UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

Maestría y Doctorado en Ciencias Bioquímicas

"Caracterización genómica de 100 individuos nativos mexicanos por secuenciación de genoma completo: Aspectos demográficos e identificación de variantes de interés biomédico"

"Whole-genome characterization in 100 Indigenous individuals from Mexico: demographic insights and variants of biomedical interest"

TESIS

QUE PARA OPTAR POR EL GRADO DE:

Doctor en Ciencias Bioquímicas

PRESENTA:

M. en C. Israel Aguilar Ordóñez

Dr. Enrique Morett Sánchez
Instituto de Biotecnología - UNAM

Dr. Adrián Ochoa Leyva
Instituto de Biotecnología - UNAM
Dr. Cei Abreu Goodger
The University of Edinburgh

Ciudad de México. Marzo, 2023

CGEP/PMDCB/1702/2022

ISRAEL AGUILAR ORDOÑEZ
Asunto: Jurado de examen
Estudiante de Doctorado en Ciencias Bioquímicas
Presente

Los miembros del Subcomité Académico en reunión ordinaria del 31 de agosto del presente año, conocieron su solicitud de asignación de **JURADO DE EXAMEN** para optar por el grado de **Doctorado en Ciencias**, con la réplica de la tesis **"Caracterización genómica de 100 individuos nativos mexicanos por secuenciación de genoma completo: Aspectos demográficos e identificación de variantes de interés biomédico",** dirigida por el Dr. Enrique Morett Sánchez.

De su análisis se acordó nombrar el siguiente jurado integrado por los doctores:

PRESIDENTE:	Dr.	Francisco Xavier Soberón Mainero
Vocal :	Dra.	Sofía Morán Ramos
Vocal :	Dra.	Sandra Romero Hidalgo
Vocal :	Dra.	María del Carmen Ávila Arcos
Secretario :	Dra.	Blanca Itzelt Taboada Ramírez

MIEMBROS DEL JURADO:
Es obligación de los tutores de este programa participar en éstas y otras actividades académicas encomendadas por nuestro Comité Académico. Sin embargo, en caso de que tenga un impedimento académico o de salud para cumplir con esta encomienda, es muy importante contar con su respuesta (Formato anexo) en un plazo no mayor a una semana. Tome en cuenta que usted tiene 30 días hábiles para emitir su voto con las rondas de revisión que considere necesarias..

Sin otro particular por el momento, aprovecho la ocasión para enviarle un cordial saludo.

A t e n t a m e n t e
"POR MI RAZA, HABLARÁ EL ESPÍRITU"
Cuernavaca, Morelos, a 31 de agosto de 2022
COORDINADOR DEL SUBCOMITÉ CAMPUS MORELOS

Dr. Enrique Rudiño Piñera

Contacto: mdcbq@posgrado.unam.mx Tel. 55-5623-7006

28/10/2022

Lic. Diana González Nieto
Directora de Certificación y Control Documental
Dirección General de Administración Escolar
UNAM
P r e s e n t e

Me es grato comunicarle que después de revisar cuidadosamente y de discutir mis sugerencias y correcciones al documento de tesis titulado "___***Caracterización genómica de 100 individuos nativos mexicanos por secuenciación de genoma completo: Aspectos demográficos e identificación de variantes de interés biomédico***___", que para obtener el grado de Doctor presenta el (la) alumno(a) __**Israel Aguilar Ordóñez**__ con número de cuenta **__511023002__**, inscrito(a) en el **Doctorado en Ciencias Bioquímicas**, considero que la tesis reúne los requisitos establecidos y por ello emito mi **VOTO APROBATORIO** para que realice la réplica oral.

Agradezco de antemano la atención que se sirva prestar a la presente.

Atentamente,

Dr. Francisco Xavier Soberon Mainero

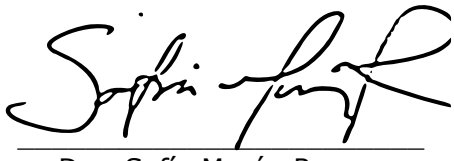
28/10/2022

Lic. Diana González Nieto
Directora de Certificación y Control Documental
Dirección General de Administración Escolar
UNAM
P r e s e n t e

Me es grato comunicarle que después de revisar cuidadosamente y de discutir mis sugerencias y correcciones al documento de tesis titulado "___***Caracterización genómica de 100 individuos nativos mexicanos por secuenciación de genoma completo: Aspectos demográficos e identificación de variantes de interés biomédico***___", que para obtener el grado de Doctor presenta el (la) alumno(a) __**Israel Aguilar Ordóñez**__ con número de cuenta **__511023002__**, inscrito(a) en el **Doctorado en Ciencias Bioquímicas**, considero que la tesis reúne los requisitos establecidos y por ello emito mi **VOTO APROBATORIO** para que realice la réplica oral.

Agradezco de antemano la atención que se sirva prestar a la presente.

Atentamente,



Dra. Sofía Morán Ramos

SALUD
SECRETARÍA DE SALUD

Instituto Nacional de Medicina Genómica
MÉXICO

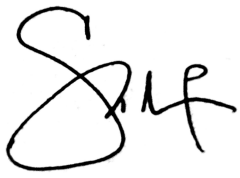
Ciudad de México a 8 de noviembre de 2022

Lic. Diana González Nieto
Directora de Certificación y Control Documental
Dirección General de Administración Escolar
UNAM
P r e s e n t e

Me es grato comunicarle que después de revisar cuidadosamente y de discutir mis sugerencias y correcciones al documento de tesis titulado "Caracterización genómica de 100 individuos nativos mexicanos por secuenciación de genoma completo: Aspectos demográficos e identificación de variantes de interés biomédico", que para obtener el grado de Doctor en Ciencias presenta el alumno Israel Aguilar Ordóñez con número de cuenta 511023002, inscrito en el Doctorado en Ciencias Bioquímicas. Considero que la tesis reúne los requisitos establecidos y por ello emito mi VOTO APROBATORIO para que realice la réplica oral.

Agradezco de antemano la atención que se sirva prestar a la presente.

Atentamente,



Dra. Sandra Romero Hidalgo
Investigadora en Ciencias Médicas "D"
Departamento de Genómica Computacional
sromero@inmegen.gob.mx

30/01/2023

Lic. Diana González Nieto
Directora de Certificación y Control Documental
Dirección General de Administración Escolar
UNAM
P r e s e n t e

Me es grato comunicarle que después de revisar cuidadosamente y de discutir mis sugerencias y correcciones al documento de tesis titulado "___***Caracterización genómica de 100 individuos nativos mexicanos por secuenciación de genoma completo: Aspectos demográficos e identificación de variantes de interés biomédico***___", que para obtener el grado de Doctor presenta el (la) alumno(a) __**Israel Aguilar Ordóñez**__ con número de cuenta __**511023002**__, inscrito(a) en el **Doctorado en Ciencias Bioquímicas**, considero que la tesis reúne los requisitos establecidos y por ello emito mi **VOTO APROBATORIO** para que realice la réplica oral.

Agradezco de antemano la atención que se sirva prestar a la presente.

Atentamente,

Mary Carmen Ávila Arcos

Dra. María del Carmen Ávila Arcos

UNAM
CAMPUS MORELOS

Instituto de Biotecnología
UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

07/11/2022

Lic. Diana González Nieto
Directora de Certificación y Control Documental
Dirección General de Administración Escolar
UNAM
P r e s e n t e

Me es grato comunicarle que después de revisar cuidadosamente y de discutir mis sugerencias y correcciones al documento de tesis titulado "***Caracterización genómica de 100 individuos nativos mexicanos por secuenciación de genoma completo: Aspectos demográficos e identificación de variantes de interés biomédico***", que para obtener el grado de Doctor presenta el (la) alumno(a) **Israel Aguilar Ordóñez** con número de cuenta **511023002**, inscrito(a) en el **Doctorado en Ciencias Bioquímicas**, considero que la tesis reúne los requisitos establecidos y por ello emito mi **VOTO APROBATORIO** para que realice la réplica oral.

Sin más por el momento quedo de usted muy atentamente,

"POR MI RAZA HABLARÁ EL ESPÍRITU"
CDMX a 04 de febrero del 2022

Blanca T.

Dra. Blanca Itzelt Taboada Ramírez
Investigador Titular A
Instituto de Biotecnología
Universidad Nacional Autónoma de México

“No me resulta molesto intentar trabajar con la imaginación, pues de esta manera puedo permanecer en casa”

- Vincent van Gogh, en alguna carta a Theo.

El presente trabajo fue realizado en el Departamento de Ingeniería Celular y Biocatálisis del Instituto de Biotecnología de la Universidad Nacional Autónoma de México y en el Instituto Nacional de Medicina Genómica, bajo la asesoría del Dr. Juan Enrique Morett Sánchez. Durante la realización de este trabajo se contó con el apoyo de la beca de doctorado otorgada por el Consejo Nacional de Ciencia y Tecnología (CONACYT ID Becario: 253835, No. apoyo: 480576).

Agradecimientos

A mis Abuelas. Por ser los pilares de la familia. Nada de lo que he visto o hecho se compara con todos sus esfuerzos para brindarme oportunidades. Todo lo que he logrado y puedo lograr es gracias a ustedes.

A Zai. Gracias por todo. Por acompañarme, por aguantarme, por amarme. Este es un paso más hacia lo que queremos construir. Y cada vez más cerca. LYA.

A mamá, papá y hermano. Porque sé que estarán ahí cuando más los necesito. Los amo.

A Enrique. Por ser mi mentor durante todos estos años. Gracias por la oportunidad de hacerme bioinformático y cambiar radicalmente mi vida profesional. Si no fuera por eso probablemente no hubiera continuado por el camino científico. También gracias por la confianza y la amistad que me permitieron crecer a mi ritmo.

A Cei Abreu y Adrían Ochoa. Por las horas de discusión en las reuniones tutorales, su experiencia, perspectiva y consejos.

A Toba. Por confiar en mí y por complementar mi educación con tu experiencia. Pero sobre todo por tu amistad.

A Judith, Ram, Fer, Josué, Edu y Pau. Por darme 6 razones para hacer buena ciencia. Sinceramente la parte más divertida y gratificante de este doctorado fue poder ser su mentor. Verlos crecer, tomar lo bueno de mis métodos, y después desarrollar sus propias personalidades científicas.

A las investigadoras e investigadores del proyecto 100G-MX. Gracias por compartirme su experiencia y por ser mis mentores en cuestiones de genómica poblacional y bioinformática.

A los sinodales revisores de esta tesis. Su experiencia y perspectiva me permitió mejorar.

A Ryosuke y el equipo Winter. Por confiar en mí, y apoyarme en el inicio de mi carrera en genómica. Mis días laborales más felices los pasé en Lindavista. Gracias por confiar y apoyar a más jóvenes como yo.

A la comunidad de facebook. Gracias por ser chidos, y por ser mis compas aunque muchos no nos conocemos en vivo. Cosa rara que me dejó la pandemia… entendí que hay formas de mostrar nuestro trabajo científico más allá de los auditorios de la UNAM. Gracias por interesarse en lo que hago, y por compartir un poco de lo que hacen ustedes también.

Y a ti que te tomarás el tiempo para leer esta tesis.

Gracias.
Totales.

Agradecimientos Técnicos

Este trabajo fue posible gracias a los individuos y comunidades de pueblos originarios que se involucraron en el proyecto, así como al laboratorio de la Dra. Lorena Orozco de INMEGEN por la recolección y resguardo de las muestras biológicas. Agradezco también al Departamento de supercómputo de INMEGEN, y a Jerome Jean Verleyen, administrador del cluster de cómputo científico en IBt-UNAM, por su apoyo para procesar la cantidad masiva de datos que demandó este proyecto durante varios años. Finalmente, agradezco a Guadalupe Ortiz Chipol por su revisión y perspectiva para la versión final de este documento.

Índice Tablas Y Figuras

Figura 1. Categorización de SNPs según la Frecuencia Alélica. 18

Figura 2. Línea del tiempo de proyectos de genómica poblacional con datos de genoma completo. 21

Figura 3. Historial de estudios genómicos en población indígena habitante del continente americano. 26

Figura 4. Los países hogar de poblaciones indígenas en América y la diversidad genómica faltante. 31

Figura 5. Filtrado del dataset 100G-MX. 32

Figura 6. Porcentaje de similitud genómica en los individuos incluidos y retirados del dataset. 33

Figura 7. Origen de las 27 poblaciones indígenas en el estudio. 33

Figura 8. Número de variantes y similitud genómica. 35

Figura 9. Resumen del efecto predicho para el catálogo de variantes detectadas en población indígena de México. 36

Figura 10. Análisis de similitud de genotipos por PCA intercontinental. 36

Figura 11. Análisis demográficos en población indígena de Mexico del proyecto 100G-MX. 39

Figura 11 bis. Sub-agrupamientos poblacionales basados en k-means. 39

Figura 12. Panorama de la variación en genomas indígenas de México en el proyecto 100G-MX. 41

Figura 13. SNVs de interés en las regiones Enhancer del gen PPARG. 42

Figura 14. Posición de las variantes en microRNAs maduros. 43

Figura 15. miRNAs variantes de la población indígena de México. 45

Figura 16. Los países involucrados en el estudio genómico de poblaciones indígenas en el continente americano. 48

Tabla 1: Resumen de las características de la secuenciación. 34

Tabla 2. Variantes miRNA en población nativa mexicana. 44

Abreviaturas

AF	*Allele Frequency*; frecuencia alélica. Proporción en la que se detecta un alelo específico en una población determinada. Se puede reportar como proporción en decimales (p. ej. 0.1), o porcentaje (p. ej. 10%).
dbSNP	*Single Nucleotide Polymorphism Database.* Es una base de datos pública que registra la variación genómica en diferentes especies. En la práctica, registra todas las variantes reportadas en la especie humana. Es mantenida por NCBI y NHGRI.
gnomAD v2.1	*Genome Aggregation Database.* Es el conjunto de datos genómicos públicos más grande a nivel mundial. Recopila 125,748 exomas y 15,708 genomas completos de 60 poblaciones diferentes. Sirve como la principal referencia para comparar frecuencias alélicas en proyectos poblacionales.
gnomAD v3.1	Similar a gnomAD v2.1, pero esta versión recopila 76,156 genomas completos de diversas poblaciones diferentes.
INDEL	*INsertion - DELetion.* Tipo de variante genómica donde se insertan o delatan nucleótidos en una determinada posición del genoma. Se considera indel cualquier cambio de entre 1 a 10,000 nucleótidos.
Microarreglo	Herramienta de laboratorio y plataforma para estudiar grandes cantidades de sondas de ADN o ARN predeterminadas. En estudios poblacionales los *arrays* o microarreglos contienen sondas para evaluar millones de variantes cortas (SNPs o INDELs). Los microarreglos incluyen posiciones predeterminadas que abarcan regiones específicas del genoma o del genoma completo.
miRNA	Un microRNA es un tipo de RNA pequeño no codificante que participa en la regulación de mRNAs blanco. La función de un microRNA es reducir la estabilidad de su mRNA blanco.
mRNA	ARN mensajero. Es el ARN de cadena sencilla que funciona como la base para la síntesis de proteínas. La estabilidad y cantidad de mRNA determina los niveles de una proteína.
SNV	*Single Nucleotide Variant.* Variante de nucleótido único donde sólo ocurre un cambio de base con respecto al genoma de referencia.
SNP	*Single Nucleotide Polymorphism.* Polimorfismo de nucleótido único. Es un SNV que aparece en una población determinada con una frecuencia alélica mayor al 1%.
WGS	*Whole Genome Sequencing,* o Secuenciación de genoma completo, por sus siglas en inglés. Es un método de laboratorio que permite digitalizar todo el ADN que constituye el genoma de un individuo.
WES	*Whole Exome Sequencing,* o Secuenciación de exoma completo, por sus siglas en inglés. Es un método de laboratorio que permite digitalizar solo el ADN que constituye los exones en los genes de un individuo.

Índice

Agradecimientos **2**

Agradecimientos Técnicos **3**

Índice de Tablas y Figuras **4**

Abreviaturas **5**

Prólogo **8**

Abstract **15**

Resumen **16**

Introducción **18**

El estudio de poblaciones indígenas en América 23

Diversidad genómica en poblaciones indígenas habitantes de América 24

Estudios sobre la diversidad genómica en México 27

Justificación **29**

Hipótesis **29**

Objetivos **29**

Resultados **30**

I. Compendio de estudios genómicos de poblaciones nativas de Ámerica 30

II. Secuenciación y catálogo de la variación genómica en poblaciones indígenas habitantes de México 31

III. Análisis demográfico de las poblaciones indígenas incluidas en el estudio 36

IV. Variantes de interés Biomédico 40

Discusión **48**

I. Sobre los estudios genómicos que involucran poblaciones indígenas de América 48

Conclusiones sobre el estudio genómico de las poblaciones indígenas de América 49

II. El aporte del proyecto 100G-MX al estudio de la genómica en America 52

Variantes de interés biomédico en el catálogo 100G-MX 52

Conclusiones 55

Perspectivas 55

Material y Métodos 57

Origen y manejo de las muestras de ADN 57

Preprocesamiento y mapeo de lecturas NGS 57

Llamado de Variantes 58

Anotación de Variantes 58

Catálogo de Variantes 59

Análisis de Componentes Principales, k-means, y ADMIXTURE 59

Análisis de F_{ST} 60

Detección de SNPs en el microRNAs 60

Anexo 1 - Artículo de Investigación 61

Anexo 2 - Artículo Review 62

Anexo 3 - Artículo de Divulgación 63

Anexo 4 - Tesis Co-Dirigidas 64

Anexo 5 - Difusión en medios de comunicación 65

Anexo 6 - Calidad de los Datos NGS 66

Referencias 67

Prólogo

En atención a las revisiones finales de mi sínodo —a quienes agradezco la perspectiva— escribo este prólogo para transmitir el contexto académico y colaborativo en el que se llevó a cabo el trabajo detrás de la tesis. Comenzando por decir que me considero un bioinformático; mi pasión científica son los datos y las computadoras. Bajo esa línea, mi camino en el doctorado estuvo enfocado en los datos genómicos humanos: ¿Cómo se procesan, organizan y analizan? Y a partir de esas sencillas preguntas surgieron varias líneas de trabajo enfocadas en la parte informática de lo bio. Sin embargo, a lo largo de estos años de doctorado, después de varias, varias revisiones —de manuscritos enviados, de seminarios, y sobre todo de este documento de tesis— he entendido por qué el programa de posgrado espera que un estudiante sea multidisciplinario. Así que aunque mi pasión científica sea la bioinformática, no puedo dejar de lado los otros aspectos importantes de un proyecto poblacional, puntualmente el aspecto antropológico.

Antes de continuar cabe aclarar que el proyecto 100G-MX (a partir del que surge esta tesis) involucró a varios grupos de investigación y diversos líderes de grupo. Mi trabajo es una parte del proyecto global, y está enfocado en la parte del análisis genómico de los datos previamente recolectados por otros colaboradores. De igual manera, las opiniones plasmadas en esta tesis no reflejan *de facto* las opiniones de los otros colaboradores del proyecto. Habiendo aclarado esto, me gustaría continuar con el prólogo.

Cuando me integré al grupo de trabajo, al inicio de mi doctorado, en INMEGEN recién estaban siendo recibidos los archivos base para el análisis (archivos fastq). Entonces planeamos este trabajo de doctorado con el objetivo de realizar análisis similares al proyecto One Thousand Genomes (detección, anotación y descripción de variantes, e inferencias sobre la historia demográfica) más algunos análisis que consideramos

relevantes y novedosos a nivel poblacional como el análisis de regiones reguladoras y microRNAs. Para lograr lo anterior había que resolver diversos problemas técnicos: almacenamiento de los datos, desarrollo de pipelines para procesamiento y análisis, incluso la configuración de hardware para poder llevar a cabo todas este trabajo en México (con la idea de desarrollar un ecosistema que permitiera autonomía tecnológica). En retrospectiva todo suele parecer sencillo, ahora analizar 100 genomas completos no me suena titánico (de hecho estamos analizando casi un ciento más de genomas públicos a un buen paso), pero por ahí de los inicios de mi trabajo de doctorado no contábamos con el conocimiento técnico ni la capacidad para llevarlo a cabo. Tomó tiempo, pero aprendimos lo necesario y lo logramos. Nuevamente debo remarcar el plural, porque no fui el único responsable de lograrlo, aunque sí participé extensivamente en cada parte del proceso.

Eventualmente se logró la publicación del análisis de los datos de genoma completo del proyecto 100G-MX. Y fue entonces que —al ser mi turno de presentarlo en seminarios y otros eventos académicos— comencé a notar mi falta de conocimiento en temas antropológicos, principalmente sobre la sensibilidad del lenguaje con que se presentan los resultados del proyecto. Se volvió claro para mí que como parte de mi formación doctoral es necesario hacer una revisión de los aspectos antropológicos alrededor de un proyecto poblacional enfocado en la población indígena que habita en el México actual. Gracias a las observaciones de mis sinodales me fue posible contar con bibliografía directamente pertinente. Dado que, reitero, mi trabajo principal es meramente bioinformático, considero que el cuerpo principal de esta tesis está completo. Así que decidí ocupar el resto de este prólogo para explorar el contexto antropológico que rodea mi trabajo, con la intención de reconocer su importancia y brindarle un espacio de reflexión propio.

El primer punto es el uso de términos para referirse a un grupo de personas que —asumimos— son representantes modernos del "pool" genético de pueblos ancestrales. Graham Coop *et*

al. (2022) menciona que "La información genómica de una población puede informarnos acerca de los ancestros genéticos que compartimos, a través de resumir las similitudes genéticas entre los individuos estudiados..."[1] y que "La estadística aplicada en la genética de poblaciones trabaja a partir de estas ideas para estudiar los procesos y la historia evolutiva. Sin embargo, muchas de las interpretaciones de estos patrones se derivan de combinar esta información con descriptores de muestra geográficos. Por lo tanto nuestra interpretación y las etiquetas aplicadas a las muestras siempre reflejan, por lo menos en parte, el contexto social de cómo se eligieron las muestras y como fueron descritas, y por lo tanto son constructos sociales parciales"[1]. Sobre esto, en el trabajo presentado en esta tesis se hicieron análisis de similitud genómica con el programa ADMIXTURE utilizando como referencia a las poblaciones etiquetadas como CEU, CHB y YRI del proyecto One Thousand Genomes, para determinar, en cada uno de los individuos, el porcentaje de los sitios en el genoma que NO se compartía con esos datos de referencia, pero sí se compartía entre los individuos del proyecto; se asumió entonces que ese porcentaje del genoma correspondía al heredado por antepasados indígenas. Dado que el objetivo técnico principal del proyecto fue describir la variación genómica de población indígena en México, se realizó un corte a partir de este porcentaje del genoma, con la intención de simplificar el análisis de variantes. Sin embargo, esto no se debe interpretar como la asignación de un gradiente de identidad indígena. Todos los participantes en el estudio del cual se derivan las muestras secuenciadas (la cohorte MAIS [2]) se auto-identificaron como pertenecientes y habitantes de una comunidad indígena, y dado el contexto social de los participantes, esa percepción siempre reflejará su realidad sin importar los resultados de un análisis de similitud. No se deben tomar los resultados de nuestro estudio genómico para determinar quién pertenece a la comunidad indígena en el México moderno. Eso jamás fue un objetivo del proyecto.

Graham Coop (2022) también propone abandonar las aseveraciones como "el individuo posee ancestría X" y reemplazarla por frases como "el individuo es similar a los

individuos X del estudio TAL", dado que esto deja más claro que se estudia la similitud de sitios en el genoma y no la identidad de los individuos estudiados. La recomendación se tomó en cuenta para la edición final de este trabajo de tesis.

Me parecen muy necesarios los cambios en el lenguaje utilizado para presentar los resultados de estudios poblacionales. Por ejemplo, en algunas revisiones (internas y externas al grupo de trabajo del proyecto) no había un consenso sobre cómo referirse a los individuos de estudio, siendo algunas de las opciones "indígenas", "nativos americanos" o "amerindios". Entiendo que algunos términos conllevan un contexto histórico en el cual no soy experto, y que el hecho de que hayan sido usados en mayor o menor medida en la bibliografía no son suficiente justificación para usarlos yo mismo. Sinceramente, antes no contaba con un marco de referencia para argumentar en favor de cualquiera de esos términos, pero ahora puedo utilizar la bibliografía citada en este prólogo —aportada por mi sínodo— para justificar por qué me refiero a las poblaciones de estudio como "poblaciones indígenas que habitan el territorio mexicano moderno". Esta terminología podría estar intercambiando brevedad por precisión, pero tal cual lo menciona Ewan Birney *et al.* (2022) "...la investigación en genética humana puede tener implicaciones extensas. Por lo tanto, el lenguaje que utilizamos para comunicar nuestros hallazgos a otros investigadores, y al público en general, es de suma importancia. [...] La antropología y la genética de poblaciones [...] actualmente rechazan firmemente ideas antiguas de la raza como una categoría biológicamente relevante y otras etiquetas que se basaron en perspectivas racistas del siglo XX en adelante"[3]. Bajo esa lógica, reitero mi compromiso y obligación de revisar y corregir el lenguaje que utilizo para la presentación de resultados en proyectos poblacionales. A veces mi predominante formación técnica hace que caiga en el error de utilizar términos comunes de los métodos y herramientas bioinformáticas que utilizo. Sin embargo, estoy en proceso contínuo de aprendizaje.

Respecto al involucramiento de las comunidades indígenas cabe mencionar que ideológicamente estoy de acuerdo en lo planteado por las tres estrategias para una investigación más ética y equitativa propuestas por Silva *et al.* (2022)[4], así como con los seis principios para la investigación ética involucrando a poblaciones indígenas propuestos por Claw *et al.* (2018)[5]. Pero debo admitir que en la práctica histórica —entendiendo este proyecto de tesis como una parte de un estudio más complejo que involucra a otros grupos de investigación, otros objetivos, etc.— , algunas de las actividades propuestas por los autores rebasaron mis posibilidades e influencia como estudiante de posgrado (por ejemplo la integración de comités locales con reuniones periódicas, o la elaboración y difusión local de materiales de difusión sobre los resultados de las investigaciones). Tal como lo propone Ávila-Arcos *et al.* (2022)[6], la investigación sustentable y respetuosa en el sur global —sin tintes de explotación académica o cultural— requiere de compromisos a largo plazo, y me atrevo a agregar explícitamente de apoyo presupuestal, logístico, e incluso del alineamiento de los objetivos institucionales en los centros de investigación que lideran estos proyectos. De manera personal yo no he estado en contacto con las poblaciones que formaron parte de este estudio de genoma completo, pero confío en que los colaboradores líderes del proyecto MAIS han estado involucrados con las comunidades indígenas desde el primer contacto durante la fase de reclutamiento.

Finalmente tengo que atender directamente una observación hecha por varios miembros del sínodo: ¿Cómo estás a favor del Open Data, si los datos de este proyecto no son públicos? Y mi respuesta nuevamente es que ideológicamente estoy convencido de que hacer disponibles los datos generados por cualquier proyecto contribuye al avance científico (entendido su objetivo final como la generación de conocimiento universal); sin embargo, en la práctica, y en este proyecto en particular, mi capacidades como estudiante de posgrado están completamente rebasadas para lograr la liberación sensible de los datos. Se buscaron alternativas para hacer disponibles los datos genotípicos de los individuos estudiados, de manera anonimizada, respetuosa

y segura; sin embargo no se logró alcanzar un acuerdo entre todas las partes responsables involucradas. A pesar de lo anterior, no lo considero un tema cerrado, y si mi carrera académica me lo permite, continuaré el diálogo con los líderes de proyectos para personalmente entender mejor y promover el Open Data pertinente. Aunque ahora también comprendo que el trabajo es aún más complejo de lo que pensaba, puesto que de acuerdo a la discusión planteada por Hudson *et al*. [7]—con la que concuerdo totalmente—, la publicación sensible de datos genómicos provenientes de poblaciones debe tener en cuenta a las comunidades en cada etapa del proceso. El punto más importante al respecto me parece la gobernanza de los datos por parte de los pueblos indígenas —entendida como la responsabilidad y capacidad de decidir sobre la generación, almacenamiento, la aplicación e incluso el borrado de la información genética y metadatos acompañantes—; desconozco si en México existen marcos reguladores de la soberanía genética, por ahora solo estoy familiarizado con la "LEY FEDERAL DE PROTECCIÓN DE DATOS PERSONALES EN POSESIÓN DE LOS PARTICULARES"[8] bajo la cual se protege la "información genética" —sin definición explícita de la misma— y se establece el mecanismo para regular la transparencia del uso y destino, así como el derecho de solicitar la eliminación de los registros. Sin embargo, esta ley no considera directamente aplicaciones en investigación científica, ni considera explícitamente la autogobernanza y el reconocimiento de usos y costumbres que pudieran existir en mayor o menor medida en poblaciones indígenas.

Espero que en el futuro próximo se logre la conjunción del conocimiento generado por este y otros proyectos con objetivos similares, si es que se logra el re-consentimiento y el involucramiento contínuo de los pueblos incluidos en el estudio. Atendiendo uno de los comentarios del sínodo que más me hizo reflexionar, el extractivismo académico genómico —entendido como la generación y explotación de datos genómicos en poblaciones vulneradas, por mero beneficio profesional— parece ser un riesgo latente en un país como México donde es imposible

garantizar el beneficio real y duradero de la participación en estudios poblaciones. Puedo decir que como bioinformático principal de los análisis presentados en esta tesis mi intención nunca fue la de explotar la información personal de los individuos incluidos en el estudio. Si por alguna razón mi trabajo o la forma en que lo presento llegara a dar esa impresión, estoy completamente receptivo a ser corregido. Si algo me enseñó el doctorado es que siempre se puede aprender para mejorar.

Israel Aguilar.

Abstract

The implementation of personalized medicine requires the development of strategies for individual diagnostics and specific treatment for each patient. To achieve this, first we must extensively study the genomic variation of the human species. In the last decade whole-genome sequencing has allowed to reveal genomic variation at a global scale. However, the study of genomic diversity in indigenous mexicans has been limited. Since 2005 there have been several projects for sampling and analyzing the genomes of native mexicans; one of those, the MAIS project (Metabolic Analysis in an Indigenous Sample) has integrated 2,596 samples from diverse mexican indigenous populations. In this work we analyzed the whole genomes of a subset of the MAIS samples, consisting of 95 individuals, from 32 indigenous populations across Mexico. Of those 5, and 13 individuals were excluded from further analyses due to close relatedness to other individuals in the study or by having less than 85% similarity to indigenous american samples, respectively. We found an average of 3.2 million single nucleotide variants and short indels. We reported a final catalog of 9,737,152 variants, of which 44,118 were not previously reported worldwide. We detected 316,577 variants located at regulatory elements of gene expression; interestingly, we found a probable haplotype spanning 3 enhancers regulating PPARG (a gene encoding a nuclear receptor / transcription factor previously implicated in health related phenotypes such as obesity, diabetes and insulin resistance). We also studied variants in microRNAs and found 4 distinctively seed-variant miRNAs: hsa-miR-4305, hsa-miR-627-5p, hsa-miR-6726-3p, and hsa-miR-6863. Our demographic analyses describe a North-Center-South axis for population stratification. The Comcaa'c population (also reported as Seri), a northern indigenous group, showed the highest genomic divergence in our study, in concordance with previous reports. The dataset published in this work describes novel variation in the world, contributing to close the genomic gap in underrepresented indigenous populations.

As a result of my Doctorate education I have co-authored an Original Research article (Annex 1), a Review (Annex 2), a science communication piece (Annex 3), and five BSc co-directed thesis (Annex 4), alongside many participations as professor and speaker in Bioinformatics (Annex 5).

Resumen

La medicina genómica está basada en estrategias para el diagnóstico y tratamiento personalizado de cada paciente a partir de la información genómica individual. Para lograrlo se requiere del estudio exhaustivo de la variación genómica poblacional. En la última década la secuenciación de genomas completos ha permitido el estudio de esta variación a nivel mundial [9–11]. Sin embargo, el estudio de la diversidad genómica en poblaciones indígenas ha sido limitado. Desde 2005 en INMEGEN se han llevado a cabo proyectos de muestreo y análisis genómico de individuos con ascendencia indígena mexicana [12]; uno de estos, el proyecto MAIS (Metabolic Analysis in an Indigenous Sample) ha integrado 2,596 muestras representativas de diversas poblaciones indígenas mexicanas [2]. En el presente trabajo realizamos el análisis de genoma completo de un subconjunto de muestras del proyecto MAIS, que consistió en 95 individuos pertenecientes a 32 poblaciones indígenas; posteriormente 5 individuos se excluyeron del análisis debido a que estaban emparentados con algún otro individuo, además 13 se excluyeron debido a que poseían menos del 85% de similitud con individuos etiquetados como indígenas habitantes del continente americano. En promedio cada individuo presentó 3.2 millones de variantes de nucleótido único e indeles cortos. En conjunto reportamos un catálogo de 9,737,152 variantes, de las cuales 44,118 no se habían reportado previamente a nivel mundial. Encontramos 316,577 variantes en posibles regiones reguladoras de la expresión génica; de manera interesante identificamos un posible haplotipo regulador del gen PPARG, que codifica un receptor nuclear previamente relacionado a

problemas de salud en el país tales como obesidad, diabetes y resistencia a insulina. También estudiamos las variantes en microRNAs y encontramos 4 miRNAs que varían distintivamente en su región seed: sa-miR-4305, hsa-miR-627-5p, hsa-miR-6726-3p, and hsa-miR-6863. Los análisis demográficos sugieren una estratificación poblacional en un eje Norte-Centro-Sur, con subestructura en la región central. La población Comcaa'c (también reportada como Seri), una población del norte del país, mostró la mayor divergencia genómica en nuestro estudio, en concordancia con reportes previos. Con el conjunto de datos publicados a partir de este trabajo describimos nueva variación a nivel poblacional, contribuyendo a reducir la brecha genómica en la representación de grupos indígenas.

Como resultado de mi formación de Doctorado se publicó un artículo de investigación original (Anexo 1) [13], un artículo de revisión (Anexo 2) [14], un artículo de divulgación (Anexo 3), y cuatro tesis de licenciatura codirigidas (Anexo 4), así como diversas participaciones como profesor y conferencista en temas de Bioinformática (Anexo 5).

Introducción

Comprender la relación entre el genotipo y el fenotipo de las poblaciones humanas es un gran reto para las ciencias biológicas, e imprescindible para la aplicación de la medicina genómica. Aunque el genoma humano de referencia proporciona un contexto base para el entendimiento de la estructura de los genes y otros elementos no codificantes, la referencia no contiene información sobre la diversidad de las diferentes poblaciones a nivel mundial [15,16]. Básicamente, el genoma de referencia humano es apenas "uno" de miles de millones. Gracias a los esfuerzos masivos por catalogar la diversidad genética humana, actualmente se conocen cerca de 113 millones de variantes (según la base de datos dbSNP b151), en su mayoría polimorfismos de nucleótido único (SNP, del inglés *Single Nucleotide Polymorphism*) respecto al genoma de referencia [17].

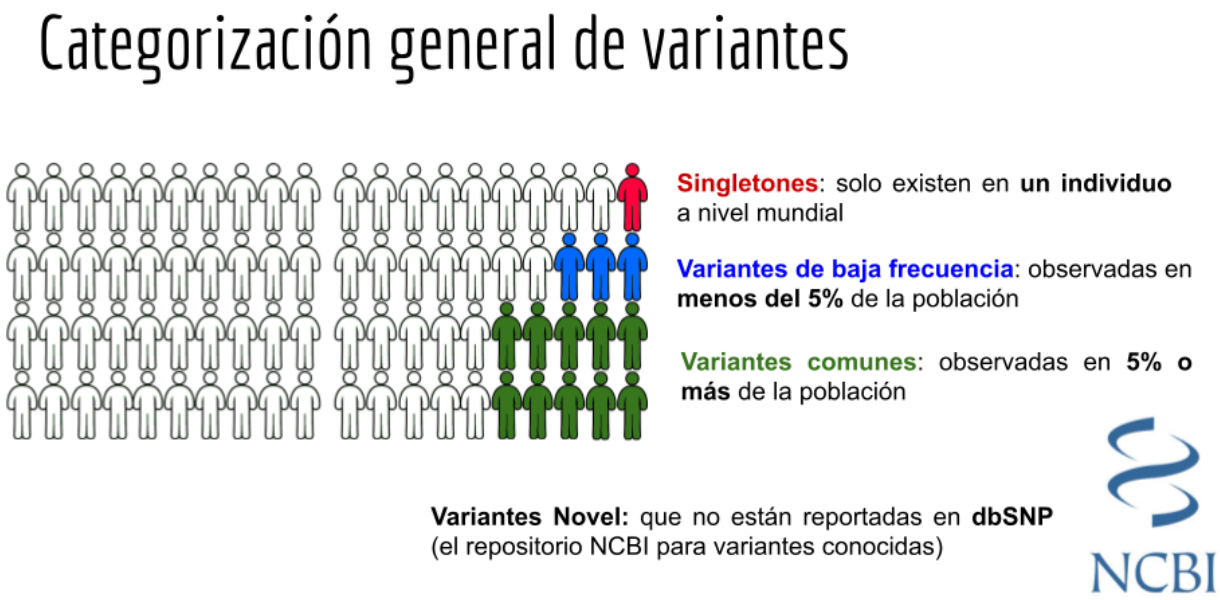


Figura 1. Categorización de SNPs según la Frecuencia Alélica.
Los SNPs encontrados a nivel poblacional se categorizar de acuerdo a la frecuencia alélica en la que se encuentran. Las variantes comunes se encuentran presentes en al menos el 5% de una población de estudio. Por otro lado, las variantes de baja frecuencia pueden representar procesos evolutivos específicos de la región geográfica. Los singletones suelen ser relevantes únicamente si se está estudiando la historia particular de un individuo o sus familiares cercanos.

Con la publicación del borrador original con la secuencia del genoma humano en 2001 (véase https://www.genome.gov/human-genome-project) inició una carrera científico-tecnológica en busca del conocimiento total de la genómica humana. A partir de ese momento comenzaron a plantearse proyectos poblacionales que —a largo plazo— permitirán conocer la secuencia genómica de la mayor parte de la población. A 20 años de ese primer genoma humano (con un costo estimado de 1,000 millones de dólares), actualmente las pruebas genéticas directas al consumidor hacen posible incluso la secuenciación a nivel personal por un costo menor al de un smartphone de moda (vease https://nebula.org/whole-genome-sequencing-dna-test/).

Algunos de los proyectos pioneros en genómica poblacional humana fueron colaboraciones internacionales, cuyos objetivos eran capturar la mayor diversidad genómica a través de la inclusión de individuos provenientes de distintas regiones geográficas. Cuatro de ellos merecen mención por ser trabajos seminales para el campo de la genotipificación poblacional: HapMap, 1000 Genomes Project, The Simons Genome Diversity Project, y el Human Genome Diversity Project. A continuación, se resumen estos proyectos:

HapMap. Fue un proyecto para describir el mapa de haplotipos a nivel mundial a través de paneles de arrays (microarreglos). Su primera versión (o fase 1) se publicó en 2005 con investigadores de USA, China, Canadá, Japón, Inglaterra, Nigeria, y Tailandia [18]. Un haplotipo es una serie de variantes SNV que se heredan en conjunto (como un bloque de ADN durante la recombinación de los gametos). El HapMap es una referencia que permite ubicar estos bloques de ADN heredables, así como los SNV mínimos para definir haplotipos conocidos. Su publicación final (fase 3) se lanzó en 2010 y por muchos años funcionó como una referencia para el estudio de variantes mediante tecnología de microarreglos [10]. En su momento fue el proyecto genómico publicado con más información poblacional, incluyendo 1.6 millones de SNVs, descritos en 1,184 individuos provenientes de

11 poblaciones: etiquetados como CEU (residentes de USA con alta similitud genómica a población Europea), CHB (personas Han residentes de China), JPT (personas Japonesas residentes de Tokio), YRI (personas Yoruba residentes de Nigeria), ASW (personas residentes del sur de USA con alta similitud genómica con población Africana), CHD (personas residentes de USA con alta similitud genómica con población de China), GIH (personas residentes de USA con alta similitud con población Indo-asiática), LWK (personas Luhya residentes de Kenya), MKK (personas Maasai residentes de Kenya), MXL (personas residentes de USA con alta similitud genómica con población mexicana), TSI (personas residentes de la toscana Italiana) [10]. Vale la pena señalar que en este estudio no se incluyeron explícitamente poblaciones indígenas.

1000 Genomes Project. Fue una iniciativa para describir el genoma de 1,000 individuos anónimos provenientes de diversas regiones del mundo. El proyecto utilizó técnicas de secuenciación de genoma completo. En diciembre de 2008 reportó los análisis de los primeros 4 individuos a través de un sitio web público para el proyecto, marcando una tendencia importante hacia el Open Data [19]. Hasta su publicación final en 2015, el proyecto incluyó 2,504 participantes de más de 26 poblaciones provenientes de distintos continentes [11], reportando la mayor diversidad genómica de la especie humana hasta el momento. En este proyecto tampoco se incluyeron poblaciones indígenas habitantes del continente americano. Cabe mencionar que a nivel bioinformático, 1000 Genomes Project también marcó el comienzo del desarrollo de herramientas computacionales para el análisis masivo de datos genómicos [20]; mucho del software que se ha vuelto estándar en las ciencias genómicas existe gracias a los retos planteados por este proyecto seminal. En este proyecto masivo participaron investigadores de México, Colombia y Perú.

The Simons Genome Diversity Project. Fue un estudio enfocado en estudiar más poblaciones indígenas y minoritarios, aunque fueran menos individuos por grupo. Reportado en

2016 como “300 genomas de 142 poblaciones diferentes” [21], incluyendo poblaciones indígenas de manera explícita (a diferencia de proyectos previos que representaron continentes más que poblaciones específicas). El estudio incluyó por primera vez poblaciones indígenas americanas tales como: Maya, Ayuuk ja’ay (Mixe), Ñuu Savi (Mixteca), Nahua, o’ob (Pima), Zapoteca, entre otras. Los datos se encuentran disponibles para quienes cumplan con filtros de acceso y acuerdos de confidencialidad.

The Human Genome Diversity Project. Es un proyecto interdisciplinario continuo que aspira a registrar toda la diversidad genética de los grupos poblacionales del mundo. La intención es poder conocer toda la variación regional, que estudiada en conjunto permitirá construir la historia evolutiva completa de la especie humana. En 2020 se publicó su primer reporte de genomas completos, incluyendo 929 genomas pertenecientes a 54 poblaciones en su geografía, lingüística, y cultura [22]. En este estudio se incluyeron las siguientes poblaciones indígenas americanas: Karitiana, Maya, o’ob (Pima), y Surui.

Además de las iniciativas internacionales ya mencionadas, se han desarrollado proyectos de secuenciación de genoma completo más “locales”. Estos han sido enfocados e impulsados en países específicos como: Turquía, Países Bajos, Japón, Australia, Dinamarca, México, Canadá, Perú y Rusia (Figura 2).

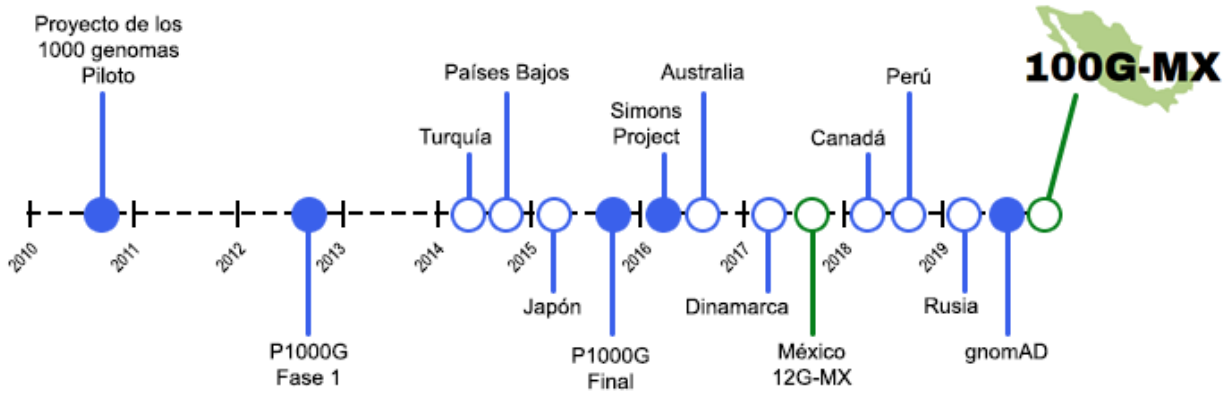


Figura 2. Línea del tiempo de proyectos de genómica poblacional con datos de genoma completo.
La década 2010 - 2020 vio el apogeo de los proyectos poblacionales con tecnología de genotipificación masiva. Algunos países llevaron a cabo estudios en su región para incrementar el conocimiento genómico local, generando catálogos de variación nacional.

Con la información genómica disponible hasta el momento se puede concluir que la mayoría de las variantes comunes (aquellas presentes en más del 5% de la población) ya han sido identificadas en la especie humana (Figura 1). En contraste, las variantes de baja frecuencia (presentes con una frecuencia menor al 5%), suelen ser alelos de aparición relativamente reciente [23], por lo cual muestran un agrupamiento geográfico y son diferenciables entre poblaciones [15]. Por lo anterior es necesario llevar a cabo estudios genómicos de poblaciones específicas para contribuir a la construcción de un catálogo más representativo de la variación en el genoma humano a nivel mundial [16].

Actualmente en México la población en general presenta un genoma heredado por la conjunción histórica de ancestros indígenas americanos, europeos y personas africanas ocurrida durante el periodo colonial [24]. Desde entonces los diferentes procesos demográficos han formado un panorama genómico complejo a lo largo del país, dando lugar a la heterogeneidad genética observada entre diferentes subpoblaciones o regiones de México. Como resultado, los mexicanos pueden presentar una similitud genómica con individuos indígenas americanos que varían dependiendo de la región geográfica [25], es decir en el norte se observa un mayor porcentaje de similitud con poblaciones europeas de referencia, en el sur de México un mayor porcentaje de similitud con población de referencia indígena americana, y en las costas un Guerrero y Veracruz un mayor porcentaje de similitud con población africana.

En México habitan diferentes poblaciones indígenas que son representantes actuales de dicha herencia genómica. Los estudios genómicos de estas poblaciones no sólo permiten conocer su diversidad genética, también aportan conocimiento sobre la diversidad genómica de la población en general.

El estudio de poblaciones indígenas en América

Las poblaciones indígenas de América están subrepresentadas en el acervo genómico global. Las bases de datos internacionales contienen un porcentaje muy bajo de representantes de estas poblaciones. A pesar de eso, el estudio de la genómica en poblaciones indígenas de América es un campo de estudio en crecimiento reciente. La poca representación de poblaciones no-europeas en ciencias genómicas es un hecho bien documentado, y un tema de discusión actual [26–29]. Menos del 1% de los individuos incluidos en estudios tipo GWAS (estudios de asociación de genoma completo, por sus siglas en inglés, enfocados en encontrar asociaciones estadísticas entre fenotipos y variantes a nivel genómico) son representantes de minorías indígenas americanas [29]. Lo anterior limita el alcance de potenciales beneficios que pudiera traer el conocimiento genómico para estas poblaciones, p. ej. aplicaciones en salud y política pública, o información antropológica derivada de las secuencias genómicas de pueblos representativos.

Estos son algunos ejemplos de beneficios traídos por proyectos genómicos realizados en poblaciones indígenas alrededor del mundo: un estudio genético de la población Gitxsan en Canadá encontró una variante genética asociada a un incremento en el riesgo de padecer arritmia y muerte súbita cardiaca, lo cual permitió mejorar el diagnóstico y la atención médica en los individuos portadores de la variante [30]; en Aotearoa (Nueva Zelanda) el estudio genómico poblacional de la tribu Ngāti Porou permitió identificar asociaciones entre variantes y la enfermedad de gota, generando evidencia necesaria para mejorar el diagnóstico y tratamiento de la enfermedad [30]; en México, los estudios genómicos contribuyeron a la identificación de un haplotipo asociado con el riesgo a padecer diabetes tipo 2, el cual afecta al gen SLC16A11 y se encontró presente en el ~50% de los individuos indígenas americanos estudiados [31].

América Latina es el territorio donde habitan diversas

poblaciones indígenas que están subrepresentadas en los estudios genómicos. Esto puede deberse a que muchos de los países latinoamericanos no cuentan con el presupuesto para proyectos poblacionales extensos, o a la limitada capacidad técnica para analizar grandes cantidades de datos genómicos (incluyendo los recursos humanos y el acceso a cómputo de alto rendimiento o cómputo en la nube necesario para el procesamiento masivo de datos). Aún así, América Latina ha sido parte de esfuerzos internacionales que buscan explorar la diversidad genómica a nivel mundial, y ha generado datos genómicos que contribuyen al acervo mundial. Particularmente en INMEGEN desde 2005 se han llevado a cabo proyectos de muestreo y análisis genómico de individuos indígenas [4]; uno de estos, el proyecto MAIS (Metabolic Analysis in an Indigenous Sample) ha integrado 2,596 muestras representativas de diversas poblaciones mexicanas [5].

Diversidad genómica en poblaciones indígenas habitantes de América

El análisis de las variantes de baja frecuencia ha revelado regiones genómicas con evidencia de señales de selección en sitios funcionalmente relevantes del genoma. La distribución de frecuencias de las variantes, desde el punto de vista de la historia de las poblaciones, puede contribuir a entender cómo los pueblos se fueron adaptando a los diferentes ambientes [15]. Dicho de otra manera, las variantes "raras" a nivel mundial, aunque restringidas geográficamente, pueden tener relevancia funcional en los distintos ecosistemas donde se han establecido los pueblos antiguos. Particularmente para el continente americano, la historia evolutiva dió origen a variantes de alta frecuencia en las poblaciones indígenas americanas. Esto puede deberse a que las poblaciones ancestrales permanecieron aproximadamente 15,000 años en el puente de Beringia [32], antes de comenzar los principales procesos de colonización del continente. Este aislamiento pudo dar origen a variantes de alta frecuencia en las poblaciones actuales.

Para la medicina genómica lo anterior implica que la interpretación clínica de variantes debe tomar en cuenta la composición genómica del paciente o al menos la región geográfica de la que es originario. Aún más, refleja la necesidad de desarrollar proyectos de secuenciación que incluyan a individuos de diversas poblaciones para tener una mejor representación de la variación genómica humana [15].

La población mexicana actual es una población mezclada compuesta principalmente por un flujo génico de poblaciones indígenas de América y poblaciones europeas, con contribuciones de población africana [25,33,34]. La variación genómica atribuida al componente europeo ha sido ampliamente estudiada a partir de los resultados de secuenciación de genomas completos en proyectos poblacionales internacionales como 1000 Genomes Project, gnomAD, Simons Genome Diversity Project y Human Genome Diversity Project. Sin embargo, el estudio de la diversidad genómica del componente indígena comienza a estudiarse con más detalle gracias al acceso a datos de exomas y genomas completos en 2010 [14].

Como se mencionó previamente, los pueblos indígenas que habitan el continente americano han sido representados de manera escasa en los estudios genómicos. Como parte de este proyecto de Doctorado revisamos 56 publicaciones científicas donde se estudió el DNA de individuos indígenas habitantes de la región actual del continente americano o restos antiguos mediante el uso de tecnologías genómicas (microarreglos, WES y WGS) (Anexo 2). Encontramos que en total 13,706 indígenas americanos han participado en estudios poblacionales, de los cuales 1,292 corresponden muestras a las cuales se secuenció el genoma completo, incluyendo personas de todo el continente (Figura 3). En contraste con estos números, por ejemplo, hay 39,345 individuos habitantes de europa incluidos en la base de datos gnomAD [35]. No sólo el número de genomas secuenciados es limitado, el acceso a los datos es restringido (comprensiblemente tomando en cuenta la complejidad de

garantizar el involucramiento y beneficio real para los pueblos estudiados); únicamente el 3.6% de los datos de ADN actuales se encuentran públicamente disponibles para su estudio. De lo anterior deducimos que el estudio de las poblaciones indígenas americanas es un área en crecimiento, aunque debe mejorar su apertura en favor del acceso abierto a los datos. Esto promoverá que se reduzca la subrepresentación de estas poblaciones en los estudios genómicos a nivel mundial. Los resultados de nuestra revisión se encuentran publicados en Aguilar-Ordoñez, et al. Diversity, 2022 (Anexo 2) [14], y se describen en extenso en la tesis de Guzman-Linares 2022 (Anexo 4).

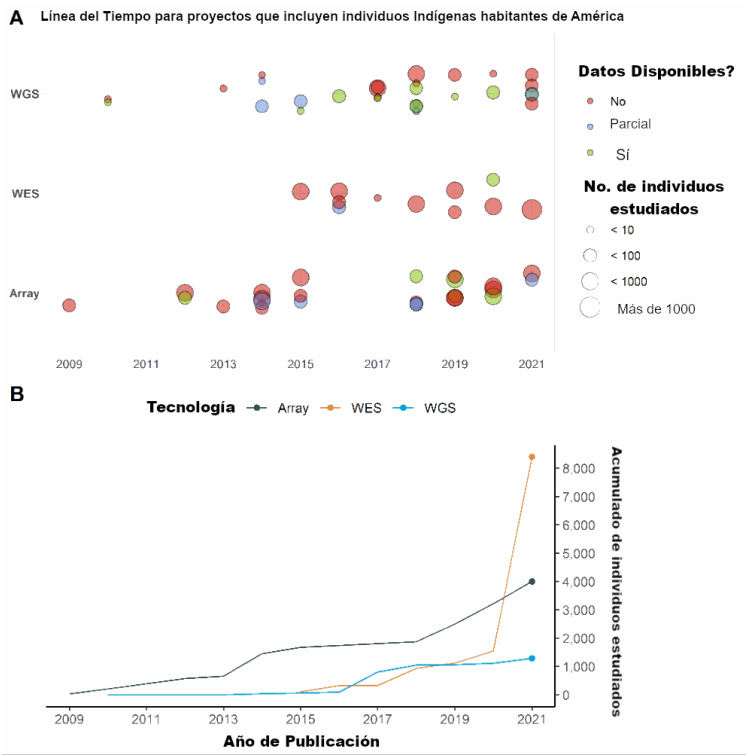


Figura 3. Historial de estudios genómicos en población indígena habitante del continente americano.
A) Línea del tiempo de proyectos de genómica poblacional que incluyen individuos indígenas americanos. Los proyectos se dividen de acuerdo a la tecnología utilizada; WGS: Whole Genome Sequencing; WES: Whole Exome Sequencing; Array: DNA arrays. El tamaño de los puntos señala la cantidad de individuos estudiados. B) La cantidad acumulada de datos genómicos indígenas habitantes del continente americano. Con el tiempo hay un mayor número de proyectos, y un mayor número de individuos por proyecto. Tomado de Aguilar-Ordoñez, *et al.*, 2022 [14].

Estudios sobre la diversidad genómica en México

La región del México contemporáneo fue formada originalmente por poblaciones ancestrales que crecieron rápidamente cuando se implementó la agricultura en el centro del México antiguo [24]. Los pobladores que se asentaron en el norte del país continuaron viviendo principalmente como cazadores-recolectores, un estilo de vida que no es compatible con poblaciones de gran tamaño [36,37]. La llegada de la conquista española en el siglo XVI cambió de manera drástica la distribución poblacional en la región [38], de modo que en el México contemporáneo la población es heterogénea, con genomas resultado de una mezcla de componentes ancestrales de diferentes orígenes geográficos, principalmente indígenas americanos, europeos, y africanos [34].

Actualmente casi 8 millones de mexicanos se reconocen como pertenecientes a uno de los 68 poblaciones indígenas a lo largo del país [39]. La diversidad cultural es consistente con la diversidad genética que se ha observado en estudios genómicos recientes [33]. La genómica en poblaciones indígenas mexicanas ha sido previamente analizada utilizando tecnologías como STR (repeticiones cortas en tándem, por sus siglas en inglés) [40], y tecnologías masivas como la secuenciación de exoma (WES) y de microarreglos de ADN [25,34,41]. Dichas tecnologías son por definición limitadas porque no alcanzan a analizar todas las posiciones del genoma. Para obtener un panorama más completo de la variación genómica es esencial generar estudios y datos de genoma completo. Previo a este trabajo de doctorado hubo esfuerzos internacionales por estudiar con genoma completo las poblaciones Mexicanas actuales [21,22,42] (incluyendo a individuos pertenecientes a los poblaciones indígenas), uno de estos se enfocó en poblaciones indígenas mexicanas [33]; hasta entonces, únicamente se había reportado 54 individuos indígenas mexicanos [21,22,33] estudiados mediante tecnología de genoma completo. En el presente trabajo se estudiaron las secuencias de 76 individuos no relacionados pertenecientes a 27

poblaciones indígenas distintas, utilizando tecnología de genoma completo. Este proyecto fue denominado 100G-MX y dio lugar a diversas publicaciones (Anexo 1 [13]).

Previamente en México, Moreno-Estrada y colaboradores exploraron los patrones de variación genómica regional en individuos pertenecientes a 20 poblaciones indígenas mexicanas [34]. Para ello utilizaron genotipificación por microarreglos de genoma completo. Además de aportar a una mejor definición de la estructura genómica, este estudio encontró una correlación entre el componente ancestral y la función pulmonar, sugiriendo la importancia clínica de estos estudios exploratorios. Posteriormente, Romero-Hidalgo y colaboradores exploraron la diversidad genómica de 12 individuos provenientes de poblaciones indígenas mexicanas utilizando datos de secuenciación de genoma completo [33].

De manera importante para este proyecto de tesis, previamente Orozco y colaboradores recolectaron y estudiaron de manera descriptiva y cuantitativa muestras aportadas por un total de 2,596 individuos indígenas mexicanos a través del estudio MAIS (Metabolic Analysis in an Indigenous Sample). Desde 2012 hasta 2017, el estudio MAIS reclutó voluntarios de 73 comunidades indígenas representativas de 60 poblaciones diferentes [2]. Es importante mencionar que durante esta labor el grupo de la Dra. Orozco trabajó de cerca con los líderes de las comunidades así como con la Comisión Nacional para el Desarrollo de Pueblos Indígenas. En 2020 se publicó un estudio descriptivo en la cohorte MAIS [2] donde se reportó una alta prevalencia del síndrome metabólico en individuos con ancestría indígena mexicana.

A partir de la cohorte MAIS se hizo una selección de muestras para llevar a cabo un análisis de genoma completo. Como se mencionó previamente, este proyecto subsecuente fue denominado 100G-MX [13], y sus resultados, alcances y limitaciones se describen a continuación como parte de mis estudios de doctorado.

Justificación

Estudiar los genomas de individuos pertenecientes a poblaciones indígenas americanas es necesario para entender su diversidad genómica, así como para identificar posibles factores genéticos relacionados con enfermedades y otros rasgos biomédicos importantes para la salud. La variación genómica indígena ha sido explorada de manera muy limitada, por lo que es fundamental incrementar el número de genomas completamente secuenciados de individuos pertenecientes a distintas poblaciones indígenas. El catálogo de variantes que resulte de nuestra investigación formará parte de un acervo genómico de interés para la población mexicana, contribuyendo así al desarrollo de herramientas para el avance de la medicina genómica del país.

Hipótesis

Existen variaciones genómicas en las poblaciones indígenas mexicanas relacionadas con fenotipos de interés clínico y aspectos demográficos.

Objetivos

General:

- Estudiar la variación genómica de poblaciones indígenas mexicanas.

Específicos:

- Analizar la variación genómica de 100 individuos de 32 poblaciones indígenas mexicanas.

- Describir la estructura genética en las poblaciones estudiadas a partir de datos de genoma completo.

- Identificar SNPs de interés biomédico.

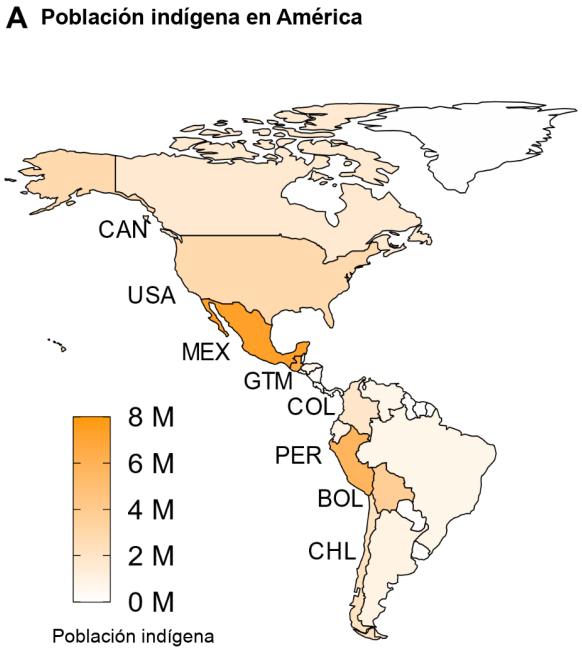
Resultados

I. Compendio de estudios genómicos de poblaciones nativas de América

Los datos genómicos pertenecientes a poblaciones indígenas del continente americano son escasos, pero los investigadores (especialmente en los países Latinoamericanos con presupuestos limitados) pueden aprovechar una cualidad importante de los datos: cuando se juntan los datos genómicos, se incrementa la información. Por lo anterior es importante contar con un compendio que recopile cada esfuerzo científico realizado para genotipificar o secuenciar poblaciones indígenas americanas. Esta información sería una buena referencia para investigadores que en el futuro busquen ampliar la representación de estas poblaciones en los estudios genómicos.

Como parte del trabajo de esta tesis de doctorado se dirigió un proyecto de licenciatura para recopilar estudios genómicos publicados que incluyen individuos indígenas habitantes del continente americano, donde se hubieran reportado datos de secuenciación o genotipificación masiva. Los resultados se encuentran publicados en Aguilar-Ordoñez, *et al. Diversity*, 2022 (Anexo 2), y se describen en extenso en la tesis de Guzman-Linares 2022 (Anexo 4). Clasificamos los estudios de acuerdo a la tecnología genómica utilizada (microarreglos, secuenciación de exoma, o secuenciación de genoma completo) y de acuerdo a la disponibilidad de los datos en cada proyecto (público o privado; los estudios con acceso restringido pero disponible a partir de la firma de convenios o responsivas se consideraron privados). En los 56 estudios revisados se estudiaron individuos actuales identificados como miembros o descendientes de poblaciones indígenas habitantes de México, y algunos también incluyeron muestras de ADN antiguo. El análisis resumido de dichos estudios y la comparación con los censos poblacionales de los países hogar (definidos como aquellos países que reconocen

a su población indígena a través de la inclusión censal), nos muestra cuáles son los grupos Indígenas que faltan por incluir en proyectos genómicos (Figura 4).



B Poblaciones indígenas no representadas en genómica

Tamaño de la Población

	menos de 100	de 100 a 1K	de 1K a 10K	de 10K a 100K	de 100K a 1M
Brazil		13	80	107	21
Colombia	3	15	32	37	20
Venezuela		8	16	12	12
USA	8	20	14		
Peru	1	6	12	13	9
Bolivia		8	13	11	2
Argentina		8	9	5	
Paraguay		4	10	3	
Mexico	1	1	5	4	4
Ecuador	1	2	5	4	1
Honduras	1	5	2		
Panama		2	4	1	
Nicaragua	1	3	2		
Chile		4			
El Salvador			4		
Canada	2				
Costa Rica			2		
Guatemala	1	1			
Dominica			1		

Figura 4. Los países hogar de poblaciones indígenas en América y la diversidad genómica faltante.
A, el mapa resalta la distribución de las poblaciones indígenas a lo largo del continente, con etiquetas en aquellos países con que son hogar de al menos un millón de habitantes indígenas (Canadá, USA, México, Guatemala, Colombia, Perú, Bolivia, y Chile). B, el mapa de calor (heatmap) muestra el número de poblaciones indígenas que no tienen inclusión en bases de datos genómicas; las poblaciones se encontraron por referencia cruzada de datos censales y las etiquetas reportadas en estudios genómicos.

II. Secuenciación y catálogo de la variación genómica en poblaciones indígenas habitantes de México

El proyecto 100G-MX fue una iniciativa conjunta entre el Instituto de Biotecnología, UNAM, y el Instituto Nacional de Medicina Genómica (INMEGEN), con el propósito de incrementar la comprensión de las características genómicas de las poblaciones que habitan México, así como para incrementar el conocimiento base requerido para el desarrollo de la medicina genómica a través del estudio de genomas completos en población mexicana. En este proyecto se llevó a cabo

la secuenciación de 95 genomas completos de personas pertenecientes a 32 poblaciones indígenas. Los individuos pertenecen a la cohorte del "Metabolic Analysis in an Indigenous Sample" (MAIS) descritos por Contreras-Cubas [43] y Mendoza-Caamal [2]. De los 95 individuos, 19 no continuaron en el análisis (Figura 5), debido a los siguientes criterios: una muestra se retiró por posible contaminación de muestra (presentaba cerca de 23% más variantes que el promedio); cinco se retiraron por estar emparentados hasta segundo grado con algún otro individuo del estudio, determinado por un coeficiente de Kinship > 0.0884 [44] (en casos de parentesco, se eliminó el individuo de menor edad); y trece se retiraron por presentar menos del 85% de similitud genómica con el grupo formado por la mayoría de los individuos estudiados, lo cual asumimos representa su similitud con antepasados indígenas.

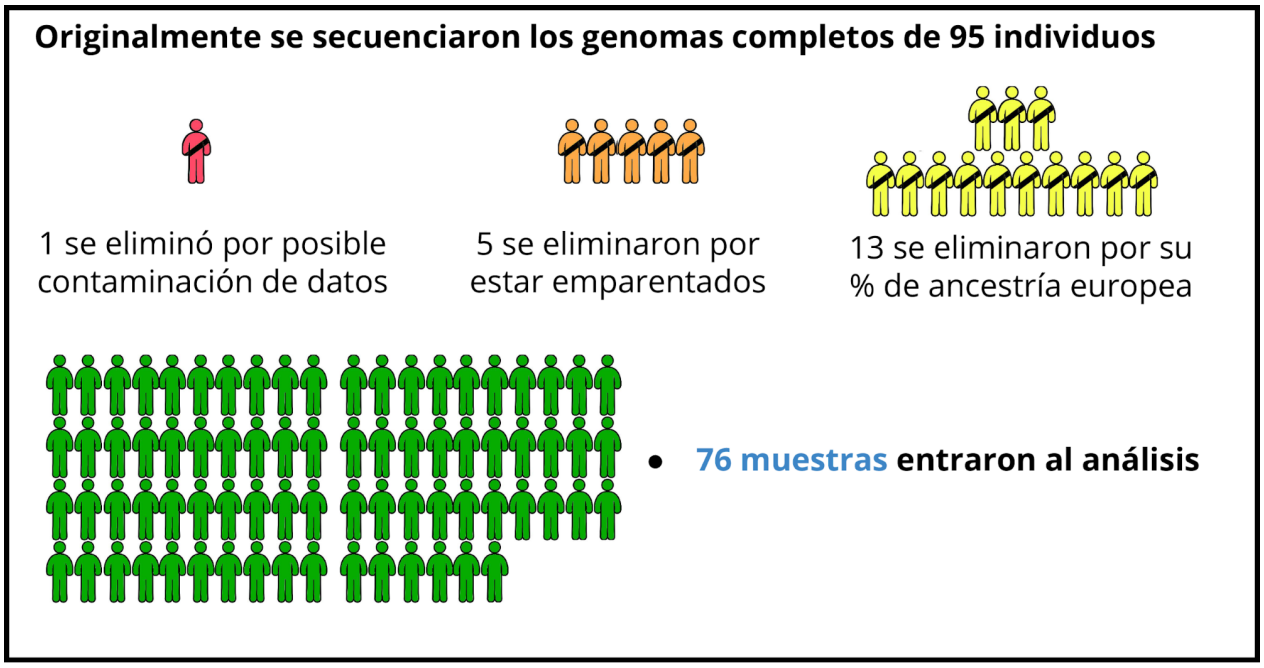


Figura 5. Filtrado del dataset 100G-MX.
A partir de un llamado de variantes preliminar se determinó parentesco y similitud genómica global (usando datos de 1000 Genomes Project como referencia). De 95 individuos originalmente secuenciados 19 no cumplieron con criterios para continuar con el análisis.

El promedio de similitud con población indígena americana en los 76 individuos determinado por ADMIXTURE fue de 97.22% (Figura 6). Este conjunto de datos final está compuesto por 40 mujeres y 36 hombres, provenientes de 27 poblaciones indígenas distintas (Figura 7).

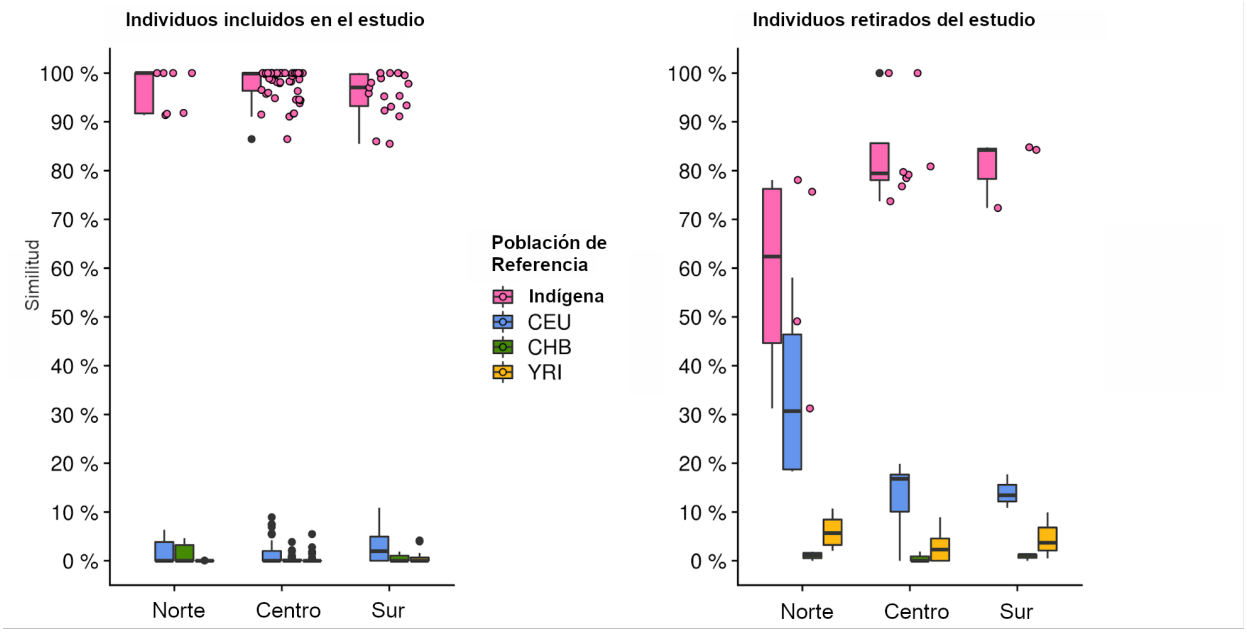


Figura 6. Porcentaje de similitud genómica en los individuos incluidos y retirados del dataset.
Distribución de la similitud genómica con individuos etiquetados por región geográfica analizada con ADMIXTURE (con datos de 1000 Genomes Project como referencia). Los individuos incluidos en el estudio presentan una proporción de similitud con población indígena americana > 85%. En contraste, los individuos retirados del estudio presentan proporciones variables de similitud con población etiquetada como Europea (CEU), Africana (AFR) y de Asia del Este (EAS). El criterio de inclusión por similitud fue de al menos 85 %. Algunos individuos de la región centro fueron eliminados a pesar de su alta similitud genómica con la población indígena debido a que estaban emparentados con alguno de los otros individuos originalmente secuenciados.

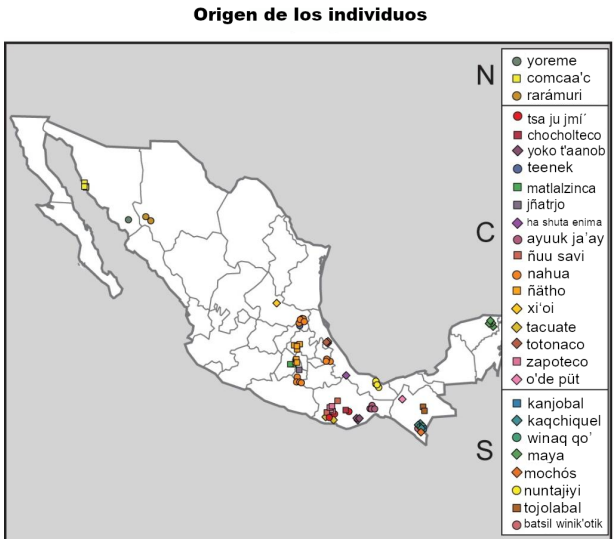


Figura 7. Origen de las 27 poblaciones indígenas en el estudio.
Cada punto representa el origen geográfico aproximado de un individuo, y los puntos idénticos aglomerados indican el número de muestras provenientes de un mismo lugar. La leyenda se separa en región Norte (N), Centro, y Sur (S) de acuerdo a los resultados de estructura poblacional que se describen más adelante en este trabajo. Para conocer las coordenadas GPS consultar el Material Suplementario del Anexo 1.

Los datos de secuenciación se procesaron como se describe en la sección MÉTODOS, utilizando la versión GRCh38 del genoma de referencia. Para identificar la variación en el conjunto de datos analizado, el llamado de variantes cortas (SNVs e Indeles cortos) se llevó a cabo con el software GATK 3.8, siguiendo los protocolos de buenas prácticas recomendados por el Broad Institute. Se realizaron los 76 llamados de variantes individualmente. Posteriormente se integraron en un llamado de variantes conjunto denominado JPVS (del inglés Joint Population Variant Set), con un call rate > 98.5%. La Tabla 1 muestra los aspectos generales del proyecto de secuenciación.

Tabla 1: Resumen de las características de la secuenciación.

Versión del genoma	GRCh38
Profundidad de cobertura promedio*	24.08 X
Porcentaje de genoma cubierto*	95.09 %
Ancestría nativa americana promedio	97.22 %
Promedio de lecturas totales¶	542,098,707
Promedio de variantes totales*†	3,262,160
Promedio de densidad de variantes*†	3.315 per kb
Promedio de transiciones/transversiones*†	2.127
Promedio de variantes biomédicamente relevantes*†§	12,871
Promedio de variantes novel *†	1,814
Promedio de singletons *†	14,841
Promedio de concordancia con SNP-Array	99 %
Promedio de genotipos faltantes (posiciones sin información de lecturas) *†	6,347

* En autosomas, y cromosomas sexuales
† Variantes que pasan los filtros de GATK, así como posiciones con call rate Z 98.5 % (AN >= 150)
§ Genotipos con registros en GWAS catalog, ClinVar o PharmGKB
¶ Lecturas de calidad después del control de calidad y trimming

En total el JPVS está compuesto por 8,638,130 SNVs y 1,099,022 indeles cortos. De este conjunto 44,118 variantes no están incluidas en dbSNP b152, por lo que se consideran novel. La precisión tanto de la secuenciación como del llamado de variantes se evaluó comparando datos de microarreglos para las mismas muestras (Affymetrix, 6.0 SNP array), donde se observó más del 99% de concordancia promedio. En promedio, cada individuo presentó 3,262,160 variantes respecto al genoma de referencia (Figura 8a), una cantidad dentro del rango observado (~ 2.7 a 3.7 millones) en otros proyectos internacionales [16,45,46]. Aproximadamente el 0.7 % de los SNVs y 0.2 % de los INDELs se localizaron sobre regiones codificantes (Figura 9). La proporción de transiciones y transversiones (Ts/Tvs) fue de 2.127, concordando con lo observado en otros proyectos. Los individuos pertenecientes al grupo Comcaa'c mostraron el menor número de singletones (variantes que no se comparten con ningún otro individuo ni de los diversos proyectos internacionales, como 1000 genomes project, o gnomAD 2.1, ni con el resto de los individuos de 100G-MX), así como el mayor número de variantes novel (Figura 8b y 8c). Ambos resultados son característicos de una población aislada, no explorada previamente por secuenciación de genoma completo.

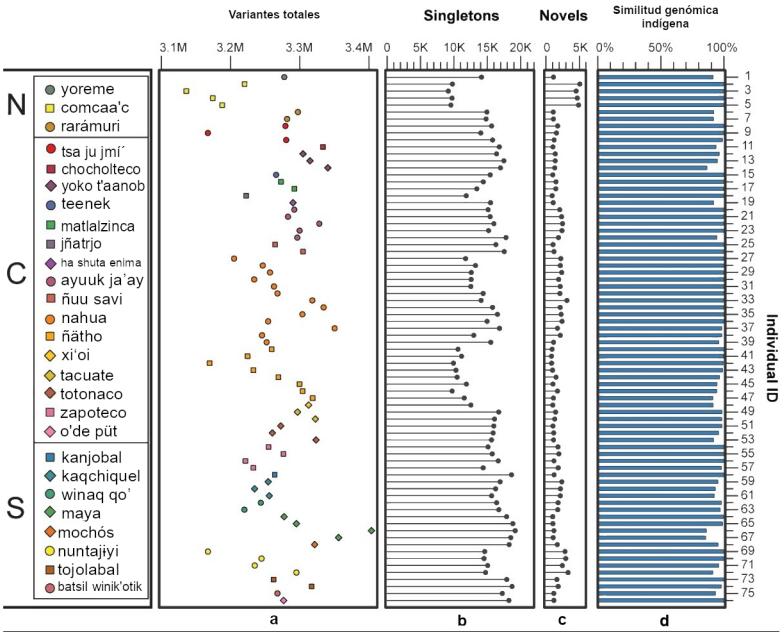


Figura 8. Número de variantes y similitud genómica.
Variantes totales (M = millones) para cada uno de los 76 individuos; el eje y enumera cada una de las muestras y se comparte entre los paneles a, b, c y d. La leyenda separa a los grupos en regiones Norte (N), Centro (C) y Sur (S). b, Número de singletones (K = miles) inferidos de la comparación mundial con gnomAD y el 1000 genomes project. c, Número de variantes novel (K = miles) no registradas en dbSNP b152. d, Porcentaje de similitud con población indígena habitante de america.

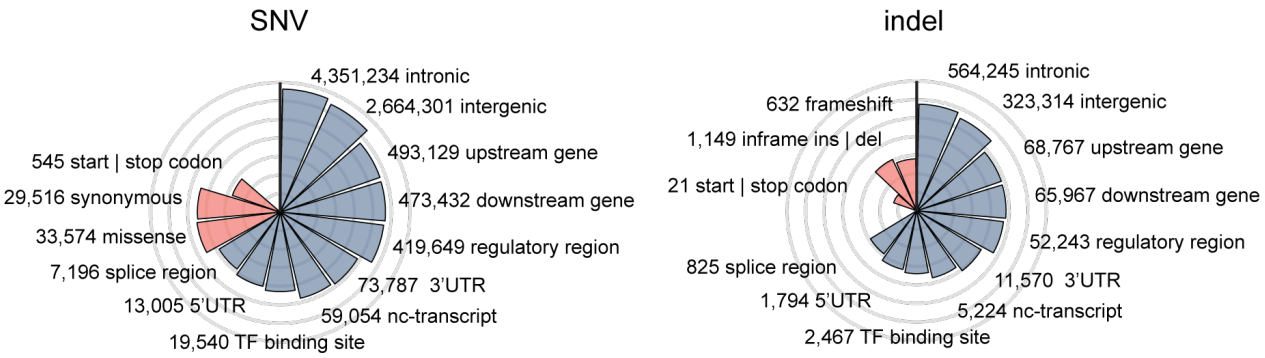


Figura 9. Resumen del efecto predicho para el catálogo de variantes detectadas en población indígena de México. Los gráficos indican el log10 del número de variantes. Se muestra el total de consecuencias en SNVs e indels.

III. Análisis demográfico de las poblaciones indígenas incluidas en el estudio

Se integró el catálogo de variantes en un nuevo conjunto de datos inter-continental (denominado IPVS, por sus siglas en inglés "Inter-population variant set") el cual incluye 358 individuos adicionales del proyecto 1000 genomes [11], incluyendo a 4 individuos de Lima Perú con alta ancestría nativa peruana (NP). Se estimó la estructura genética intercontinental mediante un análisis de PCA sobre los genotipos del IPVS, observando un patrón similar a reportes previos [47] (Figura 10). El análisis muestra que los individuos peruanos con similitud genómica indígena > 85% aparecen cerca de los individuos indígenas mexicanos estudiados en la proyección bidimensional de los primeros dos PC. Esto generó la pregunta de ¿cuál es la estructura genómico-demográfica dentro del continente americano inferida a partir de genomas públicamente disponibles?.

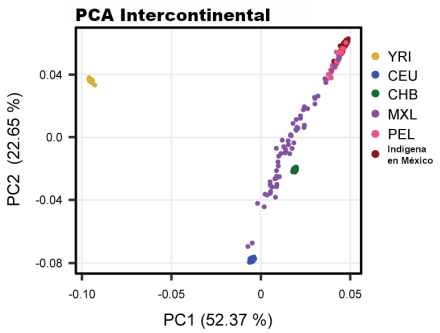


Figura 10. Análisis de similitud de genotipos por PCA intercontinental.
Se muestran representantes de las poblaciones etiquetadas como yoruba - africana (YRI), europea (CEU), china (CHB), mexicanos de Los Ángeles (MXL), indígenas peruanos (NP), e indígenas mexicanos, según fueron etiquetados por el 1000 Genomes Project.

A partir de la observación anterior se analizó la estructura local de los individuos indígenas mexicanos por PCA incluyendo a los 4 individuos indígenas peruanos (Figura 11a). El PC1 separa al a los individuos Comcaa'c del resto de las poblaciones, mientras que los individuos Rarámuri y Yoreme se agrupan juntos; los individuos del centro y sur del país también se agrupan, mientras que los 4 individuos indígenas peruanos forman su propio cluster. Se procedió a definir la regionalización de los individuos mediante un análisis de agrupamiento k-means no supervisado, inferido a partir de los 8 primeros componentes principales [48]. Con un valor de k = 5, los individuos Comcaa'c forman su propio grupo; los individuos indígenas peruanos también forman su grupo, y el resto de las muestras se agrupan en Norte, Centro y Sur (Figura 11 bis). Utilizando este agrupamiento regional no supervisado asignamos a los individuos las categorías Norte, Centro y Sur para el resto de los análisis demográficos (el grupo Comcaa'c se incluyó en el Norte debido a su localización geográfica). Para simplificar los patrones regionales observados por PCA calculamos los centroides de cada PC resumidos por región geográfica (Figura 11b). En PC2 y PC3 los individuos indígenas peruanos se separan en su propio eje de variación. En el resto de los PCs (PC4 a PC8, que acumulan el 74.92% de la varianza explicada) la distribución individual de las coordenadas muestra que los indígenas mexicanos tienen valores muy cercanos a los indígenas peruanos.

Se realizó análisis de ADMIXTURE para explorar la similitud genómica y posible estructura poblacional (Figura 11c). Al analizar la estructura utilizando k = 3, se observa un gradiente Norte-Centro-Sur con el grupo Comcaa'c nuevamente distinguiéndose de manera clara. En k = 4, se observa mayor subestructura en las regiones Centro y Sur. En k = 5, los individuos indígenas peruanos forman su propio grupo. Cabe mencionar que los individuos indígenas peruanos comparten patrones de variación cercanos con los individuos indígenas mexicanos del sur, en concordancia con los datos PC1 y PC2 del análisis PCA previo.

Para identificar subgrupos basados en similitudes genómicas aplicamos un análisis de k-means con valores desde k = 2 a k = 20. El agrupamiento óptimo determinado por el silhouette index más alto, fue de k = 2, donde los individuos Comcaa'c forman un grupo, y todo el resto de los indígenas mexicanos estudiados forman un segundo grupo (Figura 11 bis). Dado que esto se debe a la alta similitud entre los individuos del grupo Comcaa'c (que básicamente los distingue de todos los demás mexicanos en el estudio), utilizamos k = 9 (el segundo mejor silhouette index) para identificar subagrupamientos: la región del sur sólo se subdividió en los individuos Popoluca y el resto de los individuos sureños (conocidos colectivamente como población mayense debido a la familia lingüística a la que pertenece). Tomando en cuenta los resultados de análisis PCA, ADMIXTURE y el agrupamiento por k-means, la evidencia sugiere que los indígenas peruanos están más cerca del grupo mayense que de las poblaciones indígenas mexicanas del centro o del norte. Una observación similar se reportó previamente en el proyecto de genómica poblacional en Perú [49].

Se analizó la subestructura poblacional a través de la medición de divergencia genética por FST entre poblaciones seguido de un agrupamiento utilizando el método de Neighbor-joining (Figura 11d). Los resultados muestran que las poblaciones se agrupan de acuerdo a su origen geográfico, representando el gradiente Norte-Centro-Sur, en concordancia con el reporte previo realizado con datos de microarreglos a partir de 9 poblaciones indígenas en México [50].

En todos los análisis previos se distingue a la población Comcaa'c, también en concordancia con un estudio previo donde se reportaron altos niveles de divergencia entre este y otros grupos indígenas habitantes de América, utilizando genotipificación por microarreglos [34]. En el presente estudio confirmamos esta divergencia mediante secuenciación de genoma completo, lo cual además nos permitió identificar 2,496 SNVs con una frecuencia alélica de al menos el 50% en la población Comcaa'c y una frecuencia rara (menor al 5%)

en el resto de los individuos indígenas mexicanos. La existencia de estas variantes en alta frecuencia probablemente es un reflejo del aislamiento y el reducido tamaño poblacional al que históricamente se enfrentó la población Comcaa'c.

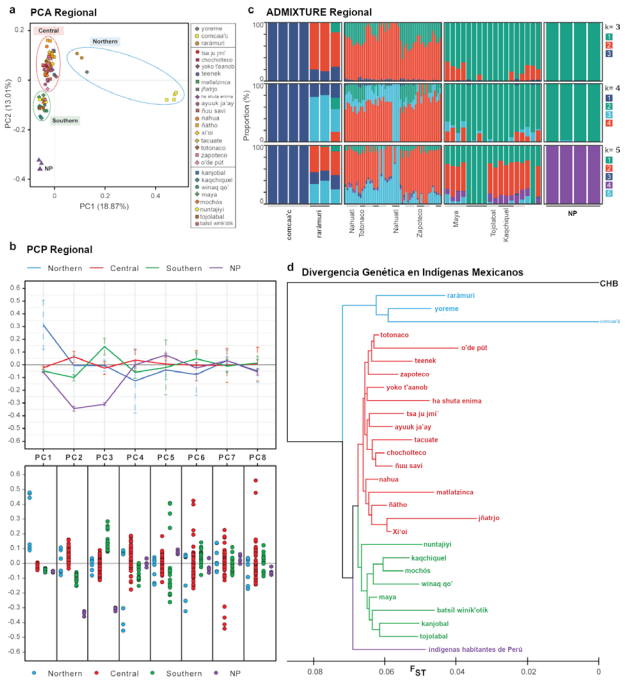


Figura 11. Análisis demográficos en población indígena de Mexico del proyecto 100G-MX.
a, Análisis PCA incluyendo a 4 individuos indigenas peruanos (NP) como grupo externo. b, Coordenadas de componente principal para los PCs significativos; panel superior, valores de PC por región, las líneas contínuas muestran el valor promedio, y las líneas punteadas muestran la desviación estándar; panel inferior, cada punto representa un individuo y sus coordenadas en cada PC. c, Análisis de ADMIXTURE para diferentes valores de k, las muestras están ordenadas de acuerdo a la latitud geográfica y la población. d, Neighbor-joining tree basado en valores de FST entre los 27 poblaciones indígenas de Mexíco estudiadas; los colores señalan la región indicada en la figura b.

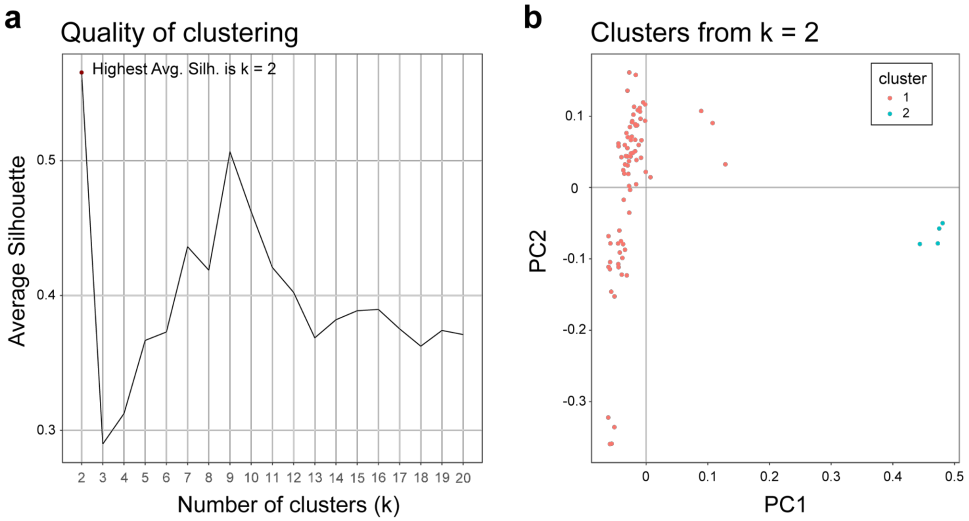


Figura 11 bis. Sub-agrupamientos poblacionales basados en k-means.
a, evaluación de clustering óptimo en diferentes valores de k. **b**, el clustering más óptimo divide el dataset en dos, un grupo Comcaa'c, y otro grupo con el resto de indígenas mexicanos e indígenas peruanos.

IV. Variantes de interés Biomédico

En promedio, por individuo encontramos casi 13 mil variantes presentes en bases de datos de asociación genotipo-fenotipo (GWAS catalog, PharmGKB, o ClinVar). En el catálogo general de los cerca de 9 millones de variantes de 100G-MX, encontramos 497 variantes comunes (presentes en más del 5% de la población de estudio) asociadas a algún gen relacionado con respuesta a medicamentos, según la base de datos PharmGKB [51]. También encontramos casi 14 mil variantes relacionadas a fenotipos clínicos de acuerdo a la base de datos ClinVar [52]. En promedio, cada individuo fue portador de 8 variantes homocigotas con significancia clínica de nivel patogénico o posiblemente patogénico (ClinVar).

Reportamos 316 mil variantes comunes localizadas en elementos tipo enhancer o promotores de la expresión a lo largo del genoma humano (Figura 12), lo cual suma casi el 6% de toda la variación común encontrada en el proyecto. Entre los principales genes con variación en elementos reguladores se encontraron: CMIP, una proteína inductora de c-Maf implicada en la señalización de células T, y NCOR2, un co-represor transcripcional relacionado con procesos de modificación de la estructura de la cromatina. Estos genes son un ejemplo del subconjunto reportado en el proyecto 100G-MX (Anexo 1 [13]), donde aprovechamos la base de datos GeneHancer [53] (una colección bien curada de elementos reguladores) para explorar la variación con posibles efectos de regulación génica, algo que comúnmente no se explora. Esperamos que en futuros trabajos se tome en cuenta la posibilidad de diferentes perfiles de expresión génica relacionada con la diversidad genética poblacional.

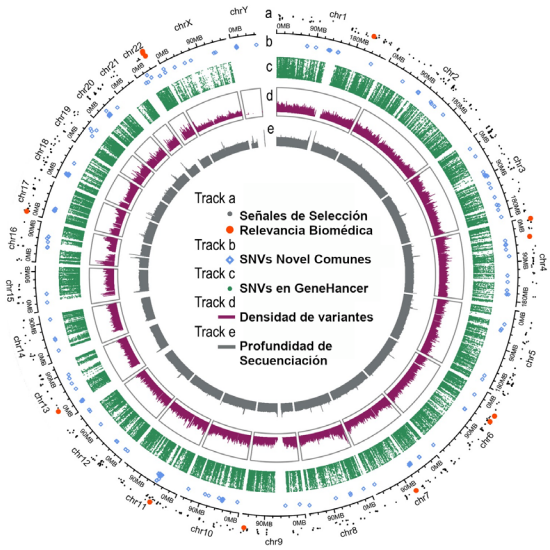


Figura 12. Panorama de la variación en genomas indígenas de México en el proyecto 100G-MX.
a, variantes bajo señales de selección, las señales relacionadas con salud (de acuerdo a GWAS catalog o ClinVar) se marcan en naranja. b, variantes novel con frecuencias mayores al 5%. c, Variantes en elementos enhancer o promotores. La altura de los puntos en a, b y c señalan la frecuencia alélica. d, Densidad de variantes a nivel poblacional. e, Cobertura promedio de secuenciación.

Uno de los genes con variación en regiones reguladoras fue PPARG, un receptor nuclear que controla el metabolismo de lípidos y adipogénesis [54]. En PPARG se encontraron 10 SNVs distribuidos en 3 elementos enhancer. Las 10 variantes muestran un patrón de frecuencias alélicas muy similar, lo cual sugiere la existencia de un haplotipo. Dicho haplotipo está ausente en individuos representantes de poblaciones de Asia del Este, y es más corto en representantes de la población europea y africana según las etiquetas del 1000 genomes project (Figura 13); este patrón de frecuencias podría deberse al "Beringia Standstill", una teoría que propone una separación entre la población ancestral de Asia y los pobladores de Beringia hace aproximadamente 30,000 años [55]. Está región de aproximadamente 20 kilobases parece ser una región de interés para futuros estudios. Dado que PPARG se ha reportado como involucrado en diferentes padecimientos como obesidad, diabetes, aterosclerosis [56], y resistencia a insulina en niños mexicanos [57], estas 10 variantes ameritan un análisis de seguimiento, incluyendo la validación en un estudio que incluya más individuos, pertenecientes a otros grupos poblacionales de México, así como la validación experimental del efecto de estas variantes sobre el gen regulado.

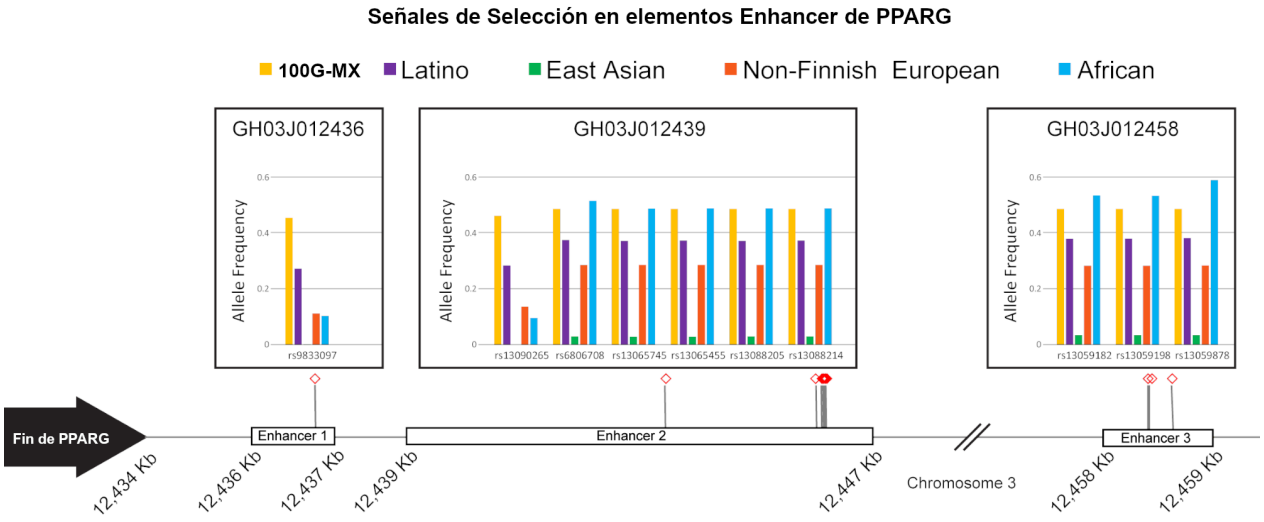


Figura 13. SNVs de interés en las regiones Enhancer del gen PPARG.
El patrón similar de frecuencias alélicas sugiere la existencia de un haplotipo en población indígena de México, ausente en la población de Asia del este, y más corto en poblaciones Africanas y Europeas. Los rombos rojos señalan el sitio donde se encontraron señales de selección. Adaptado de González Buenfil, 2020.

Con la intención de explorar posibles cambios importantes en elementos no codificantes del genoma, identificamos variantes en regiones codificantes de microRNAs utilizando la base de datos miRBase v22.1 como referencia. Encontramos 382 variantes de nucleótido único en regiones microRNA, de las cuales 68 son variantes comunes a la población (con una frecuencia alélica mayor al 5%) que alteran la secuencia de un microRNA maduro; de estas, 27 alteran la secuencia de la región semilla, o seed (una región encargada del reconocimiento entre el microRNA y su transcrito blanco) (Tabla 2). Cuando observamos el número de variantes comunes de acuerdo a su posición en el microRNA maduro (Figura 14) se observa que la región semilla es un sitio de variación frecuente. Las posiciones 9, 10, 20 y 22 parecen ser sitios poco variables.

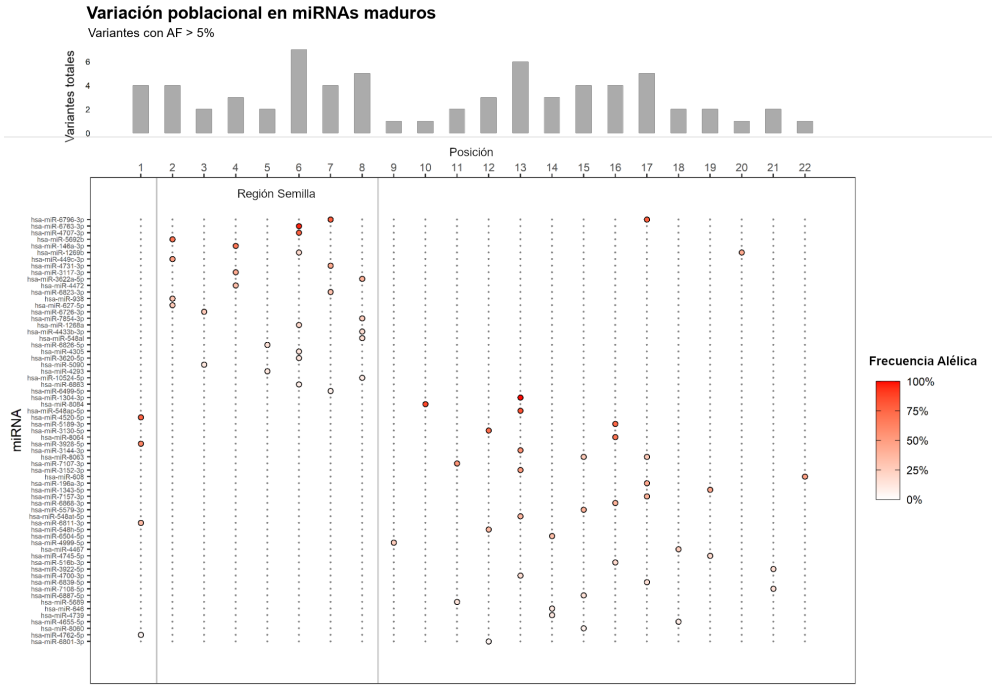


Figura 14. Posición de las variantes en microRNAs maduros.
Se muestra la distribución de 68 SNPs que alteran microRNAs maduros en población indígena que habita México; el total de variantes en cada posición se muestra como columnas grises en la parte superior de la gráfica; en la parte inferior se muestra la posición exacta en la que varía cada miRNA alterado y la intensidad del punto indica la frecuencia alélica con que se detectó la variante. Encontramos 27 miRNAs con cambios en la región seed (posiciones 2 a 8).

Cuando comparamos las frecuencias alélicas en la población de estudio con las de otras poblaciones del mundo, utilizando la base de datos gnomAD, 4 variantes miRNA llamaron nuestra atención porque se encuentran en baja frecuencia en las demás poblaciones, y en mayor frecuencia en nuestra población de estudio (Figura 15). Las poblaciones de comparación fueron los individuos etiquetados como europeos, asiáticos del este, asiáticos del sur, africanos y latinos (admixed). La variante hsa-miR-4305 es particularmente interesante porque se encuentra en el 13% de la población indígena de México, pero es rara en cualquier otra población (excepto en la población etiquetada como latina en gnomAD). Esta variante chr13_39664120_T_C provoca un cambio de A por G en la posición 6 de la región semilla del hsa-mir-4305. Las otras variantes interesantes son chr15_42199650_A_C en hsa-miR-627-5p, chr1_1296127_G_A en hsa-miR-6726-3p, y chr16_56904332_G_A en hsa-miR-6863.

Tabla 2. Variantes miRNA en población nativa mexicana.
Se muestran sólo las variantes con una frecuencia alélica mayor al 5%(AF > 0.05). En las secuencias se muestra el alelo de referencia en azul y el alelo variante en rojo. En verde se marcan los microRNAs donde la variante es menos frecuente en otras poblaciones del mundo.

miRNA	sequence	variant	AF
hsa-miR-6796-3p	GAAGCUC/GUCCCCUCCCCGCAG	chr19_40369893_C_G	0.78
hsa-miR-6763-3p	CUCCCC/UGGCCUCUGCCCCCAG	chr12_132582046_C_T	0.96
hsa-miR-4707-3p	AGCCCG/UCCCCAGCCGAGGUUCU	chr14_22956973_C_A	0.77
hsa-miR-5692b	AA/GUAAUAUCACAGUAGGUGU	chr21_42951004_T_C	0.68
hsa-miR-146a-3p	CCUC/GUGAAAUUCAGUUCUUCAG	chr5_160485411_C_G	0.67
hsa-miR-1269b	CUGGAC/GUGAGCCAUGCUACUGG	chr17_12917329_G_C	0.19
hsa-miR-449c-3p	UU/AGCUAGUUGCACUCCUCUCUGU	chr5_55172296_A_T	0.49
hsa-miR-4731-3p	CACACAA/UGUGGCCCCCAACACU	chr17_15251649_T_A	0.44
hsa-miR-3117-3p	AUAG/AGACUCAUAUAGUGCCAG	chr1_66628488_G_A	0.43
hsa-miR-3622a-5p	CAGGCACG/AGGAGCUCAGGUGAG	chr8_27701697_G_A	0.39
hsa-miR-4472	GGUG/CGGGGGUGUUGUUUU	chr8_142176399_G_C	0.34
hsa-miR-6823-3p	UGAGCCU/GCUCCUUCCCUCCAG	chr3_48549975_A_C	0.32
hsa-miR-938	UG/ACCCUUAAAGGUGAACCCAGU	chr10_29602331_C_T	0.3
hsa-miR-627-5p	GU/GGAGUCUCUAAGAAAAGAGGA	chr15_42199650_A_C	0.27
hsa-miR-6726-3p	CUC/UGCCCUGUCUCCCGCUAG	chr1_1296127_G_A	0.27
hsa-miR-7854-3p	UGAGGUGA/GCCGCAGAUGGGAA	chr16_81533949_A_G	0.25
hsa-miR-1268a	CGGGCG/AUGGUGGUGGGGG	chr15_22225320_C_T	0.19
hsa-miR-4433b-3p	CAGGAGUG/CGGGGGUGGGACGU	chr2_64340782_C_G	0.17
hsa-miR-548al	AACGGCAA/GUGACUUUUGUACCA	chr11_74399308_A_G	0.17
hsa-miR-6826-5p	UCAAU/CAGGAAAGAGGUGGGACCU	chr3_129272155_T_C	0.14
hsa-miR-4305	CCUAGA/GCACCUCCAGUUC	chr13_39664120_T_C	0.13
hsa-miR-3620-5p	GUGGGC/TUGGGCUGGGCUGGGCC	chr1_228097290_C_T	0.12
hsa-miR-4293	CAGCC/GUGACAGGAACAG	chr10_14383222_G_C	0.12
hsa-miR-5090	CCG/AGGGCAGAUUGGUGUAGGGUG	chr7_102465754_G_A	0.12
hsa-miR-10524-5p	CAGGAUGC/ACAGCAUAGU	chr6_78539301_G_T	0.1
hsa-miR-6863	UAGACG/AUGGUGAAGGAUUGAGUG	chr16_56904332_G_A	0.09
hsa-miR-6499-5p	UCGGGCG/ACAAGAGCACUGCAGU	chr5_151522138_C_T	0.07

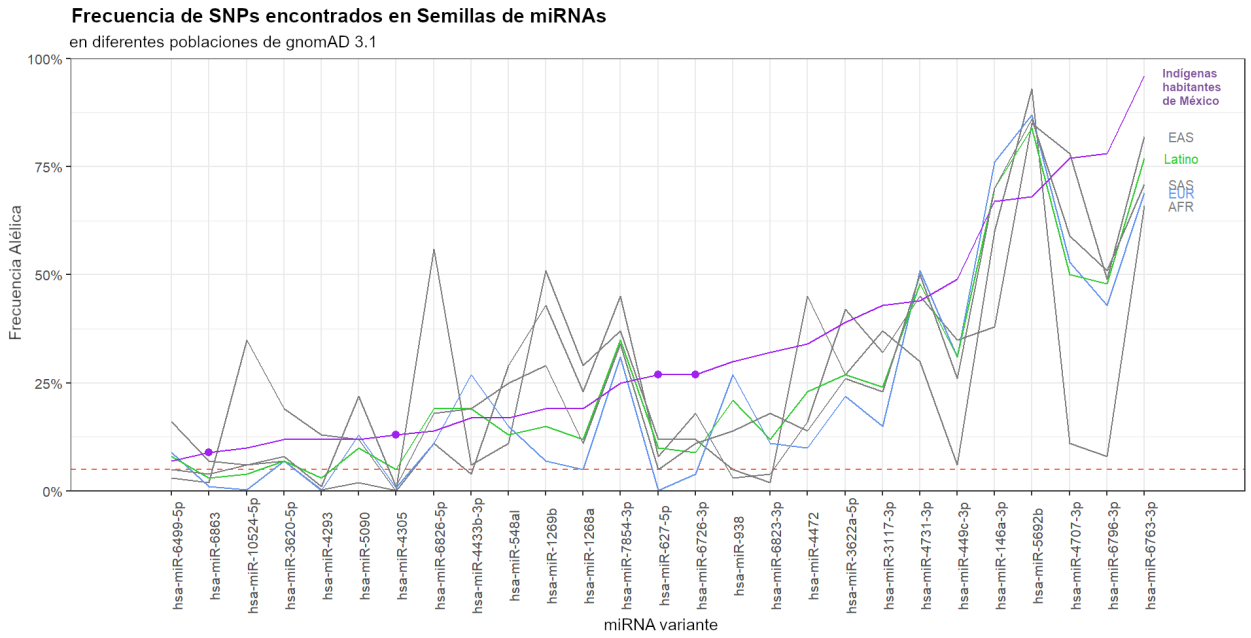


Figura 15. miRNAs variantes de la población indígena de México.
Se comparan las frecuencias alélicas para cada uno de los microRNAs variantes, contra otras poblaciones del mundo (con datos y etiquetas de gnomAD). Se destacan cuatro miRNAs (puntos morados) donde la población indígena que habita el territorio mexicano presenta la mayor frecuencia a nivel mundial mientras que en otras poblaciones la frecuencia es baja (y rara en la población etiquetada como europea).

A continuación se describe la información funcional encontrada en la literatura para los cuatro miRNAs de interés: 1) hsa-miR-4305 no se encuentra reconocido como un miRNA "real" en miRBase, debido a que no se ha recopilado suficiente evidencia de su existencia por métodos de secuenciación; sin embargo, se ha detectado en plasma circulante [58] [59] [60], también se encontró expresado diferencialmente en células dermales y se relaciona con la formación del folículo piloso [61], así como con un fenotipo de osteoporosis en pacientes de China [60]. La base de datos miRtarbase (versión 8.0 2021) reporta 52 blancos validados con evidencia molecular. La base de datos miRNASNP (-v3 2020) no reporta ningún enriquecimiento de vías causado por mutaciones conocidas en este miRNA. 2) hsa-miR-627-5p sí cuenta con suficiente evidencia en miRBase para considerarlo un miRNA "real". Se ha reportado como un regulador negativo de la proliferación de células madre derivadas de médula ósea [62] y se ha establecido como un blanco de la regulación por el receptor de vitamina D (VDR/NR1I1) [63]; también cuenta con evidencia molecular de que regula al mRNA CYP3A4 (una

monooxigenasa de la familia del citocromo P450, encargada de catalizar diversas reacciones en el metabolismo de fármacos y la síntesis de colesterol, esteroides y otros lípidos) [64], mir-627 se encontró desregulado en islotes pancreáticos expuestos a altas concentraciones de azúcar [65] y se sospecha ligeramente de su participación en las etapas 2 y 3 de cáncer colorrectal [66]. De manera interesante hay evidencia experimental para la pérdida y ganancia de targets en este miRNA cuando se insertan cambios en la región seed, aunque el cambio es diferente al que nosotros encontramos [67], esto demuestra que hsa-miR-627-5p pudiera tener cierta plasticidad de interacción determinada por la población de estudio; por último vale la pena señalar que miR-627-5p tiene un blanco validado con evidencia funcional fuerte, y 82 blancos validados con evidencia funcional débil (según miRtarbase). En miRNASNP se predice que la pérdida o ganancia de targets en este miRNA podrían alterar las siguientes vías: procesos metabólicos de glicerolípidos, glicerofosfolípidos, y fosfolípidos, biosíntesis de glicerofosfolípidos y fosfolípidos, actividad de receptores peptídicos acoplados a proteínas G y receptores peptídicos en general, vías de señalización JAK-STAT y MAPK [68]. 3) hsa-miR-6726-3p tampoco cuenta con evidencia suficiente en miRBase para reconocerlo como un miRNA "real", aunque se ha detectado en suero [69] y tiene 23 blancos validados con evidencia funcional débil (según miRtarbase). Se ha visto involucrado como parte de la respuesta al estrés oxidativo inducido por MPP+ en neuronas (aunque se desconoce la vía en la que participa) [70], y en desarrollo de enfermedad de Alzheimer [71]. En miRNASNP se predice que la pérdida o ganancia de targets en este miRNA podrían alterar las siguientes vías: actividad de correpresor transcripcional, sublocalización nuclear (nuclear speck), organización de la proyección celular y regulación del desarrollo de proyecciones neuronales, desarrollo dendrítico, regulación de la morfogénesis y morfogénesis relacionada a diferenciación neuronal, lamelipodios, densidad post sináptica, sinapsis asimétrica, vesículas recubiertas de clatrina, y especialización postsináptica [68]. 4) hsa-miR-6863 no cuenta con evidencia suficiente en miRBase para reconocerlo como un miRNA "real". Se ha propuesto que participa en el

desarrollo de enfermedad cardiovascular [72]. Según miRtarbase tiene 23 blancos validados con evidencia funcional débil. En miRNASNP se predice que la pérdida o ganancia de targets en este miRNA podrían alterar las siguientes vías: fosforilación y modificación de peptidilserina, morfogénesis celular en la diferenciación neuronal, morfogénesis dendrítica, desarrollo dendrítico, unión a GTPasa y Ras GTPasa, regulación de la transmisión sináptica glutamatérgica, membrana endosomal, y metabolismos de inositol fosfato [68].

Discusión

I. Sobre los estudios genómicos que involucran poblaciones indígenas de América

USA, Dinamarca y México han liderado la mayoría de los estudios genómicos realizados en poblaciones indígenas del continente americano (Figura 16). En USA y Dinamarca, se han centrado en el análisis de la diversidad humana (actual y ancestral), lo cual se ve reflejado en sus diversas publicaciones sobre diversidad genómica en diferentes regiones del mundo, no sólo en América; el liderazgo de estos países es algo hasta cierto punto esperado, puesto que se requiere de la experiencia para planear y llevar a buen término un proyecto poblacional. En México, en los últimos años, el gobierno y las universidades han invertido en infraestructura y capital humano para el desarrollo de la genómica [73]; la aparición de México como líder de varios proyectos es una muestra de que los grupos mexicanos que estudian la genómica poblacional se han establecido como referentes del campo.

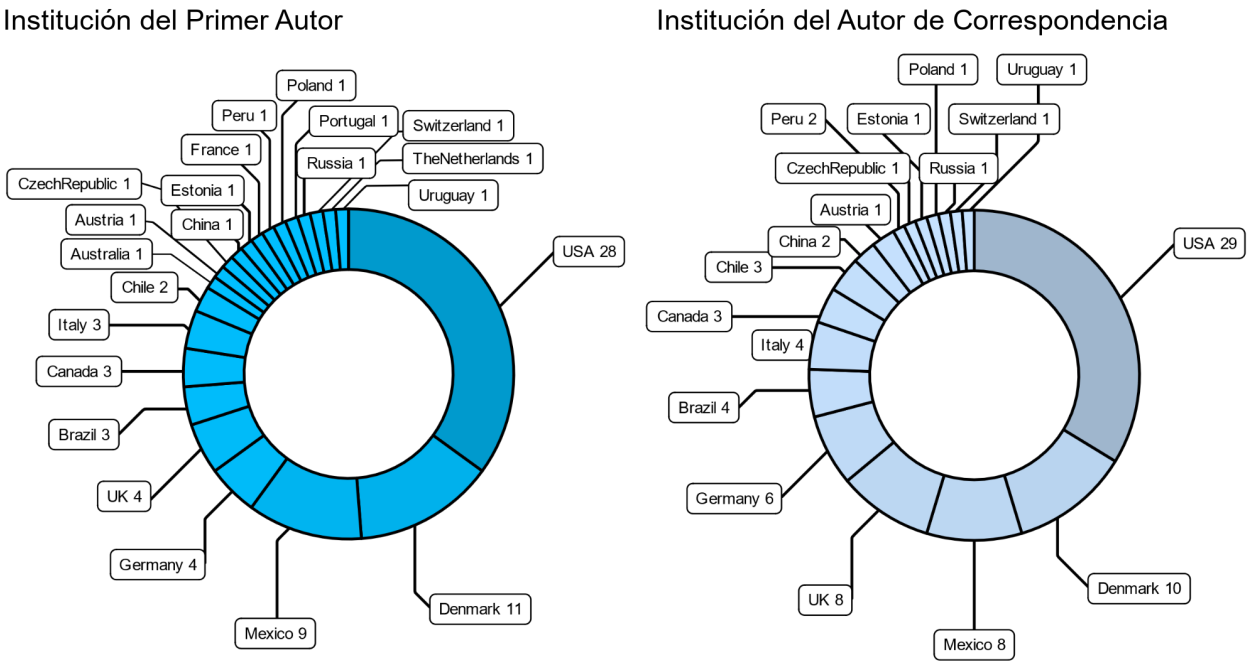


Figura 16. Los países involucrados en el estudio genómico de poblaciones indígenas en el continente americano. Existe una falta de representación de los países hogar de poblaciones indígenas como líderes de proyectos.

Si comparamos la lista de países donde habitan poblaciones indígenas de América (Figura 4) con los países líderes de los estudios (Figura 16) es evidente que varios de los países que registran a poblaciones indígenas como parte de sus censos están ausentes como representantes líderes de los proyectos de investigación de este tipo. Cabe mencionar que para que futuros proyectos genómicos sean exitosos deben considerar la voluntad de la clase política (puesto que la genómica aplicada suele tener alcances en políticas de salud a través de la promesa de medicina personalizada) y el liderazgo de instituciones de cada país [73,74]. El apoyo de las instituciones gubernamentales es crucial para adquirir financiamiento para el desarrollo de este tipo de investigaciones [75], pero los investigadores a cargo deben asegurarse de que los políticos relacionados tengan expectativas realistas de los beneficios generados por proyectos genómicos y el tiempo que toma alcanzarlos [74].

Los investigadores líderes también deben planear estrategias para fomentar el interés público en este tipo de estudios poblacionales, y lo que es aún más importante, planear cómo integrar a los grupos nativos en los proyectos [74]. La integración debe ser el eje inicial y central de los proyectos. Las inquietudes, intereses y expectativas de los pueblos indígenas deben ser el punto de partida de la planeación de este tipo de estudios. Si los países hogar en latinoamérica se vuelven líderes de proyectos, podrían beneficiarse de colaboraciones internacionales al unir esfuerzos con países desarrollados más experimentados en genómica poblacional, de quienes podrían recibir conocimiento, acceso a infraestructura, entrenamiento, y canales abiertos para comunicar rápidamente los hallazgos de la investigación [75].

Conclusiones sobre el estudio genómico de las poblaciones indígenas de América

El estudio genómico masivo de poblaciones indígenas que habitan el continente americano es un campo “joven” (con 13 años desde el primer trabajo en 2009 [25]) y en crecimiento

contínuo (Figura 3). Los grupos de investigación que abordan estos temas deben adaptarse rápidamente a los retos detrás de los aspectos éticos de los estudios genómicos. Los proyectos de investigación deben iniciar con preguntas acerca de cómo o por qué es importante realizar este tipo de estudios para las comunidades indígenas del continente [76].

¿Qué conocimiento aplicable puede obtenerse de este tipo de estudios? ¿Cómo se puede retribuir a las comunidades en una forma significativa? Las comunidades indígenas americanas en general cuentan con acceso muy limitado a servicios básicos de salud, y existe un riesgo latente de que tampoco puedan beneficiarse directamente de servicios médicos más avanzados como los que se derivan de estudios genómicos [30]. Es importante que las comunidades indígenas se involucren activamente en las conversaciones y los procesos de toma de decisión en los proyectos genómicos [77]. Aún existen cientos de poblaciones indígenas no representados en el acervo genómico global [14]. Si en el futuro los proyectos de genómica poblacional se enfocaran en estas poblaciones, quizá sería mayor el impacto potencial del conocimiento genómico adquirido.

Desafortunadamente existe poca apertura para compartir los datos de los estudios genómicos que involucran poblaciones indígenas habitantes del continente americano. Esto es comprensible desde un punto de vista ético, dado que varios de los grupos indígenas en América se enfrentan aún a la marginalización en sus países hogar. Los futuros proyectos deben de afrontar cuidadosamente los riesgos de recolectar, digitalizar, manejar y publicar datos genéticos de poblaciones indígenas. Se debe tener como prioridad el evitar errores como la falta de consentimiento informado para el uso secundario de los datos genéticos (como fue el caso de la tribu Havasupai en Arizona, USA [78], o el caso de la tribu Nuu-chah-nulth en Columbia Británica, Canadá [79]), o las incorrectas asociaciones de fenotipos debatibles (como el caso de la representación negativa del "gen guerrero" en la prensa científica respecto al pueblo Maori de Aotearoa [80]).

Futuros proyectos pueden tomar en cuenta la posibilidad de anonimizar los datos genómicos, y el uso de repositorios públicos seguros y bien establecidos con lo son SRA (https://www.ncbi.nlm.nih.gov/sra), ENA (https://www.ebi.ac.uk/ena/browser/) y EGA (https://ega-archive.org/). Estas son excelentes opciones para compartir de manera segura los datos genómicos. Es de suma importancia intentar resolver el problema para compartir datos dado que hay mucho por ganar cuando se logre integrar los esfuerzos detrás de todos los proyectos previamente realizados de manera aislada (Anexo 2). Tal como lo mostró recientemente Jiménez-Kaufmann [81], los datos generados por proyectos independientes pueden reunirse para generar una base de datos más completa.

El estudio de revisión que publicamos como parte de esta tesis doctoral (Anexo 2) se enfocó en la bibliometría y la informática incluida en los proyectos de genómica en donde participan individuos indígenas habitantes del continente americano publicados a la fecha; pero no revisamos a detalle los consentimientos informados detrás de cada proyecto. Valdría la pena analizar a profundidad las razones por las que los datos no son públicos para poder contribuir a los esfuerzos colaborativos para liberar e integrar datos genómicos, sin comprometer la seguridad y consentimiento de los participantes.

Se necesita más participación por parte de los países latinoamericanos en los estudios genómicos de sus poblaciones, preferiblemente como líderes institucionales en donde que involucren a las comunidades indígenas desde el principio y de manera activa.

Para ayudar al desarrollo de futuros proyectos genómicos que busquen cerrar la brecha de las poblaciones indígenas del continente americano no representadas, recopilamos y publicamos una lista de las poblaciones faltantes en el Anexo 2.

II. El aporte del proyecto 100G-MX al estudio de la genómica en America

Variantes de interés biomédico en el catálogo 100G-MX

El proyecto 100G-MX [13] es a la fecha la recopilación de variación genómica más grande en poblaciones indígenas habitantes de México, complementando proyectos previos de secuenciación de genoma completo [33] y exoma [41]. Entre estos tres proyectos, se han secuenciado individuos pertenecientes a 31 de los 68 poblaciones indígenas en México (tsa ju jmí, chocholteco, Yoko t'aano, teenek, wixárika, kanjobal, kaqchiquel, winaq qo, matlalzinca, maya, yoreme, jñatrjo, ha shuta enima, ayuuk ja'ay, ñuu savi, mochós, nahua, ñätho, xi'oi, nuntajɨyi', comcaa'c, tacuate, rarámuri, tepehuan, tojolabal, totonaca, yuvii chianj, batsil winik'otik, zapoteca, zapoteco, o'de püt). Se necesitan más proyectos de secuenciación que permitan catalogar la variación en los pueblos faltantes, e incrementar el número de individuos en los grupos ya estudiados por genoma o exoma completo, con la finalidad de reducir la brecha genómica.

El catálogo de variación nos permitió encontrar variantes reportadas en bases de datos genotipo-fenotipo relacionados a la salud, o localizadas en genes relacionados con fenotipos de interés. Particularmente el caso de la variación en la región reguladora de PPARG: reportamos 10 variantes en 3 enhancers a lo largo de 20 kilobases en el cromosoma 3, que son interesantes desde dos puntos de vista; el primero es el patrón de frecuencias alélicas a nivel mundial, dado que la baja frecuencia en la población etiquetada como representativa de asia del este (la menor de cualquier continente) sugiere que el haplotipo encontrado en alta frecuencia en la población indígena de México proviene de un evento migratorio, posiblemente uno relacionado con los grupos que habitaron beringia y que iniciaron el poblamiento del continente Americano. El segundo punto de vista es el posible impacto funcional de esta variación. PPARG, o PPARγ, es un factor de transcripción, y se sabe que los

sitios principales donde se expresa en el humano son el tejido adiposo subcutáneo, el epiplón, y el tejido mamario (https://www.gtexportal.org/home/). Se ha reportado que los niveles de expresión, y la actividad de PPAR□ se reducen cuando se padece obesidad, y se sospecha de cambios epigenéticos [56]. De manera interesante para este trabajo, previamente la variante rs1801282 (el cambio de C > G en la posición 12,351,626 del cromosoma 3) se ha relacionado a la resistencia a insulina en niños mexicanos [57]; este cambio afecta la región codificante del gen, provocando el cambio de una Prolina por una Alanina en el aminoácido 12 (Pro12Ala, un cambio asociado con una disminución del 30–50% en la actividad inducida por ligando). Según otros reportes, los niños mexicanos portadores de esta variante mostraron un riesgo reducido de desarrollar resistencia a la insulina cuando los niveles de colesterol LDL y colesterol total son altos. La frecuencia alélica de este cambio en la población etiquetada como Latina en gnomAD3.1 es 8.4%, y en el dataset 100G-MX es 19% (mayor incluso que la población con mayor frecuencia en gnomAD3.1, que es la etiquetada como Europea Finlandesa con 15.6%). Es de esperarse que el patrón de variación que vemos en el posible haplotipo regulador se extienda más allá de la región enhancer de PPARG.

Existen 500 variantes conocidas en el gen PPARG (gnomAD v3.1.2), la mayoría de ellas raras (solo 6 se encuentran en al menos el 1% de la población mundial); por otro lado, en el dataset 100G-MX encontramos 239 SNPs con una frecuencia alélica mayor al 5%, de los cuales 126 SNPs presentaron incluso una frecuencia alélica mayor a 50%. En general PPARG parece ser un gen con altas frecuencias en población indígena habitante de México, lo cual nos plantea las siguientes preguntas a futuro: ¿Este patrón de frecuencias alélicas se replica en otras poblaciones indígenas del continente? —una pregunta que se podría responder revisando otros datasets públicos—, y ¿si los niveles de expresión de PPARG se ven afectados por la presencia de estas variantes? Esta última pregunta se podría responder si encontramos datos de expresión en los tejidos adiposo subcutáneo, el epiplón, o el tejido mamario,

donde se pueda buscar la(s) variante(s) en los transcritos y cuantificar el nivel de expresión.

En conjunto, el catálogo de variación nos permitió explorar genotipos relacionados con fenotipos en salud y la estructura genética de la muestra de estudio, mostrando una posible distinción regional en un eje Norte-Centro-Sur. Además, identificamos que los individuos de la población comcaa'c presentan un panorama de genoma completo muy distintivo, lo cual puede ser un reflejo de una población ancestral pequeña o aislada, en concordancia con información antropológica [82] y estudios genómicos previos [34].

Nuestro reporte de la estructura genética (Anexo 1 [13]) tiene la siguiente limitación: algunos de las poblaciones indígenas de México sondeadas incluyeron más individuos que otros (p. ej. sólo pudimos incluir siete individuos del norte del país). Para poder superar esta limitación en el futuro se puede incrementar el número de individuos estudiados integrando todos los datos de genomas indígenas mexicanos disponibles, o generando nuevas secuencias de las poblaciones no exploradas. Un incremento en los números de individuos estudiados permitirá definir patrones demográficos con más detalle.

La base de datos generada en el proyecto 100G-MX es una fuente de información para proyectos futuros de medicina personalizada en México. Además de contribuir a la inclusión indígena americana en genómica. El proyecto se realizó por investigadores mexicanos, en instalaciones mexicanas. Hubo algunos retrasos y diversos retos, pero se logró realizar el análisis descriptivo de variantes en genoma completo, y se generó capital humano especializado para estudios poblacionales (Anexo 4 - Tesis Co-Dirigidas).

Conclusiones

Analizamos 76 genomas completos de personas pertenecientes de 27 poblaciones indígenas, y definimos un conjunto de variantes presentes en la población mexicana. Reportamos y catalogamos 9,737,152 variantes, de las cuales 44,118 son variantes nuevas. En cuestión de demografía todos nuestros análisis describen a la población indígena mexicana como subgrupos en un gradiente Norte-Centro-Sur. En cuanto a variantes de interés biomédico reportamos un promedio de 12,871 variantes por individuo, relacionadas a fenotipos de interés clínico según las bases de datos GWAS catalog, ClinVar o PharmGKB. Reportamos también un posible haplotipo regulador de la expresión de PPARG, un gen previamente implicado en problemas de salud en México como obesidad, aterosclerosis y resistencia a insulina. También reportamos 27 variantes que alteran la región semilla de un microRNA maduro, será interesante en el futuro evaluar si dichos cambios pudieran estar alterando la regulación transcripcional. Finalmente, el estudio y catálogo de la variación en estos individuos mexicanos contribuye también al conocimiento de la genómica aplicable a la población general, ya que un porcentaje variable pero significativo del genoma de la población actual del país tiene un componente indígena; conocer esta parte del genoma es fundamental para el desarrollo eficaz de la medicina genómica.

Perspectivas

Hay tres ejes de interés a partir de los resultados presentados en este trabajo. En conjunto se basan en análisis bioinformáticos de datos públicos. El primero consiste en realizar el análisis demográfico y descripción de la variación en las poblaciones indígenas del continente americano con datos públicamente disponibles. Gracias a la revisión bibliográfica hecha durante este trabajo de doctorado (Anexo 2) ahora sabemos cuántos son esos genomas y su disponibilidad. Cabe mencionar que aún con los

datos públicamente disponibles, en México aún faltan por incluir varias poblaciones indígenas; por lo tanto, existe la perspectiva de concretar nuevos proyectos de secuenciación con tecnología de genoma completo.

Por otro lado, la variación en elementos reguladores en el genoma sigue siendo un área poco explorada a nivel poblacional. Nuestro hallazgo del posible haplotipo regulador en PPARG debe ser validado en otros datos de poblaciones indígenas habitantes del continente americano. Muy probablemente se encuentre distribuido a lo largo del continente. Dada la relación de este gen con fenotipos de interés en salud, también vale la pena realizar la validación experimental del efecto de insertar las variantes en un contexto celular.

También consideramos que vale la pena probar funcionalmente el impacto de los 4 microRNAs con variación particularmente indígena habitante de México. La posibilidad de un miRNoma poblacional (cambios en la regulación de la red microRNA - mensajeros) es uno de los temas menos explorados en proyectos poblacionales. Sobre este último punto, sería interesante realizar el análisis de datos públicos single-cell RNA-seq en busca del comportamiento de los pares microRNA:mRNA con y sin la presencia de las variantes encontradas en este trabajo.

Finalmente sabemos que el 99.6% de las variantes novel detectadas en este proyecto no se están incluidas en los SNP-array usados comúnmente y que el 99.5% no están en desequilibrio de ligamiento con otras variantes conocidas en la población Mexicana [13]. Este conjunto novel debería ser incluido en el diseño de SNP-arrays futuros enfocados en explorar asociaciones genotipo-fenotipo en el continente americano.

Material y Métodos

Origen y manejo de las muestras de ADN

Este trabajo se llevó a cabo de conformidad con la Declaración de Helsinki, y fue aprobado por el comité de investigación, ética y bioseguridad del Instituto Nacional de Medicina Genómica (INMEGEN) en la ciudad de México (número de protocolo 31/2011/I). Se llevó a cabo con el apoyo de la Comisión Nacional para el Desarrollo de Pueblos Indígenas y en acuerdo con los líderes de las poblaciones estudiadas. Cada uno de los participantes accedió al consentimiento informado con firma autógrafa. En casos necesarios, el consentimiento informado se tradujo a la lengua nativa del participante, y algunos de ellos firmaron con su huella digital. Los 76 individuos estudiados pertenecen a 27 grupos étnicos de México. Estos individuos forman parte de la cohorte "Metabolic Analysis in an Indigenous Sample (MAIS)" integrada entre 2012 y 2018 [59].

El ADN se extrajo a partir de sangre periférica y se llevó a cabo la secuenciación de genoma completo con una librería paired-end, con fragmentos de 150 pares de bases; la secuenciación se contrató como servicio en el Beijing Genomics Institute, con un instrumento Illumina HiSeq X Ten. Se realizó control de calidad general para evaluar la calidad del servicio de secuenciación (Anexo 5).

Preprocesamiento y mapeo de lecturas NGS

Se evaluó la calidad de los datos fastq pareados con el software FastQC v0.11.4. Se usó Trimmomatic v0.39 para eliminar adaptadores y mantener secuencias con una longitud mínima de 70 pares de bases y calidad promedio de phred 28 o mayor. Se utilizaron los siguientes parámetros para Trimmomatic: LEADING:28 SLIDINGWINDOW:5:28 MINLEN:70. Las secuencias después de procesadas con Trimmomatic se mapearon contra

el genoma de referencia GRCh38 (descargado el 25/04/2017) del repositorio GATK (https://software.broadinstitute.org/gatk/download/bundle). El genoma de referencia se indexo con SNAP v1.0beta.18 utilizando una semilla de "20" y el parámetro "-exact". Para cada muestra, se utilizó SNAP aligner v1.0beta.18. Las lecturas duplicadas fueron eliminadas con Sambamba v0.6.6.

Llamado de Variantes

Los archivos BAM de cada muestra se procesaron siguiendo el protocolo para GATK 3.8 que se describe a continuación: 1) IndelRealigner; 2) BaseRecalibrator; 3) HaplotypeCaller. Lo anterior resultó en un archivo gVCF para cada muestra. Se juntaron todos los gVCF utilizando la herramienta GenotypeGVCFs seguido de recalibración de genotipos (VQSLOD) con la herramienta VQSR. Las variantes con VQSLOD > = 99.0 se marcaron como "PASS" para continuar con el análisis.

Anotación de Variantes

Acceso al pipeline en: https://www.protocols.io/view/vepextended-dm6gpr42jvzp/v1

Las variantes se anotaron con información clínicamente relevante usando un pipeline que integra bases de datos custom en Variant Effect Predictor. El pipeline completo se puede descargar de la ruta indicada en el encabezado de esta sección. En breve, la anotación de rsIDs se realizó con bcftools utilizando dbSNP b152 (fileDate = 20181015). Las bases de datos clínicas fueron: 1) GeneHancer (Enero 2019, V2) desde UCSC Table Browser, 2) PharmGKB (var_drug_ann.tsv descargado en 06/14/2019), 3) GWAS catalog (v1.0.2), 4) ClinVar (clinvar_20190403), 5) miRBase (release 22.1), 6) y gnomAD 2.1.1. De ser necesario, las bases de datos se convirtieron a formato VCF con las coordenadas de dbSNP b152 para GRCh38.

Catálogo de Variantes

Acceso al pipeline en: https://www.protocols.io/view/nf-vcf-cataloguer-ewov18ewkgr2/v1

Las variantes se catalogaron utilizando el pipeline señalado en el renglón anterior. En breve, separamos las variantes en raras (AF < = 1%), baja frecuencia (1 > AF < = 5%), y comunes (AF > 5%); cada una de estas categorías se separó en variantes de región codificante, variantes en promotores, y variantes con registros de posible asociación clínica según ClinVar, GWAS catalog o pharmGKB.

Análisis de Componentes Principales, k-means, y ADMIXTURE

Acceso al pipeline en: dx.doi.org/10.17504/protocols.io.bkwbkxan

El pipeline que realiza los análisis de agrupamiento se encuentra disponible para su descarga en el link del renglón anterior. En breve, a partir del dataset IPVS se mantuvieron los individuos nativos mexicanos, y nativos peruanos (ids del proyecto mil genomas: HG01926, HG01938, HG01961, HG02272), y se realizó un filtro de variantes utilizando bcftools v1.9-220-gc65ba41 con los siguientes parámetros MAF > 0.05 y r2 > 0.85 con el plugin +prune usando los parámetros --window 2000bp --nsites-per-win 1. Se transformaron los VCF a formato Eigenstrat, y el cálculo de PCA se realizó con Smarpca de Eigensoft v6.1.4. Utilizando las coordenadas de PCA realizamos el análisis de clustering por k-means, y utilizamos el método de Average Silhouette para definir el agrupamiento óptimo. El análisis de ADMIXTURE v1.3 mostrado se realizó utilizando el parámetro --seed 43 (se realizaron 100 replicados del análisis con diferentes --seed y se exploró la concordancia por pares de grupos. Para K=3, K=4 y K=5 se observan agrupamientos consistentes con la regionalización Norte-Centro-Sur).

Análisis de F_{ST}

A partir del dataset IPVS se mantuvieron los individuos nativos mexicanos, y nativos peruanos (ids del proyecto mil genomas: HG01926, HG01938, HG01961, HG02272), y se realizó un filtro de variantes utilizando bcftools v1.9-220-gc65ba41 con los siguientes parámetros MAF > 0.05 y r2 > 0.85 con el plugin +prune usando los parámetros --window 2000bp --nsites-per-win 1. Se calcularon valores de F_{ST} por pares utilizando Smartpca de Eigensoft v6.1.4, y los valores calculados se convirtieron en formato de matriz para construir un Neighbor-joining tree con el software MEGA X.

Detección de SNPs en el microRNAs

Se utilizó la base de datos miRbase v22.1 para obtener las coordenadas genómicas de los microRNAs en la versión GRCh38 del genoma humano (https://www.mirbase.org/) . Se utilizó la base de datos GRCh38 del UCSC genome browser para obtener las coordenadas genómicas de las regiones 3'UTR (https://genome.ucsc.edu/cgi-bin/hgTables). Como datos de variación genómica se utilizó el dataset JPVS del proyecto 100G-MX [13]. La identificación de variantes en regiones miRNA y 3'UTR se llevó a cabo de manera automatizada utilizando el pipeline creado por Eduardo García como parte de su tesis de licenciatura (Anexo 4), con los parámetros default. El repositorio de código se encuentra en: https://github.com/Ed-G655/nf-compare-miRNome. El manejo de datos y figuras para la sección "Resultados: Variación en el miRNoma de las poblaciones nativas mexicanas" y "Discusión: Definición de un miRNoma Nativo Mexicano" se realizó en el software R.

Anexo 1 - Artículo de Investigación

Aguilar-Ordonez,I., Perez-Villatoro,F., Garcia-Ortiz,H., Barajas-Olmos,F., Ballesteros-Villascan,J., Gonzalez-Buenfil,R., Fresno,C., Garciarrubio,A., Fernandez-Lopez,J.C., Tovar,H., Hernandez-Lemus,E., Orozco,L., Soberon,X., Morett,E. (2021). **Whole genome variation in 27 Mexican indigenous populations, demographic and biomedical insights.** PLoS ONE, 16 (4), e0249773. https://doi.org/10.1371/journal.pone.0249773

PLOS ONE

RESEARCH ARTICLE

Whole genome variation in 27 Mexican indigenous populations, demographic and biomedical insights

Israel Aguilar-Ordoñez[1,2☯], Fernando Pérez-Villatoro[1,2,3☯], Humberto García-Ortiz[2], Francisco Barajas-Olmos[2], Judith Ballesteros-Villascán[4], Ram González-Buenfil[4], Cristobal Fresno[2], Alejandro Garcíarrubio[1], Juan Carlos Fernández-López[2], Hugo Tovar[2], Enrique Hernández-Lemus[2], Lorena Orozco[2], Xavier Soberón[1,2], Enrique Morett[1]*

1 Instituto de Biotecnología, Universidad Nacional Autónoma de México (UNAM), Cuernavaca, Morelos, México, **2** Instituto Nacional de Medicina Genómica (INMEGEN), Mexico City, México, **3** Winter Genomics, Mexico City, México, **4** Benemérita Universidad Autónoma de Puebla (BUAP), Puebla de Zaragoza, Puebla, México

☯ These authors contributed equally to this work.
* emorett@ibt.unam.mx

OPEN ACCESS

Citation: Aguilar-Ordoñez I, Pérez-Villatoro F, García-Ortiz H, Barajas-Olmos F, Ballesteros-Villascán J, González-Buenfil R, et al. (2021) Whole genome variation in 27 Mexican indigenous populations, demographic and biomedical insights. PLoS ONE 16(4): e0249773. https://doi.org/10.1371/journal.pone.0249773

Editor: Dana C. Crawford, Case Western Reserve University, UNITED STATES

Received: July 22, 2020

Accepted: March 24, 2021

Published: April 8, 2021

Data Availability Statement: The JPVS, and subsets reported in Table 1 are available as supporting information. The datasets included are in VCF format, and cover every variant found, with the corresponding allele frequency in our studied population; it also includes spreadsheet for subsets of interest for easy exploration. In alignment with the respective Institutional Review Board approval and individual informed consents, any other BAM or VCF file will be available upon request to avoid compromising the participant's privacy. Please

Abstract

There has been limited study of Native American whole genome diversity to date, which impairs effective implementation of personalized medicine and a detailed description of its demographic history. Here we report high coverage whole genome sequencing of 76 unrelated individuals, from 27 indigenous groups across Mexico, with more than 97% average Native American ancestry. On average, each individual has 3.26 million Single Nucleotide Variants and short indels, that together comprise a catalog of 9,737,152 variants, 44,118 of which are novel. We report 497 common Single Nucleotide Variants (with allele frequency > 5%) mapped to drug responses and 316,577 in enhancer or promoter elements; interestingly we found some of these enhancer variants in PPARG, a nuclear receptor involved in highly prevalent health problems in Mexican population, such as obesity, diabetes, and insulin resistance. By detecting signals of positive selection we report 24 enriched key pathways under selection, most of them related to immune mechanisms. No missense variants in ACE2, the receptor responsible for the entry of the SARS CoV-2 virus, were found in any individual. Population genomics and phylogenetic analyses demonstrated stratification in a Northern-Central-Southern axis, with major substructure in the Central region. The Seri, a northern group with the most genetic divergence in our study, showed a distinctive genomic context with the most novel variants, and the most population specific genotypes. Genome-wide analysis showed that the average haplotype blocks are longer in Native Mexicans than in other world populations. With this dataset we describe previously undetected population level variation in Native Mexicans, helping to reduce the gap in genomic data representation of such groups.

Anexo 2 - Artículo Review

Israel Aguilar-Ordonez, Josué Guzmán-Linares, Judith Ballesteros-Villascán, Fernanda Mirón-Toruño, Alejandra Pérez-González, José García-López, Fabricio Cruz-López, and Enrique Morett. "A Tale of Native American Whole-Genome Sequencing and Other Technologies." *Diversity* 14, no. 8 (2022): 647. https://doi.org/10.3390/d14080647

diversity

MDPI

Review

A Tale of Native American Whole-Genome Sequencing and Other Technologies

Israel Aguilar-Ordoñez [1,2,*], Josué Guzmán-Linares [3], Judith Ballesteros-Villascán [4], Fernanda Mirón-Toruño [3], Alejandra Pérez-González [3], José García-López [3], Fabricio Cruz-López [3] and Enrique Morett [1,*]

1 Instituto de Biotecnología, Universidad Nacional Autónoma de México (UNAM), Cuernavaca 62210, Mexico
2 Instituto Nacional de Medicina Genómica (INMEGEN), Mexico City 14610, Mexico
3 School of Biotechnology Puebla de Zaragoza, Benemérita Universidad Autónoma de Puebla (BUAP), Puebla 72000, Mexico
4 Laboratorio Nacional de Genómica para la Biodiversidad (LANGEBIO), Centro de Investigación Y de Estudios Avanzados del IPN, Irapuato 36824, Mexico
* Correspondence: iaguilar@inmegen.gob.mx (I.A.-O.); enrique.morett@ibt.unam.mx (E.M.)

Abstract: Indigenous people from the American continent, or Native Americans, are underrepresented in the collective genomic knowledge. A minimal percentage of individuals in international databases belong to these important minority groups. Yet, the study of native American genomics is a growing field. In this work, we reviewed 56 scientific publications where ancient or contemporary DNA of Native Americans across the continent was studied by array, whole-exome, or whole-genome technologies. In total, 13,706 native Americans have been studied with genomic technologies, of which 1292 provided whole genome samples. Data availability is lacking, with barely 3.6% of the contemporary samples clearly accessible for further studies; in striking contrast, 96.3% of the ancient samples are publicly available. We compiled census data on the home countries and found that 607 indigenous groups are still missing representation in genomic datasets. By analyzing authorship of the published works, we found that there is a need for more involvement of the home countries as leads in indigenous genomic studies. We provide this review to aid in the design of future studies that aim to reduce the missing diversity of indigenous Americans.

Keywords: Native American; whole-genome sequencing; population genomics; data availability

Citation: Aguilar-Ordoñez, I.; Guzmán-Linares, J.; Ballesteros-Villascán, J.; Mirón-Toruño, F.; Pérez-González, A.; García-López, J.; Cruz-López, F.; Morett, E. A Tale of Native American Whole-Genome Sequencing and Other Technologies. *Diversity* **2022**, *14*, 647. https://doi.org/10.3390/d14080647

Academic Editor: Michael Wink

Received: 25 March 2022
Accepted: 14 May 2022
Published: 12 August 2022

1. Introduction

In 2021, humanity celebrated 20 years since the publication of the first draft of the human genome [1], a document that gave rise to an era of self-exploration for our species at a molecular level. That human genome draft, thoroughly refined up to its most recent version [2], is still being used as the reference against which we compare genomic sequences from other individuals to detect genomic variations. However, one single genomic reference will not be enough to understand the whole human species; therefore, the complete sequences of a large number of individuals from different ancestries are required to better represent the human genomic variability [3].

The promise of genomic medicine, understood as the ability to provide personalized diagnosis and treatment for each patient, is based on an exhaustive study of human genomic variation. Genomic technologies such as array genotyping, exome sequencing, and whole-genome sequencing have allowed the study of this variation at a global scale [4,5], mainly by finding and cataloging single-nucleotide polymorphisms (SNPs) and more complex structural variations (SVs) [6] on DNA. However, most of the studies have focused on describing population genomics in high-income countries, leaving a gap in the potential understanding of the genomics underlying health and disease processes in the rest of the world. The underrepresentation of non-European populations in genomic science has been well documented and is a current topic of discussion [7–10]. Less than 1% of individuals

Anexo 3 - Artículo de Divulgación

Aguilar,I., Guzman,J., Miron,F., **Morett,E.** (2021). **Conociendo el Genoma indígena Mexicano: el proyecto 100G-MX. Biotecnologia en Movimiento.Revista de divulgación del Instituto de Biotecnología de la UNAM,** 26, 8-15. https://biotecmov.files.wordpress.com/2021/09/btmv26-02_genomaindgmx-vf.pdf

Conociendo el Genoma Indígena Mexicano: el proyecto 100G-MX

Israel Aguilar, Josué Guzmán, Fernanda Mirón y Enrique Morett

Nuestro genoma contiene toda la información biológica para definirnos individualmente y además, es un gran libro —toda una biblioteca— abriéndose ampliamente, sobre la historia evolutiva de la humanidad; en él están las mejores pistas de todo el proceso de cambios que nos dio origen e identidad como especie. Los humanos hemos evolucionado a partir de un ancestro común que vivió en África hace no más de 300 mil años, desde donde eventualmente varios grupos migraron hacia el resto del mundo alrededor de 70 mil años antes de nuestra era (a.n.e.). El ser humano actual es la única especie viva del género *Homo*, aunque hace unos cuantos miles de años había en el planeta otros miembros del mismo género biológico (el hombre de Neanderthal, el Denisova, el Floresiensis y quizá, algún otro todavía no descubierto). Aunque estos parientes lejanos gradual o eventualmente se extinguieron, aún persiste su huella hereditaria, guardada en segmentos o secuencias discretas de nuestro ADN. Ahora es posible acceder a grandes cantidades de información sobre las relaciones de parentesco amplio, entre grupos e individuos de todo el mundo —y de algunos fósiles— a través del análisis del genoma.

Anexo 4 – Tesis Co-Dirigidas

BENEMÉRITA UNIVERSIDAD AUTÓNOMA DE PUEBLA

FACULTAD DE CIENCIAS BIOLÓGICAS

"DESARROLLO DE UNA HERRAMIENTA BIOINFORMÁTICA PARA EL ANÁLISIS POBLACIONAL DE LA VARIACIÓN GENÓMICA"

T E S I S

PRESENTADA PARA OBTENER EL GRADO DE:
LICENCIADA EN BIOTECNOLOGÍA

PRESENTA:
JUDITH BALLESTEROS VILLASCÁN

DIRECTORA DE TESIS: DRA. LILIANA LÓPEZ PLIEGO
CODIRECTOR DE TESIS: M.C. ISRAEL AGUILAR ORDOÑEZ

Este trabajo fue realizado en el Laboratorio del Dr. Juan Enrique Morett Sánchez del Instituto de Biotecnología de la UNAM

PUEBLA, PUE. ENERO 2020

BENEMÉRITA UNIVERSIDAD AUTÓNOMA DE PUEBLA

FACULTAD DE CIENCIAS BIOLÓGICAS

"INTEGRACIÓN DE UNA BASE DE DATOS GENÓMICOS DE NATIVOS AMERICANOS A PARTIR DE DATOS PÚBLICOS"

T E S I S

que para obtener el título de

LICENCIADO EN BIOTECNOLOGÍA

PRESENTA:

JOSUÉ GUZMÁN LINARES

DIRECTOR DE TESIS: M.C. Israel Aguilar Ordóñez
CO-DIRECTOR DE TESIS: Dr. Carlos Alberto Contreras Paredes

ASESORES DE TESIS:
M.C. Elda Carreón Moreno
Dr. Miguel Castañeda Lucio

PUEBLA, PUE. MAYO, 2022

BENEMÉRITA UNIVERSIDAD AUTÓNOMA DE PUEBLA

FACULTAD DE CIENCIAS BIOLÓGICAS

"ANÁLISIS DE LA VARIACIÓN GENÉTICA EN REGIONES ENHANCER DE GENOMAS DE POBLACIÓN NATIVO MEXICANA"

TESIS

PRESENTADA PARA OBTENER EL GRADO DE
LICENCIADO EN BIOTECNOLOGÍA

PRESENTA:
RAM GONZÁLEZ BUENFIL

DIRECTORA DE TESIS: DRA. LILIANA LÓPEZ PLIEGO
CODIRECTOR DE TESIS: M.C. ISRAEL AGUILAR ORDÓÑEZ

ASESORES DE TESIS:
DR. CARLOS ALBERTO CONTRERAS PAREDES
M.C. ELDA CARREÓN MORENO

PUEBLA, PUE. AGOSTO, 2020

BENEMÉRITA UNIVERSIDAD AUTÓNOMA DE PUEBLA

FACULTAD DE CIENCIAS BIOLOGÍA

"DETECCIÓN DE SEÑALES DE SELECCIÓN EN GENOMAS COMPLETOS DE NATIVOS AMERICANOS"

T E S I S

PRESENTADA PARA OBTENER EL TÍTULO DE:

LICENCIADO (A) EN BIOTECNOLOGÍA

PRESENTA:
MARÍA FERNANDA MIRÓN TORUÑO

DIRECTOR: DR. ISRAEL AGUILAR ORDOÑEZ
CO-DIRECTORA: DRA WENDY ARGELIA GARCÍA SUASTEGUI

MAYO 2022

Benemérita Universidad Autónoma de Puebla
Facultad de Ciencias Biológicas

Predicciones estructurales de las variantes particulares no-sinónimas en el genoma de la población Comcáac (Seri)

TESIS

PRESENTADA PARA OBTENER EL GRADO DE
LICENCIADA EN BIOTECNOLOGÍA

PRESENTA
Alejandra Paulina Pérez González

OCTUBRE 2022

DIRECTOR DE TESIS: MC. Israel Aguilar Ordóñez
CO-DIRECTORA: Dra. Norma Angélica Caballero Concha
SINODALES
Dra. Wendy Argelia García Suastegui
MSc. Ana Isabel Castillo Orozco

Este trabajo se llevó a cabo en el laboratorio del Dr. Enrique Morett en el Instituto de Biotecnología de la Universidad Nacional Autónoma de México

Anexo 5 - Difusión en medios de comunicación

Egresados BUAP estudian genomas de grupos indígenas	https://www.eluniversalpuebla.com.mx/universidades/egresados-buap-estudian-genomas-de-grupos-indigenas
Facultad de Ciencias Biológicas BUAP, Biotecnología, herramientas, técnicas y aplicaciones	https://fb.watch/cyUQ2npyZQ/
Contribuyen egresados de la BUAP a creación del catálogo de variación genómica de grupos indígenas de México	https://tribunanoticias.mx/483654-2contribuyen-egresados-de-la-buap-a-creacion-del-catalogo-de-variacion-genomica-de-grupos-indigenas-de-mexico/
De eso se Trata BUAP, con Ricardo Cartas	https://www.facebook.com/watch/live/?ref=watch_permalink&v=1499787496899170
Crean egresados BUAP gran catálogo de variación genómica de grupos indígenas de México	https://www.elsoldepuebla.com.mx/local/crean-egresados-buap-gran-catalogo-de-variacion-genomica-de-grupos-indigenas-de-mexico-6700875.html

Egresados de la BUAP demuestran vocación por la ciencia	https://pueblaonline.com.mx/2019/portal/movil/index.php/estado/item/111120-egresados-de-la-buap-demuestran-vocacion-por-la-ciencia#.YmMtK9rMLIU
Colabora UAP en catálogo de variación genómica de los grupos indígenas	https://www.milenio.com/politica/comunidad/buap-contribuye-creacion-catalogo-grupos-nativos-mexicanos
Egresados BUAP contribuyen a la creación de catálogo científico	https://boletin.buap.mx/node/1999

Anexo 6 - Calidad de los Datos NGS

A continuación se muestra el resumen de calidad de los datos NGS (Next Generation Sequencing) que se usaron para el mapeo vs el genoma humano GRCh38. En este reporte se incluyen las 95 muestras originalmente secuenciadas. El reporte completo de Qualimap se puede descargar en: https://drive.google.com/file/d/1xssmozC94JuW7-JJNnCQcPS9ZpUawPXq/view?usp=sharing

Qualimap Analysis Results

Multi-sample BAM QC analysis
Generated by Qualimap v.2.2.1
2017/07/31 16:49:22

Control de Calidad Global

Número de muestras	**95**
Promedio de contenido G-C	**42.7**
Promedio de calidad de Mapeo	**32.35**
Promedio de longitud del inserto	**322.46**

Referencias

1. Coop G. Genetic similarity versus genetic ancestry groups as sample descriptors in human genetics. arXiv [q-bio.PE]. 2022. Available: http://arxiv.org/abs/2207.11595

2. Mendoza-Caamal EC, Barajas-Olmos F, García-Ortiz H, Cicerón-Arellano I, Martínez-Hernández A, Córdova EJ, et al. Metabolic syndrome in indigenous communities in Mexico: a descriptive and cross-sectional study. BMC Public Health. 2020;20: 339.

3. Birney E, Inouye M, Raff J, Rutherford A, Scally A. The language of race, ethnicity, and ancestry in human genetic research. arXiv [q-bio.PE]. 2021. Available: http://arxiv.org/abs/2106.10041

4. Silva CP, de la Fuente Castro C, González Zarzar T, Raghavan M, Tonko-Huenucoy A, Martínez FI, et al. The Articulation of Genomics, Mestizaje, and Indigenous Identities in Chile: A Case Study of the Social Implications of Genomic Research in Light of Current Research Practices. Front Genet. 2022;13: 817318.

5. Claw KG, Anderson MZ, Begay RL, Tsosie KS, Fox K, Garrison NA, et al. A framework for enhancing ethical genomic research with Indigenous communities. Nat Commun. 2018;9: 2957.

6. Ávila-Arcos MC, de la Fuente Castro C, Nieves-Colón MA, Raghavan M. Recommendations for Sustainable Ancient DNA Research in the Global South: Voices From a New Generation of Paleogenomicists. Front Genet. 2022;13: 880170.

7. Hudson M, Garrison NA, Sterling R, Caron NR, Fox K, Yracheta J, et al. Rights, interests and expectations: Indigenous perspectives on unrestricted access to genomic data. Nat Rev Genet. 2020;21: 377–384.

8. Ley Federal de Protección de Datos Personales en Posesión de los Particulares. [cited 25 Jan 2023]. Available: https://www.diputados.gob.mx/LeyesBiblio/ref/lfpdppp.htm

9. Hinds DA, Stuve LL, Nilsen GB, Halperin E, Eskin E, Ballinger DG, et al. Whole-genome patterns of common DNA variation in three human populations. Science. 2005;307: 1072–1079.

10. The International HapMap 3 Consortium. Integrating common and rare genetic variation in diverse human populations. Nature. 2010;467: 52.

11. A global reference for human genetic variation. Nature. 2015;526: 68–74.

12. Abren en la web mapa del genoma mexicano. In: Expansión [Internet]. 14 May 2009 [cited 1 Nov 2022]. Available: https://expansion.mx/actualidad/2009/05/14/abren-en-la-web-mapa-del-genoma-mexicano

13. Aguilar-Ordoñez I, Pérez-Villatoro F, García-Ortiz H, Barajas-Olmos F, Ballesteros-Villascán J, González-Buenfil R, et al. Whole genome variation in 27 Mexican indigenous populations, demographic and biomedical insights. PLoS One. 2021;16: e0249773.

14. Aguilar-Ordoñez I, Guzmán-Linares J, Ballesteros-Villascán J, Mirón-Toruño F, Pérez-González A, García-López J, et al. A Tale of Native American Whole-Genome Sequencing and Other Technologies. Diversity . 2022;14: 647.

15. 1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, et al. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012;491: 56–65.

16. Genome of the Netherlands Consortium. Whole-genome sequence variation, population structure and demographic history of the Dutch population. Nat Genet. 2014;46: 818–825.

17. Sherry ST, Ward M, Sirotkin K. dbSNP-database for single nucleotide polymorphisms and other classes of minor genetic variation. Genome Res. 1999;9: 677–679.

18. A haplotype map of the human genome. Nature. 2005;437: 1299–1320.

19. Website. Available: https://www.internationalgenome.org/announcements/first-data-release-snp-data-downloads-and-genome-browser-representing-four-high-covera/

20. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20: 1297–1303.

21. Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, et al. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. Nature. 2016;538: 201–206.

22. Bergström A, McCarthy SA, Hui R, Almarri MA, Ayub Q, Danecek P, et al. Insights into human genetic variation and population history from 929 diverse genomes. Science. 2020;367. doi:10.1126/science.aay5012

23. Fu W, O'Connor TD, Jun G, Kang HM, Abecasis G, Leal SM, et al. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature. 2013;493: 216–220.

24. Bryc K, Velez C, Karafet T, Moreno-Estrada A, Reynolds A, Auton A, et al. Genome-wide patterns of population structure and admixture among Hispanic/Latino populations. Proceedings of the National Academy of Sciences. 2010;107: 8954–8961.

25. Silva-Zolezzi I, Hidalgo-Miranda A, Estrada-Gil J, Fernandez-Lopez JC, Uribe-Figueroa L, Contreras A, et al. Analysis of genomic diversity in Mexican Mestizo populations to develop genomic medicine in Mexico. Proc Natl Acad Sci U S A. 2009;106: 8611–8616.

26. Need AC, Goldstein DB. Next generation disparities in human genomics: concerns and remedies. Trends Genet. 2009;25: 489–494.

27. Bustamante CD, Burchard EG, De la Vega FM. Genomics for the world. Nature. 2011;475: 163–165.

28. Popejoy AB, Fullerton SM. Genomics is failing on diversity. Nature. 2016. pp. 161–164.

29. Sirugo G, Williams SM, Tishkoff SA. The Missing Diversity in Human Genetic Studies. Cell. 2019;177: 26–31.

30. Caron NR, Chongo M, Hudson M, Arbour L, Wasserman WW, Robertson S, et al. Indigenous Genomic Databases: Pragmatic Considerations and Cultural Contexts. Front Public Health. 2020;8: 111.

31. SIGMA Type 2 Diabetes Consortium, Williams AL, Jacobs SBR, Moreno-Macías H, Huerta-Chagoya A, Churchhouse C, et al. Sequence variants in SLC16A11 are a common risk factor for type 2 diabetes in Mexico. Nature. 2014;506: 97–101.

32. Pringle H. Welcome to Beringia. Science. 2014;343: 961–963.

33. Romero-Hidalgo S, Ochoa-Leyva A, Garcíarrubio A, Acuña-Alonzo V, Antúnez-Argüelles E, Balcazar-Quintero M, et al. Demographic history and biologically relevant genetic variation of Native Mexicans inferred from whole-genome sequencing. Nat Commun. 2017;8: 1005.

34. Moreno-Estrada A, Gignoux CR, Fernández-López JC, Zakharia F, Sikora M, Contreras AV, et al. Human genetics. The genetics of Mexico recapitulates Native American substructure and affects biomedical traits. Science. 2014;344: 1280–1285.

35. Tiao G, Goodrich J. GnomAD v3.1 new content, methods, annotations, and data availability. [cited 6 Nov 2022]. Available: https://gnomad.broadinstitute.org/news/2020-10-gnomad-v3-1-new-content-methods-annotations-and-data-availability/

36. Taylor WW. The hunter-gatherer nomads of northern Mexico: A comparison of the archival and archaeological records. World Archaeol. 1972;4: 167–178.

37. Martínez-Tagüeña N, Torres Cubillas LA. Walking the desert, paddling the sea: Comcaac mobility in time. Journal of Anthropological Archaeology. 2018;49: 146–160.

38. Livi-Bacci M. The depopulation of Hispanic America after the conquest. Popul Dev Rev. 2006;32: 199–232.

39. El Mundo Indígena 2020: México. [cited 6 Nov 2022]. Available: https://www.iwgia.org/es/mexico/3745-mi-2020-mexico.html

40. Rangel-Villalobos H, Martínez-Sevilla VM, Martínez-Cortés G, Aguilar-Velázquez JA, Sosa-Macías M, Rubi-Castellanos R, et al. 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. 2016;160: 298–316.

41. Ávila-Arcos MC, McManus KF, Sandoval K, Rodríguez-Rodríguez JE, Villa-Islas V, Martin AR, et al. Population History and Gene Divergence in Native Mexicans Inferred from 76 Human Exomes. Mol Biol Evol. 2020;37: 994–1006.

42. Raghavan M, Steinrücken M, Harris K, Schiffels S, Rasmussen S, DeGiorgio M, et al. POPULATION GENETICS. Genomic evidence for the Pleistocene and recent population history of Native Americans. Science. 2015;349: aab3884.

43. Contreras-Cubas C, Sánchez-Hernández BE, García-Ortiz H, Martínez-Hernández A, Barajas-Olmos F, Cid M, et al. Heterogenous Distribution of MTHFR Gene Variants among Mestizos and Diverse Amerindian Groups from Mexico. PLoS One. 2016;11: e0163248.

44. Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen W-M. Robust relationship inference in genome-wide association studies. Bioinformatics. 2010;26: 2867–2873.

45. Nagasaki M, Yasuda J, Katsuoka F, Nariai N, Kojima K, Kawai Y, et al. Rare variant discovery by deep whole-genome sequencing of 1,070 Japanese individuals. Nat Commun. 2015;6: 8018.

46. Reuter MS, Walker S, Thiruvahindrapuram B, Whitney J, Cohn I, Sondheimer N, et al. The Personal Genome Project Canada: findings from whole genome sequences of the inaugural 56 participants. CMAJ. 2018;190: E126–E136.

47. Paschou P, Ziv E, Burchard EG, Choudhry S, Rodriguez-Cintron W, Mahoney MW, et al. PCA-correlated SNPs for structure identification in worldwide human populations. PLoS Genet. 2007;3: 1672–1686.

48. Predicting geographic population using genome variants and K-Means. In: Databricks [Internet]. 24 May 2016 [cited 27 Jan 2023]. Available: https://www.databricks.com/blog/2016/05/24/predicting-geographic-population-using-genome-variants-and-k-means.html

49. Harris DN, Song W, Shetty AC, Levano KS, Cáceres O, Padilla C, et al. Evolutionary genomic dynamics of Peruvians before, during, and after the Inca Empire. Proc Natl Acad Sci U S A. 2018;115: E6526–E6535.

50. Reich D, Patterson N, Campbell D, Tandon A, Mazieres S, Ray N, et al. Reconstructing Native American population history. Nature. 2012;488: 370–374.

51. Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. 2019;47: D1005–D1012.

52. Landrum MJ, Lee JM, Benson M, Brown G, Chao C, Chitipiralla S, et al. ClinVar: public archive of interpretations of clinically relevant variants. Nucleic Acids Res. 2016;44: D862–8.

53. Fishilevich S, Nudel R, Rappaport N, Hadar R, Plaschkes I, Iny Stein T, et al. GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. Database . 2017;2017. doi:10.1093/database/bax028

54. Auwerx J. PPARgamma, the ultimate thrifty gene. Diabetologia. 1999;42: 1033–1049.

55. Kris Hirst K. The Beringian Standstill Hypothesis: An overview. In: ThoughtCo [Internet]. 27 Feb 2014 [cited 1 Nov 2022]. Available: https://www.thoughtco.com/beringian-standstill-hypothesis-first-americans-172859

56. Motawi TK, Shaker OG, Ismail MF, Sayed NH. Peroxisome Proliferator-Activated Receptor Gamma in Obesity and Colorectal Cancer: the Role of Epigenetics. Sci Rep. 2017;7: 10714.

57. Stryjecki C, Peralta-Romero J, Alyass A, Karam-Araujo R, Suarez F, Gomez-Zamudio J, et al. Association between PPAR-γ2 Pro12Ala genotype and insulin resistance is modified by circulating lipids in Mexican children. Sci Rep. 2016;6: 24472.

58. Wu S, Kim T-K, Wu X, Scherler K, Baxter D, Wang K, et al. Circulating MicroRNAs and Life Expectancy Among Identical Twins. Ann Hum Genet. 2016;80: 247–256.

59. Liu Y, Chen S, Zhang J, Shan S, Chen L, Wang R, et al. Analysis of Serum MicroRNAs as Potential Biomarker in Coronary Bifurcation Lesion. Dis Markers. 2015;2015: 351015.

60. Xu J, Li M, Pei W, Ding J, Pan Y, Peng H, et al. Reduced Circulating Levels of miR-491-5p and miR-485-3p Are Associated with the Occurrence of Vertebral Fractures in Postmenopausal Women with Osteoporosis. Genet Res . 2022;2022: 3838126.

61. Zhu N, Huang K, Liu Y, Zhang H, Lin E, Zeng Y, et al. miR-195-5p Regulates Hair Follicle Inductivity of Dermal Papilla Cells by Suppressing Wnt/ß-Catenin Activation. Biomed Res Int. 2018;2018: 4924356.

62. Mohd Ali N, Boo L, Yeap SK, Ky H, Satharasinghe DA, Liew WC, et al. Probable impact of age and hypoxia on proliferation and microRNA expression profile of bone marrow-derived human mesenchymal stem cells. PeerJ. 2016;4: e1536.

63. Campbell MJ. Vitamin D and the RNA transcriptome: more than mRNA regulation. Front Physiol. 2014;5: 181.

64. Wei Z, Jiang S, Zhang Y, Wang X, Peng X, Meng C, et al. The effect of microRNAs in the regulation of human CYP3A4: a systematic study using a mathematical model. Sci Rep. 2014;4: 4283.

65. Marchand L, Jalabert A, Meugnier E, Van den Hende K, Fabien N, Nicolino M, et al. miRNA-375 a Sensor of Glucotoxicity Is Altered in the Serum of Children with Newly Diagnosed Type 1 Diabetes. J Diabetes Res. 2016;2016: 1869082.

66. Kuo T-Y, Hsi E, Yang I-P, Tsai P-C, Wang J-Y, Juo S-HH. Computational analysis of mRNA expression profiles identifies microRNA-29a/c as predictor of colorectal cancer early recurrence. PLoS One. 2012;7: e31587.

67. Gong J, Tong Y, Zhang H-M, Wang K, Hu T, Shan G, et al. Genome-wide identification of SNPs in microRNA genes and the SNP effects on microRNA target binding and biogenesis. Hum Mutat. 2012;33: 254–263.

68. Liu C-J, Fu X, Xia M, Zhang Q, Gu Z, Guo A-Y. miRNASNP-v3: a comprehensive database for SNPs and disease-related variations in miRNAs and miRNA targets. Nucleic Acids Res. 2021;49: D1276–D1281.

69. Clark RJ. Differential microRNA Expression in Barrett's Esophagus correlates with regulation of Posterior Homeotic Genes. Wright State University. 2019. Available: https://corescholar.libraries.wright.edu/etd_all/2181/

70. Pallarès-Albanell J, Zomeño-Abellán MT, Escaramís G, Pantano L, Soriano A, Segura MF, et al. A High-Throughput Screening Identifies MicroRNA Inhibitors That Influence Neuronal Maintenance and/or Response to Oxidative Stress. Mol Ther Nucleic Acids. 2019;17: 374–387.

71. Sun C, Liu J, Duan F, Cong L, Qi X. The role of the microRNA regulatory network in Alzheimer's disease: a bioinformatics analysis. Arch Med Sci. 2022;18: 206–222.

72. Zhu Y, Xie J, Sun H. Three miRNAs cooperate with host genes involved in human cardiovascular disease. Hum Genomics. 2019;13: 40.

73. Séguin B, Hardy B-J, Singer PA, Daar AS. Genomics, public health and developing countries: the case of the Mexican National Institute of Genomic Medicine (INMEGEN). Nat Rev Genet. 2008;9 Suppl 1: S5–9.

74. Séguin B, Hardy B-J, Singer PA, Daar AS. Genomic medicine and developing countries: creating a room of their own. Nat Rev Genet. 2008;9: 487–493.

75. Helmy M, Awad M, Mosa KA. Limited resources of genome sequencing in developing countries: Challenges and solutions. Appl Transl Genom. 2016;9: 15–19.

76. Meagher KM, Lee LM. Integrating Public Health and Deliberative Public Bioethics: Lessons from the Human Genome Project Ethical, Legal, and Social Implications Program. Public Health Rep. 2016;131: 44–51.

77. de Vries J, Bull SJ, Doumbo O, Ibrahim M, Mercereau-Puijalon O, Kwiatkowski D, et al. Ethical issues in human genomics research in developing countries. BMC Med Ethics. 2011;12: 5.

78. Mello MM, Wolf LE. The Havasupai Indian tribe case--lessons for research involving stored biologic samples. N Engl J Med. 2010;363: 204–207.

79. Dalton R. Tribe blasts "exploitation" of blood samples. Nature. 2002;420: 111.

80. Merriman T, Cameron V. Risk-taking: behind the warrior gene story. N Z Med J. 2007;120: U2440.

81. Jiménez-Kaufmann A, Chong AY, Cortés A, Quinto-Cortés CD, Fernandez-Valverde SL, Ferreyra-Reyes L, et al. Imputation Performance in Latin American Populations: Improving Rare Variants Representation With the Inclusion of Native American Genomes. Front Genet. 2021;12: 719791.

82. Marlett SA. The Seris and the Comcaac: Sifting fact from fiction about the names and relationships. Work Papers of the Summer Institute of Linguistics, University of North Dakota Session. 2011;51: 1.