicas del IMBICE evaluó y aprobó las modificaciones incorporadas al texto de la Cartilla de Información para el Donante y a la organización del formulario de Consentimiento Informado, ambos elaborados en el marco del Proyecto "Aportes continentales diferenciales en la conformación de las poblaciones humanas de América Latina", que planean desarrollar en el laboratorio de Genética Molecular Poblacional del IMBICE.

Sin otro particular saludamos a uds. atentamente.

Dra. Lidia A. Vidal Rioja

Comité de Ética en Investigaciones Biomédicas
IMBICE
Dra. Lidia A Vidal Rioja - Coordinador
Dr. Ricardo S. Calandra
Prof. Jorge Asuaje
Dr. Luis Julio Couyet
Dr. Eduardo Luis Tinant

ANEXO III - Consentimiento Informado del Proyecto de Investigación

"Aportes continentales diferenciales en la conformación de las poblaciones humanas de América Latina"

Yo, ……., DNI N°……..…con domicilio legal en calle……..._Nº……_de la localidad de …… de la provincia de ……. declaro que he sido informado sobre la realización de un estudio genético para conocer el origen o procedencia de los antepasados de los habitantes actuales de América y averiguar qué proporción de ellos es originaria de este continente o bien llegó aquí desde Europa, Asia o África.

Manifiesto acceder voluntariamente a participar en este proyecto, del cual he sido informado a través de una entrevista con miembros del grupo de investigación (abajo firmantes), quienes me comunicaron ampliamente sobre las características y alcances del estudio y me entregaron una hoja de "Información para el Participante" donde se detallan los objetivos, características genéticas a investigar, metodología, beneficios, riesgos y confidencialidad de los datos. Luego de leer esta hoja de información he podido realizar libremente cualquier pregunta relacionada con el proyecto.

Comprendo las características del trabajo y acepto libre y voluntariamente que me sea tomada una muestra de sangre o de saliva para ser utilizada en el presente estudio. Acepto además aportar datos sobre el origen o procedencia de mis padres, abuelos y bisabuelos para los fines indicados en la hoja de información. He comprendido que si lo deseo puedo retirarme de la investigación sin tener que dar explicaciones, como así también solicitar los datos obtenidos con mi muestra y reclamar la muestra sobrante del estudio.

La extracción de sangre o saliva (tachar lo que no corresponda) fue realizada por……
.....……_Cargo……_DNI:……

La muestra es recibida por el investigador responsable quien la codifica y archiva el consentimiento informado en el Laboratorio de Bioantropología de la Universidad Nacional de Jujuy. La toma de la muestra y su codificación se realizó ante la presencia del/los testigos abajo firmantes.

Dado a los .....días del mes de.....de 200…. 115 ANEXO

_____

Firma del donante Aclaración y DNI

_____

Firma y cargo del entrevistador Aclaración y DNI

_____

Firma y cargo del extraccionista Aclaración y DNI

_____

Firma del testigo Aclaración y DNI

Investigadores Responsables: Dr. Claudio Bravi - Dra. Graciela Bailliet
Instituto Multidisciplinario de Biología Celular (IMBICE), Calle 526 e/ 10 y 11, La Plata.
TeleFax: (0221) 421-0112

ANEXO IV - Encuesta genealógica

Muestra N°
VARÓN
MUJER

_____

DONANTE Apellido y nombres (edad)

_____

DONANTE Lugar de nacimiento (localidad o paraje / departamento / provincia / país)

_____

1- PADRE Apellido y Nombres Lugar de nacimiento (localidad / departamento / provincia / país)

_____

2- ABUELO PATERNO Apellido y Nombres Lugar de nacimiento (localidad / departamento / provincia / país)

_____

3- ABUELA PATERNA Apellido y Nombres Lugar de nacimiento (localidad / departamento / provincia / país)

_____

4- MADRE Apellido y Nombres Lugar de nacimiento (localidad / departamento / provincia / país)

_____

5- ABUELO MATERNO Apellido y Nombres Lugar de nacimiento (localidad / departamento / provincia / país)

_____

6- ABUELA MATERNA Apellido y Nombres Lugar de nacimiento (localidad / departamento / provincia / país)

Favor de preguntar al donante si sabe que alguno de sus padres/abuelos es/era descendiente de inmigrantes o perteneciente a alguna parcialidad socio-étnica (por ejemplo **judío sefaradí**, **árabe, gringo, criollo, nativo**, **aborigen**, etc.).
Cuando el donante desconoce el lugar de nacimiento de algún ancestro, favor de consignar el lugar de residencia habitual (e indicar en la encuesta con la leyenda “vive o vivía en”)

ANEXO V - Red de haplotipos STRs Q-M3 y Q-Z780

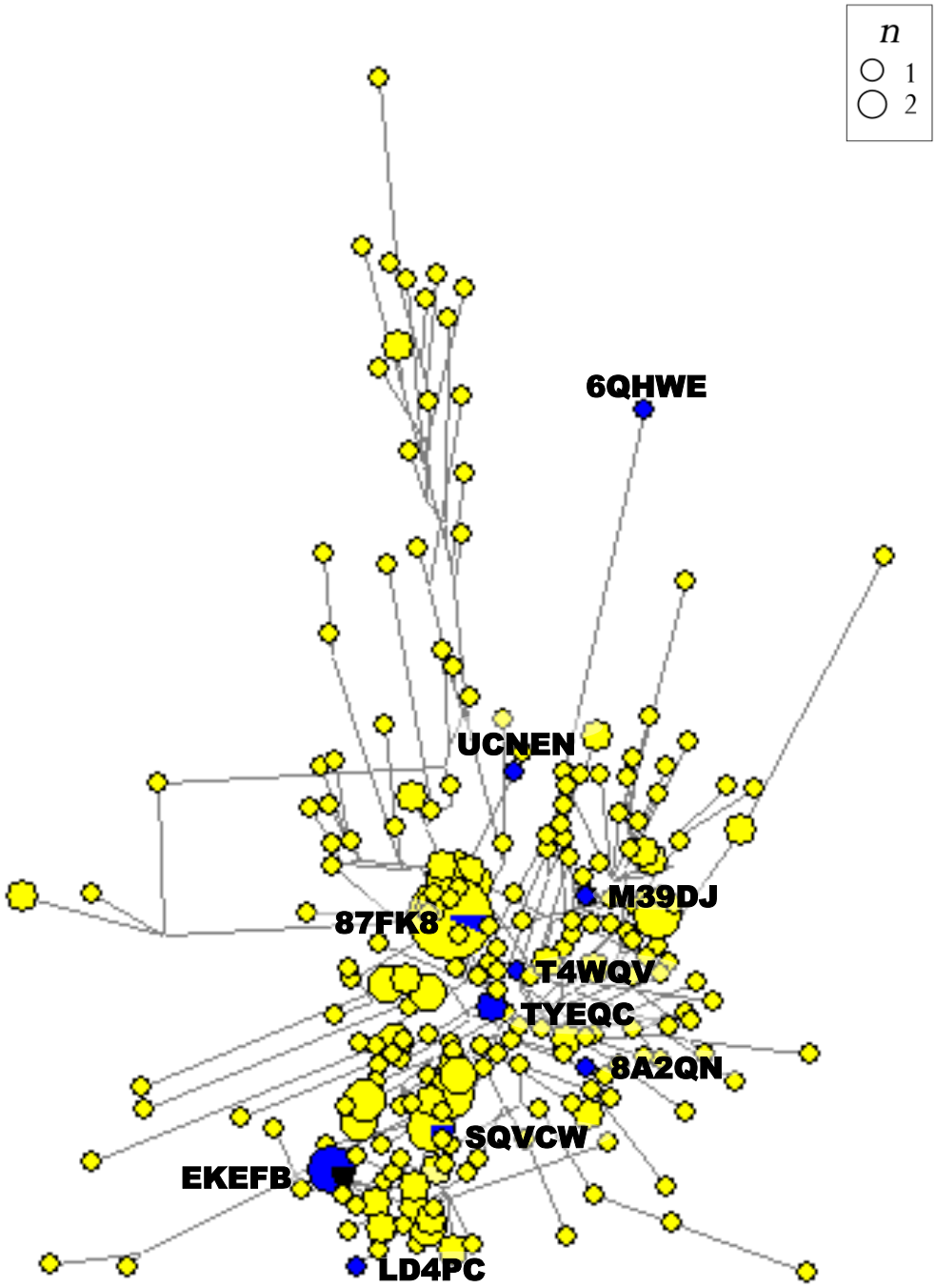


Figura anexa V: Red de haplotipos construida con 17 STR (YFiler). Esta red contiene 257 haplotipos, los puntos amarillos son haplotipos Q-M3 y Q-Z780 contempladas en la Tesis Doctoral de Jurado Medina [108]. Los puntos azules representan los haplotipos de las muestras Q-M3 seleccionadas para la secuenciación.

ANEXO VI - Red de haplotipos STR Q-Z780

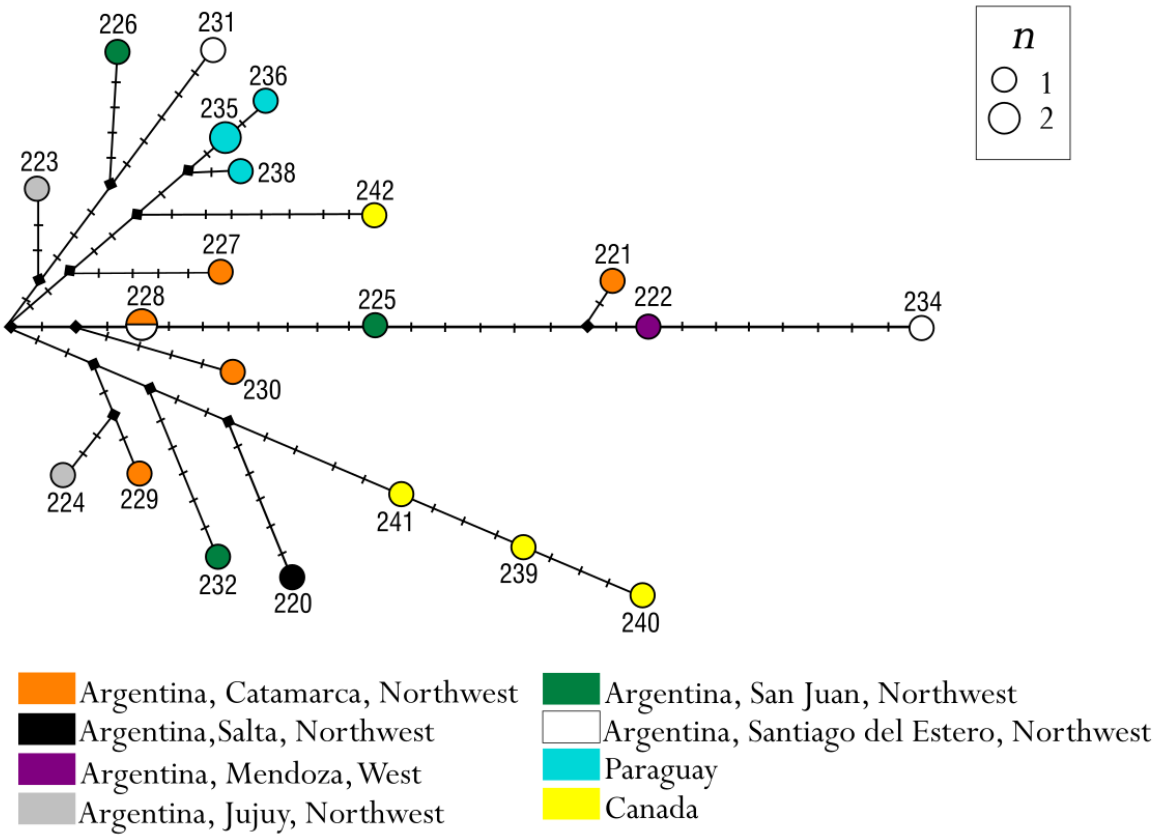


Figura anexa VI. Red de Haplotipos STR Q-Z780 construida en base a 19 muestras de Argentina y Paraguay y 4 muestras de Canadá [46, 49]. Las dos muestras de nuestra colección seleccionadas para la secuenciación, son el haplotipo 222 que corresponde a la muestra S8BAL y el haplotipo 229 corresponde a la muestra Z8ZMY.

ANEXO VII - Pipeline

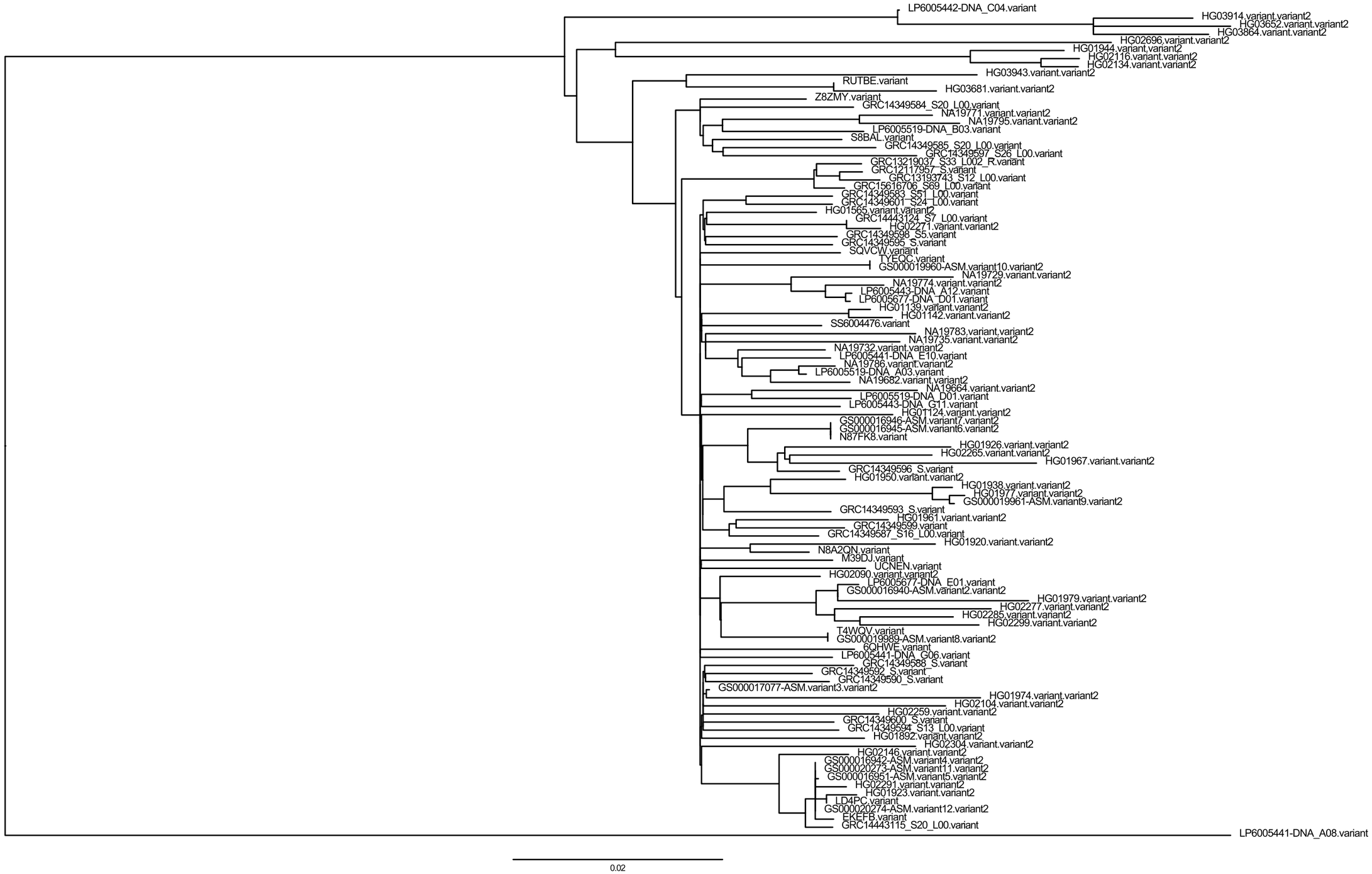
```
#!/bin/bash

#############################################################################
################################ START ######################################
#############################################################################


################################  PARAMETERS  ###############################

## Information on the sample and files ##
SAMPLENAME=LD4PC        #IMPORTANT AND REQUIRED - Sample Identifier (no spaces)
FIRSTREADS=LD4PC_HNHYNCCXX_L6_1.clean.fq        #IMPORTANT AND REQUIRED - fastq file containing the 1st reads
from the pair
MATES=LD4PC_HNHYNCCXX_L6_2.clean.fq     #IMPORTANT AND REQUIRED - fastq file containing mate pairs
REGIONS=Y       #IMPORTANT AND REQUIRED -  .bed file with regions information for WES ####### QUE PASA CUANDO
NO TENEMOS? DEBERÍA SACARSE LA OPCIÓN -L DE HAPLOTYPECALLER

PREFIX=$SAMPLENAME
PREFIX_QC=$PREFIX\_QC
LOG=$PREFIX\_log

##### READ GROUPS #####
LB=library      # RGLB=String   REQUIRED Library.
PL=Illumina     # RGPL=String   REQUIRED platform (e.g. illumina, solid). IMPORTANTE para GATK: no puede
ponerse UNKNOWN por más que lo muestre como posible. Debe especificarse una plataforma, la marca en general
(ILLUMINA y no HISEQ200, por ejemplo).
SM=LD4PC
# RGSM=String   REQUIRED sample name Required.
PU=1    # RGPU=String   REQUIRED platform unit (eg. run barcode).
ID=1    # RGID=String   ID Default value: 1. This option can be set to 'null' to clear the default value.
CN=null # RGCN=String   sequencing center name Default value: null.
DS=null # RGDS=String   description Default value: null.
DT=null # RGDT=Iso8601Date      run date Default value: null.
PI=null # RGPI=INTEGER  predicted insert size Default value: null.

### PATHs ###
HGREF=/home/marina/Paula/Reference_seq/HGREF    #Reference genome/GATK bundle directory do not change if
already set in PATH

### REFERENCE FILES USED FOR ANALYSIS ###
REF=$HGREF/human_g1k_v37_decoy.fasta    #reference genome .fasta file
DBSNP=$HGREF/dbsnp_137.b37.vcf   #dbsnp .vcf file
KGINDELS=$HGREF/1000G_phase1.indels.b37.vcf
MILLSINDELS=$HGREF/Mills_and_1000G_gold_standard.indels.b37.vcf
KSNP=$HGREF/1000G_phase1.snps.high_confidence.b37.vcf

### OPTION'S VALUES ###
QFILTER_READS=28        #Phred score value to filter reads in PrinSeq.
MINPRUN=3       # minPruning sets the kmer size for the BRUIJN tree in variant calling.
CALL=20.0       # Variant Call Confidence when creating .vcf file.

### PERFORMANCE VALUES ###
XMX=Xmx8g  #Java heap max size
GATK_NT=8 #Num of data threads sent to processor, http://gatkforums.broadinstitute.org/discussion/1975/how-can-
i-use-parallelism-to-make-gatk-tools-run-faster
GATK_NCT=4 #Num of CPU threads for each data thread, http://gatkforums.broadinstitute.org/discussion/1975/how-
can-i-use-parallelism-to-make-gatk-tools-run-faster

############ DO NOT EDIT BEYOND THIS LINE (unless you know what you are doing) ############
#-----------------------------------------------------------------------------------------

###########################    COMMANDS     ############################


mkdir -p $LOG   # make directory for log files

clear

echo "Running Sample $SM, $ID, $FASTQC"
#Quality control. Input: 2 fastq files (one per mate); output: 2 html files
perl /home/marina/Paula/FASTQC/FastQC/fastqc $FIRSTREADS $MATES --extract
```

ANEXO VII - Pipeline

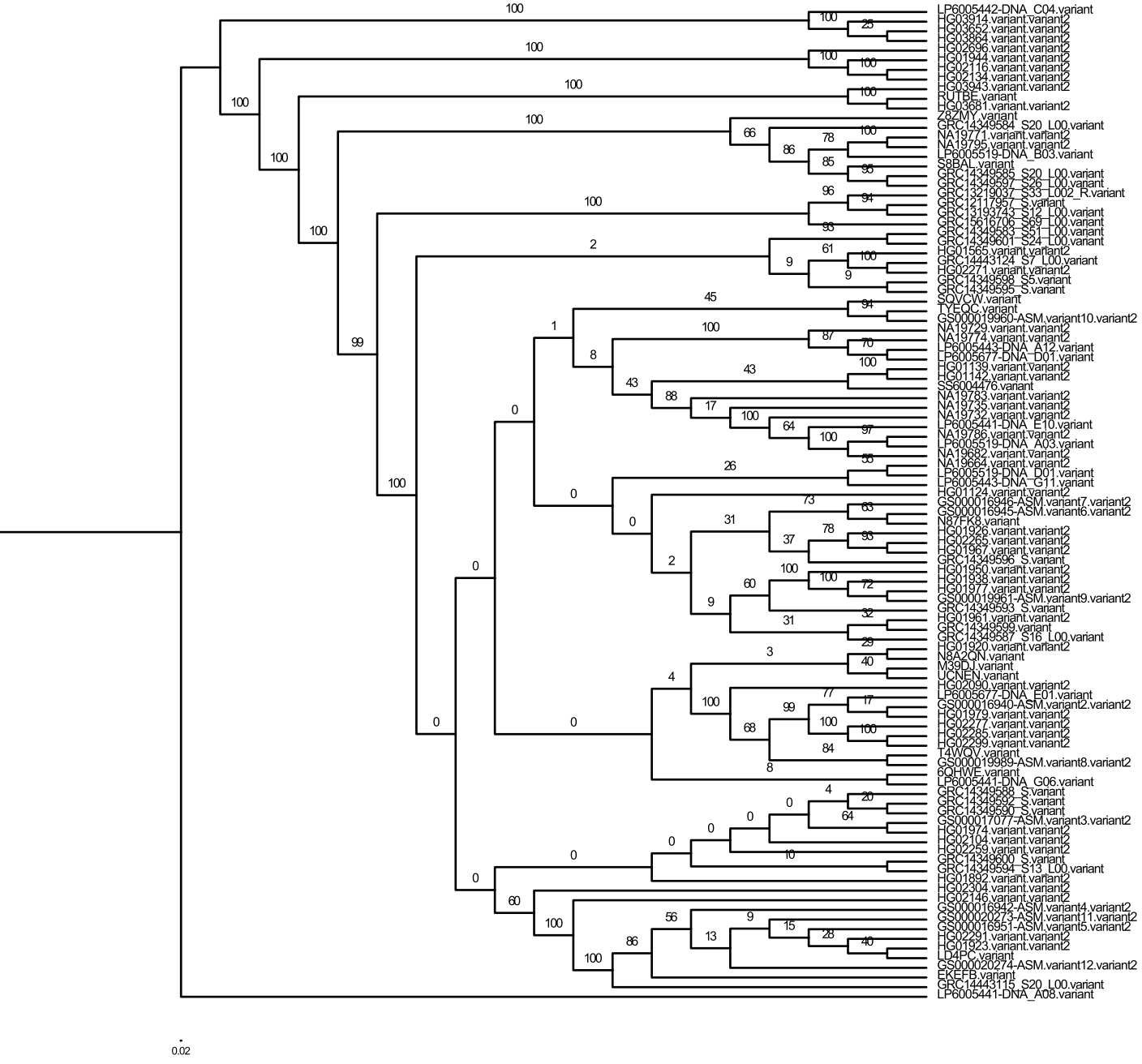
Save the current file

```
#!/bin/bash

################################################################################
################################ START #########################################
################################################################################


################################  PARAMETERS  ##################################

## Information on the sample and files ##
SAMPLENAME=LD4PC        #IMPORTANT AND REQUIRED - Sample Identifier (no spaces)
FIRSTREADS=LD4PC_HNHYNCCXX_L6_1.clean.fq        #IMPORTANT AND REQUIRED - fastq file containing the 1st reads
from the pair
MATES=LD4PC_HNHYNCCXX_L6_2.clean.fq     #IMPORTANT AND REQUIRED - fastq file containing mate pairs
REGIONS=Y       #IMPORTANT AND REQUIRED -  .bed file with regions information for WES ####### QUE PASA CUANDO
NO TENEMOS? DEBERÍA SACARSE LA OPCIÓN -L DE HAPLOTYPECALLER

PREFIX=$SAMPLENAME
PREFIX_QC=$PREFIX\_QC
LOG=$PREFIX\_log

##### READ GROUPS #####
LB=library      # RGLB=String   REQUIRED Library.
PL=Illumina     # RGPL=String   REQUIRED platform (e.g. illumina, solid). IMPORTANTE para GATK: no puede
ponerse UNKNOWN por más que lo muestre como posible. Debe especificarse una plataforma, la marca en general
(ILLUMINA y no HISEQ200, por ejemplo).
SM=LD4PC
# RGSM=String   REQUIRED sample name Required.
PU=1    # RGPU=String   REQUIRED platform unit (eg. run barcode).
ID=1    # RGID=String   ID Default value: 1. This option can be set to 'null' to clear the default value.
CN=null # RGCN=String   sequencing center name Default value: null.
DS=null # RGDS=String   description Default value: null.
DT=null # RGDT=Iso8601Date      run date Default value: null.
PI=null # RGPI=INTEGER  predicted insert size Default value: null.

### PATHs ###
HGREF=/home/marina/Paula/Reference_seq/HGREF    #Reference genome/GATK bundle directory do not change if
already set in PATH

### REFERENCE FILES USED FOR ANALYSIS ###
REF=$HGREF/human_g1k_v37_decoy.fasta    #reference genome .fasta file
DBSNP=$HGREF/dbsnp_137.b37.vcf   #dbsnp .vcf file
KGINDELS=$HGREF/1000G_phase1.indels.b37.vcf
MILLSINDELS=$HGREF/Mills_and_1000G_gold_standard.indels.b37.vcf
KSNP=$HGREF/1000G_phase1.snps.high_confidence.b37.vcf

### OPTION'S VALUES ###
QFILTER_READS=28        #Phred score value to filter reads in PrinSeq.
MINPRUN=3       # minPruning sets the kmer size for the BRUIJN tree in variant calling.
CALL=20.0       # Variant Call Confidence when creating .vcf file.

### PERFORMANCE VALUES ###
XMX=Xmx8g  #Java heap max size
GATK_NT=8 #Num of data threads sent to processor, http://gatkforums.broadinstitute.org/discussion/1975/how-can-
i-use-parallelism-to-make-gatk-tools-run-faster
GATK_NCT=4 #Num of CPU threads for each data thread, http://gatkforums.broadinstitute.org/discussion/1975/how-
can-i-use-parallelism-to-make-gatk-tools-run-faster

############ DO NOT EDIT BEYOND THIS LINE (unless you know what you are doing) ############
#-------------------------------------------------------------------------------------------

############################    COMMANDS     ############################


mkdir -p $LOG   # make directory for log files

clear

echo "Running Sample $SM, $ID, $FASTQC"
#Quality control. Input: 2 fastq files (one per mate); output: 2 html files
perl /home/marina/Paula/FASTQC/FastQC/fastqc $FIRSTREADS $MATES --extract
```

ANEXO VIII - Árbol consensus



ANEXO IV - Resultados del Bootstrap



ANEXO X - SNPs validados

	Nombre utilizado	SNP		Posición SNP		Amplificado por PCR	Confirmación por Sanger	ID
		ref	alt	GRCh37	GRCh38			
RUTBE	Rutbe_1	C	G	6931449	7063408	SI	SI	GMP1
	Rutbe_2	A	T	9023670	9186061	SI	SI	GMP2
	Rutbe_3	G	T	14169926	12049220	SI	SI	GMP3
	Rutbe_4	T	C	15479160	13367280	SI	SI	GMP4
	Rutbe_5	T	G	7540846	7672805	NO	NO	NO
	Rutbe_6	C	T	8385051	8517010	SI	SI	GMP5
	Rutbe_7	A	G	8702379	8834338	SI	SI	GMP6
	Rutbe_8	A	G	8869476	9001435	SI	SI	GMP7
	Rutbe_9	C	T	14877756	12765826	SI	SI	GMP8
	Rutbe_10	G	A	19060346	16948466	SI	SI	GMP9
Z8ZMY/S8BAL	Z8/S8	T	C	7948308	8080267	SI	SI	GMP10
	S8BAL_1	T	G	2875332	3007291	SI	SI	GMP11
	S8BAL_2	T	C	7828787	7960746	SI	SI	GMP12
	S8BAL_3	A	G	21084351	18922465	SI	NO	NO
	Z8ZMY_1	G	A	7868875	8000834	SI	SI	GMP13
	Z8ZMY_2	A	T	7905270	8037229	SI	SI	GMP14
SQVCW	SQVCW_1	G	T	6656300	6788259	SI	SI	GMP15
	SQVCW_2	C	A	6921840	7053799	SI	SI	GMP16
	SQVCW_3	A	G	7779334	7911293	SI	NO	NO
	SQVCW_4	C	T	8023779	8155738	SI	SI	GMP17
	SQVCW_5	T	C	8036434	8168393	SI	SI	GMP18
	SQVCW_6	A	T	8594266	8726225	SI	SI	GMP19
	SQVCW_7	T	G	15261236	13149320	SI	SI	GMP20
TYEQC	TYEQC_1	C	G	14246232	12125526	SI	SI	GMP21
	TYEQC_2	T	C	23583455	21421569	SI	SI	GMP22
	TYEQC_3	C	T	8601728	8733687	SI	SI	GMP23
	TYEQC_4	C	G	8030403	8162362	SI	NO	NO
	TYEQC_5	T	A	7748881	7880840	SI	SI	GMP24
	TYEQC_6	A	T	7588274	7720233	SI	NO	NO
	TYEQC_7	G	A	7383562	7515521	SI	SI	GMP25
N87FK8	N87FK8_1	T	C	14198967	12078261	SI	SI	GMP26
	N87FK8_2	C	A	23567702	21405816	SI	NO	NO
	N87FK8_3	A	G	23592805	21430919	SI	SI	GMP27
	N87FK8_4	T	C	23962077	21815930	SI	SI	GMP28
	N87FK8_5	G	T	23984584	21838437	SI	SI	GMP29
	N87FK8_6	T	C	9143566	9305957	SI	SI	GMP30
	N87FK8_7	G	C	17263815	15151935	SI	NO	NO
	N87FK8_8	C	T	23247806	21085920	SI	SI	GMP31
	N87FK8_9	C	G	23765443	21603557	SI	SI	GMP32
	N87FK8_10	T	G	23575633	21413747	SI	SI	GMP33

ANEXO X - SNPs validados

	Nombre utilizado	SNP		Posición SNP		Amplificado por PCR	Confirmación por Sanger	ID
		ref	alt	GRCh37	GRCh38			
N8A2QN	N8A2QN_1	C	T	2749149	2881108	SI	SI	GMP34
	N8A2QN_2	A	G	2804456	2936415	SI	NO	NO
	N8A2QN_3	G	C	7134535	7266494	SI	NO	NO
	N8A2QN_4	G	T	7267390	7399349	SI	SI	GMP35
	N8A2QN_5	G	T	8131788	8263747	SI	SI	GMP36
	N8A2QN_6	T	C	14886685	12774751	SI	SI	GMP37
	N8A2QN_7	C	T	15954669	13842789	SI	SI	GMP38
	N8A2QN_8	C	G	17351850	15239970	SI	SI	GMP39
	N8A2QN_9	C	T	18830601	16718721	SI	NO	NO
	N8A2QN_10	G	A	19087030	16975150	SI	SI	GMP40
M39DJ/UCNEN	MD39_1	A	T	6631920	6763879	SI	SI	GMP41
	MD39_2	C	G	7571644	7703603	SI	SI	GMP42
	MD39_3	G	A	7765120	7897079	SI	SI	GMP43
	MD39_4	A	C	8359844	8491803	SI	SI	GMP44
	MD39_5	A	G	8560447	8692406	SI	SI	GMP45
	UCNEN_1	G	T	6965772	7097731	SI	SI	GMP46
	UCNEN_2	T	G	7419588	7551547	SI	SI	GMP47
	UCNEN_3	G	C	8133490	8265449	SI	SI	GMP48
	UCNEN_4	C	T	8440075	8572034	SI	SI	GMP49
	UCNEN_5	C	T	14044033	11923327	SI	SI	GMP50
T4WQV	T4WQV_1	A	G	14587968	12476168	SI	SI	GMP51
	T4WQV_2	T	A	6678425	6810384	SI	SI	GMP52
	T4WQV_3	C	G	7353313	7485272	SI	SI	GMP53
	T4WQV_4	C	T	7566319	7698278	SI	NO	NO
	T4WQV_5	C	A	7673168	7805127	SI	SI	GMP54
	T4WQV_6	A	T	7848322	7980281	SI	SI	GMP55
	T4WQV_7	A	G	7887814	8019773	SI	SI	GMP56
	T4WQV_8	C	T	8251637	8383596	SI	SI	GMP57
	T4WQV_9	T	G	8446496	8578455	SI	SI	GMP58

ANEXO X - SNPs validados

	Nombre utilizado	SNP		Posición SNP		Amplificado por PCR	Confirmación por Sanger	ID
		ref	alt	GRCh37	GRCh38			
6QHWE	6QHWE_1	T	G	2747337	2879296	SI	SI	GMP59
	6QHWE_2	C	T	6904459	7036418	SI	SI	GMP60
	6QHWE_3	G	C	7644074	7776033	SI	SI	GMP61
	6QHWE_4	C	T	7893507	8025466	SI	NO	NO
	6QHWE_5	G	A	8051637	8183326	SI	SI	GMP62
	6QHWE_6	C	A	8539196	8671155	SI	SI	GMP63
	6QHWE_7	G	C	8833006	8964965	SI	SI	GMP64
LD4PC/EKEFB	LD/EK_1	T	C	6793301	6925260	SI	NO	NO
	LD/EK_2	A	G	7218975	7350934	SI	SI	GMP65
	LD/EK_3	C	G	14754418	12642487	SI	SI	GMP66
	LD/EK_4	T	C	16835476	14723596	SI	SI	GMP67
	LD/EK_5	G	T	23748402	21586516	SI	SI	GMP68
	LD/EK_6	C	T	23785274	21623388	SI	SI	GMP69
	LD/EK_7	C	T	8806607	8938566	SI	SI	GMP70
	LD/EK_8	A	T	14196672	12075966	SI	SI	GMP71
	LD/EK_9	T	A	23987283	21841136	SI	SI	GMP72

ANEXO XI - Dataciones utilizadas para la construcción de la figura 4.2

SNP	TMRCA (límite inferior - límite superior)	Fuente de Referencia
Q-M242	34.7 kya (30.7-39.4)	Presente Estudio
Q-CTS97	31.6 kya (27.9-35.8)	Pinotti et. al. 2019
Q-L275	26 kya (27.8 - 32.5)	Pinotti et. al. 2019
Q-Y2121	2.5 kya (1.7-3.3)	Grugni et. al. 2019
Q-L68.2	6.1 kya (4.9-7.3)	Grugni et. al. 2019
Q-F1096	19.3 kya (16.7-21.9)	Grugni et. al. 2019
Q-M25	12.4 kya (10.2-14.6)	Grugni et. al. 2019
Q-F746	20.7 kya (18.3-23.5)	Pinotti et. al. 2019
Q-Y521	1.9 kya (1.2-2.6)	Grugni et. al. 2019
Q-M346	25 kya (22-28.3)	Presente Estudio
Q-B28 / Q-F4674	16.8 kya (14.7-18.9)	Grugni et. al. 2019
Q-Z36057	asterisco verde (no se encontró datado)	
Q-L54	24.9 kya (22-28.2)	Presente Estudio
Q-Z780	19.3 kya (17-21.9)	Presente Estudio
Q-Z781	19.3 kya (17-21.9)	Presente Estudio
Q-Y2816	asterisco verde (no se encontró datado)	
Q-Z782	9.63 kya (8.51-10.9)	Pinotti et. al. 2019
Q-YP937	18.7 kya (16.5-21.2)	Presente Estudio
Q-GMP73	18.2 kya (16.1-20.6)	Presente Estudio
Q-M930	15.6 kya (13.8-17.7)	Presente Estudio
Q-L804	4.4 kya (3.9-5.0)	Presente Estudio
Q-M3	15.4kya (13.6-17.4)	Presente Estudio
Q-M848	15.4 kya (13.6-17.4)	Presente Estudio
Q-MPB118	9.7 kya (8.5-11)	Presente Estudio
Q-SK281	7.8 kya (6.3-9.3)	Grugni et. al. 2019
Q-Z35727	asterisco verde (no se encontró datado)	
Q-MPB139	14 kya (12.4-15.9)	Presente Estudio
Q-B42	14.2 kya (12.6-16.2)	Presente Estudio
Q-B46	asterisco verde (no se encontró datado)	
Q-Z35497	9.6 kya (8.4-10.8)	Presente Estudio
Q-Z6658	7.7 kya (9.2-6.2)	Grugni et. al. 2019
Q-CTS2731	9.2 kya (7.5-10.9)	Grugni et. al. 2019
Q-Y26467	0.6 kya (0.53-0.68)	Presente Estudio
Q-CTS11357	11.3 kya (10.3-13.2)	Presente Estudio
Q-CTS11330	8.4 kya (7.4-9.6)	Presente Estudio
Q-Y27993	9.6 kya (8-11.2)	Grugni et. al. 2019
Q-Z19357	8.1 kya (9.5-6.7)	Grugni et. al. 2019
Q-MPB016	11.2 kya (9.9-12.7)	Presente Estudio
Q-Z5908	13.6 kya (12.0-15.4)	Presente Estudio
Q-Z35841	asterisco verde (no se encontró datado)	
Q-Z5906	12.88 kya (11.38-14.57)	Pinotti et. al. 2019
Q-GMP70	2.4 kya (2.1-2.7)	Presente Estudio
Q-Z5907	1.7 kya (1.5-1.9)	Presente Estudio

ANEXO XII - Sub-linajes de Q-M848 que abarcan ~12.800 cal AP

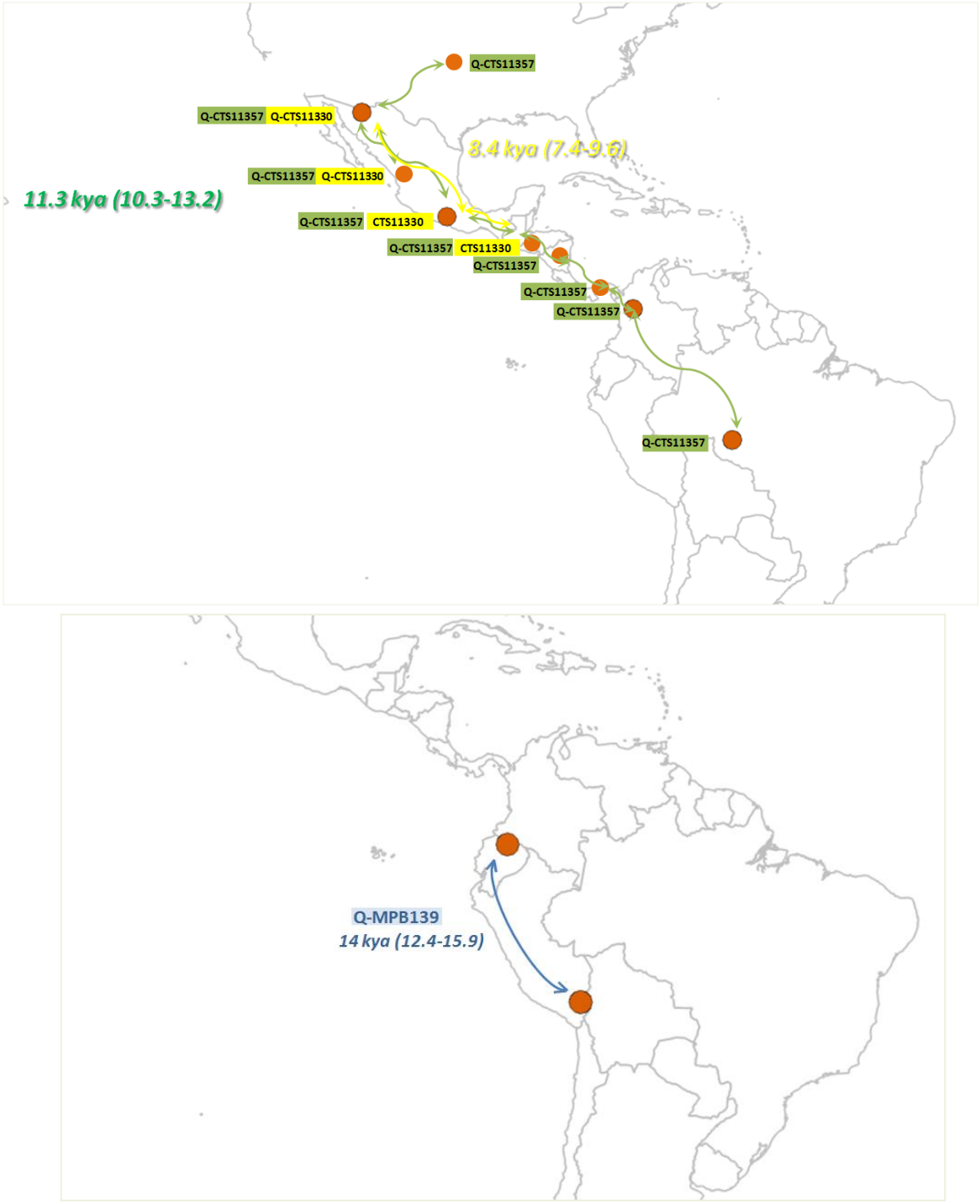


Figura anexa X: Sub-linajes de Q-M848 que abarcan ~12.800 cal AP. Profundidad temporal, dispersión geográfica y diferenciación regional. El color de las rutas migratorias coincide con color del SNP representado. Las rutas migratorias representadas en diferentes colores son tentativas. El color de la datación corresponde al SNP representado con el mismo color. Los individuos de Los Ángeles con origen mexicano fueron representados en México.

ANEXO XII - Sub-linajes de Q-M848 que abarcan ~12.800 cal AP

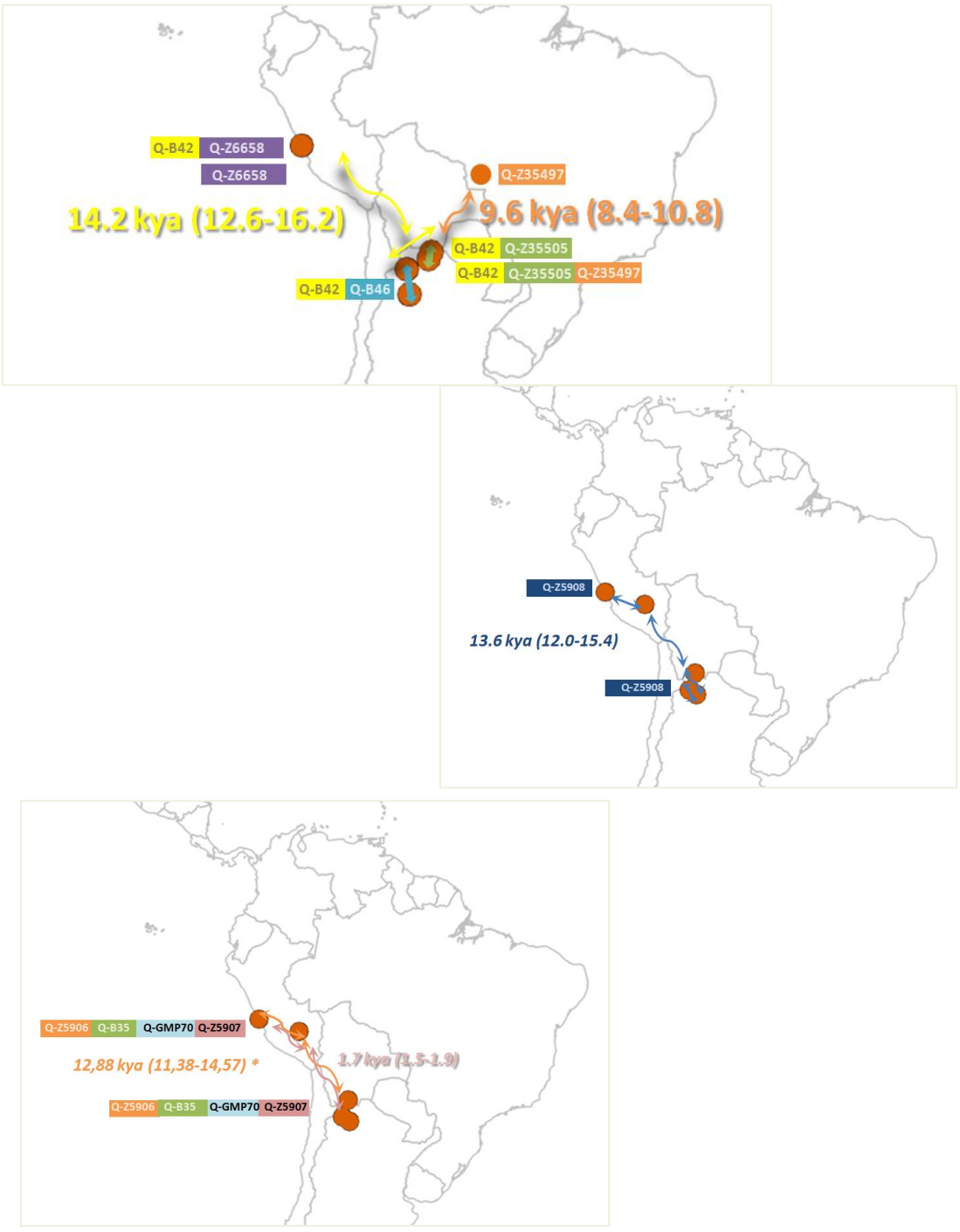


Figura anexa X: Sub-linajes de Q-M848 que abarcan ~12.800 cal AP. Profundidad temporal, dispersión geográfica y diferenciación regional. El color de las rutas migratorias coincide con color del SNP representado. Las rutas migratorias representadas en diferentes colores son tentativas. El color de la datación corresponde al SNP representado con el mismo color. El asterisco naranja representa datación extraída de Pinotti y col. 2019. Los colores con los que se representan las dataciones respetan el color del SNP indicado con el mismo color.

ANEXO XIII - Sub-linajes de Q-M848 posteriores a ~12.800 cal AP

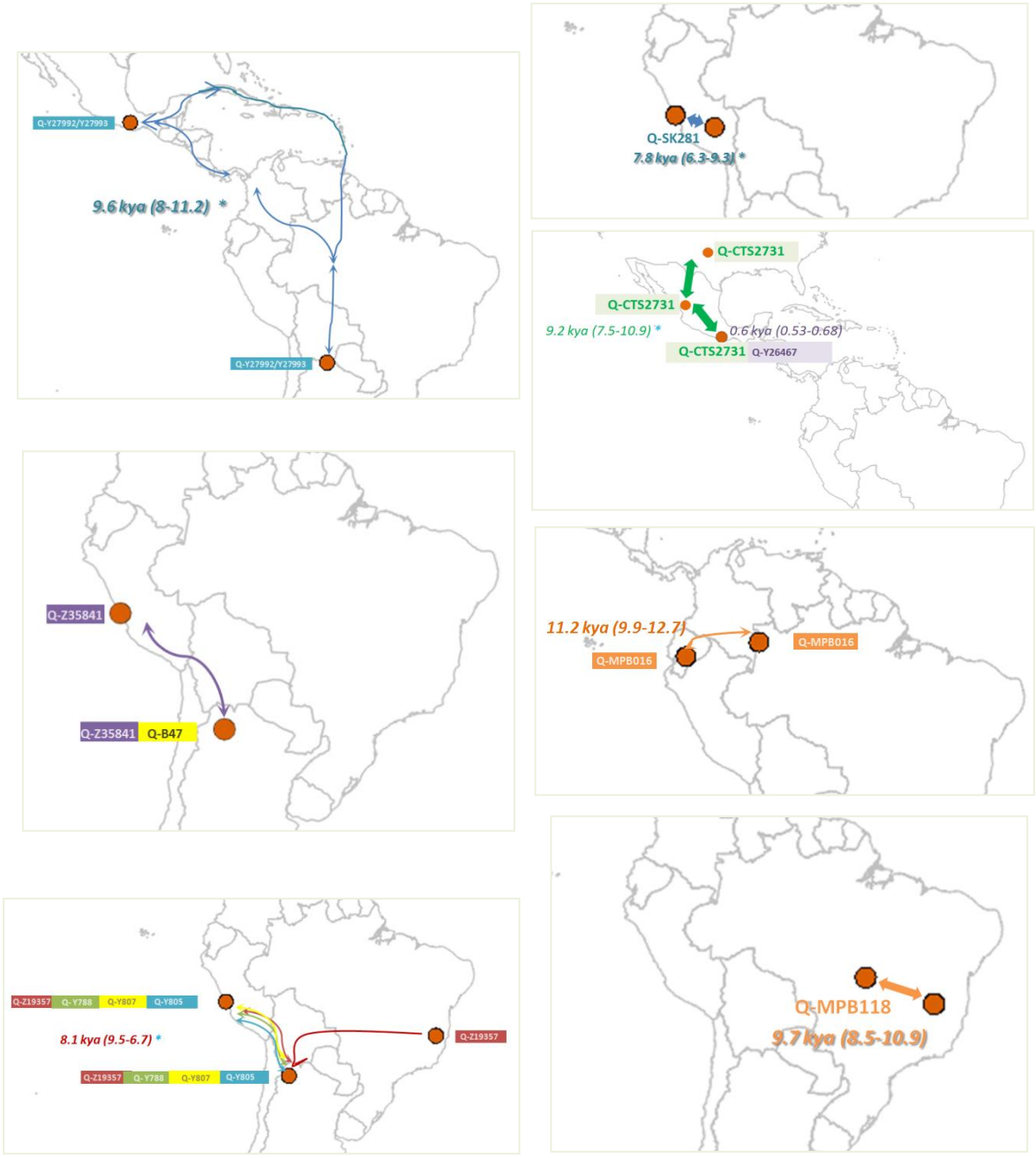


Figura anexa XI: Sub-linajes de Q-M848 posteriores ~12.800 cal AP. Profundidad temporal, dispersión geográfica y diferenciación regional. El color de las rutas migratorias coincide con color del SNP representado. Las rutas migratorias representadas en diferentes colores son tentativas. El color de la datación corresponde al SNP representado con el mismo color. El asterisco azul representa dataciones extraídas de Grugni y col. 2019. Los individuos de Los Ángeles con origen mexicano fueron representados en México. No se encontró una datación para Q-Z35841, por lo que podría ser anterior a ~12.800 cal AP o posterior.

ANEXO XIV - LISTA DE ABREVIATURAS

: cuando en este trabajo aparece para Q-M346, indica, derivado para Q-M346 y ancestral para Q-L54.

~: aproximadamente

A.C: años antes de Cristo

AP: años antes del presente

cal A.C.: años calibrados antes de Cristo

cal AP: años calibrados antes del presente

cal D.C.: años calibrados después de Cristo

D.C: años después de Cristo

GATK: Genome Analysis Toolkik (Kit de herramientas de análisis del genoma)

IPCB: Consejo de Pueblos Indígenas sobre Biocolonialismo (en inglés, Indigenous Peoples Council on Biocolonialism)

kya: (del inglés, kilo years ago, miles de años atrás)

Mb: Megabases, equivale a un millón de bases

MSA: alineación múltiple de secuencias (en inglés, multiple sequence alignment)

MSY: región masculina específica del cromosoma Y

NGS: técnica de secuenciación de próxima generación (en inglés, next-generation sequencing)

NRY: como región no recombinante del cromosoma Y

pb: pares de bases

PCR-APLP: técnica de la reacción en cadena de la polimerasa-polimorfismos de la longitud de productos

RFLP: técnica de polimorfismos en la longitud de los fragmentos de restricción (en inglés, Restriction Fragment Length Polymorphism)

SBS: secuenciación por síntesis química (en inglés, sequencing-by-synthesis)

SNP: polimorfismo de un solo nucleótido (en inglés, Single Nucleotide Polymorphism)

STR: repeticiones cortas de ADN en tándem (en inglés, Short tandem repeats)

XDG: la clase X-degenerada del cromosoma Y

XTR: la clase X-transpuesta

YCC: Consorcio del Cromosoma Y (en inglés, *Y Chromosome Consortium)*

YD: Younger Dryas

YDB: capa límite de YD (en inglés, Younger Dryas Boundary)

Y-SNPs: SNPs del cromosoma Y