



UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO
DOCTORADO EN CIENCIAS BIOMÉDICAS

Laboratorio Internacional de Investigación sobre el Genoma Humano
(LIIGH UNAM)

Estudio Paleogenómico de México Prehispánico

TESIS

QUE PARA OPTAR POR EL GRADO DE:
DOCTORA EN CIENCIAS

PRESENTA:

Viridiana Villa Islas

DIRECTORA DE TESIS:

DRA. MARÍA C. ÁVILA ARCOS (LIIGH-UNAM)

COMITÉ TUTOR:

DR. JULIO COLLADO VIDES (CCG-UNAM)

DR. SAMUEL CANIZALES QUINTEROS (FQ-CIENCIAS)

Juriquilla, Querétaro, 2023



Universidad Nacional
Autónoma de México



UNAM – Dirección General de Bibliotecas
Tesis Digitales
Restricciones de uso

DERECHOS RESERVADOS ©
PROHIBIDA SU REPRODUCCIÓN TOTAL O PARCIAL

Todo el material contenido en esta tesis esta protegido por la Ley Federal del Derecho de Autor (LFDA) de los Estados Unidos Mexicanos (México).

El uso de imágenes, fragmentos de videos, y demás material que sea objeto de protección de los derechos de autor, será exclusivamente para fines educativos e informativos y deberá citar la fuente donde la obtuvo mencionando el autor o autores. Cualquier uso distinto como el lucro, reproducción, edición o modificación, será perseguido y sancionado por el respectivo titular de los Derechos de Autor.



UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO
DOCTORADO EN CIENCIAS BIOMÉDICAS

Laboratorio Internacional de Investigación sobre el Genoma Humano
(LIIGH UNAM)

Estudio Paleogenómico de México Prehispánico

TESIS

QUE PARA OPTAR POR EL GRADO DE:
DOCTORA EN CIENCIAS

PRESENTA:

Viridiana Villa Islas

DIRECTORA DE TESIS:

DRA. MARÍA C. ÁVILA ARCOS (LIIGH-UNAM)

COMITÉ TUTOR:

DR. JULIO COLLADO VIDES (CCG-UNAM)

DR. SAMUEL CANIZALES QUINTEROS (FQ-CIENCIAS)

Juriquilla, Querétaro, 2023

Agradecimientos	1
Abreviaturas	3
Resumen	4
Abstract	6
1 Introducción	8
1.1 ADN antiguo y Paleogenómica	8
1.1.1 Características del ADN	9
1.2 Estudio de poblaciones humanas a través del ADN	11
1.2.1 Marcadores uniparentales	11
1.2.1.1 ADN mitocondrial	12
1.2.1.2 Cromosoma Y	14
1.2.2 Estudio del genoma autosomal	15
1.3 Estudios de ADN humano en México	16
2 Hipótesis y Objetivos	20
2.1 Hipótesis	20
2.2 Objetivo principal	20
2.2.1 Objetivos particulares	20
3 Métodos de laboratorio para el análisis de ADN antiguo	23
3.1 Infraestructura y medidas para prevenir la contaminación	23
3.2 Procesamiento de las muestras arqueológicas	25
3.2.1 Extracción de ADN	26
3.2.2 Construcción de bibliotecas genómicas	27
3.2.3 Procedimiento de captura genómica en solución	29
4 Revisión de métodos para el análisis de datos paleogenómicos	31
4.1 Formato FASTQ	32
4.2 Eliminación de los adaptadores de secuenciación	33
4.3 Mapeo contra genoma humano de referencia	34
4.3.1 Control de calidad del mapeo	37
4.4 Patrones de daño	37
4.5 Estimación de contenido endógeno	38
4.6 Clonalidad	39
4.7 Cobertura del genoma	40
4.8 Complejidad de las bibliotecas genómicas	40
4.9 Tuberías de análisis de datos de ADN	41
4.10 Determinación del sexo molecular	42
4.11 Análisis de marcadores uniparentales	44
4.11.1 Detección de variantes en el genoma mitocondrial	44
4.11.1.1 Llamado de variantes	44
4.11.1.2 Secuencia de consenso (fasta)	44
4.11.1.3 Archivo Hsd	45

4.11.2	Asignación del haplogrupo mitocondrial	46
4.11.3	Asignación del haplogrupo de cromosoma Y	47
4.12	Determinación de las relaciones de parentesco entre individuos	47
4.13	Detección de contaminación por ADN moderno	50
4.13.1	Métodos de contaminación basados en el genoma mitocondrial	50
4.13.2	Métodos de contaminación basados en el genoma nuclear	52
4.13.3	Métodos de contaminación basados en el cromosoma X	53
4.14	Eliminación de la contaminación de los datos de ADN	53
4.15	Comparación de genomas de ADN con paneles de referencia actuales	54
4.15.1	Afinidades genéticas	55
4.15.1.1	Análisis de componentes principales (PCA)	56
4.15.1.2	Componentes ancestrales	57
4.15.2	Deriva genética compartida	59
4.15.2.1	TreeMix	59
4.15.2.2	Prueba de outgroup-f3	59
4.15.3	Detección de mezcla genética	60
4.16	Diversidad genética	61
4.17	Análisis demográfico	64
4.18	Continuidad poblacional	65
4.19	Faseo de secuencias de ADN	66
5	El poblamiento de América estudiado a través de la Paleogenómica	68
5.1	Modelo Paleoamericano	75
5.2	Migración hacia América desde Asia	76
5.2.1	Paleo Inuit e Inuit	79
5.3	Caribe	80
5.4	Población Y	82
5.5	Contacto con Polinesios	83
5.6	Conclusión	86
6	Presentación del artículo de investigación “Demographic history and genetic structure in pre-Hispanic Central Mexico”	87
6.1	Prólogo	87
6.2	Introducción	88
6.3	Resumen de objetivos y resultados	89
6.3.1	Obtención de ADN mitocondrial y nuclear a partir de restos arqueológicos prehispanicos y secuenciación mediante NGS	89
6.3.2	Identificación de linajes uniparentales	90
6.3.3	Análisis de las afinidades genéticas en el contexto de la diversidad actual de las poblaciones indígenas de México	92
6.3.4	Estimación de la diversidad genética en las poblaciones prehispanicas	95
6.3.5	Análisis de continuidad genética poblacional en la Sierra Gorda de Querétaro (320 aC – 1521 dC)	96
6.3.6	Contribución genética fantasma en población prehispanica de México	102

6.4 Conclusiones	104
7 Implicaciones éticas en la Paleogenómica	105
7.1 Dinámicas de colaboración	105
7.2 Sustentabilidad	110
7.3 Relación con poblaciones indígenas y permiso para la obtención de las muestras	111
7.4 Accesibilidad a los datos de ADN	116
8 Discusión	118
8.1 Procesos de laboratorio	119
8.2 Análisis bioinformático de los datos de ADN prehispanico	120
8.3 Aspectos éticos de la investigación doctoral	122
9 Conclusiones	123
10 Perspectivas	125
11 Literatura citada	129
12 Anexos	143
Anexo I. Artículo de investigación como requisito para obtención de grado 'Demographic history and genetic structure in pre-Hispanic Central Mexico'	143
Anexo II. Contribución al artículo 'Population History and Gene Divergence in Native Mexicans Inferred from 76 Human Exomes'	229
Anexo III. Contribución al artículo 'Paleogenomic insights into the red complex bacteria Tannerella forsythia in Pre-Hispanic and Colonial individuals from Mexico'	230
Anexo IV. Contribución al artículo 'Ancient viral genomes reveal introduction of human pathogenic viruses into Mexico during the transatlantic slave trade'	231
Anexo V. Contribución al artículo 'Indigenous peoples in eastern Brazil: insights from 19th century genomes and metagenomes'	232

Agradecimientos

Quisiera agradecer a todas las personas que me han apoyado durante la realización de mi doctorado.

En primer lugar quiero agradecer a mis padres, a mis hermanos, a Sofi y a Caro quienes siempre han estado conmigo, me han dado su cariño y me han mostrado su apoyo incondicional para seguir mis metas. ¡Gracias por ser los mejores!

Agradezco a la Dra. María C. Ávila Arcos quién ha sido una excelente tutora y mentora. Gracias por haber confiado en mí desde el inicio para llevar a cabo este proyecto ambicioso, interesante y gratificante. Ha sido un placer trabajar contigo.

También quiero agradecer a la Dra. Alejandra Medina, la Dra. Daniela Robles, la Dra. Lucía Morales y el Dr. Diego Ortega, del LIIGH; al Dr. Federico Sánchez del INMEGEN y a la Dra. Marcela Sandoval, de quienes recibí constante retroalimentación acerca de mi trabajo desde el inicio y a lo largo de mi proyecto, ayudando así a mejorarlo.

Gracias a la Dra. Emilia Huerta por recibirme en su grupo para realizar una estancia de investigación y llevar a cabo algunos de los análisis bioinformáticos, fue muy enriquecedor. Gracias a las y los colaboradores arqueólogos que confiaron en este proyecto y nos compartieron su conocimiento y material arqueológico para llevar a cabo este estudio.

Agradezco también a los miembros de mi comité tutorial quienes a través de sus evaluaciones y retroalimentación han enriquecido mi formación como investigadora. También agradezco a mis compañeras de laboratorio Miriam Bravo y Camila Tamburrini quienes con sus consejos y apoyo hicieron más sencillas las cuestiones académicas y no-académicas.

Agradezco también a mi prima Yazmín Jiménez quien realizó las figuras de las secciones 5 y 7 de esta tesis.

Agradezco también a quienes estuvieron conmigo en el proceso de escritura de esta tesis. A Vanesa Varela y Lorenzo Lunazzi quienes no forman parte de la comunidad científica pero su energía y cariño hicieron de esta etapa un proceso muy lindo y disfrutable. Gracias por cocinar para mí, cuidarme, hacer coworking conmigo y ser la mejor compañía.

Gracias tío Raúl Villa por compartir tu casa en PDC conmigo, y gracias Lorenzo por compartir conmigo tu espacio en PDC e Italia. Fue inspirador escribir desde estos lugares.

Por último quiero agradecer a la beca que recibí por parte de Conacyt (Número de becaria: 282188), de UNAM-PAPIIT (IA203821) y de Human Frontiers Science Program.

¡Gracias a todos por ser parte de esta etapa!

Abreviaturas

ADN: Ácido desoxirribonucleico

ADNa: ADN antiguo

ADNmt: ADN mitocondrial

AP: Antes del Presente

BCE: Antes de la era común

CE: Era común

mtC: Captura de genoma mitocondrial

NGS: Secuenciación de siguiente generación

NRY: Región no recombinante del cromosoma Y

PCR: Reacción en cadena de la polimerasa

RFLP: Reacción de Restricción de Fragmentos para determinación de Polimorfismos

SNV: Variación de un sólo nucleótido

WGC: Captura de genoma completo

Resumen

El registro histórico contenido en nuestro genoma es una fuente de conocimiento invaluable que nos permite inferir eventos demográficos del pasado. Ahora con la paleogenómica, es incluso posible conocer el pasado genético directamente al analizar el ADN de individuos que vivieron hace cientos o miles de años. El vasto registro arqueológico que existe en México es evidencia de la diversidad de poblaciones que vivieron en el territorio mexicano antes de la llegada de los colonizadores europeos. Por lo que, es de especial interés conocer cómo era la genética de estas poblaciones nativas previo al contacto con poblaciones de otros continentes. Dicho conocimiento nos permitiría saber más acerca de su demografía y podríamos explorar la diversidad genética que pudo haberse perdido después de la colonización europea. La presente tesis es una contribución inicial al campo de la Paleogenómica humana en México donde se esperan muchos avances en los próximos años. La presente tesis incluye cuatro capítulos en los que se abordan temas generales de la Paleogenómica humana con el objetivo que el lector tenga acceso a los conocimientos básicos en este campo. El contenido de esta tesis comprende: 1) una introducción a la historia del campo de la paleogenómica a nivel mundial y en México, 2) una descripción de los métodos de laboratorio comúnmente utilizados en el campo, 3) la descripción del amplio repertorio de métodos de análisis bioinformático que se puede realizar con las secuencias de ADN, 4) un capítulo acerca de la evidencia que ha arrojado la paleogenómica para armar el rompecabezas del poblamiento de América y 5) un capítulo acerca de las implicaciones éticas de la investigación paleogenómica y cómo se puede mejorar para llevar a cabo investigación sustentable y soberana, al mismo tiempo que se establezcan colaboraciones equitativas.

Por otro lado, se incluye también un capítulo con los principales resultados de la investigación realizada durante mi trabajo doctoral. Como parte de éste, el lector encontrará anexo el artículo científico publicado. Mediante esta investigación se analizó el ADN de individuos de la época prehispánica de México de distintos sitios y zonas arqueológicas. Para realizar este trabajo se hizo uso de diferentes herramientas de la Paleogenómica tanto de laboratorio como análisis bioinformáticos. Como resultado, se logró: 1) una primera aproximación a cómo era la estructura y diversidad genética en la

época prehispánica tanto a nivel de ADNmt como nuclear. Encontré que la estructura genética se ha mantenido hasta la época actual; 2) se estudió una hipótesis previa de reemplazo poblacional en la Frontera Norte de Mesoamérica en la Sierra Gorda de Querétaro. Dicha hipótesis no es soportada por medio de los datos generados aquí y sugiere una continuidad poblacional y 3) se analizó la contribución genética de poblaciones desconocidas en los genomas de los individuos prehispánicos con mayor cobertura del genoma nuclear. Se encontró que hay dos contribuciones de poblaciones fantasma distinta en individuos prehispánicos de la Sierra Tarahumara y en Cañada de la Virgen. Cabe resaltar que mediante este primer estudio de paleogenómica en México se abre la puerta para seguir llevando a cabo este tipo de investigación en nuestro país cumpliendo con los estándares internacionales mientras se protege nuestro patrimonio histórico nacional.

Además, como anexos se encuentran también los aportes que realicé durante mi doctorado a otros artículos científicos.

Abstract

The historical record contained in our genome is an invaluable source of knowledge that allows us to infer past demographic events. Now with Paleogenomics, it is even possible to know the genetic past directly by analyzing the DNA of individuals who lived hundreds or thousands of years ago. The vast archaeological record in Mexico is evidence of the diversity of populations that lived in the Mexican territory before the arrival of the European colonizers. Therefore, it is of special interest to know what the genetics of these native populations were like prior to contact with populations from other continents, since it would allow us to know more about their demography and we could explore the genetic diversity that may have been lost after European colonization. This thesis is an initial contribution to the field of human Paleogenomics in Mexico which is expected to advance in the coming years. The present thesis includes five chapters. Four of these chapters address general topics in human Paleogenomics with the objective of introducing the reader to basic knowledge in this field. The content includes: 1) an introduction to the history of the field of Paleogenomics worldwide and in Mexico, 2) a description of the laboratory methods commonly used in the field, 3) a description of the vast repertoire of bioinformatics analysis methods that can be performed with aDNA sequences, 4) a chapter on the evidence that Paleogenomics has yielded for piecing together the puzzle of the peopling of America, and 5) a chapter on the ethical implications of Paleogenomics research and how it can be improved to conduct sustainable and sovereign research while establishing equitable collaborations.

On the other hand, a chapter with the main results of the research carried out during the doctoral studies is also included. As part of this chapter, readers can find the scientific article attached. During this research, the aDNA of individuals from the pre-Hispanic period in Mexico -from different archaeological sites and areas- were analyzed using Paleogenomics tools that encompass both wet lab and bioinformatic approaches. As a result we achieved: 1) a first description of the genetic structure and diversity in pre-Hispanic times, both at mtDNA and nuclear level, in which it was found that the genetic structure has been maintained up to the present time; 2) a previous hypothesis of

population replacement in the Northern Border of Mesoamerica in the Sierra Gorda of Querétaro was studied. We found that this hypothesis is not supported by the generated data and we instead find evidence for population continuity in this area, and 3) the genetic contribution of unknown ghost populations were analyzed in the genomes of pre-Hispanic individuals with the greater coverage of the nuclear genome. We found that there are two different contributions of ghost populations in prehispanic individuals from the Sierra Tarahumara and Cañada de la Virgen. It is worth noting that this is the first paleogenomics study in Mexico and opens the door to continue carrying out this type of research in our country in compliance with international standards while also protecting our national heritage.

In addition, you will also find my contributions during my doctoral studies to other scientific articles in the Appendix.

1 Introducción

1.1 ADN antiguo y Paleogenómica

El ADN antiguo (ADNa) se define como el material genético recuperado de una muestra que no se ha conservado ni tomado con el fin de extraer ácidos nucleicos. Por lo tanto, el campo del ADNa también puede incluir muestras obtenidas para el análisis forense (1).

El primer estudio publicado que reportó la obtención y secuenciación de ADNa fue realizado por investigadores de la Universidad de California, Berkeley, EUA en 1984 (2). Los investigadores tenían como objetivo comprobar si era posible obtener moléculas de ADN a partir de restos de especies extintas. En dicho trabajo extrajeron ADN de tejido momificado de 140 años de antigüedad de un quagga (*Equus quagga*), una especie extinta en 1883 que comparte un ancestro común con las cebras actuales. Higuchi y colaboradores tuvieron éxito y recuperaron ADNa mitocondrial de esta especie; éste fue clonado a través de vectores de clonación y secuenciado mediante el método de Sanger con lo cual se realizó la primera filogenia que contenía datos de ADNa (2). Al final de su artículo los autores plantean la posibilidad de extraer ADN a partir de insectos preservados en ámbar de hace 25 millones años. Además, este trabajo plantea que muchas áreas de la ciencia se beneficiarían si la preservación de ADN es un fenómeno general. Algunas de las áreas que se beneficiarían son la paleontología, biología evolutiva, arqueología y las ciencias forenses. Higuchi y colaboradores tenían mucha razón en este último punto más no en el primero. De acuerdo al modelo teórico propuesto por Allentoft y colaboradores en 2012 (3) acerca de la preservación de ADN en muestras antiguas es posible obtener ADN de especímenes de más de un millón de años que se encuentren en ambientes congelados profundos, y es extremadamente improbable recuperarlo de especímenes más antiguos de 80-85 millones de años de antigüedad. La precisión de este modelo teórico fue corroborada en el 2021, en una investigación en la que se logró recuperar el ADN más antiguo hasta el momento; obtenido a partir de un molar de mamut de hace 1.2 millones de años del este de Siberia (4). Con este trabajo se ha expandido enormemente el rango de tiempo experimental del que se pueden recuperar

moléculas de ADN y ha abierto la posibilidad de estudiar la historia evolutiva de muchas especies. Esto incluye las ramas más basales del árbol filogenético del género *Homo*, incluyendo especímenes de *Homo erectus* y *Homo heidelbergensis* de los cuales no se ha recuperado información genómica hasta el momento.

A partir de la publicación de Higuchi y colaboradores en 1984, surgieron diversas investigaciones alrededor del mundo (principalmente en EUA y Europa) para conocer el pasado genético de distintas especies (5–8), entre ellas la especie humana (9, 10) en la que se centra la presente tesis. Dos décadas más tarde, en 2005, con la llegada de las tecnologías de secuenciación de siguiente generación (NGS) se abrió una vía completamente nueva para el estudio del ADN al permitir el análisis de muestras antiguas a escala genómica, dando lugar a la era de la paleogenómica. El campo de la paleogenómica ha avanzado bastante rápido desde entonces. En el año 2010 se publicó el primer genoma completo de humano a partir de una muestra antigua a una cobertura alta (~20x) (11). Este genoma se recuperó a partir de una muestra de cabello de 4,000 años de antigüedad conservada en permafrost, en Groenlandia, lo cual favoreció la buena preservación del material genético de este individuo perteneciente a la cultura Saqqaq (12).

1.1.1 Características del ADN

Desde sus inicios el análisis del ADN ha sido un reto debido a las alteraciones que el material genético acumula *post mortem*. Tras la muerte, se inicia la degradación del material genético debido a la falta de enzimas protectoras que solían reparar el ADN cuando el organismo estaba vivo(13). Además, la exposición ambiental favorece la ruptura del ADN en fragmentos pequeños que llegan a ser de <100 pares de bases (pb) (11), lo que es considerablemente más corto en comparación con la longitud del ADN moderno, que suele ser > 1kbp después de la extracción (Figura 1). La conservación del ADN depende en gran medida de las condiciones ambientales a las que estuvo expuesta una muestra. Si la muestra ha estado en ambientes fríos y de baja humedad, la conservación del material genético se espera que sea buena, mientras que en condiciones de mayor temperatura y humedad, el material genético tiende a estar más degradado o incluso es (o suele ser) indetectable en la muestra (3). Otra característica de las moléculas de ADN es

que, debido a la fragmentación, los extremos de la molécula tienden a presentarse en cadena simple, lo que conlleva a que las bases se encuentren expuestas a daños hidrolíticos. Uno de estos daños es el que se presenta en las Citosinas, las cuales en el extremo 5' se desaminan dando lugar a un cambio de base de Citosina a Uracilo (14). Cuando se construyen bibliotecas genómicas a partir de las moléculas de ADN, durante la fase de reparación de extremos que consiste en completar la doble cadena de los extremos simples, la ADN polimerasa al encontrarse con el Uracilo lo identifica como una Timina en lugar de la Citosina original dando como resultado cambios de transición en las secuencias de C → T en el extremo 5' y de G → A en el extremo 3' (14) (Figura 1). Aunque las moléculas de ADN pueden estar expuestas a otro tipo de daños y cambios en las bases, este cambio de transición es el más común y detectable. Por lo tanto, se ha tomado como un método de autenticación de las moléculas de ADN (14, 15). Aunado a esto, el porcentaje de moléculas de ADN del organismo bajo estudio (ADN endógeno) que puede recuperarse durante la extracción de ADN es muy bajo en comparación con el ADN ambiental (Figura 1) que se asocia a la muestra debido a la diagénesis. Por ello, los laboratorios de ADN deben ser extremadamente cuidadosos durante el proceso de extracción de ADN para evitar cualquier otro tipo de contaminación con ADN moderno (11). En la sección 3 de la presente tesis se detallan las medidas preventivas que deben seguirse para trabajar con ADN.

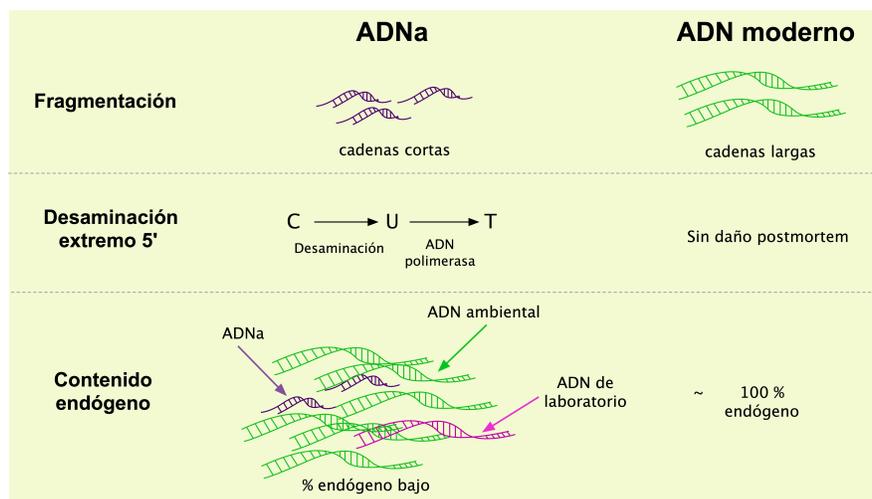


Figura 1. Características del ADNa vs ADN moderno. El ADNa es de longitud corta, presenta cambios de C a T en el extremo 5' y la correspondiente G a A en el extremo 3' (tras la secuenciación), y baja proporción endógena en comparación con el ADN ambiental recuperado durante el proceso de extracción.

1. 2 Estudio de poblaciones humanas a través del ADN

Una de las preguntas más importantes inherentes a la conciencia humana es “¿De dónde venimos?”. Por ellos el estudio de la historia de la humanidad ha tenido un papel importante desde tiempos antiguos. La historia de cada pueblo era heredada de una generación a otra por relatos, dibujos y, más adelante con la invención de la escritura, ésta se escribió en diferentes tipos de materiales (16). A partir de entonces ha sido posible estudiar la historia con este tipo de registros arqueológicos (17). Asimismo, el descubrimiento y análisis de vestigios y restos arqueológicos incluidos los huesos de humanos y sus especies cercanas ha permitido el estudio de épocas más antiguas referentes a la prehistoria, cuando no existían registros escritos. El estudio de los caracteres físicos heredables de los esqueletos humanos permitió la caracterización y clasificación taxonómica del *Homo sapiens* y de otras especies del género *Homo* (18). El análisis morfométrico de los esqueletos humanos alrededor del mundo, en especial del cráneo y la dentadura, hizo posible la identificación de rasgos distintivos en diferentes poblaciones a nivel continental. Con esta información y el registro arqueológico se propusieron algunas hipótesis acerca de migraciones humanas, como las del poblamiento de América y otras partes del mundo (19). Más tarde, con el avance científico, también ha sido posible estudiar las poblaciones humanas desde el enfoque de la herencia biológica a nivel molecular.

1.2.1 Marcadores uniparentales

El descubrimiento del ADN como la molécula de herencia abrió la puerta al estudio del registro histórico contenido en nuestro genoma. Los primeros análisis se centraron en el material genético de herencia uniparental, el ADN mitocondrial y las regiones no recombinantes del cromosoma Y (NRY). Estas moléculas se heredan prácticamente sin recombinación de madres o padres a hijas e hijos, lo que significa que pasan intactos. Específicamente, las madres heredan el ADN mitocondrial (ADNmt) a su descendencia (hombres y mujeres), y los padres heredan su cromosoma Y (chrY) a su descendencia

masculina (Fig.2). Por lo tanto, el ADNmt y el chrY son poderosas herramientas para estudiar la historia de los linajes materno y paterno, respectivamente.

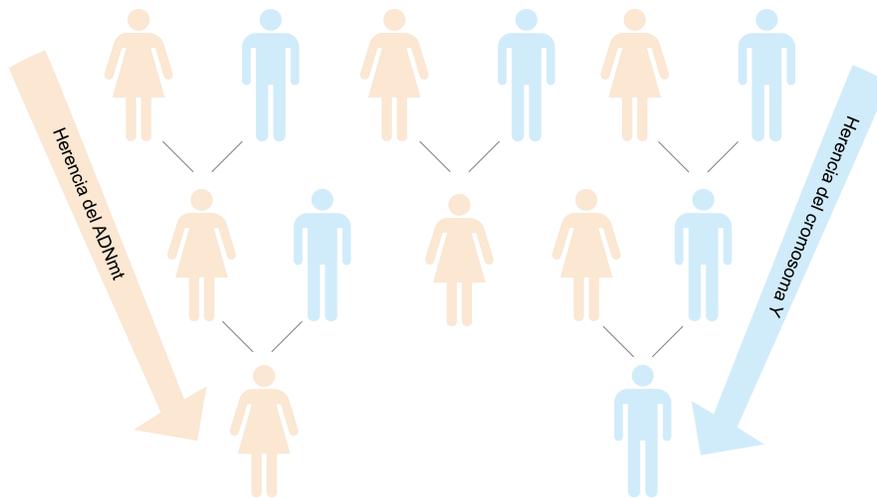


Figura 2. Herencia de marcadores uniparentales.

A través de los análisis del ADNmt y NRY se ha logrado reafirmar o proponer nuevas hipótesis acerca de la historia demográfica de algunas poblaciones e incluso proponer una antigüedad estimada para el ancestro común más reciente (MRCA) femenino que dio origen a la especie humana (Eva mitocondrial), hace 200,000 años, y del MRCA masculino (Y-MRCA) hace 275,000 años, ambos con origen en el continente africano (20); estudios más recientes han determinado que el tiempo al MRCA del linaje femenino y masculino era más reciente, 99,000-148,000 y 120,000-156,000 años, respectivamente (21).

1.2.1.1 ADN mitocondrial

Los primeros estudios que lograron la recuperación de ADN antiguo de restos humanos se centraron en el análisis del genoma mitocondrial (ADNmt), debido a ciertas características de este material genético. El ADNmt se hereda por vía materna y tiene una longitud corta de 16,569 pb comparado con los 3,200 millones pb del genoma nuclear y está presente en cientos o miles de copias dentro de las células (20). El alto número de copias del ADNmt hace más factible su recuperación en comparación con el genoma

nuclear que sólo se encuentra en dos copias por célula (Fig. 3). Asimismo, el ADNmt tiene una alta tasa de mutación (comparado con el ADN nuclear) y ausencia de recombinación (20). Además, se ha observado que la tasa de degradación del ADNmt es más lenta que la del genoma nuclear lo que hace más factible su recuperación (3).

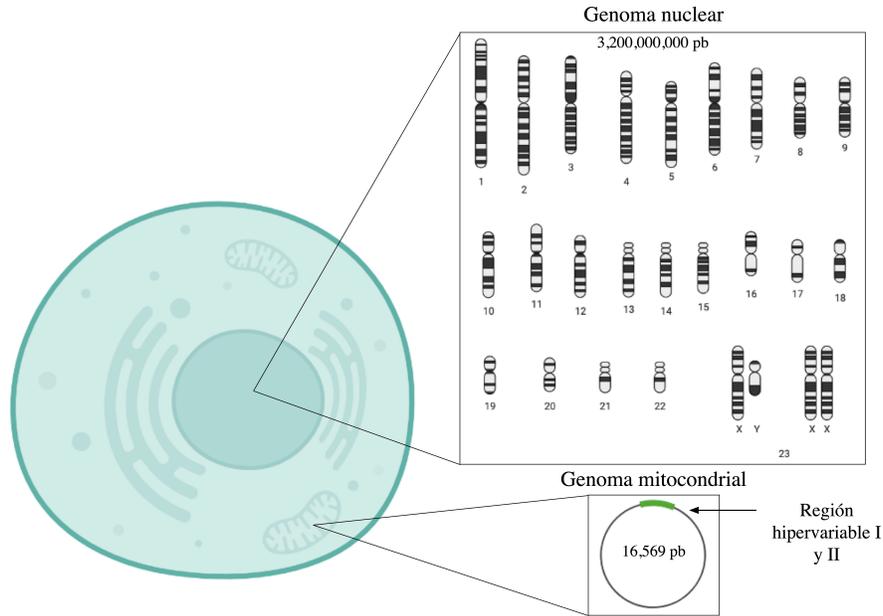


Figura 3. Genoma humano. El genoma humano está compuesto por el genoma nuclear y el genoma mitocondrial. Cada cromosoma autosomal del genoma nuclear se encuentra en dos copias. En la imagen se muestra solamente una copia (Imagen creada con BioRender).

Estas características, junto con la llegada de la reacción en cadena de la polimerasa (PCR) permitieron la amplificación de algunas regiones del genoma mitocondrial. Particularmente fue posible amplificar las regiones hipervariables HVRI y II que se encuentran en una pequeña región del genoma mitocondrial que acumula más mutaciones (Figura 3). Estas regiones HVRI y II se han utilizado ampliamente para desarrollar estudios sobre los linajes de ADNmt de poblaciones humanas antiguas. Hoy en día, gracias a toda la información que se ha generado con ADNmt, se ha propuesto un árbol filogenético bien documentado (phylo tree) (22) contra el cual pueden compararse los genomas mitocondriales antiguos.

Los haplogrupos mitocondriales son grupos de secuencias que comparten variaciones a nivel de nucleótidos. Su nomenclatura se basa en las letras del alfabeto: A-G y X para los

linajes de ADNmt americanos y asiáticos, H-K para los europeos y L para casi toda la variación encontrada en África (23).

Asimismo, el ADNmt en el campo de la paleogenómica humana permite estudiar la historia materna de los individuos y las poblaciones de interés. Estos análisis contribuyen a la comprensión de la dinámica social de las poblaciones del pasado, a dilucidar el poblamiento de continentes y regiones, a identificar las relaciones entre individuos (24), e incluso identificar algunos haplotipos que ya no están presentes en las poblaciones contemporáneas (25, 26).

1.2.1.2 Cromosoma Y

Al igual que el ADNmt, las variantes del cromosoma Y también han sido clasificadas dentro de los haplogrupos documentados en varias publicaciones e integrados en un árbol filogenético para este cromosoma ([ISOGG](#)). Sin embargo, a diferencia del ADNmt, el cromosoma Y se encuentra en una sola copia dentro de las células masculinas y es uno de los cromosomas más pequeños. A pesar de ello, el cromosoma Y es muy interesante ya que contiene la porción de ADN más largo que no se recombina en el genoma humano. Además este cromosoma alberga la historia del linaje paterno. Esto convierte al cromosoma Y en una poderosa fuente para estudiar la historia de la población humana.

El primer estudio de secuencias completas del cromosoma Y de humano (69 secuencias) publicado en 2013 propuso que la porción informativa del cromosoma Y puede reducirse a ~10 Mb (21). En el 2016 el uso de NGS permitió reconstruir un árbol filogenético más informado para el cromosoma Y, con 1,244 secuencias, el cual reveló aspectos relevantes de la historia poblacional de los linajes paternos a nivel continental (27). En el continente americano, se encuentran principalmente los haplogrupos de cromosoma-Y Q y C de forma nativa (28, 29). Sin embargo, debido a la colonización por parte de Europeos en el siglo XVI, actualmente el haplogrupo de chrY más frecuente en la población mexicana que habita en comunidades urbanas es el haplogrupo R de origen europeo (30).

1.2.2 Estudio del genoma autosomal

Los marcadores uniparentales, a pesar de ser muy informativos, sólo nos permiten reconstruir una parte de la historia genética de las poblaciones. Afortunadamente, en la actualidad es posible realizar estudios a nivel poblacional utilizando la información completa contenida en el genoma nuclear la cual nos provee de una cantidad de información más amplia y completa al ser heredada de padre y madre.

El genoma humano nuclear consta de 3,200 millones de pares de bases, distribuidas en 22 pares de cromosomas autosomales y 1 par de cromosomas sexuales (Figura 3). Cada progenitor (madre y padre) hereda a sus hijos la mitad de su información genética. Sin embargo, la información no pasa de manera intacta. Los cromosomas parentales pasan por un proceso de recombinación al formarse los gametos. Por ello la información genética de los cromosomas parentales que hereda su progenie no es idéntica a los de los padres. Este proceso se repite conforme van pasando las generaciones y, por ello, el ADN autosomal es más bien un mosaico de fragmentos que el individuo ha heredado de sus ancestros (31). Además, durante el proceso de gametogénesis, pueden ocurrir mutaciones al genoma haciéndolo también diferente al de los progenitores y generando nueva variación genética (32).

En promedio, el porcentaje de diferencias que existe entre el genoma de dos individuos es del 0.1%, lo cual equivale a ~3,200,000 sitios variantes (33). Sin embargo, para los estudios a nivel poblacional no resulta necesaria toda esta información. La variación genética que se estudia se enfoca principalmente en cientos de miles de sitios del genoma que son variantes de un solo nucleótido (SNV) las cuales se sabe que varían en los individuos estudiados (33). Este tipo de estudios a nivel del genoma nuclear se han realizado utilizando información genómica obtenida de poblaciones actuales. Con ello se ha logrado inferir cómo fue la demografía de éstas hace cientos o miles de años (34–36).

Sin embargo, ahora que es posible recuperar ADN de individuos que vivieron hace cientos o miles de años, se puede conocer directamente cómo eran genéticamente las poblaciones en el pasado, inferir cómo era su organización social e incluso inferir tamaños poblaciones. De esta manera, los estudios paleogenómicos han permitido abrir una ventana directa al pasado para analizar no sólo los marcadores uniparentales, sino

también para caracterizar la diversidad genética antigua y entender los procesos demográficos y evolutivos que han generado la diversidad genética en poblaciones humanas actuales (37, 38).

1.3 Estudios de ADN humano en México

El estudio de ADN en México inició con los trabajos de la Dra. Rocío Vargas Sanders durante su tesis de doctorado en la Facultad de Ciencias de la UNAM titulada “De los esqueletos a la doble hélice: hacia un estudio de paleoantropología molecular”. Este trabajo se realizó en la Facultad de Ciencias de la UNAM bajo la dirección del Dr. Faustino Rodríguez Romero e inició en 1986 y concluyó en 1993 (39). Durante este periodo la Dra. Vargas logró detectar núcleos celulares en tejido óseo de diversas muestras de origen prehispánico comprobando así que existía material genético recuperable. De este material también realizó la extracción de ácidos nucleicos y logró aislarlos, tanto de ADN como ARN. En su trabajo publicado en 1996 se reporta la extracción de ADN y ARN a partir de restos óseos de un individuo con una antigüedad de 650-750 años proveniente del Peñón del Marqués en Iztapalapa, DF (40). No es posible descartar que sus resultados hayan sido producto de contaminación con ADN o ARN moderno debido a que las normas para trabajo con ADN se establecieron después. A pesar de ello, con estos resultados alentadores, la Dra. Vargas planteó como perspectivas que en un futuro se podría conocer la variabilidad genética, patrones de migración, relaciones de parentesco entre individuos y relación con poblaciones indígenas actuales.

A partir de entonces, investigadores mexicanos con los mismos intereses científicos han contribuido al estudio de las poblaciones prehispánicas desde el análisis del ADN. Y tanto en la UNAM como en el Instituto Politécnico Nacional existen grupos de investigación enfocados en el estudio de ADN para resolver preguntas en torno a las poblaciones prehispánicas de México. Las investigaciones que se han realizado hasta el momento utilizando muestras de origen prehispánico se han centrado principalmente en el análisis del linaje mitocondrial y en la determinación del sexo molecular de los individuos. Estos análisis se han realizado principalmente mediante el aislamiento de ADN y su amplificación por PCR (41–46).

Uno de estos estudios fue el realizado en 2015 con muestras del sitio arqueológico de Teopancazco. El análisis del ADN mitocondrial de estas muestras ayudó a confirmar la diversidad multiétnica del sitio arqueológico a través del estudio de los linajes maternos. La evidencia arqueológica ya apuntaba a que este sitio albergaba individuos de diversas regiones, y el ADN confirmó estas observaciones. Asimismo, se logró conocer más a fondo el ritual de los entierros realizados en el sitio pues se identificó el sexo de infantes enterrados en este complejo. Este estudio demostró que el entierro de éstos tenía un acomodo especial dependiendo del sexo de los infantes (47). Cabe mencionar que la determinación de sexo en esqueletos de individuos infantes es complicada y puede llegar a ser muy imprecisa. Por ello la determinación del sexo mediante datos de NGS de ADN es un método valioso (48). De forma similar en el estudio de De la Cruz et al. 2008 se analizó el sexo de 37 infantes y 6 adultos de un contexto de sacrificio en el templo del dios Ehécatl en Tlatelolco. En este estudio se observó que la mayoría de los individuos de los que fue posible determinar el sexo pertenecían al sexo masculino. Estos resultados concuerdan con previas hipótesis que indican que las víctimas seleccionadas para el sacrificio personificaban de cierta forma al Dios al que se hacía ofrenda. En este caso, el Dios Ehécatl al ser del sexo masculino era mejor representado por individuos de este sexo (42).

En otros trabajos publicados en revistas científicas y tesis de grado realizadas en México se ha estudiado la diversidad genética del linaje mitocondrial de diferentes sitios arqueológicos. Entre los cuales se encuentran sitios del centro de México como Teotihuacán y Tlatelolco, y sitios del sureste de México (Tabla 1).

Tabla 1. Investigación de ADN en poblaciones prehispánicas de México, realizada en México.

Sitio Arqueológico	Estado	Número de individuos analizados	Datación (CE)	Tipo de análisis	Referencia	Año	Laboratorio
Xcaret	Quintana Roo	28	600-800 y 900-1521	ADNmt	41	2001	Facultad de Ciencias, UNAM, CDMX
Tlatelolco	CDMX	43	1454-1457	ADNmt y determinación de sexo	42	2008	Facultad de Ciencias, UNAM, CDMX, en colaboración con laboratorios de EUA
Teopancazco	Estado de México	21	200-650	Determinación de sexo	44	2014	LANGEBIO-CINVESTAV IPN, Guanajuato
Teopancazco	Estado de México	29	200-600	ADNmt y determinación de sexo	47	2015	LANGEBIO-CINVESTAV IPN, Guanajuato
Xcambo, Bonampak, Palenque, El Rey, Comalcalco, Tenosique, Peje Lagarto Huimanguillo, Sueños de Oro y Calicanto	Tabasco, Chiapas, Yucatán y Quintana Roo	38	250-1500	ADNmt	43	2016	CINVESTAV IPN, CDMX
Las Fuentes	Colima	1	1200-800 BCE	ADNmt	47	2016	INAH y Facultad de Ciencias UNAM
Teotihuacan	Estado de México	36	250-700	ADNmt	45	2017	Instituto de Investigaciones Antropológicas, UNAM y CINVESTAV, CDMX

Todos estos trabajos reportados hasta la fecha se han enfocado en la amplificación por PCR de distintas regiones del genoma mitocondrial y nuclear, y han sido un primer acercamiento a lo que puede conocerse genéticamente de dichas poblaciones. Investigaciones de este tipo (PCR) también han sido llevadas a cabo por grupos de investigación en el extranjero (49–54). Asimismo, al momento de escritura de esta tesis, todos los estudios con muestras humanas prehispánicas de México realizados a nivel genómico (mitocondrial o nuclear) con NGS se han llevado a cabo por grupos extranjeros (26, 52, 55) y sólo una de estas publicaciones hubo participación de autores mexicanos (55). Un ejemplo de esto es una investigación extranjera realizada en Copenhague, Dinamarca que reportó por primera vez información genómica autosomal a baja cobertura (< 0.3 x) de individuos prehispánicos de México de Pericúes del norte de México y de la Sierra Tarahumara con un rango de antigüedad de 900-300 años (Tabla 2). Este análisis reveló las afinidades genéticas de los individuos prehispánicos con poblaciones indígenas actuales y puso a prueba algunas hipótesis acerca del poblamiento del continente americano (55).

Tabla 2. Investigaciones que incluyen muestras humanas de origen prehispánico realizadas en el extranjero.

Sitio Arqueológico	Estado	Número de individuos analizados	Datación (CE)	Tipo de análisis	Investigación	Referencia	Año
Tlatelolco	CDMX	30	1350 -1400	ADNmt	Movimientos poblacionales	49	2011
Xaltocan	Estado de México	26	1240 -1521	ADNmt	Remplazo poblacional	50	2012
Hoyo Negro	Quintana Roo	1	13000 - 12000 AP	ADNmt	Poblamiento de América	51	2014
Piedra Gorda y Sierra Tarahumara	Baja California Sur y Chihuahua	8	800 - 300 AP, >500 AP	NGS genoma completo	Poblamiento de América	55	2015
Cueva La Candelaria	Coahuila	5	1000 - 1600	ADNmt	Poblamiento de América	25	2016
Paquimé	Chihuahua	14	1200 - 1450	ADNmt	Afinidades genéticas	54	2017a
Convento y Paquimé	Chihuahua	5	700 - 1450	ADNmt (NGS)	Afinidades genéticas	52	2017b
Comondú	Baja California Sur	2	ND	NGS genoma completo	Poblamiento de América	26	2018
Tlatelolco y Cholula	CDMX y Puebla	28	1325 - 1520, 900 - 1350	ADNmt y determinación de sexo. NGS de genoma completo para las muestras de Tlatelolco pero no disponible públicamente	Afinidades genéticas	53	2019

Es importante destacar que ha existido un desbalance en las publicaciones realizadas en el territorio nacional y el extranjero, siendo este último el que ha tenido acceso a mejores presupuestos y por consiguiente a mejores tecnologías para la generación de datos y su posterior análisis. Esto ha repercutido en las colaboraciones entre laboratorios a nivel nacional ya que algunos grupos prefieren la asociación con el extranjero debido a la rápida obtención de resultados pero a costo de liberar un gran número de muestras (patrimonio de la nación) para los análisis (56).

Afortunadamente, en los últimos años, el campo de la paleogenómica en México se ha fortalecido y se han establecido colaboraciones entre diferentes grupos de investigación de diferentes casas de estudios como lo son la UNAM, el IPN, el INAH y el INMEGEN (57). Esta tesis surge de colaboraciones establecidas entre estos institutos y es una primera contribución para conocer el pasado genómico de las poblaciones prehispánicas de México, su estructura genética, diversidad, pasado demográfico y ancestría. Asimismo, el presente trabajo se realizó en el primer laboratorio de Paleogenómica Humana de México que cuenta con la infraestructura y cumple con los estándares internacionales necesarios para el trabajo con ADN.

2 Hipótesis y Objetivos

2.1 Hipótesis

A través del análisis del ADN de individuos de la época prehispánica es posible observar cómo era la estructura y diversidad genética de las poblaciones prehispánicas de México previo a la colonización por europeos. Asimismo, a través del análisis de estos datos es posible contestar preguntas de interés local como lo es la continuidad poblacional en la Frontera Norte de Mesoamérica.

2.2 Objetivo principal

El pasado genético de las poblaciones prehispánicas es poco conocido, y aunque éste ha sido estudiado en diversas investigaciones, esto se ha realizado enfocándose principalmente en la descripción de los haplogrupos mitocondriales hallados en las poblaciones prehispánicas. En los trabajos realizados en el extranjero con NGS de genoma completo sólo se han explorado las afinidades genéticas de los individuos prehispánicos en el contexto del continente Americano. Por lo tanto, el objetivo principal del trabajo doctoral es contribuir con el conocimiento de la diversidad y la estructura genética que existía en las poblaciones nativas de distintas regiones de México antes de la colonización europea. Esto lo realizaremos utilizando la información genética de los marcadores uniparentales a nivel de sub-haplogrupo y del genoma autosomal de individuos de la época prehispánica.

2.2.1 Objetivos particulares

Para cumplir con el objetivo principal es necesario el establecimiento de objetivos particulares que contribuyan de manera escalonada a alcanzar el cumplimiento del trabajo doctoral.

Para ello se establecieron de manera inicial objetivos a nivel experimental y posteriormente los objetivos a nivel de análisis de los datos genómicos generados.

- 1) El Laboratorio Internacional de Investigación sobre el Genoma Humano fue fundado en 2016. En el momento en que el trabajo de la presente tesis fue iniciada (Febrero 2017), el Laboratorio de Paleogenómica del grupo de investigación de Genómica Evolutiva y de Poblaciones estaba en sus inicios y los métodos de aislamiento de ADN y construcción de bibliotecas genómicas aún no habían sido estandarizados en éste. Es por ello que el primer objetivo particular que se planteó fue la obtención exitosa de ADN mitocondrial y nuclear a partir de restos arqueológicos prehispánicos y la realización de la secuenciación mediante NGS del material genético obtenido. Todo esto cumpliendo con los estándares de calidad del campo de la paleogenómica.
- 2) El segundo objetivo particular planteado fue la identificación de marcadores uniparentales (ADNmt y chrY) de los individuos analizados para estudiar el linaje materno y paterno de los individuos. El estudio de ADNmt además de proveer información del linaje materno es útil para estimar el porcentaje de contaminación, lo que nos da un indicio de si el trabajo de laboratorio se está llevando a cabo correctamente. Asimismo, con estos primeros análisis es posible detectar los patrones de daño característicos del ADN y verificar que las secuencias obtenidas son auténticamente antiguas.
- 3) El tercer objetivo fue analizar el genoma autosomal para observar las afinidades genéticas de las poblaciones antiguas en el contexto de la diversidad actual de las poblaciones indígenas de México. Esto nos daría indicios de si los individuos de la época prehispánica y actual presentan una estructura genética similar o si ésta cambió después de la colonización por europeos.
- 4) Como cuarto objetivo se planteó la estimación de la diversidad genética en las poblaciones prehispánicas y comparar ésta con la encontrada en las poblaciones indígenas actuales de México. Este objetivo se planteó como uno de los últimos ya que para realizar las estimaciones se requiere una mayor cantidad de cobertura del genoma completo y esto se logra una vez que se han procesado todas las muestras y se ha analizado cuales son candidatas para su secuenciación a profundidad para

obtener la mayor cantidad de datos posibles. El análisis de la diversidad genética es importante porque la población nativa pasó por un declive demográfico importante después de la colonización por europeos. Hipotetizamos que el declive demográfico podría haber tenido como consecuencia la pérdida de diversidad genética y probablemente de variantes adaptativas.

- 5) El último objetivo se planteó junto con los arqueólogos Elizabeth Mejía Pérez Campos y Alberto Juan Herrera Muñoz de las zonas arqueológicas de Toluquilla y Ranas, respectivamente. Este objetivo consiste en determinar, a través del análisis genético de los individuos que habitaron en estas zonas arqueológicas en distintos períodos, si a lo largo de la ocupación prehispánica de la región existió continuidad genética poblacional (320 aC – 1521 dC). El objetivo se plantea porque estudios arqueológicos previos afirman que la frontera norte de Mesoamérica (donde se encuentra Toluquilla y Ranas, además de otros sitios arqueológicos de México) fue abandonada por los pobladores mesoamericanos a inicios del siglo XX y fue ocupada por grupos de cazadores-recolectores de Aridoamérica provenientes de la región norte de México.

Esta tesis, además del artículo científico publicado como requisito de titulación, incluye capítulos teóricos que servirán de base para la introducción al campo de la Paleogenómica para las futuras generaciones de estudiantes del presente laboratorio. También espero que sea de utilidad para estudiantes que estén interesados en el tema.

3 Métodos de laboratorio para el análisis de ADN antiguo

El entusiasmo generado en la década de los 80's por el éxito en la recuperación de ADN de especies extintas dio paso a que varios grupos de investigación iniciaran su investigación en este campo con expectativas muy elevadas (2). Al inicio de la década de los 90's se publicaron trabajos en los que se afirmaba haber obtenido ADN de muchos millones de años de antigüedad, p. ej. de la planta *Magnolia* (17 – 20 Ma) (8), de organismos del cretácico (80 Ma) (58) y de gorgojos (120-135 Ma) (59). Sin embargo, este ADN resultó ser proveniente de contaminación con ADN del laboratorio y no del organismo bajo estudio (60). Como resultado de esto, la credibilidad de la investigación realizada en el campo de ADN se vio afectada, lo que llevó a la comunidad científica interesada en estos estudios a proponer una serie de normas para evitar la contaminación y la obtención de resultados erróneos en estas investigaciones (61). Actualmente los laboratorios de paleogenómica deben seguir estas recomendaciones para llevar a cabo investigaciones reproducibles y confiables.

3.1 Infraestructura y medidas para prevenir la contaminación

Antes de trabajar con muestras antiguas es necesario tomar en cuenta que este tipo de investigación debe realizarse en un laboratorio especializado. Es de suma importancia evitar la contaminación de la muestra antigua con ADN moderno. Por ello es necesario utilizar instalaciones que se encuentren físicamente separadas de los laboratorios donde se trabaje con ADN moderno o en el que se realicen procedimientos de PCR (11). Idealmente, el laboratorio debe contar con presión positiva para reducir la entrada de aire externo contaminante. Además, es importante que el personal utilice equipo de protección de cuerpo completo para proteger las muestras de contaminación con su propio ADN (11). El equipo de protección consta de un traje completo (de pies a cabeza), cofia, cubre

bocas, guantes, mangas y zapatos desechables. La única parte que queda expuesta del cuerpo de la persona son los ojos (Fig. 4). En algunos laboratorios incluso se hace uso de careta o gafas de protección (no en todos los laboratorios se utilizan ya que éstas tienden a empañarse causando incomodidad a la hora de trabajar) (19).



Figura 4. Persona trabajando con el equipo de protección necesario dentro del Laboratorio de Paleogenómica del LIIGH-UNAM.

El material y reactivos que se utilizan deben estar estériles y las puntas que se utilizan con las micro pipetas deben contar con filtro para evitar contaminación durante el procesamiento de las muestras. Todo material que entra en el laboratorio debe ser limpiado superficialmente con hipoclorito de sodio al 5% y etanol al 75% y expuesto a luz UV (240 nm). Además, todas las superficies del laboratorio deben ser limpiadas

constantemente con dichos reactivos antes y después de cada uso y una vez a la semana o cada dos semanas dependiendo de qué tanto se utilice el laboratorio. Idealmente, el laboratorio debe también contar con una lámpara UV para irradiar todo el cuarto cuando éste no está en uso. Como medida adicional, si una persona ha estado en el laboratorio de ADN moderno y después de esto necesita entrar al laboratorio de ADN antiguo se requiere que esa persona tome una ducha y se cambie las prendas que utilizó en el laboratorio de ADN moderno para evitar introducir contaminación con moléculas de ADN o productos de PCR (11).

Al trabajar con muestras antiguas también es importante: 1) Añadir controles negativos en cada paso del procesamiento para descartar cualquier posible contaminación. El uso de controles positivos no es recomendado ya que pueden ser una fuente de contaminación y deben ser evitados. 2) Observar un comportamiento molecular esperado. Esto quiere decir, por ejemplo, que se espera que la cantidad de ADN recuperado sea mucho menor al que se obtendría de una muestra del presente, que el tamaño de las moléculas sea pequeño, y que los resultados en el análisis de datos tengan sentido filogenético (especie, población, etc.). En caso de resultados que se desvíen de lo esperado sería recomendable replicarlos con otra muestra del mismo individuo y realizando el procesamiento en otro laboratorio. Asimismo, las moléculas de ADN antiguo deberán de presentar el patrón característico de daño molecular que consiste en un aumento en la frecuencia de cambios de C → T en el extremo 5' y G → A en el extremo 3'. 3) En caso de detectar contaminación, es necesario descubrir la fuente contaminante (reactivos, material, malos protocolos de trabajo, etc.) y tomar todas las medidas necesarias para eliminarla (11, 61).

3.2 Procesamiento de las muestras arqueológicas

El personal está listo para entrar al área donde se procesan las muestras después de haberse preparado con todas las capas de protección y de tener las medidas de prevención en mente. El primer paso en el procesamiento de las muestras (dientes, hueso, cabello, tejido momificado, etc.) es retirar los residuos de suelo o cualquier otro material que pudiera encontrarse adherido a la superficie de la muestra. Una vez hecho

esto, las muestras se limpian con hipoclorito de sodio al 5% y posteriormente con etanol al 75% para retirar el exceso de cloro. Finalmente son expuestas a 240 nm de UV durante 1.5 min. Dependiendo de la muestra y el protocolo que se utilice, ésta puede ser pulverizada o fragmentada para proceder con la extracción del ADN y pueden utilizarse de 0.05 – 0.15 g de muestra (62).

3.2.1 Extracción de ADN

Actualmente existen dos principales métodos para la extracción de ADN. Estos son conocidos como método 'in silica' (62) y método de Dabney (63). La principal diferencia entre los dos es el tamaño de las moléculas recuperadas. El primero recupera principalmente moléculas > 60 pb mientras que el segundo puede recuperar moléculas más pequeñas >30 pb. El método de Dabney es preferido para la obtención de ADN porque utiliza una cantidad mucho menor de buffer de unión por lo que el costo del uso de reactivos se disminuye. Ambos protocolos de extracción se basan en la liberación del ADN de la muestra a través del uso de proteinasa K -que degrada las proteínas del tejido para ayudar a liberar el ADN-, EDTA -que evita la degradación del ADN liberado- y una pequeña concentración de detergente Tween 20 -el cual ayuda al proceso de digestión del tejido-, formando así el buffer de lisis. La solución contiene el ADN liberado después de incubar la muestra por >16 horas a 37°C con el buffer de lisis. Posteriormente las moléculas de ADN son expuestas a una solución de alta concentración de tiocianato de guanidinio junto con otros componentes (buffer de unión) con lo cual el ADN adquiere afinidad hacia partículas de sílice. Una vez hecho esto, el ADN se aísla ya sea añadiendo partículas de sílice a la solución ('in silica') o utilizando columnas especiales con sílice (Dabney) por las cuales se hace pasar el ADN con el buffer de unión. Ya que se han unido las moléculas de ADN a las partículas de sílice, el ADN es recuperado con un buffer de elución (H₂O, EDTA, Tris-HCl) que libera a las moléculas de las partículas de sílice.

Estos métodos, aunque han sido optimizados para recuperar ADN de cadena corta como es el caso del ADN, no son capaces de discriminar entre las moléculas de ADN antiguo o el ADN moderno contaminante. El ADN aislado corresponde tanto a las moléculas antiguas como a las modernas que provienen de diferentes fuentes (p. ej. microbiota en contacto con la muestra) (62) al final del procedimiento de extracción. Como paso final de

este proceso se realiza la medición de la concentración del ADN de la muestra antigua utilizando un equipo de alta sensibilidad para detectar las bajas concentraciones de ADN como el fluorómetro Qubit con el kit para cuantificación de ADN de doble cadena de alta sensibilidad. Generalmente se obtiene una cantidad de ADN de 0.1 ng/uL a 10 ng/uL (observaciones personales).

El ADN recuperado puede utilizarse ya sea como molde para amplificar regiones del genoma que sean de interés, como se hacía en los primeros estudios de ADN, o para su secuenciación mediante NGS. Actualmente lo más común es secuenciar mediante NGS una muestra del material genético recuperado; este tipo de secuenciación se conoce como secuenciación de escopeta o shotgun. Para ello es necesario construir bibliotecas genómicas de secuenciación.

3.2.2 Construcción de bibliotecas genómicas

Existen dos tipos de bibliotecas genómicas que se usan para investigación de ADNa, las de doble cadena y las de cadena sencilla (Fig.5). Ambos métodos tienen en común que a cada molécula de ADN se le deben añadir secuencias de adaptadores en cada extremo para asegurar que el ADN pueda hibridar con la plataforma de secuenciación y ser leído por ésta. Los adaptadores que se utilizan son específicos de las plataformas de secuenciación. Sin embargo, la mayoría de las investigaciones en Paleogenómica han utilizado plataformas de secuenciación illumina.

El proceso de construcción de bibliotecas genómicas de doble cadena es probablemente el más utilizado en las publicaciones de paleogenómica y consta de 3 pasos (64). El primero consiste en la reparación de los extremos de las cadenas de ADN. Esto quiere decir que los extremos de todas las moléculas de ADN se convierten en extremos romos. Seguido de este paso, se realiza la adición de dos diferentes adaptadores correspondientes a la plataforma de secuenciación en ambos extremos de la doble cadena de ADN. Los adaptadores se ligan a cada extremo al azar por lo que habrá moléculas que contengan en ambos extremos el mismo adaptador y otras que tengan un adaptador diferente en cada extremo. Éstas últimas son las que podrán ser leídas durante la secuenciación. Por último, los extremos de cadena sencilla de los adaptadores son

completados con los nucleótidos faltantes. Dependiendo del protocolo se puede requerir o no realizar pasos de limpieza entre una reacción y la siguiente. El realizar estos pasos de limpieza tiene un impacto negativo en la cantidad de moléculas de ADN que se obtienen al final del proceso de construcción de bibliotecas genómicas. Esto se debe a que en cada paso de limpieza se pierden moléculas. Para reducir esta pérdida, uno de los últimos protocolos para construcción de bibliotecas genómicas de doble cadena evita los pasos de limpieza e implementa la desnaturalización de las enzimas utilizadas en los pasos previos para poder realizar el procedimiento en un solo tubo (protocolo BEST) (65).

También existe la posibilidad de realizar bibliotecas genómicas de cadena sencilla (66–69). Los protocolos de cadena sencilla en el campo de ADNa se han desarrollado para incrementar el número de moléculas que son convertidas a bibliotecas genómicas ya que incluye el ADN de cadena sencilla que pudiera estar presente en la muestra. Estos protocolos realizan una desnaturalización de las moléculas de ADN para añadir los adaptadores tanto al ADN de cadena sencilla como al ADN de doble cadena. Posteriormente los adaptadores se adhieren a las moléculas de ADN con una orientación específica permitiendo que en cada extremo de las moléculas de ADN se encuentre el adaptador que le corresponde. Esto incrementa la eficiencia en la secuenciación. A pesar de estas ventajas, estos métodos no son ampliamente utilizados debido al alto costo de los reactivos y al tiempo de trabajo requerido que es considerablemente mayor que en los protocolos de cadena doble. Recientemente, en 2021, se publicó un protocolo de construcción de bibliotecas genómicas de cadena sencilla optimizado para ADNa. Este método utiliza enzimas de mucho menor costo y reduce los tiempos de construcción de las bibliotecas genómicas. Una ventaja de este método es que tiene una eficiencia similar al de otros protocolos de cadena sencilla (69).

Una vez construidas las bibliotecas genómicas, a éstas se les añade mediante PCR una secuencia de identificación única, llamada índice, el cual consta de 6-7 pb. Las reacciones de adición de índices se preparan en el laboratorio de ADNa y la reacción de PCR es llevada a cabo en el laboratorio de ADN moderno. Después de este proceso, las bibliotecas genómicas son trabajadas en el laboratorio de ADN moderno con la seguridad de que no se añadirá contaminación externa.

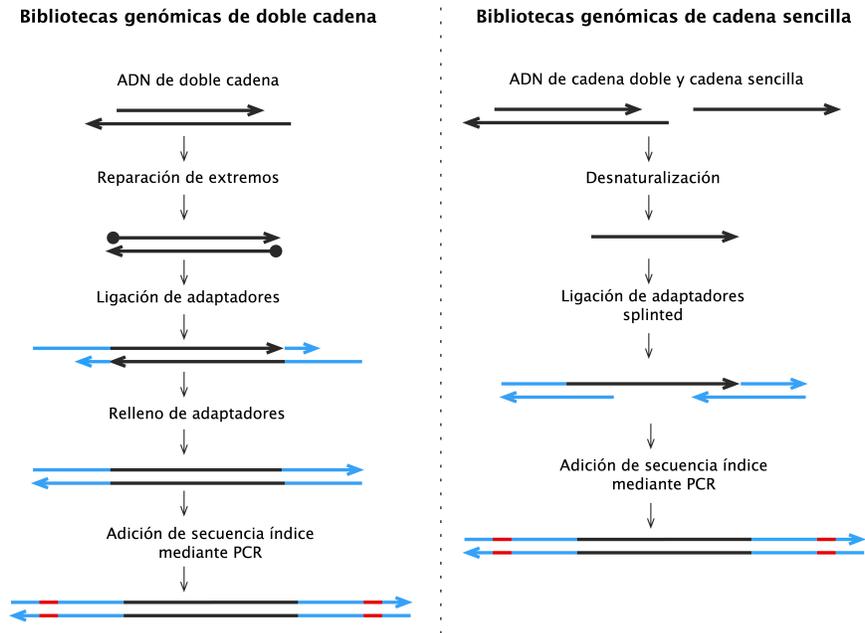


Figura 5. Proceso de construcción de bibliotecas genómicas. En negro se muestran las moléculas de ADN, en azul los adaptadores y en naranja los índices. El extremo sin punta en las moléculas de ADN indica el extremo 5', mientras que el extremo con punta indica el extremo 3' de la molécula.

Las bibliotecas de ADN son enviadas a secuenciación para realizar una primera evaluación como procedimiento estándar en el Laboratorio de Paleogenómica del LIIGH. Para llevar a cabo este proceso se requiere una cantidad aproximada de >5,000,000 de lecturas por muestra. A partir de estas lecturas se evalúa el porcentaje de secuencias que corresponden al genoma humano (contenido endógeno), el porcentaje de secuencias repetidas (clonalidad), el tamaño de las lecturas y la presencia de los patrones de daño característicos del ADN.

3.2.3 Procedimiento de captura genómica en solución

A partir de la primera secuenciación de las bibliotecas genómicas es posible evaluar cuales muestras son candidatas para realizar una secuenciación a más profundidad, o para realizar un procedimiento de captura de genoma mitocondrial (mtC) o genoma completo (WGC). Esto se hace tomando en cuenta el contenido endógeno, la clonalidad y la complejidad de la biblioteca genómica. Si el contenido endógeno es alto y la clonalidad

es baja, se podría esperar que la biblioteca genómica tenga buena complejidad (diversidad de moléculas de ADN) para cubrir la mayor parte posible del genoma humano y por lo tanto se podría secuenciar directamente a mayor profundidad (70). En el caso contrario, las bibliotecas podrían ser candidatas a WGC si tienen un contenido endógeno >1% (70, 71) o candidatas a mtC si tienen >0.01 % (observaciones personales).

El enriquecimiento de bibliotecas genómicas por medio de técnicas de captura en solución consiste en aumentar el porcentaje de secuencias de interés. Esto se hace con anzuelos de RNA biotinilados que se construyen a partir de ADN humano, ya sea autosomal o mitocondrial (71, 72). Los anzuelos de RNA pueden construirse en el laboratorio de ADN moderno o pueden ser adquiridos mediante un kit comercial y se utilizan para pescar (hibridar con) las moléculas de ADN humano. Una vez que se han pescado las moléculas de interés, éstas son recuperadas mediante perlas magnéticas con estreptavidina. Posteriormente se purifican las moléculas de ADN capturadas, se amplifican mediante PCR y se envían de nuevo a secuenciación, esperando que haya aumentado el porcentaje de ADN humano en la biblioteca genómica (Fig. 6). Estos procedimientos son muy útiles ya que incrementan considerablemente la información que se puede obtener de una muestra antigua.

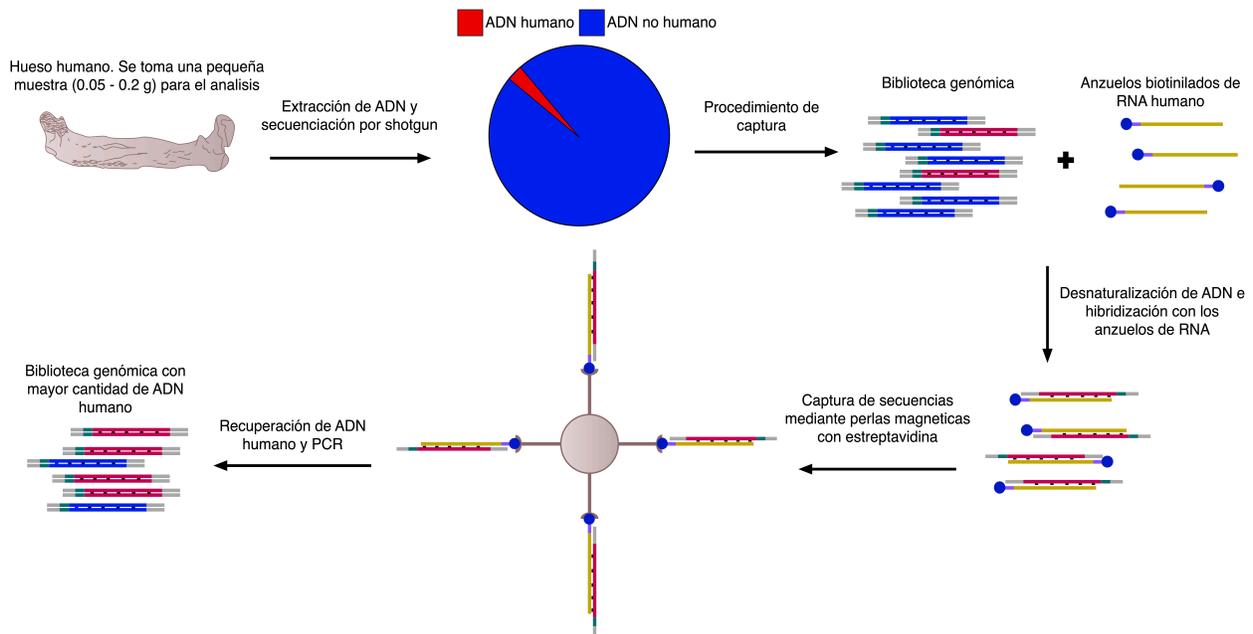


Figura 6. Proceso de captura genómica en solución. En rojo se representa el ADN humano obtenido a partir de la muestra antigua.

4 Revisión de métodos para el análisis de datos paleogenómicos

La Paleogenómica, como se indicó anteriormente, se enfoca en el estudio del genoma de individuos o especies que vivieron en el pasado. Para ello, los estudios paleogenómicos han hecho uso de las tecnologías de NGS que son capaces de producir cantidades masivas de datos a partir de bibliotecas genómicas. Las plataformas de NGS más utilizadas en los estudios paleogenómicos son las plataformas de Illumina. Asimismo, la plataforma BGI-SEQ 500 se ha utilizado recientemente en muestras de ADN resultando en una calidad comparable a la obtenida con Illumina. A pesar de ello dicha tecnología aún no se utiliza ampliamente (73).

En las plataformas de Illumina, antes de iniciar la secuenciación, el ADN con adaptadores se une a la celda de flujo (*flow cell* en inglés). Después se realiza una amplificación por PCR puente de las moléculas de ADN lo cual genera grupos de secuencias clonales o *clústers*. Esto permite a la plataforma intensificar la señal emitida en cada clúster cada vez que una polimerasa de ADN incorpora un nucleótido modificado (fluorescente) a la hebra creciente; esto se llama "secuenciación por síntesis" ([SBS](#)). La señal emitida por el nucleótido incorporado es detectada por un dispositivo óptico y se le asigna una puntuación de calidad en función de su intensidad. Con este enfoque, las plataformas Illumina son capaces de producir millones o miles de millones de lecturas cortas por corrida con una puntuación asociada a cada base leída (74). Además, la adición de las secuencias índices a las bibliotecas genómicas hace posible secuenciar varias muestras en una misma corrida de secuenciación (secuenciación múltiple - 'multiplexing en inglés').

El archivo final que contiene la información de secuenciación de cada biblioteca genómica, es un archivo de texto en formato FASTQ (75). Dependiendo del tipo de secuenciación: un extremo (*single-end*) o ambos extremos (*pair-end*), el usuario puede obtener un archivo FASTQ (R1) o dos archivos FASTQ (R1 y R2), respectivamente, para cada muestra por carril de la celda de flujo.

- Calidad de la secuencia por base y puntuaciones de calidad por secuencia: Lo ideal es que sean > 30.
- Contenido de la secuencia por base: Puede haber un poco de ruido en los extremos 5' y 3' debido a los cambios característicos C -> T y G -> A, respectivamente (Fig.8).
- Distribución de la longitud de las secuencias: Se espera que sean cortas (20 - 200 pb) dependiendo de la muestra.
- Secuencias duplicadas: Es deseable que sea baja ya que las secuencias duplicadas no serán informativas y se descartarán en los análisis posteriores.
- Secuencias sobrerrepresentadas: Idealmente no están presentes
- Contenido de adaptadores: Pueden estar presentes dependiendo de si los adaptadores han sido eliminados bioinformáticamente o no.

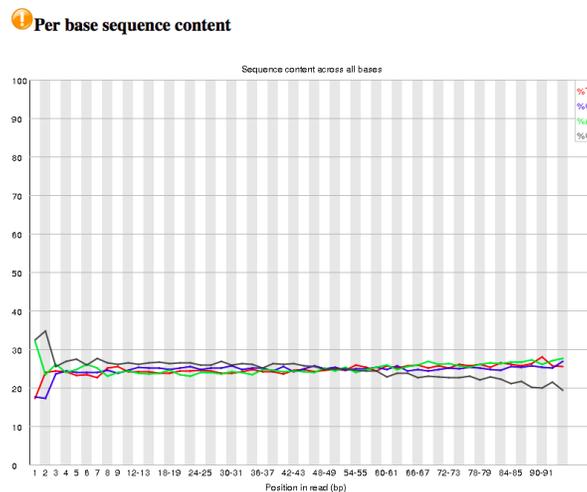


Figura 8. Ejemplo del contenido de la secuencia por base para un conjunto de secuencias de ADN. El eje de las ordenadas indica el porcentaje de cada base nucleotídica, mientras el eje de las abscisas indica la posición de éstas en la secuencia.

4.2 Eliminación de los adaptadores de secuenciación

Normalmente, las secuencias crudas de ADN en los archivos FASTQ suelen contener la secuencia de los adaptadores que se añadieron durante la preparación de las bibliotecas genómicas en el laboratorio. Los adaptadores pueden secuenciarse junto con la molécula de ADN y es importante eliminarlos bioinformáticamente (recorte o *trimming*), ya que no

forman parte de la molécula de ADN original y pueden causar problemas durante el análisis de los datos.

Existen varios programas para eliminar la secuencia de los adaptadores (76–79). Estos programas generalmente requieren la secuencia de los adaptadores y el archivo FASTQ que necesita ser recortado. El programa utilizado durante la realización de esta tesis fue Adapterremoval (77, 79), el cual por defecto ya está configurado para recortar los adaptadores de Illumina.

Otra utilidad de estos programas es la fusión (colapso) de las lecturas pareadas R1 y R2 cuando se utilizan datos de secuenciación *pair-end*. Para realizar este paso, el programa requiere ambos archivos FASTQ (R1 y R2) y devuelve un archivo que contiene las secuencias colapsadas de las lecturas R1 y R2. Para colapsar un par de secuencias, éstas deben solaparse en un número mínimo de pb (indicado por el usuario, 11 pb por defecto en Adapterremoval) y una vez colapsadas pueden ser tratadas como una sola secuencia. Las lecturas de R1 y R2 que no cuentan con este solapamiento mínimo y que por lo tanto el programa no logra colapsar, se depositan en otros dos archivos (pair1 y pair2). Para realizar los análisis posteriores de cada muestra idealmente únicamente se utilizan las secuencias colapsadas.

4.3 Mapeo contra genoma humano de referencia

El siguiente paso en el proceso de análisis es el mapeo de las lecturas contra el genoma humano de referencia. Existen varios ensamblajes del genoma humano pero la gran mayoría de las publicaciones de ADN a la fecha han utilizado el ensamblaje del genoma humano hg19 (GRCh37), generado por el Consorcio de Referencia del Genoma. El ensamblaje hg19 tiene una longitud total de secuencia de 3,101,788,170 dividida en 22 autosomas, los cromosomas sexuales "X" e "Y" y el cromosoma mitocondrial. La secuencia del genoma mitocondrial presente en este ensamblaje no es la Secuencia de Referencia Revisada de Cambridge (rCRS) (80). La secuencia rCRS es la que se usa como estándar para llamar a los haplogrupos mitocondriales. Por lo tanto, al realizar el mapeo se recomienda sustituir el cromosoma mitocondrial del hg19 por la secuencia rCRS. Asimismo, es importante mencionar que la nueva referencia del genoma humano

GRCh38 ya ha empezado a utilizarse en diversos estudios de genómica humana de aDNA por lo que es decisión del investigador si decide utilizar ésta.

Para el mapeo de las secuencias se ha utilizado el software BWA (Burrows-Wheeler Alignment), el cual fue diseñado especialmente para alinear secuencias cortas contra un genoma de referencia de gran longitud. BWA tiene la ventaja de permitir discrepancias (*mismatches*) y huecos (*gaps*), además de asignar una puntuación de calidad del mapeo y múltiples mapeos de una lectura si es necesario, y posee una buena precisión y velocidad (81). Por lo tanto, este software ha resultado de gran utilidad para el mapeo de secuencias de ADNa humano y ha sido ampliamente utilizado en el campo de la Paleogenómica humana.

Cuando se ejecuta BWA para lecturas de ADNa humano una opción útil es "aln", que es ideal para mapear lecturas illumina cortas de 70 bp o menos. Sin embargo, es importante tener en cuenta que BWA inicia el mapeo con una secuencia semilla corta que se espera que tenga pocos *mismatches* con respecto al genoma de referencia. Esto es especialmente problemático en el campo del ADNa debido a los daños estas moléculas presentan en los extremos. Esto resulta en que algunas lecturas tendrán una alta proporción de mismatches en los extremos provocando un mal mapeo o en la ausencia de mapeo de dichas lecturas. Para superar este problema se puede modificar la longitud de la semilla utilizando el indicador "-l" seguido de un número superior a la longitud esperada de las lecturas (por ejemplo, 1000). Al final del mapeo, el alineamiento es guardado en un archivo de texto delimitado por tabulaciones en formato SAM (Sequence Alignment/Map) o su versión binaria BAM (82).

El formato SAM contiene la información de las lecturas alineadas con el genoma de referencia. Comienza con un encabezado de líneas que inician con el carácter "@". El encabezado contiene información sobre la versión del formato, el genoma de referencia, la nomenclatura de los cromosomas y el número de pares de bases en cada uno, entre otros datos. La sección del alineamiento está formada por 11 campos obligatorios delimitados por tabulaciones (82) (Tabla 3) (Fig.9).

Tabla 3. Campos obligatorios en el formato SAM

Columna	Nombre	Descripción
1	QNAME	ID de la secuencia de consulta
2	FLAG	Bandera de bits con información del mapeo
3	RNAME	ID del cromosoma
4	POS	Posición de la 1ª base más a la izquierda de la lectura
5	MAPQ	Calidad del mapeo
6	CIGAR	Anotación CIGAR
7	MRNM	Nombre de referencia de la lectura pareada
8	MPOS	Posición de la 1ª base más a la izquierda de la lectura pareada
9	ISIZE	Tamaño de la lectura
10	SEQ	Secuencia de consulta en la misma cadena que la de referencia
11	QUAL	Calidad de la búsqueda

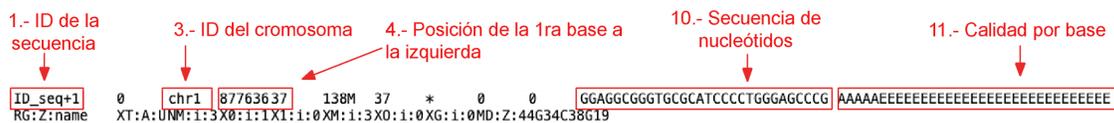


Figura 9. Ejemplo de una línea en formato SAM. En rojo se señalan algunas de sus columnas informativas.

Los archivos crudos de mapeo (SAM/BAM) deben pasar por un procesamiento previo antes de estar listos para los análisis posteriores. El procesamiento inicial se realiza mediante el software SAMtools, mediante este programa es posible (82):

- separar las lecturas mapeadas al genoma humano de las no mapeadas (las lecturas no mapeadas pueden ser analizadas, por ejemplo, para perfiles metagenómicos).
- filtrar las lecturas de acuerdo a la calidad de mapeo, una puntuación mínima de 25 ó 30 es adecuada para las lecturas de ADN.
- eliminar las secuencias duplicadas de PCR (clonas), las cuales no son informativas y deben eliminarse del archivo de mapeo.

Una vez realizado esto, también es necesario eliminar las secuencias que mapearon en múltiples lugares del genoma (parálogos) ya que no son específicas y pueden causar ruido.

Por último se realiza un re-alineamiento local mediante el programa GATK (83) con el objetivo de realinear de la mejor manera las secuencias pertenecientes a inserciones o deleciones (INDELS).

Al finalizar, los análisis posteriores se hacen con el archivo de mapeo filtrado. Éste contiene únicamente las secuencias mapeadas a humano, con buena calidad de mapeo, sin clonas, sin secuencias parálogas y con realineamiento de INDELS.

4.3.1 Control de calidad del mapeo

Una vez realizado el mapeo, es importante determinar algunos parámetros que permitan autenticar la presencia de secuencias de ADN y tener una idea de qué tan conservada se encuentra la muestra. Disponer de un perfil de calidad de la biblioteca permite identificar parámetros para mejorar la estrategia de construcción de la biblioteca en el laboratorio, así como evaluar la posibilidad de aumentar el contenido endógeno con métodos de captura en solución para el genoma completo (71) o el genoma mitocondrial (72). Los parámetros a examinar son principalmente los patrones de daño, el contenido endógeno, la clonalidad, cobertura del genoma y la complejidad de las bibliotecas genómicas.

4.4 Patrones de daño

Los cambios de base con respecto al genoma de referencia en forma de C -> T y G -> A en el extremo 5' y 3', respectivamente, de las secuencias de ADN se toman como una autenticación del ADN (como se explica en la introducción-características del ADN). La proporción de moléculas que presentan tales cambios puede cuantificarse utilizando programas como mapDamage (84), ATLAS (85), aRchaic (86), PyDamage (87), AutentiCT (88), y DamageProfiler (89). Estos programas escanean las secuencias en el archivo de mapeo (BAM) para identificar los mismatches con respecto al genoma de referencia y la posición en la lectura donde ocurren (se esperan cerca de los extremos 5' y 3'). Con esta

información generan los gráficos con la frecuencia de dichos cambios, que son característicos del ADN (Fig. 10).

Una característica útil del programa mapDamage y ATLAS es que tienen opciones que permiten al usuario bajar la calidad de base de las bases que probablemente sean producto de daño, de modo que, aplicando un filtro de calidad de base, se pueden excluir en el análisis posterior (por ejemplo, para el llamado de variantes).

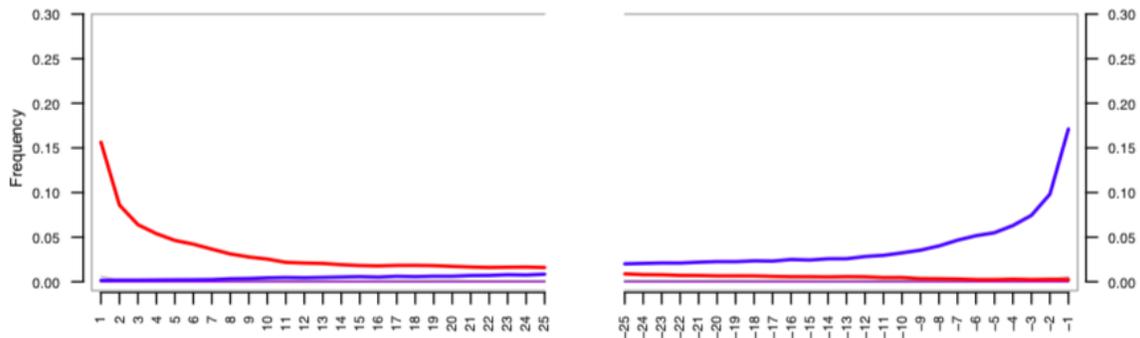


Figura 10. Gráfico de patrones de daño debidos a la frecuencia de cambios de C-> T y G -> A en el extremo 5' y 3', realizado con el programa mapDamage. El color rojo indica la frecuencia de cambios de Citosina a Timina, mientras que el color azul indica los cambios de Guanina a Adenina. El eje "x" muestra la posición de la base dentro de la lectura, dirección 5' -> 3', el eje "y" muestra la frecuencia de estos cambios.

4.5 Estimación de contenido endógeno

El contenido endógeno es el porcentaje de secuencias que mapean contra el genoma humano de referencia y se considera que pertenecen al individuo del que se obtuvo la muestra. Éste se calcula dividiendo el número de secuencias que mapearon al genoma de referencia (una vez eliminadas las secuencias duplicadas y las secuencias parálogas) entre el número total de secuencias, y se multiplica por 100.

En general, las bibliotecas de ADN tienen un contenido endógeno bajo (11), con excepciones en donde pueden alcanzar un porcentaje alto de incluso 80% (12). Debido a esto, las bibliotecas genómicas que cuentan con 1-5% pueden ser sometidas a un procedimiento de captura en solución de todo el genoma (70) para aumentar la cantidad de secuencias humanas recuperadas. La captura genómica resulta ineficiente y poco

rentable para las bibliotecas con <1% de ADN endógeno (70, 71). Sin embargo, para muestras de este tipo, con contenido endógeno bajo (0.01 – 1 %), resulta eficiente la captura en solución del genoma mitocondrial. Posiblemente esto se debe al pequeño tamaño de éste y a su elevado número de copias (observación personal).

4.6 Clonalidad

La clonalidad se refiere al porcentaje de secuencias idénticas que son producto de la PCR y, por tanto, son copias de la misma molécula. Estas moléculas no aportan información valiosa y añaden ruido a los análisis posteriores, por lo que deben ser eliminadas (90). La eliminación bioinformática de estas clonas de PCR se realiza utilizando la opción "rmdup" de SAMtools (82) o a través de "MarkDuplicates" de Picard (91). Antes de eliminar las clonas es importante estimar su porcentaje, ya que esto permite evaluar si una biblioteca es una buena candidata para ser sometida a un método de captura en solución. El porcentaje de clonalidad se calcula restando a 1 el resultado de la división del número de secuencias mapeadas al genoma humano después de eliminar las clonas, entre el número de secuencias mapeadas al genoma humano antes de eliminar las clonas. Al final se multiplica por 100 para obtener el porcentaje (70).

Cuanto menor sea la clonalidad de una biblioteca, mayor será su complejidad (diversidad de moléculas) y mejores serán los resultados del procedimiento de captura (70). En nuestra experiencia en el Laboratorio de Paleogenómica del LIIGH, las bibliotecas genómicas con valores altos de clonalidad (>8%) no son buenas candidatas para procedimientos de captura. Esto se debe a que la clonalidad aumenta también durante este proceso obteniendo en la secuenciación un alto porcentaje de moléculas no informativas. En casos especiales cuando una biblioteca genómica tiene buen contenido endógeno pero alta clonalidad se pueden realizar otros análisis para aumentar la complejidad de la biblioteca. Dichos análisis consisten en la realización de un nuevo proceso de adición de índices sobre la biblioteca genómica que sólo tiene los adaptadores utilizando menos ciclos de amplificación por PCR.

4.7 Cobertura del genoma

La cobertura del genoma humano puede expresarse en porcentaje (%) y puede estimarse calculando el porcentaje del genoma humano que fue cubierto por al menos una lectura (92).

Otra forma de representar la cobertura es mediante la profundidad (x). Se puede estimar calculando el número promedio de veces que se ha leído cada par de bases del genoma. Se expresa en cobertura de veces o x; por ejemplo, un genoma en el que, en promedio, cada par de bases se leyó cinco veces se expresaría como cobertura 5x (o cobertura de cinco veces). La mayoría de las ocasiones, las bibliotecas de ADN se examinan primero y se secuencian a baja profundidad con una cobertura de <1x.

Dependiendo de la pregunta de investigación, se puede decidir si la cobertura obtenida es suficiente o si se necesita más secuenciación para obtener la cobertura deseada. Una cobertura del genoma <1x podría ser suficiente para realizar análisis de determinación de sexo molecular, parentesco, estructura genética, deriva génica compartida, corridas de homocigosidad y diversidad nucleotídica condicional (los cuales se describirán más adelante).

Una manera simple de calcular la profundidad de la cobertura es a través de la función "coverage" del programa ATLAS (85), el cual necesita únicamente el archivo de alineación (BAM).

4.8 Complejidad de las bibliotecas genómicas

La complejidad de una biblioteca genómica se refiere a la diversidad de las moléculas de ADN contenidas en la biblioteca; a mayor diversidad, mayor complejidad. Como se ha mencionado anteriormente, la complejidad está relacionada con la clonalidad y es útil para determinar el número de lecturas de secuenciación necesarias para que una biblioteca determinada cubra todo el genoma del individuo o una fracción del mismo, y para diseñar

la estrategia de secuenciación a seguir. La estimación de la complejidad de la biblioteca se puede realizar con el programa Preseq (93) que necesita como entrada el archivo de alineamiento (BAM) previo a la eliminación de las clonas de PCR, o la información del mismo archivo en formato BED. El programa Preseq toma en cuenta la clonalidad y la diversidad de las moléculas presentes en la biblioteca y, a través de una aproximación de función racional que utiliza el modelo de Good-Toulmin (94), predice la complejidad de la biblioteca hasta más de 30 veces el conjunto de datos original. La complejidad de la biblioteca puede verse en un gráfico con el número de secuencias en el eje de las abscisas y el número de cobertura genómica en el eje de las ordenadas. En estos gráficos es posible apreciar en qué punto de secuenciación se esperaría alcanzar una meseta de saturación, es decir, el momento en que una alta proporción de las secuencias son informativamente redundantes, por lo que no tiene sentido secuenciar más allá de ese número de secuencias (Fig.11).

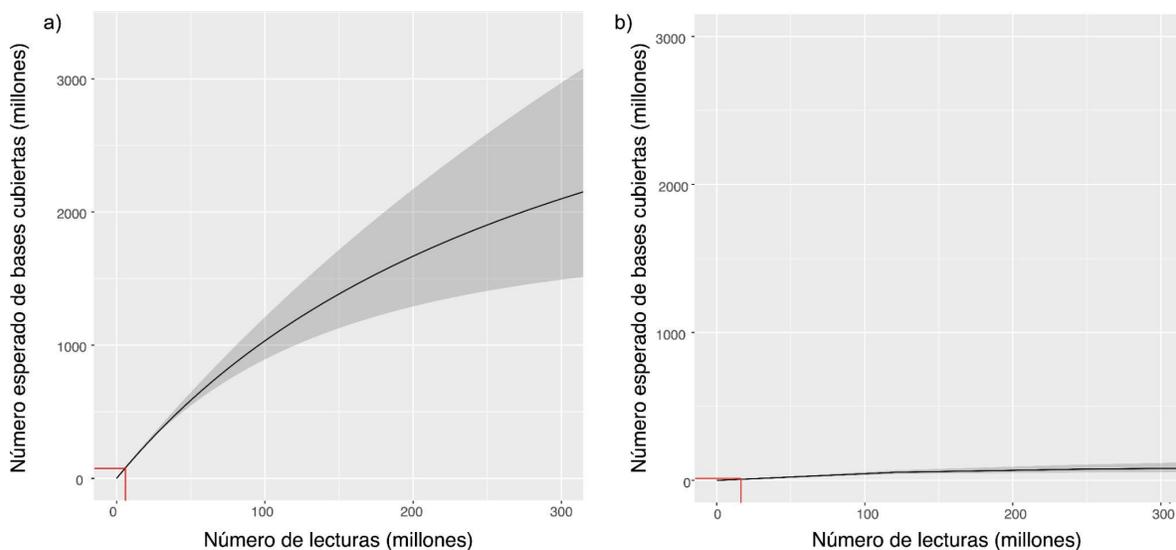


Figura 11. Ejemplo de curvas de complejidad de bibliotecas. a) Biblioteca de alta complejidad, b) Biblioteca de baja complejidad. Los intervalos de confianza se muestran en gris. La cantidad de lecturas generadas a partir de las cuales se realiza la proyección se muestra en rojo.

4.9 Tuberías de análisis de datos de ADN

La paleogenómica es un campo interdisciplinario e incluye la participación e interés de investigadores e investigadoras del área de la biología, la arqueología, la antropología, la genómica, la informática, entre otras disciplinas. El análisis de datos puede ser una tarea

abrumadora para las y los investigadores que no están familiarizadas(os) con los lenguajes de programación. Debido a ello han surgido herramientas computacionales con tuberías de análisis básicas para este tipo de secuencias. Dichas herramientas incluyen programas como PALEOMIX (95) e EAGER (nf-core) (92). Estos programas permiten a los usuarios con poca experiencia en lenguajes de programación analizar datos partiendo desde los archivos FASTQ para realizar la reconstrucción del genoma del individuo e incluso realizar análisis metagenómicos. El programa EAGER nf-core ha sido actualizado en los últimos años (2021-2022) para incluir más herramientas para el análisis de datos, generar reportes visuales con gráficas de diversos parámetros y para incluir una interfaz gráfica más amigable para los usuarios sin experiencia en bioinformática. Durante la realización de esta tesis no se utilizaron estos programas. En lugar de ello se utilizaron scripts personalizados para cada análisis.

4.10 Determinación del sexo molecular

La determinación del sexo de los restos humanos es de gran utilidad en el área de antropología. Estos datos al ser analizados junto con el contexto arqueológico, objetos asociados o fuentes etnohistóricas, permiten una mejor interpretación de la división de labores y la estratificación social en las poblaciones del pasado (48). Tradicionalmente, en antropología, se utilizan los análisis morfométricos para asignar un sexo biológico a los restos humanos; éstos se basan en las características del esqueleto y toman en cuenta la forma y el tamaño del cráneo, la pelvis y la mandíbula (96). Sin embargo, la asignación de sexo no siempre es posible para todos los restos humanos encontrados. En algunos casos debido a la ausencia de estos huesos o a la etapa de desarrollo del individuo. Los huesos no presentan las características distintivas necesarias para la asignación de sexo cuando los individuos son juveniles. Por ello es técnicamente imposible la identificación del sexo de algunos individuos. También se da el caso en que los huesos pueden caer dentro de un rango medio de variación fenotípica entre los dos sexos. Esto dificulta la determinación con precisión del sexo del individuo (48).

En este contexto, la asignación del sexo mediante métodos de análisis molecular ha sido muy útil para los restos de esqueletos humanos no completos y de individuos juveniles. Uno de los métodos para determinar el sexo molecular se basa en la amplificación por

PCR del gen de la amelogenina del cromosoma X y su comparación con el pseudogén de la amelogenina presente en el cromosoma Y (97, 98). Sin embargo, este método no es el más recomendable cuando se trabaja con ADN, ya que la amplificación por PCR puede ser ineficiente debido a la naturaleza fragmentada del ADN. Además, la amplificación por PCR es susceptible a la contaminación por fuentes humanas externas que hayan estado en contacto con la muestra (99) donde una pequeña cantidad de contaminación puede sesgar los resultados de PCR. Otra forma para determinar el sexo molecular se basa en el análisis de los datos de NGS. Esto se realiza mediante la cuantificación de la fracción de lecturas asignadas al cromosoma Y con respecto a la suma de las lecturas asignadas al cromosoma Y y al cromosoma X (R_y). De acuerdo a este método, un valor de $R_y > 0.075$ corresponde a XY mientras que $R_y < 0.016$ corresponde a XX (Fig. 12), con intervalos de confianza del 95%. Se ha probado en conjuntos de datos humanos, tanto modernos como antiguos, de varias edades y sexos y se ha demostrado que es preciso utilizando un mínimo de 100,000 lecturas mapeadas al genoma humano, y menos preciso con 10,000 lecturas. Este análisis se realiza mediante un script de python que se encuentra disponible en el material suplementario de la publicación original (48).

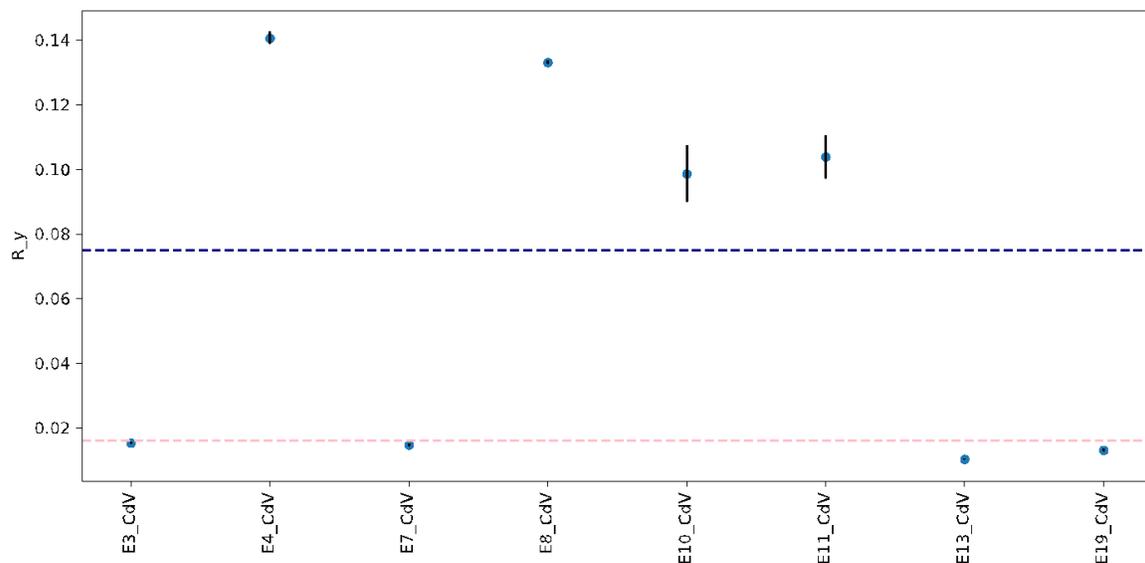


Figura 12. Determinación de sexo molecular mediante el análisis de datos NGS de genoma completo. En el eje de las abscisas se encuentran los ID de individuos antiguos, mientras que en el eje de las ordenadas se indica el valor de R_y . La línea punteada azul hace referencia al valor de $R_y > 0.075$ que corresponde a XY, mientras que la línea punteada rosa hace referencia al valor de $R_y < 0.016$ que corresponde a XX.

4.11 Análisis de marcadores uniparentales

4.11.1 Detección de variantes en el genoma mitocondrial

Como se menciona en la sección "mapeo al genoma humano", el genoma de referencia hg19 incluye el genoma mitocondrial. Dicho genoma debe ser sustituido por la referencia rCRS para realizar los análisis de ADNmt. El mapeo de lecturas del genoma mitocondrial se realiza junto con el del genoma nuclear para poder eliminar las NUMT (secuencias del ADNmt transpuestas al genoma nuclear). Esta operación evita realizar un llamado de variantes incorrecto (100). Para realizar el análisis del ADNmt se genera un nuevo archivo BAM con únicamente los alineamientos al genoma mitocondrial. Esto se realiza a partir del archivo de mapeo que pasó por todos los filtros de calidad y realineamiento.

A partir de este archivo se puede realizar un llamado de variantes, generar una secuencia fasta consenso o un archivo hsd para determinar el haplogrupo mitocondrial.

4.11.1.1 Llamado de variantes

El archivo de formato de llamado de variantes (vcf, 'variant calling format') puede ser generado usando la opción "mpileup" de SAMtools (82). SAMtools utiliza el archivo BAM y el genoma de referencia (en este caso la referencia de ADNmt rCRS). Samtools usa una selección de filtros de calidad de base y profundidad para generar un archivo vcf que contiene la información de las SNVs por posición. Para cada SNV se despliegan datos con su respectiva calidad de base, profundidad y otra información útil, como por ejemplo si la base es igual a la referencia o es una variante, si es un INDEL, en que cadena se encuentra, entre otra información.

4.11.1.2 Secuencia de consenso (fasta)

La generación de la secuencia consenso del genoma mitocondrial se puede realizar con el programa ANGSD (101) con la opción "-doFasta 2" que llama a la base más común por posición y genera un archivo de secuencia fasta.

4.11.1.3 Archivo Hsd

El formato de archivo 'hsd' fue creado para utilizarse con el programa Haplogrep (102). Este formato es un archivo de texto que contiene el rango del genoma mitocondrial que está cubierto por las lecturas. La información de las posiciones dentro de este rango que presentan una variante con respecto a la referencia se representa en forma de: posición-Base (es decir, 1650T). Con cada variante separada por tabulaciones.

Cuando el genoma mitocondrial tiene una baja profundidad de cobertura (<5x) es una buena práctica verificar las variantes visualizando el alineamiento. Esto se puede hacer usando herramientas de visualización de alineamientos como Tablet (103) que resalta las variantes de nucleótidos para que el usuario pueda identificarlas fácilmente. Esto permite al usuario decidir cuáles variantes tomar en cuenta. Esta decisión se realiza basándose en los polimorfismos que están respaldados por la mayoría de las lecturas y que no son resultado de daños en las moléculas de ADN (Fig. 13).

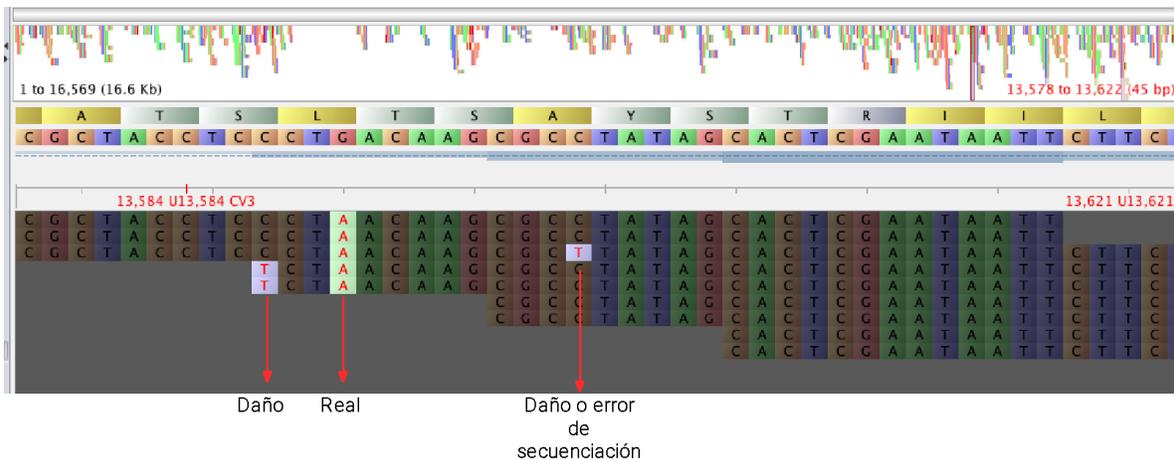


Figura 13. Visualización de los alineamientos mediante el programa Tablet (103). La parte superior muestra la cobertura seguida de la nomenclatura de aminoácidos y las bases nucleotídicas del genoma de referencia (rCRS). En la parte inferior se muestra el alineamiento de las lecturas de un genoma mitocondrial antiguo. Las bases que son diferentes a la referencia se encuentran resaltadas en otro color distinto a la referencia. Estos cambios en las bases pueden deberse a daño, errores de secuenciación o variantes reales en el genoma mitocondrial del individuo antiguo.

4.11.2 Asignación del haplogrupo mitocondrial

La asignación del haplogrupo mitocondrial se realiza mediante el uso de programas como Haplogrep (102, 104) que toma uno de los tres archivos mencionados anteriormente (vcf, fasta o hsd) y lo compara con el árbol filogenético mitocondrial de referencia llamado phylotree (22). Estos programas asignan el haplogrupo mitocondrial más probable junto con una puntuación de calidad que se calcula tomando en cuenta la presencia/ausencia de las variantes esperadas para el haplogrupo asignado. También tiene en cuenta si la secuencia presenta variantes consistentes con otros haplogrupos o variantes que no están presentes en el árbol de referencia. Haplogrep también cuenta con una interfaz visual que puede utilizarse para observar las ramas del árbol genealógico que condujeron al haplogrupo mitocondrial asignado (Fig.14).

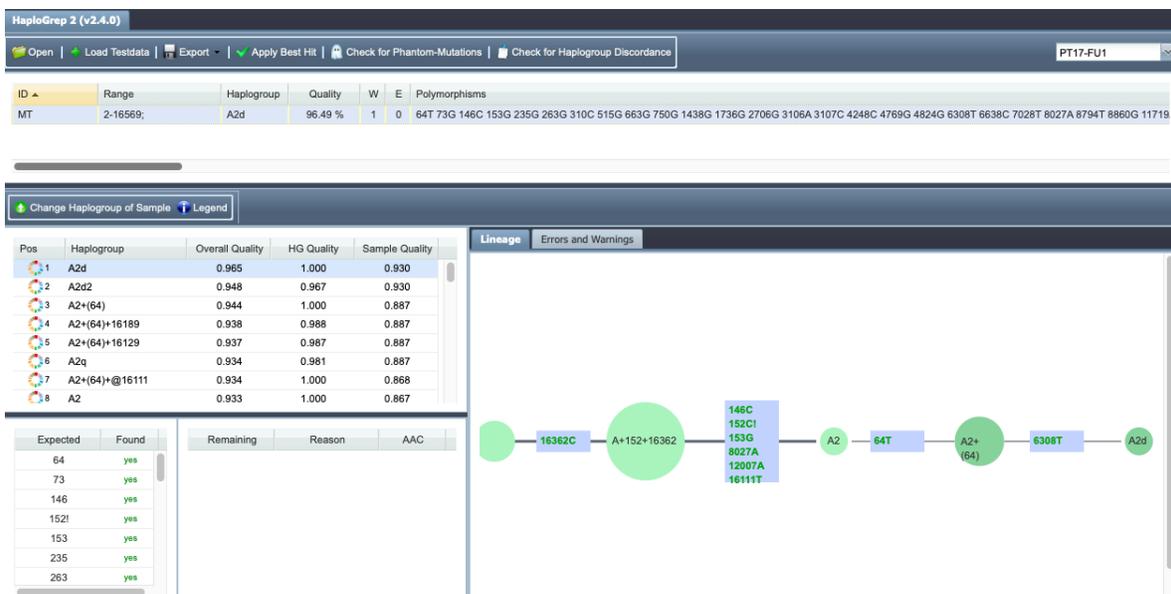


Figura 14. Visualización de la asignación de haplogrupo mitocondrial con Haplogrep para un genoma mitocondrial antiguo. La interfaz muestra la lista de variantes que se esperan para el haplogrupo asignado y cuáles se presentan en el genoma mitocondrial bajo análisis. En el panel izquierdo se muestra un árbol filogenético de estas variantes y cómo se llegó al subhaplogrupo asignado (A2d).

4.11.3 Asignación del haplogrupo de cromosoma Y

La identificación de variantes informativas del cromosoma Y a partir de los datos de NGS de ADN puede ser un reto dada la naturaleza de este cromosoma. El cromosoma Y es el cromosoma autosomal más pequeño y se encuentra en una sola copia por célula en individuos del sexo biológico masculino. Asimismo, la secuencia de este cromosoma tiene secuencias repetitivas y regiones de homología con el cromosoma X, las cuales no pueden ser analizadas adecuadamente cuando se utilizan secuencias cortas de NGS. Lo anterior resulta en que el análisis del cromosoma Y se base en un número reducido de SNVs informativos. La baja recuperación de datos del cromosoma Y debido a su pequeño tamaño y a que se encuentra en una sola copia representa un problema. Para superarlo se propuso una metodología en 2018 de captura en solución de las secuencias informativas del cromosoma Y para aumentar la cantidad de éstas en los datos de ADN (105).

El "llamado de variantes" en el cromosoma Y puede realizarse con el programa ANGSD (101) teniendo en cuenta la baja profundidad esperada para dicho cromosoma. El llamado de variantes en ANGSD se puede realizar con las opciones `-doHaploCall 1` ó `2`, muestreando una base al azar para cada posición cubierta en el cromosoma Y o tomando la base más común, respectivamente. Posteriormente, las variantes se comparan con el árbol filogenético del cromosoma Y utilizando un script previamente publicado (106). El haplogrupo Y asignado es el que comparte el mayor número de alelos derivados con la muestra.

4.12 Determinación de las relaciones de parentesco entre individuos

Determinar las relaciones entre los individuos de un mismo entierro, sitio o región es útil para comprender si la estructura social y los entierros de los individuos estaban influidos por las relaciones biológicas (dentro de una familia). Esta relación se denomina parentesco y tradicionalmente se ha estimado utilizando marcadores uniparentales y

microsatélites (STRs por sus siglas en inglés) (24, 107, 108). En este tipo de estudios se infiere que los individuos que presentan el mismo haplogrupo para cualquiera de los marcadores uniparentales pertenecen a la misma familia y los STRs pueden ayudar a confirmarlo. Estos marcadores pueden ser una buena aproximación para determinar el parentesco. Sin embargo, estos análisis podrían no ser un buen indicativo del grado de parentesco en algunos casos cuando el haplogrupo es muy común dentro de la población. Además, los STRs requieren un paso de amplificación por PCR que no es deseable en los estudios de ADN para reducir el potencial de falsos positivos debido a la contaminación.

Existen varios métodos para la identificación del parentesco a partir de datos de genotipado o secuenciación de genomas modernos (109–112). Desafortunadamente, estos métodos no pueden aplicarse directamente a los datos de ADN a menos que presenten una alta profundidad de cobertura, similar a la que se puede obtener en genomas modernos (>20x).

Por ello, recientemente se han desarrollado métodos para estimar el parentesco a partir de datos de NGS de ADN. Uno de estos métodos es GRUPS (113), el cual se basa en las distancias genéticas por pares esperadas entre individuos dependiendo del grado de parentesco. La tubería de análisis de GRUPS está disponible a través de una colección de scripts de python y R que toman como entrada un archivo vcf o BAM para todo el genoma. El método funciona bien con >3,000 SNVs de sitios sobrelapados entre los datos de ADN y el panel de referencia.

Un método similar es RECOAL (114). Este método tiene como base un modelo de verosimilitud que toma los SNV que se solapan entre todos los individuos estudiados para realizar la estimación de las frecuencias alélicas derivadas antiguas y calcular el coeficiente de parentesco.

Asimismo, el programa READ (115) se desarrolló como un método para determinar el parentesco entre individuos del pasado tomando en cuenta que la mayoría de genomas antiguos se recuperan a baja profundidad de cobertura. Para utilizar READ, se requiere un conjunto de SNVs cuidadosamente seleccionados (>1,000 SNVs). Esto implica que las SNVs no provienen de daño a la molécula de ADN, son de buena calidad y se encuentran como genotipos pseudo-haploides en formato tped/tfam. Estas SNVs deben

ser bialélicas y tomadas de un panel de referencia de una población estrechamente relacionada. El panel de referencia debe contar en su mayoría con individuos no emparentados de la misma población o población genéticamente cercana para hacer una calibración de los parámetros antes de realizar la estimación del parentesco de los individuos antiguos. READ es eficiente para detectar el parentesco de primer y segundo grado, siendo más preciso para el primero. Esto se debe a que ha sido diseñado para disminuir la tasa de falsos positivos a costa de aumentar la tasa de falsos negativos, lo que dificulta la detección del parentesco de segundo grado. Este programa fue utilizado para determinar el parentesco de los individuos prehispánicos analizados durante el trabajo de investigación doctoral de la presente tesis (Fig.15). Sin embargo, ha crecido el interés en este análisis y en los últimos dos años se han publicado otros métodos para la realización de este análisis. Estos métodos buscan mejorar la identificación de relaciones de parentesco de primer, segundo y tercer grado utilizando genomas de baja cobertura $>0.012x$ (116) y $>0.05x$ (117).

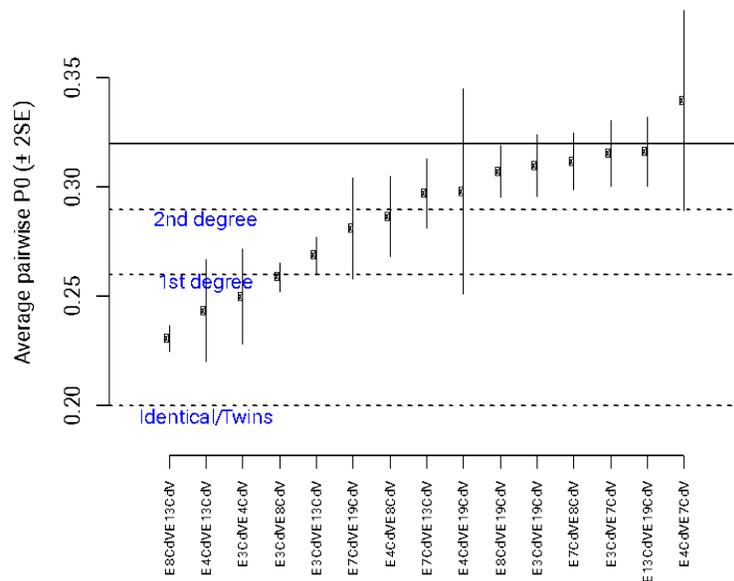


Figura 15. Gráfico que muestra las relaciones de parentesco mediante el programa READ. En las abscisas se muestran los Ids de los pares de individuos bajo análisis, mientras que en las ordenadas se muestra el promedio de alelos que son distintos en los pares de individuos. Debajo de cada línea punteada se menciona el parentesco asignado dependiendo de cada umbral. Por encima de la línea de 2do grado de parentesco, se encuentran los pares de individuos no relacionados.

4.13 Detección de contaminación por ADN moderno

Las muestras antiguas al ser extraídas de su lugar de origen tienden a estar expuestas a una potencial contaminación con ADN humano moderno. Esta contaminación puede suceder durante la excavación, la manipulación de la muestra en las colecciones antropológicas o en el procesamiento en el laboratorio si éste no cumple con los estándares internacionales del campo de la paleogenómica (11). Es importante poder determinar si los datos genómicos obtenidos de una muestra son auténticos o provenientes de la contaminación moderna debido a las múltiples fuentes de contaminación del ADN. Esto nos evita realizar inferencias erróneas acerca de los individuos o poblaciones antiguas bajo estudio. Generalmente un porcentaje de contaminación <5% es aceptable (118). La estimación de la contaminación en los datos de ADN se puede realizar mediante diversos métodos. Estos se basan en diferentes fracciones del genoma humano: genoma mitocondrial y genoma nuclear (para ambos sexos) y cromosoma X (para los individuos del sexo masculino).

4.13.1 Métodos de contaminación basados en el genoma mitocondrial

ContamMix es un método que fue desarrollado para validar la autenticidad de las lecturas mitocondriales en los datos de ADN (119). Este programa funciona con el archivo de alineamiento BAM que contiene sólo las lecturas mapeadas al genoma mitocondrial y con un archivo de la secuencia consenso fasta generada a partir del mismo BAM. Con esta información ContamMix compara las lecturas y la secuencia fasta consenso con una base de datos de 311 genomas humanos modernos de ADNmt (contaminantes potenciales) basándose en las posiciones de diagnóstico. ContamMix considera a cada lectura como proveniente de una fuente independiente y calcula la probabilidad de que el total de lecturas procedan de una única fuente o de diversas fuentes. ContamMix genera un archivo de texto con el valor de la fracción de secuencias auténticas y sus respectivos intervalos de confianza como resultado final. El método de ContamMix funciona mediante

scripts que no se encuentran disponibles públicamente. Por ello es necesario ponerse en contacto con el desarrollador Philip Johnson (plfj@umd.edu).

Otro método para estimar la contaminación basado en el genoma mitocondrial es Schmutzi (120), el cual funciona mediante un conjunto de programas. Además, es capaz de reconstruir iterativamente una secuencia consenso del genoma mitocondrial del individuo bajo estudio y del individuo que potencialmente haya contaminado la muestra. Al igual que contamMix, Schmutzi utiliza una base de datos de posibles contaminantes para realizar esta tarea. Una novedad importante de Schmutzi es que éste toma en cuenta las características del ADN (patrones de daño y lecturas cortas) para reconstruir con precisión el genoma mitocondrial. Por lo tanto, no funciona bien con bibliotecas tratadas con uracilo-ADN glicosilasa (121) o con ADN con baja frecuencia de cambios por desaminación de la citosina. Schmutzi ha sido probado en datos de ADN simulados y reales utilizando diferentes grados de contaminación y frecuencias de patrones de daño. Estas pruebas han revelado que Schmutzi funciona mejor que ContamMix en los datos simulados. El programa necesita dos archivos de entrada para realizar el análisis: el archivo de alineación (BAM) que contiene sólo las lecturas mapeadas al genoma mitocondrial y el genoma mitocondrial de referencia en fasta (rCRS).

Una limitación de ambos métodos es que no funcionan bien cuando los datos genómicos de las mitocondrias tienen una baja profundidad de cobertura (<5x). En estas situaciones ambos métodos tienden a sobreestimar la proporción de contaminación. Afortunadamente, los métodos de captura de ADNmt en solución funcionan bien y permiten la recuperación de genomas mitocondriales completos con una profundidad de cobertura >5x para la mayoría de las muestras.

Ambos métodos fueron utilizados durante la realización de la presente tesis y obtuvimos resultados similares. Las muestras que tienen discrepancias muestran una contaminación mayor estimada por ContamMix que la estimada por Schmutzi y provienen de muestras con una baja cobertura del genoma mitocondrial (Tabla 4). Este sesgo de sobreestimar la contaminación es común en ContamMix cuando la cobertura del genoma es baja. Por ello y tomando en cuenta que Schmutzi sí considera las características intrínsecas del ADN para calcular la contaminación, éste se utilizó para realizar las estimaciones reportadas (Anexo I, Tabla S1).

Tabla 4. Estimación de contaminación basada en el genoma mitocondrial, utilizando dos métodos: Schmutzi y ContamMix. En rojo se muestran los valores de contaminación por encima de 5%.

Región o sitio arqueológico	Individuo	Depth (x)	Haplogrupo	Calidad	Contaminación Schmutzi			Contammix		
					% contaminación	lim sup	lim inf	% contaminación	CI 2.5%	CI 97.5%
Sierra Tarahumara	F9	2,414.56	C1c1a	92%	8%	7%	9%	1.2%	1.5%	1.0%
	F9_pmd1	1,284.77	C1c1a	92%	2%	3%	1%	0.0%	0.0%	0.2%
	MOM6	71.67	C1b11	89%	2%	3%	1%	1.6%	3.0%	0.7%
Sierra Gorda	2417C	44.2	A2d	83%	2%	3%	1%	0.4%	2.3%	0.0%
	2417J	216.7	A2d	91%	2%	3%	1%	0.3%	0.8%	0.1%
	2417Q	222.5	B2l	73%	1%	2%	0%	0.6%	1.0%	0.3%
	333B	114.2	B2c	71%	2%	3%	1%	0.7%	1.1%	0.3%
	333C	148.8	B2c	76%	2%	3%	1%	1.3%	2.1%	0.7%
	333Q	32.4	A2d	85%	2%	3%	1%	0.9%	3.1%	0.1%
	333O	66.8	B2l	66%	2%	3%	1%	1.0%	2.3%	0.2%
	333A	7.4	B2c	62%	1%	2%	0%	19%	26%	13%
	7A	38.2	A2d	83%	2%	3%	1%	5%	8%	3%
	11R	28.2	D1m	84%	2%	3%	1%	11%	15%	8%
	G428	104.9	B2l	70%	2%	3%	1%	0.2%	1.1%	0.0%
	37AI	63.6	A2d	88%	2%	3%	1%	0.8%	2.1%	0.2%
	Peppi	46.1	A2d	78%	2%	3%	1%	3.4%	6%	2%
	Aude	2.23	ND	ND	ND	ND	ND	29%	38%	22%
	Michoacán	E1A	18.8	A2j1	84%	2%	3%	1%	2.2%	5.3%
E2		303.6	D1i2	97%	6%	5%	7%	0.4%	0.8%	0.2%
E2_pmd0		187.6	D1i2	89%	2%	3%	1%	0.3%	0.8%	0.1%
E4		168.0	D1i2	91%	1%	2%	0%	0.4%	1.0%	0.1%
M1		39.1	A2f2	85%	1%	2%	0%	2.2%	4.2%	1.0%
M2		12.1	C	74%	2%	3%	1%	37%	45%	31%
Cañada de la Virgen	E6	3.2	B2a	59%	1%	2%	0%	18.5%	26.0%	13.1%
	E7	90.0	A2+(64)	87%	1%	2%	0%	2.8%	4.3%	1.8%
	E8	98.7	B2	82%	1%	2%	0%	0.0%	0.4%	0.0%
	E10	45.5	C1b7	83%	1%	2%	0%	1.3%	2.8%	0.6%
	E11	5.7	B2	75%	1%	2%	0%	17.2%	24.6%	10.9%
	E19	169.6	D1m	95%	1%	2%	0%	4.0%	5.4%	2.8%

4.13.2 Métodos de contaminación basados en el genoma nuclear

El programa DICE (Inferencia demográfica con contaminación y error) (122), como su nombre indica es un método desarrollado para estimar los parámetros demográficos, las tasas de contaminación y las tasas de error en los datos de NGS de una muestra antigua. La inferencia de estos parámetros de este programa requiere los datos del genoma autosómico a una profundidad de cobertura >3x. Para realizar la estimación, requiere la intersección de los datos de ADN con un panel de SNVs bialélicos (>500,000) de la(s) población(es) potencial(es) contaminante(s). El método calcula la probabilidad de los genotipos del individuo antiguo, dados los parámetros demográficos y las frecuencias alélicas de los contaminantes potenciales. Inicialmente se diseñó para cuantificar la contaminación humana moderna en los datos de ADN de individuos arcaicos. Sin embargo también funciona bien para estimar la tasa de contaminación en el ADN de

humanos modernos cuando la potencial fuente contaminante procede de una población lejana genéticamente.

4.13.3 Métodos de contaminación basados en el cromosoma X

Otra forma de detectar la contaminación en muestras antiguas es a través del análisis de las variantes del cromosoma X en individuos del sexo masculino. El principio se basa en la naturaleza de los genotipos sexuales en la que los varones poseen una sola copia del cromosoma X. Por ello se espera que la información de este cromosoma sea haploide. Por lo tanto, la detección de alelos heterocigotos en el cromosoma X de individuos del sexo masculino sería indicativo de: 1) errores de secuenciación, 2) daños en el ADN o 3) contaminación humana.

Actualmente existen tres métodos para estimar la contaminación en muestras antiguas basándose en el cromosoma X. Uno forma parte de las herramientas ANGSD (101), contaminationX publicado en 2019 (123) y hapCon publicado en 2022 (124). La ventaja de los dos últimos es que han sido probados con diferentes tasas de contaminación, profundidades de cobertura y han sido comparados con ANGSD mostrando ser más precisos que éste.

4.14 Eliminación de la contaminación de los datos de ADN

La cuantificación de la contaminación en los datos de ADN es útil para decidir si éstos se utilizan o no en los análisis posteriores. Ésta es una decisión importante ya que el uso de datos contaminados podría dar lugar a sesgos en las inferencias y conclusiones realizadas sobre los individuos o sus poblaciones. Sin embargo, la opción de eliminar todo el set de datos de un individuo que presenta baja contaminación no es ideal. Esto es particularmente importante por el gran valor de las muestras antiguas, las cuales son patrimonio de México y se destruyen parcialmente para el estudio genómico.

Afortunadamente existe la herramienta computacional PMDtools (125) mediante la cual es posible separar las lecturas que provienen potencialmente de fuentes contaminantes de aquellas que son endógenas. Esta herramienta asigna una puntuación de daño post-mortem (PMDS) a cada secuencia, calculando la probabilidad de que un mismatch (C -> T o G->A) sea originado por: daño post-mortem, polimorfismo real o error de secuenciación en lugar de ser generado por una fuente de contaminación. PMDtools puede separar fácilmente las secuencias por encima de un umbral de elección en base al PMDS. Una desventaja de este método es que podría asignar un PMDS bajo a secuencias de ADN auténticas que se han conservado bien y que, por tanto, no presentan mucho daño.

4.15 Comparación de genomas de ADN con paneles de referencia actuales

Algunos de los descubrimientos más interesantes en el campo de la paleogenómica humana proceden de la comparación entre los genomas humanos antiguos y actuales. Esta comparación nos permite conocer las diferencias genéticas de las personas que vivían en una determinada región geográfica y su relación genética con las poblaciones actuales que habitan la misma región o regiones diferentes. Además, puede darnos información sobre migraciones, flujo genético, reemplazos poblacionales, cambios demográficos, cambios de diversidad, procesos de selección e introgresión arcaica.

Sin embargo, un punto a tener en cuenta es que los datos genómicos obtenidos a partir de ADN a menudo tienen una cobertura incompleta del genoma y baja profundidad. Esto implica una limitante en la cantidad de sitios del genoma que pueden ser analizados. Por lo general, para superar este problema, se analizan únicamente los sitios del genoma para los que existe un conocimiento previo de la variación genómica en las poblaciones actuales.

La comparación de los genomas antiguos con aquellos de poblaciones modernas puede realizarse con datos disponibles públicamente de diversas poblaciones humanas actuales. Así, los genomas modernos pueden utilizarse como panel de referencia para

contextualizar al individuo antiguo. Con una profundidad de cobertura media superior a 3x es factible llamar genotipos o probabilidades de genotipos (*genotype likelihoods*) (101) en los individuos antiguos. Por otro lado, cuando la cobertura es más baja, esto no es recomendable. En cambio, se puede optar por intersectar los datos de ADN con las SNV del panel de referencia. Una práctica común es muestrear aleatoriamente una lectura para cada SNV y registrar el genotipo como homocigoto para la base observada. A esto se le llama genotipo pseudohaploide. El llamado de genotipos pseudohaploides también se realiza en el panel de referencia para analizar los datos bajo las mismas condiciones (38). La intersección de los datos se puede hacer utilizando la función `mpileup` de `samtools` si se proporciona una lista de los sitios que deben llamarse, es decir, los del panel de referencia. A continuación el archivo de salida de `mpileup` debe ser formateado para ser leído por el software PLINK (109), que puede fusionar los datos de los genotipos del panel de referencia y los de ADN. Los archivos en formato `plink` pueden utilizarse para realizar varios de los análisis comparativos que se describen a continuación.

4.15.1 Afinidades genéticas

Una de las preguntas más comunes cuando se estudian individuos antiguos es a qué población actual son genéticamente más cercanos. Conocer esto permite comprender mejor la historia poblacional de dichos individuos y de la propia población actual. Además de probar hipótesis específicas sobre la historia de la población, incluyendo el poblamiento de los continentes, e inferir expansiones o cuellos de botella poblacional, migraciones, entre otros eventos demográficos.

Una forma de visualizar las afinidades genéticas de los individuos antiguos con las poblaciones actuales es a través de análisis no supervisados como el Análisis de Componentes Principales (126, 127) y ADMIXTURE (128) (modo no supervisado), que no se basan en el conocimiento previo de las etiquetas de las poblaciones dentro del análisis y realizan una agrupación de los individuos basándose únicamente en sus datos genómicos.

4.15.1.1 Análisis de componentes principales (PCA)

El análisis de componentes principales (PCA) se ha aplicado al campo de la genética de poblaciones ya que permite reducir la complejidad del análisis de genotipos de miles de SNVs pertenecientes a cientos o miles de individuos. Para realizar el análisis de PCA en datos de ADN se utiliza el archivo de genotipos pseudohaploides en formato plink que contiene la información tanto del panel de referencia como de los individuos antiguos. El PCA reduce la complejidad de los datos genéticos a menos dimensiones y puede revelar patrones imprevistos subyacentes a la variación genética. Específicamente para genética de poblaciones se desarrolló el paquete EIGENSOFT (rutina smartpca), el cual fue en principio utilizado para tomar en cuenta la estructura genética de las poblaciones en los estudios de GWAS (127).

Al momento de realizar un PCA con datos de ADN es importante tomar en cuenta que si la muestra tiene una baja cobertura existirá una cantidad elevada de genotipos faltantes. Esto afectará el resultado del análisis. Para superar este problema se ha convertido en una práctica habitual "proyectar" los datos antiguos sobre los componentes principales del panel de referencia (129, 130). Esto se hace mediante métodos como Procrustes (131) o lsqproject (dentro de smartpca de eigensoft) (127) que transforman los datos de ADN sobre los componentes principales del panel de referencia. Esto permite visualizar las afinidades genéticas del individuo antiguo en el contexto de la variación genética actual.

El PCA es un método que ha resultado útil en la genómica humana ya que permite observar las afinidades genéticas entre distintas poblaciones tanto del presente como del pasado. Sin embargo este análisis presenta limitaciones ya que no toma en cuenta las diferencias genéticas que pueden ocurrir en los individuos antiguos respecto de los actuales simplemente por los cambios en las frecuencias alélicas respecto al tiempo. Para resolver esto en 2020 se publicó el método de análisis factorial (FA, por sus siglas en inglés) el cual toma en cuenta la antigüedad de las muestras para corregir los valores de los individuos antiguos de acuerdo a la deriva génica a través del tiempo (132). Este programa fue puesto a prueba tanto en datos simulados como en datos de muestras europeas antiguas, teniendo resultados similares a aquellos obtenidos en los estadísticos F y con una mayor precisión que las proyecciones de las muestras antiguas comúnmente utilizadas en los PCA.

Asimismo, existe el método llamado “*missing DNA PCA*” (mdPCA, por sus siglas en inglés) que se ha desarrollado para mejorar los análisis de componentes principales corrigiendo el efecto negativo de los datos de genotipos faltantes en los genomas antiguos. Esta corrección se logra mediante la comparación de las distancias genéticas entre todas las muestras y funciona mejor en genomas que tienen una cobertura $>0.5x$. Este método aún no ha sido publicado pero puede encontrarse en su página de github (<https://github.com/AI-sandbox/mdPCA>).

Para la publicación del artículo científico de esta tesis (sección 6 y Anexo I), se utilizaron los tres métodos de reducción de dimensionalidad: PCA, FA y mdPCA. mdPCA y FA fueron los mejores en agrupar los individuos antiguos con el panel de referencia actual y teniendo una congruencia con los análisis de f3-outgroup.

4.15.1.2 Componentes ancestrales

El aislamiento de las poblaciones y la ausencia de apareamiento aleatorio entre diferentes regiones geográficas da lugar a la acumulación de cambios genéticos y fenotípicos entre las poblaciones que pueden ser contabilizados como estructura de la población (133). El estudio de la estructura de la población mediante el análisis de los componentes ancestrales es útil en el campo de la paleogenómica para relacionar individuos de ascendencia desconocida con poblaciones actuales genéticamente cercanas.

STRUCTURE (126) es uno de los primeros métodos desarrollados para inferir la estructura de la población a partir de datos genéticos. Se basa en el supuesto de que las poblaciones se caracterizan por un conjunto de frecuencias alélicas y que cada una de estas poblaciones podría haberse originado a partir de una o más poblaciones ancestrales (K). Dado el número K de poblaciones ancestrales y las frecuencias alélicas de cada una, STRUCTURE calcula la probabilidad de que un genotipo individual pertenezca a una o más poblaciones ancestrales. El modelo también asume que estas poblaciones ancestrales hipotéticas están en equilibrio Hardy-Weinberg y que las SNV no están en desequilibrio de ligamiento (LD). Otro programa similar es ADMIXTURE (128), el cual se desarrolló posteriormente y utiliza el modelo estadístico de STRUCTURE. La ventaja de

ADMIXTURE es que tiene la capacidad de estudiar conjuntos de datos mucho más grandes y con una mejora en los tiempos de ejecución en comparación con STRUCTURE.

Para realizar el análisis de ADMIXTURE en datos de ADN a baja cobertura es necesario usar datos de genotipos pseudohaploidizados. En estos análisis las muestras antiguas pueden tener muchos datos faltantes con respecto al panel. Para superar esto el software NGSadmix (134) se desarrolló para trabajar con datos de baja cobertura y se ha demostrado que supera a los otros métodos utilizando datos simulados y reales con profundidades de cobertura variables ($>1x$). Además, a diferencia de trabajar con genotipos pseudohaploides, NGSadmix utiliza verosimilitudes de genotipos.

Es importante señalar que estos programas utilizan un valor de K para realizar la agrupación. El valor K es proporcionado por el usuario, que se espera que tenga una idea del número de poblaciones ancestrales bajo estudio. Sin embargo, si el usuario no tiene en mente el número K que debe utilizar puede ser útil analizar el conjunto de datos con diferentes valores K y elegir el K con el menor error de validación cruzada (opción `-cv` en ADMIXTURE). Además, es una buena práctica visualizar cómo cambian los clusters de componentes ancestrales de un valor K a otro para ver si los cambios tienen sentido con el conocimiento previo sobre las poblaciones del panel de referencia. Estos programas generan un archivo de salida para cada K, el cual contiene la fracción de cada componente ancestral que presenta cada individuo. Para obtener el gráfico de ancestría, además del software R, existen otros programas que han sido desarrollados para graficar este tipo de datos como PONG (135) y AncestryPainter (136).

Es importante tener en cuenta que este tipo de análisis pueden ser una buena aproximación para explicar la variación genética entre individuos. Sin embargo, este análisis no es una buena aproximación para explicar los procesos de mezcla entre diferentes poblaciones. Por lo tanto, los resultados de componentes ancestrales deben tomarse como exploratorios y cualitativos ya que las agrupaciones observadas podrían ser el resultado de diferentes historias de mezcla que no pueden evaluarse con un único análisis. Para evitar la sobreinterpretación de los resultados se generó el programa badMIXTURE (137) con el fin de evaluar si los modelos utilizados se ajustan a los datos y tener más confianza en los resultados obtenidos.

4.15.2 Deriva genética compartida

La deriva genética se refiere al cambio aleatorio en las frecuencias alélicas dentro de una población que puede dar lugar a grandes cambios después de muchas generaciones (133). El estudio de la deriva genética compartida entre poblaciones puede ser informativo sobre su historia genética compartida, la separación de las poblaciones y las migraciones.

El análisis para evaluar la deriva genética compartida en el campo del ADN_a puede dar una idea de la historia de la población del individuo antiguo y su relación con otras poblaciones antiguas o actuales. Este análisis puede realizarse mediante los dos métodos que se describen a continuación.

4.15.2.1 TreeMix

TreeMix (138) es un modelo estadístico que tiene como objetivo encontrar un gráfico de árbol que informe sobre las relaciones entre poblaciones. Éste evalúa las divergencias y migraciones (o flujo genético) entre las poblaciones. El análisis de TreeMix inicia con un árbol filogenético de las poblaciones de máxima verosimilitud, incluyendo una población externa para enraizar el árbol. A continuación Treemix utiliza los cambios en las frecuencias alélicas con respecto a la población ancestral para realizar los cambios que mejor se ajusten al modelo (incluidas las migraciones). Con este enfoque estadístico es fácil identificar la población con la que el individuo antiguo tiene la mayor historia genética compartida y si la población antigua participó en un evento migratorio. Para medir el ajuste del modelo TreeMix puede calcular una matriz de residuos. Dicha matriz es informativa para saber si una migración o la relación entre un par de poblaciones fueron sobreestimadas o subestimadas.

4.15.2.2 Prueba de outgroup-f3

Otro método para analizar la deriva genética compartida es a través de la prueba estadística de 3 poblaciones conocida como f₃-outgroup e implementada en la función qp3pop del paquete AdmixTools (139). Esta prueba se basa en el estudio de tres

poblaciones: 1) grupo externo, 2) individuo o población antigua y 3) población de comparación (X). El grupo externo debe pertenecer a una población que se espere que tenga la misma relación genética con ambas poblaciones. Tomando como referencia las frecuencias alélicas en la población externa y los cambios en las frecuencias alélicas en la población de comparación y el individuo antiguo, el programa calcula la deriva genética compartida entre las poblaciones bajo prueba. El resultado del análisis se reporta como el valor de la longitud de la rama compartida. Por lo tanto, el tamaño de la rama es proporcional a la deriva genética compartida entre las dos poblaciones (Fig. 16).

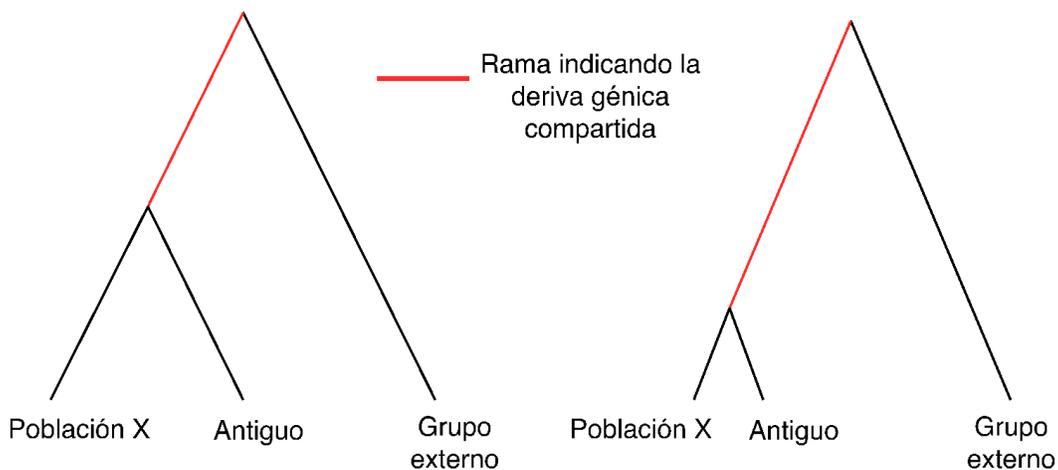


Figura 16. Simplificación de la prueba outgroup-f3. La longitud de la rama en rojo indica la cantidad de deriva genética compartida entre la población X y el individuo o población Antigua.

4.15.3 Detección de mezcla genética

La detección de mezcla en poblaciones del pasado es posible a través de la prueba del estadístico-D (D-statistics), también conocida como ABBA-BABA, descrita por primera vez en (140, 141) e implementada en el paquete AdmixTools (139). Esta prueba utiliza la información de los genotipos de 4 poblaciones (W,X,Y,Z), donde Z es un grupo externo y Y es el individuo antiguo (llamado aquí población de prueba). Mientras que W y X son poblaciones de las que se evaluará si han tenido flujo genético (mezcla) o no con la población de prueba. La prueba toma aleatoriamente muestras de alelos para cada una de las cuatro poblaciones, observando qué variantes presenta cada una en la misma posición. Cuando se tienen dos alelos en un sitio y los alelos de la población W-Z; X-Y son iguales, estos eventos son llamados ABBA. Por otro lado, en un sitio bialélico cuando

tenemos que $W-Y;X-Z$ son iguales, son llamados BABA. A continuación, se cuantifican el número de eventos ABBA y BABA y se calcula la probabilidad de encontrar alguno de estos más veces que el otro sólo por azar. Al final, se calcula el valor del estadístico D. Un valor de $D=0$ indicará que los sitios ABBA-BABA serán prácticamente similares por lo que no se ha producido ningún evento de mezcla entre W o X y la población de prueba. Por otra parte, cuando el valor de $D>0$ es un indicativo de que se produjo un evento de mezcla entre W y la población de prueba. Y por último, cuando el valor de $D<0$, indica que la mezcla se produjo entre X y la población de prueba. La significación estadística del valor D se computa mediante el cálculo de la puntuación Z, donde un valor absoluto >3 es un buen indicativo de mezcla entre W o X con Y. La prueba ABBA-BABA también se puede realizar mediante el paquete ANGSD (101). En este programa el cálculo se hace al revés. Por ello, $D>0$ significa mezcla entre X e Y mientras que $D<0$ significa mezcla entre W e Y. ABBA-BABA es útil no sólo para el evento de mezcla. También es útil para probar hipótesis específicas en las que se desea conocer si una de dos poblaciones actuales tiene mayor relación con el individuo antiguo (142) (Fig.17).

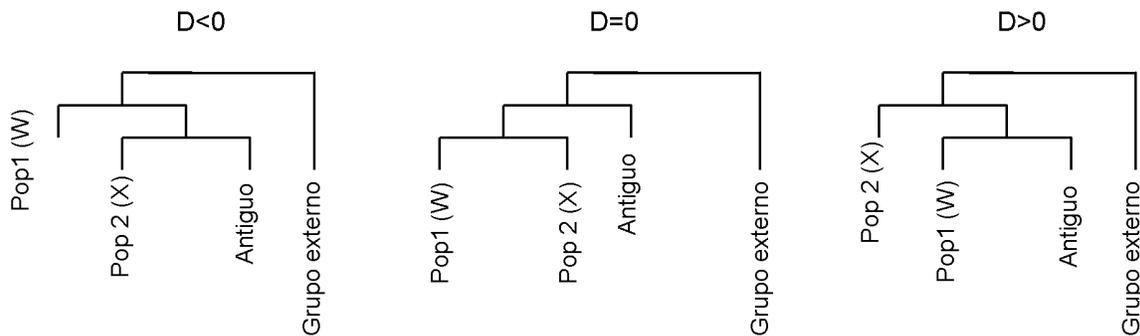


Figura 17. Análisis del estadístico D. Árboles filogenéticos esperados dependiendo del valor del estadístico D.

4.16 Diversidad genética

La diversidad genética se refiere a la variación en la secuencia de ADN entre diferentes individuos o poblaciones y tiene una influencia en la adecuación de las especies. La diversidad genética está directamente relacionada con el tamaño efectivo de la población (N_e) y puede verse afectada por varios factores como la deriva genética, los cuellos de

botella poblacionales, la fijación, los efectos fundadores, entre otros (143). La disminución de heterocigosidad puede identificarse a través de la pérdida de variantes alélicas y la aparición de tramos de homocigosidad largos (ROH).

En el campo de la paleogenómica humana es muy interesante observar los cambios en la diversidad genética entre las poblaciones antiguas y las actuales, y cotejarlas con los conocimientos arqueológicos y los registros históricos. Esta estimación se ha evaluado mediante el cálculo de los niveles de heterocigosidad, utilizando datos de secuenciación de baja profundidad ($>2x$) (144) mediante el paquete ANGSD descrito en (145). Para este análisis ANGSD excluye los sitios con transiciones para evitar tomar en cuenta cambios en las bases que hayan sido producto de daño. Asimismo, durante el llamado de variantes utiliza probabilidades de genotipo para tener en cuenta la incertidumbre. La implementación requiere un tamaño de muestra >25 y necesita un genoma de referencia ancestral. El paquete ANGSD también puede calcular varios estimadores de la diversidad genética (es decir, el estimador de Watterson (146), el estimador D de Tajima (147), el estimador de Fu & Li (148), entre otros). Sin embargo, estos requieren una mayor cobertura.

La herramienta ATLAS también cuenta con una implementación para estimar la heterocigosidad como se describe en (149). A diferencia de ANGSD, la estimación se realiza mediante el análisis directo del archivo de alineación (BAM) y no requiere un genoma de referencia ancestral. ATLAS considera el sesgo potencial que pueden introducir los errores de secuenciación y los daños *post mortem*. Atlas funciona bien con datos de cobertura $>1x$.

Como se mencionó anteriormente, una forma indirecta de estimar la diversidad genética es a través de la cuantificación de las regiones en el genoma que son homocigotas. Es decir, entre mayor sea la cantidad y longitud de estas regiones en el genoma (*runs of homocigosity* o ROHs) menor será la diversidad genética.

El programa ROHan (150) ha sido desarrollado tanto para estimar la heterocigosidad como para identificar los fragmentos del genoma en ROHs. Estos resultados se utilizan posteriormente para calcular el estimador de Watterson el cual estima el número de sitios polimórficos en una población. ROHan utiliza como entrada el archivo de alineación (BAM)

al igual que ATLAS. ROHan ha sido probado utilizando conjuntos de datos simulados con diferentes proporciones de patrones de daño, profundidad de cobertura y tamaños de población. Además ROHan ha sido probado con muestras empíricas modernas y antiguas. La comparación de este programa con ANGSD y ATLAS ha demostrado que ROHan tiene una mayor precisión en la estimación de la heterocigosidad. Sin embargo, una limitación importante de este programa es la cobertura de profundidad mínima necesaria $>8x$ para un análisis preciso. Esto no es común en datos de ADN.

La herramienta hapROH se publicó en 2021 para detectar corridas de homocigosidad en genomas antiguos (151). Este método detecta los fragmentos en ROH con longitudes mayores a cuatro centimorgans(cM) y hasta 300 cM. Esto resulta de bastante utilidad para analizar indirectamente los niveles de diversidad y estimar el tamaño de la población del pasado. Un mayor número de segmentos en ROH indica un menor tamaño de la población y menor número de sitios heterocigotos. Esto significa una menor diversidad genética. Asimismo, se puede inferir el parentesco de los padres de cada individuo ya que entre mayor sea la longitud de estos segmentos, mayor será la relación de parentesco entre el padre y la madre. Estos análisis ayudan a comprender las dinámicas sociales en la antigüedad. Para realizar el análisis, hapROH utiliza un conjunto integrado de 1240K SNVs (152) para realizar llamadas de genotipos pseudohaploides y un panel de referencia de 5,008 haplotipos humanos actuales de poblaciones de todo el mundo (153). El método requiere que los genomas antiguos tengan un solapamiento mínimo de 400K SNVs de las 1240K SNVs del panel de referencia.

Por último, la estimación de la diversidad condicional de nucleótidos (CND por sus siglas en inglés) es otra aproximación que se desarrolló para analizar los niveles relativos de diversidad genética en las poblaciones antiguas (154). Este análisis consiste en la cuantificación de las diferencias alélicas presentes entre pares de individuos de la misma población. Las posiciones del genoma que son tomadas en cuenta para esta cuantificación son aquellas que se conoce que varían en individuos de origen africano. Se excluyen aquellos sitios que presentan transiciones ($\sim 1,938,919$ SNVs) con el objetivo de excluir del análisis aquellas variantes que provengan de daño *post mortem* o errores de secuenciación. El valor de CND obtenido es más alto mientras más diversa sea una población ya que existirán mayores diferencias alélicas entre pares de individuos. El

análisis de CND llevado a cabo para la presente tesis se realizó mediante un script casero ya que la publicación del artículo que contiene el método no contiene un script disponible.

4.17 Análisis demográfico

Los análisis demográficos en la genómica de poblaciones se utilizan ampliamente para estudiar la dinámica de las poblaciones, estimar el tamaño efectivo de las mismas, identificar eventos de declive poblacional (cuellos de botella), migraciones entre poblaciones y tiempos de separación entre otros procesos evolutivos. Existen varios programas que permiten inferir la historia demográfica a partir de datos de ADN modernos. Sin embargo, en el campo de la "Paleogenómica de poblaciones" estos análisis son más complicados de implementar debido a la baja profundidad que suelen tener los genomas antiguos. Por ello, se han desarrollado programas que evalúan estos parámetros en genomas de menor cobertura.

DICE (122) es un método (mencionado anteriormente en la sección de contaminación) basado en las probabilidades del genotipo que fue desarrollado para estimar los parámetros demográficos como separación de poblaciones y tiempos de divergencia en conjunción con tasas de error de secuenciación y de contaminación. La ventaja de este método es que permite inferir estos parámetros en conjuntos de datos que contienen contaminación moderna. La población de la fuente contaminante puede ser detectada mediante un panel de posibles contaminantes para posteriormente realizar las estimaciones demográficas únicamente del individuo de interés. Este método ha probado tener buenas estimaciones de los parámetros demográficos cuando se utilizan genomas nucleares con una profundidad de cobertura $>3x$ tanto en sets simulados, genomas modernos y genomas antiguos de Neandertal. Otra ventaja de este método es que puede emplearse también en datos de genotipos de $>500,000$ SNVs.

Por otro lado, el método Fastsimcoal2 descrito en (155) se basa en el análisis del Espectro de Frecuencia por Sitio (SFS por sus siglas en inglés) para estimar parámetros demográficos como tiempos de coalescencia y tasas de migración en escenarios demográficos complejos. Fastsimcoal2 fue desarrollado para cubrir la necesidad de un

software que fuera capaz de manejar regiones genómicas largas como las que se generan actualmente tanto para genomas modernos como para genomas antiguos. Este programa ha sido implementado con éxito en un estudio donde se analiza la historia demográfica en el altiplano andino a partir de datos genómicos antiguos y actuales (156).

El análisis de ajuste de grafo de mezcla (Admixture graph fitting o qpGraph) está diseñado para estudiar la historia de la mezcla entre poblaciones a través de modelos de relaciones poblacionales que se establecen con el conocimiento previo y son puestos a prueba con este análisis (139). qpGraph analiza cada modelo dado para evaluar si éste se ajusta o no a los patrones de correlación de las frecuencias alélicas observadas en las poblaciones, asignando una puntuación del ajuste de cada árbol junto con sus valores estadísticos. Este paquete forma parte de las herramientas ADMIXTOOLS y ha sido ampliamente usado en el campo de la Paleogenómica. Nosotros lo utilizamos para tres análisis que forman parte del artículo de titulación de esta tesis (sección 6 y Anexo I).

4.18 Continuidad poblacional

Cuando se estudian restos antiguos es común preguntarse con cuál población del presente los individuos antiguos están mayormente relacionados. Por ello resulta interesante analizar mediante la genómica si la población antigua presenta o no una continuidad genética con la población actual que vive en la misma región o regiones aledañas. Para este fin, de forma inicial se pueden realizar los análisis exploratorios de PCA (127) y ADMIXTURE (128). Estos análisis nos dan una idea de con cuales poblaciones actuales o antiguas los individuos antiguos estudiados tienen mayor afinidad genética. Asimismo, de una manera más cuantitativa, se pueden utilizar los estadísticos-f como outgroup-f3 (139) para determinar las poblaciones con las que existe mayor deriva genética compartida. Por otro lado, el análisis de D-statistics y qpGraph (139) puede usarse para probar topologías de relaciones filogenéticas entre las poblaciones del pasado y/o el presente. Por otro lado, otro método para determinar si existe continuidad genética se basa en el análisis de IBD pero éste no ha sido ampliamente utilizado (157).

Otro enfoque aún más específico fue desarrollado en 2018 para estudiar directamente la continuidad genética de las poblaciones antiguas y las actuales (158). Este método de continuidad genética se basa en el análisis de las frecuencias alélicas de la población moderna y se compara con el conteo de alelos o frecuencias alélicas de los individuos antiguos. El modelo estima los tiempos de divergencia entre ambas poblaciones mediante la evolución de los alelos atrás en el tiempo para la población actual y posterior en el tiempo para el individuo antiguo. Se infiere como hipótesis nula que la población actual bajo análisis y el individuo antiguo pertenecen a una población continua, por lo que el rechazo de la hipótesis nula significa que no hay continuidad poblacional. Para llevar a cabo este análisis se puede utilizar la información genética de un solo individuo del pasado y compararla con las frecuencias alélicas de una población del presente, siempre y cuando el genoma del individuo antiguo tenga una cobertura $>0.5x$. La precisión en la estimación de los parámetros se puede mejorar al utilizar un conjunto de varios genomas de baja profundidad, o al utilizar uno o más genomas de alta profundidad, siendo el primer caso el más recomendado por el autor del método.

4.19 Faseo de secuencias de ADN

El faseo de secuencias es un método que fue desarrollado inicialmente para su uso en genomas actuales con el fin de detectar los haplotipos heredados a partir de los progenitores y de esta manera poder analizar SNVs que han sido heredadas en conjunto. El análisis de datos faseados permite realizar estudios demográficos y de selección con una mayor resolución. El faseo consta básicamente de utilizar haplotipos de un panel de referencia para a partir de estos inferir qué alelos corresponden a qué progenitor, generando así datos de haplotipos a partir de genotipos. El uso de esta aproximación debe hacerse tomando en cuenta que existen diversos parámetros que pueden alterar de manera negativa la precisión del faseo, resultando en inferencias erróneas, comúnmente cuando se usa en genomas modernos la tasa de error es de $<5\%$ (159). Y aunque es un método que no se usa comúnmente en estudios de ADN, se ha observado que los haplotipos generados mediante faseo mejoran los análisis poblacionales inclusive cuando se tiene una tasa de error en el faseo de $\sim 13\%$ (160). Los parámetros que influyen en la

calidad del faseo y que deben tomarse en cuenta son: el porcentaje de contaminación, la profundidad de cobertura, el daño de las secuencias, la historia demográfica de los individuos del pasado y el panel de referencia moderno utilizado para realizar el faseo. Se ha observado que la calidad del faseo se ve disminuida principalmente cuando: 1) se utilizan genomas de baja profundidad de cobertura ($<5x$), 2) cuando las secuencias presentan un porcentaje de daño alto, 3) se utilizan paneles de referencia de poblaciones genéticamente lejanas, 4) se utilizan datos genómicos antiguos con contaminación $\geq 5\%$ y 5) la población antigua pasó por un cuello de botella. Sin embargo, el faseo es una herramienta útil que mejora los resultados obtenidos mediante análisis poblacionales como el PCA, cuando el faseo se realiza con genomas de una cobertura de $>5x$ (160). En un futuro se espera disponer de un catálogo de haplotipos antiguos que mejore la calidad de los faseos que actualmente se llevan a cabo con los paneles de referencia actuales, resultando así en la expansión del tipo de análisis que se pueden llevar a cabo actualmente con datos de ADN.

5 El poblamiento de América estudiado a través de la Paleogenómica

El poblamiento inicial de América es un tema que ha fascinado a las y los investigadores de diversas disciplinas por décadas. Esto es en gran medida por ser el último continente en ser poblado por el humano anatómicamente moderno y por la escasez de registros arqueológicos que den indicios de cómo sucedió dicho evento. Al estudiar el poblamiento de América, las y los investigadores se han planteado principalmente estas preguntas: ¿De dónde vinieron los primeros pobladores de América?, ¿Quiénes eran?, ¿Qué rutas usaron para entrar al continente y qué rutas siguieron una vez en América? y ¿Cuándo ocurrieron estos eventos?.

Gracias a décadas de estudios arqueológicos y antropológicos existen varias hipótesis sobre las características del poblamiento de este continente. Los modelos propuestos sugieren una o varias oleadas de migración humana, así como diferentes rutas que los primeros pobladores pudieron haber seguido en su camino. Sin embargo, la historia es más compleja de lo que se pensó en un inicio. Aún falta camino por recorrer para comprender cómo ocurrieron estos eventos a pesar de las numerosas investigaciones y conocimiento generado.

En este capítulo se expone cómo han contribuido los estudios paleogenómicos, incluido el desarrollado en la presente tesis, al entendimiento del poblamiento de América. Para ello, es importante explicar primero brevemente las hipótesis que se han propuesto desde otras disciplinas y que en los últimos años han sido enriquecidas o refutadas con los estudios genómicos.

La llegada de europeos a América ocasionó cambios demográficos, sociales y culturales en las poblaciones nativas. Dentro de estos cambios, se destruyeron registros históricos

de algunos pueblos y las historias acerca de su origen. Las historias que han sobrevivido son pocas y son de origen mitológico como lo es la historia de los Aztecas provenientes de Aztlán (161) o el origen del hombre narrado en el Popol Vuh (162). Al llegar los europeos al continente Americano se encontraron con pueblos que tenían creencias religiosas distintas a las de ellos. Los europeos creían que ellos contaban con la verdad y decidieron evangelizar a los pueblos nativos, desconectándolos así de su historia, su pasado y su cultura. Por ello ahora la concepción de la historia poblacional de algunos pueblos es el resultado de la mezcla de sus creencias precolombinas y las creencias religiosas actuales.

Los conquistadores europeos en su llegada a América se plantearon la pregunta de ¿cómo habían llegado los pueblos nativos al territorio Americano territorio siendo que éste estaba tan alejado de los otros continentes?. El jesuita José de Acosta narra sus posibles explicaciones acerca del origen de los nativos en la obra “La historia Natural y Moral de las Indias” publicada en 1589. Estas explicaciones toman en cuenta que los españoles creían en la idea monogenista bíblica de que todos los hombres provienen de un mismo origen. Por ello, se infiere que los pobladores de América tendrían que ser descendientes de europeos o asiáticos. Entre las posibles explicaciones que José de Acosta narra acerca del origen de los nativos de América, menciona: 1) Llegaron en navíos, 2) provienen de Atlántida, 3) Llegaron por tierra, o 4) vienen del linaje de los judíos y fueron llevados por Dios al continente Americano. Al analizar cada una de ellas, llega a la conclusión de que la opción 3 es la más lógica. Debido a ello propone que en alguna parte de América ésta se une con los otros continentes o que está tan cercana que por medio de este camino pudieron pasar las personas a América (163).

A principios del siglo XIX se inició el estudio del poblamiento de América desde una perspectiva científica. Esa labor se inició a la par del estudio de los esqueletos de diversas regiones del mundo para tratar de entender las relaciones que había entre ellos. Algunas de estas investigaciones fueron conducidas con poca ética e incluso contribuyeron a ideas racistas. Esto tuvo un impacto negativo mundialmente, como lo fue la creencia de la supremacía blanca, basada en la creencia de una mayor capacidad craneal de individuos provenientes de ciertas regiones y la inferencia de que esto confería una mayor inteligencia (164). Una de las clasificaciones propuestas en base al tamaño de los cráneos, describe 3 tipos de ellos: “Caucasoides”, “Negroide” y “Mongoloide” (165). Estas

observaciones junto con el análisis de cientos de esqueletos de diversas regiones del mundo condujeron a la clasificación del fenotipo del cráneo de Nativos Americanos dentro del tipo Mongoloide. Esta inferencia implica que los Nativos Americanos eran descendientes de poblaciones asiáticas. Una de las principales características observadas en la dentadura de Nativos Americanos fue que los incisivos presentaban forma de pala, y que este rasgo era compartido con poblaciones del Este de Asia. Esta evidencia y otras observaciones de caracteres fenotípicos como lo son el cabello, la piel y el color de los ojos, respaldaron los argumentos de que los Nativos Americanos tienen muchas semejanzas con poblaciones de Asia. Esto implica que la región más cercana a América de donde podrían haber provenido es el Noreste de Asia y que los Nativos Americanos son semejantes pero pueden observarse subtipos entre ellos (19). Por lo anterior, se propuso un origen único desde Asia pero en diferentes oleadas de migración. Sin embargo, la fecha de estos eventos no podía ser deducida aún con la evidencia que se contaba hasta la fecha. El rango propuesto iba desde los 5,000 hasta los 200,000 años de antigüedad dependiendo de las creencias del investigador(a) (19).

Por ello, una de las preguntas a contestar a partir de entonces fue hace cuánto tiempo había ocurrido el poblamiento de América. En cuanto a este tema existen un número limitado de sitios arqueológicos que proveen evidencia sólida de presencia humana en las Américas hacia finales del Pleistoceno (Pleistoceno: 2.59 millones de años AP – 11,700 años AP).

En 1926 en Nuevo México, EUA, se excavó un sitio arqueológico que resultó ser el más antiguo descubierto hasta ese momento, el sitio Folsom. Allí, el grupo de investigación encontró una punta de lanza de piedra junto con el esqueleto de un bisonte extinto. De acuerdo a los reportes geoarqueológicos esto indicaba una presencia humana antigua a finales del Pleistoceno. Este hallazgo permitió proponer un período aproximado para el poblamiento de América de finales del Pleistoceno, es decir de > 11,700 años AP. Lamentablemente, en este sitio no se encontraron restos humanos que pudieran ser analizados (19).

En los años posteriores, en varios sitios de Norte América, como Nuevo México, Montana, Arizona y Florida, se halló evidencia de presencia humana aún más antigua que la del sitio Folsom, asociada a megafauna como mamuts, tigres diente de sable, lobos y

perezosos. Una de las características principales de estos sitios fueron las puntas de flecha encontradas, las cuales evidentemente eran cuidadosamente fabricadas y de un estilo no antes visto. A la cultura que habitó en esos sitios se le denominó cultura Clovis, debido a su cercanía con el pueblo Clovis en Nuevo México. Se estableció que la cultura Clovis se asentó a finales del Pleistoceno (19) de acuerdo al estrato de las excavaciones donde se encontró evidencia de esta cultura. Posteriormente, gracias a la invención de la técnica de datación por carbono 14 en 1949, fue posible establecer una fecha para las ocupaciones de los sitios Clovis. Éstas se establecieron con un rango de 13,050 a 12,750 años AP (166). A partir de esto se sugirió la hipótesis que los Clovis habían sido los primeros pobladores de las Américas y que estos habían dado origen a la mayoría de pueblos nativos de América. Además se obtuvo una datación estimada para la llegada al continente de hace poco más de 13,000 años. A este modelo se le llamó modelo del poblamiento tardío o teoría Clovis (19).

Los estudios de reconstrucción paleoclimática indican que existió una masa de tierra, mejor conocida como Beringia, que unió a América con Asia durante el Último Máximo Glacial (LGM por sus siglas en inglés) hace aproximadamente 23,000 - 19,000 años. Por lo que se propuso que los humanos ancestros de Clovis empezaron a poblar América a través de este puente una vez que inició el derretimiento de las grandes masas de hielo que existieron en el Norte de América. Esto abrió el camino hacia nuevas tierras hace 14,000- 11,000 años. Después de esto los ancestros de Clovis poblaron todo el continente rápidamente ya que no encontraron otras poblaciones humanas con las cuales hubiera competencia por las tierras y los recursos (167).

El poblamiento de América también ha sido estudiado mediante la lingüística. Con ello se identificaron tres lenguas distintas en el continente: Amerindia, Na-Dene e Inuit-Aleut. Esta evidencia, junto con la teoría del poblamiento tardío y el estudio morfológico dental de individuos de distintas regiones de América, dio lugar a la hipótesis de que el poblamiento de América se llevó a cabo en tres olas de migración. La primera proviene de los Clovis que dieron origen a la mayoría de pueblos nativos de América. La segunda es la de los hablantes de Na-Dene (Norte América) y la tercera es la de los Inuit-Aleut (quienes eran llamados comúnmente como Esquimales, pero este término se encuentra en desuso ya que es considerado un término despectivo para la población de Inuit). Cabe mencionar que en este mismo estudio en donde se propusieron las tres olas de migración se

incorporaron datos genéticos para reforzar esta teoría. Sin embargo, la evidencia genética no apuntaba claramente a una separación de los tres grupos por lo que los autores indicaron que en un futuro, cuando fuera posible, sería deseable analizar más datos de Na-Dene e Inuit-Aleut para poder confirmar su hipótesis (168).

Estas hipótesis del poblamiento tardío y de las tres olas de migración fueron las más aceptadas durante las últimas décadas del siglo XX. Sin embargo, a la par también hubo hallazgos que apuntaban a que el poblamiento de América podría haber iniciado mucho antes de lo establecido. En 1977 se descubrió el sitio arqueológico Monte Verde en Chile. Este sitio presentaba características distintas a aquellas encontradas en los sitios Clovis e indicaba que se trataba de un pueblo de cazadores-recolectores semi sedentarios. Uno de los aspectos más relevantes de Monte Verde es que se estimó una antigüedad mayor que la de los sitios Clovis en Norteamérica, es decir de más de 13,000 años de antigüedad (19). Además, años más tarde, también se encontró registro de ocupación humana temprana en otros sitios en distintas regiones de América como Paisley Caves (~14,000 años AP), Pages Ladson (~14,500 años AP) y Cactus Hill (~16,900 – 15,000 años AP) en EUA, Huaca Prieta (~14,500 – 13,500 años AP) en Perú, Taima Taima (~14,000 – 13,500 años AP) en Venezuela, entre otros. Sin embargo, debido a que hasta ese momento en la arqueología se favorecía la teoría de que la cultura Clovis fue la primera, la autenticidad de estos sitios más antiguos fue puesta en duda y estos no fueron aceptados inmediatamente como sitios pre-Clovis (19). Fue hasta dos décadas más tarde que se pudo verificar la datación del sitio Monte Verde de 14,500 años AP (169). Este hallazgo cambió completamente la visión que se tenía del poblamiento de América ya que implicaba que los primeros pobladores entraron a América antes del derretimiento de las grandes masas de hielo del LGM. Este hecho condujo a proponer una nueva ruta de entrada a las Américas. Se hipotetizó que si la ruta terrestre de entrada a América se había mantenido cerrada durante el poblamiento temprano entonces los humanos solamente podrían haber entrado a las Américas a través de una ruta marítima. Esto les habría permitido también desplazarse más rápido (19, 167).

A principios de la década de los 90's se llevaron a cabo diversas investigaciones para caracterizar el ADNmt de distintas poblaciones mediante PCR y la identificación de polimorfismos con RFLP (Restriction Fragment Length Polymorphisms). Esto se llevó a cabo mientras los sitios pre-Clovis estaban siendo estudiados y se encontraban en espera

de validación. Fue así como en 1993 se encontró que la variación genética del ADNmt de nativos americanos estaba ampliamente distribuida en todo América. Se encontró que el ADNmt poseía sub-variantes entre 'tribus' lo que apuntaba a que tenían un origen común en Asia y la diversificación de estos había iniciado después del poblamiento de América. Los haplotipos encontrados se clasificaron en cuatro haplogrupos mitocondriales fundadores: A, B, C y D, los cuales presentaban un tiempo de coalescencia de hace 16,750- 33,500 años (170). Dicho hallazgo fue muy importante ya que no era compatible con el modelo Clovis. En cambio respaldaba la aceptación de los sitios pre-Clovis y la teoría del poblamiento temprano de América .

Hoy en día, gracias a los estudios de reconstrucción paleoclimática mediante ADN, sabemos que el corredor o pasillo libre de hielo que se formó entre las dos grandes capas de hielo (Laurentino y de la Cordillera) al final del LGM no fue viable para el paso de los humanos sino hasta hace 12,500 años (19, 167). Esto debido a que el corredor se formó hace 13,000 años. Sin embargo, la aparición de vegetación y animales necesarios para la subsistencia humana fue 400 y 500 años más tarde, respectivamente. Estas estimaciones respaldan la hipótesis del poblamiento de América a través de una ruta marítima ya que el poblamiento debió haber iniciado antes, al menos hace 15,000 – 17,000 años de acuerdo al registro arqueológico de los sitios pre-Clovis (19, 167).

Existe evidencia también de sitios arqueológicos putativos aún más antiguos que los sitios pre-Clovis. Estos han dado lugar al modelo del poblamiento muy temprano, el cual sugiere que el humano llegó al continente Americano hace > 35,000 años. Esta hipótesis está basada en investigaciones que proponen la existencia de sitios arqueológicos muy antiguos, mucho antes del final de la LGM. Algunos ejemplos de estos sitios son: Calico (de hace 200,00 a 50,000 años) y Mastodonte Cerutti (de hace 130,000 años), ambos en California, EUA (171).

Estos sitios hasta el momento no son aceptados por la comunidad científica ya que implican que el humano moderno debió entrar al continente Americano antes de la mayor migración de humanos anatómicamente modernos fuera de África (hace ~100,000 años), o que la evidencia encontrada no esté relacionada con humanos modernos, sino con otras especies de Homo, como los neandertales, denisovanos u *Homo erectus*. En todo caso, la evidencia encontrada en ambos sitios podría no estar relacionada con el género Homo, ya

que en ambos casos se encontraron herramientas de roca como hojuelas roca, martillos y yunques las cuales no se encuentran trabajadas y su forma ambigua podría también ser el resultado de la ruptura de rocas debido a algún evento natural (19). Debido a esto, y aunado a que no existe evidencia filogenética que respalde estos hallazgos, estos sitios por el momento no son considerados evidencia de una ocupación humana muy temprana en América. De manera similar, existen sitios arqueológicos en América que apuntan a una ocupación del continente pre-LGM. De acuerdo al tiempo en el que ocurrió el LGM (hace 23,000 - 19,000 años), los humanos modernos podrían haber llegado a América antes de que se formara la gran masa de hielo o una vez que empezó a derretirse, hace 26,000 y 19,000 años, respectivamente (19). Ejemplos de sitios arqueológicos con una datación previa al LGM o durante el LGM son: el sitio del Parque Nacional de White Sands en Nuevo México (172), EUA; en donde se encontraron huellas humanas con una datación de hace 23,000 – 21,000 años y el Cerro del Chiquihuite en Zacatecas, México; en donde se encontraron herramientas de roca con residuos químicos que sugieren una presencia humana hace 26,500–19,000 años (173). De este último sitio, los investigadores realizaron un análisis de ADN de sedimentos de la cueva donde se encontraron los restos. Aunque no encontraron evidencia de ADN humano en el sitio, la evidencia arqueológica no está descartada. Dichos pre-LGM fueron descritos en publicaciones muy recientes (2020-2021), por lo que aún no se ha propuesto un nuevo modelo para el poblamiento de América a partir de estos. Además, el hecho de que estos no cuenten con restos humanos hace imposible por el momento hacer análisis filogenéticos de estos supuestos primeros habitantes. Sin embargo, investigaciones en otras regiones del mundo han demostrado la posibilidad de recuperar ADN de humano a partir de sedimento asociado a un lugar de ocupación (174). Es posible que eventualmente se lleven a cabo investigaciones de este tipo en otros sitios pre-LGM que permitan esclarecer si existe evidencia genética de ocupación humana más antigua. Por el momento, con la evidencia que se cuenta los datos genómicos tanto actuales como antiguos del continente apuntan a que la ocupación se inició a finales del pleistoceno o inicios del holoceno (este punto se describe con mayor detalle más adelante). Y en caso de haber una ocupación más temprana (pre-LGM), estos primeros habitantes no han dejado huella genética detectable hasta el momento en los nativos americanos antiguos y actuales analizados.

A continuación se resume lo que se sabe hasta el momento acerca del poblamiento de América y que ha sido respaldado o refutado mediante evidencia paleogenómica del genoma nuclear.

5.1 Modelo Paleoamericano

El modelo paleoamericano proponía que los primeros pobladores de las Américas fueron los 'paleoamericanos', los cuales más tarde fueron reemplazados por los Nativos Americanos más tardíos y que estos últimos son los que han habitado América a partir de entonces. Esta hipótesis se basó en la morfología craneal de individuos antiguos de América del Pleistoceno y principios del Holoceno que presentaban dolico-hipsicefalia y dientes grandes, similar a la que presentan individuos nativos de Australia y Melanesia. Por otro lado, la mayoría de los Nativos Americanos más tardíos y actuales presentaban una bóveda craneal más ancha y dientes más pequeños (175–177). Por medio del análisis paleogenómico una investigación en 2015 refutó la hipótesis paleoamericana analizando genomas antiguos de América de individuos que presentaban morfología paleoamericana (Pericúes y Tierra de Fuego Patagonia) y de individuos antiguos que no la presentaban (Sierra Tarahumara) (55). Los individuos Pericúes y de Tierra de Fuego Patagonia se consideraban remanentes de los primeros pobladores de América, quienes conservaron el fenotipo craneal paleoamericano al haber permanecido aislados genéticamente de los nativos americanos más tardíos. Los genomas de estas dos poblaciones se compararon con un panel de referencia de individuos del presente de América y de Oceanía. Los resultados revelaron que tanto los individuos Pericúes y de Tierra de Fuego con morfología 'paleoamericana' y los Nativos Americanos sin esta morfología compartían las mismas afinidades genéticas y pertenecían a la misma población, independientemente del fenotipo craneal. Esto rechaza el modelo Paleoamericano. Asimismo, se ha propuesto que la morfología craneal/facial es un rasgo fenotípico con plasticidad que no necesariamente refleja la ancestría de un individuo, y que puede estar relacionado a un proceso relacionado al medio ambiente (178).

5.2 Migración hacia América desde Asia

La historia de los nativos Americanos se remonta hacia finales del Pleistoceno en el noreste del continente asiático donde habitaron sus ancestros. Por lo tanto para entender el poblamiento de América se hace necesario estudiar el registro arqueológico del Noreste de Asia.

En el noreste de Asia la evidencia arqueológica de ocupación por parte de humanos modernos es escasa. El sitio más antiguo reportado es el yacimiento del 'cuerno de rinoceronte' mejor conocido como Yana RHS por sus siglas en inglés. Este sitio presenta una datación de hace 31,600 años y en él se encontraron diversos artefactos de piedra y huesos de animales. Entre ellos se encontró un hueso de rinoceronte con el que se fabricó un mango de lanza que da su nombre al sitio (179, 180). Un dato curioso es que después de la ocupación de Yana RHS en la región no se encuentra registro arqueológico durante un período largo de 15,000 años. Probablemente, esto se debe a las condiciones climáticas del LGM, al escaso registro arqueológico que dejaron los antiguos habitantes de esa zona o a la escasa exploración arqueológica (167). Otros sitios antiguos de esta región son los pertenecientes a la cultura Malt'a con una datación de 22,050 CE y el complejo de sitios arqueológicos de Afontova Gora con una datación de 17,000 años. Se cree que estas poblaciones ocuparon la mayor parte de Siberia y tal vez Beringia y han sido denominados 'Antiguos Siberianos del Norte' (ANS, por sus siglas en inglés), quienes estaban bien adaptados para las condiciones climáticas hostiles del ártico Siberiano (167). Estudios genéticos de individuos de estos sitios revelaron que estos pertenecían a una población con estructura genética, o dicho de otro modo, no pertenecían a una población homogénea, más bien estaban divididas en subpoblaciones (181, 182). Un dato curioso es que los ANS presentaban relación genética con población europea y con nativos americanos. Por otro lado, los ANS no son cercanos a la población actual del Este de Asia. Esto llevó a una estimación de la separación de los ANS de la población euroasiática occidental hace 39,000 años. De acuerdo a esta hipótesis, los ANS desaparecieron pero antes de hacerlo dieron origen a dos linajes: el Paleosiberiano y la rama basal de los nativos americanos. Esto sucedió mediante el flujo genético de ANS con otras poblaciones del Este de Asia en eventos independientes hace aproximadamente 23,000 – 20,000 años. El linaje Paleosiberiano está representado por el individuo Kolyma

del yacimiento Duvanny del Noreste de Siberia, de hace 9,800 años. Se considera ancestral a las poblaciones Chukchi, Koryak e Itelmen, todas hablantes de las lenguas Paleosiberianas (167).

No se sabe con certeza dónde y cuándo surgió la rama basal de los Nativos Americanos. Sin embargo, se sugiere que ésta es el resultado de la mezcla entre población del Este de Asia y ANS (Fig. 18). Dentro de los ANS, los Nativos Americanos presentan una mayor afinidad genética con el individuo Afontova Gora 3, seguido de Malt'a y finalmente con Yana RHS, lo cual sugiere que el flujo genético de ANS con la población de Asia del Este debió suceder después de que divergieron estos tres linajes de ANS: Afontova Gora, Malt'a y Yana RHS (167).

De acuerdo a los datos genéticos tanto de genoma nuclear como mitocondrial tanto de individuos actuales como antiguos de América se ha propuesto que la población ancestral de los Nativos Americanos (ANA) permaneció aislada genéticamente de otras poblaciones por un periodo <8,000 años antes de entrar a América como una sola ola de migración (19, 25, 55). Durante este periodo de aislamiento surgieron los haplogrupos mitocondriales fundadores de las Américas (183) y los distintos linajes poblacionales. Algunos de estos linajes eran desconocidos hasta hace poco, como la población 'UPopA'. La población UPopA es una población 'fantasma' para la cual no se tiene registro arqueológico y no se conoce algún individuo del pasado que sea representante de esta ascendencia. Sin embargo, esta población ha dejado una huella genética en algunas poblaciones de América (184). UPopA se identificó por primera vez en población actual Mixe y en la población antigua de Lagoa Santa, Brasil. Asimismo, un estudio reciente sugiere que otras poblaciones indígenas actuales de México también presentan esta señal: Akimel O'odham, Guarijo, Rarámuri, Cora, Mexicanero, Tepehuanos, Mixe, Totonacas, Nahuas de Puebla, Otomí, Chocholtecos y Mochos (185). Un aporte importante del presente trabajo doctoral fue el hallazgo de la contribución genética de UPopA a la población prehispánica de la Sierra Tarahumara (Anexo I, Fig. 5 del artículo). Asimismo, se encontró que en la población prehispánica del sitio Cañada de la Virgen, en Guanajuato, hay una contribución genética de otra población fantasma que no es UPopA. A esta población la nombramos UPopA2 (Anexo I, Fig. 5 del artículo) y es la primera vez que se reporta. Hasta el momento se desconoce el origen de UPopA y UPopA2 (Fig. 18).

Cabe resaltar que en el presente trabajo no se exploró la contribución de UPopA2 en otras poblaciones actuales de México.

Los linajes que surgieron durante el periodo de aislamiento no sólo fueron ANA y las poblaciones fantasma UPopA sino que también surgió el linaje de los Antiguos Beringianos (AB) de los cuales no se conocen descendientes actuales. De estos linajes existe la evidencia genética por el individuo USR1 de río 'Upward Sun' en Alaska de hace 11,500 años (142) y de otro individuo de hace 9,000 años de la cueva Trail Creek en la Reserva Nacional del puente de Bering en Alaska (184). Estudios de los AB han revelado que esta población también presentaba estructura genética. Esto nos habla de que posiblemente estas poblaciones mantuvieron escaso flujo genético entre ellas y que probablemente permanecieron sin contacto una vez dentro de América.

El lugar preciso donde ocurrió el periodo de aislamiento que dio origen a estos linajes de ANA, AB, UPopA y UPopA2 no se puede delimitar con certeza. Esto pudo ocurrir en Beringia o en el Noreste asiático, ya que después del LGM, Beringia quedó cubierta por el océano Ártico, lo cual impide encontrar evidencia arqueológica que indique presencia humana. Posterior al periodo de aislamiento, es decir a finales del LGM hace 19,000-17,000 años, los ANA empezaron a migrar hacia América, probablemente por la costa, la cual es la ruta más aceptada. Posterior a la entrada a América, ANA divergió en dos linajes: Nativos Americanos del Norte (NNA, por sus siglas en inglés) y Nativos Americanos del Sur (SNA, por sus siglas en inglés). De acuerdo al registro arqueológico, se cree que esta divergencia ocurrió en el Noroeste de América, al sur de las masas de hielo Laurentino y de la cordillera (167). El linaje NNA se dispersó por el Norte de América y se ha mantenido en esa región, mientras que el linaje SNA es el que migró a regiones más sureñas y se dispersó a través de todo el continente llegando hasta Sudamérica (55) (Fig. 18). Uno de los grupos representantes del linaje NNA que se usa comúnmente en genética de poblaciones son los Atabascanos quienes habitan en el norte de Norteamérica, principalmente en Canadá y los Estados Unidos de América (55). Para los Atabascanos se cuenta con datos genéticos públicos tanto antiguos como actuales. En cuanto al linaje SNA, el representante más antiguo del que se haya secuenciado su genoma, es el individuo Anzick-1 de hace 12,000 años perteneciente a la cultura Clovis (186). El estudio de su genoma reveló que éste tenía mayor relación genética con todos los Nativos Americanos (a excepción de los de Norteamérica) que con cualquier otra

población del mundo. Esto indicó que el linaje SNA se dispersó rápidamente poco después de su divergencia de NNA. Un estudio del 2018 encontró evidencia genética de que podrían haber ocurrido al menos dos olas de migración de SNA hacia el sur ya que individuos de diferentes regiones de Brasil, Argentina y Chile presentan distintos grados de afinidad genética con Clovis comparado con el resto de Sudamérica (187).

Asimismo, otro estudio propone que podría haber ocurrido flujo genético en distintas ocasiones entre los linajes NNA y SNA, ya que algunas poblaciones de SNA, como individuos de las Islas del Canal del Norte en California, los Akimel O'odham y Surui son modelados genéticamente mejor cuando se introduce una contribución de NNA (26). Esto aún se encuentra debatido (167). En el presente trabajo de tesis encontramos que las poblaciones prehispánicas de México presentan una composición genética distinta de estos linajes NNA y SNA (Anexo I, Fig. S26 del material suplementario del artículo). A partir de la separación de NNA y SNA no se sabe con certeza cómo se fue poblando cada parte del continente. Se espera que en un futuro estas historias locales sean reveladas con el registro arqueológico y el genómico.

5.2.1 Paleo Inuit e Inuit

De acuerdo a evidencia arqueológica, el poblamiento de la parte ártica del continente Americano se llevó a cabo en un período más tardío, hace ca. 5,500 años. Para aquel entonces había terminado el LGM y la tierra que conectaba a Siberia con América se encontraba cubierta por el mar. Por ello esta ola de migración desde Siberia tuvo que haber llegado por medio de navegación marítima o durante inviernos sobre el mar congelado. Al llegar a América, estos pobladores se dispersaron en la región ártica de América poblando por primera vez parte de Alaska, el norte de Canadá y Groenlandia. A los habitantes de esta región se les ha denominado Paleo-Inuit, quienes estaban adaptados a las condiciones extremas del clima ártico (19). Como dato curioso, el primer genoma humano antiguo secuenciado pertenece a un individuo Paleo-Inuit de la cultura Saqqaq con una antigüedad de 4,000 años (12). El análisis de este genoma reveló que los Paleo-Inuit representaban una migración distinta a la de los ANA y a la de los Inuit actuales. Se desconoce cómo, pero los Paleo-Inuit desaparecieron después de la llegada de los Thule a la región ártica de América (167). Los Thule provenían de una cultura

distinta que migró desde Alaska a Groenlandia hace ca. 800 años y se consideran como los ancestros de los Inuit actuales. El reemplazo poblacional en esta región es confirmado con datos genéticos de ADNmt de individuos antiguos Paleo-Inuit, y antiguos o actuales Thule, Birnirk de Siberia e individuos de Groenlandia (19). El estudio genético de estos individuos demostró que los Paleo-Inuit solamente presentaban el haplogrupo mitocondrial D2a (188), identificado previamente en el individuo Saqqaq, y un tamaño poblacional reducido (189). Por otro lado, los ancestros de los Inuit presentaban una variedad más amplia de haplogrupos mitocondriales: A2a, A2b y D4 (19). De acuerdo a la evidencia genética observada en los antiguos Ekven de Chukotka de ca. 2,000 años y los actuales Chukchi se sabe que en algún punto (2,200 - 500 años) hubo una migración desde América hacia Siberia. En la actualidad sigue habiendo contacto entre estas dos regiones. Esto se observa en la migración invernal actual de personas de las poblaciones Iñupiat de Alaska y Yup'ik de la península de Chukotka, Siberia, se reúnen con sus familiares y amigos viajando a través de esta región con vehículos para la nieve (19).

5.3 Caribe

El poblamiento del Caribe es otro tema de interés que se ha explorado recientemente con estudios paleogenómicos. Cuando llegaron los europeos al continente Americano, el primer contacto que tuvieron con la población nativa fue en el Caribe, en las Bahamas. Los Taínos habitaban gran parte de las islas del Caribe y fueron el primer pueblo en tener contacto con los colonizadores europeos (19). La colonización y las enfermedades durante este primer contacto afectaron de manera grave a la población y ésta pasó por un declive demográfico drástico (118). Debido a ello se habló de una extinción de las poblaciones nativas de esta región (190). Empero, no existió una extinción de los taínos y actualmente existen descendientes de esta población indígena (118). Para conocer los procesos históricos previos a la colonización europea ha sido necesario el estudio de restos antiguos precolombinos. Esto se debe a que los procesos demográficos post-colonización europea involucraron movimientos (muchas veces forzados) de personas de una isla a otra o hacia regiones continentales complicando así el estudio de la historia previa. Sin embargo, llevar a cabo estudios de ADN con muestras del Caribe

ha presentado un verdadero reto para el campo ya que las condiciones cálidas y húmedas tropicales de esta región favorecen la degradación del material genético. Esto hace que el éxito de recuperación de ADN sea mucho menor que en otras regiones del mundo (118).

De acuerdo al registro arqueológico, en el Caribe se han observado dos ocupaciones distintas en cuanto a la cultura. Éstas están representadas en periodos diferentes y que en algunas islas, como Cuba, presentan un solapamiento. El primer periodo de ocupación es el llamado 'pre-Cerámico', para el cual el registro más antiguo se encuentra en la isla de Trinidad con evidencia de ocupación hace 8,000 años. En otras islas, se observa una ocupación a partir de hace 6,000 años y hasta ca. hace 2,500 años. Para este periodo el registro arqueológico es escaso. Se sabe que estos primeros pobladores utilizaban herramientas de piedra y se alimentaban de animales y plantas de la región. Además no practicaban el uso de la cerámica. A partir de hace 2,500 años se observa un cambio evidente en el registro arqueológico. Este cambio es más amplio y está caracterizado por el uso de cerámica parecida a la encontrada en Sudamérica, el inicio de la presencia de objetos sagrados, así como el uso de la agricultura y de plantas provenientes de Sudamérica (19).

Los primeros estudios de ADN, principalmente basados en marcadores uniparentales, observaron que los grupos pre-Cerámicos y Cerámicos eran distintos también genéticamente y provenían de olas independientes de migración hacia el Caribe (19). Debido a los pocos datos generados para el periodo pre-Cerámico, no se sabía con certeza de dónde provenían esos primeros pobladores. Por ello esta cuestión permaneció sin contestarse un par de años. Por otro lado, los datos genómicos del periodo Cerámico mostraron que la ola de poblamiento durante este periodo siguió una ruta de entrada desde Sudamérica partiendo desde la cuenca del río Orinoco en Venezuela hacia el norte, primero a través de las Antillas Menores hacia Puerto Rico y expandiéndose por las islas llegando a las del norte como Cuba (19).

Más adelante, en el año 2020 y 2021, se publicaron dos estudios acerca del poblamiento del Caribe. Uno de estos estudios encontró: 1) que los pueblos pre-Cerámicos y Cerámicos mostraron ascendencias diferentes y poco o nulo flujo genético entre ellos. Por lo que a pesar de que en algunas regiones co-existieron, estas poblaciones de algún modo, tal vez por diferencias culturales, se mantuvieron aisladas una de la otra, 2)

probablemente hubo dos dispersiones tempranas antes del periodo Cerámico, ya que encontraron dos ascendencias diferentes en los individuos pre-Cerámicos en Cuba (191). En el otro estudio encontraron que: 1) los individuos del periodo Cerámico del Caribe podrían modelarse como parientes directos de los Arawaka hablantes del idioma piapoco, los cuales actualmente habitan en Colombia y Venezuela, 2) los habitantes de las Antillas Mayores durante el periodo pre-Cerámico, presentan un componente ancestral en mayor proporción relacionado con población Quechua, junto con una menor proporción de componente genético compartido con poblaciones amazónicas como Surui o Karitiana (192).

Los estudios genéticos tanto de población actual como antigua han demostrado que el legado genético y cultural de los antiguos Tainos aún se encuentra en las poblaciones actuales de estas islas. Esto demuestra la resiliencia de estas poblaciones a pesar de las lamentables prácticas de colonialismo. Observamos prácticas similares en otras regiones de América (118).

5.4 Población Y

Motivados por investigar el modelo paleoamericano, un grupo de investigación en 2014 inició el análisis de diversas poblaciones de Nativos Americanos. Para ello seleccionaron 21 poblaciones actuales de distintas regiones de América, en su mayoría de Sudamérica, y unas pocas de Centro y Norteamérica, con un total de 63 individuos sin ancestría europea ni africana. Para su comparación utilizaron datos genómicos de poblaciones de todo el mundo incluyendo aquellos que presentan morfología similar a la paleoamericana como lo son los "negritos" del sudeste asiático y poblaciones indígenas actuales de Australia y Melanesia (193).

En esa investigación encontraron que las poblaciones de amazonia como los Surui, Karitiana y Xavante presentan una ancestría de 2-85% de una población desconocida que está relacionada con los Onge de Andamán, indígenas de Papúa Nueva Guinea, de Australia y los mamánwa 'negritos' de las Filipinas. Esto es diferente de otras poblaciones,

como las Mesoamericanas, donde no se observó esto. En el estudio se decidió llamar a esta población ancestral 'Población Y' haciendo referencia a la palabra 'Ypykuéra' que en la lengua tupí de los Surui y Karitianos significa 'antepasado'. El modelo propuesto en el estudio sugiere que antes de llegar a la Amazonia y a la meseta central brasileña, la Población Y ya se encontraba mezclada con un linaje relacionado con los Primeros Americanos. Asimismo, indican que la contribución genética de la Población Y debió ocurrir en tiempos antiguos e incluso sugieren que los sitios Pre-Clovis podrían haber sido ocupados por ellos. Sin embargo, eso permanecerá como una hipótesis hasta que sea posible recuperar ADN de dichos sitios (193).

Acerca del modelo paleoamericano, el estudio concluyó que debido a que no contaban con datos genómicos de esqueletos con morfología paleoamericana. Sus resultados no eran suficientes para concluir que la Población Y presentará dicha morfología. Esta investigación se publicó el mismo año que aquella que refutó el modelo paleoamericano utilizando datos genómicos de esqueletos con morfología paleoamericana. Sin embargo, pese a que este no era el objetivo, para su sorpresa en este último, también encontraron una ancestría desconocida no descrita previamente en las Américas, la cual incluye la ancestría compartida entre la población Surui y los Australo-Melanesians y para la cual aún no se tiene una respuesta de cómo y cuando llegó al continente Americano (55).

5.5 Contacto con Polinesios

Los habitantes de la Polinesia, la cual se encuentra en el Pacífico central y meridional, se han caracterizado por sus habilidades marítimas. Gracias a su sistema eficaz y único de navegación lograron colonizar las islas del Triángulo Polinésico, pero no sólo eso, sino que también llegaron a la isla Rapa Nui en el año ca. 1,200 CE (19). Esta isla actualmente pertenece a Chile y se encuentra a 3,700 km del continente Americano. De manera interesante, existen registros arqueológicos de que previo al contacto con europeos, en la isla Rapa Nui se han encontrado restos de camote, el cual es una planta nativa del continente Americano, lo que indica un contacto entre estas dos regiones (194). Además, análisis genéticos de la población actual Rapanui mostraron que su composición genética presenta un bajo porcentaje de ancestría de Nativos Americanos, y se estimó que este

evento de mezcla no pertenece a años recientes, sino que ocurrió entre el año 1280 y 1425 d.C (194). Una de las interrogantes que sigue sin respuesta es si los polinesios no se detuvieron en la isla Rapa Nui y continuaron su exploración marítima hasta América (Fig. 18). De acuerdo a las corrientes oceánicas, esta ruta de navegación presenta condiciones desfavorables por lo que se cree que podrían haber sido los Nativos Americanos quienes navegaron hacia la isla de Rapa Nui (19) (Fig. 18). Sin embargo, hay otras líneas de evidencia que sugieren que los polinesios, a pesar de la existencia de las corrientes oceánicas desfavorables, podrían haber llegado a Sudamérica (19). Esto se sostiene por, un estudio de dos individuos antiguos (1479–1708 CE y 1730–1804 CE) de la población de Botocudos de Brasil, encontró que dichos individuos tenían únicamente ancestría genética polinesia y no nativa americana (195). Este hallazgo respalda la idea de que los polinesios llegaron a Sudamérica. No obstante, un estudio más reciente con un tamaño de muestra mayor de Botocudos antiguos no encontró rastros de ancestría polinesia en ellos, por lo que los resultados de la investigación previa podrían corresponder precisamente a aquellos individuos polinesios que llegaron a Sudamérica pero que no se mezclaron con la población nativa (196). Las autoras de este último estudio plantean la posibilidad de que los dos botocudos que presentaron ancestría polinesia en el primer estudio hayan estado mal etiquetados y por lo tanto se llegó a una mala conclusión (comunicado personal de la primera autora Dra. Diana Ivette Cruz Dávalos y de la autora de correspondencia Dra. C. María Ávila Arcos). Otra alternativa es que los polinesios hayan llegado a América, se hayan mezclado con Nativos Americanos, y posteriormente hayan regresado a Rapa Nui o que hayan viajado con individuos Nativos Americanos en su regreso (19).

Asimismo, otra investigación posterior a la que se llevó a cabo en Rapanui actuales analizó el ADN de tres individuos Rapanui antiguos con una datación de ca. 1,500-1,900 CE (197). Los resultados mostraron que no presentaban ancestría genética de Nativos Americanos y concluyeron que el componente genético americano que presentan los Rapanui actuales debe ser el resultado de una mezcla posterior al contacto de las Américas con los europeos. Sin embargo, las conclusiones de esta investigación no toman en cuenta el limitado tamaño de la muestra. Esto indica que el hecho de que no hayan encontrado ancestría Nativo Americana en los tres individuos no indica que en otros individuos antiguos de Rapa Nui fuera inexistente. Además, es posible que haya existido un contacto entre estas poblaciones sin tener como resultado un flujo genético.

Estudios con un mayor número de muestras antiguas podrían contribuir a resolver esta cuestión.



Figura 18. Mapa de América representando el resumen de los linajes que dieron origen a los Nativos Americanos. La ilustración de la mezcla de los linajes ANS y del Este de Asia está basado en la figura 2 de Willerslev y Meltzer, 2021(167). Con estrella roja se marcan las ubicaciones sólo de algunos individuos antiguos de América estudiados mediante paleogenómica. En círculo verde se indica el sitio Pre-Clovis 'Monte Verde'. En cuadrado morado se indica la ubicación de algunas poblaciones actuales mencionadas en el capítulo.

5.6 Conclusión

Como se ha visto a lo largo del capítulo, la historia del poblamiento de América es más complicada de lo que se creyó en un inicio cuando llegaron los colonizadores europeos y se preguntaron cómo habían llegado los nativos Americanos a estas tierras. Cabe resaltar que el continente Americano presenta considerablemente menos registros arqueológicos a diferencia de los otros continentes donde ocurrió un poblamiento más temprano y por ende tienen un mayor número de registros arqueológicos. Por lo anterior, las investigaciones genéticas tanto con ADN moderno como de ADN antiguo han jugado un papel muy importante en la generación de conocimiento de cómo fue poblado. Durante este capítulo se han resumido las principales teorías acerca del poblamiento de América y cómo éstas han sido respaldadas o refutadas mediante la evidencia arqueológica y genómica. Cierro este capítulo haciendo hincapié en que el campo de la Paleogenómica ha avanzado rápidamente en los últimos años y cada año existe nueva evidencia que pone a prueba o afirma las interpretaciones hechas acerca del poblamiento de América. Debido a ello la información contenida aquí podría desactualizarse rápidamente.

6 Presentación del artículo de investigación “Demographic history and genetic structure in pre-Hispanic Central Mexico”

6.1 Prólogo

Este capítulo incluye un resumen de los principales resultados del artículo de investigación publicado como parte de los requisitos para la obtención de grado. También se describen los resultados alineados con los objetivos que se plantearon para la investigación doctoral.

El artículo científico se encuentra como un anexo (Anexo I) de este capítulo en su formato original en inglés. Ha sido publicado en la revista Science y es la pieza central del trabajo de tesis. Dentro del texto principal del artículo y el material suplementario se exponen las preguntas de investigación que se plantearon a lo largo de este recorrido. De igual manera, se describe el contexto arqueológico de los individuos prehispánicos analizados, se indica cuál protocolo de extracción de ADN se utilizó para cada muestra, cuántas bibliotecas genómicas se generaron, si se realizaron o no capturas de genoma completo o mitocondrial y el resumen de cada corrida de secuenciación. Además, dentro del mismo se incluyen en detalle los métodos utilizados para responder cada pregunta, los resultados obtenidos y la conclusión de estos.

Los resultados obtenidos abarcan distintos enfoques. Estos incluyen un enfoque particular en el que se muestra por primera vez la estructura y diversidad genética de las

poblaciones prehispánicas de México de una manera directa mediante el estudio de ADN. Posteriormente evaluamos una hipótesis previa de reemplazo poblacional en la Frontera Norte de Mesoamérica. También se presenta un enfoque más general en el que hubo una contribución al conocimiento del poblamiento de América mediante la evaluación de ancestría genética “fantasma” en algunas de las poblaciones prehispánicas estudiadas.

6.2 Introducción

El campo de la Paleogenómica humana desde sus inicios ha estado dominado por estudios llevados a cabo en laboratorios del norte global. Este hecho ha tenido un impacto en la sub-representación de las poblaciones de México, Centroamérica y Sudamérica en estos estudios. A la fecha en la que se inició el presente trabajo de tesis, en Febrero del 2017, existían pocos genomas antiguos de América y la mayoría de estos estaban enfocados en responder preguntas acerca del poblamiento de América. Ahora ya se cuenta con un mayor número de investigaciones que han contribuido a este tema y a preguntas demográficas de regiones específicas del continente o de mecanismos de adaptación a diferentes ambientes. En cuanto a la representación de México, hasta el momento han sido publicados dos artículos (en 2015 y 2018) en los que se incluyen muestras prehispánicas del norte de México con genomas a baja cobertura para su estudio en el contexto del poblamiento de América (26, 55).

Por otro lado, los trabajos realizados con ADN en México se han enfocado casi exclusivamente en el estudio de haplogrupos mitocondriales dejando de lado la riqueza de información genética que alberga el genoma nuclear. Por lo que el campo de la Paleogenómica en nuestro país es un área de especial interés y amplio potencial que se manifiesta en nuestra riqueza cultural y lingüística, así como en nuestro extenso registro histórico y arqueológico.

De acuerdo a los registros históricos, la llegada de los españoles a México en el siglo XVI produjo un declive demográfico en la población nativa asociado a enfermedades, masacres, desplazamiento y explotación (198). Posteriormente, el proceso de mestizaje

entre la población nativa, españoles y africanos tuvo un segundo impacto en la población, cambiando su demografía y estructura genética. Esto dio lugar a la población actual de México en la que sólo el 19.4% de la población dice pertenecer a alguna población indígena (199, 200).

En los últimos años, la caracterización genética de las poblaciones indígenas actuales de México ha proporcionado información valiosa acerca de la diversidad y estructura genética de estas poblaciones y ha ayudado a inferir eventos demográficos pasados (185, 201, 202). Sin embargo, una de las hipótesis de la presente tesis es que el declive demográfico causado por la colonización redujo el nivel de diversidad genética de estas poblaciones. Aquí muestro que una reconstrucción precisa de su historia genética y demográfica es posible solamente mediante el estudio del ADN de muestras arqueológicas previas a la colonización.

6.3 Resumen de objetivos y resultados

6.3.1 Obtención de ADN mitocondrial y nuclear a partir de restos arqueológicos prehispánicos y secuenciación mediante NGS

El primer objetivo de la presente tesis estableció la obtención exitosa de ADN mitocondrial y nuclear a partir de restos arqueológicos prehispánicos y la realización de la secuenciación mediante NGS del material genético obtenido. De manera inicial se planteó la obtención de restos arqueológicos tanto del Norte, Centro y Sureste de México. Sin embargo, debido a la disponibilidad y colaboraciones establecidas, se procesaron solamente muestras del centro de México. Durante el trabajo doctoral se realizó el procesamiento de las muestras arqueológicas obtenidas, entre ellas huesos o dientes de 37 individuos de la época prehispánica de siete sitios arqueológicos de México. Tres sitios pertenecen a la Sierra Gorda de Querétaro: Toluquilla, Ranas y una cueva del municipio Cadereyta de Montes; tres más del estado de Michoacán: Zaragoza, Tanhuato y La Mina; y el sitio arqueológico de Cañada de la Virgen en Guanajuato. Desde el primer procesamiento de muestras, de individuos de Toluquilla y Ranas, el proceso de extracción

de ADN, la construcción de bibliotecas genómicas y la secuenciación mediante NGS fue exitoso. Asimismo, esto se observó en el resto de muestras que se procesaron posteriormente. De esta manera cumplí con el primer objetivo de este trabajo. Las primeras extracciones de ADN se realizaron mediante el método de *in silico* (62). Posteriormente todas las extracciones se realizaron con el método de Dabney (63) debido a la practicidad y a una mejora en la recuperación para secuencias cortas. Las bibliotecas genómicas construidas fueron de doble cadena (64) y dependiendo de la complejidad de éstas (70, 93) se realizaron o no capturas de genoma completo (WGC) o de genoma mitocondrial (mtC). Por último, las secuenciaciones de *screening* (secuenciación inicial para evaluación de las bibliotecas genómicas) y algunas de saturación se realizaron en su mayoría con el equipo NextSeq500. Por otro lado se utilizó el equipo NovaSeq para la saturación de las bibliotecas genómicas con mejor contenido endógeno. Los detalles de las muestras procesadas, métodos de procesamiento, cantidad de datos generados, contenido endógeno y clonalidad de cada biblioteca genómica pueden encontrarse como material anexo del artículo de investigación en la Tabla S2.

6.3.2 Identificación de linajes uniparentales

El segundo objetivo de la presente tesis fue la identificación de los linajes uniparentales (ADNmt y chrY) de los individuos prehispánicos. Además de la estimación de contaminación en base al genoma mitocondrial y el análisis de confirmación de los patrones de daño para verificar la autenticidad de las secuencias antiguas. Con los datos genómicos generados se realizó la reconstrucción del ADNmt mediante el programa Schmutzi (120). Esto nos permitió la reconstrucción del genoma mitocondrial de 25 de los 37 individuos analizados. Aunado a esto, en el presente trabajo de tesis se incluyeron datos genómicos de dos individuos prehispánicos de la Sierra Tarahumara, en Chihuahua, los cuales habían sido reportados anteriormente en una investigación del extranjero (55). La Dra. María Ávila Arcos junto con la Dra. Rosa Fregel realizaron capturas de genoma completo de estos individuos y las secuenciaron para la generación de una mayor cantidad de datos llegando a la saturación de las bibliotecas genómicas, en la Universidad de Stanford y la Universidad de Upsala. Los datos de estos individuos fueron incluidos dentro de los análisis del presente proyecto de investigación. Con ello se alcanzó un total de 27 genomas mitocondriales para su inclusión y análisis en el artículo de investigación.

Todos estos ADNmt fueron asignados mediante Haplogrep2 (104) a alguno de los cuatro haplogrupos mitocondriales fundadores de las Américas: A B, C o D. Con base en las proporciones encontradas en cada sitio arqueológico observamos que la estructura genética a nivel mitocondrial se ha mantenido a lo largo de más de 2,300 años tomando en cuenta la muestra prehispánica más antigua y las proporciones actuales para cada haplogrupo observadas en poblaciones modernas y antiguas de México en estudios previos (203).

Para el análisis del cromosoma Y fue necesaria la identificación de los individuos masculinos de acuerdo a un análisis de determinación de sexo molecular. Esto se realizó mediante el método descrito por Skoglund *et al.* 2013 (48) (véase sección 4.10, Determinación del sexo molecular) siendo posible la identificación del sexo molecular de 20 de los 39 individuos (debido a las bajas coberturas del genoma); 13 de estos correspondieron al sexo masculino (XY) y 7 al sexo femenino (XX). Una vez identificado el sexo de los individuos se prosiguió al análisis del cromosoma Y de los individuos del sexo masculino. Debido a la baja cobertura en la mayoría de los individuos, solamente fue posible la asignación del haplogrupo de cromosoma Y para 5 individuos, 2 de la Sierra Tarahumara, 2 de la Sierra Gorda y 1 de Cañada de la Virgen. Como se esperaba, el haplogrupo de cromosoma Y identificado correspondió al haplogrupo Q masculino fundador de las Américas, encontrando dos sub-haplogrupos: Q1a2a1-L54 y Q1a2a1a1a1-M3 sin aparente diferenciación geográfica.

El análisis del ADNmt mediante Schmutzi (120) también permitió la estimación de la contaminación de las muestras antiguas, siendo para la mayoría < 3 %. Sólo en el caso de una muestra de Michoacán (E2_Mich_b) la estimación fue del 6 % por lo que se utilizó la herramienta pmdtools (125) para remover las posibles secuencias contaminantes. Posterior a esta remoción la estimación de la contaminación fue de 2 %. De esta manera cumplimos con los estándares del campo de la Paleogenómica en los cuales se utilizan datos con contaminación < 5 %.

Asimismo, como medida de autenticidad de las secuencias antiguas, se realizó el análisis de la presencia de los patrones de daño característicos del ADN antiguo mediante el programa mapDamage (15). En todos los casos se demostró que las secuencias obtenidas

presentaban los daños característicos de las moléculas de ADN, verificando así su autenticidad (Anexo I, Figuras S1- S4).

6.3.3 Análisis de las afinidades genéticas en el contexto de la diversidad actual de las poblaciones indígenas de México

El tercer objetivo fue analizar las afinidades y estructura genética de las poblaciones prehispánicas en el contexto de las poblaciones indígenas actuales de México. Para este objetivo se utilizaron análisis de agrupamiento no supervisados como el PCA (127) y ADMIXTURE (128), además de dos variantes del PCA: FA (132) y mdPCA (<https://github.com/AI-sandbox/mdPCA>) (sección 4.15.1.1 - Análisis de componentes principales (PCA)). Como panel de referencia se utilizaron los datos de genotipado de diversas poblaciones indígenas de México (NMDP) (201). Asimismo se añadió la información genética generada por otros proyectos: HGDP (36), SGDP (34) y (55) en los cuales se secuenciaron genomas completos de individuos indígenas de México. De estos genomas completos se extrajo la información genética específicamente de las SNVs del panel NMDP para formar un panel de referencia más completo y analizarlos de manera conjunta; en total se analizaron 561,327 SNVs.

En los tres análisis de componentes principales se observó que las poblaciones prehispánicas mantenían afinidades genéticas con poblaciones actuales de regiones geográficas cercanas a su lugar de origen. Los individuos de la Sierra Gorda, Cañada de la Virgen y Michoacán presentaron afinidades genéticas con poblaciones del centro de México. Un individuo de la Sierra Tarahumara (F9_ST_a) mostró una clara afinidad hacia poblaciones actuales del Norte de México mientras que el otro (MOM6_ST_a) se mostró un poco más cercano a poblaciones actuales del Centro-oeste de México (Purépechas y Nahuas de Jalisco). En los análisis de FA y mdPCA se observó aún más clara la afinidad del individuo F9_ST_a con las poblaciones del norte de México (Anexo I, Figuras 3, S9 y S10). Para conocer la composición genética de los individuos prehispánicos se llevó a cabo el análisis de ADMIXTURE; se eligió una $k=6$ para detectar los componentes ancestrales previamente observados en otras publicaciones de poblaciones indígenas actuales. Con este análisis se observaron algunos componentes ancestrales representativos de cada región. Por ejemplo, para las poblaciones del norte se observaron 3 componentes ancestrales, uno con mayor representación en población Konkaak, otro en

población Akimel O’odham y Rarámuri y otro en población Wixárika. Para el sureste se observaron dos, uno en Tojolabales y otro en Lacandones. Mientras que para el sur, un componente con mayor representación en Triquis. En cambio, las poblaciones del centro mostraron una composición formada por los 6 componentes. Esto mismo se observó para los individuos de la época prehispánica, a excepción del individuo MOM6_ST_a que, a pesar de tener una proporción importante de un componente norteño, también muestra componentes de otras regiones similar a los del centro de México (Anexo I, Figura 3).

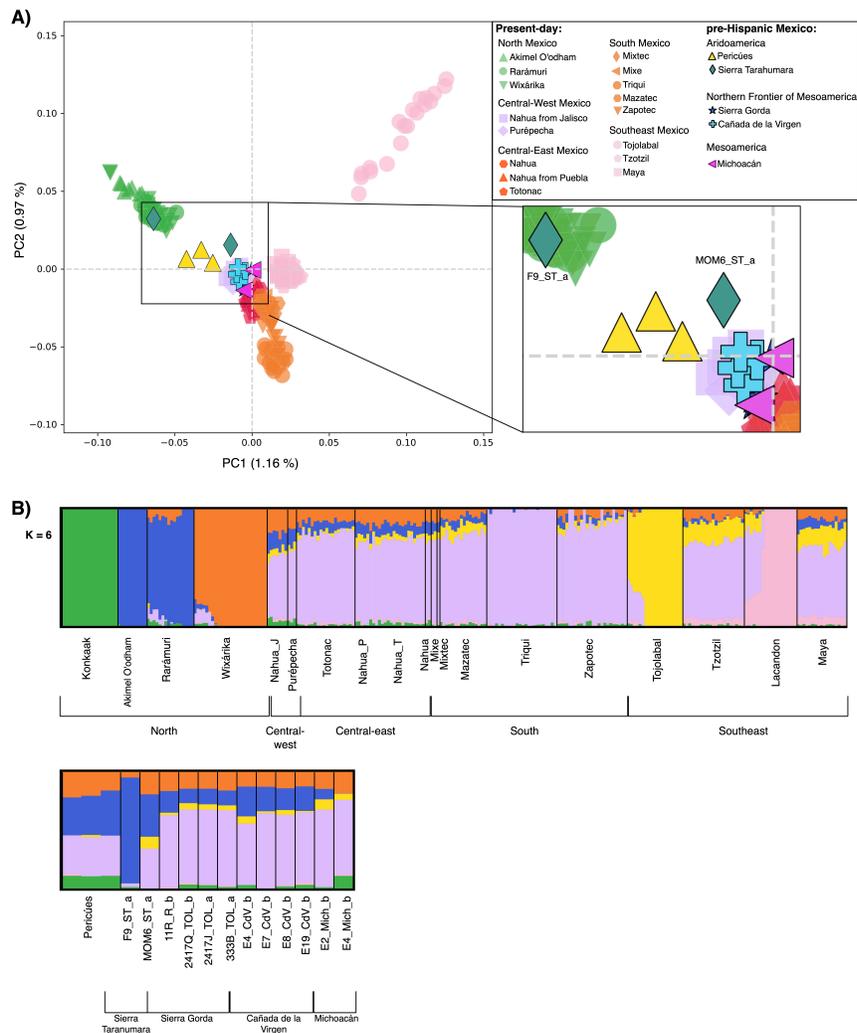


Figura 3 del Anexo I. Se muestra la estructura genética de las poblaciones indígenas de México y de los individuos prehispánicos. A) mdPCA las poblaciones indígenas de México y de los individuos prehispánicos. B) Análisis de ADMIXTURE con k=6 de las poblaciones indígenas de México y de los individuos prehispánicos (Villa-Islas et al., 2023).

Con el fin de identificar las poblaciones actuales con las que los individuos prehispánicos tienen mayor deriva genética compartida, se utilizó el estadístico *outgroup-f3* y el estadístico *D* (139). Utilizando el mismo panel de referencia, se observó que el individuo prehispánico F9_ST_a tiene una mayor relación genética con la población actual Rarámuri que con el resto de las poblaciones. Para el resto de los individuos prehispánicos no se observó una afinidad clara con ninguna población actual, lo cual tiene sentido si tomamos en cuenta que las poblaciones del centro han tenido más contacto y flujo genético tanto entre ellas como con poblaciones distantes de acuerdo al registro arqueológico (204).

Los resultados de estos distintos análisis, PCA, ADMIXTURE, *outgroup-f3* y estadístico *D*, demuestran que en el genoma nuclear también se ha conservado la estructura genética prehispánica hasta la fecha, siendo éste el primer reporte de cómo era la estructura genética en la época prehispánica a nivel de genoma nuclear (Anexo I, Figuras 3, S9-S18).

Posteriormente, la cercanía genética entre el individuo F9_ST_a y los Rarámuri actuales nos llevó a cuestionarnos si podríamos probar la existencia de una continuidad poblacional entre estos dos. Para lo cual se llevó a cabo el análisis de continuidad poblacional descrito por (158) (sección 4.18 - Continuidad poblacional), utilizando las frecuencias alélicas de Rarámuris del NMDP y el conteo de alelos de F9_ST_a para las SNVs presentes en los Rarámuri. Los resultados indicaron que no existe una continuidad poblacional entre estos dos. Sin embargo, esto podría deberse al rigor del método, ya que el método reporta que no hay continuidad genética si existe un poco de flujo genético con otras poblaciones.

6.3.4 Estimación de la diversidad genética en las poblaciones prehispánicas

El siguiente objetivo fue comparar la diversidad genética antigua con la actual para observar posibles cambios debido a la colonización europea de México. Esto se analizó mediante la estimación de diversidad genética con CND (154) y hapROH (151) (sección 4.16 – Diversidad genética). Las poblaciones actuales incluidas en estos análisis fueron aquellas de los proyectos HGDP y SGDP ya que para éstas se contaba con información de genoma completo. Asimismo, se integró la información genética de individuos prehispánicos Pericúes (55) y de Patagonia (205) como referencia de poblaciones aisladas, las cuales se esperaba que tuvieran baja diversidad. Para llevar a cabo el análisis de CND, los individuos de la época prehispánica de un mismo sitio debían cumplir con el requisito de solapamiento de al menos 10,000 SNVs entre ellos, por lo que este análisis no fue posible para los individuos de Michoacán por el bajo solapamiento entre ellos. Para el análisis de hapROH los genomas tanto antiguos como actuales se interseccionan con un panel de 1240K SNVs y deben contar con un mínimo de 400,000 SNVs de intersección. Por ello, el análisis de hapROH de los individuos prehispánicos se realizó solamente en 6 de ellos, uno de la Sierra Tarahumara, tres de la Sierra Gorda, uno de Cañada de la Virgen y un individuo Pericú.

Los resultados de diversidad en base a CND mostraron que las poblaciones aisladas como los Pericúes, los individuos de Patagonia y la población actual Akimel O' odham obtuvieron los valores más bajos de CND. Mientras que el resto de las poblaciones prehispánicas mostraron valores similares al resto de las poblaciones indígenas actuales. En cuanto al análisis de hapROH para identificar las regiones del genoma en ROH, se observó de la misma forma que el único Pericú analizado (el único que tuvo el solapamiento de SNVs mínimo), los individuos de Patagonia y la población actual Akimel O' odham mostraron mayor cantidad de segmentos en ROH, así como una mayor longitud de estos segmentos. Esto muestra que dichos individuos provienen de poblaciones pequeñas donde la diversidad genética tiende a reducirse. En cuanto a los individuos prehispánicos de la Sierra Tarahumara y el centro de México, se observó que presentaban una cantidad mucho menor de segmentos en ROH, así como una menor longitud de estos segmentos. El resto de poblaciones actuales del centro y sureste de México presentaron en general

menor cantidad de segmentos en ROH y menor longitud de segmentos en ROH. Sin embargo, también presentan en ocasiones individuos con mayor cantidad de ROH. En ambos análisis no se logró una comparación adecuada entre individuos o poblaciones pre y post contacto con europeos ya que las poblaciones del presente analizadas no son de la misma localidad o vecinas al sitio que habitaron los individuos de la época prehispánica (Anexo I, Figuras S19 y S20).

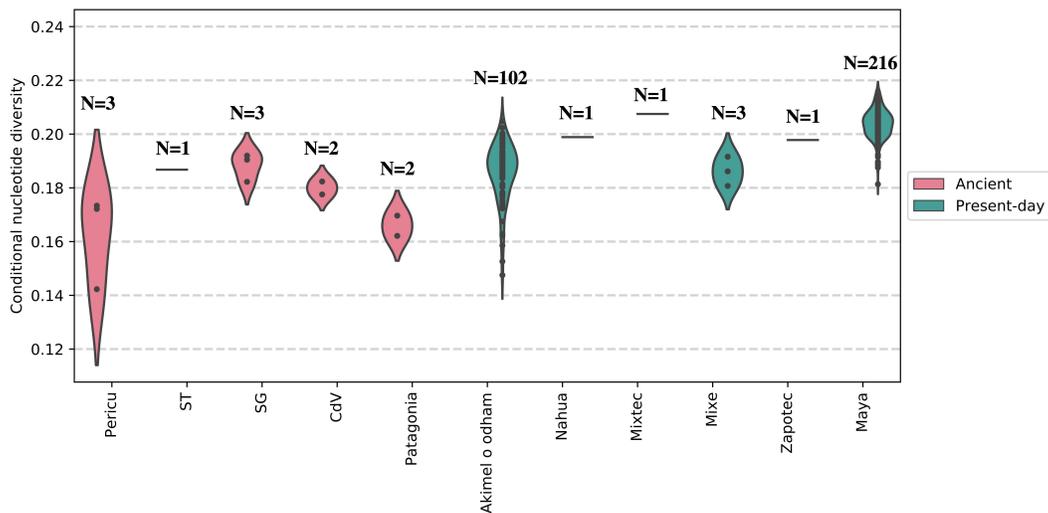


Figura S19 del Anexo I. Diversidad genética condicional (Villa-Islas et al., 2023).

6.3.5 Análisis de continuidad genética poblacional en la Sierra Gorda de Querétaro (320 aC – 1521 dC)

Para finalizar, junto con la Dra. Elizabeth Mejía Pérez Campos y el Mtro. Alberto Juan Herrera Muñoz, ambos arqueólogos del INAH, Querétaro, se planteó el objetivo de analizar un posible reemplazo poblacional ocurrido entre los años 900-1,300 CE en las zonas arqueológicas de la Sierra Gorda, Querétaro.

Las zonas arqueológicas de Querétaro estudiadas: Toluquilla, Ranas y una cueva en Cadereyta de Montes (en adelante llamadas Sierra Gorda por practicidad) estuvieron

ubicadas en la frontera entre dos regiones bioculturales en la época prehispánica. Estas zonas arqueológicas estudiadas pertenecieron a lo que se conoce como Mesoamérica. Dicha zona se caracterizó por ser habitada por poblaciones sedentarias con la agricultura como principal medio de subsistencia, creencias cosmológicas compartidas y estructura política similar entre ellas (206). La ubicación de estas zonas arqueológicas de la Sierra Gorda fue en la frontera norte de Mesoamérica colindando con Aridoamérica. La hipótesis del reemplazo poblacional supone que estos sitios fueron abandonados tras un período largo de sequías ocurrido entre 900-1300 CE, siendo 1,000-1,100 la época más seca y que posteriormente fueron ocupados por grupos de cazadores recolectores provenientes de Aridoamérica (207).

Durante el trabajo de tesis se analizaron un total de 13 individuos de la Sierra Gorda de Querétaro de diferentes dataciones, de antes y después del supuesto reemplazo poblacional para poner a prueba esta hipótesis utilizando los datos genómicos.

El primer análisis se enfocó en observar si existía una continuidad genética a nivel de ADNmt. Se logró reconstruir el ADNmt de todos los individuos de la Sierra Gorda con una cobertura de 28-222x, con lo cual fue posible asignar un haplogrupo mitocondrial para cada individuo analizado. De manera interesante, se observó que el sub-haplogrupo A2d estuvo presente en 6 de los 13 individuos, desde el más antiguo con una datación de 320 BCE hasta uno de los más recientes, del año 1,160 CE. Además, en el análisis de redes de haplotipo (ANEXO I, Figura 2) se observó que 5 de estos individuos formaron un mismo clado excluyendo solamente aquel que presentaba una mayor cantidad de sitios faltantes (Ns). Este resultado fue el primero que sugirió una continuidad genética a nivel de ADNmt. En cuanto a los otros haplogrupos identificados, se asignaron también los haplogrupos B2l (n=3) y D1m (n=1) en el periodo pre-sequía y el haplogrupo B2c (n=3, dos de ellos emparentados de 1er grado) en el post-sequía (Tabla 5). Sin embargo, debido a que los cambios en el ADNmt dentro de una población son más proclives a deriva génica o pueden ser afectados por las costumbres sociales y culturales, los resultados de continuidad poblacional en base al ADNmt no son concluyentes y deben ser soportados con otras líneas de evidencia a nivel del genoma nuclear.

Tabla 5. Datación e información del ADNmt de los individuos prehispánicos de la Sierra Gorda de Querétaro.

Individuo	Sitio arqueológico	Datación (CE)	Sub-haplogru po mitocondrial	Calidad de la asignación del haplogru po mitocondrial (%)	Cobertura del ADNmt (x)
P_CCM_b	Cadereyta de Montes	320 ± 40 BCE	A2d	78.11%	46.1488
2417Q_TOL_b	Toluquilla	671-867	B2l	73.34%	222.469
333O_TOL_b	Toluquilla	700 ± 50	B2l	65.94%	66.8072
11R_R_b	Ranas	769 ± 40	D1m	84.00%	28.2146
6428A_TOL_b	Toluquilla	784 ± 50	B2l	69.98%	104.88
7A_R_b	Ranas	ND, Pre-sequía	A2d	82.93%	38.1544
37AI_R_b	Ranas	ND, Pre-sequía	A2d	88.06%	63.6186
333Q_TOL_b	Toluquilla	520 ± 40	A2d	84.96%	32.4223
2417C_TOL_a	Toluquilla	1,160 ± 60	A2d	82.90%	44.2487
2417J_TOL_a	Toluquilla	1,161 ± 60	A2d	90.53%	216.664
333A_TOL_a	Toluquilla	1,351 ± 50	B2c	61.53%	7.37009
333B_TOL_a	Toluquilla	1,351 ± 50	B2c	78.54%	529.856
333C_TOL_a	Toluquilla	1,351 ± 50	B2c	76.36%	148.775

ND: No determinado

Posteriormente, se llevaron a cabo los análisis del genoma nuclear, parte de los cuales se mencionaron anteriormente, en los análisis de PCA (127), FA (132) y mdPCA (<https://github.com/AI-sandbox/mdPCA>) (Anexo I, Figuras 3, S9 y S10) en los que los individuos de la Sierra Gorda se agrupaban entre ellos independientemente de su datación, mostrando que compartían las mismas afinidades genéticas con las poblaciones

actuales del centro, contrario a lo que se esperaría en un caso de reemplazo poblacional. En el caso de reemplazo poblacional se esperaría observar una mayor afinidad genética de los individuos post-sequía con poblaciones prehispánicas o actuales del norte de México.

En el análisis de ADMIXTURE también se observó que todos los individuos de la Sierra Gorda presentaron una composición genética similar a la observada en población actual del centro de México (Anexo I, Figura 3).

En cuanto a los análisis de outgroup-f3 (139) no se observó que los individuos de la Sierra Gorda presentaran una mayor afinidad genética con una población indígena en específico. Posiblemente debido a que las poblaciones del centro de México se han caracterizado por estar en constante flujo genético entre ellas y en ocasiones con poblaciones distantes como se ha observado en estudios previos (201), por lo que su composición genética es el resultado de estos múltiples eventos de mezcla que no permiten ver un patrón claro (Anexo I, Figura S13). Asimismo, se llevó a cabo un análisis de outgroup-f3 incluyendo solamente a los individuos de la época prehispánica. Este análisis reveló que independientemente del período (pre-sequía, post-sequía) los individuos de la Sierra Gorda compartían una mayor deriva génica entre ellos que con cualquier otro individuo de la época prehispánica. Esto no se cumplió para el individuo 11R_R_b de Ranas, el cual por tener un número limitado de SNVs presentó afinidades genéticas similares con distintos individuos y barras de error muy amplias (Anexo I, Figura 4).

Para evaluar estadísticamente la hipótesis de reemplazo poblacional en la Sierra Gorda se llevó a cabo un análisis de D-statistics (139) en el que se evaluó si los individuos de la Sierra Gorda podrían ser modelados dentro de un mismo clado a exclusión de otras poblaciones prehispánicas y actuales de México. El análisis de D-statistics se realizó de la forma de $D(\text{Sierra_Gorda_pre-sequía}, \text{Sierra_Gorda_post-sequía}; \text{Pop3}, \text{YRI})$, donde los individuos de la Sierra Gorda de ambos períodos son modelados dentro de un mismo clado, a exclusión de 'Pop3' que puede ser cualquier población prehispánica o actual de México y 'YRI', que corresponde a Yoruba de Nigeria, África, como grupo externo en el árbol modelado. En esta topología de árbol una desviación significativa ($|Z| > 3$) del valor $D=0$ indicaría que los individuos pre-sequía son más similares a 'Pop3' ($D>0$) o que los individuos post-sequía son más similares a 'Pop3' ($D<0$). En ninguna de las comparaciones realizadas se observó una desviación significativa de $D=0$, indicando que

los individuos de Toluquilla forman un mismo clado y que son más similares entre ellos que con cualquier otra población prehispánica o actual (Anexo I, Figura 4).

Posteriormente, para evaluar la hipótesis de continuidad poblacional se realizó un modelado demográfico mediante el método de ajuste de grafo de mezcla (qpGraph) (139). En el modelo se incluyó el individuo de Toluquilla 2417Q_TOL_b del período pre-sequía en un clado dando origen a los individuos 2417J_TOL_a y 333B_TOL_a del período post-sequía. Además se incluyó el individuo F9_ST_a en un clado distinto representando el componente genético de Aridoamerica. En el caso del modelado demográfico con qpGraph, un valor ($|Z| < 3$) indica la aceptación del modelo. De esta manera, de acuerdo a los datos genéticos, el modelo propuesto para la continuidad poblacional fue aceptado, indicando así una continuidad genética entre la población pre-sequía y post-sequía del sitio Toluquilla en la Sierra Gorda (Anexo I, Figura 4).

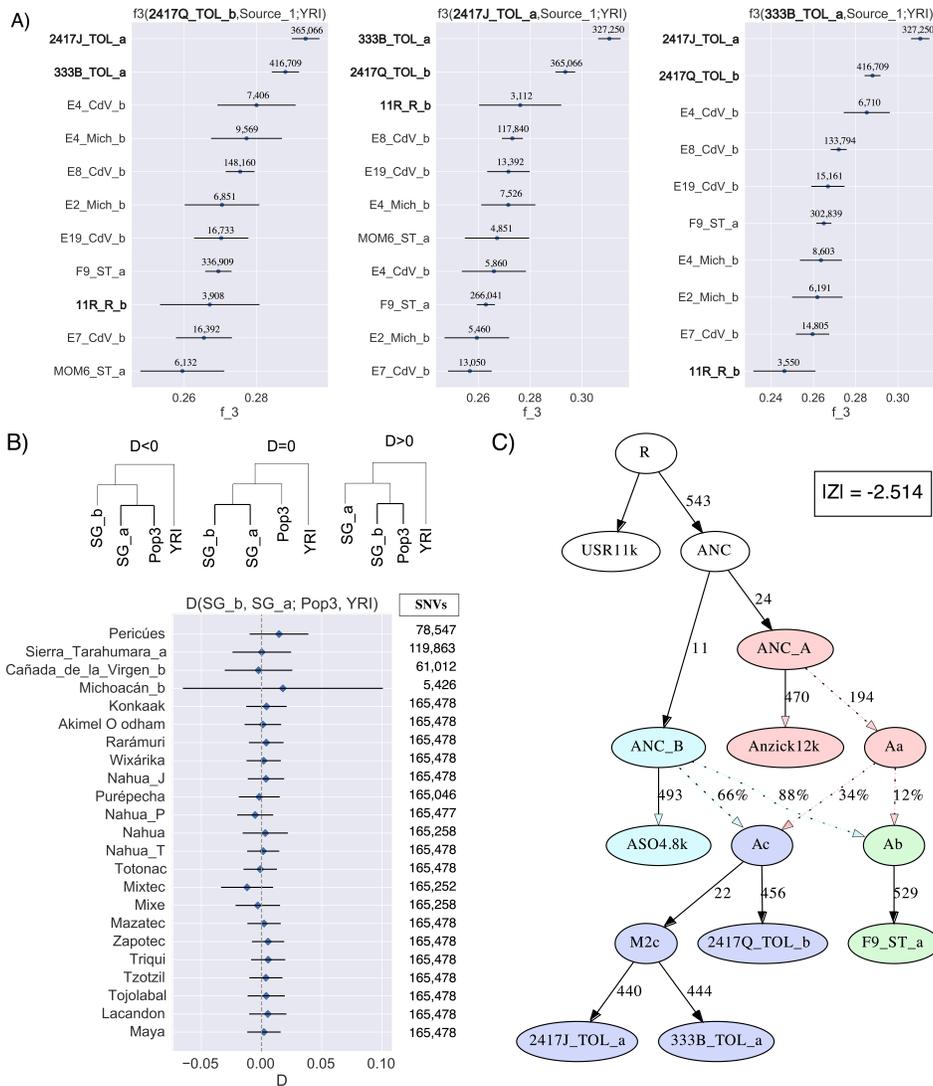


Figura 4 del Anexo I. Continuidad poblacional en la Sierra Gorda de Querétaro, analizada mediante outgroup-f3, estadístico D y qpGraph (Villa-Islas et al., 2023).

El hecho de que exista una continuidad poblacional en esta región de la Sierra Gorda, podría deberse a que, a diferencia de otras poblaciones, la población de Toluquilla y Ranas además de dedicarse a la agricultura, otra de sus fuentes de subsistencia fue la minería del cinabrio, un compuesto de mercurio altamente apreciado en la época prehispánica (208–210). Lo que nos hace suponer que esto pudo permitirles continuar con el intercambio de bienes y alimentos con otras poblaciones. Asimismo, actualmente las zonas arqueológicas estudiadas de la Sierra Gorda se encuentran dentro de un bosque mixto y de coníferas por donde pasan los vientos húmedos del Golfo de México. Esto nos da indicios de que esta zona podría no haber sido tan afectada por las sequías como otras

regiones de la Frontera Norte de Mesoamérica. Como por ejemplo, la zona arqueológica de Cañada de la Virgen, donde actualmente se presenta un clima semiárido y donde se sabe que el sitio fue abandonado entre los años 1,000-1,100 CE (211).

En conclusión, los análisis realizados indican una continuidad genética en la población analizada de la Sierra Gorda y abren el camino para el estudio de otros sitios arqueológicos de la Frontera Norte de Mesoamérica donde sí podría haber ocurrido un reemplazo poblacional por parte de grupos cazadores-recolectores provenientes de Aridoamérica.

6.3.6 Contribución genética fantasma en población prehispánica de México

Otro aporte del artículo de investigación fue la detección de dos distintas contribuciones genéticas 'fantasmas' en la población prehispánica de México. A finales de 2018 una investigación detectó que la población Mixe, de México y de Lagoa Santa, de Brasil presentaban una contribución genética desconocida a la cual se denominó UPopA (véase, sección 5.2). A pesar de que este no era uno de los objetivos principales del estudio, decidimos explorar si alguna de las poblaciones prehispánicas estudiadas de: Sierra Gorda, Cañada de la Virgen, Michoacán o de la Sierra Tarahumara presentaban esta contribución fantasma. Esto se realizó mediante el análisis de ajuste de grafo de la mezcla (qpGraph) tomando como modelo base aquel utilizado en la publicación de 2018. De manera interesante, en las poblaciones de la Sierra Gorda y Michoacán no detectamos una contribución genética fantasma, pero en Cañada de la Virgen y la Sierra Tarahumara sí. Explorando más a fondo con diferentes modelos, asesorados por Víctor Moreno Mayar (primer autor del artículo donde se detectó UPopA), encontramos que las contribuciones genéticas pertenecían a dos poblaciones fantasma distintas. La población prehispánica de la Sierra Tarahumara representada por el individuo F9_ST_a presentó una contribución de UPopA, al igual que los Mixes. Mientras que en Cañada de la Virgen detectamos una contribución no reportada anteriormente a la que denominamos UPopA2 y que no ha sido explorada en otras poblaciones actuales (Anexo I, Figura 5).

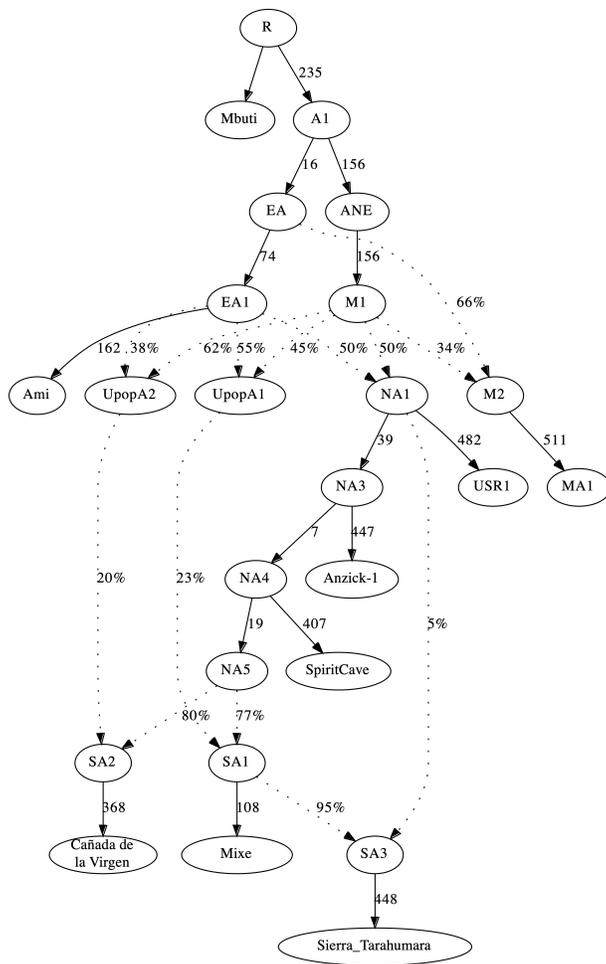


Figura 5 del Anexo I. Contribución de ancestría genética fantasma en los individuos prehispánicos de México (Villa-Islas et al., 2023).

6.4 Conclusiones

Como conclusiones generales de la investigación, tenemos que:

1. La distribución de los linajes maternos se ha conservado desde la época prehispánica hasta la fecha.
2. De manera similar, observamos que la estructura genética a nivel del genoma nuclear se ha conservado a través del tiempo, hasta el presente.
3. Observamos continuidad genética en la Sierra Gorda de Querétaro desde la época pre-sequía y post-sequía, que se sustenta a través de la presencia del linaje materno A2d y nivel del genoma nuclear.
4. Encontramos contribución genética de dos poblaciones fantasma en los individuos prehispánicos de México de la Sierra Tarahumara y Cañada de la Virgen, UPopA y UPopA2, respectivamente. Siendo UPopA la población fantasma reportada previamente en Mixe. Mientras que UPopA2 corresponde a una nueva contribución genética no reportada previamente.

7 Implicaciones éticas en la Paleogenómica

Los avances en el campo de la paleogenómica han traído muchos beneficios para el conocimiento de la especie humana, de su evolución, procesos adaptativos y de su historia demográfica tanto local como global. A pesar de los beneficios de la investigación paleogenómica, ésta en ocasiones ha caído en malas prácticas científicas que se reproducen globalmente. Las cuales incluyen la escasa colaboración entre grupos de investigación del norte global con laboratorios locales en los sitios de procedencia de las muestras bajo estudio. Además de la ausencia de: 1) permisos para el procesamiento destructivo de las muestras y el análisis de ADN, 2) vinculación con poblaciones de pueblos originarios o descendientes, y 3) ética en la comunicación de los resultados de las investigaciones.

7.1 Dinámicas de colaboración

Es importante hacer énfasis en que la investigación paleogenómica se trata de un trabajo multidisciplinario. Lamentablemente, las colaboraciones entre grupos de investigación se han dado históricamente de forma asimétrica remarcando las diferencias entre los países desarrollados del norte global y aquellos en desarrollo del sur global. Dichas diferencias han continuado con una práctica neocolonialista, por lo que es importante que existan buenas prácticas de investigación no sólo a nivel científico sino también a nivel ético y legal (56, 212). Como se mencionó en la introducción de esta tesis, la investigación paleogenómica con muestras de México en su mayoría se realiza en el extranjero. Esto no solamente aplica para el caso de México, sino también en general para los países del sur global. Además, en ocasiones, las investigaciones que reportan el análisis genómico de muestras de países del sur global no contienen coautores del país de origen de las muestras, o en otras ocasiones sí existen coautores de dicha región pero únicamente como proveedores de muestras (56, 212).

Actualmente, no existen normas globales ni una estructura jurídica y ética que se puedan aplicar en general a la investigación Paleogenómica ya que depende mucho de los países involucrados en este tipo de proyectos. Sin embargo, se han discutido varias recomendaciones para promover la investigación ética en paleogenómica (56, 213, 214).

Para resolver las asimetrías en las colaboraciones es importante primero reconocer que existe este problema y saber cómo afecta a los grupos locales. El hecho de que existan asimetrías perpetúa el subdesarrollo científico de los países con menos recursos, a la vez que pierden parte de su patrimonio nacional y el estudio e interpretación de su historia queda en manos de grupos de investigación externos que sólo ofrecen coautoría o ningún beneficio para el desarrollo local del campo (56, 212). Cuando, por el contrario, estas colaboraciones deberían de resultar en la formación de recursos humanos a nivel local para que los grupos de investigación puedan realizar este tipo de investigación de forma soberana. Asimismo, se debe establecer una retroalimentación entre los diferentes grupos de investigación que enriquezca la interpretación de los hallazgos. En este punto el grupo de investigación local podría tener un mejor conocimiento de su historia, de cómo pueden interpretarse los hallazgos en base al registro arqueológico y la historia de su país o región considerando a las poblaciones descendientes. Además, el grupo de investigación local conoce más acerca de la situación sociopolítica, las normas y las legislaciones de su país en torno a la investigación (56). Por lo anterior, es importante que las colaboraciones se den de forma equitativa y respetuosa entre las partes involucradas para así evitar que la colaboración se limite a una transacción de muestras a cambio de autoría en publicaciones científicas (212).

Cabe mencionar que las malas prácticas no sólo son promovidas por los laboratorios extranjeros. También se da el caso de grupos de investigación locales que incurren en estas prácticas solicitando muestras de colecciones ajenas, de las cuales no tienen el permiso de la autoridad competente ni el conocimiento del contexto arqueológico, ya que no formaron parte del proyecto en el que se recuperaron las muestras del sitio arqueológico. Y a pesar de esto, solicitan las muestras para proporcionarlas a otros grupos (generalmente extranjeros), fungiendo así como intermediarios y teniendo como resultado el beneficio de la coautoría (212).

Otro aspecto en el que se evidencia la desigualdad del norte global comparado con el sur global en cuanto a la investigación paleogenómica es la diferencia en los presupuestos asignados a este tipo de investigación y los apoyos que reciben los grupos de investigación. En los países del sur global es más común que se atiendan otras necesidades del país más urgentes y que el presupuesto asignado a la ciencia básica y a la formación de recursos humanos científicos sea menor. En cambio, en los países del norte global se apoya más a este tipo de iniciativas. Por ello cuentan con mayores recursos financieros para llevar a cabo la investigación. Cabe destacar que hacer investigación de ADN es costoso debido a varios aspectos: 1) Infraestructura especial para trabajar con restos antiguos. Como se mencionó en la introducción, el laboratorio ADN debe encontrarse aislado de aquellos laboratorios en los que se trabaje con ADN moderno y PCR; 2) Uso de equipo de protección personal desechable. 3) Adquisición constante de insumos, equipo y reactivos de laboratorio costosos; y 4) Secuenciación NGS. En varios países del sur global se debe pagar el precio de importación del equipo y los insumos, resultando así en costos más elevados de procesamiento por muestra, y todo esto con un presupuesto menor. Si esto se compara con los laboratorios del norte global, en los que se habla incluso de procesamiento de muestras a escala industrial, la desigualdad se hace aún más evidente (56).

En consecuencia, tanto grupos de investigación del norte global como el sur global deben establecer prácticas de colaboración e intercambio de conocimiento horizontal con atención a los diferenciales de poder y dejando de lado las prácticas neocolonialistas.

Existen varias medidas para dar solución a las colaboraciones asimétricas (56), por ejemplo:

-Capacitación de recursos humanos para el grupo local mediante cursos en los que se instruya acerca de las técnicas de laboratorio para ADN y de análisis de datos genómicos, ya sea en la sede del grupo extranjero o en la sede del grupo local.

-Al diseñar el proyecto de investigación, asignar un presupuesto que considere dentro de éste el apoyo para los gastos que implica el traslado del personal que se va a capacitar. Asimismo, se puede considerar también que si el grupo de investigación local ya cuenta con instalaciones para el trabajo con ADN, podría establecerse dentro de este

presupuesto la adquisición de equipo faltante o insumos para llevar a cabo la investigación en dicho laboratorio.

-Discusión de los resultados en los que participen ambos grupos de investigación de tal manera que los resultados sean interpretados adecuadamente desde ambas perspectivas. También dicha discusión podría incluir el enfoque de preguntas de interés local o el planteamiento de preguntas para futuras investigaciones.

En este sentido, desde el año 2011 existe el consorcio de Prácticas de verano en Genómica para personas indígenas (Summer Internship for Indigenous peoples in Genomics), 'SING', por sus siglas en inglés (<https://singconsortium.org>). Con el cual no solo se promueven mejores colaboraciones entre grupos de investigación sino que se promueve también la participación y capacitación de las comunidades indígenas. El SING tiene como objetivo abrir un espacio para jóvenes indígenas en el que exista la discusión de temas de soberanía genómica así como la capacitación en el área de genómica, y de reconocimiento y defensa de sus derechos. Éste surgió de la necesidad de tener una mayor representación de la comunidad indígena en la ciencia. Esto debido al hecho de que existen jóvenes indígenas científicos o con potencial de ser científicos que pueden no tener acceso a los medios para desarrollar su carrera profesional. Abrir este tipo de espacios beneficia a que la comunidad científica se expanda y tenga representantes indígenas que puedan ayudar a mejorar la comunicación y dirección de las investigaciones genómicas enfocadas en comunidades indígenas o de ADN de ancestros de comunidades actuales. El consorcio SING fue fundado en EUA y la mayoría de sus eventos anuales se han llevado a cabo en dicho país. Sin embargo, en los últimos años se ha expandido a otros países incluyendo Canadá, Australia, Aotearoa (Nueva Zelanda) y en 2023 se ha llevado a cabo la primera edición de SING México. Durante estos encuentros se han impartido cursos acerca de la investigación paleogenómica desde el trabajo de laboratorio, el análisis bioinformático y hasta sus implicaciones éticas. Iniciativas como el SING son clave para equilibrar la balanza en las ciencias genómicas y lograr una investigación más ética y en condiciones de igualdad. Un aspecto importante es que uno de los objetivos del SING es llegar al mayor número de comunidades indígenas posible. Por ello en los próximos años se espera ver la expansión de este consorcio alrededor del mundo, en especial del sur global.

-Formación de recursos humanos en Paleogenómica, LIIGH

Asimismo, el Laboratorio de Paleogenómica del LIIGH tiene dentro de sus objetivos el establecimiento de colaboraciones equitativas y respetuosas y la formación de recursos humanos locales para disminuir las asimetrías en el campo de la Paleogenómica en México y en el sur global. En acuerdo con estos principios y por voluntad propia, durante la realización de esta tesis participé como co-instructora en diferentes cursos de capacitación en el área de Paleogenómica para estudiantes de la Escuela Nacional de Antropología e Historia en los años 2017, 2018 y 2019. También, impartí dos cursos introductorios a la genómica y Paleogenómica humana, así como de microorganismos, para estudiantes de bachillerato y licenciatura en los cursos de verano 'Clubes de Ciencia México', con título: 1) 'My genome and its hidden history', en el 2018 y 2) 'El gran mundo de los pequeños organismos' en el 2020. Además, participé como asistente en tres cursos de Bioinformática de los cuales fui alumna en un inicio: 1) Escuela de Verano en Bioinformática III, Querétaro en 2018, 2) 'Workshop in Advanced Bioinformatics: Comparative Genomics and Evolution Using Data Science', Querétaro en 2019, y 3) '2nd Workshop in Advanced Bioinformatics', *online* en 2021. Ninguno de estos cursos formó parte de mis créditos del programa de doctorado pero los considero importantes para la inclusión en la ciencia y la transferencia de conocimientos a otros estudiantes. Por otro lado, las colaboraciones que se establecieron con otros grupos de investigación se dieron de la forma más equitativa posible durante la realización de esta tesis. En algunos casos se trabajó con los colaboradores dentro del laboratorio para el procesamiento de las muestras y se llevaron a cabo reuniones de seguimiento de los resultados obtenidos. También, en el caso de los/as arqueólogos/as, se llevaron a cabo reuniones para la discusión de los resultados. De esta manera se obtuvo su retroalimentación durante todo el proceso, la cual enriqueció de manera importante la interpretación de los hallazgos y la dirección de nuevas preguntas de investigación.

Por lo anterior estoy convencida que, aunque la investigación Paleogenómica en México sigue siendo un campo relativamente nuevo, es buen momento para establecer redes de colaboración entre institutos y grupos de investigación mexicanos, que funcionen como una red apoyo para llevar a cabo investigaciones de calidad internacional, con ética, equitativas y soberanas (Fig. 19).



Figura 19. Colaboración equitativa y soberana en la investigación paleogenómica en México.

7.2 Sustentabilidad

Otra característica cuestionable dentro de la investigación paleogenómica es la 'prisa por publicar' (215). El avance en las tecnologías de secuenciación permitió escalar el análisis del ADN de pequeñas amplificaciones mediante PCR a la obtención del genoma completo de un individuo. A partir de la publicación del primer genoma humano antiguo en 2010 (12), la cifra de genomas humanos antiguos se incrementó exponencialmente y hasta 2018 se contaba ya con más de 600 genomas (216). En este momento (inicios de 2023) no se cuenta con una cifra estimada de cuantos genomas antiguos se han publicado hasta la fecha, ya que cada año se publican diversas investigaciones de paleogenómica humana, algunas de las cuales cuentan con cientos de muestras. Inclusive, un grupo líder en paleogenómica en Harvard ha declarado que dentro de sus investigaciones han secuenciado más de 16,000 individuos del pasado de distintas partes del mundo (<https://www.harvardmagazine.com/2022/07/feature-ancient-dna>, febrero 2023). En este sentido, se debe prestar especial atención a evitar la destrucción

injustificada de un alto número de muestras ya que estas muestras podrían ser útiles también para otros estudios, como en el caso del hueso petroso. El hueso petroso es un hueso del cráneo que se ha observado que presenta niveles altos de contenido endógeno. Sin embargo, este hueso incluye el canal auditivo por lo que puede ser importante también para estudios de audición en individuos del pasado (216). Así que, antes de destruir una muestra arqueológica, es importante cuestionarse si ésta puede servir para otros estudios y en la medida de lo posible implementar protocolos que minimicen la destrucción de la muestra. Una opción útil es la preservación digital por medio de escaneo de la muestra en tercera dimensión para conservar su registro morfológico tridimensional para estudios posteriores. En el Laboratorio de Paleogenómica del LIIGH, en los estudios que se iniciaron hace un año (2022), se ha implementado la construcción de moldes y réplicas de las piezas dentales con la finalidad de preservar el patrimonio para futuros estudios. También es importante considerar que las metodologías de extracción de ADN podrían mejorar en un futuro. Esto podría permitir obtener ADN sin destrucción de la muestra, por lo que se debe pensar en el avance de la ciencia y en las futuras generaciones de estudiantes e investigadoras e investigadores. Si terminamos con el material arqueológico por presión de publicar o por querer contestar preguntas de investigación desde una perspectiva egoísta entonces las siguientes generaciones no tendrán material para llevar a cabo mejores investigaciones en un futuro. Por lo anterior es importante promover que la investigación paleogenómica sea sustentable y evitar la destrucción injustificada del patrimonio histórico.

7.3 Relación con poblaciones indígenas y permiso para la obtención de las muestras

Otro aspecto importante es que los restos humanos no deben considerarse como objetos ya que pertenecieron a un ser humano que vivió en el pasado y deben tratarse con el debido respeto (213). Asimismo, en ocasiones estos restos humanos pueden ser considerados como ancestros de comunidades indígenas del presente. En estos casos es necesaria la consulta con representantes de las comunidades para integrarlos como colaboradores de la investigación. Mediante dicha colaboración se debe tomar en cuenta

la disposición de la comunidad para que los restos de sus antepasados sean o no estudiados genéticamente (217, 218). Las comunidades indígenas no han tenido una participación activa en las decisiones para la investigación Paleogenómica en México y en otras regiones del sur global. Esto puede deberse a múltiples aspectos, principalmente por la falta de comunicación entre la comunidad científica y la indígena, o porque la comunidad indígena no se siente vinculada con los restos humanos antiguos (56).

Los permisos para llevar a cabo estudios sobre material arqueológico son otorgados por el Instituto Nacional de Antropología e Historia (INAH) de manera legal en México. El INAH es la autoridad competente para proteger el patrimonio nacional. En este proceso no se ha involucrado a las comunidades indígenas.

Sin embargo, hasta hace unos años no era clara la normatividad al respecto y las publicaciones de ADN que incluían datos genéticos de muestras de México pocas veces mencionaban la obtención de permisos legales. En estos casos es probable que el permiso fuese otorgado verbalmente de común acuerdo entre arqueólogos/as y genetistas. En octubre del 2019 se incluyó el capítulo “Manejo y resguardo de restos humanos durante el proceso de análisis y toma de muestras” en los ‘Lineamientos generales para el manejo y resguardo de restos humanos’ del INAH (<https://www.normateca.inah.gob.mx/pdf/01574791311.PDF>). Este capítulo menciona los requisitos para llevar a cabo estudios genéticos con muestras arqueológicas de colecciones mexicanas. En dichos lineamientos se establece cuál es el proceso a seguir y quiénes son las autoridades competentes que deben otorgar dichos permisos. Entre los requisitos se encuentran: 1) presentar ante el Consejo de Arqueología del INAH la propuesta de un proyecto académico detallado previo a la autorización de la toma de las muestras, 2) que el personal que llevará a cabo los estudios genéticos esté debidamente calificado para tales fines, y 3) que las muestras que se planea analizar estén bien documentadas dentro de la colección (procedencia, datación, entierro, contexto, pieza, etc.). Una parte importante de este capítulo es la recomendación de llevar a cabo dichos análisis en laboratorios e institutos nacionales para salvaguardar el patrimonio osteológico. Como requisitos finales se pide la devolución del material osteológico restante a la colección, y la entrega de los resultados obtenidos de la investigación. Estos incluyen los datos y bibliotecas genómicas generadas y que pasarán a ser propiedad del INAH para ser salvaguardados por dicha institución. Para este último punto se hace

necesario el desarrollo de una infraestructura en el INAH para salvaguardar tanto los datos genómicos como el material molecular. Los datos genómicos deberían ser salvaguardados en un servidor computacional y las bibliotecas genómicas en congeladores de - 20°C.

Un escenario totalmente diferente es el de EUA en el que desde 1990 existe la Ley de Protección y Repatriación de Sepulturas de Nativos Americanos (NAGPRA, por sus siglas en inglés), la cual exige a los grupos de investigación la identificación de las comunidades indígenas que pudieran estar relacionadas con los restos antiguos que se desea estudiar, y consultar con ellas si obtienen su aprobación para tales fines (<https://www.nps.gov/subjects/nagpra/upload/SR101-473.pdf>). Sin embargo, algunos autores sugieren que esto también es problemático desde el punto de vista científico ya que consideran que si obtienen una negativa por parte de la población indígena para llevar a cabo los estudios de ADN o para hacer públicos los datos se limita el avance del conocimiento (214). Dado el contexto histórico, esto no debería interpretarse así y se debería respetar la decisión de las comunidades acerca de sus antepasados; de no hacerlo, se podría dañar a la comunidad profundamente debido a sus creencias y cosmovisión. Asimismo, es importante resaltar que las poblaciones indígenas, desde la colonización por parte de europeos hace 5 siglos, han sufrido un desplazamiento de sus territorios muchas veces a través de la violencia y marginación. Por ello se deben eliminar las prácticas científicas que perpetúen la afectación de éstas.

En una publicación reciente acerca de las prácticas éticas en paleogenómica se exponen opiniones similares en las que los autores sugieren no consultar en algunos casos. Por ejemplo cuando: 1) existan restos humanos que podrían estar relacionados con una o más comunidades indígenas, lo cual resultaría en avivar los conflictos entre éstas, 2) no se encuentra afiliación a una comunidad en específico (214). Estas perspectivas de eliminar la consulta con las poblaciones indígenas podrían traer más problemas al dañar a las comunidades cuando no son consultadas y también tienen repercusiones sobre el campo de la paleogenómica, ya que las comunidades pierden la confianza hacia la comunidad científica en general (219).

Por otro lado, actualmente muchos de los países del sur global carecen de la normatividad para regular la investigación genética de restos humanos. Por lo que es importante que

independientemente de que exista una normativa o no, en la medida de lo posible, las investigadoras e investigadores establezcan colaboraciones simétricas. Para lo cual deben hacer del conocimiento de las personas encargadas de las colecciones y de las comunidades involucradas que se desea llevar a cabo una investigación genética, sus objetivos y el alcance de la investigación. Esto evita crear un daño a las colecciones arqueológicas y a las comunidades relacionadas (56, 212). Estos principios deben aplicarse tanto para las muestras humanas como las no humanas (ej. sedimento) de las que se pueda extraer información genética humana antigua. Asimismo, se contemplan las muestras que se encuentran fuera de su país de origen debido a la apropiación de estos restos durante los procesos extractivos de la colonización, o por saqueos o robo. Aunque estos materiales no se encuentren dentro de las bases de datos del registro arqueológico y no se hayan repatriado, sí forman parte del patrimonio del país de origen.

Existen casos en que las personas a cargo de una investigación de ADN antiguo no han consultado a las comunidades relacionadas con los restos antiguos causando un daño en estas. Tal es el ejemplo de una investigación sobre la población taína, que se realizó sin que se estableciera comunicación con poblaciones indígenas taínas. Como se describe en el artículo (219) en una charla científica 'Genomic Reconstruction of an Extinct Population from Next-Generation Sequence Data' en el congreso de 'Human Genetics' en 2011, los autores del estudio anunciaron la reconstrucción de partes del genoma de los "extintos" taínos. Dicha charla causó molestia y enojo en la comunidad indígena de taínos, quienes manifestaron sentirse altamente ofendidos al haberles llamado públicamente como 'extintos'. Esto pudo haberse evitado si se hubiera establecido una colaboración real con grupos de investigación y comunidades locales, quienes están más informados de su historia y contexto social. Más tarde, en esta misma región se ha desarrollado investigación genética por investigadoras e investigadores locales que conocen bien su historia y han procurado informar los resultados de su investigación con respeto y ética (118).

Por otro lado, existen ejemplos de buenas prácticas, como la que involucró el estudio del individuo llamado Shuká Káa, de hace 10,000 años de antigüedad encontrado en Alaska (220). Desde el descubrimiento de los restos humanos de Shuká Káa, el grupo de investigación se apegó a la ley NAGPRA para identificar a las comunidades indígenas que pudieran estar relacionadas con dicho individuo. Las comunidades acordaron continuar

con la investigación arqueológica y genética, y se estableció que se llevarían a cabo reuniones periódicas para mantener al tanto a la comunidad acerca de los resultados de la investigación. Asimismo, antes de su publicación, los resultados deberían consultarse con la comunidad. Los hallazgos tanto arqueológicos como genéticos fueron de interés para la comunidad para conocer cómo había vivido y fallecido su ancestro Shuká Káa. Además, los hallazgos se alinearon con las historias orales acerca de su origen, en las que contaban que ellos provenían de ancestros que estaban adaptados a la costa, que estaban involucrados en comercios a larga distancia y que se habían mantenido por milenios en esas tierras (19). Estudios como este sirven para resaltar que las colaboraciones pueden resultar en beneficio, no solo para el conocimiento científico, sino también para las comunidades.

Para el trabajo realizado durante esta tesis, las arqueólogas y los arqueólogos fueron quienes se encargaron de los informes y solicitud de permisos ante el INAH, previo a la realización de los trabajos del laboratorio. Al momento de la publicación del capítulo del “Manejo y resguardo de restos humanos durante el proceso de análisis y toma de muestras” en los ‘Lineamientos generales para el manejo y resguardo de restos humanos’ del INAH, el trabajo experimental de esta tesis estaba prácticamente terminado. Sin embargo, previo al envío del artículo científico para su publicación, se preparó un informe con todos los resultados para el Consejo de Arqueología del INAH. Una vez que el informe fue aprobado, y después de haber compartido los datos genómicos con dicho instituto, se envió el artículo a la revista científica para su evaluación y publicación. El remanente de las bibliotecas genómicas generadas para el proyecto hasta el momento sigue en resguardo del laboratorio de Paleogenómica del LIIGH y serán entregadas al INAH cuando éste lo solicite y ya cuente con la infraestructura necesaria para su adecuado resguardo. En cuanto a la consulta con las comunidades indígenas, las arqueólogas y arqueólogos del proyecto, quienes conocen las localidades aledañas a los sitios estudiados y a su población, nos informaron que las comunidades cercanas no los identifican como sus ancestros directos por lo que no hubo una aprobación o rechazo de su parte para los estudios genómicos. Sin embargo, la Dra. Elizabeth Mejía Pérez Campos y el Dr. Alberto Herrera han impartido charlas al público en general de Querétaro y en San Joaquín (la localidad más cercana a los sitios de Toluquilla y Ranas) acerca de los avances de la investigación arqueológica y genética de los sitios. También, en el 2020 en las ‘II Jornadas culturales de arqueología: Cañada de la Virgen’ en San Miguel de

Allende, integrantes del laboratorio de Paleogenómica del LIIGH impartimos charlas de divulgación de los resultados de genómica obtenidos durante la investigación. Entre el público del encuentro se encontraban representantes de las comunidades aledañas quienes recibieron con entusiasmo y curiosidad los resultados obtenidos. Otra opción en México para acercar a las poblaciones indígenas en futuros estudios paleogenómicos sería incluirlas desde el inicio de un proyecto, consultando con ellas la dirección de la investigación y manteniéndolas al tanto de los avances del proyecto.

7.4 Accesibilidad a los datos de ADN

La accesibilidad de los datos es un tema que aún se encuentra en debate (213, 214, 219). Comúnmente para las publicaciones científicas de paleogenómica se pide como requisito que los datos generados se publiquen libremente para que cualquier grupo de investigación sea capaz de replicar los resultados obtenidos en la publicación. Sin embargo, cuando hay comunidades vinculadas a los restos humanos analizados es importante que se tome en cuenta si éstas aprueban o no la publicación de los resultados. En ocasiones, para los datos genómicos modernos, los participantes de dichos estudios tienen la libertad de decidir si sus datos son públicos o si deben permanecer privados y sólo en el caso de que el grupo de investigación se comprometa a no compartirlos y hacer buen uso de ellos, se les comparten. Algo similar podría llevarse a cabo en el ámbito de la paleogenómica, ya que los datos al ser publicados libremente pueden ser utilizados por cualquier persona o empresa en el mundo para los fines que considere pertinentes. Sin embargo, actualmente no hay una normatividad al respecto que se pueda aplicar globalmente, pero es algo que se deberá discutir en los próximos años.

Como se mencionó en párrafos anteriores, actualmente no existe una base de datos completa o un registro donde se encuentren todos los genomas humanos antiguos publicados hasta la fecha, por lo que uno de los laboratorios líderes en el campo de la paleogenómica se ha dado a la tarea de juntar dichos datos en un mismo sitio web para ser descargados en conjunto. No obstante, los datos que ofrecen descargar son en formato de genotipos para 1.23 millones de SNVs y no del genoma completo. A pesar de

ello, pueden ser de mucha utilidad para quienes deseen analizar los SNVs proporcionados, los cuales han sido utilizados en muchas publicaciones. Esta base de datos es actualizada aproximadamente cada dos meses (<https://reich.hms.harvard.edu/allen-ancient-dna-resource-aadr-downloadable-genotypes-present-day-and-ancient-dna-data>, enero 2023). Hasta el momento, este conjunto de datos cuenta con datos genómicos de 9990 genomas antiguos.

8 Discusión

La investigación realizada durante mi doctorado ha contribuido a abrir las puertas a la investigación de Paleogenómica humana en México. Durante este proyecto analicé los restos arqueológicos de 37 individuos de la época prehispánica de México provenientes de distintas zonas arqueológicas, ubicadas en los actuales estados de Guanajuato, Querétaro y Michoacán. Asimismo, tuve acceso a la información de datos de secuenciación de dos momias de la Sierra Tarahumara que habían sido reportadas en una publicación previa a una menor cobertura de profundidad (55). El entrenamiento bioinformático inició precisamente con estos datos de las momias Tarahumaras y fue de gran ayuda para iniciar con los análisis exploratorios desde el mapeo hasta los análisis de PCA (127) y ADMIXTURE (128). Los hallazgos encontrados en este primer estudio Paleogenómico de México nos ha permitido observar por primera vez directamente cómo era la genómica de las poblaciones en la época prehispánica de México y explorar algunos escenarios demográficos de interés.

Es importante mencionar que el reto de realizar esta investigación, además de la curva de aprendizaje, recae en el hecho de que el campo de la Paleogenómica en los últimos años ha avanzado rápidamente y cada año, desde el inicio de este trabajo, se han publicado un número importante de investigaciones. Estas incluyen nuevos métodos de laboratorio, nuevos métodos de análisis y nuevos conjuntos de datos genómicos tanto antiguos como actuales que se pueden incluir en los análisis. Como ejemplo cito el caso del programa hapROH (151) utilizado durante mi investigación doctoral, el cual en 2021 fue publicado para realizar la detección de corridas de homocigosidad en datos de ADN_a. Para mediados del 2022 se publicó una nueva función de este mismo programa, mediante la cual era posible la estimación de la contaminación en bibliotecas genómicas antiguas, basada en el cromosoma X para individuos con genotipo XY (124). Esta última función también la utilicé durante la etapa de revisión del artículo. Por lo que es de esperarse que para el momento de la publicación de esta tesis existan nuevos métodos que no han sido citados aquí.

8.1 Procesos de laboratorio

El trabajo de laboratorio de las muestras inició en abril del 2017 con muestras dentales pertenecientes a individuos de la Sierra Gorda de Querétaro. El procesamiento de estas muestras fue todo un reto y aventura ya que eran las primeras muestras del proyecto que se procesaban en el Laboratorio de Paleogenómica del LIIGH. Gracias a la capacitación impartida por la Dra. Marcela Sandoval Velasco dicho procesamiento fue todo un éxito y con la primera ronda de secuenciación se obtuvieron resultados alentadores. Fue posible verificar la autenticidad de las secuencias antiguas mediante la observación de los patrones de daño en los extremos de las moléculas e inclusive se logró determinar el haplogrupo mitocondrial de algunos de los individuos (Artículo de investigación, Tabla S2, individuos con 'Library ID': MA1GH1SSXXX). De manera inicial se trabajó con el protocolo de extracción de ADNa de *in silico* (62). Posteriormente se adoptó el protocolo de Dabney para el resto de muestras debido a su sencillez, menor costo de reactivos y mejora en la recuperación de secuencias cortas (63).

A pesar de los resultados alentadores, al tratarse de ADNa, en la mayoría de los casos fue necesario aplicar métodos para incrementar la cantidad y diversidad de secuencias recuperadas de cada individuo. Es por ello que dependiendo de la complejidad de las bibliotecas genómicas generadas (70, 93) se decidió realizar capturas de ADNmt (mtC) y de genoma completo (WGC). Un punto importante observado en la realización de las capturas de ADNmt es que éstas pueden tener éxito al realizarse en bibliotecas genómicas con un contenido endógeno realmente bajo, siempre y cuando sea mayor a $>0.01\%$. Anteriormente se había sugerido que el límite mínimo ideal de contenido endógeno era de 1% para realizar capturas genómicas (71); observamos que esto sí se cumple para las WGC. Gracias a las capturas genómicas realizadas se lograron reconstruir 26 genomas mitocondriales a una cobertura de $> 5x$ y 12 genomas nucleares a una cobertura $> 0.01x$, cinco de los cuales fueron de: $0.5x$, $1.4x$, $1.4x$, $2.3x$ y $4.7x$. Estos son hasta el momento los genomas prehispánicos de México con mayor profundidad de cobertura publicados y serán de utilidad para futuras investigaciones sobre la genómica antigua de México.

Cabe resaltar que el proceso de capacitación inicial impartido por la Dra. Marcela Sandoval Velasco, así como la estancia de investigación realizada en el Laboratorio del Dr. Tom Gilbert en agosto-octubre 2017 en la Universidad de Copenhague—líder en el campo de la Paleogenómica—, jugó un papel crucial en la adopción de las técnicas y prácticas del laboratorio. Lo cual a mi criterio puede considerarse una buena colaboración entre el norte global y sur global. Estas técnicas y prácticas de laboratorio se tomaron de base para reproducirlas de la mejor manera posible en el Laboratorio de Paleogenómica del LIIGH. Estas técnicas de laboratorio incluyen: la extracción de ADN, construcción e indexado de bibliotecas genómicas y métodos de captura; las cuales fueron empleadas para el presente trabajo y estandarizadas en el Laboratorio de Paleogenómica del LIIGH y han sido utilizadas por los nuevos estudiantes en los posteriores proyectos que han surgido en el laboratorio.

8.2 Análisis bioinformático de los datos de ADN prehispanico

El análisis bioinformático de las secuencias de ADN inició con los datos de secuenciación de las dos momias de la Sierra Tarahumara previamente generados. Gracias a esto, el análisis de las secuencias generadas en el laboratorio a partir de las muestras prehispanicas fue mucho más sencillo ya que la curva de aprendizaje en bioinformática ya había iniciado.

Los análisis empleados pueden dividirse en dos, aquellos realizados con el genoma mitocondrial y aquellos realizados con el genoma nuclear. Los análisis del genoma mitocondrial permitieron, además de la reconstrucción de este genoma, realizar una primera estimación del porcentaje de contaminación de las bibliotecas genómicas generadas. Esto se realizó mediante dos métodos, el de ContamMix (119) y el de Schmutzi (120), resultando mejor éste último ya que ContamMix tiende a inflar los valores de contaminación cuando se tienen coberturas del genoma mitocondrial de < 30x. Esto no ocurre con Schmutzi que se diseñó tomando en cuenta los patrones de daño de las secuencias antiguas. Asimismo, se logró observar que la estructura genética a nivel del genoma mitocondrial se ha mantenido desde la época prehispanica hasta la fecha, similar

a lo que se había observado en estudios previos de ADN_a en los que se reportaban datos de las regiones hipervariables I y II (203). Además, la determinación de sub-haplogrupos en la Sierra Gorda dio indicios de que existía una continuidad genética en esta región en la línea materna del sub-haplogrupo A2d encontrado en 6 individuos. Uno de ellos era el más antiguo del proyecto (320 BCE) y uno de los individuos más tardíos analizado (1,161 CE).

En cuanto al análisis del genoma nuclear, éste permitió realizar una batería de análisis mucho más amplia. Esto incluyó la determinación de sexo molecular, estimación de la contaminación basado en el cromosoma X, determinación de relaciones de parentesco entre los individuos, análisis de la estructura genética, diversidad genética, continuidad poblacional y contribución de ancestría genética fantasma. A nivel de genoma nuclear también observamos una conservación de la estructura genética desde la época prehispánica hasta la actual. Mientras que la diversidad genética de las poblaciones prehispánicas estudiadas aquí nos indica que éstas presentaban un tamaño efectivo de la población pequeño. Sin embargo, es importante estudiar otros sitios arqueológicos y diferentes periodos para describir de mejor manera cómo era la diversidad genética de cada población. Asimismo se encontró que en la Sierra Gorda de Querétaro hubo una continuidad genética respaldada por diversos análisis tanto exploratorios: PCA y ADMIXTURE así como los estadísticos: outgroup-f₃, D-statistics y qpGraph. Mostrando así que no en todos los sitios de la Frontera Norte de Mesoamérica hubo un reemplazo poblacional debido al periodo de décadas de sequía que obligó a otras poblaciones mesoamericanas a migrar a regiones con mejores condiciones para su subsistencia. Por último, la exploración de la posible contribución de ancestría fantasma en las poblaciones prehispánicas permitió observar por primera vez la contribución de la ancestría genética de U_{popA} en población prehispánica de México, en la Sierra Tarahumara. Por otro lado, este análisis también reveló que en Cañada de la Virgen la ancestría fantasma encontrada es diferente a U_{popA}, por lo que decidimos referirnos a ella como U_{PopA2}.

8.3 Aspectos éticos de la investigación doctoral

En la paleogenómica, al igual que en otras ramas de la ciencia, se ha caído en prácticas neocolonialistas. Grupos de investigación del norte global han tomado ventaja de su privilegio tanto en infraestructura como en presupuestos para obtener muestras del sur global que benefician su investigación pero no aportan beneficios a nivel local. Es por ello que la presente tesis aborda las implicaciones éticas de la investigación paleogenómica ya que en los últimos años se ha hecho un llamado a detener estas prácticas. El presente trabajo, pionero de Paleogenómica humana en México, se condujo en la medida de lo posible para cumplir con una investigación ética, sustentable y soberana. Además de ser creador de recursos humanos capaces de realizar investigación Paleogenómica.

Las muestras arqueológicas estudiadas se mantuvieron en el territorio nacional, y todas ellas se procesaron en el Laboratorio de Paleogenómica del LIIGH para la extracción de ADN y construcción de bibliotecas genómicas. Cuando fue posible se trabajó junto con colaboradores de otros grupos de investigación dentro del laboratorio o con nuevos integrantes del Laboratorio de Paleogenómica del LIIGH para la transferencia de conocimiento, formando así recursos humanos a nivel local. Asimismo, como parte de esta formación de recursos humanos, se estableció la estructura de la presente tesis, la cual incluye diferentes capítulos teóricos explicados de manera sencilla para servir de guía a estudiantes con interés en este campo.

A medida que han pasado los años, desde el inicio de este trabajo, también se han modificado los requisitos para llevar a cabo investigación genómica con restos óseos de México. La creación de un apartado con respecto a esto, en los 'Lineamientos generales para el manejo y resguardo de restos humanos' del INAH, en 2019, ha sido importante para salvaguardar el patrimonio de la nación y establecer colaboraciones bidireccionales en las que tanto grupos locales como extranjeros se benefician de la investigación Paleogenómica, y no sólo estos últimos. Para cumplir con dichos lineamientos, la presente investigación fue compartida con el INAH para su evaluación y aprobación, previo al envío del artículo científico.

9 Conclusiones

Durante el trabajo de investigación de esta tesis fue posible recuperar ADN de restos arqueológicos de la época prehispánica de México. El ADN recuperado comprendió tanto secuencias del genoma nuclear como mitocondrial y presentó los patrones de daño característicos de ADN, con lo que se verificó su autenticidad. La información genética recuperada permitió el estudio del linaje mitocondrial de 26 individuos, para todos ellos el haplogrupo asignado fue uno de los 4 haplogrupos fundadores del poblamiento de América. A nivel del linaje materno observamos que la estructura genética se ha mantenido desde la época prehispánica hasta la actual. En cuanto al linaje paterno, sólo fue posible analizarlo para 5 individuos, todos ellos presentaron el haplogrupo Q, el cual es el linaje fundador encontrado en América.

Los análisis del genoma nuclear mostraron que también la estructura genética a este nivel se ha mantenido desde la época prehispánica hasta la actual. En cuanto a la diversidad genética observamos niveles de diversidad similares entre los individuos de la época prehispánica y las poblaciones indígenas actuales de México, sin embargo, es necesaria una exploración a futuro en otras poblaciones prehispánicas y actuales que no están representadas en este estudio.

A través de distintos análisis tanto exploratorios como estadísticos observamos que existe continuidad genética en la Sierra Gorda de Querétaro la cual se encontraba en la Frontera Norte de Mesoamérica, y en la cual se creía que había existido un reemplazo poblacional. Mediante los análisis de ADN demostramos en los sitios estudiados no existió un reemplazo poblacional, por lo que la población de esta región fue capaz de adaptarse al periodo de sequías prolongadas que afectó de manera negativa a otros pueblos, como por ejemplo a los habitantes de Cañada de la Virgen, quienes sí tuvieron que abandonar su región probablemente en busca de mejores condiciones ambientales.

El análisis de la contribución de ancestría genética fantasma muestra que la población prehispánica de la Sierra Tarahumara comparte con los Mixe la ancestría genética proveniente de UtopA. Por otro lado, en Cañada de la Virgen encontramos también una contribución genética fantasma, pero en este caso no se comparte con los Mixe, sino que,

es una contribución no descrita previamente. Esta información en conjunto aporta nuevas pistas acerca del poblamiento de América.

10 Perspectivas

En el presente trabajo se realizaron hallazgos a distintos niveles.

En cuanto al genoma mitocondrial, en futuros estudios sería interesante recuperar una mayor cantidad de genomas mitocondriales que abarquen un rango geográfico más amplio: del sur, sureste de México, Noreste, centro-oeste y centro-este y observar la distribución de los linajes maternos en diferentes periodos. Esto sería útil para comprender la continuidad genética de los linajes maternos y observar posibles cambios que sugieran desplazamientos poblacionales o mezclas, las cuales pueden ser soportadas con los registros arqueológicos.

En cuanto al linaje paterno estudiado a partir del cromosoma Y se espera que sea posible la captura de las regiones informativas de este cromosoma en el ADN de individuos prehispánicos para comprender mejor la diversidad de éste. Este tópico es de importancia porque la población masculina nativa se vió severamente afectada al ocurrir la colonización por parte de europeos. Esto lo respalda el cruzamiento asimétrico que ocurrió en la colonia y que observamos hoy en día en los marcadores uniparentales en México. La mayoría de la población mexicana presenta algún haplogrupo mitocondrial de linaje nativo americano. Mientras que los varones en su linaje paterno muestran una gran representación de haplogrupo del cromosoma Y europeo.

En cuanto al genoma nuclear, los análisis realizados en la presente tesis pueden ser refinados al obtener genomas con profundidad de cobertura más amplios. Los análisis que se podrían mejorar con genomas de mayor cobertura serían los de diversidad genética y los de modelado demográfico.

El escenario de continuidad poblacional podría explorarse en mayor detalle en caso de contar con un mayor número de muestras. Con estas exploraciones a más profundidad

podría evaluarse la presencia o ausencia de otros haplogrupos mitocondriales distintos al A2d.

Para el hallazgo de la contribución genética fantasma en las poblaciones prehispánicas sería interesante explorar en qué otras poblaciones antiguas de México se presentan y tratar de dilucidar los eventos demográficos que dieron lugar a la dispersión de este componente genético.

11 Literatura citada

1. B. Hermann, S. Hummel, Ancient DNA: recovery and analysis of genetic material from paleontological, archaeological, museum, medical, and forensic specimens (1996).
2. R. Higuchi, B. Bowman, M. Freiberger, O. A. Ryder, A. C. Wilson, DNA sequences from the quagga, an extinct member of the horse family. *Nature*. **312**, 282–284 (1984).
3. M. E. Allentoft, M. Collins, D. Harker, J. Haile, C. L. Oskam, M. L. Hale, *et al.*, The half-life of DNA in bone: measuring decay kinetics in 158 dated fossils. *Proc. Biol. Sci.* **279**, 4724–4733 (2012).
4. T. van der Valk, P. Pečnerová, D. Díez-Del-Molino, A. Bergström, J. Oppenheimer, S. Hartmann, *et al.*, Million-year-old DNA sheds light on the genomic history of mammoths. *Nature*. **591**, 265–269 (2021).
5. M. D. Sorenson, A. Cooper, E. E. Paxinos, T. W. Quinn, H. F. James, S. L. Olson, *et al.*, Relationships of the extinct moa-nalos, flightless Hawaiian waterfowl, based on ancient DNA. *Proc. Biol. Sci.* **266**, 2187–2193 (1999).
6. J. F. Bailey, M. B. Richards, V. A. Macaulay, I. B. Colson, I. T. James, D. G. Bradley, *et al.*, Ancient DNA suggests a recent expansion of European cattle from a diverse wild progenitor species. *Proc. Biol. Sci.* **263**, 1467–1473 (1996).
7. H. Yang, Ancient dna from pleistocene fossils: Preservation, recovery, and utility of ancient genetic information for quaternary research. *Quat. Sci. Rev.* **16**, 1145–1161 (1997).
8. E. M. Golenberg, D. E. Giannasi, M. T. Clegg, C. J. Smiley, M. Durbin, D. Henderson, *et al.*, Chloroplast DNA sequence from a miocene Magnolia species. *Nature*. **344**, 656–658 (1990).
9. S. Pääbo, J. A. Gifford, A. C. Wilson, Mitochondrial DNA sequences from a 7000-year old brain. *Nucleic Acids Res.* **16**, 9775–9787 (1988).
10. A. C. Stone, M. Stoneking, Ancient DNA from a pre-Columbian Amerindian population. *American journal of physical* (1993) (available at <https://onlinelibrary.wiley.com/doi/abs/10.1002/ajpa.1330920405>).
11. M. Hofreiter, B. Shapiro, *Ancient DNA: methods and protocols* (Humana Press Incorporated, 2012).
12. 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*. **463**, 757–762 (2010).
13. Z. Darzynkiewicz, G. Juan, X. Li, W. Gorczyca, T. Murakami, F. Traganos, Cytometry in cell necrobiology: analysis of apoptosis and accidental cell death (necrosis). *Cytometry*. **27**, 1–20 (1997).
14. A. W. Briggs, U. Stenzel, P. L. F. Johnson, R. E. Green, J. Kelso, K. Prüfer, *et al.*, Patterns of damage in genomic DNA sequences from a Neandertal. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 14616–14621 (2007).
15. A. Ginolhac, M. Rasmussen, M. T. P. Gilbert, E. Willerslev, L. Orlando, mapDamage: testing for

- damage patterns in ancient DNA sequences. *Bioinformatics*. **27**, 2153–2155 (2011).
16. J. Goody, *The Interface Between the Written and the Oral* (Cambridge University Press, 1987).
 17. K. W. Arafat, Fact and Artefact: Texts and Archaeology. *Hermathena*, 45–67 (1990).
 18. S. L. Washburn, in *Evolutionary Biology: Volume 6*, T. Dobzhansky, M. K. Hecht, W. C. Steere, Eds. (Springer US, New York, NY, 1972), pp. 349–361.
 19. J. Raff, *Origin: A Genetic History of the Americas* (Grand Central Publishing, 2022).
 20. B. Pakendorf, M. Stoneking, Mitochondrial DNA and human evolution. *Annu. Rev. Genomics Hum. Genet.* **6**, 165–183 (2005).
 21. 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*. **341**, 562–565 (2013).
 22. M. van Oven, M. Kayser, Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. *Hum. Mutat.* **30**, E386–94 (2009).
 23. T. Kivisild, Maternal ancestry and population history from whole mitochondrial genomes. *Investig. Genet.* **6**, 3 (2015).
 24. K. Drosou, C. Price, T. A. Brown, The kinship of two 12th Dynasty mummies revealed by ancient DNA sequencing. *Journal of Archaeological Science: Reports*. **17**, 793–797 (2018).
 25. B. Llamas, L. Fehren-Schmitz, G. Valverde, J. Soubrier, S. Mallick, N. Rohland, *et al.*, Ancient mitochondrial DNA provides high-resolution time scale of the peopling of the Americas. *Sci Adv.* **2**, e1501385 (2016).
 26. C. L. Scheib, H. Li, T. Desai, V. Link, C. Kendall, G. Dewar, *et al.*, Ancient human parallel lineages within North America contributed to a coastal expansion. *Science*. **360**, 1024–1027 (2018).
 27. G. D. Poznik, Y. Xue, F. L. Mendez, T. F. Willems, A. Massaia, M. A. Wilson Sayres, *et al.*, Punctuated bursts in human male demography inferred from 1,244 worldwide Y-chromosome sequences. *Nat. Genet.* **48**, 593–599 (2016).
 28. S. L. Zegura, T. M. Karafet, L. A. Zhivotovsky, M. F. Hammer, High-resolution SNPs and microsatellite haplotypes point to a single, recent entry of Native American Y chromosomes into the Americas. *Mol. Biol. Evol.* **21**, 164–175 (2004).
 29. 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*. **8**, e71390 (2013).
 30. B. Z. González-Sobrinó, A. P. Pintado-Cortina, L. Sebastián-Medina, F. Morales-Mandujano, A. V. Contreras, Y. E. Aguilar, *et al.*, Genetic Diversity and Differentiation in Urban and Indigenous Populations of Mexico: Patterns of Mitochondrial DNA and Y-Chromosome Lineages. *Biodemography Soc. Biol.* **62**, 53–72 (2016).
 31. David Reich (Of Harvard Medical School), *Who We are and how We Got Here: Ancient DNA and the New Science of the Human Past* (Oxford University Press, 2018).
 32. R. Bergero, P. Ellis, W. Haerty, L. Larcombe, I. Macaulay, T. Mehta, *et al.*, Meiosis and beyond - understanding the mechanistic and evolutionary processes shaping the germline genome. *Biol. Rev. Camb. Philos. Soc.* **96**, 822–841 (2021).

33. National Institutes of Health (US), Biological Sciences Curriculum Study, *Understanding Human Genetic Variation* (National Institutes of Health (US), 2007).
34. 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*. **538**, 201–206 (2016).
35. 1000 Genomes Project Consortium, G. R. Abecasis, A. Auton, L. D. Brooks, M. A. DePristo, R. M. Durbin, *et al.*, An integrated map of genetic variation from 1,092 human genomes. *Nature*. **491**, 56–65 (2012).
36. A. Bergström, S. A. McCarthy, R. Hui, M. A. Almarri, Q. Ayub, P. Danecek, *et al.*, Insights into human genetic variation and population history from 929 diverse genomes. *Science*. **367** (2020), doi:10.1126/science.aay5012.
37. P. Skoglund, I. Mathieson, Ancient Genomics of Modern Humans: The First Decade. *Annu. Rev. Genomics Hum. Genet.* **19**, 381–404 (2018).
38. L. Orlando, R. Allaby, P. Skoglund, C. Der Sarkissian, P. W. Stockhammer, M. C. Ávila-Arcos, *et al.*, Ancient DNA analysis. *Nature Reviews Methods Primers*. **1**, 1–26 (2021).
39. Vargas Sanders, Rocio. Tesis 'De los esqueletos a la doble hélice : hacia un estudio de paleoantropología molecular'. UNAM. (1993).
https://ru.dgb.unam.mx/handle/DGB_UNAM/TES01000193007.
40. R. Vargas-Sanders, Z. Salazar, M. C. Enriquez, in *Archaeological Chemistry* (American Chemical Society, 1996), vol. 625 of *ACS Symposium Series*, pp. 391–400.
41. A. González-Oliver, L. Márquez-Morfin, J. C. Jiménez, A. Torre-Blanco, Founding Amerindian mitochondrial DNA lineages in ancient Maya from Xcaret, Quintana Roo. *Am. J. Phys. Anthropol.* **116**, 230–235 (2001).
42. I. D. L. Cruz, I. De La Cruz, A. González-Oliver, B. M. Kemp, J. A. Román, D. G. Smith, *et al.*, Sex Identification of Children Sacrificed to the Ancient Aztec Rain Gods in Tlatelolco. *Current Anthropology*. **49** (2008), pp. 519–526.
43. M. I. Ochoa-Lugo, M. de L. Muñoz, G. Pérez-Ramírez, K. G. Beaty, M. López-Armenta, J. Cervini-Silva, *et al.*, Genetic Affiliation of Pre-Hispanic and Contemporary Mayas Through Maternal Lineage. *Hum. Biol.* **88**, 136–167 (2016).
44. B. A. Álvarez-Sandoval, L. R. Manzanilla, R. Montiel, Sex determination in highly fragmented human DNA by high-resolution melting (HRM) analysis. *PLoS One*. **9**, e104629 (2014).
45. A. J. Aguirre-Samudio, B. Z. González-Sobrino, B. A. Álvarez-Sandoval, R. Montiel, C. Serrano-Sánchez, A. Meza-Peñaloza, Genetic history of classic period Teotihuacan burials in Central Mexico. *Revista argentina de antropología biológica*. **19**, 7–7 (2017).
46. J. A. R. Berelleza, A. S. A. Salinas, A. G. Oliver, Identificación de la presencia de ADN antiguo en restos óseos de la cultura capacha de Colima. *Diario de Campo*, 94–104 (2015).
47. B. A. Álvarez-Sandoval, L. R. Manzanilla, M. González-Ruiz, A. Malgosa, R. Montiel, Genetic Evidence Supports the Multiethnic Character of Teopancazco, a Neighborhood Center of Teotihuacan, Mexico (AD 200-600). *PLoS One*. **10**, e0132371 (2015).
48. P. Skoglund, J. Storå, A. Götherström, M. Jakobsson, Accurate sex identification of ancient human remains using DNA shotgun sequencing. *Journal of Archaeological Science*. **40** (2013), pp. 4477–4482.
49. E. Solórzano, N. Díaz, R. Montiel, A. Malgosa, Análisis del ADN mitocondrial de tres series

- antiguas mexicanas. Estudios de antropología biológica. 14 (2011).<http://revistas.unam.mx/index.php/eab/article/view/27243>.
50. J. Mata-Míguez, L. Overholtzer, E. Rodríguez-Alegría, B. M. Kemp, D. A. Bolnick, The genetic impact of Aztec imperialism: ancient mitochondrial DNA evidence from Xaltocan, Mexico. *Am. J. Phys. Anthropol.* **149**, 504–516 (2012).
 51. J. C. Chatters, D. J. Kennett, Y. Asmerom, B. M. Kemp, V. Polyak, A. N. Blank, *et al.*, Late Pleistocene human skeleton and mtDNA link Paleoamericans and modern Native Americans. *Science*. **344**, 750–754 (2014).
 52. A. Y. Morales-Arce, C. A. Hofman, A. T. Duggan, A. K. Benfer, M. Anne Katzenberg, G. McCafferty, *et al.*, Successful reconstruction of whole mitochondrial genomes from ancient Central America and Mexico. *Scientific Reports*. **7** (2017), , doi:10.1038/s41598-017-18356-0.
 53. A. Y. Morales-Arce, G. McCafferty, J. Hand, N. Schmill, K. McGrath, C. Speller, Ancient mitochondrial DNA and population dynamics in postclassic Central Mexico: Tlatelolco (ad 1325–1520) and Cholula (ad 900–1350). *Archaeol. Anthropol. Sci.* **11**, 3459–3475 (2019).
 54. A. Y. Morales-Arce, M. H. Snow, J. H. Kelley, M. Anne Katzenberg, Ancient mitochondrial DNA and ancestry of Paquimé inhabitants, Casas Grandes (A.D. 1200-1450). *Am. J. Phys. Anthropol.* **163**, 616–626 (2017).
 55. M. Raghavan, M. Steinrücken, K. Harris, S. Schiffels, S. Rasmussen, M. DeGiorgio, *et al.*, POPULATION GENETICS. Genomic evidence for the Pleistocene and recent population history of Native Americans. *Science*. **349**, aab3884 (2015).
 56. M. C. Ávila-Arcos, C. de la Fuente Castro, M. A. Nieves-Colón, M. Raghavan, Recommendations for Sustainable Ancient DNA Research in the Global South: Voices From a New Generation of Paleogenomicists. *Frontiers in Genetics*. **13** (2022), , doi:10.3389/fgene.2022.880170.
 57. M. A. Nieves-Colón, K. E. Blevins, Paleogenómica y bioarqueología en México. *Cuicuilco. Revista de* (2021) (available at https://www.scielo.org.mx/scielo.php?pid=S2448-84882021000200009&script=sci_arttext).
 58. S. R. Woodward, N. J. Weyand, M. Bunnell, DNA Sequence from Cretaceous Period Bone Fragments. *Science*. **266** (1994), pp. 1229–1232.
 59. R. J. Cano, H. N. Poinar, N. J. Pieniazek, A. Acra, G. O. Poinar Jr, Amplification and sequencing of DNA from a 120-135-million-year-old weevil. *Nature*. **363**, 536–538 (1993).
 60. J. J. Austin, A. J. Ross, A. B. Smith, R. A. Fortey, R. H. Thomas, Problems of reproducibility – does geologically ancient DNA survive in amber–preserved insects? *Proceedings of the Royal Society of London. Series B: Biological Sciences*. **264** (1997), pp. 467–474.
 61. M. Stoneking, Ancient DNA: how do you know when you have it and what can you do with it? *Am. J. Hum. Genet.* **57**, 1259–1262 (1995).
 62. N. Rohland, M. Hofreiter, Ancient DNA extraction from bones and teeth. *Nature Protocols*. **2** (2007), pp. 1756–1762.
 63. J. Dabney, M. Knapp, I. Glocke, M.-T. Gansauge, A. Weihmann, B. Nickel, *et al.*, Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. *Proceedings of the National Academy of Sciences*. **110** (2013), pp. 15758–15763.
 64. M. Meyer, M. Kircher, Illumina sequencing library preparation for highly multiplexed target

- capture and sequencing. *Cold Spring Harb. Protoc.* **2010**, db.prot5448 (2010).
65. C. Carøe, S. Gopalakrishnan, L. Vinner, S. S. T. Mak, M. H. S. Sinding, J. A. Samaniego, *et al.*, Single-tube library preparation for degraded DNA. *Methods in Ecology and Evolution.* **9** (2018), pp. 410–419.
 66. M.-T. Gansauge, M. Meyer, Single-stranded DNA library preparation for the sequencing of ancient or damaged DNA. *Nature Protocols.* **8** (2013), pp. 737–748.
 67. M.-T. Gansauge, T. Gerber, I. Glocke, P. Korlević, L. Lippik, S. Nagel, *et al.*, Single-stranded DNA library preparation from highly degraded DNA using T4DNA ligase. *Nucleic Acids Research* (2017), p. gkx033.
 68. C. J. Troll, J. Kapp, V. Rao, K. M. Harkins, C. Cole, C. Naughton, *et al.*, A ligation-based single-stranded library preparation method to analyze cell-free DNA and synthetic oligos. *BMC Genomics.* **20** (2019), , doi:10.1186/s12864-019-6355-0.
 69. J. D. Kapp, R. E. Green, B. Shapiro, A Fast and Efficient Single-stranded Genomic Library Preparation Method Optimized for Ancient DNA. *Journal of Heredity.* **112** (2021), pp. 241–249.
 70. M. C. Ávila-Arcos, M. Sandoval-Velasco, H. Schroeder, M. L. Carpenter, A. Malaspinas, N. Wales, *et al.*, Comparative performance of two whole-genome capture methodologies on ancient DNA Illumina libraries. *Methods in Ecology and Evolution.* **6** (2015), pp. 725–734.
 71. M. L. Carpenter, J. D. Buenrostro, C. Valdiosera, H. Schroeder, M. E. Allentoft, M. Sikora, *et al.*, Pulling out the 1%: Whole-Genome Capture for the Targeted Enrichment of Ancient DNA Sequencing Libraries. *The American Journal of Human Genetics.* **93** (2013), pp. 852–864.
 72. T. Maricic, M. Whitten, S. Pääbo, Multiplexed DNA Sequence Capture of Mitochondrial Genomes Using PCR Products. *PLoS ONE.* **5** (2010), p. e14004.
 73. S. S. T. Mak, S. Gopalakrishnan, C. Carøe, C. Geng, S. Liu, M.-H. S. Sinding, *et al.*, Comparative performance of the BGISEQ-500 vs Illumina HiSeq2500 sequencing platforms for palaeogenomic sequencing. *GigaScience.* **6** (2017), , doi:10.1093/gigascience/gix049.
 74. S. Goodwin, J. D. McPherson, W. Richard McCombie, Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics.* **17** (2016), pp. 333–351.
 75. P. J. A. Cock, C. J. Fields, N. Goto, M. L. Heuer, P. M. Rice, The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Research.* **38** (2010), pp. 1767–1771.
 76. M. Martin, Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal.* **17** (2011), p. 10.
 77. S. Lindgreen, AdapterRemoval: easy cleaning of next-generation sequencing reads. *BMC Research Notes.* **5** (2012), , doi:10.1186/1756-0500-5-337.
 78. A. M. Bolger, M. Lohse, B. Usadel, Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics.* **30** (2014), pp. 2114–2120.
 79. M. Schubert, S. Lindgreen, L. Orlando, AdapterRemoval v2: rapid adapter trimming, identification, and read merging. *BMC Research Notes.* **9** (2016), , doi:10.1186/s13104-016-1900-2.
 80. R. M. Andrews, I. Kubacka, P. F. Chinnery, R. N. Lightowlers, D. M. Turnbull, N. Howell, Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nat. Genet.* **23**, 147 (1999).

81. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*. **25**, 1754–1760 (2009).
82. H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, *et al.*, The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. **25**, 2078–2079 (2009).
83. 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 Res.* **20**, 1297–1303 (2010).
84. H. Jónsson, A. Ginolhac, M. Schubert, P. L. F. Johnson, L. Orlando, mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics*. **29** (2013), pp. 1682–1684.
85. V. Link, A. Kousathanas, K. Veeramah, C. Sell, A. Scheu, D. Wegmann, ATLAS: Analysis Tools for Low-depth and Ancient Samples. *bioRxiv* (2017), p. 105346.
86. H. Al-Asadi, K. K. Dey, J. Novembre, M. Stephens, Inference and visualization of DNA damage patterns using a grade of membership model. *Bioinformatics*. **35**, 1292–1298 (2019).
87. M. Borry, A. Hübner, A. B. Rohrlach, C. Warinner, PyDamage: automated ancient damage identification and estimation for contigs in ancient DNA de novo assembly. *PeerJ*. **9**, e11845 (2021).
88. S. Peyrégne, B. M. Peter, AuthentiCT: a model of ancient DNA damage to estimate the proportion of present-day DNA contamination. *Genome Biol.* **21**, 246 (2020).
89. J. Neukamm, A. Peltzer, K. Nieselt, DamageProfiler: Fast damage pattern calculation for ancient DNA. *Bioinformatics* (2021), doi:10.1093/bioinformatics/btab190.
90. M. T. W. Ebbert, M. E. Wadsworth, L. A. Staley, K. L. Hoyt, B. Pickett, J. Miller, *et al.*, Evaluating the necessity of PCR duplicate removal from next-generation sequencing data and a comparison of approaches. *BMC Bioinformatics*. **17 Suppl 7**, 239 (2016).
91. Picard, (available at <https://broadinstitute.github.io/picard/>).
92. A. Peltzer, G. Jäger, A. Herbig, A. Seitz, C. Kniep, J. Krause, *et al.*, EAGER: efficient ancient genome reconstruction. *Genome Biol.* **17**, 60 (2016).
93. T. Daley, A. D. Smith, Predicting the molecular complexity of sequencing libraries. *Nat. Methods*. **10**, 325–327 (2013).
94. I. J. Good, G. H. Toulmin, THE NUMBER OF NEW SPECIES, AND THE INCREASE IN POPULATION COVERAGE, WHEN A SAMPLE IS INCREASED. *Biometrika*. **43**, 45–63 (1956).
95. M. Schubert, L. Ermini, C. Der Sarkissian, H. Jónsson, A. Ginolhac, R. Schaefer, *et al.*, Characterization of ancient and modern genomes by SNP detection and phylogenomic and metagenomic analysis using PALEOMIX. *Nat. Protoc.* **9**, 1056–1082 (2014).
96. A. R. Kales, *Sex Estimation of the Human Skeleton: History, Methods, and Emerging Techniques* (Academic Press, 2020).
97. M. Faerman, D. Filon, G. Kahila, C. L. Greenblatt, P. Smith, A. Oppenheim, Sex identification of archaeological human remains based on amplification of the X and Y amelogenin alleles. *Gene*. **167**, 327–332 (1995).
98. A. C. Stone, G. R. Milner, S. Pääbo, M. Stoneking, Sex determination of ancient human

- skeletons using DNA. *Am. J. Phys. Anthropol.* **99**, 231–238 (1996).
99. J. Krause, A. W. Briggs, M. Kircher, T. Maricic, N. Zwyns, A. Derevianko, *et al.*, A complete mtDNA genome of an early modern human from Kostenki, Russia. *Curr. Biol.* **20**, 231–236 (2010).
 100. H. Maude, M. Davidson, N. Charitakis, L. Diaz, W. H. T. Bowers, E. Gradovich, *et al.*, NUMT Confounding Biases Mitochondrial Heteroplasmy Calls in Favor of the Reference Allele. *Frontiers in Cell and Developmental Biology.* **7** (2019), , doi:10.3389/fcell.2019.00201.
 101. T. S. Korneliussen, A. Albrechtsen, R. Nielsen, ANGSD: Analysis of Next Generation Sequencing Data. *BMC Bioinformatics.* **15**, 356 (2014).
 102. A. Kloss-Brandstätter, D. Pacher, S. Schönherr, H. Weissensteiner, R. Binna, G. Specht, *et al.*, HaploGrep: a fast and reliable algorithm for automatic classification of mitochondrial DNA haplogroups. *Hum. Mutat.* **32**, 25–32 (2011).
 103. I. Milne, M. Bayer, L. Cardle, P. Shaw, G. Stephen, F. Wright, *et al.*, Tablet—next generation sequence assembly visualization. *Bioinformatics.* **26**, 401–402 (2009).
 104. H. Weissensteiner, D. Pacher, A. Kloss-Brandstätter, L. Forer, G. Specht, H.-J. Bandelt, *et al.*, HaploGrep 2: mitochondrial haplogroup classification in the era of high-throughput sequencing. *Nucleic Acids Res.* **44**, W58–63 (2016).
 105. D. I. Cruz-Dávalos, M. A. Nieves-Colón, A. Sockell, G. D. Poznik, H. Schroeder, A. C. Stone, *et al.*, In-solution Y-chromosome capture-enrichment on ancient DNA libraries. *BMC Genomics.* **19**, 608 (2018).
 106. G. D. Poznik, Identifying Y-chromosome haplogroups in arbitrarily large samples of sequenced or genotyped men. *bioRxiv* (2016), , doi:10.1101/088716.
 107. T. Schultes, S. Hummel, B. Herrmann, Ancient DNA-typing approaches for the determination of kinship in a disturbed collective burial site. *Anthropol. Anz.* **58**, 37–44 (2000).
 108. M. Baca, K. Doan, M. Sobczyk, A. Stankovic, P. Węgleński, Ancient DNA reveals kinship burial patterns of a pre-Columbian Andean community. *BMC Genet.* **13**, 30 (2012).
 109. S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira, D. Bender, *et al.*, PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
 110. T. Thornton, H. Tang, T. J. Hoffmann, H. M. Ochs-Balcom, B. J. Caan, N. Risch, Estimating kinship in admixed populations. *Am. J. Hum. Genet.* **91**, 122–138 (2012).
 111. H. Li, G. Glusman, C. Huff, J. Caballero, J. C. Roach, Accurate and robust prediction of genetic relationship from whole-genome sequences. *PLoS One.* **9**, e85437 (2014).
 112. A. Ko, R. Nielsen, Joint Estimation of Pedigrees and Effective Population Size Using Markov Chain Monte Carlo. *Genetics.* **212**, 855–868 (2019).
 113. M. D. Martin, F. Jay, S. Castellano, M. Slatkin, Determination of genetic relatedness from low-coverage human genome sequences using pedigree simulations. *Mol. Ecol.* **26**, 4145–4157 (2017).
 114. C. Theunert, F. Racimo, M. Slatkin, Joint Estimation of Relatedness Coefficients and Allele Frequencies from Ancient Samples. *Genetics.* **206**, 1025–1035 (2017).
 115. J. M. Monroy Kuhn, M. Jakobsson, T. Günther, Estimating genetic kin relationships in

- prehistoric populations. *PLoS One*. **13**, e0195491 (2018).
116. D. M. Fernandes, O. Cheronet, P. Gelabert, R. Pinhasi, TKGWV2: an ancient DNA relatedness pipeline for ultra-low coverage whole genome shotgun data. *Sci. Rep.* **11**, 21262 (2021).
 117. D. Popli, S. Peyrégne, B. M. Peter, KIN: a method to infer relatedness from low-coverage ancient DNA. *Genome Biol.* **24**, 10 (2023).
 118. M. A. Nieves-Colón, W. J. Pestle, A. W. Reynolds, B. Llamas, C. de la Fuente, K. Fowler, *et al.*, Ancient DNA Reconstructs the Genetic Legacies of Precontact Puerto Rico Communities. *Mol. Biol. Evol.* **37**, 611–626 (2020).
 119. Q. Fu, A. Mittnik, P. L. F. Johnson, K. Bos, M. Lari, R. Bollongino, *et al.*, A revised timescale for human evolution based on ancient mitochondrial genomes. *Curr. Biol.* **23**, 553–559 (2013).
 120. G. Renaud, V. Slon, A. T. Duggan, J. Kelso, Schmutzi: estimation of contamination and endogenous mitochondrial consensus calling for ancient DNA. *Genome Biol.* **16**, 224 (2015).
 121. N. Rohland, E. Harney, S. Mallick, S. Nordenfelt, D. Reich, Partial uracil–DNA–glycosylase treatment for screening of ancient DNA. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **370**, 20130624 (2015).
 122. F. Racimo, G. Renaud, M. Slatkin, Joint Estimation of Contamination, Error and Demography for Nuclear DNA from Ancient Humans. *PLoS Genet.* **12**, e1005972 (2016).
 123. J. V. Moreno-Mayar, T. S. Korneliussen, J. Dalal, G. Renaud, A. Albrechtsen, R. Nielsen, *et al.*, A likelihood method for estimating present-day human contamination in ancient male samples using low-depth X-chromosome data. *Bioinformatics.* **36**, 828–841 (2020).
 124. Y. Huang, H. Ringbauer, hapCon: estimating contamination of ancient genomes by copying from reference haplotypes. *Bioinformatics.* **38**, 3768–3777 (2022).
 125. P. Skoglund, B. H. Northoff, M. V. Shunkov, A. P. Derevianko, S. Pääbo, J. Krause, *et al.*, Separating endogenous ancient DNA from modern day contamination in a Siberian Neandertal. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 2229–2234 (2014).
 126. J. K. Pritchard, M. Stephens, P. Donnelly, Inference of population structure using multilocus genotype data. *Genetics.* **155**, 945–959 (2000).
 127. N. Patterson, A. L. Price, D. Reich, Population structure and eigenanalysis. *PLoS Genet.* **2**, e190 (2006).
 128. D. H. Alexander, J. Novembre, K. Lange, Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
 129. P. Skoglund, M. Jakobsson, Archaic human ancestry in East Asia. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 18301–18306 (2011).
 130. P. Skoglund, H. Malmström, M. Raghavan, J. Storå, P. Hall, E. Willerslev, *et al.*, Origins and genetic legacy of Neolithic farmers and hunter-gatherers in Europe. *Science.* **336**, 466–469 (2012).
 131. C. Wang, Z. A. Szpiech, J. H. Degnan, M. Jakobsson, T. J. Pemberton, J. A. Hardy, *et al.*, *Stat. Appl. Genet. Mol. Biol.*, in press.
 132. O. François, F. Jay, Factor analysis of ancient population genomic samples. *Nat. Commun.* **11**, 4661 (2020).

133. R. Nielsen, M. Slatkin, *An Introduction to Population Genetics: Theory and Applications* (Sinauer, 2013).
134. L. Skotte, T. S. Korneliussen, A. Albrechtsen, Estimating individual admixture proportions from next generation sequencing data. *Genetics*. **195**, 693–702 (2013).
135. A. A. Behr, K. Z. Liu, G. Liu-Fang, P. Nakka, S. Ramachandran, pong: fast analysis and visualization of latent clusters in population genetic data. *Bioinformatics*. **32**, 2817–2823 (2016).
136. Q. Feng, D. Lu, S. Xu, AncestryPainter: A Graphic Program for Displaying Ancestry Composition of Populations and Individuals. *Genomics Proteomics Bioinformatics*. **16**, 382–385 (2018).
137. D. J. Lawson, L. van Dorp, D. Falush, A tutorial on how not to over-interpret STRUCTURE and ADMIXTURE bar plots. *Nat. Commun.* **9**, 3258 (2018).
138. J. K. Pickrell, J. K. Pritchard, Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet.* **8**, e1002967 (2012).
139. N. Patterson, P. Moorjani, Y. Luo, S. Mallick, N. Rohland, Y. Zhan, *et al.*, Ancient admixture in human history. *Genetics*. **192**, 1065–1093 (2012).
140. R. E. Green, J. Krause, A. W. Briggs, T. Maricic, U. Stenzel, M. Kircher, *et al.*, A draft sequence of the Neandertal genome. *Science*. **328**, 710–722 (2010).
141. E. Y. Durand, N. Patterson, D. Reich, M. Slatkin, Testing for ancient admixture between closely related populations. *Mol. Biol. Evol.* **28**, 2239–2252 (2011).
142. J. V. Moreno-Mayar, B. A. Potter, L. Vinner, M. Steinrücken, S. Rasmussen, J. Terhorst, *et al.*, Terminal Pleistocene Alaskan genome reveals first founding population of Native Americans. *Nature*. **553**, 203–207 (2018).
143. H. Ellegren, N. Galtier, Determinants of genetic diversity. *Nat. Rev. Genet.* **17**, 422–433 (2016).
144. J. Lindo, M. Rogers, E. K. Mallott, B. Petzelt, J. Mitchell, D. Archer, *et al.*, Patterns of Genetic Coding Variation in a Native American Population before and after European Contact. *Am. J. Hum. Genet.* **102**, 806–815 (2018).
145. T. S. Korneliussen, I. Moltke, A. Albrechtsen, R. Nielsen, Calculation of Tajima's D and other neutrality test statistics from low depth next-generation sequencing data. *BMC Bioinformatics*. **14** (2013), , doi:10.1186/1471-2105-14-289.
146. G. A. Watterson, On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* **7**, 256–276 (1975).
147. F. Tajima, Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*. **123**, 585–595 (1989).
148. Y. X. Fu, W. H. Li, Statistical tests of neutrality of mutations. *Genetics*. **133** (1993), pp. 693–709.
149. A. Kousathanas, C. Leuenberger, V. Link, C. Sell, J. Burger, D. Wegmann, Inferring Heterozygosity from Ancient and Low Coverage Genomes. *Genetics*. **205**, 317–332 (2017).
150. G. Renaud, K. Hanghøj, T. S. Korneliussen, E. Willerslev, L. Orlando, Joint Estimates of Heterozygosity and Runs of Homozygosity for Modern and Ancient Samples. *Genetics*. **212**,

- 587–614 (2019).
151. H. Ringbauer, J. Novembre, M. Steinrücken, Parental relatedness through time revealed by runs of homozygosity in ancient DNA. *Nat. Commun.* **12**, 5425 (2021).
 152. Q. Fu, M. Hajdinjak, O. T. Moldovan, S. Constantin, S. Mallick, P. Skoglund, *et al.*, An early modern human from Romania with a recent Neanderthal ancestor. *Nature*. **524**, 216–219 (2015).
 153. 1000 Genomes Project Consortium, A. Auton, L. D. Brooks, R. M. Durbin, E. P. Garrison, H. M. Kang, *et al.*, A global reference for human genetic variation. *Nature*. **526**, 68–74 (2015).
 154. P. Skoglund, H. Malmström, A. Omrak, M. Raghavan, C. Valdiosera, T. Günther, *et al.*, Genomic diversity and admixture differs for Stone-Age Scandinavian foragers and farmers. *Science*. **344**, 747–750 (2014).
 155. L. Excoffier, I. Dupanloup, E. Huerta-Sánchez, V. C. Sousa, M. Foll, Robust demographic inference from genomic and SNP data. *PLoS Genet.* **9**, e1003905 (2013).
 156. J. Lindo, R. Haas, C. Hofman, M. Apatá, M. Moraga, R. A. Verdugo, *et al.*, The genetic prehistory of the Andean highlands 7000 years BP through European contact. *Science Advances*. **4** (2018), , doi:10.1126/sciadv.aau4921.
 157. M. Ferrando-Bernal, C. Morcillo-Suarez, T. de-Dios, P. Gelabert, S. Civit, A. Díaz-Carvajal, *et al.*, Mapping co-ancestry connections between the genome of a Medieval individual and modern Europeans. *Sci. Rep.* **10**, 6843 (2020).
 158. J. G. Schraiber, Assessing the Relationship of Ancient and Modern Populations. *Genetics*. **208**, 383–398 (2018).
 159. R. J. Hofmeister, D. M. Ribeiro, S. Rubinacci, O. Delaneau, Accurate rare variant phasing of whole-genome and whole-exome sequencing data in the UK Biobank. *bioRxiv* (2022), p. 2022.10.19.512867.
 160. Medina Tretmanis, Jazeps. Tesis 'Evaluación de métodos de faseo en datos genómicos de ADN antiguo'. UNAM. (2021). (available at https://ru.dgb.unam.mx/handle/DGB_UNAM/TES01000810645).
 161. M. E. Smith, The Aztlan Migrations of the Nahuatl Chronicles: Myth or History? *Ethnohistory*. **31**, 153–186 (1984).
 162. Recinos, Adrián, *Popol Vuh: Las antiguas historias del Quiché* (Fondo de Cultura Económica, 2022).
 163. J. de Acosta, *Historia natural y moral de las Indias: en que se tratan las cosas notables del cielo, elementos, metales, plantas y animales de ellas; y los ritos, ceremonias, leyes, gobierno y guerras de los Indios* (P. Aznar, 1792).
 164. L. Wade, The ghosts in the museum. *Science*. **373**, 148–152 (2021).
 165. S. G. Morton, *Crania americana, or, A comparative view of the skulls of various aboriginal nations of North and South America to which is prefixed an essay on the varieties of the human species* (1839), , doi:10.5962/bhl.title.48501.
 166. M. R. Waters, T. W. Stafford Jr, D. L. Carlson, The age of Clovis-13,050 to 12,750 cal yr B.P. *Sci. Adv.* **6**, eaaz0455 (2020).
 167. E. Willerslev, D. J. Meltzer, Peopling of the Americas as inferred from ancient genomics.

- Nature*. **594**, 356–364 (2021).
168. J. H. Greenberg, C. G. Turner, S. L. Zegura, L. Campbell, J. A. Fox, W. S. Laughlin, *et al.*, The Settlement of the Americas: A Comparison of the Linguistic, Dental, and Genetic Evidence [and Comments and Reply]. *Curr. Anthropol.* **27**, 477–497 (1986).
 169. D. J. Meltzer, D. K. Grayson, G. Ardila, A. W. Barker, D. F. Dincauze, C. Vance Haynes, *et al.*, On the Pleistocene Antiquity of Monte Verde, Southern Chile. *Am. Antiq.* **62**, 659–663 (1997).
 170. A. Torroni, T. G. Schurr, M. F. Cabell, M. D. Brown, J. V. Neel, M. Larsen, *et al.*, Asian affinities and continental radiation of the four founding Native American mtDNAs. *Am. J. Hum. Genet.* **53**, 563–590 (1993).
 171. 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*. **544**, 479–483 (2017).
 172. M. R. Bennett, D. Bustos, J. S. Pigati, K. B. Springer, T. M. Urban, V. T. Holliday, *et al.*, Evidence of humans in North America during the Last Glacial Maximum. *Science*. **373**, 1528–1531 (2021).
 173. 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*. **584**, 87–92 (2020).
 174. 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*. **356**, 605–608 (2017).
 175. M. M. Lahr, Patterns of modern human diversification: Implications for Amerindian origins. *Am. J. Phys. Anthropol.* **38**, 163–198 (1995).
 176. R. González-José, A. González-Martín, M. Hernández, H. M. Pucciarelli, M. Sardi, A. Rosales, *et al.*, Craniometric evidence for Palaeoamerican survival in Baja California. *Nature*. **425**, 62–65 (2003).
 177. W. A. Neves, M. Hubbe, Cranial morphology of early Americans from Lagoa Santa, Brazil: implications for the settlement of the New World. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 18309–18314 (2005).
 178. R. González-José, M. C. Bortolini, F. R. Santos, S. L. Bonatto, The peopling of America: craniofacial shape variation on a continental scale and its interpretation from an interdisciplinary view. *Am. J. Phys. Anthropol.* **137**, 175–187 (2008).
 179. V. V. Pitulko, P. A. Nikolsky, E. Y. Giryay, A. E. Basilyan, V. E. Tumskoy, S. A. Koulakov, *et al.*, The Yana RHS site: humans in the Arctic before the last glacial maximum. *Science*. **303**, 52–56 (2004).
 180. V. Pitulko, E. Pavlova, P. Nikolskiy, Revising the archaeological record of the Upper Pleistocene Arctic Siberia: Human dispersal and adaptations in MIS 3 and 2. *Quat. Sci. Rev.* **165**, 127–148 (2017).
 181. M. Raghavan, P. Skoglund, K. E. Graf, M. Metspalu, A. Albrechtsen, I. Moltke, *et al.*, Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. *Nature*. **505**, 87–91 (2014).
 182. M. Lipson, D. Reich, A Working Model of the Deep Relationships of Diverse Modern Human Genetic Lineages Outside of Africa. *Mol. Biol. Evol.* **34**, 889–902 (2017).

183. E. Tamm, T. Kivisild, M. Reidla, M. Metspalu, D. G. Smith, C. J. Mulligan, *et al.*, Beringian standstill and spread of Native American founders. *PLoS One*. **2**, e829 (2007).
184. J. V. Moreno-Mayar, L. Vinner, P. de Barros Damgaard, C. de la Fuente, J. Chan, J. P. Spence, *et al.*, Early human dispersals within the Americas. *Science*. **362** (2018), doi:10.1126/science.aav2621.
185. H. García-Ortiz, F. Barajas-Olmos, C. Contreras-Cubas, M. Á. Cid-Soto, E. J. Córdova, F. Centeno-Cruz, *et al.*, The genomic landscape of Mexican Indigenous populations brings insights into the peopling of the Americas. *Nat. Commun.* **12**, 5942 (2021).
186. 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*. **506**, 225–229 (2014).
187. C. Posth, N. Nakatsuka, I. Lazaridis, P. Skoglund, S. Mallick, T. C. Lamnidis, *et al.*, Reconstructing the Deep Population History of Central and South America. *Cell*. **175**, 1185–1197.e22 (2018).
188. A. T. Duggan, A. J. T. Harris, S. Marciniak, I. Marshall, M. Kuch, A. Kitchen, *et al.*, Genetic Discontinuity between the Maritime Archaic and Beothuk Populations in Newfoundland, Canada. *Curr. Biol.* **27**, 3149–3156.e11 (2017).
189. P. Flegontov, P. Changmai, A. Zidkova, M. D. Logacheva, N. E. Altinisik, O. Flegontova, *et al.*, Genomic study of the Ket: a Paleo-Eskimo-related ethnic group with significant ancient North Eurasian ancestry, , doi:10.1101/024554.
190. C. Lalueza-Fox, F. L. Calderón, F. Calafell, B. Morera, J. Bertranpetit, MtDNA from extinct Tainos and the peopling of the Caribbean. *Ann. Hum. Genet.* **65**, 137–151 (2001).
191. D. M. Fernandes, K. A. Sirak, H. Ringbauer, J. Sedig, N. Rohland, O. Cheronet, *et al.*, A genetic history of the pre-contact Caribbean. *Nature*. **590**, 103–110 (2021).
192. K. Nägele, C. Posth, M. Iraeta Orbegozo, Y. Chinique de Armas, S. T. Hernández Godoy, U. M. González Herrera, *et al.*, Genomic insights into the early peopling of the Caribbean. *Science*. **369**, 456–460 (2020).
193. P. Skoglund, S. Mallick, M. C. Bortolini, N. Chennagiri, T. Hünemeier, M. L. Petzl-Erler, *et al.*, Genetic evidence for two founding populations of the Americas. *Nature*. **525**, 104–108 (2015).
194. J. V. Moreno-Mayar, S. Rasmussen, A. Seguin-Orlando, M. Rasmussen, M. Liang, S. T. Flâm, *et al.*, Genome-wide ancestry patterns in Rapanui suggest pre-European admixture with Native Americans. *Curr. Biol.* **24**, 2518–2525 (2014).
195. A.-S. Malaspinas, O. Lao, H. Schroeder, M. Rasmussen, M. Raghavan, I. Moltke, *et al.*, Two ancient human genomes reveal Polynesian ancestry among the indigenous Botocudos of Brazil. *Curr. Biol.* **24**, R1035–7 (2014).
196. D. I. C. Dávalos, Y. O. A. Cárdenas, M. J. Bravo-Lopez, S. Neuenschwander, S. Reis, M. Q. R. Bastos, *et al.*, Indigenous peoples in eastern Brazil: insights from 19th century genomes and metagenomes. *bioRxiv* (2022), p. 2022.01.27.477466.
197. L. Fehren-Schmitz, C. L. Jarman, K. M. Harkins, M. Kayser, B. N. Popp, P. Skoglund, Genetic Ancestry of Rapanui before and after European Contact. *Curr. Biol.* **27**, 3209–3215.e6 (2017).
198. R. Acuna-Soto, Megadrought and Megadeath in 16th Century Mexico. *Emerging Infectious*

- Diseases*. **8** (2002), pp. 360–362.
199. A. Moreno, K. Sandoval, Diversidad genómica en México: Pasado indígena y mestizaje. *Cuicuilco* (2013) (available at https://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S0185-16592013000300013).
 200. F. J. Álvarez García, Carnaval y música en Salamanca en la primera década del s. XX a través de la prensa local. *Futuro Pasado*. **7**, 459–476 (2016).
 201. A. Moreno-Estrada, C. R. Gignoux, J. C. Fernández-López, F. Zakharia, M. Sikora, A. V. Contreras, *et al.*, Human genetics. The genetics of Mexico recapitulates Native American substructure and affects biomedical traits. *Science*. **344**, 1280–1285 (2014).
 202. S. Romero-Hidalgo, A. Ochoa-Leyva, A. Garcíarrubio, V. Acuña-Alonzo, E. Antúnez-Argüelles, M. Balcazar-Quintero, *et al.*, Demographic history and biologically relevant genetic variation of Native Mexicans inferred from whole-genome sequencing. *Nat. Commun.* **8**, 1005 (2017).
 203. A. Gorostiza, V. Acunha-Alonzo, L. Regalado-Liu, S. Tirado, J. Granados, D. Sámano, *et al.*, Reconstructing the history of Mesoamerican populations through the study of the mitochondrial DNA control region. *PLoS One*. **7**, e44666 (2012).
 204. M. C. Solanes, E. Vela, Atlas del México prehispánico. *Arqueol. Mex.*
 205. C. de la Fuente, M. C. Ávila-Arcos, J. Galimany, M. L. Carpenter, J. R. Homburger, A. Blanco, *et al.*, Genomic insights into the origin and diversification of late maritime hunter-gatherers from the Chilean Patagonia. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E4006–E4012 (2018).
 206. A. L. Austin, L. L. Luján, *El pasado indígena* (Fondo de Cultura Económica, 2019).
 207. P. Kirchhoff, Mesoamerica: its geographic limits, ethnic composition and cultural characteristics. *Heritage of conquest*.
 208. A. J. Herrera Muñoz, E. Mejía Campos, Un minero en la Sierra Gorda: caso de contaminación ocupacional multielemental de metales pesados a finales del periodo Clásico. *Arqueología (México, D.F.)*, 127–135 (2017).
 209. E. M. P. Campos, J. M. M. P. Campos, A. H. Muñoz, Mercurialism Determination in Fetuses Bone Remains from Toluquilla, Queretaro, Mexico (available at https://www.academia.edu/download/58788921/Mejia_et_al-822015JSRR18043.pdf).
 210. E. M. Pérez Campos, A. J. Herrera Muñoz, MINAS Y MINEROS: PRESENCIA DE METALES EN SEDIMENTOS Y RESTOS HUMANOS AL SUR DE LA SIERRA GORDA DE QUERÉTARO EN MÉXICO. *Chungará (Arica)*. **45**, 161–176 (2013).
 211. E. Nalda, La arqueología de Guanajuato: trabajos recientes. *Arqueología mexicana*. **16**, 36–43 (2008).
 212. J. M. Argüelles, A. Fuentes, B. Yáñez, Analyzing asymmetries and praxis in aDNA research: A bioanthropological critique. *Am. Anthropol.* **124**, 130–140 (2022).
 213. J. Bardill, A. C. Bader, N. A. Garrison, D. A. Bolnick, J. A. Raff, A. Walker, *et al.*, Advancing the ethics of paleogenomics. *Science*. **360**, 384–385 (2018).
 214. S. Alpaslan-Roodenberg, D. Anthony, H. Babiker, E. Bánffy, T. Booth, P. Capone, *et al.*, Ethics of DNA research on human remains: five globally applicable guidelines. *Nature*. **599**, 41–46 (2021).

215. G. Lewis-Kraus, Is ancient DNA research revealing new truths—or falling into old traps. *NY Times* .
216. K. Fox, J. Hawks, Use ancient remains more wisely. *Nature*. **572**, 581–583 (2019).
217. K. S. Tsosie, R. L. Begay, K. Fox, N. A. Garrison, Generations of genomes: advances in paleogenomics technology and engagement for Indigenous people of the Americas. *Curr. Opin. Genet. Dev.* **62**, 91–96 (2020).
218. A. D. Cortez, D. Lippert, J. L. Davis, G. Nicholas, R. S. Malhi, L. S. Weyrich, *et al.*, Extracting the practices of paleogenomics: A study of ancient DNA labs and research in relation to Native Americans and Indigenous peoples. *Am J Biol Anthropol* (2023), doi:10.1002/ajpa.24714.
219. E. Kowal, L. S. Weyrich, J. M. Argüelles, A. C. Bader, C. Colwell, A. D. Cortez, *et al.*, Community partnerships are fundamental to ethical ancient DNA research. *Human Genetics and Genomics Advances*, 100161 (2023).
220. J. Lindo, A. Achilli, U. A. Perego, D. Archer, C. Valdiosera, B. Petzelt, *et al.*, Ancient individuals from the North American Northwest Coast reveal 10,000 years of regional genetic continuity. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 4093–4098 (2017).

12 Anexos

Anexo I. Artículo de investigación como requisito para obtención de grado ‘Demographic history and genetic structure in pre-Hispanic Central Mexico’



RESEARCH ARTICLE SUMMARY

HUMAN GENETICS

Demographic history and genetic structure in pre-Hispanic Central Mexico

Viridiana Villa-Islas, Alan Izarraras-Gomez, Maximilian Larena, Elizabeth Mejía Perez Campos, Marcela Sandoval-Velasco, Juan Esteban Rodríguez-Rodríguez, Miriam Bravo-Lopez, Barbara Moguel, Rosa Fregel, Ernesto Garfias-Morales, Jazeps Medina Tretmanis, David Alberto Velázquez-Ramírez, Alberto Herrera-Muñoz, Karla Sandoval, María A. Nieves-Colón, Gabriela Zepeda García Moreno, Fernando A. Villanea, Eugenia Fernández Villanueva Medina, Ramiro Aguayo-Haro, Cristina Valdiosera, Alexander G. Ioannidis, Andrés Moreno-Estrada, Flora Jay, Emilia Huerta-Sanchez, J. Víctor Moreno-Mayar, Federico Sánchez-Quinto, María C. Ávila-Arcos*

INTRODUCTION: Between 1100 and 900 years ago, a global warming episode affected numerous civilizations worldwide. In the Americas, severe droughts reconfigured the demography of pre-Hispanic civilizations as well as the ecological landscape. We report ancient genomic data from 12 pre-Hispanic individuals from before and after this climate change episode to investigate the population dynamics at the limit between the two biocultural regions of Aridoamerica in the north and Mesoamerica in central and south Mexico.

RATIONALE: The archaeological evidence indicates that the droughts shifted the border southward between desertic Aridoamerica and the verdant and culturally rich Mesoamerica, which was home to large civilizations like the Aztecs and Mayans. This climate change allegedly led to a population replacement in the northern frontier of Mesoamerica by seminomadic hunter-gatherers from Aridoamerica. However, this hypothesis relies solely on archaeological data, so we generated

ancient genomic and mitochondrial DNA data to test it. In addition, we leveraged these data to describe the ancient population structure of Mexico and to investigate the contribution of unsampled genetic lineages to the ancient genomes.

RESULTS: The pre-Hispanic population structure closely resembles the structure observed in present-day Mexico, which clearly differentiates northern and central Indigenous populations. This reflects an overall conservation of the genetic structure of the populations inhabiting the Mexican territory for at least 1400 years (which is the date of the most ancient individual in our study).

We found genetic continuity in the ancient individuals from before and after the climate change episode. This contradicts the hypothesis of population replacement by Aridoamerican groups in this region and suggests that the local population stayed in their homeland despite the long-standing droughts. The population continuity at the studied site can be

explained by the favorable location in the and humid Sierra Gorda mountain range the fact that cinnabar mining was the main economic activity, not agriculture.

We identified the contribution of two distinct unsampled “ghost” genetic ancestries to pre-Hispanic populations of northern and central Mexico, respectively. Whereas the unsampled genetic ancestry contributing to the northern genome coincides with one previously identified in a present-day population from south Mexico, the second ghost genetic ancestry was previously unknown.

CONCLUSION: The ancient genomes revealed a conservation of the genetic structure in Mexico in the past 1400 years and population continuity in the northern frontier of Mesoamerica despite the severe droughts 1100 years ago. It is likely that the mining-based economy allowed the population to subsist in their homeland during this climate change period when the border between Aridoamerica and Mesoamerica shifted southward. The identification of a new ghost genetic ancestry (UPopA2), along with that observed in ancient Sierra Tarahumara and present-day Mixe (UPopA), reveals a complex population history in the late Pleistocene in the Americas. The recovery and study of ancient genomes from Mexico, conducted ethically, can help fill important gaps in our understanding of the deep population history of the Americas. ■

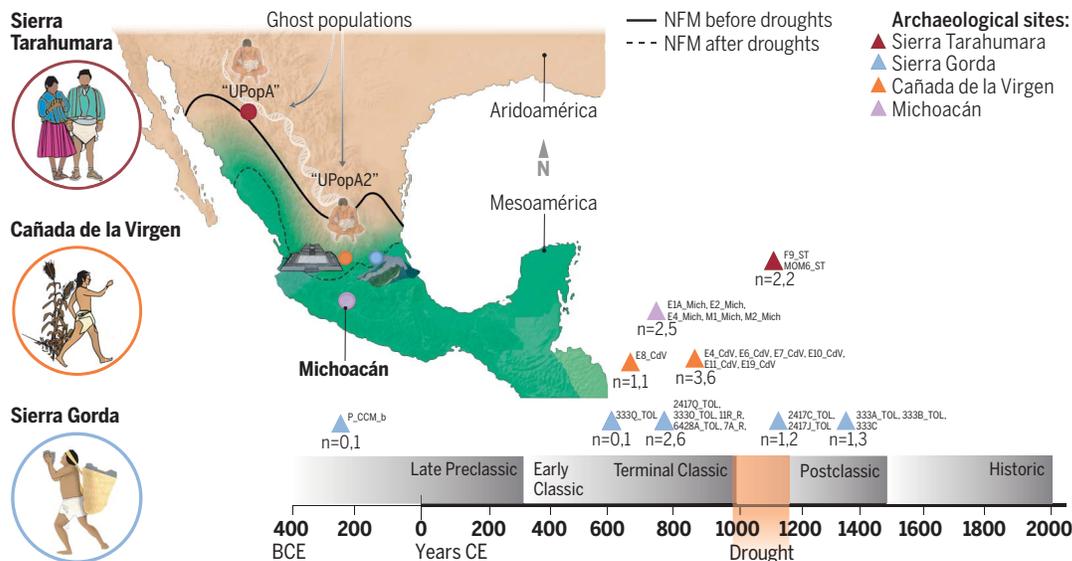
The list of author affiliations is available in the full article online.

*Corresponding author. Email: mavila@ligh.unam.mx
Cite this article as V. Villa-Islas et al., *Science* 380, eadd6142 (2023). DOI: 10.1126/science.add6142

READ THE FULL ARTICLE AT
<https://doi.org/10.1126/science.add6142>

Population continuity and ghost genetic ancestries in pre-Hispanic Mexico.

Ancient individuals who inhabited the northern frontier of Mesoamerica (NFM) before and after a 200-year period of severe droughts (shown in timeline) belong to a continuous population. This is in contrast to a previous hypothesis suggesting that hunter-gatherers from Aridoamerica replaced the populations at the NFM following the southward shift of its limits with Mesoamerica (solid and dashed lines). Individuals from Sierra Tarahumara and Cañada de la Virgen show genetic ancestry from two distinct ghost populations (UPopA and UPopA2).



RESEARCH ARTICLE

HUMAN GENETICS

Demographic history and genetic structure in pre-Hispanic Central Mexico

Viridiana Villa-Islas¹, Alan Izarraras-Gomez¹, Maximilian Larena², Elizabeth Mejía Perez Campos³, Marcela Sandoval-Velasco^{4,5,6}, Juan Esteban Rodríguez-Rodríguez¹, Miriam Bravo-Lopez¹, Barbara Moguel^{1,7}, Rosa Fregel⁸, Ernesto Garfias-Morales¹, Jazeps Medina Tretmanis⁹, David Alberto Velázquez-Ramírez¹⁰, Alberto Herrera-Muñoz³, Karla Sandoval¹¹, Maria A. Nieves-Colón^{12,13}, Gabriela Zepeda García Moreno¹⁴, Fernando A. Villanea¹⁵, Eugenia Fernández Villanueva Medina¹⁶, Ramiro Aguayo-Haro¹⁶, Cristina Valdiosera^{17,18}, Alexander G. Ioannidis¹⁹, Andrés Moreno-Estrada¹², Flora Jay²⁰, Emilia Huerta-Sanchez⁹, J. Víctor Moreno-Mayar²¹, Federico Sánchez-Quinto²², María C. Ávila-Arcos^{1*}

Aridoamerica and Mesoamerica are two distinct cultural areas in northern and central Mexico, respectively, that hosted numerous pre-Hispanic civilizations between 2500 BCE and 1521 CE. The division between these regions shifted southward because of severe droughts ~1100 years ago, which allegedly drove a population replacement in central Mexico by Aridoamerican peoples. In this study, we present shotgun genome-wide data from 12 individuals and 27 mitochondrial genomes from eight pre-Hispanic archaeological sites across Mexico, including two at the shifting border of Aridoamerica and Mesoamerica. We find population continuity that spans the climate change episode and a broad preservation of the genetic structure across present-day Mexico for the past 2300 years. Lastly, we identify a contribution to pre-Hispanic populations of northern and central Mexico from two ancient unsampled “ghost” populations.

Before European colonization, the territory that is present-day Mexico was home to numerous civilizations that occupied two main cultural areas: Aridoamerica in northern Mexico, inhabited mainly by hunter-gatherers, and Mesoamerica in central and southern Mexico (Fig. 1), where some of the largest pre-Hispanic agriculture-based civilizations in the Americas flourished between 2500 BCE and 1521 CE (1, 2). The distinction between Aridoamerica and Mesoamerica is based on the cultural characteristics and subsistence strategies of the peoples that inhabited them, as well as the ecological features of each region (3, 4). Archaeological evidence indicates that the border between these two areas shifted southward between 900–1300 CE after multidecadal droughts (5). This period is also known as the Medieval Warm Period in other regions of the world (6). The droughts allegedly led to population replacements in the northern frontier of Mesoamerica by semino-

madic hunter-gatherers (“Chichimecas”) from Aridoamerica (7) and precipitated the fall of some pre-Hispanic societies and the abandonment of Mesoamerican cities in central and southeast Mexico (5, 8–10).

Evidence for the population replacement at the northern frontier of Mesoamerica comes solely from the archaeological record. Whether this change was the product of migration or acculturation has been debated by archaeologists for years (8, 11–13). Studying the genetic variation of ancient populations that span this period of climate change across these two regions is thus necessary to illuminate the regional population dynamics in response to this drastic environmental change. However, ancient genomic data for pre-Hispanic populations from Mexico is very limited, with only two studies reporting 11 low-depth genomes (<0.3×) for a few individuals restricted to northern Mexico (14, 15), and no available genomes from central and southern Mexico.

In this study, we report the most extensive set of complete mitochondrial genomes (mitogenomes) and shotgun genome-wide data to date from pre-Hispanic individuals from Mexico. We analyze our data jointly with publicly available datasets of other ancient Native Americans and present-day Indigenous populations from Mexico. This large compendium of ancient genomic data allows us to study the pre-Hispanic genetic structure in the territory occupied by Mexico, helping to answer long-standing questions regarding population dynamics at the northern frontier of Mesoamerica, disentangling the complex structure of central Mexico populations, and revealing a previously unknown ancient contribution of unsampled genetic lineages to Mexican populations.

Results

Sampling and pre-Hispanic genomic data generation

We screened archaeological samples from 37 pre-Hispanic individuals excavated at seven sites from within Mesoamerica in central Mexico. Three are in the Sierra Gorda in Querétaro state (Toluquilla and Ranas sites and a cave in Cadereyta de Montes), three in Michoacán state (Zaragoza, Tanhuato, and La Mina sites), and one in Guanajuato state (Cañada de la Virgen site) (Fig. 1). Furthermore, we produced additional shotgun data for two mummies from the Sierra Tarahumara from Aridoamerica, northern Mexico, who were sequenced at lower depths in a previous study (14) (Fig. 1). The sampling of the archeological human remains was made after approval by the Archaeology Council of the Instituto Nacional de Antropología e Historia with permit numbers 401.3S.16-2017/990 and 401.3S.16-2019/222. We sampled the minimum possible amount of tissue to avoid unnecessary destruction. For teeth, we tried to separate the root without damaging the crown and returned all leftover material to the archaeologists responsible for each collection. Although consultation with Indigenous populations for destructive analysis of human archaeological material is not a requirement by law in Mexico and the studied sites are not located within Indigenous communities, we acknowledge the need to involve Indigenous perspectives in delineating the

¹International Laboratory for Human Genome Research, Universidad Nacional Autónoma de México (UNAM), Querétaro, México. ²Department of Organismal Biology, Uppsala University, Uppsala, Sweden. ³National Institute of Anthropology and History, Querétaro, México. ⁴Section for Evolutionary Genomics, GLOBE Institute, University of Copenhagen, Copenhagen, Denmark. ⁵Department of Anthropology, National Museum of Natural History, Smithsonian Institution, Washington, DC 20560, USA. ⁶Departamento de Ecología Evolutiva, Instituto de Ecología, Universidad Nacional Autónoma de México, Ciudad de México, México. ⁷Centro de Geociencias, UNAM Juriquilla, Juriquilla, Querétaro, México. ⁸Department of Biochemistry, Microbiology, Cell Biology and Genetics, Universidad de La Laguna, San Cristóbal de La Laguna, Spain. ⁹Center for Computational Molecular Biology, Brown University, Providence, RI, USA. ¹⁰Institute for Medical Microbiology and Hospital Epidemiology, Hannover Medical School, 30625 Hannover, Germany. ¹¹Equity and Gender Office of the Centre for Research and Advanced Studies (CODIGO-C), CINVESTAV, Mexico City, México. ¹²Unit of Advanced Genomics, National Laboratory of Genomics for Biodiversity (LANGEBO), CINVESTAV, Irapuato, Guanajuato, México. ¹³Department of Anthropology, University of Minnesota Twin Cities, Minneapolis, MN, USA. ¹⁴National Institute of Anthropology and History, Guanajuato, México. ¹⁵Department of Anthropology, University of Colorado Boulder, Boulder, CO, USA. ¹⁶National Institute of Anthropology and History, Michoacán, México. ¹⁷Departamento de Historia, Geografía y Comunicaciones, Universidad de Burgos, Burgos, Spain. ¹⁸Department of History and Archaeology, La Trobe University, Melbourne, Australia. ¹⁹Institute for Computational and Mathematical Engineering, Stanford University, Stanford, CA, USA. ²⁰Laboratoire Interdisciplinaire des Sciences du Numérique, Université Paris-Saclay, CNRS, INRIA, 91400 Orsay, France. ²¹Lundbeck Foundation GeoGenetics Centre, GLOBE Institute, University of Copenhagen, Copenhagen, Denmark. ²²Computational Genomics, Instituto Nacional de Medicina Genómica, Ciudad de México, México.

*Corresponding author. Email: mavila@igh.unam.mx

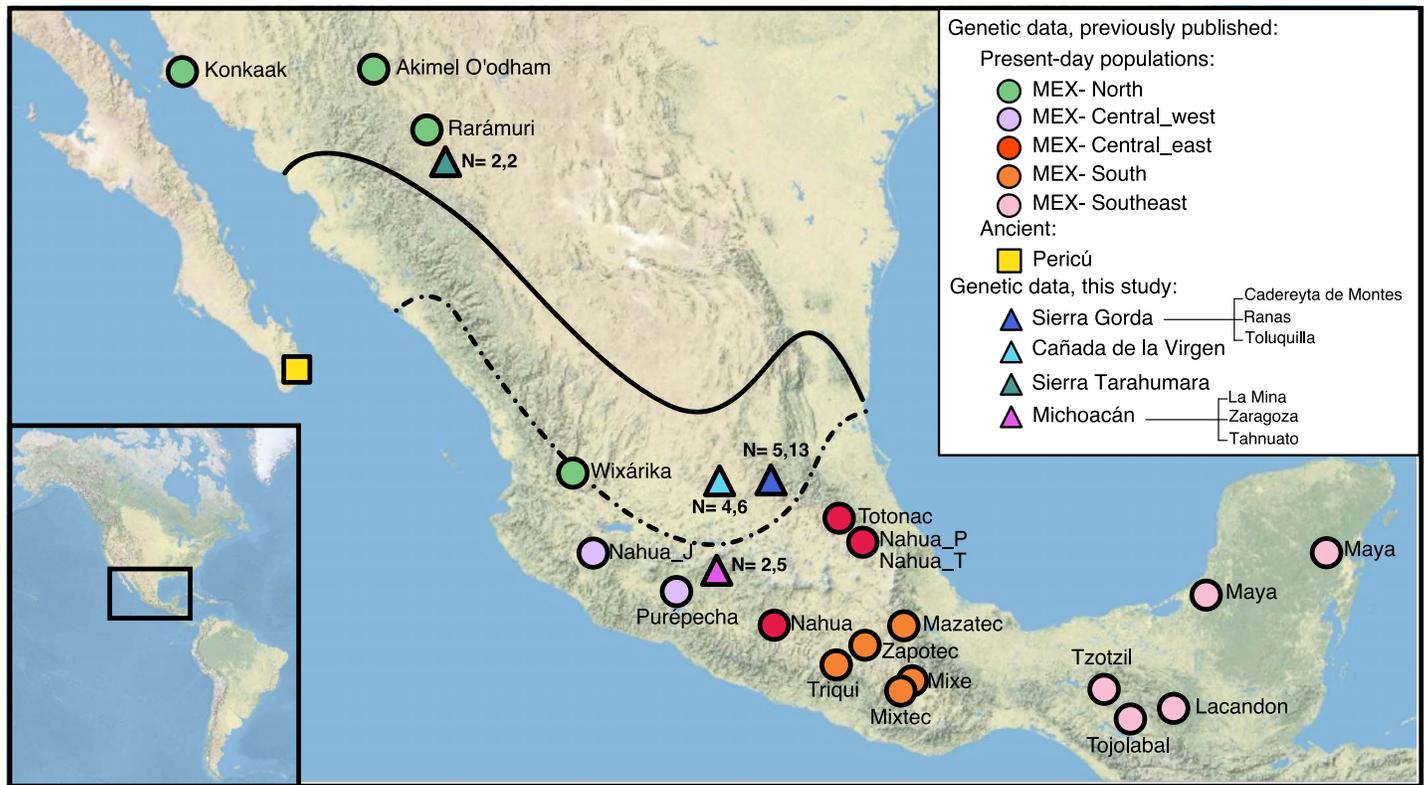


Fig. 1. Map of archeological sites and modern populations. Map of Mexico indicating the location of the pre-Hispanic sites analyzed in this study (triangles) and in previous studies (squares) (14, 15), and the approximate locations of present-day Indigenous populations sampled in previous studies (circles) (14, 18, 22, 36). The continuous line indicates the northern frontier of Mesoamerica, with Aridoamerica to the north and Mesoamerica to the south during the years ~800–900 CE. The dashed line indicates this border after the year ~1500 CE, based on (2). Numbers next to archaeological sites from this study indicate the number of whole-genome data and mitochondrial genomes recovered per site (separated by a comma).

regulations for these types of studies [see (16) for a deeper discussion on the subject]. To engage the neighboring communities, we delivered public lectures to communicate the genetic results to the village closest to the sites in Sierra Gorda.

After processing and sequencing samples, we obtained informative amounts of shotgun sequence data (0.01 to 4.7× genomes) for 12 individuals spanning a time transect of 750 years (600–1351 CE). These include 10 low-depth genomes from within Mesoamerica in central Mexico, and two from Aridoamerica, in the Sierra Tarahumara (table S1). In addition, we reconstructed the mitochondrial genomes for 27 individuals across these sites (5.7 to 1284.8x), spanning a time transect of 1600 years (320 BCE to 1351 CE). (tables S1 and S2). We included in the dataset three previously published ancient genomes (table S3) as well as nine ancient mitogenomes from northern Mexico (table S4), yielding a total of 15 ancient genomes and 36 ancient mitogenomes from the Mexican territory.

We named each sample according to burial and individual identifiers provided by archaeologists and added two additional labels. The first label includes one to three letters that refer to the archaeological site, whereas the

second refers to whether the individual dates to before (“b”) or after (“a”) the drought period. The three pieces of information are divided by an underscore character. Throughout the text we refer to all individuals using these identifiers.

Genetic structure and diversity in uniparental markers

All chromosomal sex assignments matched the morphological sex except one (XX chromosomally and male on the basis of morphology) (table S1) (17). It was possible to assign Y-DNA haplogroups to 5 out of the 13 individuals assigned as XY, all of which had a Native American Q lineage. This is in agreement with previous studies on ancient and present-day Indigenous Mexican individuals (18–23). At the subhaplogroup level, we found Q1a2a1-L54 and Q1a2a1a1a1-M3, with no apparent differences between Aridoamerica and Mesoamerica (table S1).

All 27 reconstructed mitogenomes carried one of the mitochondrial DNA (mtDNA) haplogroups found in Indigenous populations of the Americas: A ($n = 10$), B ($n = 9$), C ($n = 4$), and D ($n = 4$) (24) (tables S1 and S4) (17). After merging with nine additional pre-Hispanic mitogenomes from Mexico from other studies

(14, 15, 25), we found all four haplogroups in Mesoamerica ($n = 25$) and only C and B in Aridoamerica ($n = 11$) (table S4). Consistently, we observed higher nucleotide diversity (π) values in Mesoamerican populations compared with Aridoamerica (17).

The pre-Hispanic mtDNA haplogroup distribution resembles the one in present-day individuals (23, 25, 26), showing an overall continuity of the matrilineal genetic structure for at least 2300 years. Of the available complete ancient mtDNA genomes from Aridoamerica, 36% harbor haplogroup C, which decreases in frequency to 8% in Mesoamerica, whereas the opposite occurs for haplogroup A, with 40% of presence in Mesoamerica and 0% in Aridoamerica (table S4). This gradient has been previously observed in ancient and contemporary populations from Mexico (18, 19, 23, 25–31). Moreover, we detected unreported variants in the mtDNA PhyloTree (32) for seven individuals assigned to the subhaplogroups A2d, B2c, and B2l (table S5) (17).

In Sierra Gorda, we identified subhaplogroup A2d in six individuals—four predrought and two postdrought. After quality filtering (17), five of them were included in a median-joining network (table S4). Four (two pre- and two postdrought) clustered together in a single

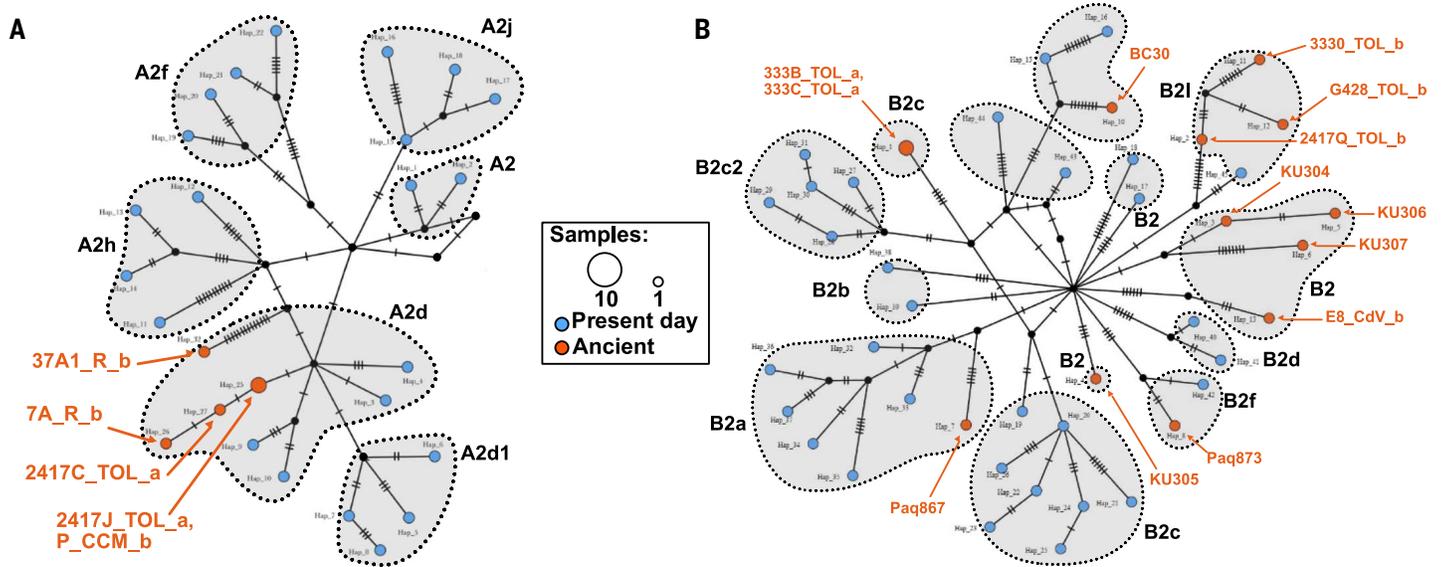


Fig. 2. mtDNA haplotype networks. (A) Haplotype network of mitochondrial subhaplogroups A2 including present-day and pre-Hispanic mitochondrial haplotypes from Mexico. Arrows point to individuals from Sierra Gorda within haplogroup A2d from before and after the drought period. (B) Haplotype network of mitochondrial subhaplogroups of the lineage B2 including present-day and pre-Hispanic mitochondrial haplotypes from Mexico. Arrows point to pre-Hispanic individuals from Cueva Candelaria, Paquimé, Sierra Gorda, and Cañada de la Virgen.

clade despite spanning a time transect of 1480 years, whereas the fifth individual remained in a different clade within A2d (Fig. 2A). Notably, P_CCM_b and 2417J_TOL_a are found in the same node within A2d. This reflects a continuity of this maternal lineage despite the severe droughts in Sierra Gorda. By contrast, the haplotype network for haplogroup B, including five Sierra Gorda individuals, shows that the three predrought individuals share subhaplogroup B2I, whereas the two postdrought individuals are related and found in a node representing subhaplogroup B2c. Because mtDNA is subject to genetic drift that leads to lineage loss, it cannot accurately inform about genetic structure or admixture events. The discrepant patterns between haplogroups A and B could be explained by a number of scenarios, which are difficult to test because of the small sample size. This highlights the need to explore autosomal data to test the population-replacement hypothesis reliably.

We further leveraged the mtDNA data to estimate the past female effective population size (N_e) using Extended Bayesian Skyline Plots (EBS) in pre-Hispanic ($n = 29$) and present-day Mexico ($n = 232$) (tables S4 and S6), per haplogroup and all haplogroups together (fig. S6, A to H). All EBS runs (table S7) showed wide (95%) confidence intervals (CIs) except the EBS of only present-day Mexico, in which we observed a population expansion ~5 kyr B.P. (thousand years before present) and a population decline ~500 yr B.P., though the median values do not support the latter (fig. S6, G and H). Despite the wide CI, there is a clear distinction between the estimated female

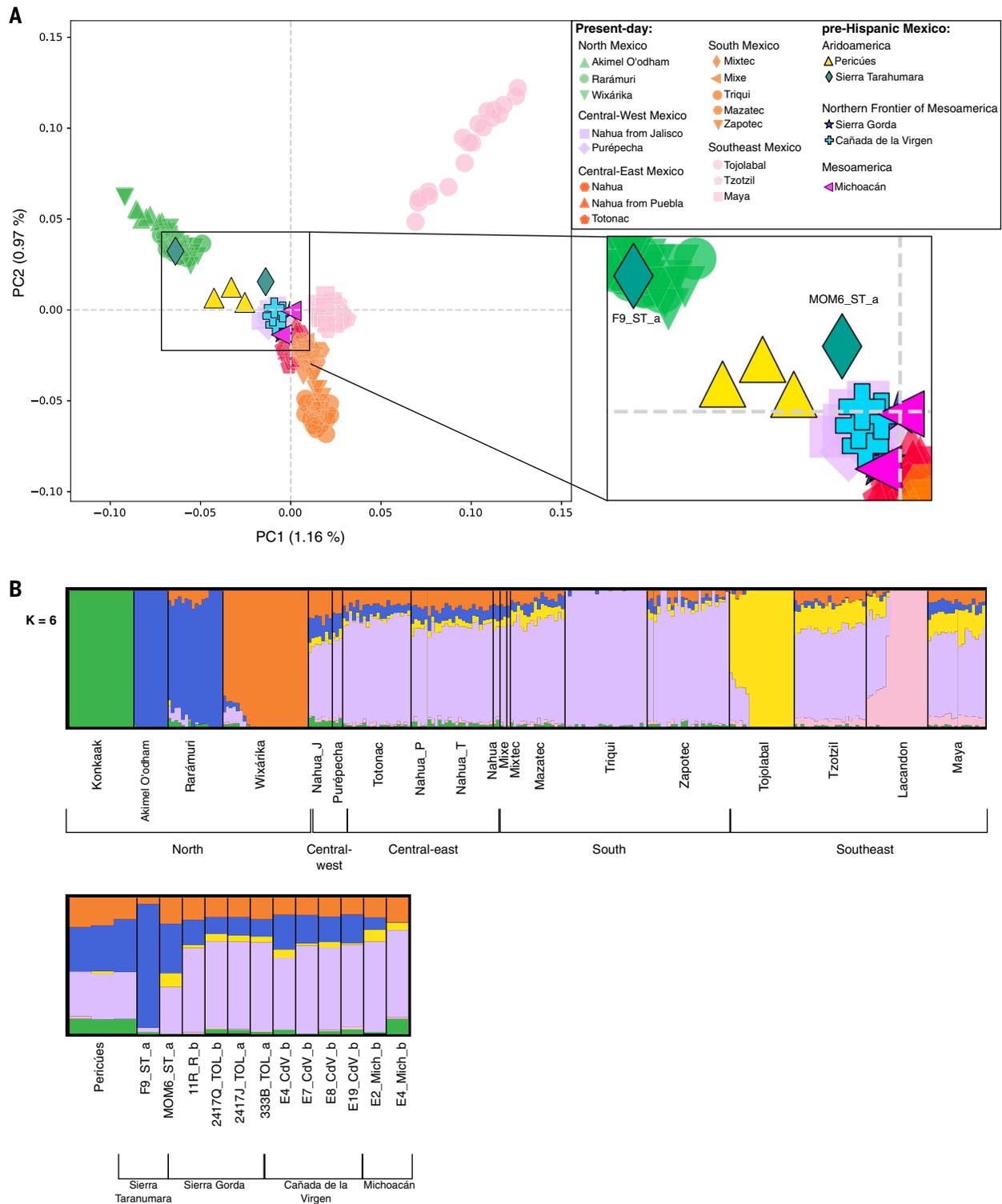
N_e per haplogroup, with haplogroup A reaching the highest present-day values (104,048), followed by D (100,812), and decreasing almost an order of magnitude for haplogroups B and C (37,155 and 7226, respectively). To improve the estimations and gain insights about the past female N_e in a wider geographical range, we merged our dataset with available ancient mtDNA sequences from South America ($n = 137$) (table S6). The EBSs for haplogroups A, B, and D, respectively, also show wide CIs (fig. S6, I, J, and N), whereas haplogroup C shows a clearer pattern of population expansion starting ~15 kyr B.P. and a population decline starting ~5 kyr B.P. (fig. S6K). Consistent with our previous observation, the estimated female N_e is notably higher for haplogroups A and D than for B and C in the merged dataset (fig. S6, I to L).

Autosomal genetic structure in pre-Hispanic Mexico

We performed ADMIXTURE (33) and principal component analysis (PCA) (34, 35) to visualize the genetic relationship and genetic structure at the autosomal level between the pre-Hispanic individuals and other ancient individuals representative of North, Central, and South America (table S3), along with present-day Native American individuals from Mexican and continental populations (table S8) (14, 18, 22, 36, 37). All pre-Hispanic individuals from Mexico, as well as all ancient individuals from California, Belize, and Patagonia, clustered together in PCA with present-day Indigenous populations from Mexico (fig. S7) and shared similar genetic composition in the ADMIXTURE analysis (fig. S8).

We projected the pre-Hispanic individuals onto a principal-components space of only pre-Hispanic and present-day Indigenous populations from Mexico (14, 18, 22, 36) using an approach named “missing DNA” PCA (mdPCA) that corrects for the high missing fraction of low-depth ancient genomes (17). All ancient individuals clustered closely in the mdPCA with present-day populations from the same geographical region except for individual MOM6_ST_a, who is intermediate between northern and central present-day populations. Both pre- and postdrought Sierra Gorda individuals clustered together with present-day Nahua from Jalisco, Purépecha, Totonac, Nahua from Puebla, and Nahua, all of which are from central Mexico (Fig. 3A). This pattern was replicated when using traditional PCA (fig. S9) and a temporal factor analysis approach (38), which corrected the ancestral relationships with the dates of the pre-Hispanic individuals (fig. S10) (17). In agreement, the ADMIXTURE analysis assuming six ancestral components revealed that pre-Hispanic individuals show similar genetic ancestry proportions to those observed in present-day populations from the same region (Fig. 3B) and that individuals from Sierra Gorda show a homogeneous genetic composition independent of the period in which they lived (Fig. 3B).

We then performed several combinations of D- and outgroup-f3 statistics to further explore the genetic structure of ancient and present-day populations (17). Outgroup-f3 statistics were of the form $f_3(\text{Test}, \text{Source1}; \text{YRI})$, where the Test is the pre-Hispanic individual under analysis and Source1 is a present-day Indigenous population. D-statistics were of the form



Downloaded from https://www.science.org at UNAM Campus - Cuernavaca, Morelos on May 25, 2023

Fig. 3. Genetic structure of Indigenous and pre-Hispanic individuals. (A) Missing DNA Principal Component Analysis (mdPCA) of present-day Indigenous Populations with the ancient and pre-Hispanic genomes from Mexico. (B) Unsupervised clustering analysis of same data by ADMIXTURE with $k = 6$, showing green, blue, and orange components in northern populations, yellow and pink components in southeast populations, and a violet component in south and central Mexico populations.

D(Pop1, Pop2; Test, YRI), where Test is the pre-Hispanic individual under analysis, and Pop1 and Pop2 are all possible combinations of present-day Indigenous populations. YRI represents Yoruba as the outgroup population in both tests. We found that F9_ST_a shared

higher genetic drift with present-day northern Mexican populations than with any other present-day population (fig. S11 and table S9) and was significantly more related to the present-day Rarámuri than to any other population when tested in the form D(Pop1,

Rarámuri; F9_ST_a, YRI) (fig. S12 and table S10). However, a population continuity test that considers the read counts in the ancient individual and allele frequencies in the present-day population (39) rejected the null hypothesis of population continuity between F9_ST_a and

present-day Rarámuri, which had a P value of $10^{-499.3}$ (17).

The genetic ancestries of pre-Hispanic individuals from central Mexico are more complex. They do not seem to have a higher shared genetic drift with any present-day Indigenous population according to the f_3 -outgroup values, for which almost all standard errors overlapped (figs. S13 to S15 and table S9). Furthermore, in D-statistics analysis, we found mostly $D = 0$ in all possible combinations of Pop1 and Pop2 (figs. S16 to S18 and table S10), which was expected in populations with extensive gene flow between them and not completely diverged from each other.

Genetic diversity in pre-Hispanic Mexico

We estimated the conditional nucleotide diversity (CND) (40) and runs of homozygosity (ROH) (41) in pre-Hispanic and present-day Indigenous individuals from Mexico (18, 22) to make inferences about past genetic diversity and effective population sizes and investigate changes in patterns of genetic variation that may have arisen after the drought period (17). Both the CND values and the ROH distribution show that pre-Hispanic and present-day individuals from northern Mexico (Pericúes and Akimel O'odham) have the lowest CND values and the highest sum of inferred ROH [>4 centimorgan (cM)] compared with other populations of their respective periods (tables S11 to S13). Pericú hunter-gatherers had the lowest CND values (fig. S19 and table S12), but the ROH proportion revealed a larger population size than another ancient hunter-gatherer population from Patagonia (fig. S19 and table S12) (17). These results agree with previous studies that report lower genetic diversity and longer ROH in hunter-gatherer populations from South America than those observed in other regions of the Americas (42).

Pre-Hispanic individuals from Sierra Tarahumara, Toluquilla, and Cañada de la Virgen show similar values of CND and sum of inferred ROH segments. In Toluquilla, we found that the CND value increases with the date difference between pairs of individuals, which probably reflects the accumulation of new mutations during the 489 to 680 years of difference between pre- (2417Q_TOL_b) and postdrought (2417J_TOL_a and 333B_TOL_a) individuals. Lastly, the segment-size distribution of ROHs suggests that the pre-Hispanic individuals studied here belonged to populations with small N_e ($2N_e = 1600$ to 6400) (fig. S20) (17), which is in agreement with N_e values previously calculated for present-day northern and southern populations (43).

Genetic continuity before and after the 900–1300 CE droughts in the Sierra Gorda

To formally test the hypothesis of population replacement in the Sierra Gorda from the

northern frontier of Mesoamerica during the 900–1300 CE droughts, we applied different combinations of outgroup- f_3 and D-statistics. We used individuals from Sierra Gorda from the pre- (TOL_b and R_b) or postdrought (TOL_a) periods as the Test population in an outgroup- f_3 of the form $f_3(\text{Test}, \text{Source}; \text{YRI})$; for the Source, we used another ancient individual from Mexico or ancient representatives of other regions in the Americas (North, Central, and South America) (table S3).

Under a population replacement scenario, we would expect the Toluquillan individuals who lived after the drought to be more closely related to pre-Hispanic individuals with higher proportions of Aridoamerican genetic ancestry (Pericúes and F9_ST_a) than to Toluquillan individuals who lived before the drought (2417Q_TOL_b). Instead, what we observed is that the pre-Hispanic individuals from Sierra Gorda before and after the climate change episode share higher genetic drift between them than with any other pre-Hispanic individual (Fig. 4A), except for the individual IIR_R_b, who shows higher shared genetic drift with an individual from Cañada de la Virgen (fig. S21 and table S14). However, standard errors for IIR_R_b are high because of its low coverage (fig. S21). In the case of the two individuals from Sierra Tarahumara, the highest value of genetic drift is between them (fig. S22). In contrast, outgroup- f_3 tests with individuals from Cañada de la Virgen (CdV) as the Test population in $f_3(\text{CdV}, \text{Source}; \text{YRI})$ (Fig. S23) have the highest outgroup- f_3 values when a related individual [inferred with a relatedness analysis, (17)] from the same site is used as the Source, but the Source with the second-highest value does not belong to Cañada de la Virgen.

We performed D-statistic tests in the form $D(\text{SG}_b, \text{SG}_a; \text{Pop3}, \text{YRI})$, with the individuals from Sierra Gorda (SG) of each period as the Source—first without collapsing them per period and then merging the individuals in their respective group—and Pop3 being any pre-Hispanic or present-day individual or population from Mexico. SG_b represents pre-drought individuals 2417Q_TOL_b and IIR_R_b, whereas SG_a includes postdrought individuals 2417J_TOL_a and 333B_TOL_a. Notably, 2417Q_TOL_b shared more alleles with California or Pericú individuals than with the postdrought individuals, but this sharing was not observed when only transversions were used (fig. S24 and table S15). When merging the individuals per period, none of the comparison under this model deviates significantly from $D = 0$, meaning that SG_b and SG_a form a clade to the exclusion of other individuals and are more closely related to each other than with any other pre-Hispanic or present-day population. The only exception was when Pericúes were

used as Pop3, where $D > 0$ with a z -score = 3.625, which indicates that the predrought individuals are more closely related to Pericúes (fig. S25 and table S16). However, this significant deviation disappears when repeating the analysis with only transversion sites (Fig. 4B and table S17).

We then used f -statistic-based admixture graphs (qpGraph) (44) to further evaluate the population continuity scenario in Sierra Gorda by testing the fit of models that grouped individuals from before and after the drought period in the same clade. For the model, we used the 11,500-kyr-old USR1 (45) as the surrogate for ancient Beringian, the 10,700-kyr-old Anzick (46) individual as the surrogate for ancestral Southern Native American (SNA) (14), and the 4200-kyr-old Ancient Southwestern Ontario individual (ASO) (15) as the surrogate for Northern Native American (NNA) (14). In the analysis, we included the pre-Hispanic individuals with the highest coverage: 2417Q_TOL_b, 2417J_TOL_a, and 333B_TOL_a from Toluquilla, and F9_ST_a as the surrogate for Aridoamerica. The three surrogates can be modeled in the same clade separated from F9_ST_a, with no contribution to this (Fig. 4C).

Finally, we used qpWave (44) to assess whether pre- and postdrought Sierra Gorda individuals can share a demographic history without the need of an additional wave of genetic ancestry from Aridoamerica or another population. Indeed, we found that pre- and postdrought Sierra Gorda individuals can be modeled in the same clade without the need of an admixture event ($P > 0.12$). By contrast, an additional genetic source is needed to reflect genetic variation in F9_ST_a, Pericúes, Akimel O'odham, and Maya (table S18).

Admixture model for Central Mexico populations

Given the complex relationships observed for ancient individuals in central Mexico, we tested additional demographic models using qpGraph. The base model that fit the pre-Hispanic populations includes the Anzick (46) individual as the ancestral SNA (14), present-day Athabascan (14) as the representative for NNA (14), and present-day Indigenous populations in Mexico (36) that represent the northern (Konkaak), central (Nahua from Puebla), and southern (Triqui) regions. Furthermore, this model involves a split of the SNA into S2A and S2B sources. The pre-Hispanic populations from Sierra Gorda and Cañada de la Virgen show different levels of genetic ancestry from SNA and NNA. We found that the population from Sierra Gorda had a higher percentage of the SNA branch from S2B (64%) than the population from Cañada de la Virgen (37%), which had 62% genetic ancestry shared with the NNA branch, indicating a different demographic history for each of these populations

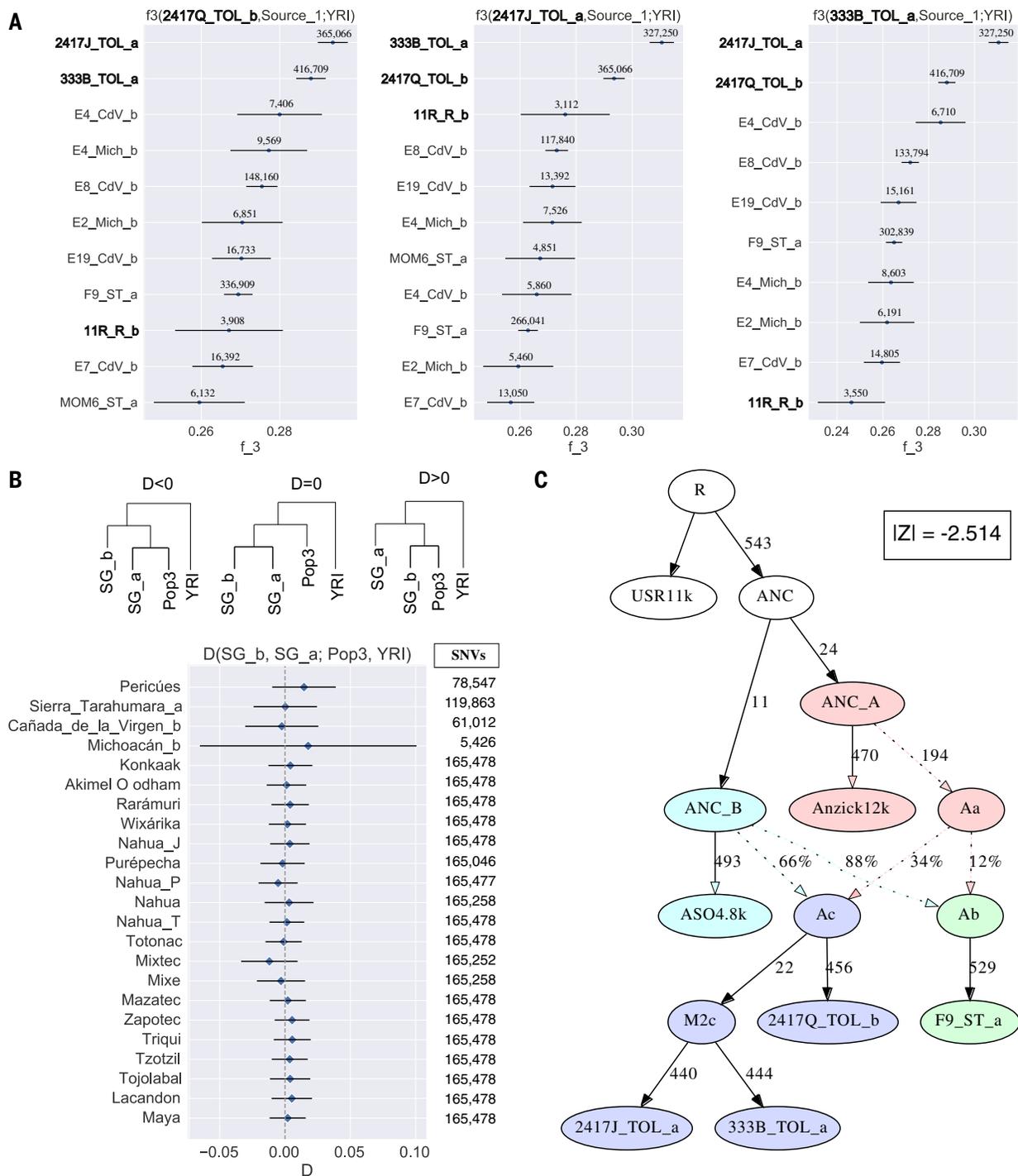


Fig. 4. Population continuity in Sierra Gorda. (A) Shared genetic drift and genetic affinities of pre-Hispanic individuals from Toluquilla, 2417Q_TOL_b, 2417J_TOL_a, and 333B_TOL_a, compared with pre-Hispanic individuals from Mexico. Higher values of f_3 indicate higher shared genetic drift. Point estimates and one standard error are shown. (B) D statistics in the form $D(SG_b, SG_a; Pop3, YRI)$, including transversions only. SG_b includes 2417Q_TOL_b and 11R_R_b, and SG_a includes 2417J_TOL_b and 333B_TOL_a. Pop3 is any of the other pre-Hispanic or

despite their relative geographic proximity (fig. S26). This model supports previous studies that report several admixture events between the two branches that have given rise to Central and South American populations

(15). For the Michoacán population, we obtained terminal branches with no support (zeros); thus, we did not make conclusions with this model for the Michoacán pre-Hispanic population.

present-day populations from Mexico and is shown in the y axis. YRI as an outgroup corresponds to Yoruba from Africa from the 1000 Genomes Project (37, 95). Expected tree topologies according to different D values are drawn on the top of the plot. Blue dots indicate no significant deviations from $D = 0$ (z -score > 3). The number of SNVs used in each test is shown in the right. (C) qpGraph model of individuals from Sierra Gorda from before and after the drought is shown in the same clade, without contribution of F9_ST_a to the postdrought individuals.

Ghost population contribution to pre-Hispanic Mexico

Earlier studies report the contribution of “ghost” genetic ancestry from an unsampled group, designated as UpopA, among the present-day

Mixe from Mexico (47). We tested the presence of this UpopA in the pre-Hispanic individuals using combinations of admixture graph models. Notably, Sierra Tarahumara (represented by F9_ST_a) and Cañada de la Virgen (represented by E8_CdV_b) showed a genetic ancestry contribution from a ghost population at 28% and 17%, respectively (fig. S27, A and B). This was consistent with the ghost genetic ancestry, UpopA, that was previously reported in Mixe (fig. S27C), whereas models including individuals from the other archaeological sites were rejected. When we modeled the pre-Hispanic individuals and Mixe together, we found that Sierra Tarahumara and Mixe share the same ghost genetic ancestry, UPopA (with a better *z*-score when Sierra Tarahumara received NA1 genetic ancestry) (fig. S28), whereas Cañada de la Virgen required the contribution of a second ghost genetic ancestry from that found in Mixe (models without this second ghost genetic ancestry are rejected, *z*-score > 3), which we named UPopA2 (fig. S29). Finally, we included Sierra Tarahumara, Cañada de la Virgen, Mixe, and the two ghost populations in the same model and confirmed that F9_ST_a and Mixe share the contribution from UpopA1, whereas Cañada de la Virgen requires the contribution from an additional unsampled group, UpopA2 (Fig. 5) (18).

Discussion

We have generated an ancient genome dataset from Mexico to address long-standing questions about population dynamics between Aridoamerica and Mesoamerica. The data allowed us to (i) describe levels of population structure and genetic diversity in Mexico before European colonization, (ii) test a previous hypothesis of population replacement in central Mexico following a drastic climate change between 900–1300 CE, (iii) demonstrate a complex admixture model for central Mexico populations, and (iv) detect a contribution from unsampled population A (UpopA) to some northern and central Mexico populations.

Our findings show a geographical structure in pre-Hispanic individuals that differentiates northern and central populations from Mexico, which is consistent with the northwest-southeast cline observed in the genetic structure of present-day Indigenous populations (36). The pre-Hispanic individuals cluster in proximity to present-day Indigenous populations from the same geographical area in the mdPCA and exhibit similar ancestral components based on ADMIXTURE analysis, except for individual MOM6_ST_a (Fig. 3 and figs. S9 and S10). This reflects an overall conservation of the genetic structure of the populations inhabiting the Mexican territory (36) for at least 1400 years (which is the date of the most ancient individual in the dataset). This is

consistent with demographic models based on present-day Indigenous populations (43), (48) which propose a northern-southern population split between 4000 and 10,000 years ago followed by multiple waves of admixture events

between them (48). This geographical structure is also reflected in the maternal lineages (table S5). The spatial distribution of the haplogroups found—namely A, B, C, and D—closely resembles that of present-day Mexico (31, 49).

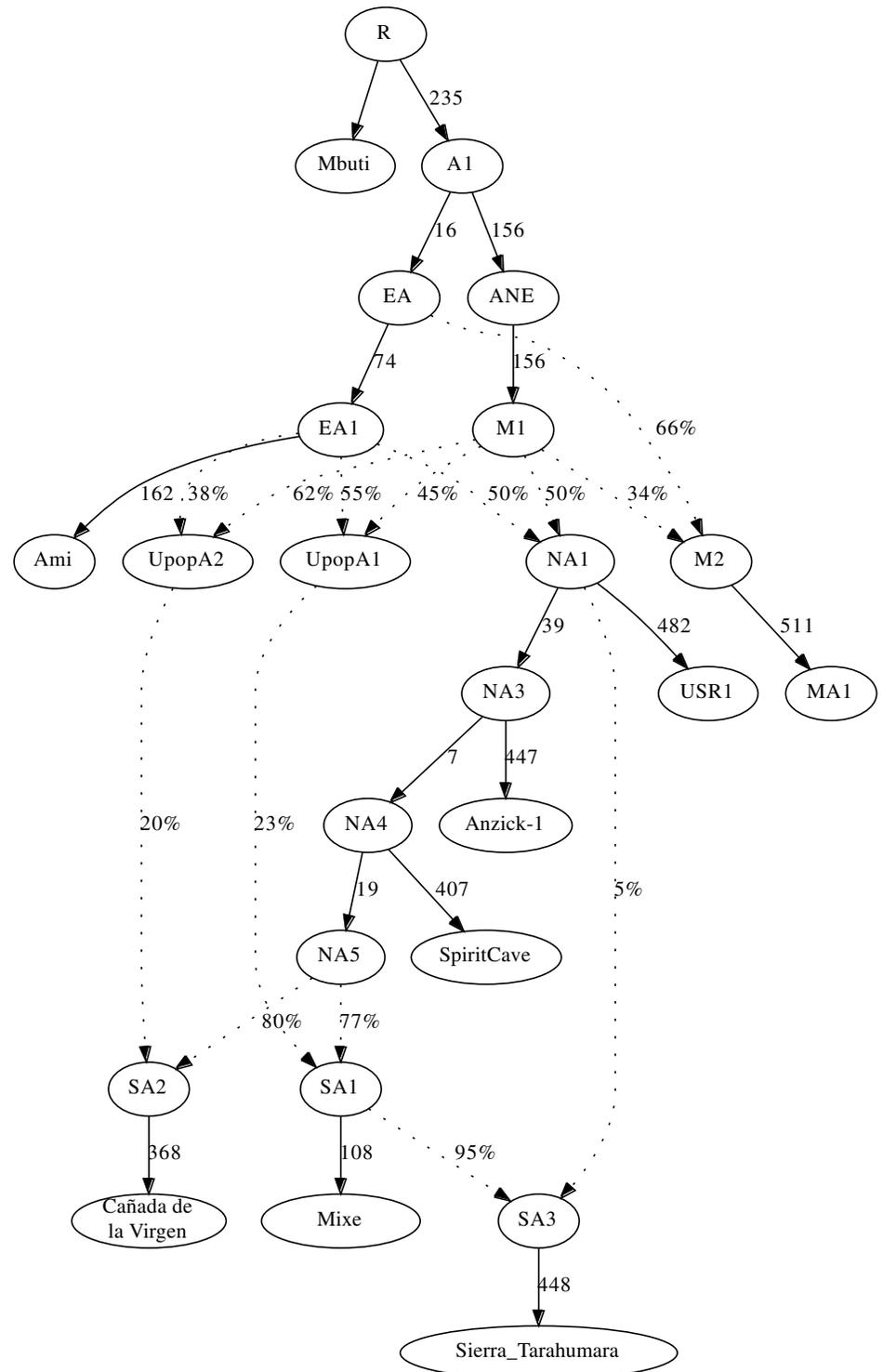


Fig. 5. Ghost ancestries in pre-Hispanic individuals from Mexico. qpGraph model for Sierra Tarahumara and Cañada de la Virgen including present-day Mixe. The tree has no inner O branches and a *z*-score of 2.415. The model shows different ghost ancestries, UPopA1 and UPopA2, in Sierra Tarahumara and Cañada de la Virgen individuals, respectively. UPopA1 indicates the previously unsampled population genetic ancestry reported in Mixe, whereas UPopA2 is a previously unidentified genetic ancestry. All *f* statistics are within a standard error of 1.75.

Downloaded from https://www.science.org at UNAM Campus - Cuernavaca, Morelos on May 25, 2023

Although we lacked the resolution to make accurate inferences about past female population size based solely on the ancient mitogenomes in the E BSP analyses, we were able to recapitulate a population expansion ~5 kyr B.P. when using modern mitogenomes, as it has been previously observed with whole-genome data from present-day Indigenous people from Mexico (50). This expansion is likely related to the domestication and propagation of maize cultivation in Mexico (51). Notably, we did not recover a signal of a population bottleneck ~500 yr B.P. after European colonization in the mitogenome data. We speculate that the demographic impact of European colonization on the female population was variable across Mexico and less drastic than in insular and other isolated regions of the Americas (52–54). The recovery of additional ancient mitogenomes from a wider temporal and spatial range will provide resolution to paint a more accurate picture of the female population history.

When we focused on the genetic diversity at the autosomal level, our results revealed that ancient Aridoamericans and the populations at the northern frontier of Mesoamerica have similar levels of conditional nucleotide diversity (CND) between them and those observed in present-day Indigenous populations from northern (Akimel O'odham) and southern Mexico (Mixe and some Mayan) (fig. S19). However, we caution that we lack genome-wide genetic information of present-day individuals from the exact same site as the pre-Hispanic individuals in central Mexico analyzed here to make a direct temporal comparison. Yet, we observed that the lowest CND and the highest total number of segments in ROH were found in the ancient hunter-gatherer Pericúes and the present-day Akimel O'odham population, both from northern Mexico. In contrast, pre-Hispanic individuals from Sierra Gorda and Cañada de la Virgen and present-day populations from central, southern, and southeastern Mexico had the lowest total number of segments in ROH (fig. S20). Altogether, these results are reflective of the lower population sizes maintained in isolated Aridoamerica populations and the larger population sizes in pre-Hispanic and present-day populations in Mesoamerica. Notably, the pre-Hispanic individuals from this study came from small-sized villages (55, 56) and do not necessarily reflect the demography of other Mesoamerican pre-Hispanic populations from larger and multiethnic pre-Hispanic metropolises (e.g., Teotihuacán, Tenochtitlán, or Palenque) (57, 58), where higher levels of diversity and lower ROH would be expected. Ancient DNA studies on individuals from these sites would be needed to reveal the extent of genetic diversity in these sites before the population collapse inflicted by European colonization.

To directly test the hypothesis of a population replacement at the northern frontier of Mesoamerica by Aridoamerican hunter-gatherers after the 900–1300 CE droughts, we studied Sierra Gorda individuals from the times before and after the climate change episode. Outgroup-f3 analyses showed that individuals from Sierra Gorda pre- and postdrought shared higher genetic drift between them than with any other pre-Hispanic individual. Using D statistics, we observed that pre- and postdrought individuals from Sierra Gorda always form a clade to the exclusion of the rest of pre-Hispanic or present-day populations from Mexico. Then, qpGraphs showed that pre- and postdrought individuals from Sierra Gorda can be modeled in the same clade, which is different from the one formed by F9_ST_a as an Aridoamerican surrogate. Regardless, F9_ST_a might not belong to the population that allegedly replaced the population in Sierra Gorda following the drought; we used F9_ST_a safely as a surrogate for northern genetic ancestry because previous studies have estimated that the split-time between the northern and southern ancestries occurred ~7200 years ago (43). Therefore, even though F9_ST_a likely did not belong to the population that allegedly replaced inhabitants of Sierra Gorda following the drought, we can confidently use this genome as a surrogate for northern genetic ancestry because, by the time of the climate change episode, both ancestries were already well differentiated. Furthermore, with qpWave, we confirmed that the genetic makeup of pre- and postdrought individuals from Sierra Gorda can be explained within the same genetic history without an additional source of genetic variation. Even if this replacement had happened by an unsampled northern population, we would still expect to see signals that significantly differentiated the postdrought individuals with an input from Aridoamerica in the qualitative (mdPCA and ADMIXTURE) and the quantitative (Outgroup-f3, D statistics, qpGraph, and qpWave) analyses, which is not the case. Instead, the evidence points to population continuity in Sierra Gorda after the climate change episode.

A possible explanation for a population continuity despite the droughts is that the favorable climatic conditions at the northern Sierra Gorda maintained higher humidity than in other arid sites of the northern frontier of Mesoamerica, such as Cañada de la Virgen. Because agriculture was the main subsistence strategy in Cañada de la Virgen, the heavy droughts forced its inhabitants to migrate to other regions, which resulted in an abandonment of the site between 1000–1100 CE (56). By contrast, the main subsistence strategy in Toluquilla and Ranas was the mining and trade of cinnabar, a valuable mineral of sacred value in pre-Hispanic cultures (55, 59, 60). We

hypothesize that the cinnabar trade and the landscape of the Sierra Gorda allowed the peoples of Toluquilla and Ranas to subsist despite low rainfall conditions during the drought. Notably, we found shorter ROH segments in the predrought pre-Hispanic individual than in the postdrought individuals. This could indicate a reduction in population size after the climate change episode and suggest a possible demographic impact in the same population. Additional assessment of this hypothesis at other sites at the northern frontier of Mesoamerica will help explain the population migrations and dynamics in a wider geographical range.

We were also interested in obtaining insights into the demography of central Mexico populations because previous attempts of demographic modeling have been hampered by the high genetic heterogeneity observed in these populations (43, 50). Our outgroup-f3 and D statistics results show that, although pre-Hispanic individuals from Mesoamerica in central Mexico share higher genetic drift with present-day populations from that region, none of these relationships were statistically significant. Using qpGraph, we found that the pre-Hispanic populations from central Mexico all have different genetic-ancestry sharing with the NNA and SNA branches. This was expected given previous studies on ancient genomes from Central and South America, which report multiple admixture events between these two branches after their split ~15,000 years ago (15, 48). Together, these observations point to a scenario in which populations from central Mexico have not completely diverged from one another, possibly because of extensive gene flow, as expected on the basis of the active commercial exchange between different Mesoamerican populations that has occurred for centuries (61). This interaction was mainly through trade routes and alliances between different nations (61), as revealed in present-day Indigenous populations who share identity-by-descent (IBD) segments (36, 48). The study of IBD segments in present-day Indigenous populations from Mexico has also evidenced gene flow between Mesoamerican and Aridoamerican populations in pre-Hispanic times (48). However, this gene flow occurred less frequently than within Mesoamerica. The individual MOM6_ST_a from Aridoamerica may have been the result of such admixture between Mesoamerican and Aridoamerican ancestors.

Furthermore, the qpGraph admixture models that we explored for the pre-Hispanic populations showed that F9_ST_a and the ancient individuals from Cañada de la Virgen have genetic ancestry from a ghost, or unsampled, population. A contribution from an unsampled population named UpopA was previously identified in present-day Mixe

(47) as well as in present-day northern and central Indigenous populations from Mexico (48). UpopA was estimated to have diverged ~24,700 years ago from Native Americans (47). Notably, although Sierra Tarahumara shares the contribution of UpopA with Mixe, a second ghost population (UpopA2) was needed to model Cañada de la Virgen in the admixture model together with Sierra Tarahumara and Mixe (Fig. 5). This observation reveals a complex population history in the Americas during the late Pleistocene that needs to be further characterized. Additional ancestral DNA studies from the Americas could help identify the source of both ghost ancestries, which seemingly contribute to many present-day Indigenous populations from Mexico.

We find that the pre-Hispanic population structure from over a thousand years ago can still be observed today. Our work, together with previous studies (15, 36, 43, 48, 50), shows that the demographic events that gave rise to Aridoamerican and Mesoamerican populations are more complex than previously thought. Commercial trade routes may have contributed to increased mobility, facilitating gene flow between different populations within and between various cultural areas. Furthermore, we found genetic continuity in the Sierra Gorda region at the northern frontier of Mesoamerica, which suggests that the local population stayed in their homeland despite the long-standing droughts that forced other populations to abandon their cities. The identification of a second ghost genetic ancestry contribution to some pre-Hispanic central Mexico populations reveals a complex past that needs to be characterized through the ethical study of ancient genomes from Mexico. Our study opens the door for further research to address the questions of the unknown genetic past and population dynamics of Mexican pre-Hispanic populations, whose genetic legacy is retained today among Indigenous and admixed populations.

Materials and Methods

Laboratory procedures

DNA extraction and library preparation (before amplification) were performed in the Human Paleogenomics Laboratory, a clean lab facility at the International Laboratory for Human Genome Research, Universidad Nacional Autónoma de México (LIIGH-UNAM). Bones and teeth surfaces were cleaned with a 1% sodium hypochlorite solution, followed by a solution of 75% ethanol. Then, the surface was UV irradiated (256 nm) for 1.5 min on each side using a UVP CL-1000 crosslinker. A Dremel tool was used to remove the outer surfaces of bones and teeth. Bones were cut to get an inner sample of around 100-200 mg for DNA extraction. Teeth were cut at the cemento-enamel junction, the roots were wrapped in alumi-

num foil and pulverized using a hammer. Around 100-200 mg of the root was used for DNA extraction. DNA extraction was performed using the methods described in (62, 63), as described in table S2.

DNA extraction from individual P_CCM_b was performed taking 150 mg of mummified skin. The sample was washed in deionized sterile water and dried. Then, the sample was UV irradiated (256 nm) for 1.5 min using a UVP CL-1000 crosslinker. Epithelium was cut in small pieces and incubated at 50°C for 24 hours in lysis buffer (10mM Tris-HCl, 10mM NaCl, 5mM CaCl₂, 2.5 mM EDTA, 1% SDS, 10 mg/ml proteinase K, 10 mg/mL DTT). Followed by centrifugation at 16,100 × g for 5 min and recovery of the supernatant. Then, DNA purification was performed according to the protocol in (63).

Double-stranded libraries for Illumina sequencing were prepared according to (64), using single-indexed adapters with 6-base pair (bp) barcodes or double-indexed with 7-bp barcodes, depending on the sequencing platform, NextSeq550 or NovaSeq, respectively. Libraries were analyzed with qPCR to determine the optimum number of cycles during the indexing PCR step. Barcoded libraries were sequenced for a first screening on the NextSeq550 equipment from Illumina using a 2x75 run at either LANGEBIO's genomics core facility (National Laboratory of Genomics for Biodiversity, Irapuato, Guanajuato) or INMEGEN genomics facilities (National Institute of Genomic Medicine, Mexico City). Depending on the quality of the libraries (% endogenous and % clonality), some were chosen to be subjected to mitochondrial genome capture or whole-genome capture using Daicel Arbor Biosciences (Ann Arbor, MI, USA) commercial kits (table S2). Captured libraries were sequenced to assess complexity and yield with the tool preseq (65) and sequenced to higher depth using the NextSeq 550 (2x75 cycles) at LANGEBIO's core facility or in the NovaSeq at the Centro Internacional de Mejoramiento de Maíz y Trigo (CIMMYT) with an S1, 2x100 cycles run (table S2). Sequencing runs for Sierra Tarahumara individuals were carried out at the Sci LifeLab Uppsala Sequencing Center and at the Bustamante Lab in the Department of Genetics at Stanford University (table S2).

Sequence data processing

Raw reads were processed using Adapter-removal (66) for trimming Illumina adapter sequences and collapsing of pairs of reads with an overlap of at least 11 bp (-trimns-trimqualities-qualitybase 33-minlength 30-collapse). Collapsed reads with > 30 bp and quality above 33 were retained for downstream processing.

The retained reads were mapped to the human reference genome b37 (hg19), with the mitochondrial sequence replaced by the

revised Cambridge reference sequence (rCRS, NC_012920). Mapping was done with bwa 0.7.13, aln algorithm, seed disabled (-1 500), and keeping reads with mapping quality > 25. Clonal duplicates and reads mapping to more than one place in the genome were removed with SAMtools rmdup (67) and eliminating reads with labels 'XT:A:R' and 'XA:Z', respectively. Then, reads were realigned using GATK with RealignerTargetCreator and IndelRealigner (68). To authenticate ancient DNA sequence data, mapDamage2 (69) was used to assess the damage patterns and distribution of reads length, using default parameters (figs. S1 to S4). Alignments to rCRS were separated into new bam files using SAMtools (67) to analyze the mitochondrial genome for haplogroup assignment, haplotype network analysis and Extended Bayesian Skyline Plot (EBS). For the analysis of autosomal variants, base qualities were rescaled with mapDamage2 (69). Depth of coverage was estimated using ATLAS (70).

mtDNA analysis

The program Schmutzi (71) was used to estimate mtDNA contamination and to obtain consensus sequences. The reads aligned to the Cambridge reference sequence (rCRS, NC_012920) were analyzed with the script contDeam.pl with parameters-library double and the default-uselength, and schmutzi.pl was then run with 8 threads. Haplogrep 2 (72) was used to assign a mitochondrial haplogroup and quality of assignment for each library using the fasta file with the consensus mitochondrial genome generated by Schmutzi (table S1).

For the estimation of the mitochondrial genetic diversity, we aligned the 27 mitochondrial genomes reconstructed in this study along with the eight public mitochondrial genomes from Mexico (table S4), using the BioEdit sequence alignment editor v. 7.0.5.3 (73). The haplotype list was obtained with DNA Sequence Polymorphism software (DNAsp) v. 6.12.03x64 (74). The sites with gaps were not considered. Then, we estimated the nucleotide diversity π through the genetic distance method of Tajima and Nei, with the software Arlequin v. 3.5.1.3 using default parameters (75).

Consensus fasta sequences of the ancient mtDNA genomes from this and previous studies were aligned together with previously published modern sequences (14, 15, 18, 25, 32, 53, 76-85) (tables S4 and S6) using MEGA X (86). Multiple sequence alignments of mitochondrial genomes of the same haplogroup (A, B, C, or D) were further visually inspected and edited (i.e., removing common indels, as well as to confirm variants and making sure sequences remained at the same length), to assure an accurate alignment. Common hypervariable sites that are phylogenetically uninformative were excluded from the analysis (309, 315,

515-522 AC indels, 3107, 16182-16183, 16193, and 16519) (32). mtDNA sequences were excluded from the alignments based on two criteria: 1) > 200 “N’s” as observed in fasta files or 2) Visual evaluation of an excess of C > T mutations that cannot be explained by random mutation. Four of the mtDNA sequences (333B_TOL_a, 333O_TOL_b, E8_CdV_b and, BC30) were excluded from EBSB analysis but not from haplotype network analysis (table S4), as their high number of N’s, would have likely resulted in overestimation of the mutation rate and biased inferences in EBSB analyses.

Haplotype median-joining networks were constructed in Popart (87) after collapsing aligned full mtDNA sequences into haplotypes with DnaSP v. 6.12.03x64 (74). Haplotype networks were constructed with both ancient (from this study and previously published) ($n = 25$) and present-day ($n = 78$) mtDNA of individuals in the Mexican territory (tables S4 and S6). All present-day samples were retrieved from the PhyloTree (version 17) database (32, 76). Only sequences that were reported as being of the same sub haplogroup or neighboring sub-haplogroup as our ancient samples were chosen for sequence comparison.

Past female effective population sizes were reconstructed using a Bayesian skyline approach (88) in BEAST 2 (89). The ancient mtDNA sequences, mentioned above were merged with present-day mtDNA sequences collected from published public databases (tables S4 and S6). Alignments were partitioned into five concatenated regions in the following order (Dloop, Coding, rRNA, and tRNA) as proposed by (53). HKY+G was found to be the best substitution model for this arrangement of the data using PartitionFinder 2 software (90). Substitution rates were estimated following a strict molecular clock model starting from point estimations as in (91). Additionally, we used tip calibrations for ancient samples using dating estimates (table S1). We ran Markov Chain Monte Carlo (MCMC) chains of 100 million steps for each haplogroup alignment with a sampling of parameters every 10,000 generations, discarding the first 10 million steps as burn-in. Extended Bayesian skyline analysis was plotted using SkyViz reported in (54). Two independent runs of EBSB were run, showing a similar behavior (see table S7 for the statistics of each run). Figure S6, A to N, corresponds to the first of the two EBSB runs.

Sex assignment

Determination of biological sex was made using the tool reported in (92). This approach computes the proportion of reads mapped to the Y chromosome with respect to the reads mapping to the X chromosome (R_y) and Y chromosomes. According to the method, $R_y > 0.075$

corresponds to XY, while $R_y < 0.016$ corresponds to XX (table S1).

Contamination estimation based on the X chromosome

To account for the contamination present in the nuclear DNA, we estimated the contamination with the software hapCon (93), which is based on detecting polymorphic sites on the X chromosome of male individuals. The estimation of contamination was performed only in three individuals meeting the inclusion criteria (XY assignment and $> 0.02x$ on the X chromosome). Estimates with hapCon are reported in Table S1 and were lower ($< 1\%$) than the ones estimated with the mitochondrial genome (1-2%) (table S2).

Y-chromosome haplogroup inference

Y chromosome genotype calling and haplogroup assignments were made as reported in (94). Genotype calling was performed sampling one random base at each site of the Y chromosome covered at least once, with ANGSD (-dohaplocall 1 -doCounts 1 -r chrY: -minMinor 0 -maxMis 4). Then, haplogroup assignments were made using the phylogenetic tree of Y chromosome single nucleotide polymorphisms constructed from the 1000 Genomes Project Phase 3, as in (94). The most derived haplogroup was assigned.

Autosomal reference panels

For analysis at a continental level, we constructed a reference panel with modern populations (Supplementary Table S5) that includes genetic information of: Yoruba (YRI), European genetic ancestry (CEU), and Chinese (CHB) from the 1000 Genomes Project Phase 3 (37, 95), and available genome-wide information for Indigenous populations from Mexico previously published by the Simons Genome Diversity Project (SGDP) and the Human Genome Diversity Project (HGDP) (18, 22). These genome-wide data were intersected using plink (96) with previously published reference panels of genotype information of Indigenous populations from Mexico (36), after masking non-Native American sites as in (97) and keeping individuals with $> 85\%$ genotype information (mind 0.15) and single-nucleotide variants (SNVs) with a minor allele frequency of 0.01 (maf 0.01), and with a 90% genotyping rate (geno 0.1) using the software plink (96). The final reference panel includes genotype information of 576,409 SNVs from 596 individuals from 4 continental populations (YRI, CEU, CHB, and Indigenous from Mexico) (table S8).

For the analysis within Mexico, we used the reference panel including only the Indigenous populations from Mexico (without the other three continental populations: YRI, CEU, and CHB) and keeping the individuals with $> 90\%$ of Native American genetic ancestry. In total,

this reference panel included genetic information of 561,327 SNVs from 268 individuals (table S5).

We also integrated our generated ancient genome-wide data with previously reported low-depth genomes (0.09 – 0.3 x) from three ancient Pericúes from Aridoamerica reported in (14, 15), resulting in a total of 15 pre-Hispanic individuals from Mexico. Our analysis also included ancient genomic data from previous studies for 21 individuals from across the Americas (table S3) (14, 15, 45, 46, 98–101).

Pseudo-haploid calls

Pseudo-haploid calls of the ancient genomic data were made for the positions included in the reference panel. This was done by randomly sampling one read (when more than one read covered the site) at each site present in the reference panel and keeping the observed base at that site if it had a minimum base quality of 30. If only one read covered the site, the observed base was kept if it had a base quality above 30. Then, the calls were turned into homozygous genotypes. Similarly, for the genotype data in the reference panel, one allele was randomly sampled at heterozygous sites and turned into a homozygous genotype. For downstream analyses, we kept ancient individuals with 0.01x genome-wide coverage (table S1), all having > 4650 SNVs intersected with the reference panel.

Relatedness inference

Relatedness between ancient individuals from the same site was assessed using READ (102). A normalization step with a panel of non-related individuals is required previous to estimating the relationship in the ancient individuals, this step was performed with all Nahuas from the reference panel as they are the biggest group in the reference panel. Then, the normalization value was used to run READ with the ancient individuals. Two individuals from Toluquilla (333C_TOL_a and 333B_TOL_a) were identified as first-degree relatives. Thus, the one with the highest coverage (333B_TOL_a) was used for downstream analyses. Furthermore, individuals E4_CdV_b and E8_CdV_b were identified as second-degree relatives, as well as E7_CdV_b and E19_CdV_b.

ADMIXTURE analysis

For each dataset analyzed with ADMIXTURE, one hundred replicates were run with randomly generated seed values and calculating the cross-validation error with parameter -cv. The run with the best likelihood for each k was plotted using pong software (103). A first ADMIXTURE (33) analysis was carried out on the genotype reference panel including Indigenous individuals from Mexico and the three continental populations (YRI, CEU, and CHB)

and $k=2$ through $k=9$ to identify present-day Indigenous individuals with <90% Native American genetic ancestry. We found $k=4$ to be the k with the lowest mean cv error (fig. S8A), separating YRI, CEU, CHB, and Native American genetic ancestry. The run of $k=4$ with the best likelihood was chosen and the Indigenous individuals with <90% Native American genetic ancestry were removed from the panel. A second ADMIXTURE analysis was carried out with only the present-day Indigenous populations with >90% Native American genetic ancestry, from $k=2$ to $k=8$, being $k=2$ the one with the lowest cv error (fig. S8B) and separating populations with northern and southern ancestries. Then, a third ADMIXTURE run was carried out including the present-day Indigenous populations and pre-Hispanic individuals (Fig. 3).

Missing DNA PCA

Missing DNA PCA (mdPCA) is a principal component analysis that corrects for genotypes missing due to genetic ancestry masking or degradation in ancient DNA samples. The correction is performed by comparing genetic distances between all samples. The genetic distance is computed as the average number of pairwise differences (π as defined in (104)). The manual of the method and the description of the parameters can be found in the mdPCA GitHub repository <https://github.com/Al-sandbox/mdPCA>. The dataset included 15 ancient samples (12 from this study and 3 Pericúes previously published) and the reference panel including present-day Indigenous individuals from Mexico. Genotyping data in plink format was converted to an unphased VCF format with Plink version 1.90 beta (96). The method used this VCF file as an input with the following parameters: each individual was plotted as an average of both parental haplotypes (AVERAGE_PARENTS=True), weights for each individual were used inversely proportional to the number of samples from the corresponding population (IS_WEIGHTED=True), the simple weighted of covariance PCA without any optimization was used (METHOD=1), no genotype was masked based on local genetic ancestry calls (IS_MASKED=False), and only the three first PC's were calculated (NUM_DIMS=3). Individuals with <0.15x whole genome coverage (19,851 SNVs or less) were further down-weighted to 0.1x to avoid having the principal components be selected based on noise variance stemming from these samples. These included BC30, MOM6_ST_a, IIR_R_b, E4_CdV_b, E7_CdV_b, E19_CdV_b, E2_Mich_b, and E4_Mich_b.

Principal component analysis

We performed principal component analyses (PCA) using smartpca from the software eigensoft v6.0.1 (34) and projecting the indi-

viduals (lsqproject: YES) on the PCs estimated for the modern populations. The PCA (fig. S7) with only the ancient individuals as well as the present-day Indigenous populations from Mexico, only included present-day individuals masked and with >90% Native American genetic ancestry.

Temporal factor analysis

Temporal factor analysis (TFA) (38) were run with the reference panel that includes only the Indigenous populations from Mexico. The ancient samples included in the TFA consisted exclusively of individuals from Mexican territory from this study with >1x genome-wide coverage: one individual from Sierra Tarahumara (F9_ST_b), and three from Toluquilla (2417Q_TOL_b, 2417J_TOL_a, and 333B_TOL_a). Genotyping data was manipulated with Plink version 1.90 beta (96). The dataset was converted from Plink format to Eigenstrat's geno format with the command convert from Eigensoft version 6.0.1.1 (34). Missing genotypes were imputed with the snmf function from the LEA package, using the parameters $K=2$, entropy= TRUE, and repetitions=5. The runs with the lowest cross-entropy value were considered for posterior steps. Genotypes were corrected according to the autosomal coverage depth of each sample. We assigned the average depth from the HGDP dataset mentioned in the publication: 35X (22) (table S8). We assigned the respective coverage read depth for all 11 individuals from the Simons Genome Diversity Project (18) (table S8). Modern samples genotyped with microarrays (36) did not have a coverage depth value as they represent a different technology without sequencing reads, therefore, we assigned the maximum value from the dataset to all microarray samples, i.e., 56.19X. TFA was applied to these imputed and corrected genotypes with the following parameters: $\lambda = 5e-1$, and $K=2$. Imputation, coverage correction, TFA, and the manipulation of the genotyping matrix in geno format were performed with R version 4.0.2. Variance explained by λ did not change considerably across logarithmic values, thus we chose the default λ value $5e-1$. We plotted factor 1 and 2 as negative when required to match geography and facilitate the comparison between TFA plots (fig. S10).

Outgroup f3 and D statistics

To estimate each pre-Hispanic individual's genetic relatedness with a particular pre-Hispanic or present-day Indigenous population, we performed outgroup f_3 statistics using ADMIXTOOLS v5.0 (44). Outgroup f_3 was calculated in the form (Test, Source1; YRI), where the "Test" is the pre-Hispanic individual under analysis, and Source1 being another pre-Hispanic individual or present-day Indigenous popula-

tion. We used all YRI from the 1000 Genomes Project (37, 95) as an outgroup population. Assuming there was no admixture in the tree, the f_3 value is proportional to the shared genetic drift between Test and Source1.

We performed D statistics using ADMIXTOOLS v5.0 (44) to identify whether a pre-Hispanic individual had higher genetic affinities with: i) other pre-Hispanic individuals from the same archaeological site than to pre-Hispanic individuals from other sites; or ii) a specific present-day indigenous population than to other present-day population. D values were computed in the form (Pop1, Pop2; Test, YRI). The "Test" refers to the pre-Hispanic individual under analysis, while Pop1 and Pop2 were all possible combinations between pre-Hispanic individuals and present-day Indigenous populations. Under the null hypothesis, Test individual would be equally related to Pop1 and Pop2, and we expect no significant deviations from $D=0$. Significant $D<0$ indicates closer relation between Test individual and Pop2. Moreover, a significant $D>0$ indicates closer relation between Test individual and Pop1. Only absolute values of Z-score > 3 were considered as significant (which corresponds to a p-value of ~ 0.0027).

Population continuity test between F9_ST_a and present-day Rarámuri

According to the PCA, TFA, ADMIXTURE, outgroup- f_3 and D statistics, individual F9_ST_a from north Mexico is significantly closer to present-day Rarámuri. Thus, we analyzed whether individual F9 belonged to the present-day Rarámuri ancestral population using the method published in (39). Derived alleles in individual F9 and present-day Rarámuri were identified using ancestral alleles' information from (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/pilot_data/technical/reference/ancestral_alignments). Then, allele counts were performed for each site in the present-day Rarámuri (reference population), excluding alleles with frequency 0 or 1. We performed allele counts for the pre-Hispanic individual F9. Parameters alpha and beta necessary for the analysis were estimated using $\text{min_samples} = 14$, which refers to the minimum number of individuals required in the reference panel that have information for each SNV. Under the null hypothesis, it is assumed that reference population and ancient individual belong to a continuous population, rejection of the null hypothesis would mean no population continuity.

qpGraph

We used the qpGraph tool of the ADMIXTOOLS v5.0 software package (44) to construct admixture graphs and test which various models of demographic history best fit the data. We applied different models according to the

analysis: (i) For the model of population continuity, we used the 11,500-year-old USR1 (45) as the surrogate for ancient Beringian, 10,700 year-old Anzick (46) individual as the surrogate for ancestral Southern Native American (SNA) (14), and the 4200 year-old Ancient Southwestern Ontario individual (ASO) (15) as the surrogate for Northern Native American (NNA) (14). (ii) For the admixture model for Central Mexico populations, the base model included the pre-Hispanic populations, Anzick (46) individual as the ancestral SNA (14), present-day Athabascan as the representative for NNA (14), and present-day Indigenous populations in Mexico representing the northern (Konkaak), central (Nahua from Puebla), and southern (Triqui) regions. (iii) For UpopA we used the different models described in main text and (17). pre-Hispanic individuals were added to each model in various combinations to test the fit. The following parameters were applied: outpop (NULL), diag (0.0001), hires (YES), blgsize (0.05), and lsqmode (YES). A worst fitting f statistical test of z -score > 3 is set as a threshold for model rejection.

qpWave

We used qpWave version 410 (44) to test whether pre- and post-drought Sierra Gorda (SG) individuals could be modeled using either SG individuals or other ancient or modern Mexican indigenous population relative to a set of reference groups (Mbuti, Onge, USR1; Papuan, Anzick, Han, Mixe, MAI, Zapotec, Karitiana, Piapoco) as in (105). In order to maximize the power for this test we increased the number of SNVs to 2,224,359k transversion sites ascertained in IKGP and having a 5% MAF. Present-day human sequencing data from HGDP genomes (22) was used to represent Mexican indigenous genetic variation. With these data, we tested whether we could model the genetic variation of pre (2417Q_TOL_b) and post-drought Sierra Gorda (2417J_TOL_a and 333B_TOL_a) individually, using as a source either another Sierra Gorda individual or another Mexican indigenous population. For the latter case we used as references the pre-Hispanic individual F9_ST_a, and Pericúes, along with present-day Akimel O'odham and Mayan. If pre- and post-drought Sierra Gorda individuals arise from different demographic histories to each other, then when analyzed together it would be required to add one or more different sources of genetic variation to explain the data, which would be denoted by a p -value < 0.05 . We also performed qpWave analysis with the panel of Indigenous populations from Mexico reported in (36) to increase the number of populations, though at the cost of reducing the number of overlapping sites. Results are consistent with those obtained using the HGDP Akimel O'odham

and Mayan in that post-drought genetic variation can be explained with pre-drought genetic variation, suggesting genetic continuity. However, using the Moreno-Estrada reference panel resulted in a low number of overlapping SNVs for the qpWave analysis and, consequently, some p -values could not be computed by the program (e.g., when using the pre and post-drought samples as both target (Pop1) and sources (Pop2) (table S18).

Conditional nucleotide diversity

CND, a relative measure of genetic diversity, was estimated by genetic pairwise comparison of the genomic positions that are variable in African individuals (40). For these comparisons we grouped the individuals per archeological site. This analysis included 16 pre-Hispanic individuals (Pericúes, Sierra Tarahumara, Sierra Gorda, Cañada de la Virgen and Patagonia) with at least 10,000 SNVs overlapping between them (tables S11 and S12) and present-day populations from Mexico (18, 22).

This consists of counting the differences between the ascertained alleles present in one pair of individuals from the same population. For this analysis, we used a panel of 1,938,919 transversions SNVs with a minor allele frequency of 0.01 in the YRI population. In all cases, pseudohaploid genotype calls were performed before the comparisons. CND values were calculated for all pairs of individuals for each present-day Indigenous population and for all pairs of pre-Hispanic individuals within the same site. For present-day populations we masked the sites with no Native American genetic ancestry to avoid counting levels of diversity from a different genetic ancestry. Mean and standard error values were estimated 100 times using a window size of 50 Kb even for pairs of ancient individuals. Results are reported in a violin plot, where each dot inside the violin corresponds to the CND value for a unique pair of individuals (fig. S19).

Runs of homozygosity

We used the tool hapROH (41) to detect runs of homozygosity in pre-Hispanic individuals. The distribution of ROHs was used as an indirect way of measuring genetic diversity and effective population size. This method detects the ROH's lengths longer than four centimorgans (cM) and up to 300 cM. The sum of the ROH detected gives insights into the relatedness of the parents of each individual and informs about population sizes. Length of ROH and relatedness between parents are positively correlated. While the higher the number of ROH segments, the smaller the population size and the smaller the number of heterozygote sites, which means less genetic diversity. hapROH uses a built-in set of 1240K SNVs from to perform pseudohaploid genotype calls and a reference panel of 5,008 present-day

human haplotypes from (37). We performed a pseudohaploid variant calling for the 1240K SNVs for each pre-Hispanic genome. hapROH's manual recommends using samples with a minimum of 400K SNVs intersected with the 1240K SNVs. However, we also included the individual E8_CdV_b, which intersected 369,504 SNVs, and the Pericú individual B03, which intersected 370,026 SNVs, to have at least one individual of these sites in the analyses and considering the number of SNVs is not too far from 400K (table S13). Detection of runs of homozygosity was made using the set of 5,008 haplotypes mentioned above.

REFERENCES AND NOTES

1. A. L. Austin, L. L. Luján, *El pasado indígena* (Fondo de Cultura Económica, 2019).
2. B. B. Cornejo, *La frontera septentrional de Mesoamérica* (Instituto Nacional de Antropología e Historia, México, ed. 2, 2000), vol. 1 of *Historia Antigua de México*.
3. P. Kirchhoff, Mesoamerica: its geographic limits, ethnic composition and cultural characteristics. *Heritage of conquest*, pp. 17–30 (1952).
4. M. del C. Solanes, E. Vela, Atlas del México prehispánico. *Arqueol. Mex.* 5, 5–18 (2000).
5. D. W. Stahle *et al.*, Major Mesoamerican droughts of the past millennium. *Geophys. Res. Lett.* 38, L05703 (2011). doi: 10.1029/2010GL046472
6. M. K. Hughes, H. F. Diaz, Was there a 'medieval warm period', and if so, where and when? *Clim. Change* 26, 109–142 (1994). doi: 10.1007/BF01092410
7. P. Armillas, Condiciones ambientales y movimientos de pueblos en la frontera septentrional de Mesoamérica. *Homenaje a Fernando Márquez-Miranda*, 62–82 (1964).
8. K. H. Wogau, H. W. Arz, H. N. Böhm, N. R. Nowaczyk, J. Park, High resolution paleoclimate and paleoenvironmental reconstruction in the Northern Mesoamerican Frontier for Prehistory to Historical times. *Quat. Sci. Rev.* 226, 106001 (2019). doi: 10.1016/j.quascirev.2019.106001
9. D. J. Kennett *et al.*, Development and disintegration of Maya political systems in response to climate change. *Science* 338, 788–791 (2012). doi: 10.1126/science.1226299; pmid: 23139330
10. J. A. Hoggarth *et al.*, The political collapse of Chichén Itzá in climatic and cultural context. *Global Planet. Change* 138, 25–42 (2016). doi: 10.1016/j.gloplacha.2015.12.007
11. S. L. O'Hara, S. E. Metcalfe, F. A. Street-Perrott, On the arid margin: The relationship between climate, humans and the environment. A review of evidence from the highlands of central Mexico. *Chemosphere* 29, 965–981 (1994). doi: 10.1016/0045-6535(94)90163-5; pmid: 7953466
12. B. Braniff, Oscilación de la frontera norte mesoamericana: un nuevo ensayo. *Arqueología, segunda época*, pp. 99–114 (1989).
13. A. Gómez Aiza, S. Sánchez Vázquez, La Frontera Cultural Meso-Aridoamericana: Construcción De Imaginarios Nacionalistas En La Historia Mexicana. *Xihmai* 9, (2014). doi: 10.37646/xihmai.v9i18.244
14. M. Raghavan *et al.*, Genomic evidence for the Pleistocene and recent population history of Native Americans. *Science* 349, aab3884–aab3884 (2015). doi: 10.1126/science.aab3884; pmid: 26198033
15. C. L. Scheib *et al.*, Ancient human parallel lineages within North America contributed to a coastal expansion. *Science* 360, 1024–1027 (2018). doi: 10.1126/science.aar6851; pmid: 29853687
16. M. C. Ávila-Arcos, C. de la Fuente Castro, M. A. Nieves-Colón, M. Raghavan, Recommendations for Sustainable Ancient DNA Research in the Global South: Voices From a New Generation of Paleogenomicists. *Front. Genet.* 13, 880170 (2022). doi: 10.3389/fgene.2022.880170; pmid: 35559028
17. Materials and methods are available as supplementary materials.
18. S. Mallick *et al.*, The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* 538, 201–206 (2016). doi: 10.1038/nature18964; pmid: 27654912

19. I. De La Cruz *et al.*, Sex Identification of Children Sacrificed to the Ancient Aztec Rain Gods in Tlatelolco. *Curr. Anthropol.* **49**, 519–526 (2008). doi: [10.1086/587642](https://doi.org/10.1086/587642)
20. D. Perez-Benedico *et al.*, Mayans: A Y chromosome perspective. *Eur. J. Hum. Genet.* **24**, 1352–1358 (2016). doi: [10.1038/ejhg.2016.18](https://doi.org/10.1038/ejhg.2016.18); pmid: 26956252
21. K. Sandoval *et al.*, Y-chromosome diversity in Native Mexicans reveals continental transition of genetic structure in the Americas. *Am. J. Phys. Anthropol.* **148**, 395–405 (2012). doi: [10.1002/ajpa.22062](https://doi.org/10.1002/ajpa.22062); pmid: 22576278
22. A. Bergström *et al.*, Insights into human genetic variation and population history from 929 diverse genomes. *Science* **367**, eaay5012 (2020). doi: [10.1126/science.aay5012](https://doi.org/10.1126/science.aay5012); pmid: 32193295
23. B. Z. González-Sobrinho *et al.*, Genetic Diversity and Differentiation in Urban and Indigenous Populations of Mexico: Patterns of Mitochondrial DNA and Y-Chromosome Lineage. *Biodemography Soc. Biol.* **62**, 53–72 (2016).
24. H.-J. Bandelt *et al.*, Identification of Native American founder mtDNAs through the analysis of complete mtDNA sequences: Some caveats. *Ann. Hum. Genet.* **67**, 512–524 (2003). doi: [10.1046/j.1469-1809.2003.00049.x](https://doi.org/10.1046/j.1469-1809.2003.00049.x); pmid: 14641239
25. A. Y. Morales-Arce, M. H. Snow, J. H. Kelley, M. Anne Katzenberg, Ancient mitochondrial DNA and ancestry of Paquimé inhabitants, Casas Grandes (A.D. 1200–1450). *Am. J. Phys. Anthropol.* **163**, 616–626 (2017). doi: [10.1002/ajpa.23223](https://doi.org/10.1002/ajpa.23223); pmid: 28382719
26. J. Mata-Míguez, L. Overholtzer, E. Rodríguez-Alegria, B. M. Kemp, D. A. Bolnick, The genetic impact of Aztec imperialism: Ancient mitochondrial DNA evidence from Xaltocan, Mexico. *Am. J. Phys. Anthropol.* **149**, 504–516 (2012). doi: [10.1002/ajpa.22152](https://doi.org/10.1002/ajpa.22152); pmid: 23076995
27. A. Y. Morales-Arce *et al.*, Ancient mitochondrial DNA and population dynamics in postclassic Central Mexico: Tlatelolco (ad 1325–1520) and Cholula (ad 900–1350). *Archaeol. Anthropol. Sci.* **11**, 3459–3475 (2019). doi: [10.1007/s12520-018-00771-7](https://doi.org/10.1007/s12520-018-00771-7)
28. A. González-Oliver, L. Márquez-Morfin, J. C. Jiménez, A. Torre-Blanco, Founding Amerindian mitochondrial DNA lineages in ancient Maya from Xcaret, Quintana Roo. *Am. J. Phys. Anthropol.* **116**, 230–235 (2001). doi: [10.1002/ajpa.1118](https://doi.org/10.1002/ajpa.1118); pmid: 11596002
29. A. Y. Morales-Arce *et al.*, Successful reconstruction of whole mitochondrial genomes from ancient Central America and Mexico. *Sci. Rep.* **7**, 18100 (2017). doi: [10.1038/s41598-017-18356-0](https://doi.org/10.1038/s41598-017-18356-0); pmid: 29273718
30. B. M. Kemp *et al.*, Evaluating the Farming/Language Dispersal Hypothesis with genetic variation exhibited by populations in the Southwest and Mesoamerica. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 6759–6764 (2010). doi: [10.1073/pnas.0905753107](https://doi.org/10.1073/pnas.0905753107); pmid: 20351276
31. M. Bodner *et al.*, The Mitochondrial DNA Landscape of Modern Mexico. *Genes* **12**, 1453 (2021). doi: [10.3390/genes12091453](https://doi.org/10.3390/genes12091453); pmid: 34573435
32. M. van Oven, M. Kayser, Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. *Hum. Mutat.* **30**, E386–E394 (2009). doi: [10.1002/humu.20921](https://doi.org/10.1002/humu.20921); pmid: 18853457
33. D. H. Alexander, J. Novembre, K. Lange, Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009). doi: [10.1101/gr.094052.109](https://doi.org/10.1101/gr.094052.109); pmid: 19648217
34. N. Patterson, A. L. Price, D. Reich, Population structure and eigenanalysis. *PLoS Genet.* **2**, e190 (2006). doi: [10.1371/journal.pgen.0020190](https://doi.org/10.1371/journal.pgen.0020190); pmid: 17194218
35. D. Reich, A. L. Price, N. Patterson, Principal component analysis of genetic data. *Nat. Genet.* **40**, 491–492 (2008). doi: [10.1038/ng0508-491](https://doi.org/10.1038/ng0508-491); pmid: 18443580
36. A. Moreno-Estrada *et al.*, The genetics of Mexico recapitulates Native American substructure and affects biomedical traits. *Science* **344**, 1280–1285 (2014). doi: [10.1126/science.1251688](https://doi.org/10.1126/science.1251688); pmid: 24926019
37. A. Auton *et al.*, A global reference for human genetic variation. *Nature* **526**, 68–74 (2015). doi: [10.1038/nature15393](https://doi.org/10.1038/nature15393); pmid: 26432245
38. O. François, F. Jay, Factor analysis of ancient population genomic samples. *Nat. Commun.* **11**, 4661 (2020). doi: [10.1038/s41467-020-18335-6](https://doi.org/10.1038/s41467-020-18335-6); pmid: 32938925
39. J. G. Schraiber, Assessing the Relationship of Ancient and Modern Populations. *Genetics* **208**, 383–398 (2018). doi: [10.1534/genetics.117.300448](https://doi.org/10.1534/genetics.117.300448); pmid: 29167200
40. P. Skoglund *et al.*, Genomic diversity and admixture differs for Stone-Age Scandinavian foragers and farmers. *Science* **344**, 747–750 (2014). doi: [10.1126/science.1253448](https://doi.org/10.1126/science.1253448); pmid: 24762536
41. H. Ringbauer, J. Novembre, M. Steinrücken, Parental relatedness through time revealed by runs of homozygosity in ancient DNA. *Nat. Commun.* **12**, 5425 (2021). doi: [10.1038/s41467-021-25289-w](https://doi.org/10.1038/s41467-021-25289-w); pmid: 34521843
42. D. I. Cruz Dávalos *et al.*, Indigenous peoples in eastern Brazil: Insights from 19th century genomes and metagenomes. bioRxiv 448717 [Preprint] (2022) [cited 2023]. doi: [10.1101/2022.01.27.477466](https://doi.org/10.1101/2022.01.27.477466)
43. M. C. Ávila-Arcos *et al.*, Population History and Gene Divergence in Native Mexicans Inferred from 76 Human Exomes. *Mol. Biol. Evol.* **37**, 994–1006 (2020). doi: [10.1093/molbev/msz282](https://doi.org/10.1093/molbev/msz282); pmid: 31848607
44. N. Patterson *et al.*, Ancient admixture in human history. *Genetics* **192**, 1065–1093 (2012). doi: [10.1534/genetics.112.145037](https://doi.org/10.1534/genetics.112.145037); pmid: 22960212
45. J. V. Moreno-Mayar *et al.*, Terminal Pleistocene Alaskan genome reveals first founding population of Native Americans. *Nature* **553**, 203–207 (2018). doi: [10.1038/nature25173](https://doi.org/10.1038/nature25173); pmid: 29323294
46. M. Rasmussen *et al.*, The genome of a Late Pleistocene human from a Clovis burial site in western Montana. *Nature* **506**, 225–229 (2014). doi: [10.1038/nature13025](https://doi.org/10.1038/nature13025); pmid: 24522598
47. J. V. Moreno-Mayar *et al.*, Early human dispersals within the Americas. *Science* **362**, eaav2621 (2018). doi: [10.1126/science.aav2621](https://doi.org/10.1126/science.aav2621); pmid: 30409807
48. H. García-Ortiz *et al.*, The genomic landscape of Mexican Indigenous populations brings insights into the peopling of the Americas. *Nat. Commun.* **12**, 5942 (2021). doi: [10.1038/s41467-021-26188-w](https://doi.org/10.1038/s41467-021-26188-w); pmid: 34642312
49. A. Gorostiza *et al.*, Reconstructing the history of Mesoamerican populations through the study of the mitochondrial DNA control region. *PLOS ONE* **7**, e44666 (2012). doi: [10.1371/journal.pone.0044666](https://doi.org/10.1371/journal.pone.0044666); pmid: 23028577
50. S. Romero-Hidalgo *et al.*, Demographic history and biologically relevant genetic variation of Native Mexicans inferred from whole-genome sequencing. *Nat. Commun.* **8**, 1005 (2017). doi: [10.1038/s41467-017-01194-z](https://doi.org/10.1038/s41467-017-01194-z); pmid: 29044207
51. J. Ramos-Madriral *et al.*, Genome Sequence of a 5,310-Year-Old Maize Cob Provides Insights into the Early Stages of Maize Domestication. *Curr. Biol.* **26**, 3195–3201 (2016). doi: [10.1016/j.cub.2016.09.036](https://doi.org/10.1016/j.cub.2016.09.036); pmid: 27866890
52. M. A. Nieves-Colón *et al.*, Ancient DNA Reconstructs the Genetic Legacies of Precontact Puerto Rico Communities. *Mol. Biol. Evol.* **37**, 611–626 (2020). doi: [10.1093/molbev/msz267](https://doi.org/10.1093/molbev/msz267); pmid: 31710665
53. B. Llamas *et al.*, Ancient mitochondrial DNA provides high-resolution time scale of the peopling of the Americas. *Sci. Adv.* **2**, e1501385 (2016). doi: [10.1126/sciadv.1501385](https://doi.org/10.1126/sciadv.1501385); pmid: 27051878
54. B. D. O'Fallon, L. Fehren-Schmitz, Native Americans experienced a strong population bottleneck coincident with European contact. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 20444–20448 (2011). doi: [10.1073/pnas.1112563108](https://doi.org/10.1073/pnas.1112563108); pmid: 22143784
55. A. J. Herrera Muñoz, E. Mejía Campos, Un minero en la Sierra Gorda: caso de contaminación ocupacional multielemental de metales pesados a finales del período Clásico. *Arqueología (México, DF)*, 127–135 (2017).
56. E. Nalda, La arqueología de Guanajuato: Trabajos recientes. *Arqueol. Mex.* **16**, 36–43 (2008).
57. M. E. Smith, Mesoamerica's First World City: Teotihuacan in Comparative Context. *Teotihuacan: The World Beyond the City* (2020).
58. B. A. Álvarez-Sandoval, L. R. Manzanilla, M. González-Ruiz, A. Malgosa, R. Montiel, Genetic Evidence Supports the Multiethnic Character of Teopancazco, a Neighborhood Center of Teotihuacan, Mexico (AD 200–600). *PLOS ONE* **10**, e0132371 (2015). doi: [10.1371/journal.pone.0132371](https://doi.org/10.1371/journal.pone.0132371); pmid: 26200455
59. E. Campos, J. Campos, A. Muñoz, Mercurialism Determination in Fetuses Bone Remains from Toluquilla, Queretaro, Mexico. *J. Sci. Res. Rep.* **8**, 1–10 (2015). doi: [10.9734/JSRR/2015/18043](https://doi.org/10.9734/JSRR/2015/18043)
60. E. M. Pérez Campos, A. J. Herrera Muñoz, Minas y mineros: presencia de metales en sedimentos y restos humanos al sur de la Sierra Gorda de Querétaro en México. *Chungara (Arica)* **45**, 161–176 (2013). doi: [10.4067/S0717-73562013000100008](https://doi.org/10.4067/S0717-73562013000100008)
61. D. A. C. Velázquez, Expansión Territorial Comercial en Mesoamérica y Mesoamérica Septentrional por medio del Intercambio. *Geografía Enseña & Pesquisa* **19**, 59–68 (2015).
62. N. Rohland, M. Hofreiter, Ancient DNA extraction from bones and teeth. *Nat. Protoc.* **2**, 1756–1762 (2007). doi: [10.1038/nprot.2007.247](https://doi.org/10.1038/nprot.2007.247); pmid: 17641642
63. J. Dabney *et al.*, Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 15758–15763 (2013). doi: [10.1073/pnas.1314445110](https://doi.org/10.1073/pnas.1314445110); pmid: 24019490
64. M. Meyer, M. Kircher, Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harb. Protoc.* **2010**, prot5448 (2010). doi: [10.1101/pdb.prot5448](https://doi.org/10.1101/pdb.prot5448); pmid: 20516186
65. T. Daley, A. D. Smith, Predicting the molecular complexity of sequencing libraries. *Nat. Methods* **10**, 325–327 (2013). doi: [10.1038/nmeth.2375](https://doi.org/10.1038/nmeth.2375); pmid: 23435259
66. M. Schubert, S. Lindgreen, L. Orlando, AdapterRemoval v2: Rapid adapter trimming, identification, and read merging. *BMC Res. Notes* **9**, 88 (2016). doi: [10.1186/s13104-016-1900-2](https://doi.org/10.1186/s13104-016-1900-2); pmid: 26868221
67. H. Li *et al.*, The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009). doi: [10.1093/bioinformatics/btp352](https://doi.org/10.1093/bioinformatics/btp352); pmid: 19505943
68. A. McKenna *et al.*, The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010). doi: [10.1101/gr.107524.110](https://doi.org/10.1101/gr.107524.110); pmid: 20644199
69. H. Jönsson, A. Ginolhac, M. Schubert, P. L. F. Johnson, L. Orlando, mapDamage2.0: Fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics* **29**, 1682–1684 (2013). doi: [10.1093/bioinformatics/btt193](https://doi.org/10.1093/bioinformatics/btt193); pmid: 23613487
70. V. Link *et al.*, ATLAS: Analysis Tools for Low-depth and Ancient Samples, bioRxiv 105346 [Preprint] (2017) [cited 2023]. doi: [10.1101/105346](https://doi.org/10.1101/105346)
71. G. Renaud, V. Slon, A. T. Duggan, J. Kelso, Schmutzi: Estimation of contamination and endogenous mitochondrial consensus calling for ancient DNA. *Genome Biol.* **16**, 224 (2015). doi: [10.1186/s13059-015-0776-0](https://doi.org/10.1186/s13059-015-0776-0); pmid: 26458810
72. H. Weissensteiner *et al.*, HaploGrep 2: Mitochondrial haplogroup classification in the era of high-throughput sequencing. *Nucleic Acids Res.* **44**, W58–63 (2016). doi: [10.1093/nar/gkw233](https://doi.org/10.1093/nar/gkw233); pmid: 27084951
73. T. A. Hall, BioEdit: A User-Friendly Biological Sequence Alignment Editor and Analysis Program for Windows 95/98/NT. *Nucleic Acids Symp Ser* **41**, 95–98 (1999).
74. J. Rozas *et al.*, DnaSP 6: DNA Sequence Polymorphism Analysis of Large Data Sets. *Mol. Biol. Evol.* **34**, 3299–3302 (2017). doi: [10.1093/molbev/msz248](https://doi.org/10.1093/molbev/msz248); pmid: 29029172
75. L. Excoffier, H. E. Lischer, Arlequin suite ver 3.5: A new series of programs to perform population genetics analyses under Linux and Windows. *Mol. Ecol. Resour.* **10**, 564–567 (2010). doi: [10.1111/j.1755-0998.2010.02847.x](https://doi.org/10.1111/j.1755-0998.2010.02847.x); pmid: 21565059
76. M. van Oven, PhyloTree Build 17: Growing the human mitochondrial DNA tree. *Forensic Sci. International. Genet. Suppl. Ser.* **5**, e392–e394 (2015). doi: [10.1016/j.fsigss.2015.09.155](https://doi.org/10.1016/j.fsigss.2015.09.155)
77. S. Kumar *et al.*, Large scale mitochondrial sequencing in Mexican Americans suggests a reappraisal of Native American origins. *BMC Evol. Biol.* **11**, 293 (2011). doi: [10.1186/1471-2148-11-293](https://doi.org/10.1186/1471-2148-11-293); pmid: 21978175
78. U. A. Perego *et al.*, Distinctive Paleo-Indian migration routes from Beringia marked by two rare mtDNA haplogroups. *Curr. Biol.* **19**, 1–8 (2009). doi: [10.1016/j.cub.2008.11.058](https://doi.org/10.1016/j.cub.2008.11.058); pmid: 19135370
79. A. Achilli *et al.*, Reconciling migration models to the Americas with the variation of North American native mitochondrial genomes. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 14308–14313 (2013). doi: [10.1073/pnas.1306290110](https://doi.org/10.1073/pnas.1306290110); pmid: 23940335
80. A. Achilli *et al.*, Mitochondrial DNA backgrounds might modulate diabetes complications rather than T2DM as a whole. *PLOS ONE* **6**, e21029 (2011). doi: [10.1371/journal.pone.0021029](https://doi.org/10.1371/journal.pone.0021029); pmid: 21695278
81. U. A. Perego *et al.*, The initial peopling of the Americas: A growing number of founding mitochondrial genomes from Beringia. *Genome Res.* **20**, 1174–1179 (2010). doi: [10.1101/gr.109231.110](https://doi.org/10.1101/gr.109231.110); pmid: 20587512
82. D. M. Behar *et al.*, A “Copernican” reassessment of the human mitochondrial DNA tree from its root. *Am. J. Hum.*

- Genet.* **90**, 675–684 (2012). doi: [10.1016/j.ajhg.2012.03.002](https://doi.org/10.1016/j.ajhg.2012.03.002); pmid: [22482806](https://pubmed.ncbi.nlm.nih.gov/22482806/)
83. A. Hartmann *et al.*, Validation of microarray-based resequencing of 93 worldwide mitochondrial genomes. *Hum. Mutat.* **30**, 115–122 (2009). doi: [10.1002/humu.20816](https://doi.org/10.1002/humu.20816); pmid: [18623076](https://pubmed.ncbi.nlm.nih.gov/18623076/)
84. R. S. Just, T. M. Diegoli, J. L. Saunier, J. A. Irwin, T. J. Parsons, Complete mitochondrial genome sequences for 265 African American and U.S. “Hispanic” individuals. *Forensic Sci. Int. Genet.* **2**, e45–e48 (2008). doi: [10.1016/j.fsigen.2007.12.001](https://doi.org/10.1016/j.fsigen.2007.12.001); pmid: [19083815](https://pubmed.ncbi.nlm.nih.gov/19083815/)
85. M. Ingman, H. Kaessmann, S. Pääbo, U. Gyllensten, Mitochondrial genome variation and the origin of modern humans. *Nature* **408**, 708–713 (2000). doi: [10.1038/35047064](https://doi.org/10.1038/35047064); pmid: [11130070](https://pubmed.ncbi.nlm.nih.gov/11130070/)
86. S. Kumar, G. Stecher, M. Li, C. Knyaz, K. Tamura, MEGA X: Molecular evolutionary genetics analysis across computing platforms. *Mol. Biol. Evol.* **35**, 1547–1549 (2018). doi: [10.1093/molbev/msy096](https://doi.org/10.1093/molbev/msy096); pmid: [29722887](https://pubmed.ncbi.nlm.nih.gov/29722887/)
87. J. W. Leigh, D. Bryant, POPART: Full-feature software for haplotype network construction. *Methods Ecol. Evol.* **6**, 1110–1116 (2015). doi: [10.1111/2041-210X.12410](https://doi.org/10.1111/2041-210X.12410)
88. J. Heled, A. J. Drummond, Bayesian inference of population size history from multiple loci. *BMC Evol. Biol.* **8**, 289 (2008). doi: [10.1186/1471-2148-8-289](https://doi.org/10.1186/1471-2148-8-289); pmid: [18947398](https://pubmed.ncbi.nlm.nih.gov/18947398/)
89. R. Bouckaert *et al.*, BEAST 2: A software platform for Bayesian evolutionary analysis. *PLoS Comput. Biol.* **10**, e1003537 (2014). doi: [10.1371/journal.pcbi.1003537](https://doi.org/10.1371/journal.pcbi.1003537); pmid: [24722319](https://pubmed.ncbi.nlm.nih.gov/24722319/)
90. R. Lanfear, P. B. Frandsen, A. M. Wright, T. Senfeld, B. Calcott, PartitionFinder 2: New Methods for Selecting Partitioned Models of Evolution for Molecular and Morphological Phylogenetic Analyses. *Mol. Biol. Evol.* **34**, 772–773 (2017). doi: [10.1093/molbev/msw260](https://doi.org/10.1093/molbev/msw260); pmid: [28013191](https://pubmed.ncbi.nlm.nih.gov/28013191/)
91. P. Soares *et al.*, Correcting for purifying selection: An improved human mitochondrial molecular clock. *Am. J. Hum. Genet.* **84**, 740–759 (2009). doi: [10.1016/j.ajhg.2009.05.001](https://doi.org/10.1016/j.ajhg.2009.05.001); pmid: [19500773](https://pubmed.ncbi.nlm.nih.gov/19500773/)
92. P. Skoglund, J. Storå, A. Götherström, M. Jakobsson, Accurate sex identification of ancient human remains using DNA shotgun sequencing. *J. Archaeol. Sci.* **40**, 4477–4482 (2013). doi: [10.1016/j.jas.2013.07.004](https://doi.org/10.1016/j.jas.2013.07.004)
93. Y. Huang, H. Ringbauer, hapCon: Estimating contamination of ancient genomes by copying from reference haplotypes. *Bioinformatics* **38**, 3768–3777 (2022). doi: [10.1093/bioinformatics/btac390](https://doi.org/10.1093/bioinformatics/btac390); pmid: [35695771](https://pubmed.ncbi.nlm.nih.gov/35695771/)
94. The 1000 Genomes Project Consortium, D. G. Poznik, et al., Punctuated bursts in human male demography inferred from 1,244 worldwide Y-chromosome sequences. *Nature Genetics* **48**, 593–599 (2016).
95. G. R. Abecasis *et al.*, An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012). doi: [10.1038/nature11632](https://doi.org/10.1038/nature11632); pmid: [23128226](https://pubmed.ncbi.nlm.nih.gov/23128226/)
96. S. Purcell *et al.*, PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007). doi: [10.1086/519795](https://doi.org/10.1086/519795); pmid: [17701901](https://pubmed.ncbi.nlm.nih.gov/17701901/)
97. B. K. Maples, S. Gravel, E. E. Kenny, C. D. Bustamante, RFMix: A discriminative modeling approach for rapid and robust local-ancestry inference. *Am. J. Hum. Genet.* **93**, 278–288 (2013). doi: [10.1016/j.ajhg.2013.06.020](https://doi.org/10.1016/j.ajhg.2013.06.020); pmid: [23910464](https://pubmed.ncbi.nlm.nih.gov/23910464/)
98. M. Rasmussen *et al.*, Ancient human genome sequence of an extinct Palaeo-Eskimo. *Nature* **463**, 757–762 (2010). doi: [10.1038/nature08835](https://doi.org/10.1038/nature08835); pmid: [20148029](https://pubmed.ncbi.nlm.nih.gov/20148029/)
99. P. Flegontov *et al.*, Palaeo-Eskimo genetic ancestry and the peopling of Chukotka and North America. *Nature* **570**, 236–240 (2019). doi: [10.1038/s41586-019-1251-y](https://doi.org/10.1038/s41586-019-1251-y); pmid: [31168094](https://pubmed.ncbi.nlm.nih.gov/31168094/)
100. C. Posth *et al.*, Reconstructing the Deep Population History of Central and South America. *Cell* **175**, 1185–1197.e22 (2018). doi: [10.1016/j.cell.2018.10.027](https://doi.org/10.1016/j.cell.2018.10.027); pmid: [30415837](https://pubmed.ncbi.nlm.nih.gov/30415837/)
101. C. de la Fuente *et al.*, Genomic insights into the origin and diversification of late maritime hunter-gatherers from the Chilean Patagonia. *Proc. Natl. Acad. Sci. U.S.A.* **115**, E4006–E4012 (2018). doi: [10.1073/pnas.1715688115](https://doi.org/10.1073/pnas.1715688115); pmid: [29632188](https://pubmed.ncbi.nlm.nih.gov/29632188/)
102. J. M. Monroy Kuhn, M. Jakobsson, T. Günther, Estimating genetic kin relationships in prehistoric populations. *PLoS ONE* **13**, e0195491 (2018). doi: [10.1371/journal.pone.0195491](https://doi.org/10.1371/journal.pone.0195491); pmid: [29684051](https://pubmed.ncbi.nlm.nih.gov/29684051/)
103. A. A. Behr, K. Z. Liu, G. Liu-Fang, P. Nakka, S. Ramachandran, pong: Fast analysis and visualization of latent clusters in population genetic data. *Bioinformatics* **32**, 2817–2823 (2016). doi: [10.1093/bioinformatics/btw327](https://doi.org/10.1093/bioinformatics/btw327); pmid: [27283948](https://pubmed.ncbi.nlm.nih.gov/27283948/)
104. M. Nei, *Molecular Evolutionary Genetics* (Columbia University Press, New York, 1987).
105. K. Nägele *et al.*, Genomic insights into the early peopling of the Caribbean. *Science* **369**, 456–460 (2020). doi: [10.1126/science.aba8697](https://doi.org/10.1126/science.aba8697); pmid: [32499399](https://pubmed.ncbi.nlm.nih.gov/32499399/)

ACKNOWLEDGMENTS

We greatly acknowledge the IT support of L. A. A. Bautista, A. de León Cuevas, C. S. F. Bautista, and J. G. Sotelo of the Laboratorio Nacional de Visualización Científica Avanzada (LAVIS UNAM); A. C. Carbajal and C. U. Díaz from LIIGH-UNAM for wet-lab technical support during the project; A. Sockell for technical support at the Bustamante Lab; R. M. Duarte for his early contribution in the Cañada de la Virgen project; P. S. Fernández for help in the collection of samples from Cañada de la Virgen; L. Correa-Mendoza for technical support with qpWave; D. Ortega-Del Vecchyo and A. G. Oliver for enriching comments throughout the project on earlier versions of the manuscripts; and Y. J. Islas for her help in the summary figure. **Funding:** V.V.-I. is a

doctoral student from the Programa de Doctorado en Ciencias Biomédicas, Universidad Nacional Autónoma de México (UNAM), and has received the CONACyT fellowship (número de becaria 282188). M.C.A.-A. received support from UNAM-PAPIIT (project nos. IA201219 and IA203821), the Wellcome Trust Seed Award in science (grant no. 208934/Z/17/Z), and the International Centre for Genetic Engineering and Biotechnology (project no. CRP/MEX10-03). M.C.A.-A., E.H.-S., and F.J. received support from the Human Frontiers Science Program (grant no. RGY0075/201). R.F. was supported by a grant from Fondo Europeo de Desarrollo Regional (FEDER)/Ministerio de Ciencia e Innovación—Agencia Estatal de Investigación (PALEUCOL; PGC2018-094101-A-I00) and from CajaCanarias and La Caixa Banking foundations (GENPAC; 2018PATRI16). Data generation for the Sierra Tarahumara individuals was funded with a Trainee Research Grant to M.C.A.-A. by Stanford Center for Computational, Evolutionary, and Human Genomics. E.H.-S. received support from NIH R35GM128946 and the Alfred P. Sloan Foundation. Data generation for the Cañada de la Virgen individuals was supported by The Mexican Biobank Project (grants FONCICYT/50/2016 and MR/N028937/1 from CONACyT and Newton Fund, respectively) awarded to A.M.-E. This work has been partially supported by the government of Guanajuato (grant no. 093/2016 CINVESTAV) awarded to K.S. Data generation for the Aridoamerican individuals was supported in part by a Leakey Foundation grant to F.A.V. Additionally, M.L. received Swedish Research Council grant 2020-04789. **Author contributions:** M.C.A.-A. and V.V.-I. conceived and designed the study with input from E.M.P.C. M.S.-V., M.B.-L., B.M., R.F., K.S., and M.A.N.-C. performed the wetlab work. V.V.-I. processed all sequence data. V.V.-I. carried out most analyses with help from, A.I.-G., M.L., J.M.T., J.E.R.-R., D.A.V.-R., and F.S.-Q. E.M.P.C., A.H.-M., K.S., G.Z.G.M., E.G.-M., R.A.-H., and C.V. contributed samples, provided the archaeological context, and discussed results. F.A.V., A.M.-E., F.J., A.G.I., E.H.-S., J.V.M.-M., and F.S.-Q. suggested analyses, discussed results, and provided feedback throughout the project. V.V.-I. and M.C.A.-A. wrote and edited the manuscript with input from all authors. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** Data are available from the European Nucleotide Archive under project accession PRJEB51440. **License information:** Copyright © 2023 the authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original US government works. <https://www.science.org/about/science-licenses-journal-article-reuse>

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.add6142
Supplementary Text
Figs. S1 to S33
References (106–125)

Submitted 25 June 2022; accepted 17 March 2023
10.1126/science.add6142



Demographic history and genetic structure in pre-Hispanic Central Mexico

Viridiana Villa-Islas, Alan Izarraras-Gomez, Maximilian Larena, Elizabeth Meja Perez Campos, Marcela Sandoval-Velasco, Juan Esteban Rodriguez-Rodriguez, Miriam Bravo-Lopez, Barbara Moguel, Rosa Fregel, Ernesto Garfias-Morales, Jazeps Medina Tretmanis, David Alberto Velzquez-Ramrez, Alberto Herrera-Muz, Karla Sandoval, Maria A. Nieves-Coln, Gabriela Zepeda Garca Moreno, Fernando A. Villanea, Eugenia Fernandez Villanueva Medina, Ramiro Aguayo-Haro, Cristina Valdiosera, Alexander G. Ioannidis, Andrs Moreno-Estrada, Flora Jay, Emilia Huerta-Sanchez, J. Vctor Moreno-Mayar, Federico Snchez-Quinto, and Mara C. vila-Arcos

Science, **380** (6645), eadd6142.
DOI: 10.1126/science.add6142

Editor's summary

Ancient DNA has revealed much about the movement and genetic legacy of human populations. However, DNA is notoriously sensitive to degradation in hot climates, resulting in a greatly imbalanced study of regions and ancestries because of climatological restrictions. Villa-Islas *et al.* analyzed ancient DNA from pre-Hispanic individuals in northern and central Mexico, revealing contributions from an unknown “ghost” population (see the Perspective by Llamas and Roca-Rada). They found that some populations endured in regions of central Mexico despite changing climate beginning in the 10th century. Much of the ancient population’s genetic structure has withstood population declines and is represented in present-day Indigenous populations. —Corinne Simonti

View the article online

<https://www.science.org/doi/10.1126/science.add6142>

Permissions

<https://www.science.org/help/reprints-and-permissions>

Use of this article is subject to the [Terms of service](#)

Science (ISSN) is published by the American Association for the Advancement of Science. 1200 New York Avenue NW, Washington, DC 20005. The title *Science* is a registered trademark of AAAS.

Copyright © 2023 The Authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original U.S. Government Works



Supplementary Materials for

Demographic history and genetic structure in pre-Hispanic Central Mexico

Viridiana Villa-Islas *et al.*

Corresponding author: María C. Ávila-Arcos, mavila@liigh.unam.mx

Science **380**, eadd6142 (2023)
DOI: 10.1126/science.add6142

The PDF file includes:

Supplementary Text
Figs. S1 to S33
References

Other Supplementary Material for this manuscript includes the following:

Tables S1 to S18
MDAR Reproducibility Checklist

Supplementary Text

Full Ethics statement

The sampling of the archeological human remains was made upon approval by the Consejo de Arqueología (Archaeology Council) of the Instituto Nacional de Antropología e Historia (National Institute of Anthropology and History) of each collection. The permit numbers are: 401.3S.16-2017/990, 401.3S.16-2019/222 and 401.1S.3-2022/1152.

We sampled the minimum possible amount of tissue to avoid unnecessary destruction of these precious materials. For teeth, we tried to separate the root from the crown without damaging the crown. We returned all leftover material to the archeologists responsible for each collection. In addition, we took photographs of each sample during several steps of the sample processing to have a visual archive. Because of historical and socio-political reasons, together with the prominent admixture of different continental ancestries in the territory and the imposed institutionalized narrative of “mestizaje” (108, 109) to minimize the contribution of present-day Indigenous peoples to our national identity (something utterly condemnable), it becomes quite complex to assign a present-day population in Mexico as the single direct descendant of the remains of human ancestors recovered from archaeological contexts. This complexity raises several questions regarding consultation of descendant populations for this kind of research (e.g. should admixed "mestizo" populations who share the geographical area be consulted? who should be considered a representative for them? under what capacity? In part because of these uncertainties, consultation with Indigenous populations for destructive analysis of archeological remains is not a standard procedure in the country. This is on its own an emerging focus of discussion that warrants further investigation and debate beyond the scope of this research (but see Ávila-Arcos et. al 2022). However, a series of efforts to communicate genetic findings resulting from this work to local communities from the sites of Toluquilla and Ranas have taken place over the course of the project.

Sample overview

Samples include skeletal remains from Northern Mesoamerica: Sierra Gorda - Querétaro (n=13), Cañada de la Virgen - Guanajuato (n=16) and Central Mesoamerica: Michoacán (n=8); along with updated genomic data from Aridoamerica: Sierra Tarahumara – Chihuahua (n=2), previously reported at less resolution (1) (Table S1). The most extended period covered here belongs to Sierra Gorda samples from 320 BCE to 1,351 CE (Classic-Postclassic). While the rest of the archaeological sites correspond to Classic (Cañada de la Virgen and Michoacán), and Postclassic (Sierra Tarahumara).

Archaeological context

Sierra Tarahumara (Mummy-F9 and Mummy-MOM6)

These individuals and their archaeological context were previously published in (1)

Sierra Gorda, Querétaro (Mummy-P)

Cadereyta cave, Sierra Gorda, Querétaro (P_CCM_b)

In 2002, the discovery of the mummified remains of individual P_CCM_b was reported in the vicinity of the village of Altamira, in the municipality of Cadereyta, Querétaro. The remains were removed and taken to the public prosecutor's offices in the municipal capital. Once the remains were handed over to INAH and archaeologist Elizabeth Mejia Perez Campos, the site was visited and cleaned. Fragments of textiles, organic matter such as bird feathers and wood formed a bracelet, and human hair braids were found. The piece was taken to the Templo Mayor Museum in Mexico City, where computerized axial tomography, necropsy with an endoscope to see internal organs and samples were taken. Likewise, mtDNA studies were performed by PCR amplification of the HVR at the Centro de Investigación y Estudios Avanzados of the Instituto Politécnico Nacional (CINVESTAV - IPN) in Mexico City and the textile and skin were dated by C13 and C14 at Beta Analytical in Florida, USA. The results showed a twenty-months-old infant who lived in 340 BC, shrouded with a cotton blanket with human hair interwoven and with an offering of maguey quills, stork feathers, pine, oak, and sotol leaves (110).

Toluquilla, Sierra Gorda, Querétaro

Toluquilla archaeological site is located in the Sierra Gorda, in the State of Querétaro. The occupation of Toluquilla spans ca. 2,000 years, from 400 BCE to 1,550 CE.

Eight individuals, 333A_TOL_a, 333B_TOL_a, 333C_TOL_a, 333O_TOL_b, 333Q_TOL_b, 2417C_TOL_a, 2417Q_TOL_b and, 2417J_TOL_a, from the archaeological site of Toluquilla were sampled for this study. We analyzed samples from individuals from two ritual burials at the site. The burials were analyzed by Dr. Elizabeth Mejia Perez Campos as part of her doctoral thesis (6) and were named based on the relevance of the offerings. These burials correspond to the 'Manatee context', and the 'Mirror context'.

Manatee Context (Individuals 333x)

The manatee context has a great quantity, quality, and variety of offerings. It includes shells and snails from both coasts (Gulf of Mexico and the Pacific Ocean), obsidian from the Sierra de las Navajas (Hidalgo, Mexico), local animals such as river fish lynx, armadillo plates, and foreign aquatic species from the deep sea.

Individuals 333A_TOL_a, 333B_TOL_a and 333C_TOL_a 333O_TOL_b, 333Q_TOL_b from this context were sampled for this study. This context is located in building 33 for which dates are available, either by C14 dating, obsidian hydration dating, or the analysis of ceramic materials.

In this context, the more recent burial is dated between 1,300 and 1,400 CE. In this burial individuals 333A_TOL_a, 333B_TOL_a, and 333C_TOL_a were found. Charcoal was deposited among the sediment that covered the bodies and has been C14 dated to 1,350 ± 50 CE.

333A_TOL_a was a 25-30-year-old female. She had a pendant made from a lynx tusk placed as a trousseau on her chest, possibly hung with knots since it has no orifice, in addition to a necklace of Pacific olive shells and a human jaw with a perforation as a pendant. As an offering, a piece in the shape of a star or flower made with a spondylus snail, a stingray caudal spine, a natural shell, a fragment of river fish, an obsidian flake, a yew, and a bean-shaped piece with a mixture of cinnabar and red ocher were deposited under the body (*III*).

333B_TOL_a was an infant individual; he was found carefully arranged, leaving the arm in almost total order, first the humerus, radius, and ulna. However, the latter was placed upside down, and in the place of the wrist, a necklace was placed as a trousseau. The necklace comprises four pieces that form a snake, the first of a cylindrical clay bead, another similar one of bone, a flat and square bead of the mother-of-pearl shell to finish with an elongated piece also of mother-of-pearl shell, which represents the rattlesnake's rattle. It is deduced that this burial is of ritual type, since the coxal bones of the child were placed around the circle, and as an offering, there is a complete shell, an earring made with the incisor of a six or seven-year-old child, and a necklace of shells.

Individual 333C_TOL_a was a woman between 20 and 21 years old. A number of pieces were found with her. Her trousseau was a necklace of 1,150 beads and three snail earrings. Also, as offerings on the body there was a button-like circular piece, an armadillo plaque, an obsidian flake, shuffleboard or ceramic circle, and, next to her, a pectoral made on a manatee rib, which gives its name to this archaeological context.

Other older burials were found in this building. Individual 333O_TOL_b was found next to the platform wall, C14 dated to 700 ± 50 CE. Individual 333Q_TOL_b was found in an older burial C14 dated to 520 ± 40 CE. The platform was to the west of the altar, and the burial was 50 cm deep, it corresponds to the first phase of occupation (*III*).

Mirror context (Individuals 2417x)

Individuals 2417Q_TOL_b, 2417C_TOL_a and 2417J_TOL_a from this context were sampled for this study. This archaeological context is composed of sixteen individuals; it is located in front of the altar of building 24. Among the fill, scattered bones of additional thirty-five people were found. The bones in the fill and the closeness of the primary individuals suggest that the deposition of the primary bodies was at a single time, occupying the bones of previous individuals as fill.

A total of four levels of human body deposition were observed. There is at least one primary body surrounded by secondary bodies, remnants of previous burials, and filler bones in each level. In the first and most superficial level, in front of the altar, a group formed by a woman's primary body (2417C_TOL_a) and the secondary remains of three children (not analyzed in this study) were placed. The remains form a U shape, with the open portion pointing west. The bones of the woman 2417C_TOL_a seem to have received scalding liquids or some other substance, which produces an effect similar to boiling the bones, as they appear crystallized, although additional studies are required to define what was the precise taphonomic process. In addition, she presented pathologies such as spondylitis (the fusion of the last vertebra and the sacrum) and poor dental health, orbital sieve and marks at the insertions of the arms. The woman had as trousseau a necklace formed by

sixteen river mollusks of the genus *Unio*, and as offerings a snail button and obsidian flake and razor. The offering of individual 2417C_TOL_a is significant for its quality, variety, abundance, and treatment given to the corpse.

Five individuals were found in the third level of the northeast of the assemblage. The most relevant is individual 2417J_TOL_a, who presents pathologies such as spondylitis, evident in the sacrum, iliac and lumbar vertebrae. In addition to poor dental health, he also performed exercises that developed his muscles and left a mark on his arms and legs.

In the fourth and last level, the remains of four more individuals were located, all adults, two secondary, and two remnant primary portions. Individual 2417Q_TOL_b corresponds to a remnant primary portion that was placed in the ventral decubitus position (face down) of a male between forty-five and fifty-five years old. Under individual 2417Q_TOL_b, there were two isolated bones, a coxal, and a vertebrae. Also under this individual, a pyrite mosaic mounted on a ceramic base was found, thought to serve as a mirror. The mirror is a significant offering; it is related to the priests and the god Tezcatlipoca, the deity of the underworld and the night, characterized because it carries a mirror, which is also used in the arts of divination, to obtain omens and to provoke evils (*III*).

Individual 6428A_TOL_b

The body of individual 6428A_TOL_b was found in building 64, being part of burial 28. It corresponds to an adult male in a flexed dorsal decubitus position with a significant offering including a flint bifacial, two jade beads, and ten amulets carved in bone (*III*).

Ranas, Sierra Gorda, Querétaro

Ranas archaeological site is located in the Sierra Gorda, in the State of Querétaro, and the occupation of the site is recorded from 400 BCE to 1,300 CE.

Individual 11R_R_b

Individual 11R_R_b was found in a burial that contained three individuals, one primary (11R_R_b) and two secondary arranged on the sides of the primary. 11R_R_b corresponds to an adult male over 35 years old. He was found deposited in dorsal decubitus, with the trunk slightly flexed upwards. The arms were placed in front of the body and the wrists were touching each other. His hands were on the pelvic region and in the right hand he was holding a ceramic pipe. Two bone circles forming a truncated cone were found in the area of the pubic symphysis. He presented osteological characteristics of repetitive work and high stress to the body. The enteropathies in the long bones of the legs, arm, vertebrae and calcaneus may be directly related to mining. Thus, individual 11R_R_b is inferred to have worked in mining for several years. The measurement of high values of heavy metals in a combination of Pb, Hg, As and Sb indicate chronic exposure to a wide variety of heavy metals present in geological contexts, which were systematically deposited in the bones of the individual (*55*).

The relative chronology of the excavated materials and those directly associated with a more superficial burial context suggested a chronology between 450 and 700 CE. In terms of

Mesoamerican periodicity, it corresponds to the transition between the Classic and Epiclassic periods (55).

Individuals 7A_R_b and 3A7I_R_b were found in structure 3A, in burial 7 at the archaeological site of Ranas. Burial 7 is a complex burial, where initially five areas of intrusion are defined, and inside each of them, there is evidence of skeletal remains of one or more individuals.

Individual 7A_R_b

This individual is the most important since it is located just below the altar. It is a direct primary burial with a very flexed seated disposition. The individual was deposited as a mortuary bundle strongly tied (at the height of the shoulders and waist), so when the body decomposed in its taphonomic process, the head fell between the legs. The care taken in the deposition and the fact that limestone slabs covered the individual from its lowest level helped to keep most of the skeleton in its primary position, where the only element of post-depositional alteration was a rodent burrow that altered the metatarsal bones and phalanges of the left foot and altered the distal phalanges of the right foot.

Individual 3A7I_R_b

The case of individual 3A7I_R_b is another burial of an important dignitary since it is the only one deposited in an extended form with the head to the south, its face looking to the east, with the right arm flexed on its chest. The hand was deposited on its left shoulder, while its left arm rested on the belly, and its hand is extended between the right ribs and the right humerus, with the torso slightly leaning forward. The pelvic region was horizontal on the ground while the legs were separated with the right knee to the east, the left knee to the west, and the feet under the coxae. Therefore, this is a primary burial, in dorsal decubitus slightly raised from the back and head, with legs flexed on themselves. As it is a primary individual, the different body portions could be recognized and recovered, although in many cases, they were found in regular and poor preservation.

Cañada de la Virgen, Guanajuato

Individual E2_CdV_b

Individual E2_CdV_b is a direct secondary burial, located in layer III of Table V-4 (although it partially intrudes up to layer IV). The burial presents only proximal fragments of incomplete femurs corresponding to an adult individual of undetermined age. Although it is possible that its partial intrusion into floor 1 and layer IV is accidental, the only difference with respect to the rest of the skeletons from Room 3 is the fact that it does not have contact with floor 2, but is found right at the level of floor 1. It seems more likely that this burial, due to its position and secondary character, functioned as a ritual seal for floor 2, in a termination ritual aimed at "killing the funerary space between the two floors (112).

Individual E4_CdV_b

Located in layer IV of Table X-4. It is an indirect primary burial deposited in a cist in extended dorsal decubitus facing east. The estimated age is 5 years +/- 24 months and indeterminate sex. It presents 60% of the skeleton with absence of bones of the feet, patella, sacrum and coccyx (112).

Individual E6 CdV b

Located in layer IV of Table V-3. Indirect secondary burial in cist, corresponding to a male individual of 40-45 years of age. It presents only fragmented iliacs, complete right radius, proximal part of both femurs at 50% and part of the proximal portion of the right tibia. Position and orientation not discernible (112).

Individual E7 CdV b

Located in layer IV of Table X-3. Indirect primary burial deposited in cist in dorsal decubitus flexed oriented to the east. Corresponds to a female of 25 to 30 years of age. Feet and hands are missing (112).

Individual E8 CdV b

Located in layer IV of table U-3. Direct primary burial in flexed dorsal decubitus oriented to the east. It corresponds to a male individual between 35 and 40 years old. He presents 90% of the skeleton with absence of feet (112).

Individual E9 CdV b

Located in layer IV of Table T-4. It is an indirect primary burial deposited in flexed dorsal decubitus with orientation to the east. It corresponds to a male subject of 20 to 25 years of age. It presents 80% of the skeleton with absence of feet (112).

Individual E10 CdV b

Located in layer IV of box V-3. This is a direct secondary burial with no discernible anatomical position. It consists only of a left femur diaphysis and a mandible (112).

Individual E11 CdV b

Located in layer IV of squares U/V-3 and 4. It is a direct secondary burial formed by three fragments of long bones, fragments of ribs, a clavicle, some phalanges, fragments of pelvis and skull, as well as some molars (112).

Individual E14 CdV b

Located in complex C, which is a circular structure, with an association to the wind and related to the North Star and Ursa Minor, which is linked to Tezcatlipoca. This burial was found at the foot of the access ramp to the second structure (112).

Individual E15_CdV_b

This individual was called the "girl of the rain". This subadult individual, through osteological method was estimated to be approximately 7 years old and possible a female, but we take this with caution since the individual is a subadult. She was found in complex B, and linked to the pluvial drainage, which is large and well-constructed with superimposed slabs. She was 1.10 m tall and is thought to be part of the ancestral veneration practiced in the place, since she was accompanied by a coyote and placed in such a way that it was necessary to decompose her body in order to accommodate her seated and compressed. Possible postmortem treatment in the temazcal that is at the top of the pyramidal base. The "girl of the rain", carried in her neck a small necklace of beads that includes a bead in the form of butterflies in the central part of the necklace. The offering is completed with a ceramic spindle and a fragment of marine shell. Next to the girl and as a companion and vestige of the practices of ancestral veneration, a small coyote was placed - a complete specimen of *Canis latrans*, an animal of totemic importance in the region. Forensic studies performed on the girl of the rain provided her height, and some pathologies associated with infectious processes such as periostitis or infection of the periosteum, as well as cranial lesions probably diagnostic of tuberculosis (112).

Individual E18_CdV_b

"The decapitated one". Part of his burial offering consisted of an elongated handle smoker. He was found associated with the enclosed patio located on the elevated platform on one side of the pyramidal base of Complex B. The individual is male, aged between 20 and 30 years old and with an unusual height of 1.73 centimeters - the left femur itself measured 45 centimeters. Forensic analysis indicates evidence of disease, specifically ankylosing arthritis, as well as apparent postmortem decapitation (112).

Individual E19_CdV_b

According to the skeleton, the burial of E19_CdV_b corresponds to a female individual, 20-25 years old. It was identified as an indirect primary burial in a dorsal decubitus position (slightly to the right) flexed. This burial presented a series of alterations regarding the original position, and the hands were absent. Because there are no marks of violence or dismemberment, archaeologists suggest that the most likely hypothesis is that these elements were removed during some looting and were subsequently lost by agricultural work or by the action of rodents. The remains of this burial were found under a layer of killed pottery and associated remains of a canid (*Canis familiaris*) (113).

Michoacán

Eight individuals were retrieved from the state of Michoacán. Individuals E1A_Mich_b, E2_Mich_b and E4_Mich_b from the site La Mina; individual M1_Mich from the site Tanhuato and Individuals M2_Mich_b, M3_Mich_b, M4_Mich_b and M5_Mich_b from the site Zaragoza.

La Mina, Michoacán

At the end of December 2014, the INAH Michoacán Center was notified of the discovery of a series of skeletons located in the construction of a perimeter wall in the Telesecundaria #133 school, in the town of La Mina, municipality of Álvaro Obregón, Michoacán.

Upon arriving at Telesecundaria School #133, the archaeologists detected that the funerary context had been seriously altered due to the construction of a trench for the foundation of the perimeter wall. From the initial reports obtained from the masons, it was possible to determine that the bodies were concentrated in the same sector. During the first visit, towards the end of December 2014, a collection of those osteological elements extracted due to the construction work was carried out.

So far, the extension of the site of La Mina is unknown due to the lack of monumental architecture and the growth of the current settlement, which occupies much of the area with vestiges. Excavations have shown that La Mina appears to be on what was once the shores of a village founded on the southwestern shore of Lake Cuitzeo.

The absence of architectural elements on the surface, in the intervened area, and the surroundings, suggests an architecture based on perishable materials. The bioarchaeological remains were located in the northern sector of an embankment covering an area of approximately 1200 m². The ceramic materials on the surface indicate that the settlement was occupied between 200 B.C. and 900 A.D. Although this data indicates an extensive temporal extension, the stratigraphic unit indicates a single moment of occupation. The excavations carried out at this site were only aimed at recovering the bioarchaeological materials affected by the construction work, so the archaeologists do not rule out the existence of more burials at the site.

E1A_Mich_b

Individual E1A_Mich_b was in a poor state of preservation. Only part of the lower limbs, two-thirds of the mandible, and the upper dental arcade were located. This individual corresponded to a male young adult placed in a right lateral decubitus position and was oriented in a west-east direction.

The grave goods consisted of a vessel of the Paso Ancho/graffito type placed at the level of the lumbar vertebrae and showed clear traces of use. Two reused sherds deposited in front of the face completed the offering. The intentional placement of this ceramic fragment leads us to suppose that it is an object used in the development of the funerary ritual and later placed as part of the trousseau. These two ceramic fragments were dated together with another sample from another burial using the archaeomagnetic technique. The results indicated that the burial corresponded to 647-825 CE (114) during the so-called Epiclassic period. A third element that made up the offering consisted of a dog in its juvenile development phase. Despite its poor state of preservation, it was possible to identify that the animal was not more than eight months old, considering the process of ossification of the epiphyses (115).

E2_Mich_b

Individual E2_Mich_b corresponds to an early infant who was approximately six months old. This infant did not present any artifact or other element that could be identified as part of an offering.

The infant was buried at a greater depth than the rest of the individuals, approximately 1m, even surpassing the tezontle layer that formed the leveling embankment.

E4_Mich_b

The individual E4_Mich_b corresponds to a skull collected at the site without context because it was removed by personnel building the school's perimeter fence. This skull was placed next to the burial of individual E1A_Mich_b.

Tanhuato, Michoacán

M1_Mich corresponds to the burial excavated in a rescue excavation in the town of Tanhuato, Michoacán, during the construction of a canal in a street inside the town. Tanhuato is located in the municipality of the same name in the state of Michoacán. The burial was found at 1,543 meters above sea level without a greater context, since it is located in an urbanized area. The surrounding area is almost flat and there is a large amount of cultivated land. The burial was mixed, and the inhumations were made in an enclosure dug in the soil matrix, whose bottom is one meter deep with respect to the level of the street; it covers an area of about 5 m². Primary and secondary burials were found in the pit and two incomplete individuals, two ossuaries and a funerary pot could be identified. Unfortunately, both bones and materials had already been removed by construction workers, so the context was very altered. Absolute dates for this individual are not yet available.

Zaragoza, Michoacán

The Zaragoza site is located in the municipality of La Piedad, on a flat-topped, isolated elevation called Mesa Acuitzio. This site presents a large rocky front that the settlers of Zaragoza took advantage of as a support for their settlement. The monumental area of the site was built on a relatively flat part located more or less halfway up the mesa, around 1,753 meters above sea level, while the living and cultivation area is distributed downwards, both to the east and north, along the slopes through artificial terraces.

Samples M2_Mich_b and M3_Mich_b come from a mixed burial, excavated in 2004 by Anyul Cuellar and Diana Bustos, called Burial 1. The space was used in successive events, that is, not all the bodies were deposited there simultaneously. It is located outside, on the east side, of an architectural unit that we have identified as a possible temascal (low heat sweat lodge), located at the southern end of one of the main group of structures. Among the buildings that make up this complex, a ball court stands out, almost 57 meters long (including the headers) and 14 meters wide.

A total of 14 individuals were located in the funerary context, identified from the number of skulls, i.e., not necessarily complete bodies were found. The remains were in a very poor state of preservation, but at least 5 primary deposits were identified. M2_Mich_b comes from individual 7, located at the western end of the pit, next to the wall that delimits the east side of the temascal. M3_Mich_b corresponds to a second molar of skull 13. Both skulls were found without anatomical connection, although they were part of the group of bodies deposited in the burial space.

M4_Mich_b and M5_Mich_b, also from the Zaragoza site, but from Burial 2, were also mixed and excavated in 2012 by Alfredo Salas. This burial is located to the south of structure 1, which is part of the same complex as the court, to the east of it. The pit, of 24 m², presents a total of 8 or 9 individuals, which are not specified yet. The samples correspond to a first molar of M4_Mich_b and a third molar of M5_Mich_b, respectively. Individual M4_Mich_b belongs to an individual placed in extended dorsal decubitus with a south-north axis, slightly deviated from 6 to 8 degrees to the east; it is important to note that this is the only individual with this orientation inside the pit. This individual had an associated offering composed of a cajete (ceramic bowl), a pot and a "Capiral" lid. Individual M5_Mich_b was the most complete skeleton, although in a poor state of preservation; M5_Mich_b was found in an extended dorsal decubitus position, with arms flexed and with an east-west orientation with a 287° deviation with respect to north.

Extended Results

Sample selection and sequencing strategy

We screened the DNA content of archeological samples from 39 pre-Hispanic individuals (Table S1). The central Mexico sites include three in the Sierra Gorda (SG) in Querétaro state, namely Toluquilla (TOL, n=9), Ranas (R, n=3) and a cave in Cadereyta de Montes (CCM, n=1); Cañada de la Virgen (CdV) in Guanajuato state (n=16); and the Zaragoza (n=4), Tanhuato (n=1) and La Mina (n=3) sites in Michoacán state (Mich). Furthermore, we produced additional genome-wide data for two mummies from the Sierra Tarahumara in northern Mexico (Table S1), which were sequenced at lower depths in a previous study (14) (Fig. 1).

For each sample, we evaluated the extent of endogenous ancient DNA (aDNA) via shotgun sequencing. Out of the 39 samples, only three had endogenous content above 10% (13 – 39%) (Table S2). We then generated additional sequence data for these three samples using shotgun sequencing, while for the remaining libraries, we performed whole-genome or mitochondrial DNA (mtDNA) capture (116, 117) (Table S2). For all of these, we verified the presence of the characteristic misincorporation patterns of aDNA using mapDamage (69) (Fig. S1-S4). We obtained informative amounts of genome-wide data for 12 individuals (0.01 – 4.7 x). Ten were from Mesoamerica: Sierra Gorda (n=4), Cañada de la Virgen (n=4), and Michoacán (n=2); and two from the Sierra Tarahumara in Aridoamerica (Table S1). We integrated these data with previously reported low-depth genomes (0.09 – 0.3 x) from three ancient Pericúes from Aridoamerica reported in (14, 15), resulting in a total of 15 pre-Hispanic individuals from Mexico for which we had genome-wide data. Our analysis also included ancient genomic data from previous studies for 21 individuals from across the Americas (Table S3)(14, 15, 45, 46, 98–102). Additionally, we reconstructed the mitochondrial genomes for 27 individuals across Mexico (3.2 – 1284.8x), spanning 1,600 years (320 BCE to 1,351 CE). Of these, 25 were from Mesoamerica: SG (n=13), CdV (n=7), and Michoacán (n=5); and two individuals were from Aridoamerica: Sierra Tarahumara (n=2) (14) (Table S1 and S2). These mitochondrial genomes were analyzed along with previously published mitochondrial genomes from Mexico and South America (14, 18, 25, 53, 77–79, 81–85, 101, 118) (14, 15, 29) (Table S4 and S6).

Mitochondrial haplogroup structure and diversity

We observed mitochondrial DNA (mtDNA) haplogroup C to be predominant in Aridoamerica in the Sierra Tarahumara, with decreasing frequency toward Mesoamerica. This gradient has been

previously observed in ancient and contemporary populations from Mexico (18, 25–31). Specifically for central Mexico, we found the presence of all four autochthonous haplogroups, A, B, C, and D, similar to previous observations on ancient and present-day populations inhabiting this region (19, 26, 27, 49, 119, 120). This highlights the greater genetic diversity and heterogeneity in this particular area (43, 50), and is in agreement with larger Pre-Hispanic agriculture-based civilizations and greater population mobility compared to Aridoamerica. Furthermore, we detected unreported variants in the mtDNA PhyloTree (32) for seven individuals assigned to the sub-haplogroups A2d (n=4), B2c (n=2), and B2I (n=1). The absence of these variants could reflect a poor sampling of Native American mitochondrial genomes in the PhyloTree database, or a loss of these haplotypes because of drift. Accordingly, we propose adding these new unreported variants to the PhyloTree database (Table S5).

In Sierra Tarahumara, we found two sub-lineages of the mtDNA haplogroup C (C 100%), as previously reported. In contrast, Central Mexico shows a higher diversity of mtDNA haplogroups. Cañada de la Virgen was the site with the highest number of sub-haplogroups, distributed in B 42.3%, A 28.6%, C 14.3%, and D 14.3%. Michoacán presented four different sub-haplogroups (A 40%, D 40%, and C 20%). Sierra Gorda also presented four different sub-haplogroups but corresponding to major lineages A 46.15%, B 46.15%, and D 7.7%, respectively.

In Sierra Gorda, out of six A2d individuals four belong to the pre-drought period according to dating based either on C14 or the archaeological context, and two to the drought period. The haplotype median-joining network for the haplogroup A2 shows a clustering of five Sierra Gorda individuals carrying the sub-haplogroup A2d (the sixth A2d individual, 333Q_TOLb, was not included in the network as it did not pass the quality filters for inclusion) despite a time transect of 1,480 years (Fig. 2A). One of the five individuals, 37A1_R_b, harbors several polymorphisms with respect to the other four. Unfortunately, this individual barely passed our filtering criteria (high presence of ‘N’s’ and excess C > T mutations). Because haplotype networks can be biased by sequence uncertainty, most of this individual's polymorphisms are likely from sequencing errors and/or ancient damage. This makes this individual difficult to place and compare with confidence. However, the other four individuals (7A_R_b, 2417C_TOL_a, 2417J_TOL_a and P_CCM_b) form a unique clade with each other showing genetic continuity. The clade that they form is striking given their high sequence similarity, location and time span. This observation supports a continuity of this sub-haplogroup, despite the severe droughts between 900 – 1,300 CE in this region of the Northern frontier of Mesoamerica. In contrast, the haplotype network for haplogroup B, shows three pre-drought individuals sharing sub-haplogroup B2I, while the two post-drought are found in the same node representing sub-haplogroup B2c. No individuals from Sierra Gorda carried C haplogroups and only one individual (11R_R_b) had a D sub-haplogroup (D1) (Fig. S5).

We estimated the diversity of the mitochondrial genomes per pre-Hispanic site and period using the 27 mtDNA genomes, along with eight public mitochondrial genomes from Mexico (Table S4). We calculated nucleotide diversity π through the genetic distance method of Tajima and Nei, with the software Arlequin v. 3.5.1.3 (75). As expected, we find that the pre-Hispanic populations from Mesoamerica have higher π values (Sierra Gorda pre-drought: 0.003022, Sierra Gorda post-drought: 0.003219, Cañada de la Virgen: 0.005609, Michoacán: 0.004394) than those from Aridoamerica (Sierra Tarahumara: 0.001895, Cueva Candelaria: 0.000413, Paquimé: 0.002118).

Past female effective population size inference

To estimate the past female effective population size, we merged our dataset of pre-Hispanic mtDNA sequences from (n=13, selected after alignments, see methods) with 164 publicly available ancient (n=8) (Table S4), modern (n=232) mtDNA sequences from Mexico and ancient from South America (n= 129) (Table S6).

When performing EBSP on the Pre-Hispanic mitogenome data, we faced a lack of resolution due to the small sample size. However, when using present-day mitogenomes from Mexico, the EBSP showed a marked population expansion ca. 5 kybp likely related to the domestication and propagation of maize cultivation in Mexico (51). Notably, we were not able to recapitulate with the modern mitogenome data the drastic population bottleneck experienced ca. 500 ybp after European colonization. Some possible explanations are: 1) Lack of power to infer coalescent events in recent times, 2) Need of a larger sample size, 3) European colonization affected the female lineage in a lower extent than in other regions in the Americas, 4) a combination of all the three. This is intriguing considering previous reports of massive Indigenous population decline in New Spain during European conquest (121, 122). However, these estimates were based primarily on historical accounts of the early colonial Mexican population through taxation lists of most settlements across Mexico but not including all of them (122). Our results suggest that the demographic impact of European colonization on the female population was variable across Mexico, and less drastic than in insular and other isolated regions of the Americas where a bottleneck in ca. 500 ybp is clearly observed in the EBSPs (52, 53, 54). When using mtDNA from South America in the EBSP we could not improve the resolution at the haplogroup level, except for haplogroup C, in which we could observe a population decline ca. 5 kybp., in agreement with a previous study analyzing haplogroup C in South America.

Sex determination and Y chromosome haplogroups

We could assign a sex for 16/39 individuals (females=3, males= 13). Determination of biological sex was made using the tool reported in (92). This approach computes the proportion of reads mapped to the Y chromosome with respect to the reads mapping to the X chromosome and Y chromosomes (R_y). According to the method, $R_y > 0.075$ corresponds to XY, while $R_y < 0.016$ corresponds to XX. For another four individuals, although the R_y values were above the required threshold for XX assignment, they were very close (E2_Mich_b: 0.0218, E4_Mich_b: 0.022, E7_CdV_b: 0.0339 and E19_CdV_b: 0.0312). We followed the recommendation by Skoglund et al. (98) of restricting the sex estimation to the sequences with evidence of the characteristic aDNA damage patterns when deviations from expected R_y values are observed. Consequently, sex assignment for these individuals was corrected after applying *post-mortem* damage filters with pmdtools (123) using thresholds of 0 – 3 (see Table S1). After pmdtools, all these four individuals were assigned as XX (Table S1). We speculate that the whole-genome in-solution enrichment (WGE kit by Arbor Biosciences) performed in these female individuals biased the proportion of reads in the sex chromosomes, as the baits are built from male individuals (as described by the company Arbor Biosciences). Even with the low contamination estimates (<3% in mtDNA), we assume that the few contaminating Y chromosome sequences are captured and enriched when using a WGE kit that includes Y chromosome probes. Moreover, according to a previous study, captured libraries tend to be enriched in pseudo-autosomal and repetitive regions, also due to the use of DNA of male individuals for preparing the baits (124).

After applying the filters, a total of 20 individuals got a molecular sex assignment; nine agreed with the previous morphological assignment, one was not in agreement (individual 2417J_TOL_a), and ten corresponded to individuals for which there was not a morphological assignment due to poor preservation of the skull or youngness of the individual (Table S1).

The identification of Y-DNA haplogroups was only possible in five out of the thirteen individuals genetically assigned as males, due to the low coverage (Table S1). All individuals possessed the Native American Q lineage as expected and in agreement with previous studies on ancient and present-day Indigenous Mexican individuals (18, 19–22, 120). Specifically, we identified the sub-haplogroups Q1a2a1-L54 in the Sierra Tarahumara (n=1) and Ranas (n=1) and Q1a2a1a1a1-M3 in the Sierra Tarahumara (n=1), Toluquilla (n=1) and Cañada de la Virgen (n=1), with no apparent differences between Aridoamerica and Mesoamerica (Table S1). The information on the paternal lineage was limited, since, despite confirming the presence of the Native American chromosomal haplogroup Q in the pre-Hispanic males, there is no clear differential spatial distribution of the two identified sub-haplogroups, Q1a2a1-L54 and Q1a2a1a1a1-M3, in both Aridoamerica and Mesoamerica. Few studies have recovered information of ancient Y chromosome haplogroups across the Americas and, given the sex-biased admixture events after European colonization, the frequency of haplogroup Q in the present-day population is low. Thus, further studies on more ancient individuals from the Americas at higher coverage will likely shed light on the demography of the male lineages.

Autosomal genetic structure

All pre-Hispanic individuals from Mexico, as well as all ancient individuals from California (15), Belize (101) and Patagonia (102) clustered together in PCA with present-day Indigenous populations from Mexico (Fig. S7) and share similar genetic composition between themselves as evidenced in ADMIXTURE analysis (Fig. S8). The same analyses were repeated including only the pre-Hispanic individuals and the present-day Indigenous populations from Mexico (14, 18, 22, 36) as a reference.

We applied a novel approach named “missing DNA” PCA (mdPCA) on the 15 pre-Hispanic individuals from Mexico, 12 from this study (Table S1) and the 3 Pericúes previously published (Table S3). Ancient individuals clustered closely with present-day populations from the same geographical region. Individual F9_ST_a from Aridoamerica clusters closely with present-day Rarámuri populations from north Mexico. Notably, MOM6_ST_a, also from Aridoamerica, clustered closely with present-day Indigenous populations from central-west Mexico (Purépecha and Nahuas from Jalisco). Pericúes from Baja California (14, 15) cluster closely between Wixárika, and central west Mexico Nahuas from Jalisco and Purépechas. Regarding the ancient individuals from central Mexico (Sierra Gorda, Cañada de la Virgen, and Michoacán), these cluster closely with present-day Indigenous populations from this same region (Nahua from Jalisco, Purépecha, Totonac, Nahua from Puebla, and Nahua) (Fig. 3A). Both pre-, and post-drought Sierra Gorda individuals cluster together with central Mexico present-day populations, contrary to what would be expected if a population replacement from Aridoamerica in the Sierra Gorda would have taken place (Fig. 3A). A similar structure was observed with standard PCA, except that here F9_ST_a clusters closer with Wixárika than with Rarámuri (Fig S9).

We contrasted the placement of individuals in mdPCA and standard space with that obtained when applying a Temporal Factor Analysis (TFA) (38) to correct the ancestral relationships using the

dates of the pre-Hispanic individuals. For this, we included only four pre-Hispanic individuals with >1x genome-wide coverage (F9_ST_a, 2417Q_TOL_b, 2417J_TOL_a, and 333B_TOL_a). This analysis confirmed the closer genetic affinity of individual F9_ST_b with populations from Aridoamerica, while Sierra Gorda individuals from Toluquilla were slightly shifted towards the central-east populations in contrast with their position in the mdPCA between central-west and central-east populations. Most notably, despite the different periods of the individuals from Toluquilla, they cluster together with all three dimension reduction approaches.

In agreement, ADMIXTURE analysis (K=6) reveals that pre-Hispanic individuals from Sierra Tarahumara and Pericúes have a bigger proportion of genetic ancestry shared with present-day northern populations Akimel O'odham and Rarámuri (blue component in Fig. 3B). Pre-Hispanic individuals from central Mexico also display a similar genetic composition to present-day central-west and central-east Mexico populations (Nahua from Jalisco, Purépecha, Totonac and, Nahua from Puebla). Interestingly, individuals from Sierra Gorda show a homogeneous genetic composition independently of the period in which they lived. The genetic composition of individuals from Cañada de la Virgen and Sierra Gorda is similar, but Cañada de la Virgen individuals have a slightly higher proportion of the blue component shared with northern populations. In addition, the two individuals from Michoacán have a different proportion of the components shared with northern populations, E2_Mich_b presents the orange and blue components while E4_Mich_b presents orange and green. However, due to the low-coverage of these individuals (0.03x and 0.04x, respectively), we take this observation with caution.

Outgroup-f3 and D-statistics

We computed outgroup-f3 statistics of the form $f_3(\text{Test}, \text{Source1}; \text{YRI})$, where the Test refers to the pre-Hispanic individual under analysis, and Source1 being a present-day Indigenous population. D-statistics were of the form $D(\text{Pop1}, \text{Pop2}; \text{Test}, \text{YRI})$, where Test refers to the pre-Hispanic individual under analysis, while Pop1 and Pop2 were all possible combinations of present-day Indigenous populations. For both analysis we considered only the combinations that have >1,000 overlapping Single Nucleotide Variants (SNVs), being 1,181 the lowest amount of SNVs plotted.

We found F9_ST_a to share higher genetic drift with present-day northern Mexico populations: Rarámuri, Akimel O'odham and Konkaak (0.291, 0.282, 0.277), than with any other present-day population (Fig. S11) (Table S9). Consistently, D-statistics showed F9_ST_a to be significantly more related to the present-day Rarámuri than to any other population (Pop1) when tested in the form $D(\text{Pop1}, \text{Rarámuri}; \text{F9_ST_a}, \text{YRI})$ ($D = -0.0763$ to -0.0210 , $|Z| = 6.913$ to 20.281) (Fig. S12) (Table S10).

In contrast, the genetic ancestries of pre-Hispanic individuals from central Mexico are more complex. They do not seem to have a higher shared genetic drift with any present-day Indigenous population according to the f3 outgroup values, where almost all standard errors overlap. Furthermore, in D-statistics analysis, we find mostly $D=0$ in all possible combinations of Pop1 and Pop2 (Fig. S13 – S18) (Table S9-S10). These results are consistent with populations from central Mexico having extensive gene flow between them and not completely diverged from each other.

Conditional Nucleotide Diversity (CND) and Runs of Homozygosity (ROH)

CND, a relative measure of genetic diversity, was estimated by genetic pairwise comparison of the genomic positions that are variable in African individuals (40). For these comparisons we grouped the individuals per archeological site and calculated CND by pairs of individuals. This analysis included 16 pre-Hispanic individuals (Pericúes, Sierra Tarahumara, Sierra Gorda, Cañada de la Virgen and Patagonia) with at least 10,000 SNVs overlapping between each pair of individuals (Table S11 and S12). We found similar values of CND in pairs of pre-Hispanic individuals to those observed in present-day Indigenous populations from Mexico, except for Pericúes from Baja California, who had the lowest CND values (Fig. S19) (Table S12). This is in line with Pericúes being hunter gatherers, likely with small population sizes and living in isolation in Baja California (125).

We calculated ROH in six pre-Hispanic individuals with the highest coverage; namely two from Aridoamerica (the Pericú B03 and F9_ST_a from Sierra Tarahumara) and four from central Mexico (three from Toluquilla and one from Cañada de la Virgen) (Table S13). Of the three Toluquilla individuals included in this analysis, one was dated pre-drought (2417Q_TOL_b), and two post-drought (2417J_TOL_a and 333B_TOL_a). Pre-Hispanic individuals display fewer segments in ROH than some present-day Indigenous individuals, especially present-day individuals from northern Mexico. The segment size distribution of ROHs suggests that the pre-Hispanic individuals studied here belonged to populations with small effective population (N_e) sizes ($2N_e = 1,600-6,400$) (Fig. S20), which is similar to N_e previously calculated for present-day northern (Rarámuri, $N_e = 2,419$; Wixárika, $N_e = 2,193$) and southern (Triqui, $N_e = 2,407$; Maya, $N_e = 2,750$) populations (43). We contrasted the CND values and ROH estimations of Pericúes from Mexico with another isolated pre-Hispanic population from Patagonian hunter gatherers (1,275 – 895 BP) (102), expecting similarly low levels of genetic diversity. We found that two pairs of Pericúes had similar CND levels as Patagonia individuals ($\sim 0.16-0.175$), and the third pair of Pericúes presented a lower CND value (~ 0.14). However, in ROH analysis, we found that the four individuals from Patagonia have a higher sum of segments in ROH than the B03 Pericú individual, suggesting that the Pericú population had a bigger population size than that of South American Patagonian individuals (Fig. S20). In summary, the CND and ROH information showed that pre-Hispanic and present-day individuals from northern Mexico (Pericúes and Akimel O'odham, respectively) present the lowest CND values and the highest sum of inferred ROH >4 cM compared to other populations of their respective period. Individuals from the Sierra Tarahumara, Toluquilla and Cañada de la Virgen show similar values of genetic diversity and sum of inferred ROH segments. Furthermore, in the case of Toluquilla, where we have three individuals from different dates, we found that the CND value increases with the date difference between pairs of individuals, probably reflective of the accumulation of new mutations or genetic drift through time during the 489-680 years of difference between 2417Q_TOL_b and the individuals of a later period (2417J_TOL_a and 333B_TOL_a).

Genetic continuity from 1,100 to the present in Sierra Tarahumara, northern Mexico

We decided to evaluate the possible scenario of population continuity in the Sierra Tarahumara from northern Mexico, considering that the individual F9_ST_a was significantly more related to present-day Rarámuri than to any other indigenous population from Mexico in f_3 and D statistics. We used the population continuity test described in (39), with the null hypothesis of population continuity tested by using read counts in the ancient individual and allele frequencies in the

reference population (Rarámuri). Using this method, we found no population continuity between F9_ST_a and present-day Rarámuri and rejected the null hypothesis with a p-value of $10^{-499.3}$.

The northern region of Mexico was one of the most difficult to access and the last areas to be invaded by the Europeans in the seventeenth century, due to its geography and the resistance movements by the local people. Rarámuri still preserve their pre-Hispanic cultural elements today in language, music, dresses, and handcraft. We hypothesize that this genetic population discontinuity observed from this analysis can result from the changes in allele frequencies derived from the admixture events, population bottleneck, and isolation experienced by Rarámuri since pre-Hispanic times and after the European colonization (48).

Interestingly, Sierra Tarahumara mummy MOM6_ST_a has a higher shared genetic drift with F9_ST_a, as evidenced by outgroup-f3 tests, but mdPCA and ADMIXTURE analyses show that he has also genetic ancestry components found in central west present-day populations. Using D statistics, we found MOM6_ST_a to be equally related to all individuals from Aridoamerica and Mesoamerica, supporting the idea of gene flow between these two areas.

Ghost population contribution to pre-Hispanic Mexico

We tested the presence of UPopA in the pre-Hispanic individuals, using an admixture graph model with the same topological framework but some modifications compared to (47). We used Mbuti as an outgroup, Ami as the East Asian, MA1 as the ancient north Eurasian, USR1 as the ancient Beringian (45), Anzick-1 (46) and Spirit Cave (47) as the SNA, and Athabascan as the NNA. Interestingly, the Mummy F9_ST_a from the north and the individuals from Cañada de la Virgen from central Mexico show genetic ancestry from a ‘ghost’ population, at 28% and 17%, respectively ($|Z\text{-score}|$: 2.142 and 1.655, respectively) (Fig. S27A-B), consistent with the “ghost” UPopA genetic ancestry reported in Mixe. Models including individuals from other archaeological sites had $|Z\text{-score}|$ values over 3 and branches with inner zeros, thus we rejected them. We repeated the test with the Mixe population to verify the contribution of the ghost population in our tested admixture graph structures, and found Mixe to have a genetic ancestry coming from an unsampled UPopA population at 13% ($|Z\text{-score}|$: 2.34) (Fig. S27C), which is similar to previous estimates of 11% (47). To confirm these results, we tested whether the ghost genetic ancestry observed in pre-Hispanic individuals was the same found in Mixe, by modeling the pre-Hispanic individuals along with Mixe, applying two different scenarios: 1) pre-Hispanic modeled in a same clade with Mixe, receiving UPopA genetic ancestry, and 2) pre-Hispanic and Mixe receiving UPopA genetic ancestry, with an extra contribution of NNA in the pre-Hispanic. Interestingly, Results show that Sierra Tarahumara and Mixe share the same ghost genetic ancestry UPopA (with F9_ST_a better modeled when having a higher contribution of NNA genetic ancestry) (Fig. S28), while in both models Cañada de la Virgen does not share the same UPopA contribution ($|Z| > 3$) (Fig. S29). Then we tested a third model with Mixe and Cañada de la Virgen to see whether the genetic ancestry in the latter, comes from a secondary unsampled population (UPopA2) and not the previously observed in Mixe (UPopA1). Results show Cañada de la Virgen can be modeled receiving genetic ancestry from a different unsampled population “UPopA2” suggesting the existence of another ghost population not sampled yet (Fig S29). We then modeled the two pre-Hispanic, Mixe and the two unsampled populations within the same model to confirm and summarize the results (Fig. 5). We attempted different combinations for the placement of UPopA, as well as adding and removing USR1 to confirm whether Mixe and Cañada de la Virgen could be

modeled together with only one UpopA contribution, but all these models were rejected (Fig. S30-S33).

Extended Discussion

Population continuity in Sierra Gorda

The observed population continuity in the Sierra Gorda might have been supported by the favorable climatic conditions of the northern Sierra Gorda Mountain range. The landscape could have helped maintain higher humidity conditions than in other regions of the northern frontier of Mesoamerica. In this part of the Sierra, the humid winds from the Gulf of Mexico course through. The area is characterized by coniferous and mixed forests, where the average maximum temperature is 25°C, and the minimum temperature is 6-7°C (126). This landscape is very different from the conditions found in Cañada de la Virgen and Guanajuato, where the climate is considerably more arid. Cañada de la Virgen is located north of the Lerma River, where rainfall is known to be scarce. The inhabitants of that region had to develop systems for the capture and maintenance of water needed for cultivation, which probably became impossible after the prolonged droughts. This forced them to abandon the site during 1,000-1,100 CE and migrate to areas with better climatic conditions suitable for agriculture, which was the core of their primary subsistence activity (56). In contrast, the main subsistence strategy in Toluquilla and Ranas was mining. Residents at these sites exploited cinnabar for trade purposes with other villages and metropolises (55, 59). Cinnabar was of great sacred value to pre-Hispanic cultures. Some ancient communities used it to anoint their dead, especially in cases of highly regarded individuals, and as part of ritual offerings (59, 60). We hypothesize that the cinnabar trade and the landscape of the Sierra Gorda allowed the peoples of Toluquilla and Ranas to subsist despite low rainfall conditions during the drought period as their economy did not depend strongly on agriculture.

Demography of central Mexico populations

We were also interested in obtaining insights into the demography of central Mexico populations since previous attempts of demographic modeling have been hampered by the high genetic heterogeneity observed in these populations (43, 50). Our outgroup- f_3 and D statistics results show that, although pre-Hispanic individuals from central Mexico share higher genetic drift with present-day populations from that region, none of these relationships are statistically significant. We thus leveraged the ancient genome data to propose an admixture model for central Mexico populations. Using qpGraph we found that the pre-Hispanic populations from central Mexico, all have different genetic ancestry sharing with the NNA and SNA branches. This was expected given previous studies on ancient genomes from Central and South America reporting multiple admixture events between these two branches after their split ca. 15,000 years ago (15, 48). Together, these observations point to a scenario in which populations from central Mexico have not completely diverged from one another, possibly due to extensive gene flow as expected based on the active commercial exchange between different Mesoamerican populations for centuries (61). This interaction was mainly through trade routes and alliances between different nations (61), as observed in present-day Indigenous populations who share identity-by-descent (IBD) segments (36, 48) and through the analysis of STR loci (127). There was also gene flow between Mesoamerican and Aridoamerican populations in pre-Hispanic times, as evidenced by the study of IBD segments in present-day Indigenous populations from Mexico (48). However, this gene

flow occurred less frequently than within Mesoamerica. The individual MOM6_ST_a from Aridoamerica may have been the result of such admixture between Mesoamerican and Aridoamerican ancestors. This individual has a higher shared genetic drift with F9_ST_a in outgroup-f3 tests. However, ADMIXTURE analysis show that MOM6_ST_a has genetic ancestry components found in central west present-day populations. Also, using \bar{D} statistics, we found MOM6_ST_a to be equally related to all individuals from Aridoamerica and Mesoamerica, supporting the idea of gene flow between these two areas.

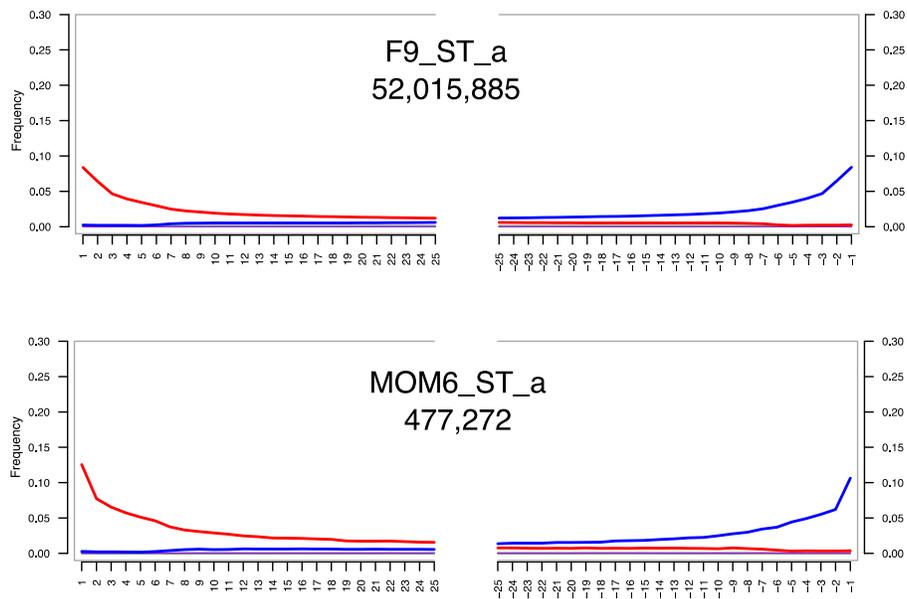


Fig S1. Damage patterns of the genomic data from Sierra Tarahumara pre-Hispanic individuals. The number of reads included in the analysis is indicated below the individual's ID. The red color indicates the frequency of cytosine to thymine changes, while the blue color indicates guanine to adenine changes. The x-axis shows the position of the base within the read, in the direction 5' to 3'. The y-axis shows the frequency of such changes.

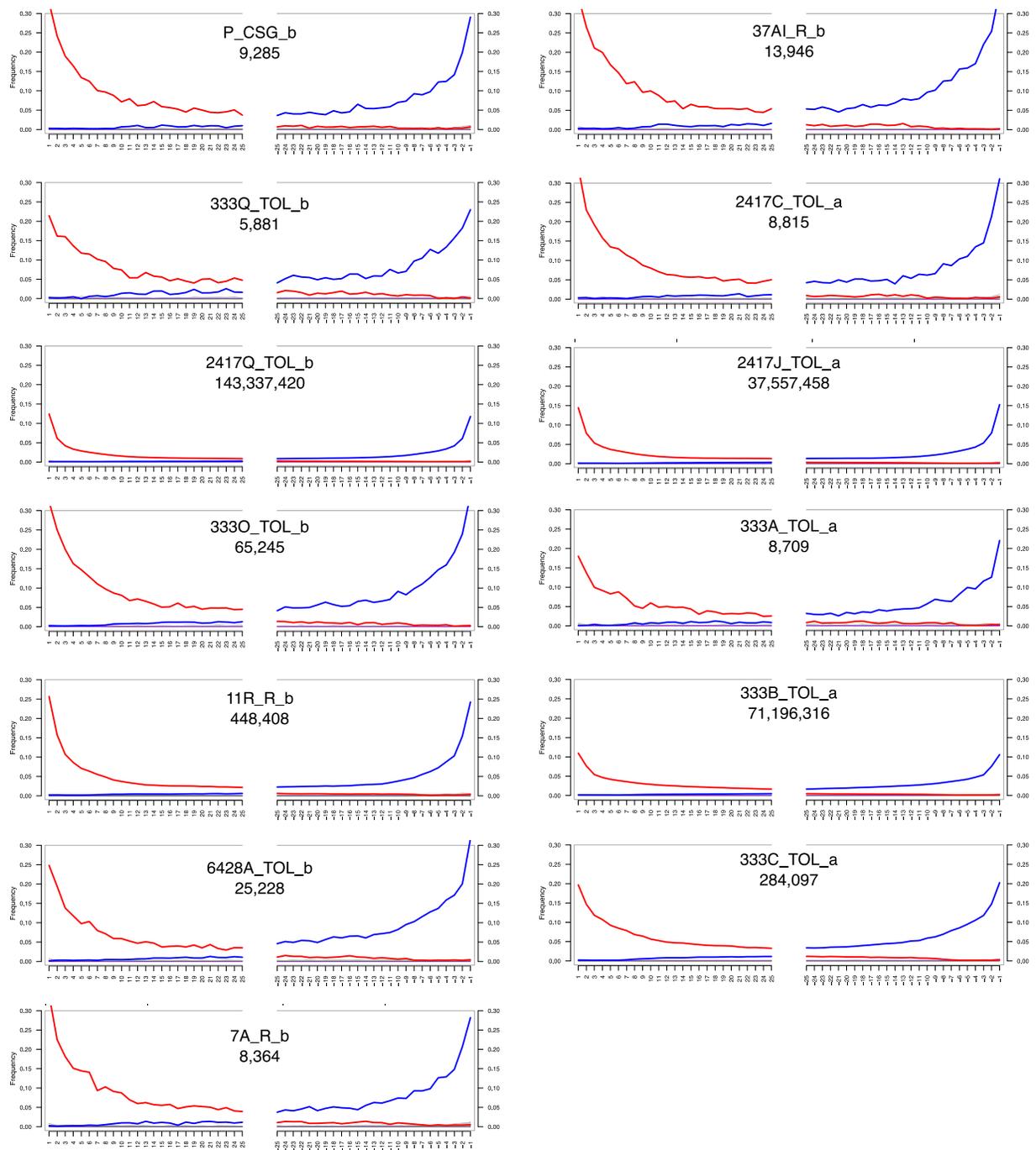


Fig S2. Damage patterns of the genomic data from Sierra Gorda pre-Hispanic individuals. The number of reads included in the analysis is indicated below the individual's ID. The red color indicates the frequency of cytosine to thymine changes, while the blue color indicates guanine to adenine changes. The x-axis shows the position of the base within the read, in the direction 5' to 3'. The y-axis shows the frequency of such changes.

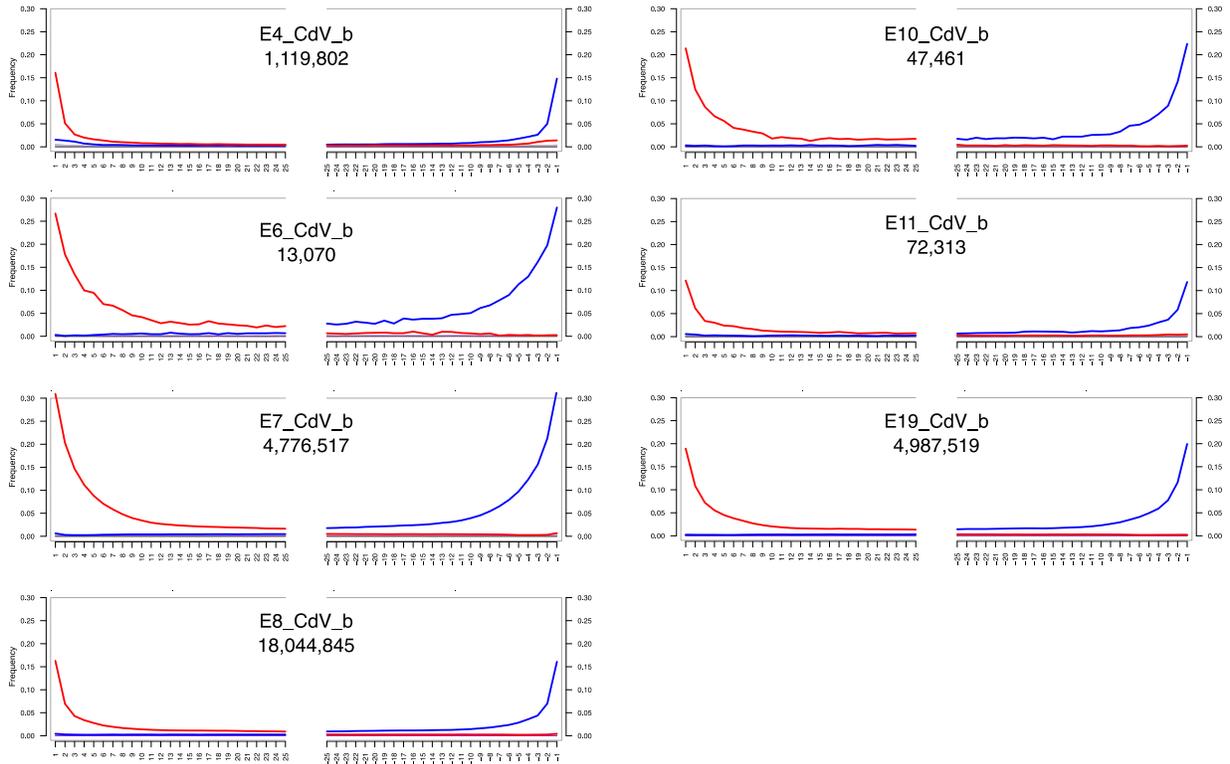


Fig S3. Damage patterns of the genomic data from Cañada de la Virgen pre-Hispanic individuals. The number of reads included in the analysis is indicated below the individual's ID. The red color indicates the frequency of cytosine to thymine changes, while the blue color indicates guanine to adenine changes. The x-axis shows the position of the base within the read, in the direction 5' to 3'. The y-axis shows the frequency of such changes.

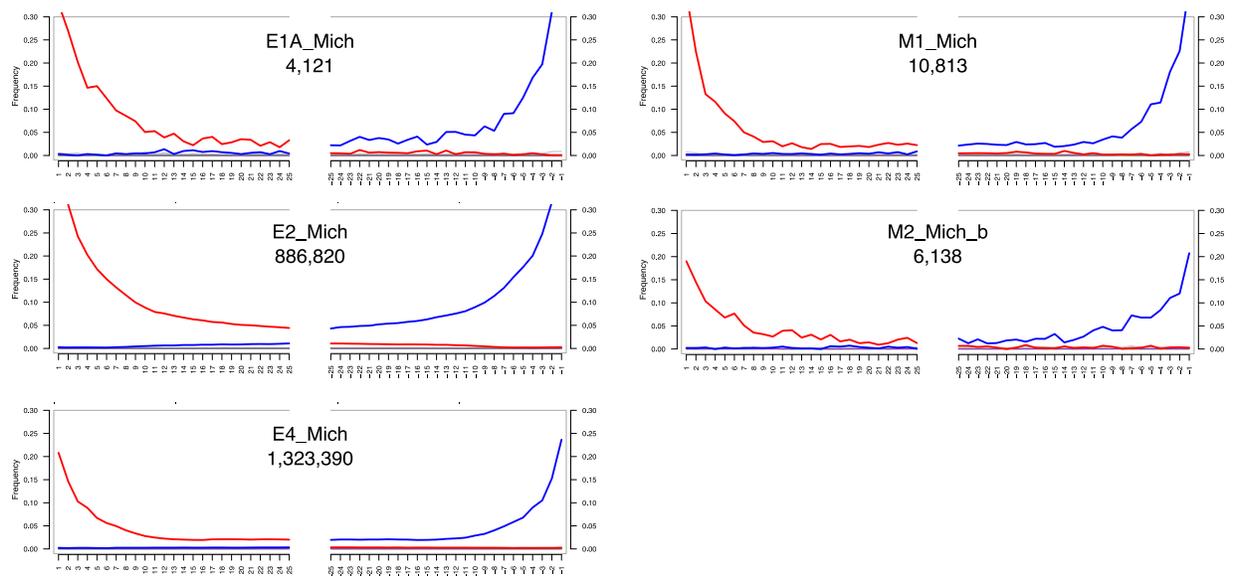


Fig S4. Damage patterns of the genomic data from Michoacán pre-Hispanic individuals. The number of reads included in the analysis is indicated below the individual's ID. The red color indicates the frequency of cytosine to thymine changes, while the blue color indicates guanine to adenine changes. The x-axis shows the position of the base within the read, in the direction 5' to 3'. The y-axis shows the frequency of such changes.

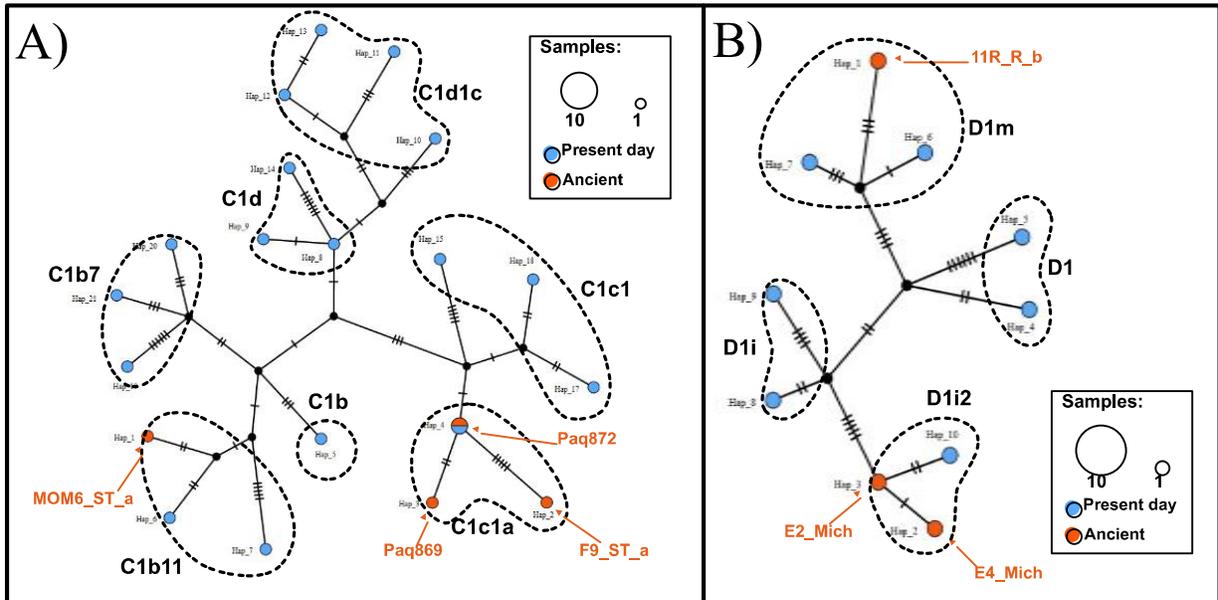


Fig S5. Haplotype Networks for mitochondrial haplogroups of pre-Hispanic individuals from Mexico. (A) Haplogroup C, (B) Haplogroup D. Sub-haplogroups are shaded with gray. Orange arrows point to ancient individuals from the present study and public available data used for the analysis (Table S4).

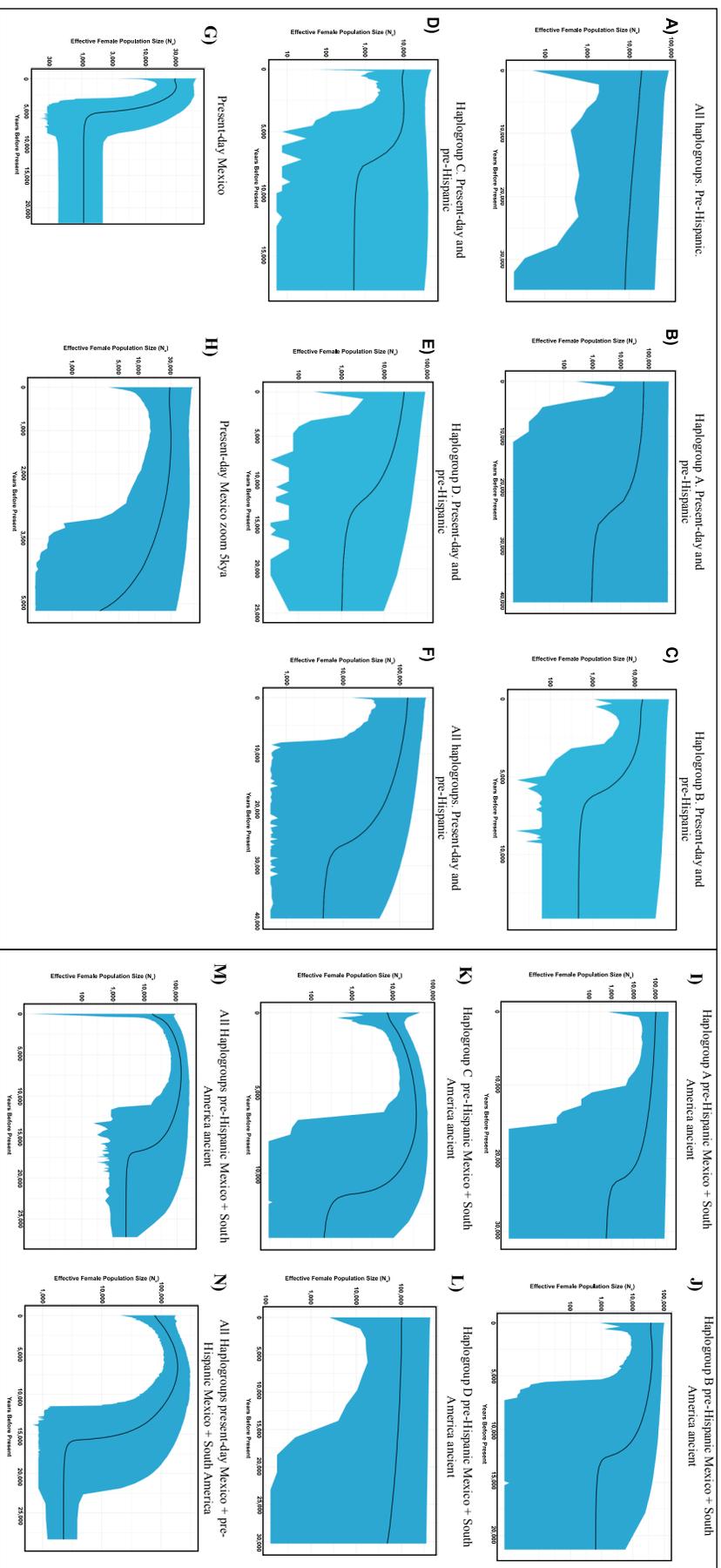


Fig S6. Inference of past effective female population size through Extended Bayesian skyline plots (EBSP). The EBSPs were constructed with ancient and present-day mitochondrial DNA. The median is shown as a black line, and the credibility interval of 95% is shown in blue.

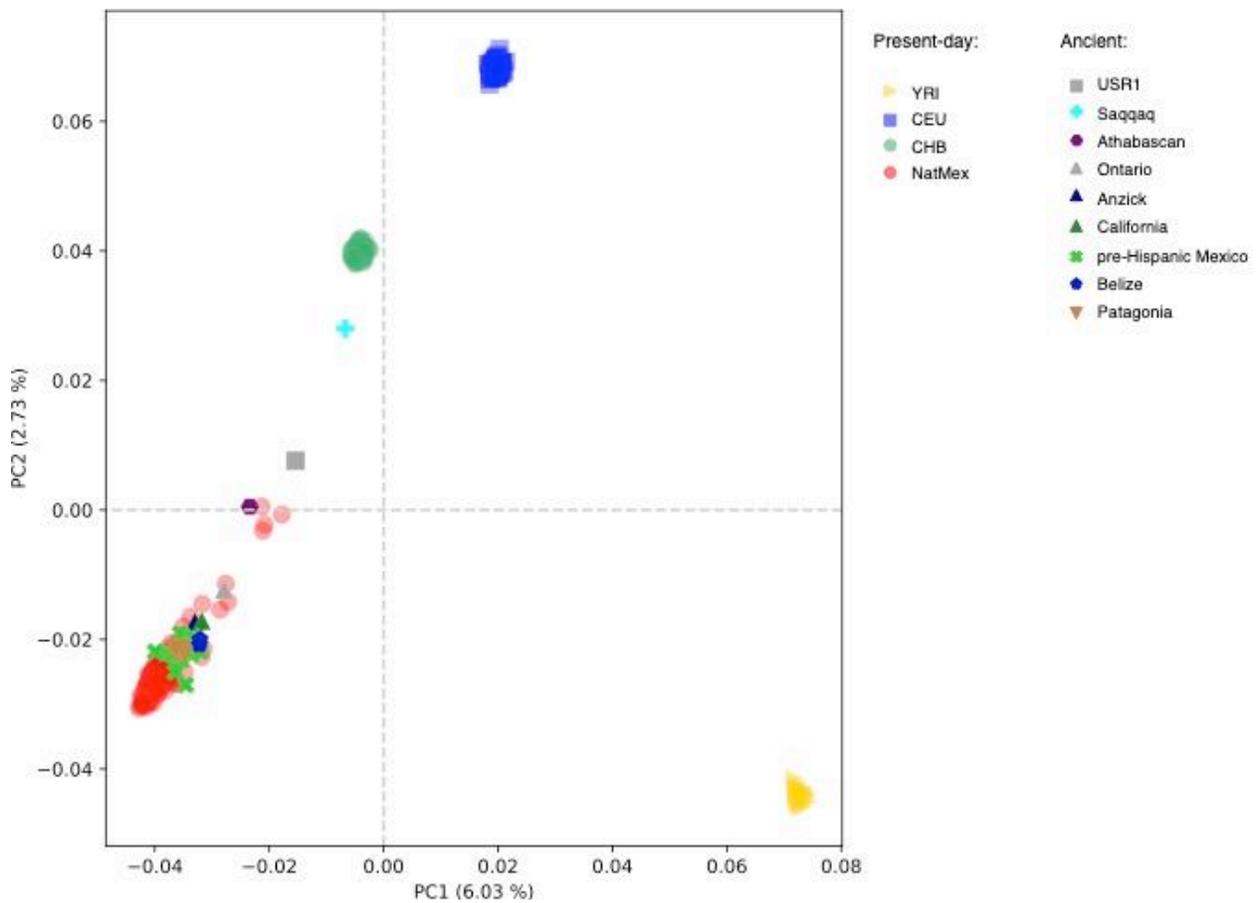


Fig. S7. Global Principal Components Analysis (PCA) including pre-Hispanic individuals from this study. The PCA includes ancient individuals Mexico from this and other studies, as well as ancient individuals from the Americas (Supplementary Table S1 and S3), Indigenous populations from Mexico (NatMex), and continental reference populations from Africa (YRI), Europe (CEU), and Asia (CHB) from 1000 Genomes.

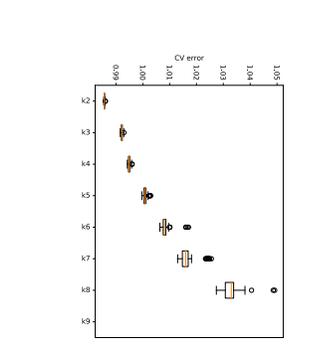
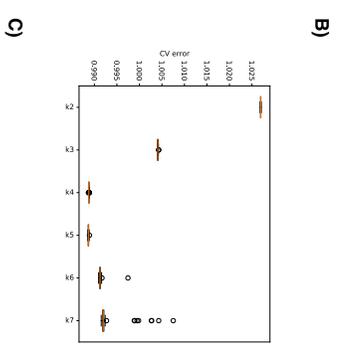
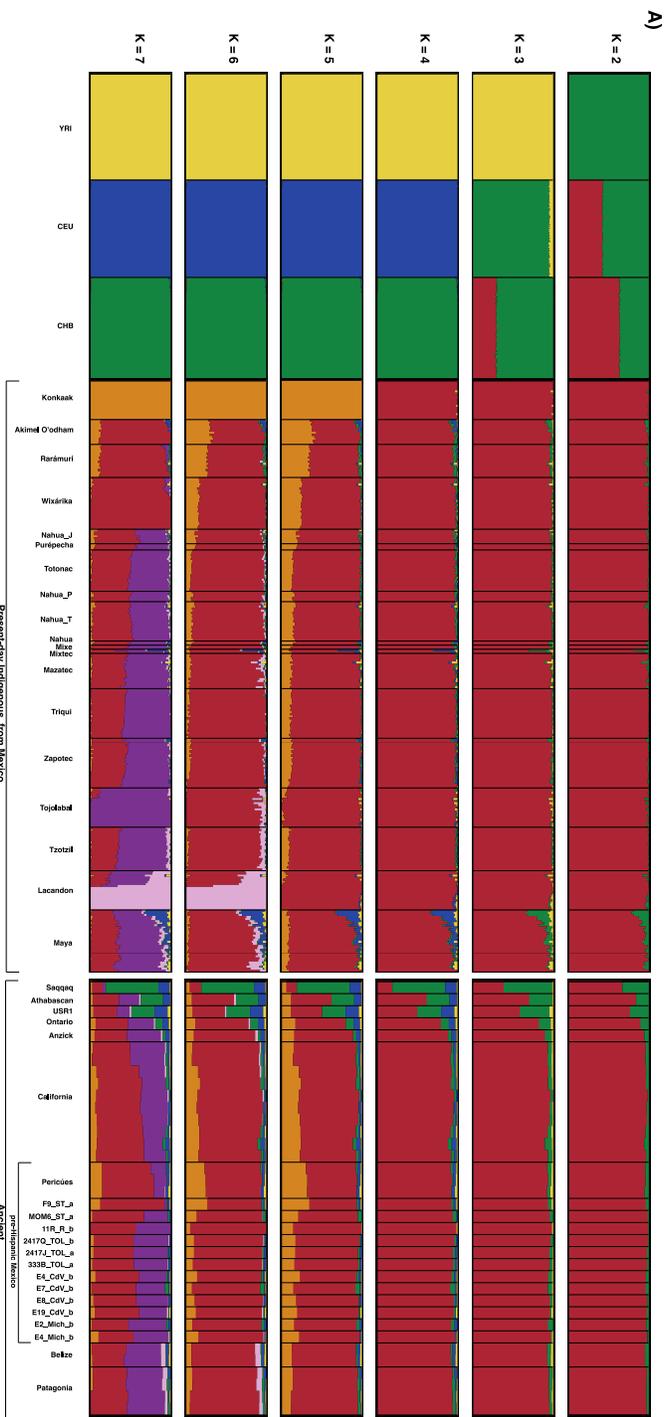


Fig. S8. ADMIXTURE analysis including pre-Hispanic individuals from Mexico. (A) Continental ADMIXTURE plot shown for K=2 through K=7. Individuals include continental references from Africa (YRI), Europe (CEU), and Asia (CHB) from 1000 Genomes as well as present-day Indigenous populations from Mexico, and ancient individuals from the Americas. Ancient populations are sorted from north to south. (B) Cross validation (cv) error values for the Continental ADMIXTURE. (C) Cross validation (cv) error values for the ADMIXTURE analysis with the reference panel including only Indigenous individuals from Mexico.

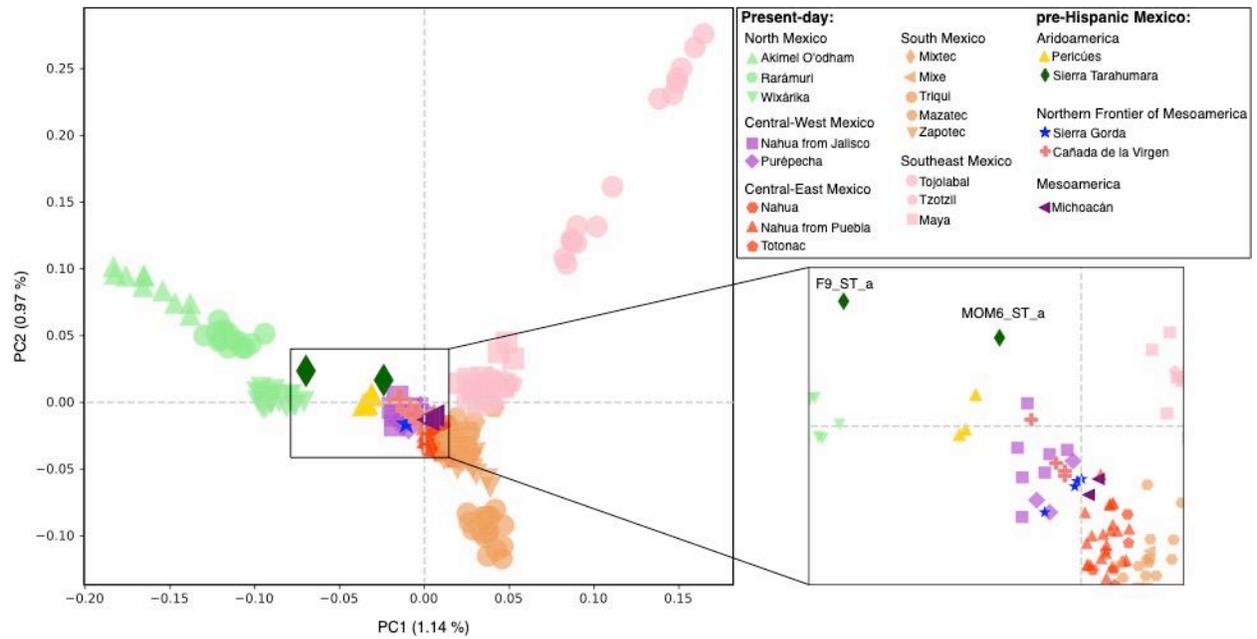


Fig. S9. PCA of populations from Mexico. The analysis includes the reference panel of present-day Indigenous Populations (see methods) and the genomic data of the pre-Hispanic individuals from Mexico. Individual F9_ST_a from Sierra Tarahumara clusters with present-day populations from northern Mexico, while individuals from Cañada de la Virgen and MOM6 cluster close to central-west present-day populations. Sierra Gorda and Michoacán individuals cluster between central-west and central-east present-day populations.

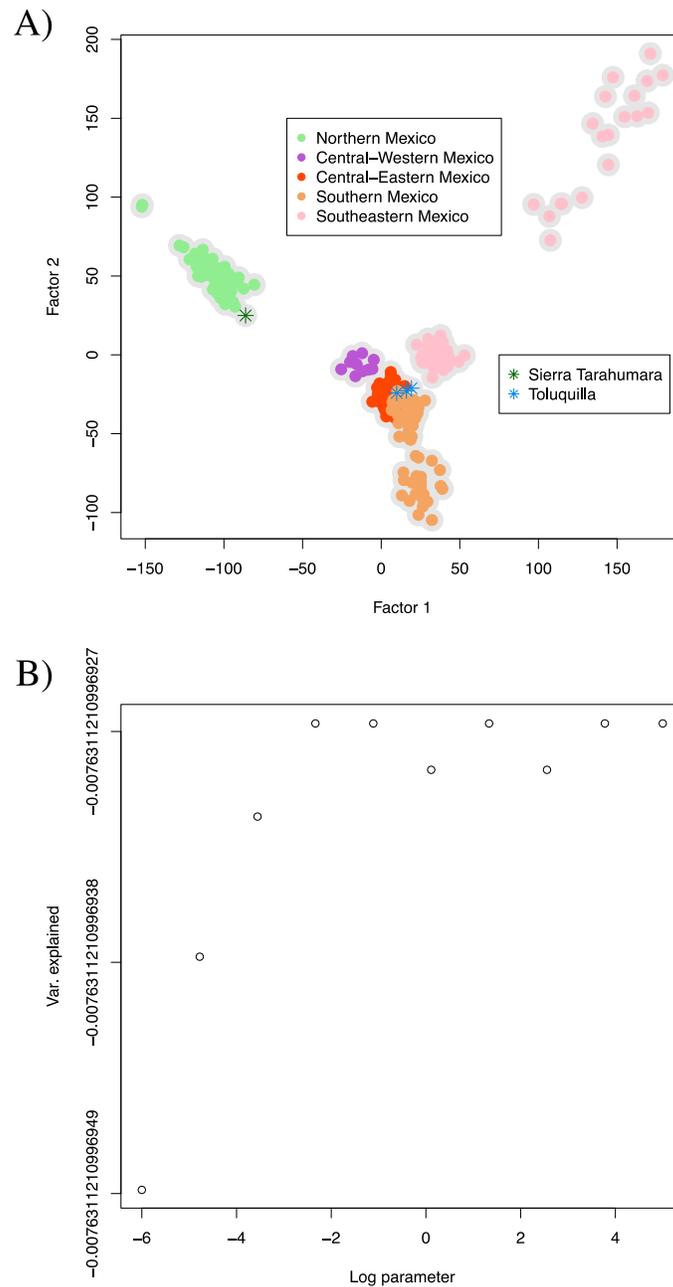


Fig. S10. Temporal factor analysis (TFA) of ancient and present-day Indigenous individuals from Mexico. (A) TFA includes the four pre-Hispanic individuals, F9_ST_a from Sierra Tarahumara and 2417Q_TOL_b, 2417J_TOL_a 333B_TOL_a from Toluquilla, Sierra Gorda. (B) Percentage of variance explained in the y-axis by different lambda values. The x-axis shows the lambda values in a log scale. A decrease in variance from right to left represents the lambda value used to correct temporality.

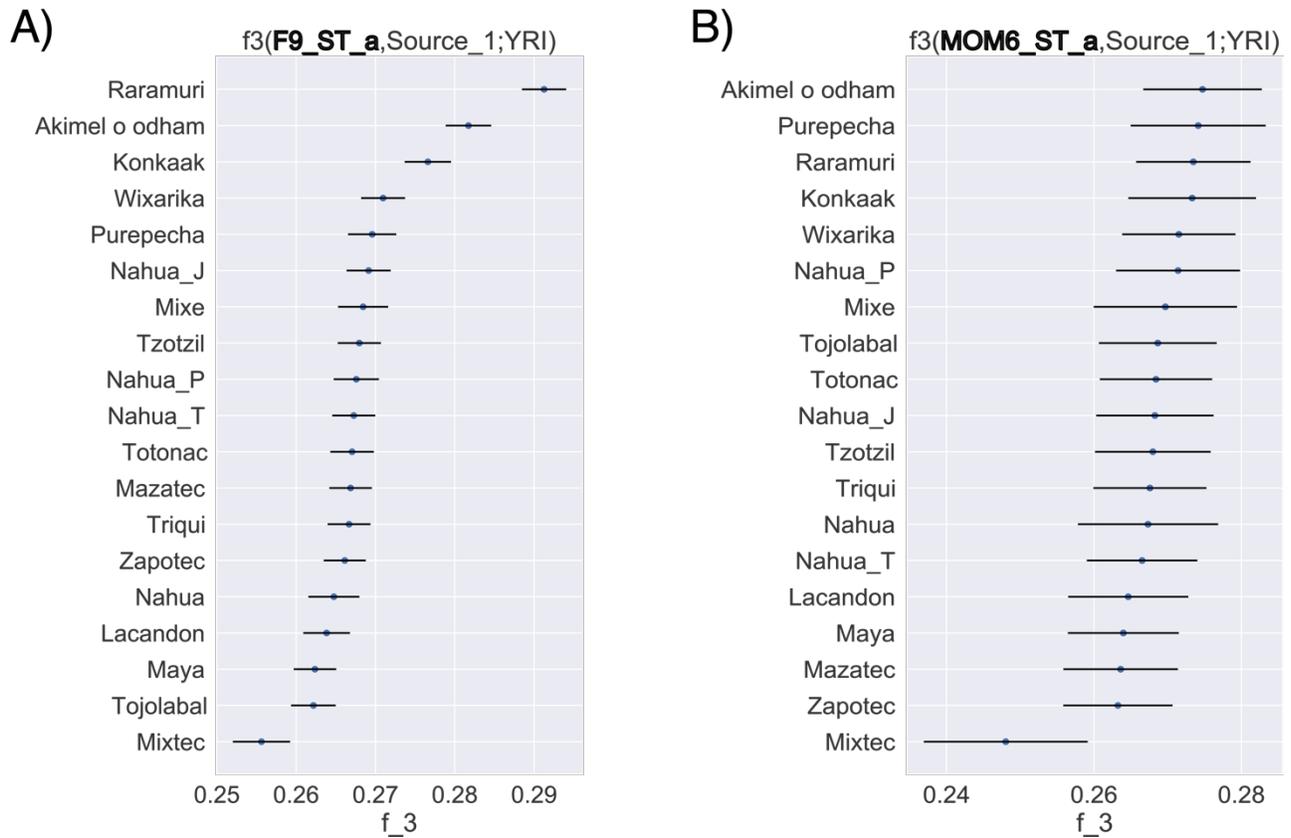
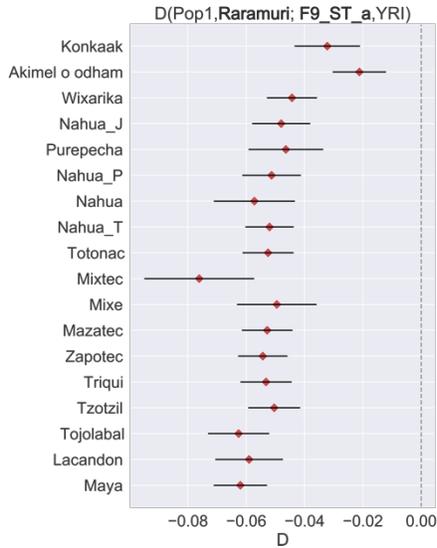
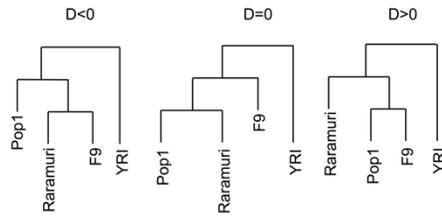


Fig. S11. Shared generic drift between Sierra Tarahumara mummies and other present-day populations of Mexico. Outgroups f3 statistics are shown for (A) Mummy F9_ST_a and (B) Mummy MOM6_ST_a. Higher values of f3 indicate higher shared genetic drift. Point estimates and one standard error are shown. The individual F9_ST_a have a higher genetic drift shared with present-day Rarámuri population from Northern Mexico. While individual MOM6_ST_a have a similar genetic drift shared with all present-day Indigenous populations from Mexico except with Mixtec.

A)



B)

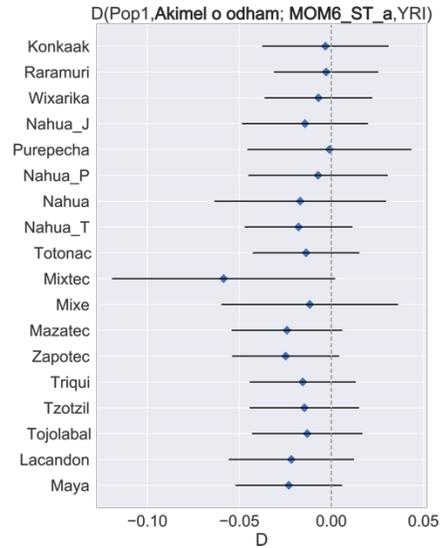
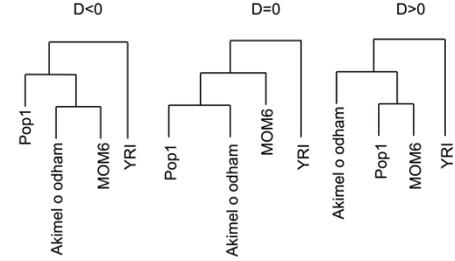


Fig. S12. D statistics for pre-Hispanic individuals from Sierra Tarahumara and other present-day populations. (A) D statistics in the form $D(\text{Pop1}, \text{Rarámuri}; \text{F9_ST_a}, \text{YRI})$, (B) D statistics in the form $D(\text{Pop1}, \text{Akimel o'odham}; \text{MOM6_ST_a}, \text{YRI})$. Expected tree topologies according to D value are drawn on the top of the plot, individual IDs in the trees are indicated with no suffixes. Red dots indicate significant deviations from $D=0$ ($|Z| > 3$). The individual F9_ST_a shows a significantly higher relationship with present-day Raramuri than with any other present-day population. The individual MOM6_ST_a seems to be equally related to all present-day Indigenous populations from Mexico.

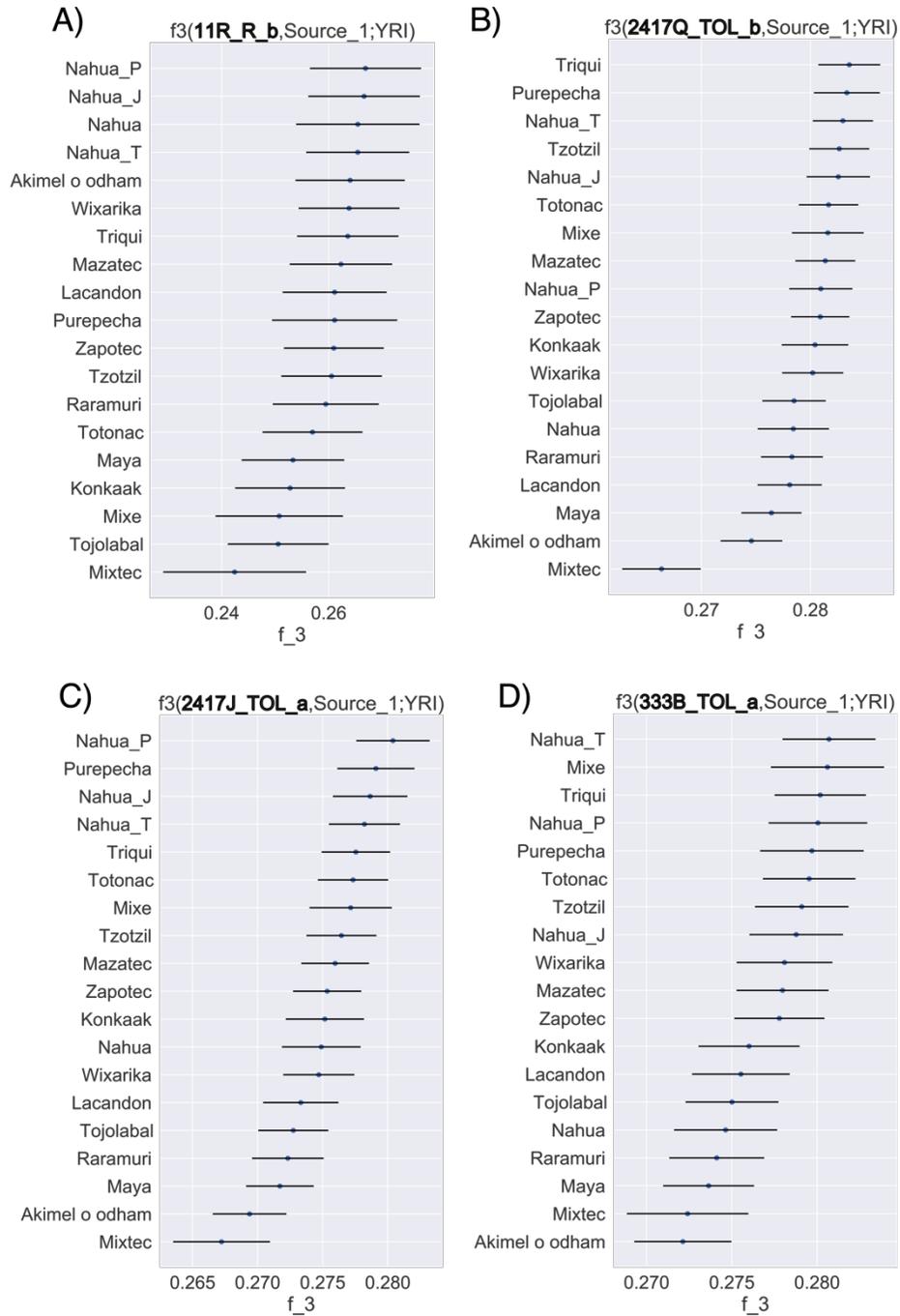


Fig. S13. Outgroup f₃ statistics for pre-Hispanic individuals from Sierra Gorda and present-day Indigenous populations. (A) 11R_R_b, (B) 2417Q_TOL_b, (C) 2417J_TOL_a, and (D) 333B_TOL_a. Higher values of f₃ indicate higher shared genetic drift. Point estimates and one standard error are shown. All individuals show a similar genetic drift shared with all present-day Indigenous populations from Mexico.

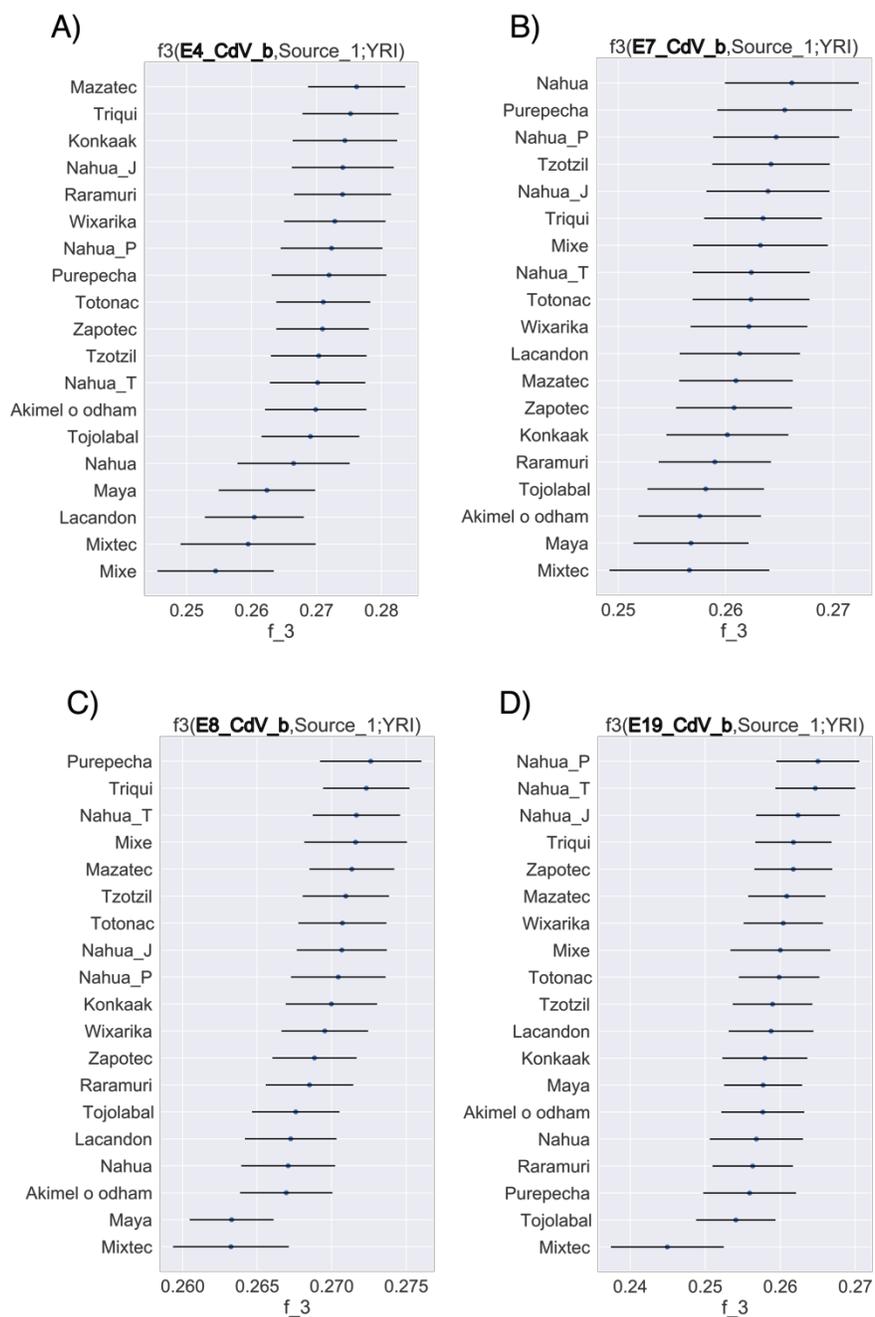


Fig. S14. Outgroups f3 statistics for pre-Hispanic individuals from Cañada de la Virgen (CdV) and present-day Indigenous populations. (A) E4_CdV_b, (B) E7_CdV_b, (C) E8_CdV_b, and (D) E19_CdV_b. Higher values of f3 indicate higher shared genetic drift. Point estimates and one standard error are shown. All pre-Hispanic CdV individuals show a similar genetic drift shared with all present-day Indigenous populations from Mexico, except with Mixtec.

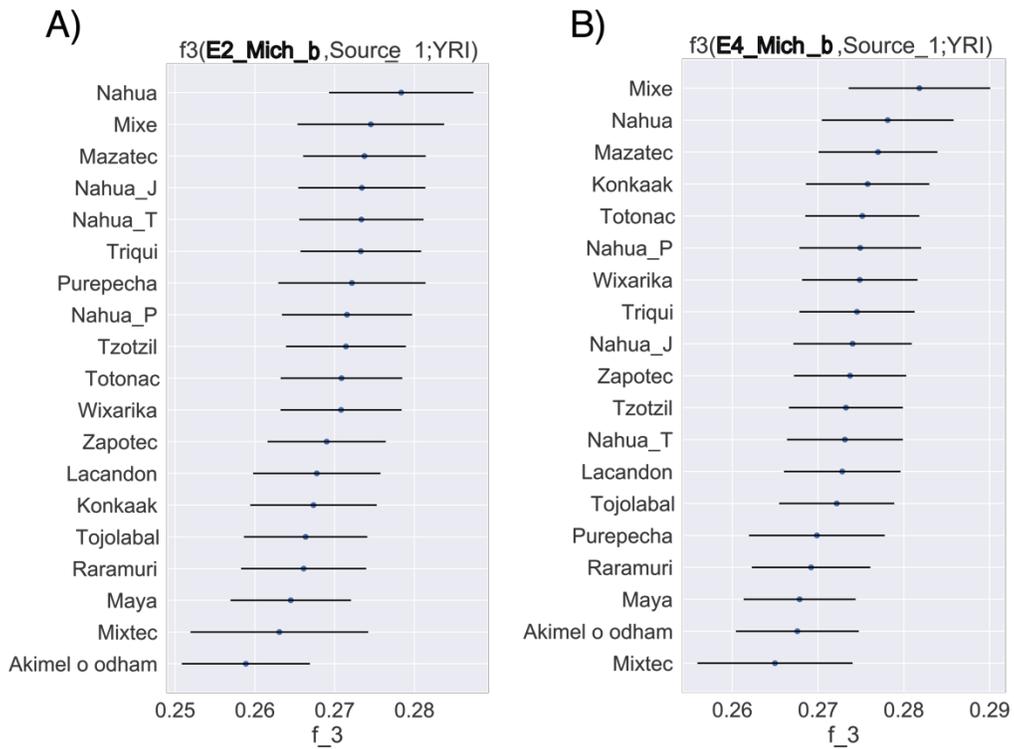


Fig. S15. Outgroups f_3 statistics for pre-Hispanic individuals from Michoacán and present-day Indigenous populations. (A) E2_Mich_b, and (B) E4_Mich_b. Higher values of f_3 indicate higher shared genetic drift. Point estimates and one standard error are shown. Both individuals show a similar genetic drift shared with all present-day Indigenous populations from Mexico.

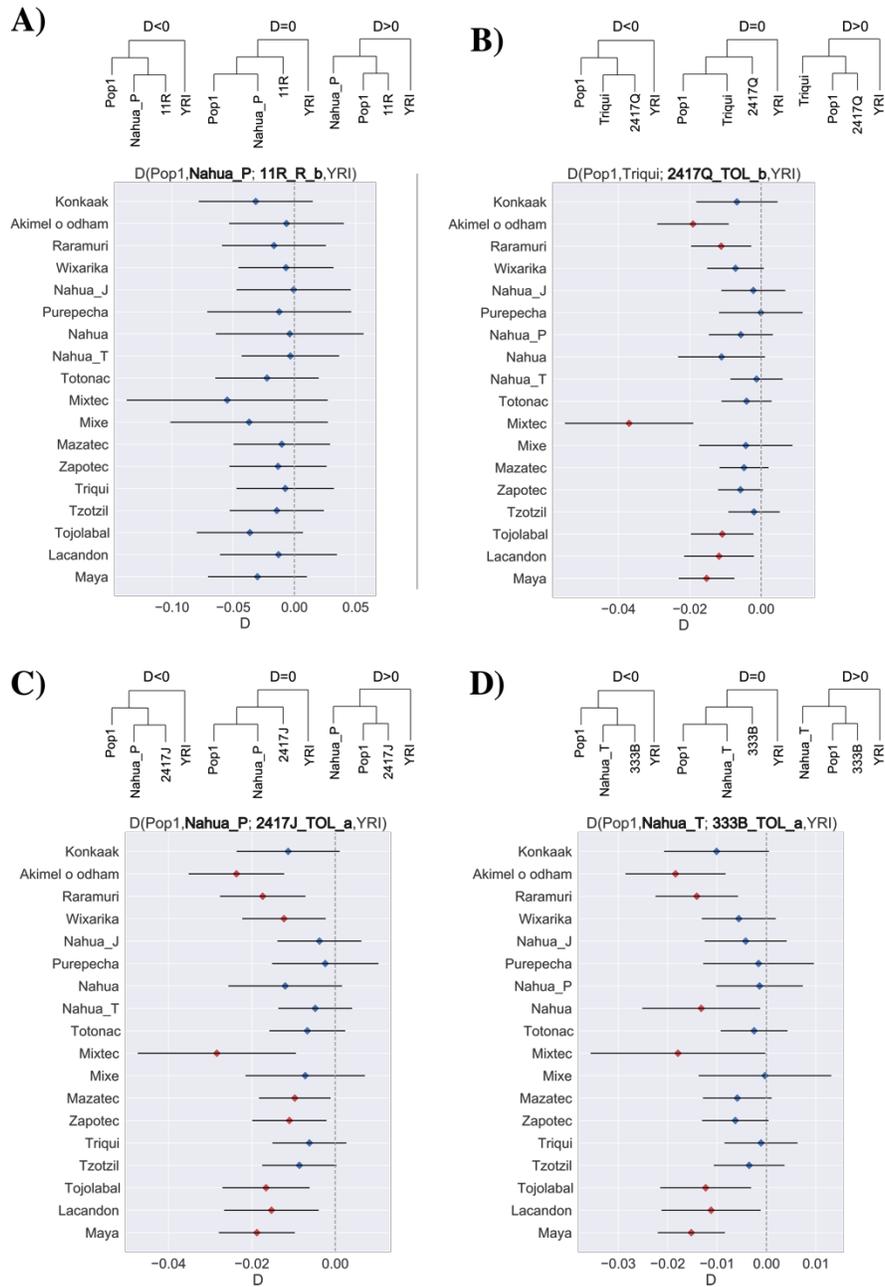


Fig. S16. D statistics for pre-Hispanic individuals from Sierra Gorda and present-day Indigenous populations. (A) $D(\text{Pop1, Nahua_P; 11R_R_b, YRI})$, (B) $D(\text{Pop1, Triqui; 2417Q_TOL_b, YRI})$, (C) $D(\text{Pop1, Nahua_P; 2417J_TOL_a, YRI})$ and (D) $D(\text{Pop1, Nahua_T; 333B_TOL_a, YRI})$. Expected tree topologies according to D values are drawn on the top of the plot, individual IDs in the trees are indicated with no suffixes. Red dots indicate significant deviations from $D=0$ ($|Z| > 3$). Individuals from Sierra Gorda tend to have a significantly higher relationship with the present-day population used as Pop2 when Pop1 is a population from Northern or Southeast Mexico, but not when they are compared with other present-day populations from Central Mexico.

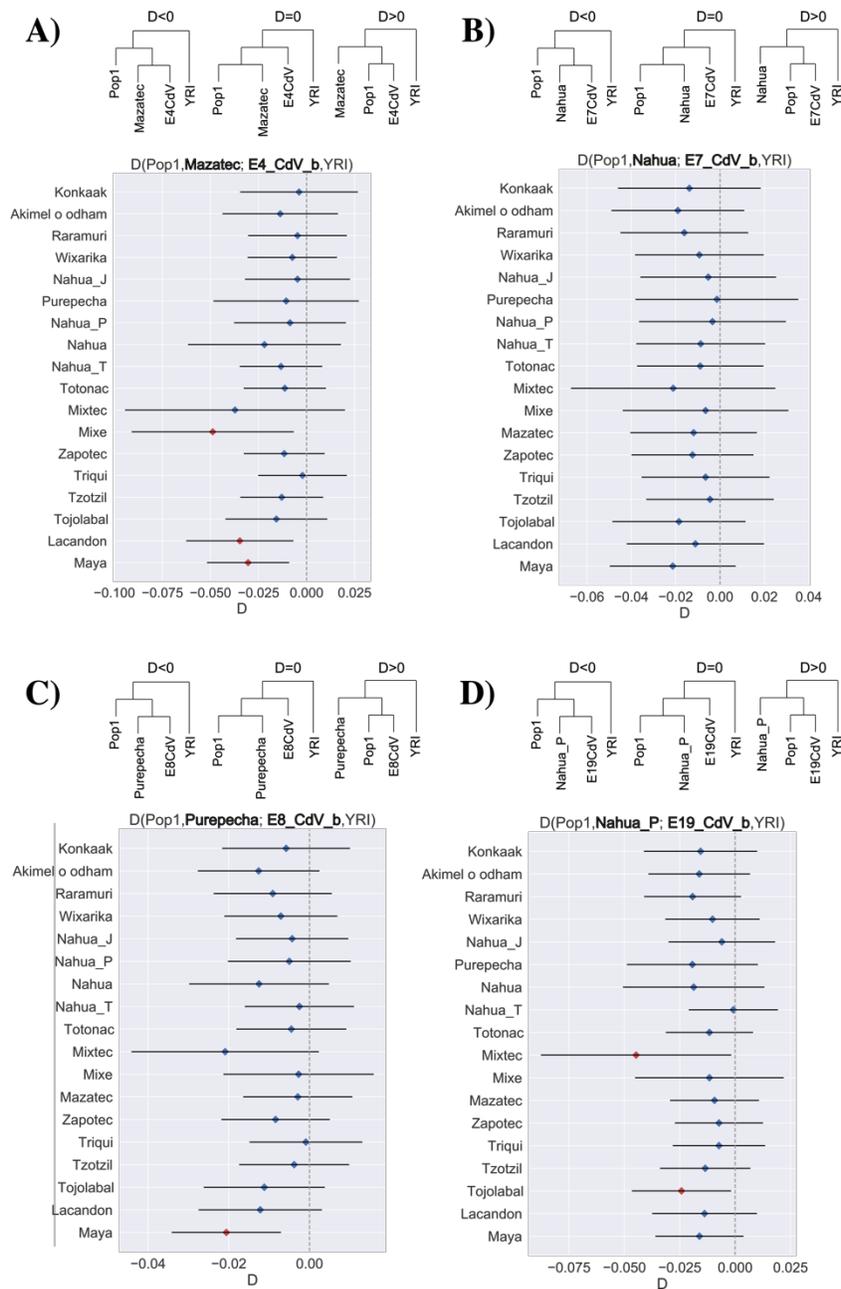


Fig. S17. D statistics for pre-Hispanic individuals from Cañada de la Virgen and present-day Indigenous populations. (A) $D(\text{Pop1, Mazatec; E4_CdV_b, YRI})$, (B) $D(\text{Pop1, Nahua; E7_CdV_b, YRI})$, (C) $D(\text{Pop1, Purepecha; E8_CdV_b, YRI})$ and (D) $D(\text{Pop1, Nahua_P; E19_CdV_b, YRI})$. Expected tree topologies according to D value are drawn on the top of the plot, individual IDs in the trees are indicated with no suffixes. Red dots indicate significant deviations from $D=0$ ($|Z| > 3$). Individuals from Cañada de la Virgen seem to be equally related to all present-day Indigenous populations except when Pop1 is Mixtec, Mixe or Southeastern populations, where they show a significant higher relationship with Pop2.

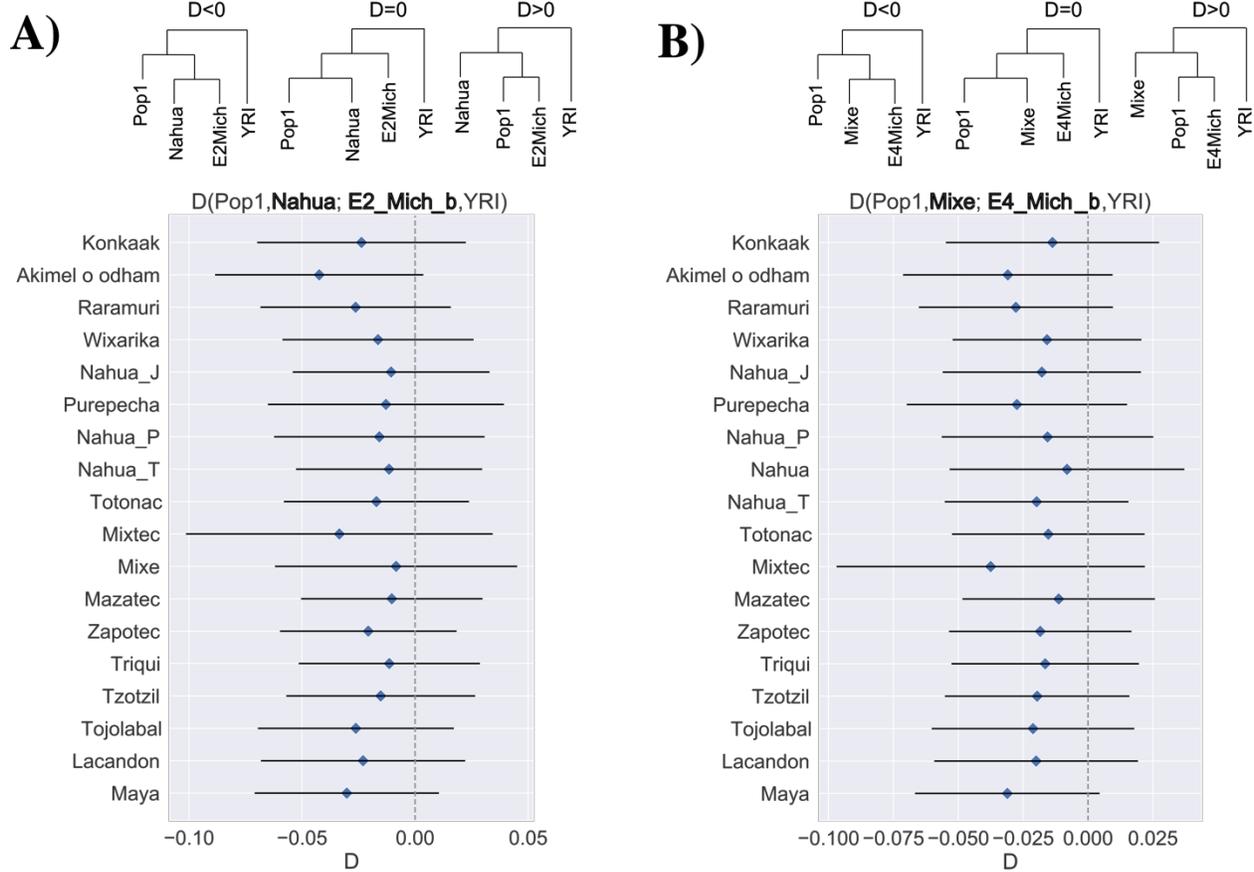


Fig. S18. D statistics for pre-Hispanic individuals from Michoacán and present-day Indigenous populations. (A) $D(\text{Pop1, Nahua; E2_Mich_b, YRI})$, (B) $D(\text{Pop1, Mixe; E4_Mich_b, YRI})$. Expected tree topologies according to D value are drawn on the top of the plot, individual IDs in the trees are indicated with no suffixes. Red dots indicate significant deviations from $D=0$ ($|Z|>3$). Individuals from Michoacán seem to be equally related to all present-day Indigenous populations.

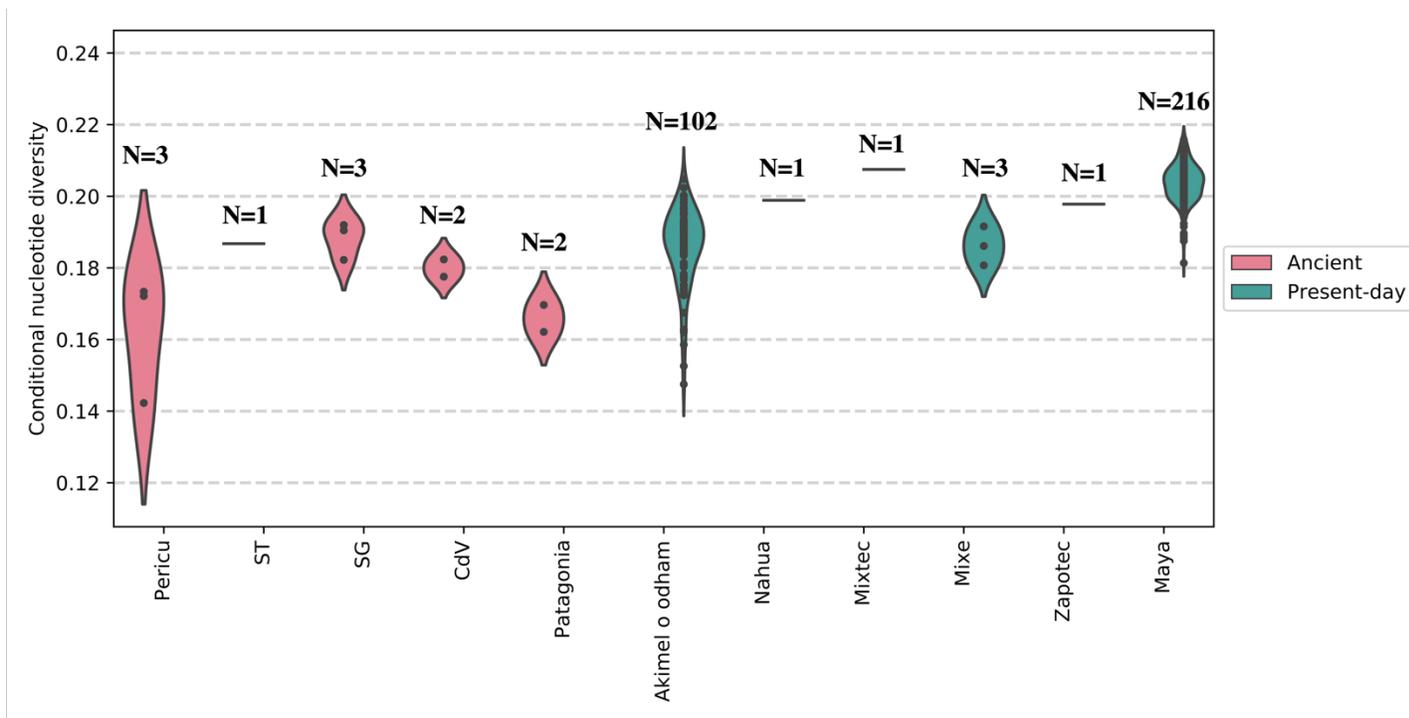
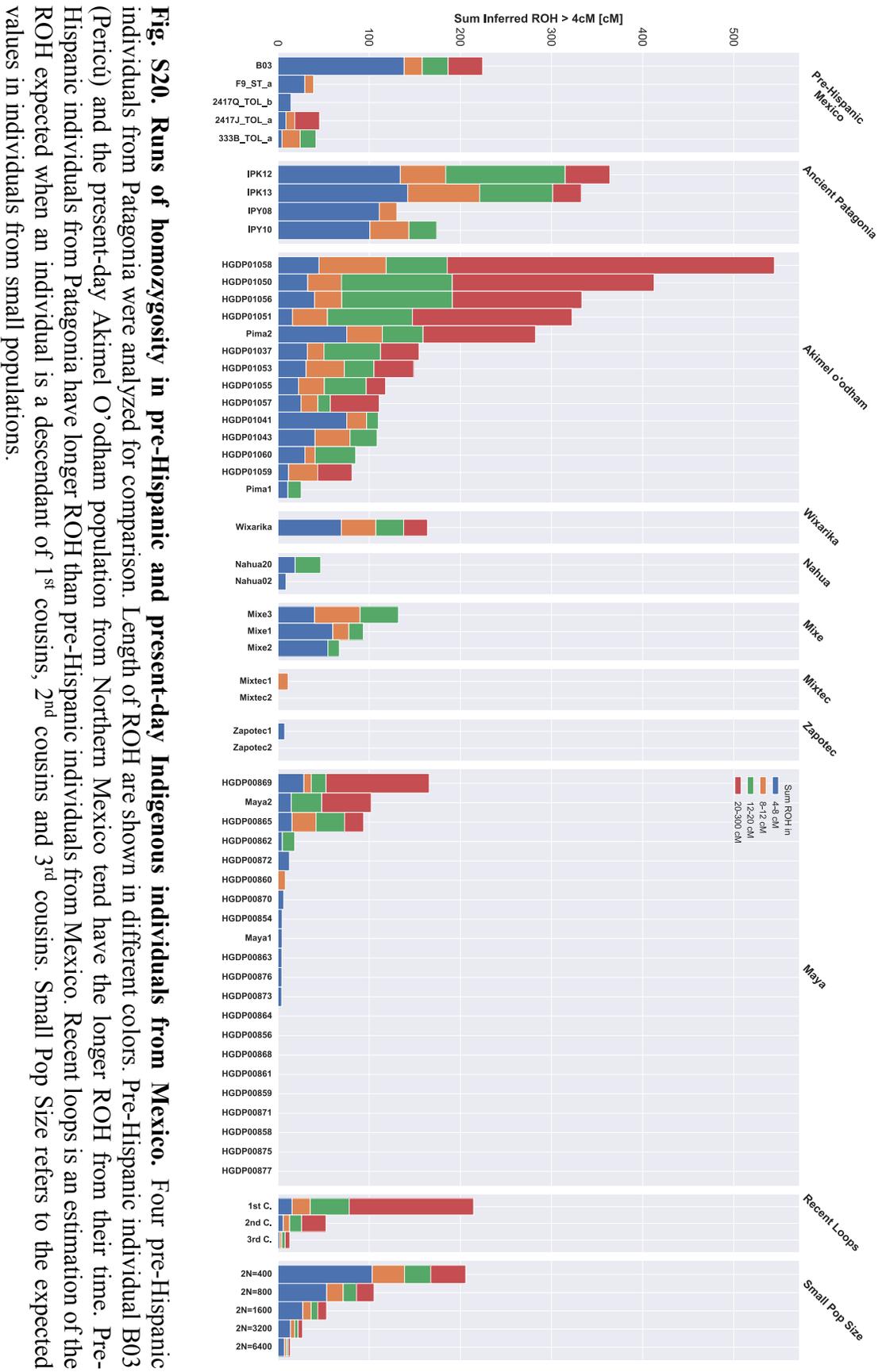


Fig. S19. Conditional Nucleotide Diversity of pairs of ancient pre-Hispanic individuals and present-day Indigenous populations from Mexico. Violin plot indicating the CND values estimated by pairs of individuals from the same archaeological site or population. The N value above each violin indicates the number of pairwise comparisons. Ancient pre-Hispanic populations are indicated in pink while present-day populations are in green. Up to each violin plot are indicated the number of comparisons (CND values estimated) for each population or archaeological site. ST (Sierra Tarahumara), SG (Sierra Gorda) and CdV (Cañada de la Virgen).



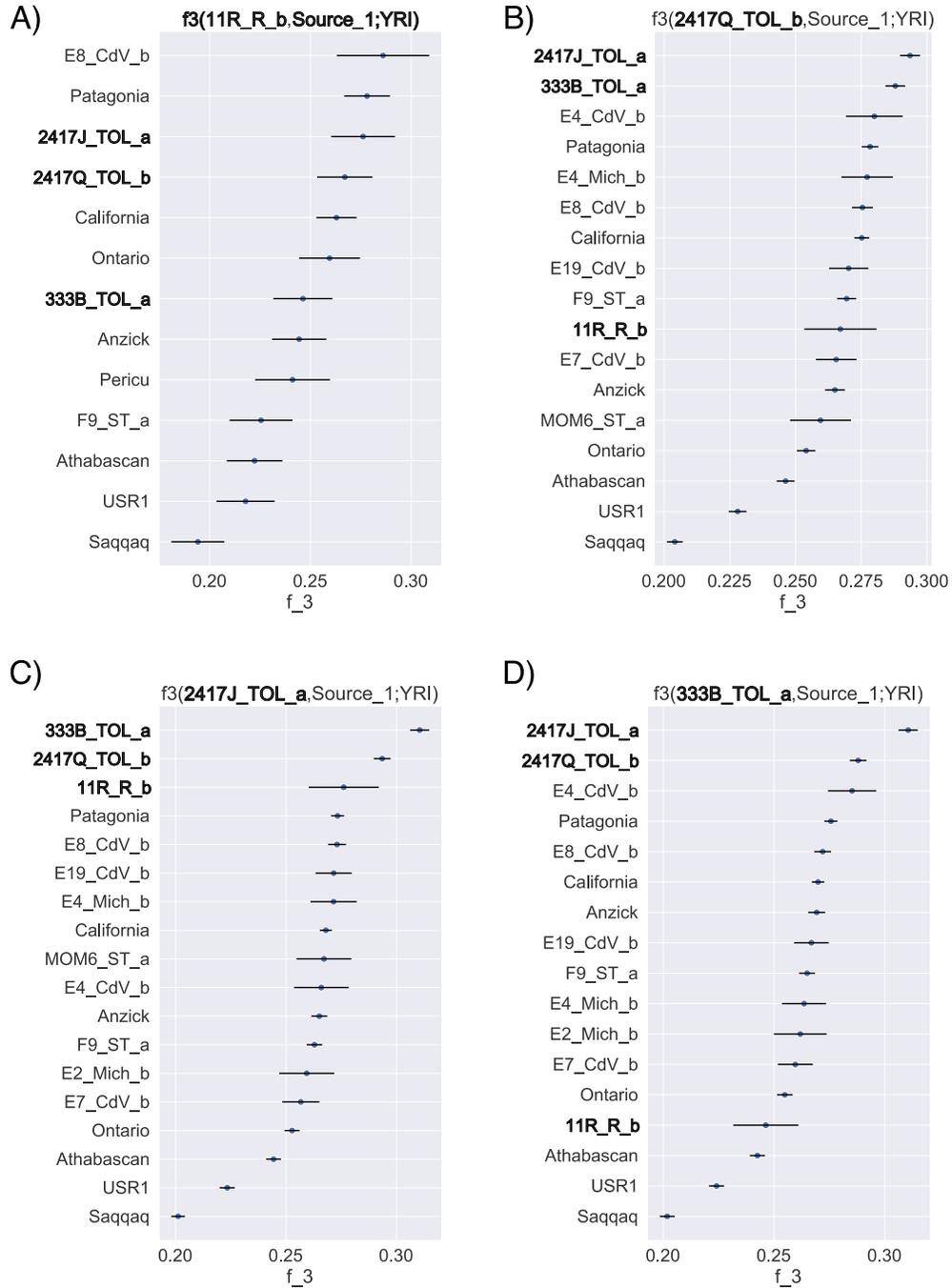


Fig. S21. Outgroup f_3 statistics for pre-Hispanic individuals from Sierra Gorda. (A) 11R_R_b, (B) 2417Q_TOL_b, (C) 2417J_TOL_a, and (D) 333B_TOL_a compared with ancient individuals from Mexico and the Americas. Sierra Gorda individuals are shown in bold. Higher values of f_3 indicate higher shared genetic drift. Point estimates and one standard error are shown.

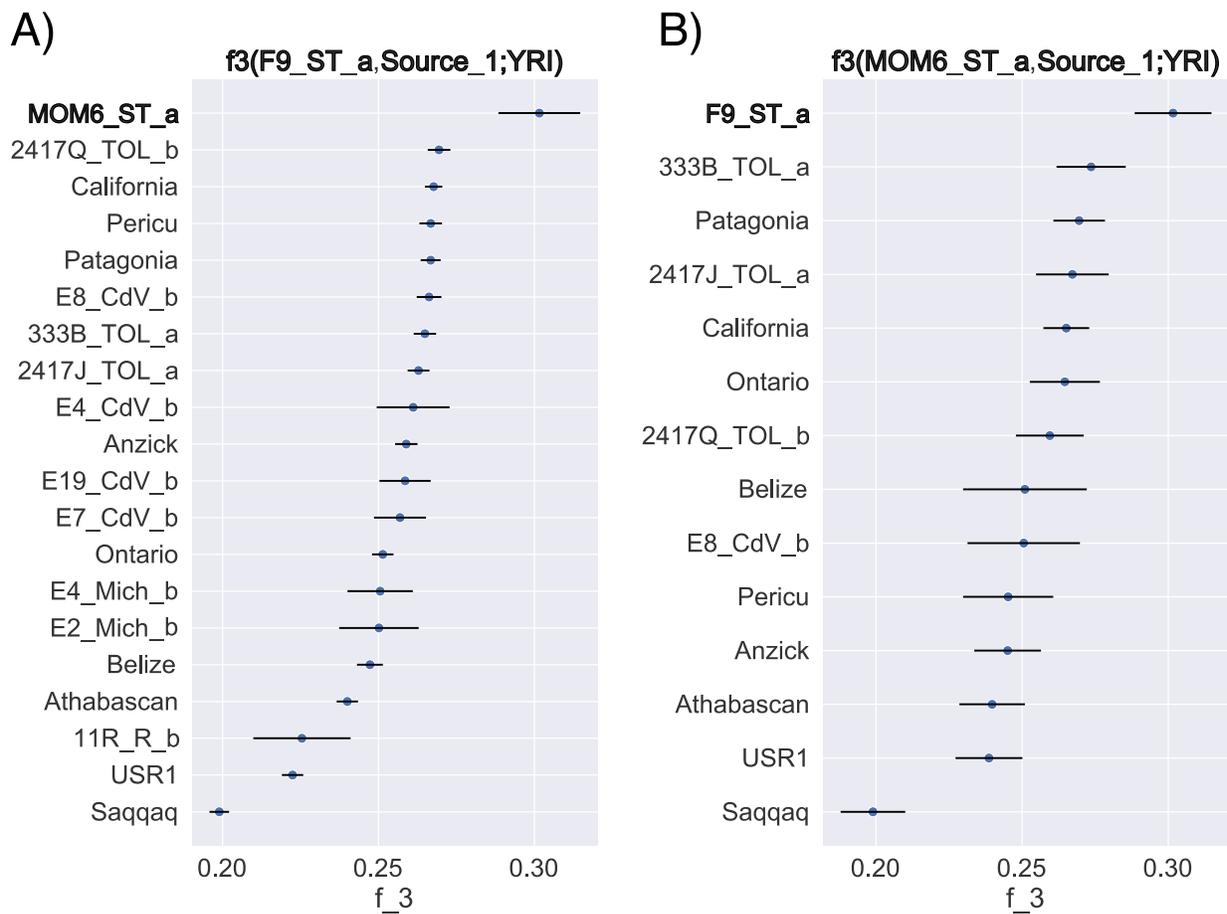


Fig. S22. Outgroup f_3 statistics for pre-Hispanic individuals from Sierra Tarahumara and other ancient Native Americans. (A) Individual **F9_ST_a** and (B) Mummy **MOM6_ST_a** compared with pre-Hispanic individuals from Mexico and the Americas. Individuals from Sierra Tarahumara are highlighted in bold. Higher values of f_3 indicate higher shared genetic drift.

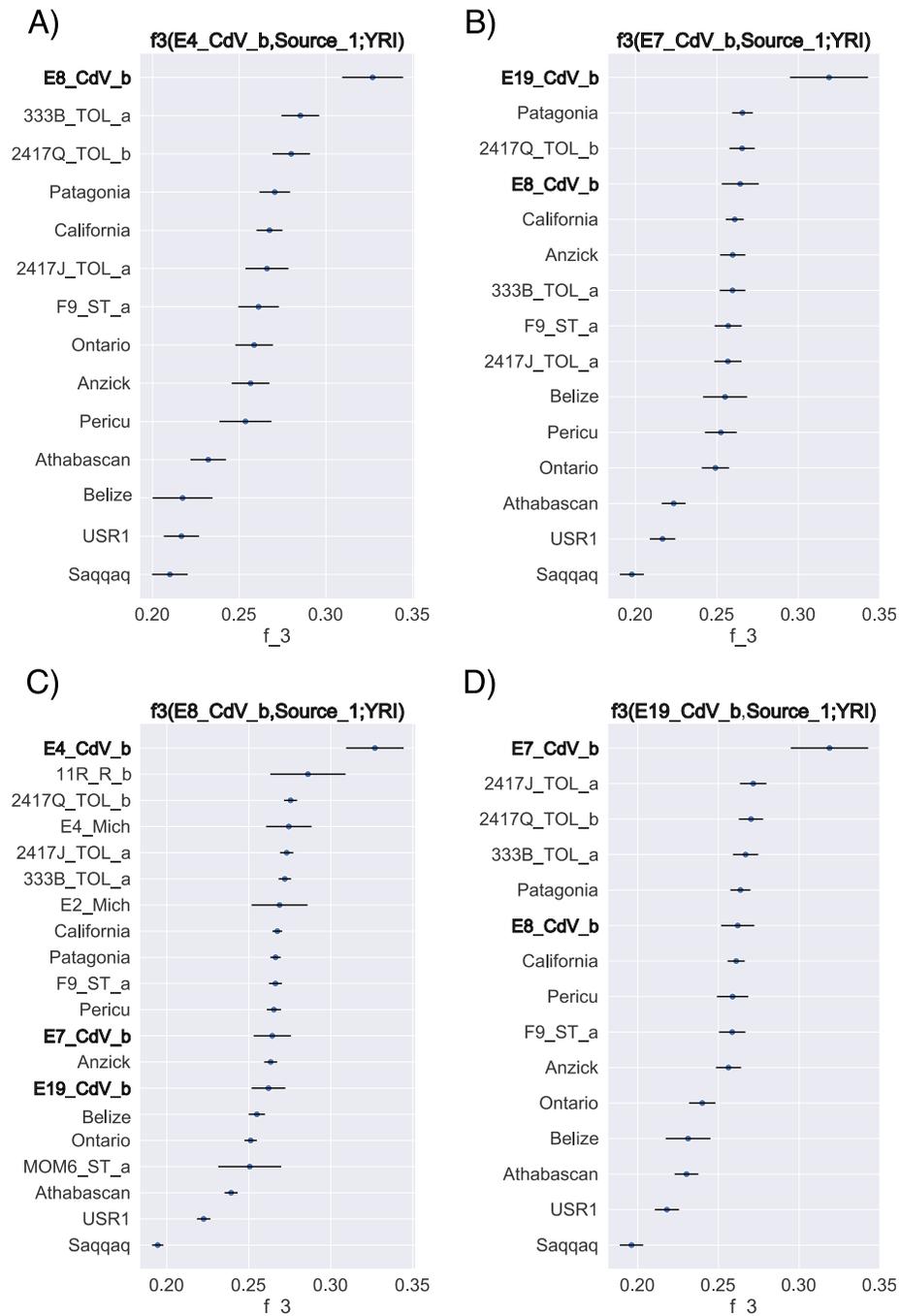


Fig. S23. Outgroup f_3 statistics for pre-Hispanic individuals from Cañada de la Virgen. (A) E4_CdV_b, (B) E7_CdV_b, (C) E8_CdV_b, and (D) E19_CdV_b compared with ancient individuals from Mexico and the Americas. Individuals from Cañada de la Virgen are highlighted in bold. Higher values of f_3 indicate higher shared genetic drift. Point estimates and one standard error are shown.

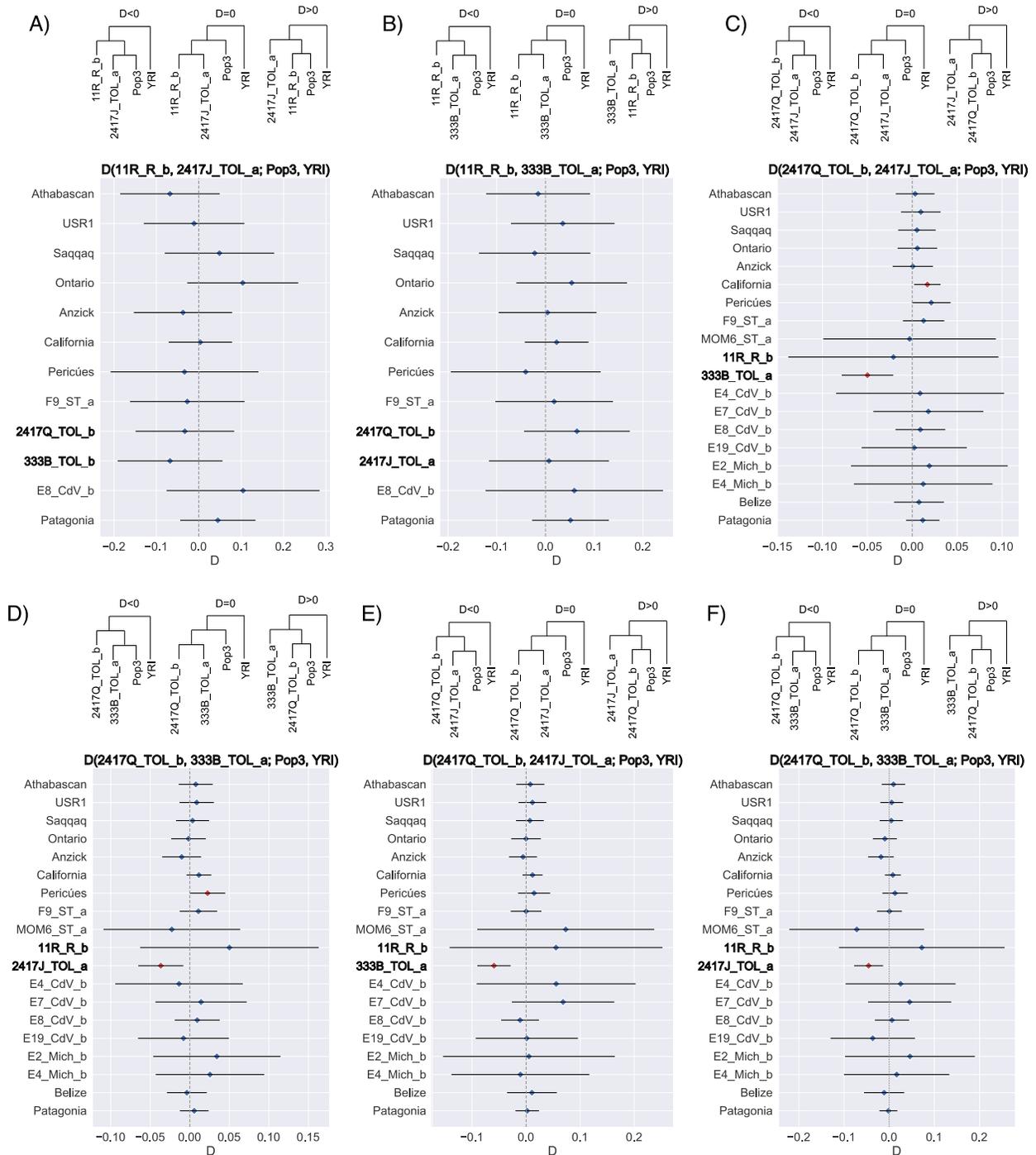


Fig. S24. D statistic values between the pre-Hispanic individuals from Sierra Gorda in the form D (SG_b, SG_a; Pop3, YRI). SG_b is a pre-drought pre-Hispanic individual from Sierra Gorda, SG_a is a post-drought pre-Hispanic individual from Sierra Gorda. Pop3 is any ancient individual from Mexico or the Americas and are shown in the Y axis. (A) Dstats in the form D (11R_R_b, 2417J_TOL_a; Pop3, YRI). (B) Dstats in the form D (11R_R_b, 333B_TOL_a; Pop3, YRI). (C) Dstats in the form D (2417Q_TOL_b, 2417J_TOL_a; Pop3, YRI). (D) Dstats in the

form D (2417Q_TOL_b, 333B_TOL_a; Pop3, YRI). (E) Dstats in the form D (2417Q_TOL_b, 2417J_TOL_a; Pop3, YRI), transversions only. (F) Dstats in the form D (2417Q_TOL_b, 333B_TOL_a; Pop3, YRI), transversions only. Expected tree topologies according to D value are drawn on the top of the plot. Red dots indicate significant deviations from $D=0$ ($|Z|>3$). Individuals from Sierra Gorda are highlighted in bold.

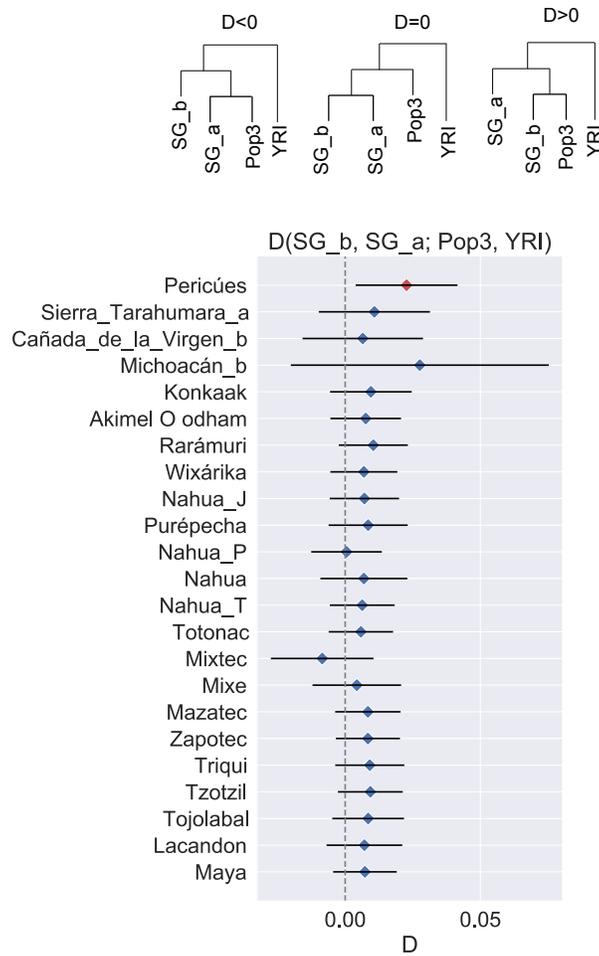
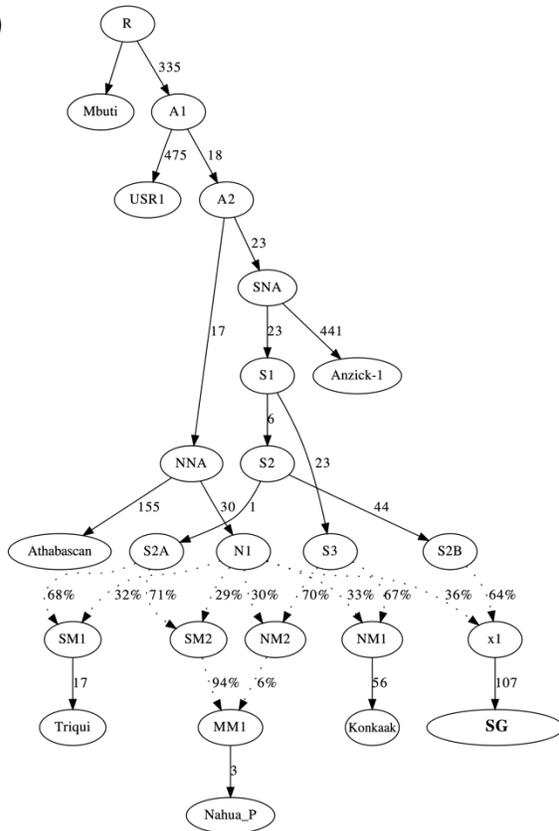


Fig. S25. D statistic values between the pre-Hispanic individuals from Sierra Gorda and pre-Hispanic/present-day Indigenous populations from Mexico. D statistics in the form $D(SG_b, SG_a; Pop3, YRI)$. SG_b includes the pre-Hispanic individuals 2417Q_TOL_b and 11R_R_b, while SG_a include the pre-Hispanic individuals 2417J_TOL_b and 333B_TOL_a, Pop3 is any of the other pre-Hispanic or present-day populations from Mexico and is shown in the Y axis. YRI as outgroup corresponds to Yoruba from Africa from the 1000 Genomes Project (37, 95). Expected tree topologies according to different D values are drawn on the top of the plot. Red dot indicates significant deviation from $D=0$ ($|Z|>3$).

A)



B)

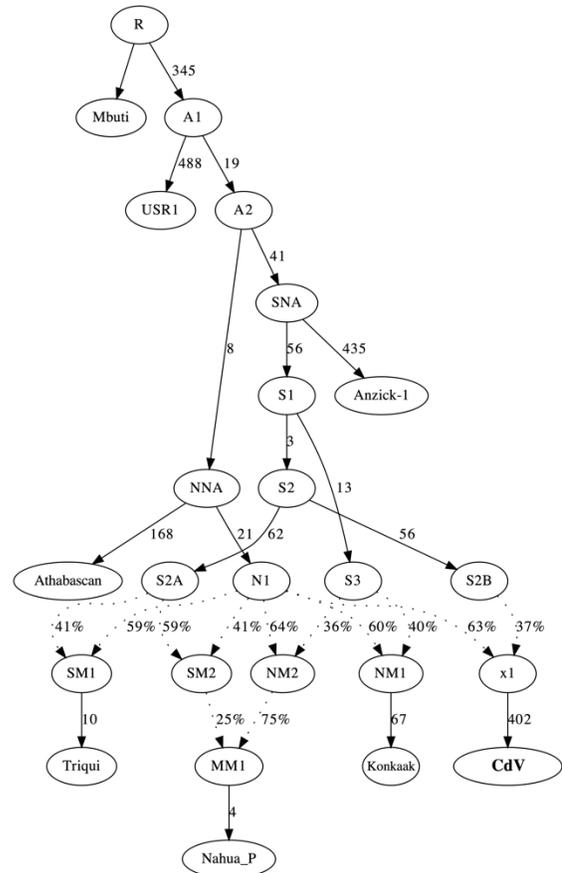


Fig. S26. Genetic ancestry proportions for the individuals from Central Mexico, inferred with qpGraph. Base model includes Anzick as a reference for the SNA and Athabaskan as a reference for the NNA branch. Konkaak was used a proxy for Northern Indigenous population from Mexico, Triqui as South proxy and Nahua_P as Central Mexico proxy. (A) Model for SG individuals, it includes the genetic information of 2417Q_TOL_b, 2417J_TOL_a and, 333B_TOL_a. The tree has a $|Z\text{-score}|$ of 1.319. (B) Model for CdV individuals, it includes the genetic information of E7_CdV_b, E8_CdV_b and E19_CdV_b. Individuals from Sierra Gorda (SG) and Cañada de la Virgen (CdV) show very different proportions of S2B and N1 ancestries. Tree with $|Z\text{-score}|$ of 2.821. All f statistics are within 1.75 standard error.

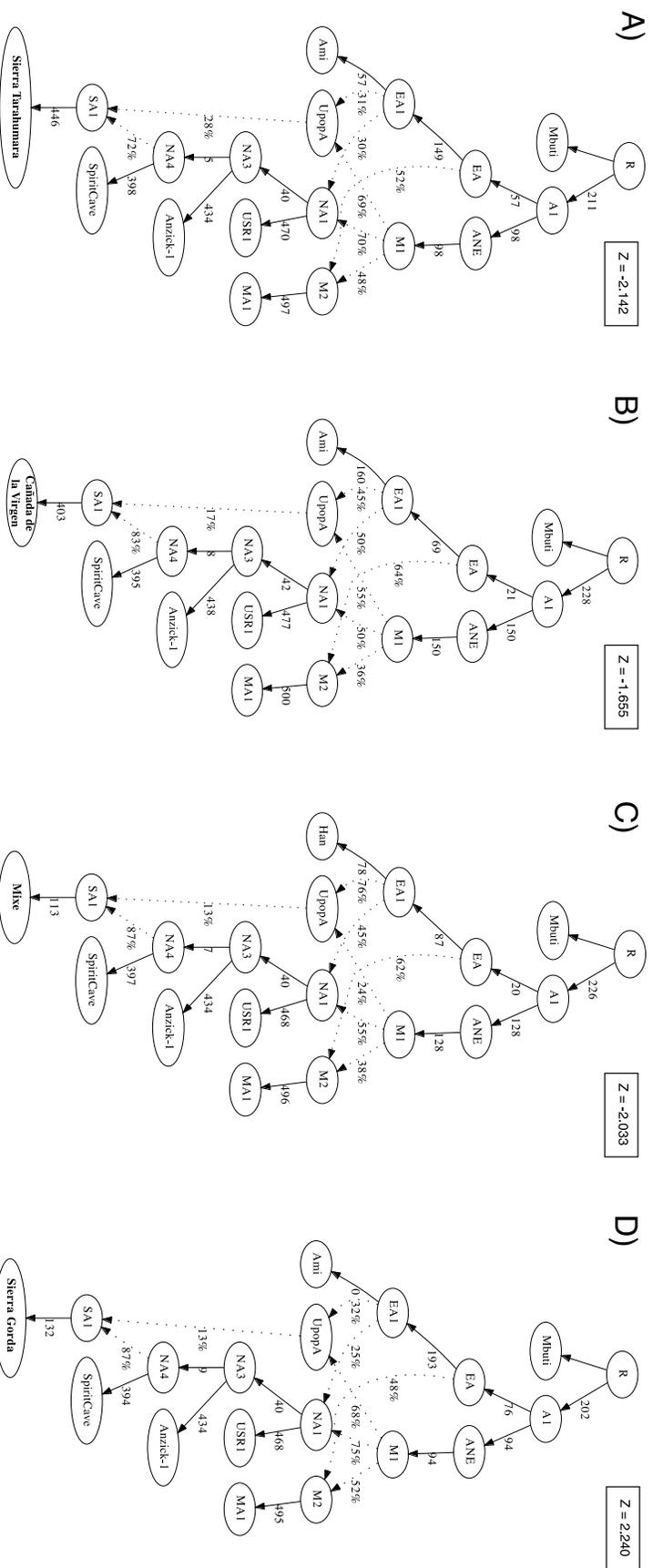


Fig. S27. Presence of ghost population genetic ancestry in pre-Hispanic from Sierra Tarahumara and Cañada de la Virgen, and present-day Mixe population inferred with qpGraph. Base model includes Mbuti as an outgroup, Ami as the East Asian, MA1 as the ancient North Eurasian, USR1 as the ancient Beringian, Anzick-1 and SpiritCave as the NNA. (A) Model with UpopA contribution to Sierra Tarahumara. (B) Model with UpopA contribution to Cañada de la Virgen. (C) Model with UpopA contribution to present-day Mixe. (D) Model rejected of UpopA contribution to present-day Sierra Gorda, with inner zeros in the branch to Ami.

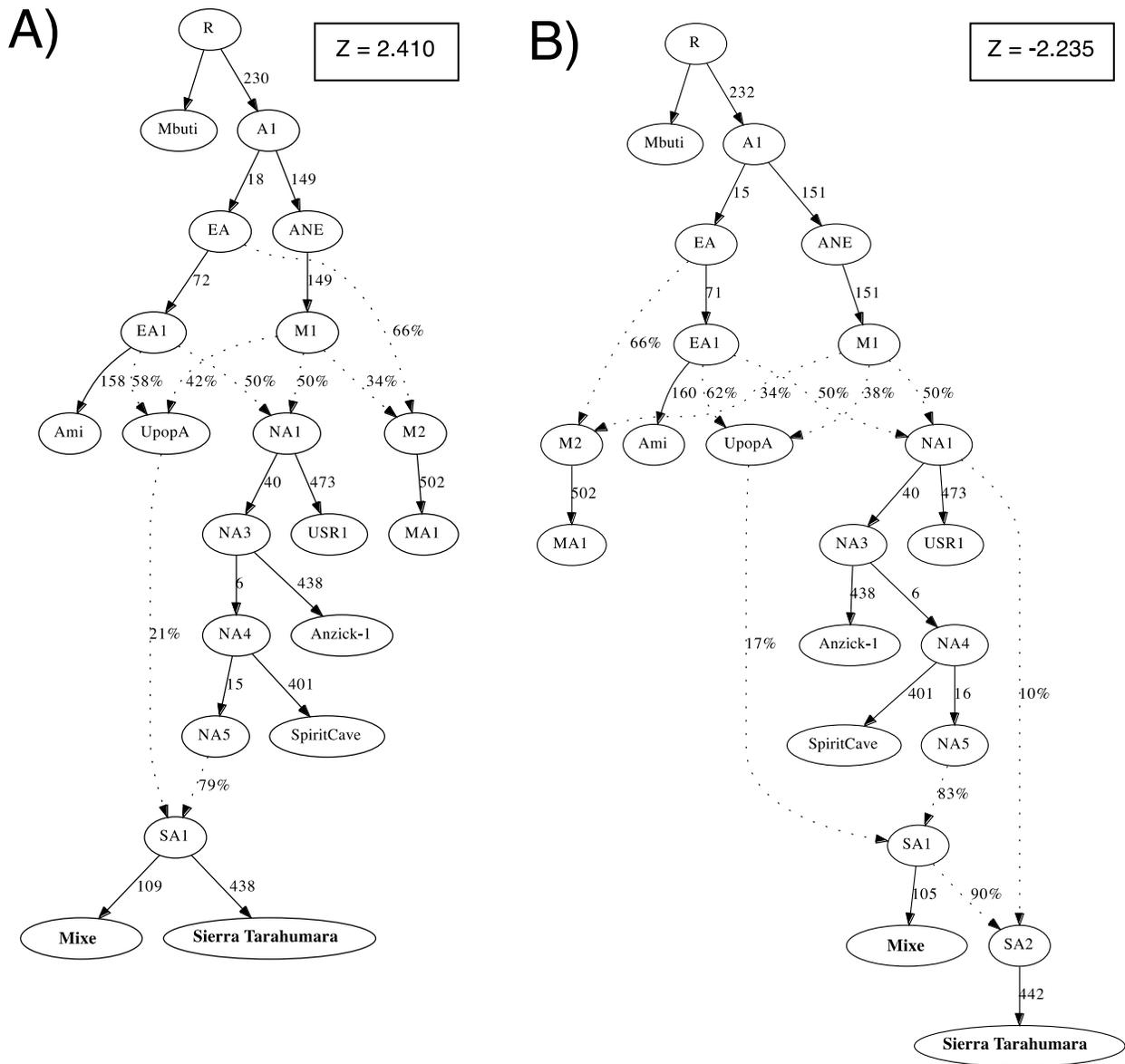


Fig. S28. Presence of ghost population genetic ancestry in pre-Hispanic from Sierra Tarahumara and present-day Mixe. Base model includes Mbuti as an outgroup, Ami as the East Asian, MA1 as the ancient North Eurasian, USR1 as the ancient Beringian, Anzick-1 and Spirit Cave as the NNA. (A) Model with UpopA contribution to present-day Mixe and Sierra Tarahumara in a single clade. (B) Model with UpopA contribution to both present-day Mixe and Sierra Tarahumara in a different clade.

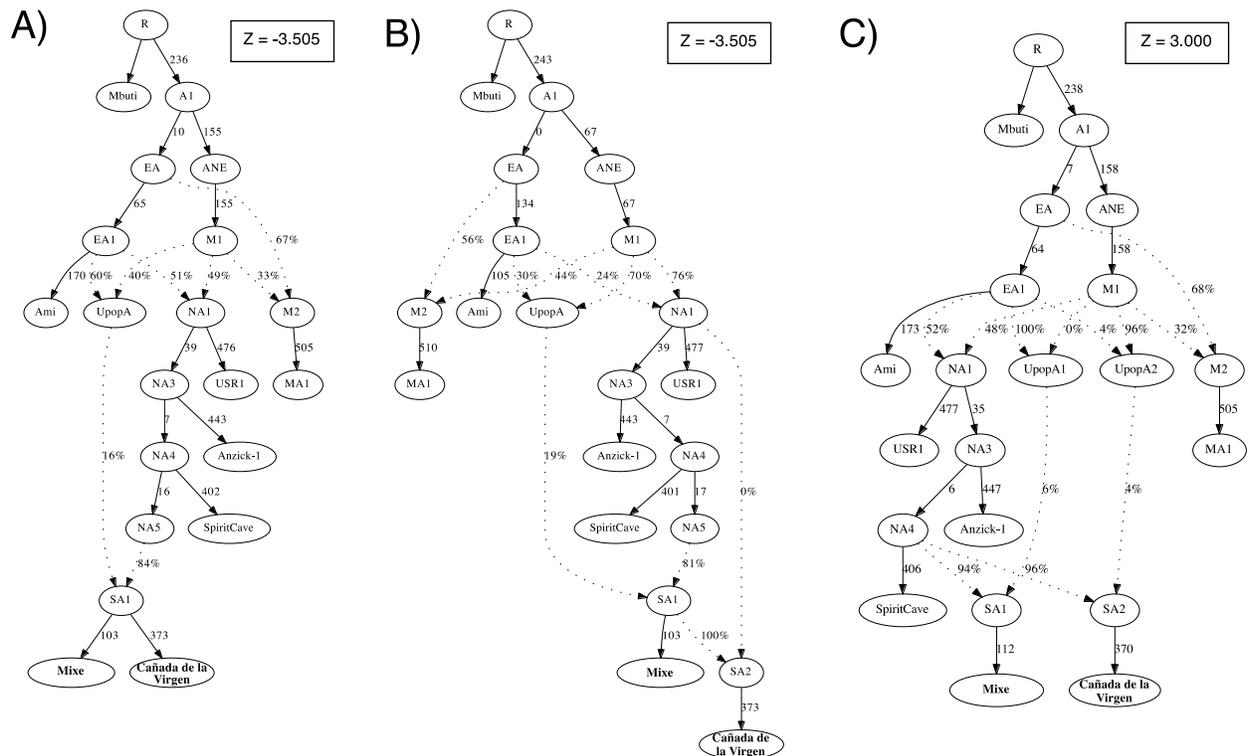


Fig. S29. Presence of ghost population genetic ancestry in pre-Hispanic from Cañada de la Virgen. Base model includes Mbuti as an outgroup, Ami as the East Asian, MA1 as the ancient North Eurasian, USR1 as the ancient Beringian, Anzick-1 and Spirit Cave as the NNA. (A) The model with UpopA contribution to present-day Mixe and Cañada de la Virgen in a single clade is rejected. (B) The model with UpopA contribution to both present-day Mixe and Cañada de la Virgen in a different clade is also rejected. (C) The model with UpopA contributing to Mixe and a second "ghost" population (UpopA2) contributing to Cañada de la Virgen is accepted.

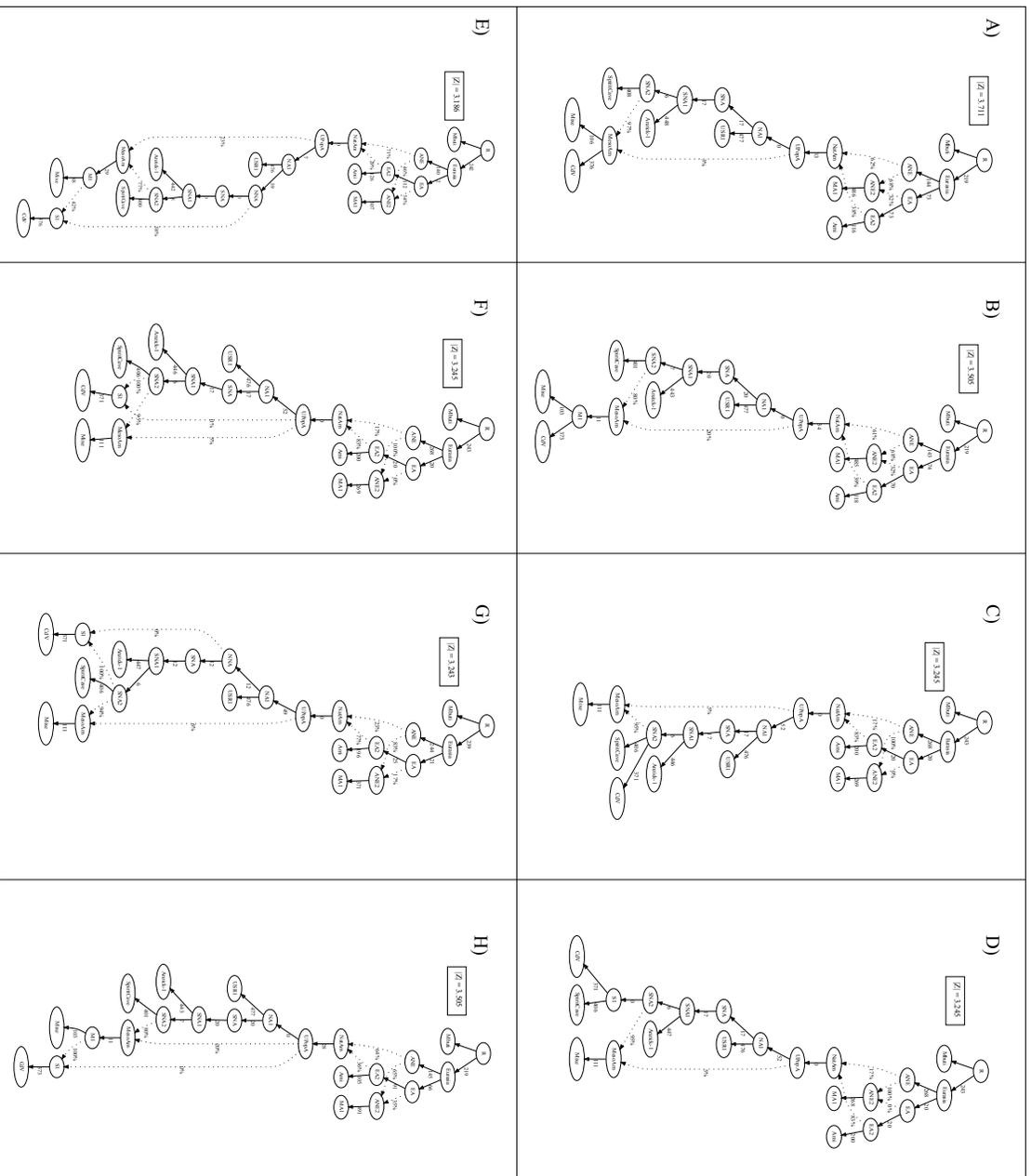


Fig. S30. qpGraphs modeling the ghost genetic ancestry in Cañada de la Virgen (CdV) and UPopA modeled with Native American clade as its ancestor. (A) Model with Cañada de la Virgen and Mixe in the same clade. (B) Same as A but including an intermediate lineage 'MI' between Mesoamerica, Mixe and Cañada de la Virgen. (C) Model with Cañada de la Virgen and SpiritCave in the same clade. (D) Same as C but including an intermediate

lineage 'S1' between SNA2, SpiritCave and Cañada de la Virgen. (E) Model with Cañada de la Virgen coming from 'M1' with a contribution of NNA. (F) Model with Cañada de la Virgen and SpiritCave in the same clade with a contribution of UPopA. (G) Model with Cañada de la Virgen and SpiritCave in the same clade with a contribution of NNA. (H) Model with Cañada de la Virgen coming from the 'M1' lineage with an intermediate 'S1' and receiving contribution of UPopA. All models are rejected.

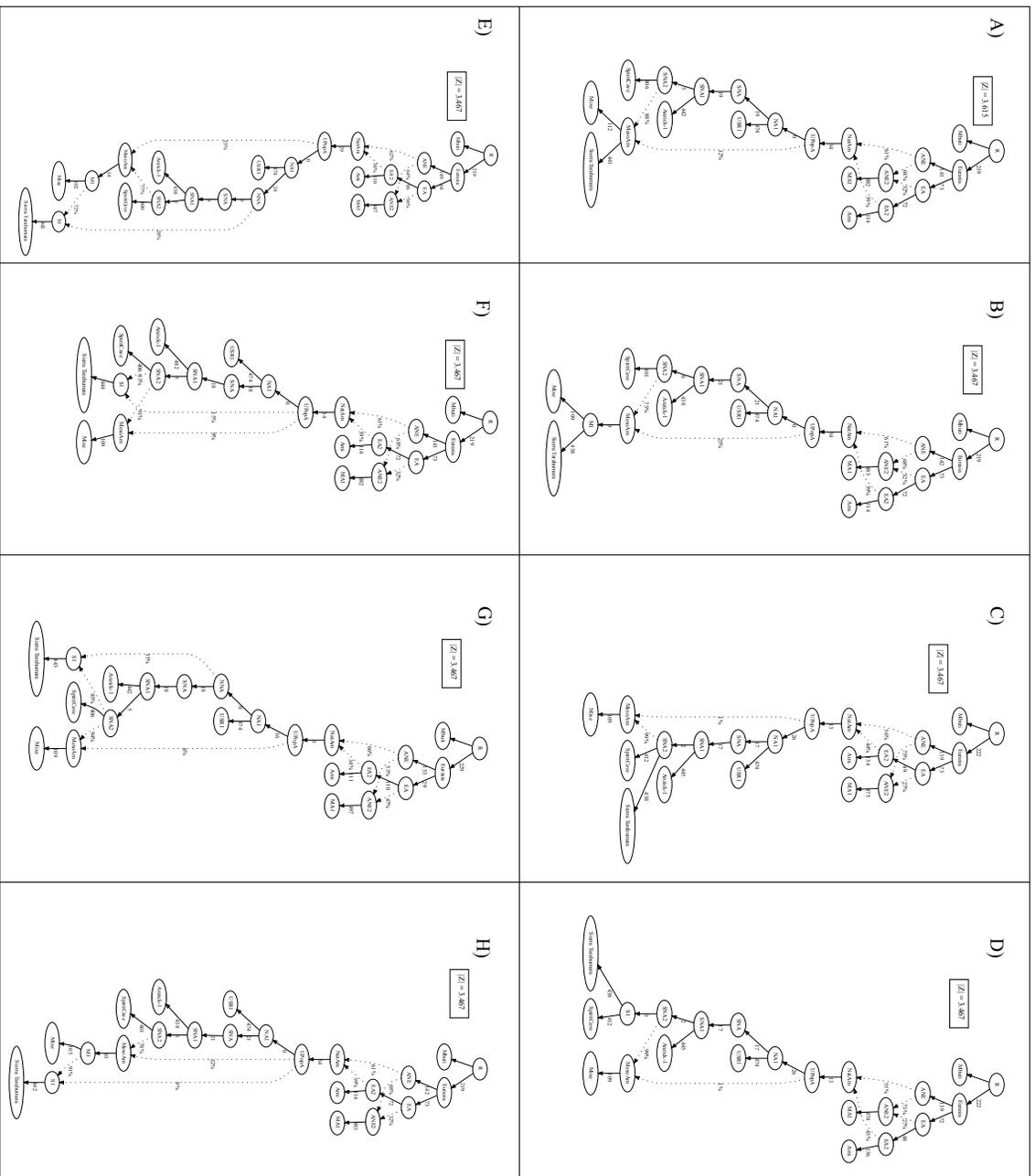


Fig.S31. qpGraphs modeling the ghost genetic ancestry in Sierra Tarahumara and U'PopeA modeled with Native American clade as its ancestor. (A) Model with Sierra Tarahumara and Mixe in the same clade. (B) Same as A but including an intermediate lineage 'MI' between Mesoamerica, Mixe and Sierra Tarahumara. (C) Model with Sierra Tarahumara and SpiritCave in the same clade. (D) Same as C but including an intermediate lineage 'SI'

between SNA2, SpiritCave and Sierra Tarahumara. (E) Model with Sierra Tarahumara coming from 'M1' with a contribution of NNA. (F) Model with Sierra Tarahumara and SpiritCave in the same clade with a contribution of UPopA. (G) Model with Sierra Tarahumara and SpiritCave in the same clade with a contribution of NNA. (H) Model with Sierra Tarahumara coming from the 'M1' lineage with an intermediate 'S1' and receiving contribution of UPopA. All models are rejected.

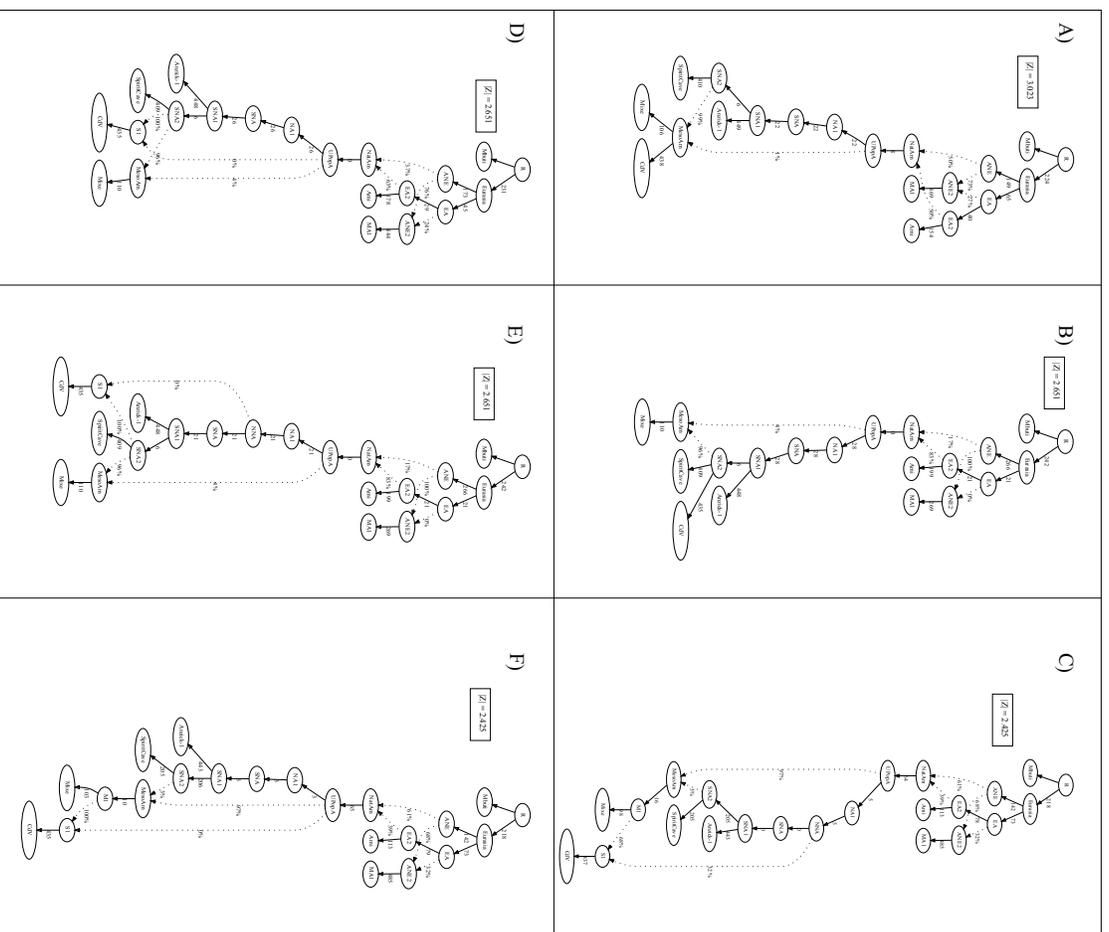


Fig. S32. qpGraphs modeling the ghost genetic ancestry in Cañada de la Virgen (CdV), UPopA modeled with Native American clade as its ancestor and excluding USR1. (A) Cañada de la Virgen and Mixe in the same clade. (B) Cañada de la Virgen and SpirtCave in the same clade. (C) Cañada de la Virgen coming from 'MI' with a

contribution of NNA. (D) Cañada de la Virgen and SpiritCave in the same clade with a contribution of UPopA. (E) Cañada de la Virgen and SpiritCave in the same clade with a contribution of NNA. (F) Cañada de la Virgen coming from the 'M1' lineage with an intermediate 'S1' and receiving contribution of UPopA.

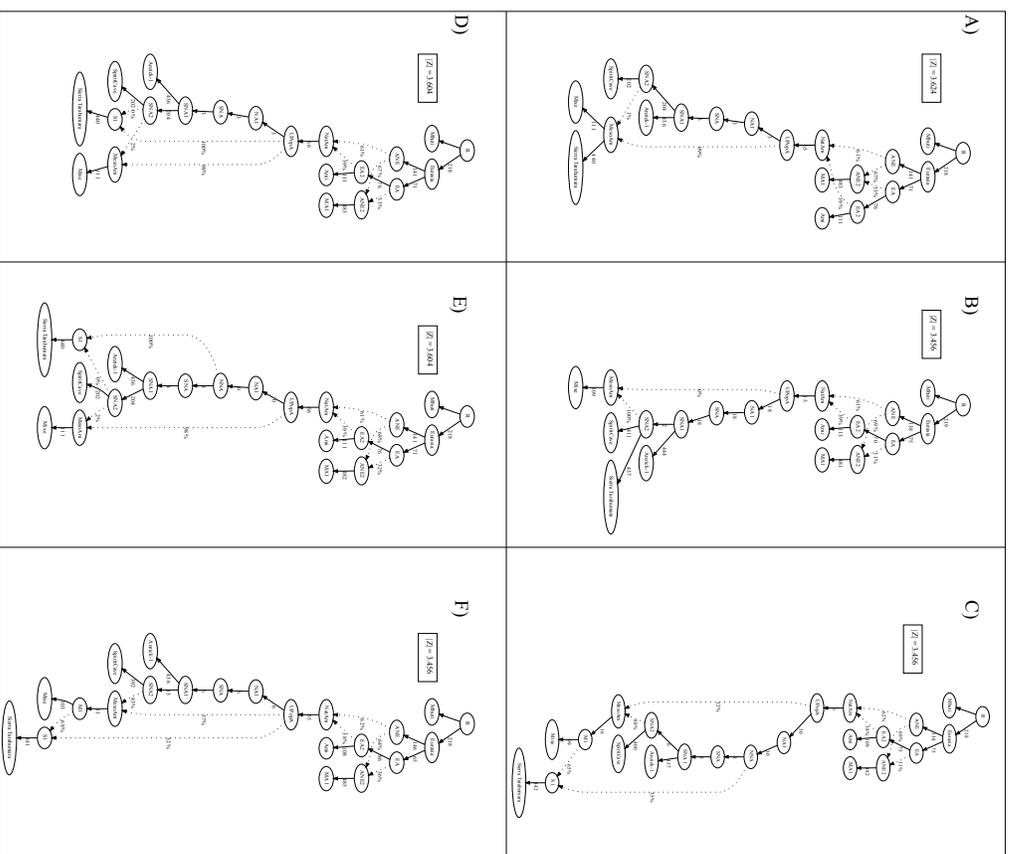


Fig. S33. qpGraphs modeling the ghost genetic ancestry in Sierra Tarahumara, UPopA modeled with Native American clade as its ancestor and excluding USR1. (A) Sierra Tarahumara and Mixe in the same clade. (B) Sierra Tarahumara and SpiritCave in the same clade. (C) Sierra Tarahumara coming from 'M1' with a contribution of NNA. (D) Sierra Tarahumara and SpiritCave in the same clade with a contribution of UPopA. (E) Sierra Tarahumara and SpiritCave in the same clade with a contribution of NNA. (F) Sierra Tarahumara coming from the 'M1' lineage with an intermediate 'SI' and receiving contribution of UPopA.

References and Notes

1. A. L. Austin, L. L. Luján, *El pasado indígena* (Fondo de Cultura Económica, 2019).
2. B. B. Cornejo, *La frontera septentrional de Mesoamérica* (Instituto Nacional de Antropología e Historia, México, ed. 2, 2000), vol. 1 of *Historia Antigua de México*.
3. P. Kirchhoff, Mesoamerica: its geographic limits, ethnic composition and cultural characteristics. *Heritage of conquest*, pp. 17–30 (1952).
4. M. del C. Solanes, E. Vela, Atlas del México prehispánico. *Arqueol. Mex.* **5**, 5–18 (2000).
5. D. W. Stahle, J. V. Diaz, D. J. Burnette, J. C. Paredes, R. R. Heim Jr., F. K. Fye, R. Acuna Soto, M. D. Therrell, M. K. Cleaveland, D. K. Stahle, Major Mesoamerican droughts of the past millennium. *Geophys. Res. Lett.* **38**, L05703 (2011). [doi:10.1029/2010GL046472](https://doi.org/10.1029/2010GL046472)
6. M. K. Hughes, H. F. Diaz, Was there a ‘medieval warm period’, and if so, where and when? *Clim. Change* **26**, 109–142 (1994). [doi:10.1007/BF01092410](https://doi.org/10.1007/BF01092410)
7. P. Armillas, Condiciones ambientales y movimientos de pueblos en la frontera septentrional de Mesoamérica. *Homenaje a Fernando Márquez-Miranda*, 62–82 (1964).
8. K. H. Wogau, H. W. Arz, H. N. Böhnel, N. R. Nowaczyk, J. Park, High resolution paleoclimate and paleoenvironmental reconstruction in the Northern Mesoamerican Frontier for Prehistory to Historical times. *Quat. Sci. Rev.* **226**, 106001 (2019). [doi:10.1016/j.quascirev.2019.106001](https://doi.org/10.1016/j.quascirev.2019.106001)
9. D. J. Kennett, S. F. M. Breitenbach, V. V. Aquino, Y. Asmerom, J. Awe, J. U. L. Baldini, P. Bartlein, B. J. Culleton, C. Ebert, C. Jazwa, M. J. Macri, N. Marwan, V. Polyak, K. M. Prufer, H. E. Ridley, H. Sodemann, B. Winterhalder, G. H. Haug, Development and disintegration of Maya political systems in response to climate change. *Science* **338**, 788–791 (2012). [doi:10.1126/science.1226299](https://doi.org/10.1126/science.1226299) [Medline](#)
10. J. A. Hoggarth, S. F. M. Breitenbach, B. J. Culleton, C. E. Ebert, M. A. Masson, D. J. Kennett, The political collapse of Chichén Itzá in climatic and cultural context. *Global Planet. Change* **138**, 25–42 (2016). [doi:10.1016/j.gloplacha.2015.12.007](https://doi.org/10.1016/j.gloplacha.2015.12.007)
11. S. L. O’Hara, S. E. Metcalfe, F. A. Street-Perrott, On the arid margin: The relationship between climate, humans and the environment. A review of evidence from the highlands of central Mexico. *Chemosphere* **29**, 965–981 (1994). [doi:10.1016/0045-6535\(94\)90163-5](https://doi.org/10.1016/0045-6535(94)90163-5) [Medline](#)
12. B. Braniff, Oscilación de la frontera norte mesoamericana: un nuevo ensayo. *Arqueología, segunda época*, pp. 99–114 (1989).
13. A. Gómez Aíza, S. Sánchez Vázquez, La Frontera Cultural Meso-Aridoamericana: Construcción De Imaginarios Nacionalistas En La Historia Mexicana. *Xihmai* **9**, (2014). [doi:10.37646/xihmai.v9i18.244](https://doi.org/10.37646/xihmai.v9i18.244)
14. M. Raghavan, M. Steinrücken, K. Harris, S. Schiffels, S. Rasmussen, M. DeGiorgio, A. Albrechtsen, C. Valdiosera, M. C. Ávila-Arcos, A.-S. Malaspina, A. Eriksson, I. Moltke, M. Metspalu, J. R. Homburger, J. Wall, O. E. Cornejo, J. V. Moreno-Mayar, T. S. Korneliusen, T. Pierre, M. Rasmussen, P. F. Campos, P. de Barros Damgaard, M. E. Allentoft, J. Lindo, E. Metspalu, R. Rodríguez-Varela, J. Mansilla, C. Henriksson, A.

- Seguin-Orlando, H. Malmström, T. Stafford Jr., S. S. Shringarpure, A. Moreno-Estrada, M. Karmin, K. Tambets, A. Bergström, Y. Xue, V. Warmuth, A. D. Friend, J. Singarayer, P. Valdes, F. Balloux, I. Lebreiro, J. L. Vera, H. Rangel-Villalobos, D. Pettener, D. Luiselli, L. G. Davis, E. Heyer, C. P. E. Zollikofer, M. S. Ponce de León, C. I. Smith, V. Grimes, K.-A. Pike, M. Deal, B. T. Fuller, B. Arriaza, V. Standen, M. F. Luz, F. Ricaut, N. Guidon, L. Osipova, M. I. Voevoda, O. L. Posukh, O. Balanovsky, M. Lavryashina, Y. Bogunov, E. Khusnutdinova, M. Gubina, E. Balanovska, S. Fedorova, S. Litvinov, B. Malyarchuk, M. Derenko, M. J. Mosher, D. Archer, J. Cybulski, B. Petzelt, J. Mitchell, R. Worl, P. J. Norman, P. Parham, B. M. Kemp, T. Kivisild, C. Tyler-Smith, M. S. Sandhu, M. Crawford, R. Villems, D. G. Smith, M. R. Waters, T. Goebel, J. R. Johnson, R. S. Malhi, M. Jakobsson, D. J. Meltzer, A. Manica, R. Durbin, C. D. Bustamante, Y. S. Song, R. Nielsen, E. Willerslev, Genomic evidence for the Pleistocene and recent population history of Native Americans. *Science* **349**, aab3884–aab3884 (2015). [doi:10.1126/science.aab3884](https://doi.org/10.1126/science.aab3884) [Medline](#)
15. C. L. Scheib, H. Li, T. Desai, V. Link, C. Kendall, G. Dewar, P. W. Griffith, A. Mörseburg, J. R. Johnson, A. Potter, S. L. Kerr, P. Endicott, J. Lindo, M. Haber, Y. Xue, C. Tyler-Smith, M. S. Sandhu, J. G. Lorenz, T. D. Randall, Z. Faltyskova, L. Pagani, P. Danecek, T. C. O’Connell, P. Martz, A. S. Boraas, B. F. Byrd, A. Leventhal, R. Cambra, R. Williamson, L. Lesage, B. Holguin, E. Ygnacio-De Soto, J. Rosas, M. Metspalu, J. T. Stock, A. Manica, A. Scally, D. Wegmann, R. S. Malhi, T. Kivisild, Ancient human parallel lineages within North America contributed to a coastal expansion. *Science* **360**, 1024–1027 (2018). [doi:10.1126/science.aar6851](https://doi.org/10.1126/science.aar6851) [Medline](#)
16. M. C. Ávila-Arcos, C. de la Fuente Castro, M. A. Nieves-Colón, M. Raghavan, Recommendations for Sustainable Ancient DNA Research in the Global South: Voices From a New Generation of Paleogenomicists. *Front. Genet.* **13**, 880170 (2022). [doi:10.3389/fgene.2022.880170](https://doi.org/10.3389/fgene.2022.880170) [Medline](#)
17. Materials and methods are available as supplementary materials.
18. S. Mallick, H. Li, M. Lipson, I. Mathieson, M. Gymrek, F. Racimo, M. Zhao, N. Chennagiri, S. Nordenfelt, A. Tandon, P. Skoglund, I. Lazaridis, S. Sankararaman, Q. Fu, N. Rohland, G. Renaud, Y. Erlich, T. Willems, C. Gallo, J. P. Spence, Y. S. Song, G. Poletti, F. Balloux, G. van Driem, P. de Knijff, I. G. Romero, A. R. Jha, D. M. Behar, C. M. Bravi, C. Capelli, T. Hervig, A. Moreno-Estrada, O. L. Posukh, E. Balanovska, O. Balanovsky, S. Karachanak-Yankova, H. Sahakyan, D. Toncheva, L. Yepiskoposyan, C. Tyler-Smith, Y. Xue, M. S. Abdullah, A. Ruiz-Linares, C. M. Beall, A. Di Rienzo, C. Jeong, E. B. Starikovskaya, E. Metspalu, J. Parik, R. Villems, B. M. Henn, U. Hodoglugil, R. Mahley, A. Sajantila, G. Stamatoyannopoulos, J. T. S. Wee, R. Khusainova, E. Khusnutdinova, S. Litvinov, G. Ayodo, D. Comas, M. F. Hammer, T. Kivisild, W. Klitz, C. A. Winkler, D. Labuda, M. Bamshad, L. B. Jorde, S. A. Tishkoff, W. S. Watkins, M. Metspalu, S. Dryomov, R. Sukernik, L. Singh, K. Thangaraj, S. Pääbo, J. Kelso, N. Patterson, D. Reich, The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* **538**, 201–206 (2016). [doi:10.1038/nature18964](https://doi.org/10.1038/nature18964) [Medline](#)

19. I. De La Cruz, A. González-Oliver, B. M. Kemp, J. A. Román, D. G. Smith, A. Torre-Blanco, Sex Identification of Children Sacrificed to the Ancient Aztec Rain Gods in Tlatelolco. *Curr. Anthropol.* **49**, 519–526 (2008). [doi:10.1086/587642](https://doi.org/10.1086/587642)
20. D. Perez-Benedico, J. La Salvia, Z. Zeng, G. A. Herrera, R. Garcia-Bertrand, R. J. Herrera, Mayans: A Y chromosome perspective. *Eur. J. Hum. Genet.* **24**, 1352–1358 (2016). [doi:10.1038/ejhg.2016.18](https://doi.org/10.1038/ejhg.2016.18) [Medline](#)
21. K. Sandoval, A. Moreno-Estrada, I. Mendizabal, P. A. Underhill, M. Lopez-Valenzuela, R. Peñaloza-Espinosa, M. Lopez-Lopez, L. Buentello-Malo, H. Avelino, F. Calafell, D. Comas, Y-chromosome diversity in Native Mexicans reveals continental transition of genetic structure in the Americas. *Am. J. Phys. Anthropol.* **148**, 395–405 (2012). [doi:10.1002/ajpa.22062](https://doi.org/10.1002/ajpa.22062) [Medline](#)
22. A. Bergström, S. A. McCarthy, R. Hui, M. A. Almarri, Q. Ayub, P. Danecek, Y. Chen, S. Felkel, P. Hallast, J. Kamm, H. Blanché, J.-F. Deleuze, H. Cann, S. Mallick, D. Reich, M. S. Sandhu, P. Skoglund, A. Scally, Y. Xue, R. Durbin, C. Tyler-Smith, Insights into human genetic variation and population history from 929 diverse genomes. *Science* **367**, eaay5012 (2020). [doi:10.1126/science.aay5012](https://doi.org/10.1126/science.aay5012) [Medline](#)
23. B. Z. González-Sobrino, A. P. Pintado-Cortina, L. Sebastián, F. Morales-Mandujano, A. V. Contreras, J. Chávez-Benavides, *et al.*, Genetic Diversity and Differentiation in Urban and Indigenous Populations of Mexico: Patterns of Mitochondrial DNA and Y-Chromosome Lineage. *Biodemography Soc. Biol.* **62**, 53–72 (2016).
24. H.-J. Bandelt, C. Herrnstadt, Y.-G. Yao, Q.-P. Kong, T. Kivisild, C. Rengo, R. Scozzari, M. Richards, R. Villems, V. Macaulay, N. Howell, A. Torroni, Y.-P. Zhang, Identification of Native American founder mtDNAs through the analysis of complete mtDNA sequences: Some caveats. *Ann. Hum. Genet.* **67**, 512–524 (2003). [doi:10.1046/j.1469-1809.2003.00049.x](https://doi.org/10.1046/j.1469-1809.2003.00049.x) [Medline](#)
25. A. Y. Morales-Arce, M. H. Snow, J. H. Kelley, M. Anne Katzenberg, Ancient mitochondrial DNA and ancestry of Paquimé inhabitants, Casas Grandes (A.D. 1200-1450). *Am. J. Phys. Anthropol.* **163**, 616–626 (2017). [doi:10.1002/ajpa.23223](https://doi.org/10.1002/ajpa.23223) [Medline](#)
26. J. Mata-Míguez, L. Overholtzer, E. Rodríguez-Alegría, B. M. Kemp, D. A. Bolnick, The genetic impact of Aztec imperialism: Ancient mitochondrial DNA evidence from Xaltocan, Mexico. *Am. J. Phys. Anthropol.* **149**, 504–516 (2012). [doi:10.1002/ajpa.22152](https://doi.org/10.1002/ajpa.22152) [Medline](#)
27. A. Y. Morales-Arce, G. McCafferty, J. Hand, N. Schmill, K. McGrath, C. Speller, Ancient mitochondrial DNA and population dynamics in postclassic Central Mexico: Tlatelolco (ad 1325–1520) and Cholula (ad 900–1350). *Archaeol. Anthropol. Sci.* **11**, 3459–3475 (2019). [doi:10.1007/s12520-018-00771-7](https://doi.org/10.1007/s12520-018-00771-7)
28. A. González-Oliver, L. Márquez-Morfin, J. C. Jiménez, A. Torre-Blanco, Founding Amerindian mitochondrial DNA lineages in ancient Maya from Xcaret, Quintana Roo. *Am. J. Phys. Anthropol.* **116**, 230–235 (2001). [doi:10.1002/ajpa.1118](https://doi.org/10.1002/ajpa.1118) [Medline](#)
29. A. Y. Morales-Arce, C. A. Hofman, A. T. Duggan, A. K. Benfer, M. A. Katzenberg, G. McCafferty, C. Warinner, Successful reconstruction of whole mitochondrial genomes

- from ancient Central America and Mexico. *Sci. Rep.* **7**, 18100 (2017). [doi:10.1038/s41598-017-18356-0](https://doi.org/10.1038/s41598-017-18356-0) [Medline](#)
30. B. M. Kemp, A. González-Oliver, R. S. Malhi, C. Monroe, K. B. Schroeder, J. McDonough, G. Rhett, A. Resendéz, R. I. Peñaloza-Espinosa, L. Buentello-Malo, C. Gorodesky, D. G. Smith, Evaluating the Farming/Language Dispersal Hypothesis with genetic variation exhibited by populations in the Southwest and Mesoamerica. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 6759–6764 (2010). [doi:10.1073/pnas.0905753107](https://doi.org/10.1073/pnas.0905753107) [Medline](#)
 31. M. Bodner, U. A. Perego, J. E. Gomez, R. M. Cerda-Flores, N. Rambaldi Migliore, S. R. Woodward, W. Parson, A. Achilli, The Mitochondrial DNA Landscape of Modern Mexico. *Genes* **12**, 1453 (2021). [doi:10.3390/genes12091453](https://doi.org/10.3390/genes12091453) [Medline](#)
 32. M. van Oven, M. Kayser, Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. *Hum. Mutat.* **30**, E386–E394 (2009). [doi:10.1002/humu.20921](https://doi.org/10.1002/humu.20921) [Medline](#)
 33. D. H. Alexander, J. Novembre, K. Lange, Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009). [doi:10.1101/gr.094052.109](https://doi.org/10.1101/gr.094052.109) [Medline](#)
 34. N. Patterson, A. L. Price, D. Reich, Population structure and eigenanalysis. *PLOS Genet.* **2**, e190 (2006). [doi:10.1371/journal.pgen.0020190](https://doi.org/10.1371/journal.pgen.0020190) [Medline](#)
 35. D. Reich, A. L. Price, N. Patterson, Principal component analysis of genetic data. *Nat. Genet.* **40**, 491–492 (2008). [doi:10.1038/ng0508-491](https://doi.org/10.1038/ng0508-491) [Medline](#)
 36. A. Moreno-Estrada, C. R. Gignoux, J. C. Fernández-López, F. Zakharia, M. Sikora, A. V. Contreras, V. Acuña-Alonzo, K. Sandoval, C. Eng, S. Romero-Hidalgo, P. Ortiz-Tello, V. Robles, E. E. Kenny, I. Nuño-Arana, R. Barquera-Lozano, G. Macín-Pérez, J. Granados-Arriola, S. Huntsman, J. M. Galanter, M. Via, J. G. Ford, R. Chapela, W. Rodriguez-Cintron, J. R. Rodríguez-Santana, I. Romieu, J. J. Sienra-Monge, B. del Rio Navarro, S. J. London, A. Ruiz-Linares, R. Garcia-Herrera, K. Estrada, A. Hidalgo-Miranda, G. Jimenez-Sanchez, A. Carnevale, X. Soberón, S. Canizales-Quinteros, H. Rangel-Villalobos, I. Silva-Zolezzi, E. G. Burchard, C. D. Bustamante, The genetics of Mexico recapitulates Native American substructure and affects biomedical traits. *Science* **344**, 1280–1285 (2014). [doi:10.1126/science.1251688](https://doi.org/10.1126/science.1251688) [Medline](#)
 37. A. Auton, L. D. Brooks, R. M. Durbin, E. P. Garrison, H. M. Kang, J. O. Korbel, J. L. Marchini, S. McCarthy, G. A. McVean, G. R. Abecasis; 1000 Genomes Project Consortium, A global reference for human genetic variation. *Nature* **526**, 68–74 (2015). [doi:10.1038/nature15393](https://doi.org/10.1038/nature15393) [Medline](#)
 38. O. François, F. Jay, Factor analysis of ancient population genomic samples. *Nat. Commun.* **11**, 4661 (2020). [doi:10.1038/s41467-020-18335-6](https://doi.org/10.1038/s41467-020-18335-6) [Medline](#)
 39. J. G. Schraiber, Assessing the Relationship of Ancient and Modern Populations. *Genetics* **208**, 383–398 (2018). [doi:10.1534/genetics.117.300448](https://doi.org/10.1534/genetics.117.300448) [Medline](#)
 40. P. Skoglund, H. Malmström, A. Omrak, M. Raghavan, C. Valdiosera, T. Günther, P. Hall, K. Tambets, J. Parik, K.-G. Sjögren, J. Apel, E. Willerslev, J. Storå, A. Götherström, M. Jakobsson, Genomic diversity and admixture differs for Stone-Age Scandinavian foragers and farmers. *Science* **344**, 747–750 (2014). [doi:10.1126/science.1253448](https://doi.org/10.1126/science.1253448) [Medline](#)

41. H. Ringbauer, J. Novembre, M. Steinrücken, Parental relatedness through time revealed by runs of homozygosity in ancient DNA. *Nat. Commun.* **12**, 5425 (2021).
[doi:10.1038/s41467-021-25289-w](https://doi.org/10.1038/s41467-021-25289-w) [Medline](#)
42. D. I. Cruz Dávalos, Y. O. Arizmendi Cárdenas, M. J. Bravo-Lopez, S. Neuenschwander, S. Reis, M. Q. R. Bastos, J. Stenderup, F. E. Yediay, V. Villa-Islas, C. S. Reyna-Blanco, C. Rodrigues-Carvalho, T. Hünemeier, M. E. Allentoft, C. E. G. Amorim, J. V. Moreno-Mayar, M. C. Ávila-Arcos, A.-S. Malaspinas, Indigenous peoples in eastern Brazil: Insights from 19th century genomes and metagenomes. *bioRxiv* 448717 [Preprint] (2022) [cited 2023]. [doi:10.1101/2022.01.27.477466](https://doi.org/10.1101/2022.01.27.477466)
43. M. C. Ávila-Arcos, K. F. McManus, K. Sandoval, J. E. Rodríguez-Rodríguez, V. Villa-Islas, A. R. Martin, P. Luisi, R. I. Peñaloza-Espinosa, C. Eng, S. Huntsman, E. G. Burchard, C. R. Gignoux, C. D. Bustamante, A. Moreno-Estrada, Population History and Gene Divergence in Native Mexicans Inferred from 76 Human Exomes. *Mol. Biol. Evol.* **37**, 994–1006 (2020). [doi:10.1093/molbev/msz282](https://doi.org/10.1093/molbev/msz282) [Medline](#)
44. N. Patterson, P. Moorjani, Y. Luo, S. Mallick, N. Rohland, Y. Zhan, T. Genschoreck, T. Webster, D. Reich, Ancient admixture in human history. *Genetics* **192**, 1065–1093 (2012). [doi:10.1534/genetics.112.145037](https://doi.org/10.1534/genetics.112.145037) [Medline](#)
45. J. V. Moreno-Mayar, B. A. Potter, L. Vinner, M. Steinrücken, S. Rasmussen, J. Terhorst, J. A. Kamm, A. Albrechtsen, A.-S. Malaspinas, M. Sikora, J. D. Reuther, J. D. Irish, R. S. Malhi, L. Orlando, Y. S. Song, R. Nielsen, D. J. Meltzer, E. Willerslev, Terminal Pleistocene Alaskan genome reveals first founding population of Native Americans. *Nature* **553**, 203–207 (2018). [doi:10.1038/nature25173](https://doi.org/10.1038/nature25173) [Medline](#)
46. M. Rasmussen, S. L. Anzick, M. R. Waters, P. Skoglund, M. DeGiorgio, T. W. Stafford Jr., S. Rasmussen, I. Moltke, A. Albrechtsen, S. M. Doyle, G. D. Poznik, V. Gudmundsdottir, R. Yadav, A.-S. Malaspinas, S. S. White 5th, M. E. Allentoft, O. E. Cornejo, K. Tambets, A. Eriksson, P. D. Heintzman, M. Karmin, T. S. Korneliusson, D. J. Meltzer, T. L. Pierre, J. Stenderup, L. Saag, V. M. Warmuth, M. C. Lopes, R. S. Malhi, S. Brunak, T. Sicheritz-Ponten, I. Barnes, M. Collins, L. Orlando, F. Balloux, A. Manica, R. Gupta, M. Metspalu, C. D. Bustamante, M. Jakobsson, R. Nielsen, E. Willerslev, The genome of a Late Pleistocene human from a Clovis burial site in western Montana. *Nature* **506**, 225–229 (2014). [doi:10.1038/nature13025](https://doi.org/10.1038/nature13025) [Medline](#)
47. J. V. Moreno-Mayar, L. Vinner, P. de Barros Damgaard, C. de la Fuente, J. Chan, J. P. Spence, M. E. Allentoft, T. Vimala, F. Racimo, T. Pinotti, S. Rasmussen, A. Margaryan, M. Iraeta Orbegozo, D. Mylopotamitaki, M. Wooller, C. Bataille, L. Becerra-Valdivia, D. Chivall, D. Comeskey, T. Devière, D. K. Grayson, L. George, H. Harry, V. Alexandersen, C. Primeau, J. Erlandson, C. Rodrigues-Carvalho, S. Reis, M. Q. R. Bastos, J. Cybulski, C. Vullo, F. Morello, M. Vilar, S. Wells, K. Gregersen, K. L. Hansen, N. Lynnerup, M. Mirazón Lahr, K. Kjær, A. Strauss, M. Alfonso-Durruty, A. Salas, H. Schroeder, T. Higham, R. S. Malhi, J. T. Rasic, L. Souza, F. R. Santos, A.-S. Malaspinas, M. Sikora, R. Nielsen, Y. S. Song, D. J. Meltzer, E. Willerslev, Early human dispersals within the Americas. *Science* **362**, eaav2621 (2018).
[doi:10.1126/science.aav2621](https://doi.org/10.1126/science.aav2621) [Medline](#)

48. H. García-Ortiz, F. Barajas-Olmos, C. Contreras-Cubas, M. Á. Cid-Soto, E. J. Córdova, F. Centeno-Cruz, E. Mendoza-Caamal, I. Cicerón-Arellano, M. Flores-Huacuja, P. Baca, D. A. Bolnick, M. Snow, S. E. Flores-Martínez, R. Ortiz-Lopez, A. W. Reynolds, A. Blanchet, M. Morales-Marín, R. Velázquez-Cruz, A. D. Kostic, C. Galaviz-Hernández, A. G. García-Zapién, J. C. Jiménez-López, G. León-Reyes, E. G. Salas-Bautista, B. P. Lazalde-Ramos, J. L. Jiménez-Ruíz, G. Salas-Martínez, J. Ramos-Madrigal, E. Mirzaeicheshmeh, Y. Saldaña-Alvarez, M. Del Carmen Abrahantes-Pérez, F. Loeza-Becerra, R. Mojica-Espinosa, F. Sánchez-Quinto, H. Rangel-Villalobos, M. Sosa-Macías, J. Sánchez-Corona, A. Rojas-Martinez, A. Martínez-Hernández, L. Orozco, The genomic landscape of Mexican Indigenous populations brings insights into the peopling of the Americas. *Nat. Commun.* **12**, 5942 (2021). [doi:10.1038/s41467-021-26188-w](https://doi.org/10.1038/s41467-021-26188-w) [Medline](#)
49. A. Gorostiza, V. Acunha-Alonzo, L. Regalado-Liu, S. Tirado, J. Granados, D. Sámano, H. Rangel-Villalobos, A. González-Martín, Reconstructing the history of Mesoamerican populations through the study of the mitochondrial DNA control region. *PLOS ONE* **7**, e44666 (2012). [doi:10.1371/journal.pone.0044666](https://doi.org/10.1371/journal.pone.0044666) [Medline](#)
50. S. Romero-Hidalgo, A. Ochoa-Leyva, A. Garcíarrubio, V. Acuña-Alonzo, E. Antúnez-Argüelles, M. Balcazar-Quintero, R. Barquera-Lozano, A. Carnevale, F. Cornejo-Granados, J. C. Fernández-López, R. García-Herrera, H. García-Ortiz, Á. Granados-Silvestre, J. Granados, F. Guerrero-Romero, E. Hernández-Lemus, P. León-Mimila, G. Macín-Pérez, A. Martínez-Hernández, M. Menjivar, E. Morett, L. Orozco, G. Ortíz-López, F. Pérez-Villatoro, J. Rivera-Morales, F. Riveros-McKay, M. Villalobos-Comparán, H. Villamil-Ramírez, T. Villarreal-Molina, S. Canizales-Quinteros, X. Soberón, Demographic history and biologically relevant genetic variation of Native Mexicans inferred from whole-genome sequencing. *Nat. Commun.* **8**, 1005 (2017). [doi:10.1038/s41467-017-01194-z](https://doi.org/10.1038/s41467-017-01194-z) [Medline](#)
51. J. Ramos-Madrigal, B. D. Smith, J. V. Moreno-Mayar, S. Gopalakrishnan, J. Ross-Ibarra, M. T. P. Gilbert, N. Wales, Genome Sequence of a 5,310-Year-Old Maize Cob Provides Insights into the Early Stages of Maize Domestication. *Curr. Biol.* **26**, 3195–3201 (2016). [doi:10.1016/j.cub.2016.09.036](https://doi.org/10.1016/j.cub.2016.09.036) [Medline](#)
52. M. A. Nieves-Colón, W. J. Pestle, A. W. Reynolds, B. Llamas, C. de la Fuente, K. Fowler, K. M. Skerry, E. Crespo-Torres, C. D. Bustamante, A. C. Stone, Ancient DNA Reconstructs the Genetic Legacies of Precontact Puerto Rico Communities. *Mol. Biol. Evol.* **37**, 611–626 (2020). [doi:10.1093/molbev/msz267](https://doi.org/10.1093/molbev/msz267) [Medline](#)
53. B. Llamas, L. Fehren-Schmitz, G. Valverde, J. Soubrier, S. Mallick, N. Rohland, S. Nordenfelt, C. Valdiosera, S. M. Richards, A. Rohrlach, M. I. B. Romero, I. F. Espinoza, E. T. Cagigao, L. W. Jiménez, K. Makowski, I. S. L. Reyna, J. M. Lory, J. A. B. Torrez, M. A. Rivera, R. L. Burger, M. C. Ceruti, J. Reinhard, R. S. Wells, G. Politis, C. M. Santoro, V. G. Standen, C. Smith, D. Reich, S. Y. W. Ho, A. Cooper, W. Haak, Ancient mitochondrial DNA provides high-resolution time scale of the peopling of the Americas. *Sci. Adv.* **2**, e1501385 (2016). [doi:10.1126/sciadv.1501385](https://doi.org/10.1126/sciadv.1501385) [Medline](#)
54. B. D. O’Fallon, L. Fehren-Schmitz, Native Americans experienced a strong population bottleneck coincident with European contact. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 20444–20448 (2011). [doi:10.1073/pnas.1112563108](https://doi.org/10.1073/pnas.1112563108) [Medline](#)

55. A. J. Herrera Muñoz, E. Mejía Campos, Un minero en la Sierra Gorda: caso de contaminación ocupacional multielemental de metales pesados a finales del periodo Clásico. *Arqueología (México, DF)*, 127–135 (2017).
56. E. Nalda, La arqueología de Guanajuato: Trabajos recientes. *Arqueol. Mex.* **16**, 36–43 (2008).
57. M. E. Smith, Mesoamerica's First World City: Teotihuacan in Comparative Context. *Teotihuacan: The World Beyond the City* (2020).
58. B. A. Álvarez-Sandoval, L. R. Manzanilla, M. González-Ruiz, A. Malgosa, R. Montiel, Genetic Evidence Supports the Multiethnic Character of Teopancazco, a Neighborhood Center of Teotihuacan, Mexico (AD 200-600). *PLOS ONE* **10**, e0132371 (2015).
[doi:10.1371/journal.pone.0132371](https://doi.org/10.1371/journal.pone.0132371) [Medline](#)
59. E. Campos, J. Campos, A. Muñoz, Mercurialism Determination in Fetuses Bone Remains from Toluquilla, Queretaro, Mexico. *J. Sci. Res. Rep.* **8**, 1–10 (2015).
[doi:10.9734/JSRR/2015/18043](https://doi.org/10.9734/JSRR/2015/18043)
60. E. M. Pérez Campos, A. J. Herrera Muñoz, Minas y mineros: presencia de metales en sedimentos y restos humanos al sur de la Sierra Gorda de Querétaro en México. *Chungara (Arica)* **45**, 161–176 (2013). [doi:10.4067/S0717-73562013000100008](https://doi.org/10.4067/S0717-73562013000100008)
61. D. A. C. Velázquez, Expansión Territorial Comercial en Mesoamérica y Mesoamérica Septentrional por medio del Intercambio, *Geografía Ensino & Pesquisa* **19**, 59–68 (2015).
62. N. Rohland, M. Hofreiter, Ancient DNA extraction from bones and teeth. *Nat. Protoc.* **2**, 1756–1762 (2007). [doi:10.1038/nprot.2007.247](https://doi.org/10.1038/nprot.2007.247) [Medline](#)
63. J. Dabney, M. Knapp, I. Glocke, M.-T. Gansauge, A. Weihmann, B. Nickel, C. Valdiosera, N. García, S. Pääbo, J.-L. Arsuaga, M. Meyer, Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 15758–15763 (2013). [doi:10.1073/pnas.1314445110](https://doi.org/10.1073/pnas.1314445110) [Medline](#)
64. M. Meyer, M. Kircher, Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harb. Protoc.* **2010**, prot5448 (2010).
[doi:10.1101/pdb.prot5448](https://doi.org/10.1101/pdb.prot5448) [Medline](#)
65. T. Daley, A. D. Smith, Predicting the molecular complexity of sequencing libraries. *Nat. Methods* **10**, 325–327 (2013). [doi:10.1038/nmeth.2375](https://doi.org/10.1038/nmeth.2375) [Medline](#)
66. M. Schubert, S. Lindgreen, L. Orlando, AdapterRemoval v2: Rapid adapter trimming, identification, and read merging. *BMC Res. Notes* **9**, 88 (2016). [doi:10.1186/s13104-016-1900-2](https://doi.org/10.1186/s13104-016-1900-2) [Medline](#)
67. H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin; 1000 Genome Project Data Processing Subgroup, The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
[doi:10.1093/bioinformatics/btp352](https://doi.org/10.1093/bioinformatics/btp352) [Medline](#)
68. A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, M. A. DePristo, The Genome Analysis

- Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010). [doi:10.1101/gr.107524.110](https://doi.org/10.1101/gr.107524.110) [Medline](#)
69. H. Jónsson, A. Ginolhac, M. Schubert, P. L. F. Johnson, L. Orlando, mapDamage2.0: Fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics* **29**, 1682–1684 (2013). [doi:10.1093/bioinformatics/btt193](https://doi.org/10.1093/bioinformatics/btt193) [Medline](#)
70. V. Link, A. Kousathanas, K. Veeramah, C. Sell, A. Scheu, D. Wegmann, ATLAS: Analysis Tools for Low-depth and Ancient Samples, bioRxiv 105346 [Preprint] (2017) [cited 2023]. [doi:10.1101/105346](https://doi.org/10.1101/105346)
71. G. Renaud, V. Slon, A. T. Duggan, J. Kelso, Schmutzi: Estimation of contamination and endogenous mitochondrial consensus calling for ancient DNA. *Genome Biol.* **16**, 224 (2015). [doi:10.1186/s13059-015-0776-0](https://doi.org/10.1186/s13059-015-0776-0) [Medline](#)
72. H. Weissensteiner, D. Pacher, A. Kloss-Brandstätter, L. Forer, G. Specht, H.-J. Bandelt, F. Kronenberg, A. Salas, S. Schönherr, HaploGrep 2: Mitochondrial haplogroup classification in the era of high-throughput sequencing. *Nucleic Acids Res.* **44**, W58-63 (2016). [doi:10.1093/nar/gkw233](https://doi.org/10.1093/nar/gkw233) [Medline](#)
73. T. A. Hall, BioEdit: A User-Friendly Biological Sequence Alignment Editor and Analysis Program for Windows 95/98/NT. *Nucleic Acids Symp. Ser.* **41**, 95–98 (1999).
74. J. Rozas, A. Ferrer-Mata, J. C. Sánchez-DelBarrio, S. Guirao-Rico, P. Librado, S. E. Ramos-Onsins, A. Sánchez-Gracia, DnaSP 6: DNA Sequence Polymorphism Analysis of Large Data Sets. *Mol. Biol. Evol.* **34**, 3299–3302 (2017). [doi:10.1093/molbev/msx248](https://doi.org/10.1093/molbev/msx248) [Medline](#)
75. L. Excoffier, H. E. Lischer, Arlequin suite ver 3.5: A new series of programs to perform population genetics analyses under Linux and Windows. *Mol. Ecol. Resour.* **10**, 564–567 (2010). [doi:10.1111/j.1755-0998.2010.02847.x](https://doi.org/10.1111/j.1755-0998.2010.02847.x) [Medline](#)
76. M. van Oven, PhyloTree Build 17: Growing the human mitochondrial DNA tree. *Forensic Sci. International. Genet. Suppl. Ser.* **5**, e392–e394 (2015). [doi:10.1016/j.fsigss.2015.09.155](https://doi.org/10.1016/j.fsigss.2015.09.155)
77. S. Kumar, C. Bellis, M. Zlojutro, P. E. Melton, J. Blangero, J. E. Curran, Large scale mitochondrial sequencing in Mexican Americans suggests a reappraisal of Native American origins. *BMC Evol. Biol.* **11**, 293 (2011). [doi:10.1186/1471-2148-11-293](https://doi.org/10.1186/1471-2148-11-293) [Medline](#)
78. U. A. Perego, A. Achilli, N. Angerhofer, M. Accetturo, M. Pala, A. Olivieri, B. Hooshiar Kashani, K. H. Ritchie, R. Scozzari, Q.-P. Kong, N. M. Myres, A. Salas, O. Semino, H.-J. Bandelt, S. R. Woodward, A. Torroni, Distinctive Paleo-Indian migration routes from Beringia marked by two rare mtDNA haplogroups. *Curr. Biol.* **19**, 1–8 (2009). [doi:10.1016/j.cub.2008.11.058](https://doi.org/10.1016/j.cub.2008.11.058) [Medline](#)
79. A. Achilli, U. A. Perego, H. Lancioni, A. Olivieri, F. Gandini, B. Hooshiar Kashani, V. Battaglia, V. Grugni, N. Angerhofer, M. P. Rogers, R. J. Herrera, S. R. Woodward, D. Labuda, D. G. Smith, J. S. Cybulski, O. Semino, R. S. Malhi, A. Torroni, Reconciling migration models to the Americas with the variation of North American native mitogenomes. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 14308–14313 (2013). [doi:10.1073/pnas.1306290110](https://doi.org/10.1073/pnas.1306290110) [Medline](#)

80. A. Achilli, A. Olivieri, M. Pala, B. Hooshiar Kashani, V. Carossa, U. A. Perego, F. Gandini, A. Santoro, V. Battaglia, V. Grugni, H. Lancioni, C. Sirolla, A. R. Bonfigli, A. Cormio, M. Boemi, I. Testa, O. Semino, A. Ceriello, L. Spazzafumo, M. N. Gadaleta, M. Marra, R. Testa, C. Franceschi, A. Torroni, Mitochondrial DNA backgrounds might modulate diabetes complications rather than T2DM as a whole. *PLOS ONE* **6**, e21029 (2011). [doi:10.1371/journal.pone.0021029](https://doi.org/10.1371/journal.pone.0021029) [Medline](#)
81. U. A. Perego, N. Angerhofer, M. Pala, A. Olivieri, H. Lancioni, B. Hooshiar Kashani, V. Carossa, J. E. Ekins, A. Gómez-Carballa, G. Huber, B. Zimmermann, D. Corach, N. Babudri, F. Panara, N. M. Myres, W. Parson, O. Semino, A. Salas, S. R. Woodward, A. Achilli, A. Torroni, The initial peopling of the Americas: A growing number of founding mitochondrial genomes from Beringia. *Genome Res.* **20**, 1174–1179 (2010). [doi:10.1101/gr.109231.110](https://doi.org/10.1101/gr.109231.110) [Medline](#)
82. D. M. Behar, M. van Oven, S. Rosset, M. Metspalu, E.-L. Loogväli, N. M. Silva, T. Kivisild, A. Torroni, R. Villems, A “Copernican” reassessment of the human mitochondrial DNA tree from its root. *Am. J. Hum. Genet.* **90**, 675–684 (2012). [doi:10.1016/j.ajhg.2012.03.002](https://doi.org/10.1016/j.ajhg.2012.03.002) [Medline](#)
83. A. Hartmann, M. Thieme, L. K. Nanduri, T. Stempfl, C. Moehle, T. Kivisild, P. J. Oefner, Validation of microarray-based resequencing of 93 worldwide mitochondrial genomes. *Hum. Mutat.* **30**, 115–122 (2009). [doi:10.1002/humu.20816](https://doi.org/10.1002/humu.20816) [Medline](#)
84. R. S. Just, T. M. Diegoli, J. L. Saunier, J. A. Irwin, T. J. Parsons, Complete mitochondrial genome sequences for 265 African American and U.S. “Hispanic” individuals. *Forensic Sci. Int. Genet.* **2**, e45–e48 (2008). [doi:10.1016/j.fsigen.2007.12.001](https://doi.org/10.1016/j.fsigen.2007.12.001) [Medline](#)
85. M. Ingman, H. Kaessmann, S. Pääbo, U. Gyllensten, Mitochondrial genome variation and the origin of modern humans. *Nature* **408**, 708–713 (2000). [doi:10.1038/35047064](https://doi.org/10.1038/35047064) [Medline](#)
86. S. Kumar, G. Stecher, M. Li, C. Knyaz, K. Tamura, MEGA X: Molecular evolutionary genetics analysis across computing platforms. *Mol. Biol. Evol.* **35**, 1547–1549 (2018). [doi:10.1093/molbev/msy096](https://doi.org/10.1093/molbev/msy096) [Medline](#)
87. J. W. Leigh, D. Bryant, POPART : Full-feature software for haplotype network construction. *Methods Ecol. Evol.* **6**, 1110–1116 (2015). [doi:10.1111/2041-210X.12410](https://doi.org/10.1111/2041-210X.12410)
88. J. Heled, A. J. Drummond, Bayesian inference of population size history from multiple loci. *BMC Evol. Biol.* **8**, 289 (2008). [doi:10.1186/1471-2148-8-289](https://doi.org/10.1186/1471-2148-8-289) [Medline](#)
89. R. Bouckaert, J. Heled, D. Kühnert, T. Vaughan, C.-H. Wu, D. Xie, M. A. Suchard, A. Rambaut, A. J. Drummond, BEAST 2: A software platform for Bayesian evolutionary analysis. *PLOS Comput. Biol.* **10**, e1003537 (2014). [doi:10.1371/journal.pcbi.1003537](https://doi.org/10.1371/journal.pcbi.1003537) [Medline](#)
90. R. Lanfear, P. B. Frandsen, A. M. Wright, T. Senfeld, B. Calcott, PartitionFinder 2: New Methods for Selecting Partitioned Models of Evolution for Molecular and Morphological Phylogenetic Analyses. *Mol. Biol. Evol.* **34**, 772–773 (2017). [doi:10.1093/molbev/msw260](https://doi.org/10.1093/molbev/msw260) [Medline](#)
91. P. Soares, L. Ermini, N. Thomson, M. Mormina, T. Rito, A. Röhl, A. Salas, S. Oppenheimer, V. Macaulay, M. B. Richards, Correcting for purifying selection: An improved human

- mitochondrial molecular clock. *Am. J. Hum. Genet.* **84**, 740–759 (2009).
[doi:10.1016/j.ajhg.2009.05.001](https://doi.org/10.1016/j.ajhg.2009.05.001) [Medline](#)
92. P. Skoglund, J. Storå, A. Götherström, M. Jakobsson, Accurate sex identification of ancient human remains using DNA shotgun sequencing. *J. Archaeol. Sci.* **40**, 4477–4482 (2013).
[doi:10.1016/j.jas.2013.07.004](https://doi.org/10.1016/j.jas.2013.07.004)
93. Y. Huang, H. Ringbauer, hapCon: Estimating contamination of ancient genomes by copying from reference haplotypes. *Bioinformatics* **38**, 3768–3777 (2022).
[doi:10.1093/bioinformatics/btac390](https://doi.org/10.1093/bioinformatics/btac390) [Medline](#)
94. The 1000 Genomes Project Consortium, D. G. Poznik, et al., Punctuated bursts in human male demography inferred from 1,244 worldwide Y-chromosome sequences. *Nature Genetics* **48**, 593–599 (2016).
95. G. R. Abecasis, A. Auton, L. D. Brooks, M. A. DePristo, R. M. Durbin, R. E. Handsaker, H. M. Kang, G. T. Marth, G. A. McVean; 1000 Genomes Project Consortium, An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).
[doi:10.1038/nature11632](https://doi.org/10.1038/nature11632) [Medline](#)
96. S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. W. de Bakker, M. J. Daly, P. C. Sham, PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007). [doi:10.1086/519795](https://doi.org/10.1086/519795) [Medline](#)
97. B. K. Maples, S. Gravel, E. E. Kenny, C. D. Bustamante, RFMix: A discriminative modeling approach for rapid and robust local-ancestry inference. *Am. J. Hum. Genet.* **93**, 278–288 (2013). [doi:10.1016/j.ajhg.2013.06.020](https://doi.org/10.1016/j.ajhg.2013.06.020) [Medline](#)
98. M. Rasmussen, Y. Li, S. Lindgreen, J. S. Pedersen, A. Albrechtsen, I. Moltke, M. Metspalu, E. Metspalu, T. Kivisild, R. Gupta, M. Bertalan, K. Nielsen, M. T. P. Gilbert, Y. Wang, M. Raghavan, P. F. Campos, H. M. Kamp, A. S. Wilson, A. Gledhill, S. Tridico, M. Bunce, E. D. Lorenzen, J. Binladen, X. Guo, J. Zhao, X. Zhang, H. Zhang, Z. Li, M. Chen, L. Orlando, K. Kristiansen, M. Bak, N. Tommerup, C. Bendixen, T. L. Pierre, B. Grønnow, M. Meldgaard, C. Andreasen, S. A. Fedorova, L. P. Osipova, T. F. G. Higham, C. B. Ramsey, T. V. Hansen, F. C. Nielsen, M. H. Crawford, S. Brunak, T. Sicheritz-Pontén, R. Villems, R. Nielsen, A. Krogh, J. Wang, E. Willerslev, Ancient human genome sequence of an extinct Palaeo-Eskimo. *Nature* **463**, 757–762 (2010).
[doi:10.1038/nature08835](https://doi.org/10.1038/nature08835) [Medline](#)
99. P. Flegontov, N. E. Altınışık, P. Changmai, N. Rohland, S. Mallick, N. Adamski, D. A. Bolnick, N. Broomandkoshbacht, F. Candilio, B. J. Culleton, O. Flegontova, T. M. Friesen, C. Jeong, T. K. Harper, D. Keating, D. J. Kennett, A. M. Kim, T. C. Lamnidis, A. M. Lawson, I. Olalde, J. Oppenheimer, B. A. Potter, J. Raff, R. A. Sattler, P. Skoglund, K. Stewardson, E. J. Vajda, S. Vasilyev, E. Veselovskaya, M. G. Hayes, D. H. O’Rourke, J. Krause, R. Pinhasi, D. Reich, S. Schiffels, Palaeo-Eskimo genetic ancestry and the peopling of Chukotka and North America. *Nature* **570**, 236–240 (2019).
[doi:10.1038/s41586-019-1251-y](https://doi.org/10.1038/s41586-019-1251-y) [Medline](#)
100. C. Posth, N. Nakatsuka, I. Lazaridis, P. Skoglund, S. Mallick, T. C. Lamnidis, N. Rohland, K. Nägele, N. Adamski, E. Bertolini, N. Broomandkoshbacht, A. Cooper, B. J. Culleton, T. Ferraz, M. Ferry, A. Furtwängler, W. Haak, K. Harkins, T. K. Harper, T. Hünemeier,

- A. M. Lawson, B. Llamas, M. Michel, E. Nelson, J. Oppenheimer, N. Patterson, S. Schiffels, J. Sedig, K. Stewardson, S. Talamo, C.-C. Wang, J.-J. Hublin, M. Hubbe, K. Harvati, A. Nuevo Delaunay, J. Beier, M. Francken, P. Kaulicke, H. Reyes-Centeno, K. Rademaker, W. R. Trask, M. Robinson, S. M. Gutierrez, K. M. Prufer, D. C. Salazar-García, E. N. Chim, L. Müller Plumm Gomes, M. L. Alves, A. Liryo, M. Inglez, R. E. Oliveira, D. V. Bernardo, A. Barioni, V. Wesolowski, N. A. Scheifler, M. A. Rivera, C. R. Plens, P. G. Messineo, L. Figuti, D. Corach, C. Scabuzzo, S. Eggers, P. DeBlasis, M. Reindel, C. Méndez, G. Politis, E. Tomasto-Cagigao, D. J. Kennett, A. Strauss, L. Fehren-Schmitz, J. Krause, D. Reich, Reconstructing the Deep Population History of Central and South America. *Cell* **175**, 1185–1197.e22 (2018). [doi:10.1016/j.cell.2018.10.027](https://doi.org/10.1016/j.cell.2018.10.027) [Medline](#)
101. C. de la Fuente, M. C. Ávila-Arcos, J. Galimany, M. L. Carpenter, J. R. Homburger, A. Blanco, P. Contreras, D. Cruz Dávalos, O. Reyes, M. San Roman, A. Moreno-Estrada, P. F. Campos, C. Eng, S. Huntsman, E. G. Burchard, A.-S. Malaspinas, C. D. Bustamante, E. Willerslev, E. Llop, R. A. Verdugo, M. Moraga, Genomic insights into the origin and diversification of late maritime hunter-gatherers from the Chilean Patagonia. *Proc. Natl. Acad. Sci. U.S.A.* **115**, E4006–E4012 (2018). [doi:10.1073/pnas.1715688115](https://doi.org/10.1073/pnas.1715688115) [Medline](#)
102. J. M. Monroy Kuhn, M. Jakobsson, T. Günther, Estimating genetic kin relationships in prehistoric populations. *PLOS ONE* **13**, e0195491 (2018). [doi:10.1371/journal.pone.0195491](https://doi.org/10.1371/journal.pone.0195491) [Medline](#)
103. A. A. Behr, K. Z. Liu, G. Liu-Fang, P. Nakka, S. Ramachandran, pong: Fast analysis and visualization of latent clusters in population genetic data. *Bioinformatics* **32**, 2817–2823 (2016). [doi:10.1093/bioinformatics/btw327](https://doi.org/10.1093/bioinformatics/btw327) [Medline](#)
104. M. Nei, *Molecular Evolutionary Genetics* (Columbia University Press, New York, 1987).
105. K. Nägele, C. Posth, M. Iraeta Orbegozo, Y. Chinique de Armas, S. T. Hernández Godoy, U. M. González Herrera, M. A. Nieves-Colón, M. Sandoval-Velasco, D. Mylopotamitaki, R. Radzeviciute, J. Laffoon, W. J. Pestle, J. Ramos-Madrigal, T. C. Lamnidis, W. C. Schaffer, R. S. Carr, J. S. Day, C. Arredondo Antúnez, A. Rangel Rivero, A. J. Martínez-Fuentes, E. Crespo-Torres, I. Roksandic, A. C. Stone, C. Lalueza-Fox, M. Hoogland, M. Roksandic, C. L. Hofman, J. Krause, H. Schroeder, Genomic insights into the early peopling of the Caribbean. *Science* **369**, 456–460 (2020). [doi:10.1126/science.aba8697](https://doi.org/10.1126/science.aba8697) [Medline](#)
106. J. Bracho, Narrativa e identidad. El mestizaje y su representación historiográfica. *LREL*, 55 (2009).
107. P. Wade, in *The Companion to Latin American Studies* (Routledge, 2014), pp. 197–211.
108. E. Mejía Pérez Campos, X. Chávez Balderas, R. Chávez Sánchez, Pepita. La momia de la Sierra Gorda de Querétaro. *Arqueol. Mex.* **179**, 70–75 (2009).
109. E. Mejía Pérez Campos, thesis, Escuela Nacional de Antropología e Historia, Ciudad de México, Méx (2010).
110. G. Zepeda García Moreno, Cañada de la Virgen, refugio de los muertos y los ancestros. *Guanajuato, Ediciones La Rana* (2010).

111. I. D. Lara Barajas, G. Islas Estrada, G. Zepeda García Moreno, “Informe Final del Análisis Antropofísico del Entierro 19,” *Proyecto Arqueológico Cañada de la Virgen* (Centro INAH-Querétaro, 2011).
112. A. Goguitchaichvili, V. Ortega, J. Archer, J. Morales, A. T. Guerrero, Absolute geomagnetic intensity record from pre-Columbian pottery dates elite Tlailotlacan Woman in ancient Teotihuacan. *J. Archaeol. Sci. Rep.* **14**, 146–151 (2017).
[doi:10.1016/j.jasrep.2017.05.030](https://doi.org/10.1016/j.jasrep.2017.05.030)
113. A. Brito-Mayor, A dos y a cuatro patas: El Occidente de Mesoamérica y la relación entre el perro, el humano y viceversa. Una aproximación a través de estudios de caso. *Archeofauna* **28**, 169 (2019). [doi:10.15366/archaeofauna2019.28.014](https://doi.org/10.15366/archaeofauna2019.28.014)
114. T. Maricic, M. Whitten, S. Pääbo, Multiplexed DNA sequence capture of mitochondrial genomes using PCR products. *PLOS ONE* **5**, e14004 (2010).
[doi:10.1371/journal.pone.0014004](https://doi.org/10.1371/journal.pone.0014004) [Medline](#)
115. M. L. Carpenter, J. D. Buenrostro, C. Valdiosera, H. Schroeder, M. E. Allentoft, M. Sikora, M. Rasmussen, S. Gravel, S. Guillén, G. Nekhrizov, K. Leshtakov, D. Dimitrova, N. Theodossiev, D. Pettener, D. Luiselli, K. Sandoval, A. Moreno-Estrada, Y. Li, J. Wang, M. T. P. Gilbert, E. Willerslev, W. J. Greenleaf, C. D. Bustamante, Pulling out the 1%: Whole-genome capture for the targeted enrichment of ancient DNA sequencing libraries. *Am. J. Hum. Genet.* **93**, 852–864 (2013). [doi:10.1016/j.ajhg.2013.10.002](https://doi.org/10.1016/j.ajhg.2013.10.002) [Medline](#)
116. A. García, R. Nores, J. M. B. Motti, M. Pauro, P. Luisi, C. M. Bravi, M. Fabra, A. L. Gosling, O. Kardailsky, J. Boocock, N. Solé-Morata, E. A. Matisoo-Smith, D. Comas, D. A. Demarchi, Ancient and modern mitogenomes from Central Argentina: New insights into population continuity, temporal depth and migration in South America. *Hum. Mol. Genet.* **30**, 1200–1217 (2021). [doi:10.1093/hmg/ddab105](https://doi.org/10.1093/hmg/ddab105) [Medline](#)
117. A. González-Martín, A. Gorostiza, L. Regalado-Liu, S. Arroyo-Peña, S. Tirado, I. Nuño-Arana, R. Rubi-Castellanos, K. Sandoval, M. D. Coble, H. Rangel-Villalobos, Demographic History of Indigenous Populations in Mesoamerica Based on mtDNA Sequence Data. *PLOS ONE* **10**, e0131791 (2015). [doi:10.1371/journal.pone.0131791](https://doi.org/10.1371/journal.pone.0131791) [Medline](#)
118. B. Z. González-Sobrinho, A. P. Pintado-Cortina, L. Sebastián-Medina, F. Morales-Mandujano, A. V. Contreras, Y. E. Aguilar, J. Chávez-Benavides, A. Carrillo-Rodríguez, I. Silva-Zolezzi, L. Medrano-González, Genetic Diversity and Differentiation in Urban and Indigenous Populations of Mexico: Patterns of Mitochondrial DNA and Y-Chromosome Lineages. *Biodemogr. Soc. Biol.* **62**, 53–72 (2016).
[doi:10.1080/19485565.2015.1117938](https://doi.org/10.1080/19485565.2015.1117938) [Medline](#)
119. R. Acuna-Soto, D. W. Stahle, M. K. Cleaveland, M. D. Therrell, Megadrought and megadeath in 16th century Mexico. *Emerg. Infect. Dis.* **8**, 360–362 (2002).
[doi:10.3201/eid0804.010175](https://doi.org/10.3201/eid0804.010175) [Medline](#)
120. G. Kubler, *The Population of Central Mexico in the Sixteenth Century*. (JSTOR, 1948).
121. P. Skoglund, B. H. Northoff, M. V. Shunkov, A. P. Derevianko, S. Pääbo, J. Krause, M. Jakobsson, Separating endogenous ancient DNA from modern day contamination in a

Siberian Neandertal. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 2229–2234 (2014).
[doi:10.1073/pnas.1318934111](https://doi.org/10.1073/pnas.1318934111) [Medline](#)

122. R. Fregel, F. L. Méndez, Y. Bokbot, D. Martín-Socas, M. D. Camalich-Massieu, J. Santana, J. Morales, M. C. Ávila-Arcos, P. A. Underhill, B. Shapiro, G. Wojcik, M. Rasmussen, A. E. R. Soares, J. Kapp, A. Sockell, F. J. Rodríguez-Santos, A. Mikdad, A. Trujillo-Mederos, C. D. Bustamante, Ancient genomes from North Africa evidence prehistoric migrations to the Maghreb from both the Levant and Europe. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 6774–6779 (2018). [doi:10.1073/pnas.1800851115](https://doi.org/10.1073/pnas.1800851115) [Medline](#)
123. S. J. Macfarlan, C. N. Henrickson, Inferring Relationships Between Indigenous Baja California Sur and Seri/Comcáac Populations Through Cultural Traits. *J. Calif. Gt. Basin Anthropol.* **30**, 18 (2021).
124. E. M. P. Campos, *Interpretación preliminar respecto a la temporalidad de Toluquilla, Querétaro*. Annick Daneels, Ed. (V Coloquio Bosch Gimpera, Instituto de Investigaciones Antropológicas, Universidad Nacional Autónoma de México, México, 2009)
125. H. Rangel-Villalobos, V. M. Martínez-Sevilla, G. Martínez-Cortés, J. A. Aguilar-Velázquez, M. Sosa-Macías, R. Rubi-Castellanos, A. González-Martín, Importance of the geographic barriers to promote gene drift and avoid pre- and post-Columbian gene flow in Mexican native groups: Evidence from forensic STR Loci. *Am. J. Phys. Anthropol.* **160**, 298–316 (2016). [doi:10.1002/ajpa.22969](https://doi.org/10.1002/ajpa.22969) [Medline](#)

Anexo II. Contribución al artículo ‘Population History and Gene Divergence in Native Mexicans Inferred from 76 Human Exomes’

El artículo ‘Population History and Gene Divergence in Native Mexicans Inferred from 76 Human Exomes’ fue publicado en la revista *Molecular Biology and Evolution* en Diciembre del 2019. Mi aporte a este artículo fue una serie de análisis de D-statistics que se realizó con la finalidad de evaluar la relación de las ramas filogenética de Nativos Americanos del Norte (representada por individuos Athabascanos (55)) y de Nativos Americanos del Sur (representada por el individuo Anzick-1 (186)) respecto a cinco poblaciones indígenas de México (Rarámuri, Wixarika, Nahuatl, Triqui y Maya). En resumen, estos análisis mostraron que estas poblaciones indígenas de México están mayormente relacionadas con la rama filogenética de Nativos Americanos del Sur. Además de esto, realicé la figura 1 del artículo la cual muestra las ubicaciones geográficas de las poblaciones indígenas analizadas así como una representación de la historia demográfica de éstas. La figura muestra los tiempos de divergencia y los tamaños poblacionales estimados para cada una.

La lista de autores del artículo se muestra a continuación y mi nombre se muestra subrayado:

María C Ávila-Arcos, Kimberly F McManus, Karla Sandoval, Juan Esteban Rodríguez-Rodríguez, Viridiana Villa-Islas, Alicia R Martin, Pierre Luisi, Rosenda I Peñaloza-Espinosa, Celeste Eng, Scott Huntsman, Esteban G Burchard, Christopher R Gignoux, Carlos D Bustamante, Andrés Moreno-Estrada

El artículo puede encontrarse en: <https://doi.org/10.1093/molbev/msz282>, marzo 2023.

Anexo III. Contribución al artículo ‘Paleogenomic insights into the red complex bacteria *Tannerella forsythia* in Pre-Hispanic and Colonial individuals from Mexico’

El artículo ‘Paleogenomic insights into the red complex bacteria *Tannerella forsythia* in Pre-Hispanic and Colonial individuals from Mexico’ fue publicado en la revista *Philosophical Transactions of the Royal Society B* en octubre del 2020. Mi aporte a este artículo fue en la parte experimental. Realice tanto sola como junto con la primera autora Miriam Bravo-López las extracciones de ADN, la construcción de las bibliotecas genómicas, indexado de bibliotecas genómicas y preparación de éstas para su envío a secuenciación.

La lista de autores del artículo se muestra a continuación y mi nombre se muestra subrayado:

Miriam Bravo-Lopez , Viridiana Villa-Islas , Carolina Rocha Arriaga , Ana B. Villaseñor-Altamirano , Axel Guzmán-Solís , Marcela Sandoval-Velasco , Julie K. Wesp , Keitlyn Alcantara , Aurelio López-Corral , Jorge Gómez-Valdés , Elizabeth Mejía , Alberto Herrera , Alejandro Meraz-Moreno , Maria de la Luz Moreno-Cabrera , Andrés Moreno-Estrada , Maria A. Nieves-Colón , Joel Olvera , Julia Pérez-Pérez , Katrine Højholt Iversen , Simon Rasmussen , Karla Sandoval , Gabriela Zepeda and María C. Ávila-Arcos

El artículo puede encontrarse en: <https://doi.org/10.1098/rstb.2019.0580>, marzo 2023.

Anexo IV. Contribución al artículo ‘Ancient viral genomes reveal introduction of human pathogenic viruses into Mexico during the transatlantic slave trade’

El artículo ‘Ancient viral genomes reveal introduction of human pathogenic viruses into Mexico during the transatlantic slave trade’ fue publicado en la revista *eLife* en Agosto del 2021. Mi aporte a este artículo incluyó una parte experimental y una de análisis de datos. Durante la parte experimental realicé junto con el primer autor Axel A Guzmán-Solís las extracciones de ADN, la construcción de las bibliotecas genómicas, indexado de bibliotecas genómicas y preparación de éstas para su envío a secuenciación. En la parte de análisis de datos llevé a cabo el análisis de las secuencias asignadas al genoma humano. Los análisis realizados fueron PCA y ADMIXTURE para determinar las afinidades genéticas de los individuos ‘antiguos’ con poblaciones continentales de ascendencia africana, europea, asiática y nativo americana.

La lista de autores del artículo se muestra a continuación y mi nombre se muestra subrayado:

Axel A Guzmán-Solís, Viridiana Villa-Islas, Miriam J Bravo-López, Marcela Sandoval-Velasco , Julie K Wesp, Jorge A Gómez-Valdés, María de la Luz Moreno-Cabrera, Alejandro Meraz, Gabriela Solís-Pichardo, Peter Schaaf, Benjamin R TenOever, Daniel Blanco-Melo, María C Ávila Arcos

El artículo puede encontrarse en: <https://doi.org/10.7554/eLife.68612>, marzo 2023.

Anexo V. Contribución al artículo ‘Indigenous peoples in eastern Brazil: insights from 19th century genomes and metagenomes’

El artículo ‘Indigenous peoples in eastern Brazil: insights from 19th century genomes and metagenomes’ se encuentra como preprint en *bioRxiv* desde enero de 2022. Mi aporte a este artículo incluyó una parte experimental en la que se realizaron las capturas de genoma mitocondrial de los individuos Botocudos analizados en el artículo. Además, como parte del análisis de datos, se analizaron las secuencias del genoma mitocondrial y se observó que estos individuos presentaban un linaje materno característico de nativos americanos y no de alguna otra parte del mundo.

La lista de autores del artículo se muestra a continuación y mi nombre se muestra subrayado:

Diana Ivette Cruz Dávalos, Yami Ommar Arizmendi Cárdenas, Miriam Jetzabel Bravo-Lopez, Samuel Neuenschwander, Silvia Reis, Murilo Q. R. Bastos, Jesper Stenderup, Fulya Eylem Yediay, Viridiana Villa-Islas, Carlos S. Reyna-Blanco, Claudia Rodrigues-Carvalho, Tábita Hünemeier, Morten E. Allentoft, Carlos Eduardo G. Amorim, J. Víctor Moreno-Mayar, María C. Ávila-Arcos, Anna-Sapfo Malaspinas

El artículo puede encontrarse en: <https://doi.org/10.1101/2022.01.27.477466>, marzo 2023.

