



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. JUNIO, 2022



## AGRADECIMIENTOS

Quiero expresar mi gratitud a mi familia por su incondicional apoyo y cariño en mi vida. Todos sus esfuerzos y sacrificios han hecho posible este logro.

A mi tutor de tesis, **Dr. Israel Aguilar Ordóñez**, por darme la oportunidad de formar parte de tu laboratorio virtual, por el tiempo cada semana (o el día que lo requiriese) de revisiones y enseñanzas en genética, genómica, bioinformática, y también de vida. Gracias por confiar en mí y por tu amistad.

A mi cotutor de tesis, **Dr. Carlos A. Contreras Paredes**, por la oportunidad de trabajar en su laboratorio aun sin tener mucha experiencia, por tenerme paciencia y enseñarme y corregirme de la mejor manera posible. Agradezco su mentoría en este proyecto.

A mis asesores de tesis, **M.C. Elda Carreón Moreno y Dr. Miguel Castañeda Lucio**, pues sus contribuciones a este trabajo han sido de gran ayuda.

A mis compañeros del Laboratorio de Bioinformática, **Fernanda, Paulina, Eduardo, Judith y Ram**, gracias por su amistad y contribución a este proyecto.

A mis mejores amigas y amigos de la universidad, **Michael C., Yael G., Isaac H., Kenneth I., Sergio I., Pablo M., Jesús O., Krystel M., Sofía M., Elena R., Natalia S., Eva G., Dulce S., Ahibi I.**, por ser parte esencial de mi vida universitaria, por los momentos, aventuras, comidas, pláticas y fiestas que hicieron de una gran experiencia estos años de universidad. Ustedes son mi segunda familia.

A los amigos que he conservado a través del tiempo, **Manuel F., Saúl R., Alfonso U., Nila P., Jessica M., Andrea M., Berenice G., y Sandra G.**, gracias por los buenos momentos que hemos vivido a través de muchos años, por su apoyo, ánimos y buenos deseos; ustedes también forman parte de mi segunda familia.

## ÍNDICE GENERAL

<b>AGRADECIMIENTOS</b>	<b>2</b>
<b>ÍNDICE GENERAL</b>	<b>3</b>
<b>LISTA DE FIGURAS</b>	<b>6</b>
<b>LISTA DE ESQUEMAS</b>	<b>6</b>
<b>LISTA DE ABREVIATURAS</b>	<b>7</b>
<b>RESUMEN</b>	<b>8</b>
<b>1. INTRODUCCIÓN</b>	<b>9</b>
<b>2. ANTECEDENTES</b>	<b>11</b>
2.1 El inicio de la era genómica	11
2.2 Las tecnologías genómicas: El nuevo microscopio del siglo XXI	13
2.2.1 Secuenciación de genoma completo (WGS)	14
2.2.2 Secuenciación de exoma completo (WES)	14
2.2.3 Genotipado por microarreglos (Arrays)	15
2.2.4 Variantes genéticas	16
2.2.4.1 Clasificación de variantes	17
2.2.5 La importancia de la bioinformática y la ciencia de datos en genómica	18
2.3 ¿Cuáles son los beneficios de estudiar genomas?: La promesa de la medicina genómica	20
2.4 Proyectos Genómicos Poblacionales	20
2.4.1 Hapmap	21
2.4.2 1000G	21
2.4.3 SGDP	21
2.4.4 ExAC	22
2.4.5 HGDP	22
2.4.6 gnomAD	22
2.4.7 La importancia de la diversidad genómica: la subrepresentación de grupos nativos americanos	23

2.5 Proyectos genómicos de la población nativa americana	24
2.6 La disponibilidad de datos en la genómica	25
2.6.2 Datos compartidos mediante acceso controlado	26
2.6.3 Datos compartidos bajo petición	26
2.6.4 Datos compartidos exclusivamente en un grupo de investigadores	26
<b>3. MARCO TEÓRICO</b>	<b>27</b>
3.1 Nativos americanos contemporáneos, o restos antiguos de nativos americanos	28
3.1.1 ADN de nativos americanos contemporáneos	28
3.1.2 ADN de restos de nativos americanos (o ADN antiguo)	28
3.2 Distribución de la población nativa en América	28
3.3 ¿En dónde se almacenan los datos genómicos?	29
3.3.1 NCBI SRA	30
3.3.2 EBI ENA	30
3.3.3 The international genome sample resource	30
3.3.4 EBI EGA	30
3.3.5 Otras bases de datos	31
3.3.5.1 Base de datos de David Reich	31
3.3.5.2 Figshare	31
<b>4. PLANTEAMIENTO DEL PROBLEMA</b>	<b>31</b>
<b>5. PREGUNTA DE INVESTIGACIÓN</b>	<b>32</b>
<b>6. OBJETIVOS</b>	<b>32</b>
6.1 Objetivo general	32
6.2 Objetivos específicos	32
<b>7. MATERIALES Y MÉTODOS</b>	<b>32</b>
7.1 Búsqueda bibliográfica.	33
7.1.1 Criterios de inclusión de los artículos revisados en este trabajo.	34
7.1.1.1 Tecnología utilizada	34

7.1.1.2 Tipo de ADN estudiado	34
7.1.1.3 Disponibilidad de datos:	34
7.2 Selección de metadatos de los artículos e integración del compilado.	35
7.2.1 Integración de la base de datos	36
7.3 Recopilación de información censal de NatAm en los países de América.	37
<b>8. RESULTADOS</b>	<b>37</b>
8.1 Análisis general de estudios genómicos NatAm	40
8.2 Los grupos NatAm estudiados	41
8.2.1 Comparación de la población NatAm representada en genómica contra censos de población NatAm	43
8.3 Disponibilidad de datos	45
8.4 El interés en la investigación genómica nativa americana	47
<b>9. DISCUSIÓN</b>	<b>48</b>
<b>10. CONCLUSIONES</b>	<b>52</b>
<b>11. PERSPECTIVAS</b>	<b>53</b>
<b>12. REFERENCIAS</b>	<b>54</b>
<b>13. MATERIAL SUPLEMENTARIO</b>	<b>76</b>

## **LISTA DE FIGURAS**

Figura 1. Hitos de la era genómica.

Figura 2. Tecnologías genómicas (Secuenciación y microarreglos de ADN).

Figura 3. Variantes genéticas

Figura 4. La cantidad de datos producidos en genómica.

Figura 5. Formas de compartir datos en la genómica.

Figura 6. Países de origen y diversidad de nativos americanos.

Figura 7. Los estudios genómicos de nativos americanos a través del tiempo.

Figura 8. Vista general de los grupos NatAm representados en genómica.

Figura 9. Grupos NatAm no representados en genómica.

Figura 10. Comparación de individuos NatAm representados en genómica con censos poblacionales.

Figura 11. La disponibilidad de datos de los estudios de NatAm.

Figura 12. Tipos de Acceso reportados en los estudios de NatAm.

Figura 13. Participación de países en la investigación genómica de la población.

## **LISTA DE ESQUEMAS**

Esquema 1. Metodología general para la integración de la base de datos de nativos americanos.

Esquema 2. Metodología para la búsqueda bibliográfica de artículos.

Esquema 3. Metodología para la recopilación de metadata y la integración de la base de datos.

Esquema 4. Metodología para la recopilación de censos poblacionales de América.

## **LISTA DE ABREVIATURAS**

GWAS: Estudio de asociación de genoma completo

NatAm: Nativo(s) Americano(s)

PGH: Proyecto del Genoma Humano

SNP: Polimorfismo de Nucleótido Único

SNV: Variación de Nucleótido Único

WGS: Secuenciación de Genoma completo

WES: Secuenciación de Exoma completo

## RESUMEN

La población nativa de América está subrepresentada dentro de los estudios de la genómica humana. En los estudios y bases de datos genómicas existentes, los individuos nativos americanos representan un porcentaje muy bajo en comparación con individuos pertenecientes a poblaciones europeas, lo que representa un sesgo en la investigación genómica. Como consecuencia, los conocimientos y las aplicaciones de la genómica han beneficiado muy poco a los individuos nativos de América y sus descendientes.

Para contribuir a integrar la escasa información genómica en un solo lugar de referencia, en el presente trabajo se integró una base de datos con información de estudios genómicos de nativos americanos a partir de la revisión de 56 publicaciones científicas, en las que el ADN antiguo o contemporáneo de nativos americanos de todo el continente fue secuenciado o genotipado.

Nuestros resultados muestran que 13,706 nativos americanos han sido estudiados con tecnologías genómicas, de los cuales 1,292 proporcionaron muestras de genoma completo. La disponibilidad de datos es escasa, ya que apenas el 4.4 % de las muestras de nativos contemporáneos son accesibles para estudios posteriores. Se recopilaron reportes censales de los países en los que se ha reportado presencia de grupos nativos, y al comparar la información con los grupos ya estudiados en las publicaciones revisadas, se descubrió que 607 grupos indígenas siguen sin estar representados en los conjuntos de datos genómicos. Al analizar la autoría de los estudios revisados se encontró que es necesaria una mayor participación de los países en los que habitan individuos nativos americanos en los estudios genómicos.

La base de datos generada ayudará en el diseño de investigaciones que tengan como objetivo impulsar el estudio de la diversidad de la población nativa de América. Además, resaltamos la importancia y necesidad de tener más proyectos de secuenciación de genomas de poblaciones nativas en América que contribuyan al desarrollo y aplicación del conocimiento de la genómica en el continente.

## 1. INTRODUCCIÓN

A nivel mundial los estudios genómicos poblacionales han centrado su atención en países de ingresos altos lo que ha provocado que casi 80% de los genomas estudiados pertenezcan a individuos con componente ancestral europeo, y ha dejado en porcentajes muy bajos (menos del 1%) a otras poblaciones como la nativa de América (Sirugo et al., 2019), a pesar de que la diversidad genética del continente americano representa, en cierto modo, la diversidad genética del mundo, debido a los contactos que se han producido entre los grupos nativos de América y otras poblaciones del mundo (Adhikari et al., 2017). Por lo tanto, es entendible que los conocimientos y las aplicaciones de la genómica hasta ahora han beneficiado poco al continente americano debido a las desigualdades en los sistemas de atención sanitaria, la escasa mano de obra local para la investigación y la falta de financiación en genómica.

Es necesario hacer la siguiente aclaración: los términos "nativo americano" o "indígena" hacen referencia a todos los individuos pertenecientes a grupos, pueblos o comunidades indígenas que habitan todo el continente de América (Bolnick et al., 2016). A lo largo de este trabajo se usará el término nativo americano para referirse a estos pueblos.

El continente americano alberga un gran número de grupos nativos americanos (NatAm), y algunos han sido estudiados ya sea mediante secuenciación de genoma completo (WGS), de exoma completo (WES) o genotipado por microarreglos de ADN (Arrays). Sin embargo, en comparación con países de mayores recursos, los estudios y los datos genómicos generados de estos individuos son mucho menores (Aguilar-Ordoñez et al., 2021). Además, la disponibilidad de datos de los estudios de NatAm sigue lidiando con dificultades debido a cuestiones técnicas, sociales y culturales. En la investigación genómica de NatAm hay que considerar los aspectos éticos, legales y de privacidad que rodean a los participantes de estos estudios para promover un mejor desarrollo científico y cultural que beneficie a todas las partes involucradas.

Para lograr capturar todo el alcance de la diversidad genética y mejorar la atención sanitaria, la equidad y la investigación médica en América, se han realizado algunos estudios sobre la diversidad genética de la población NatAm que están presentes en el continente a través de sus genomas. Lo anterior ha permitido generar conocimiento base que podría impactar en diversos campos como la investigación biomédica, antropológica y evolutiva (Aguilar-Ordoñez et al., 2021; Moreno-Estrada et al., 2014; Reich et al., 2012; Romero-Hidalgo et al., 2017).

Los estudios genómicos en poblaciones NatAm son pocos en comparación con otras regiones más desarrolladas. Pero incluso si la cantidad de datos genómicos son escasos, los investigadores de genómica humana pueden aprovechar la ventaja de que si los datos generados se comparten de manera libre, se pueden analizar y plantear más investigaciones para seguir descubriendo la diversidad genética de las poblaciones de América. Por lo tanto es importante tener una recopilación de cada investigación ya sea de secuenciación WGS, WES o Arrays que se haya hecho en América para estudiar la diversidad genética de los grupos NatAm. Esta información puede volverse una referencia integral para futuras investigaciones de la diversidad genética de los grupos NatAm en genómica, por ejemplo, al permitir conocer de una manera sencilla qué cantidad de individuos nativos ya han sido representados o qué grupos nativos no han sido estudiados aún o tienen muy pocos individuos representados.

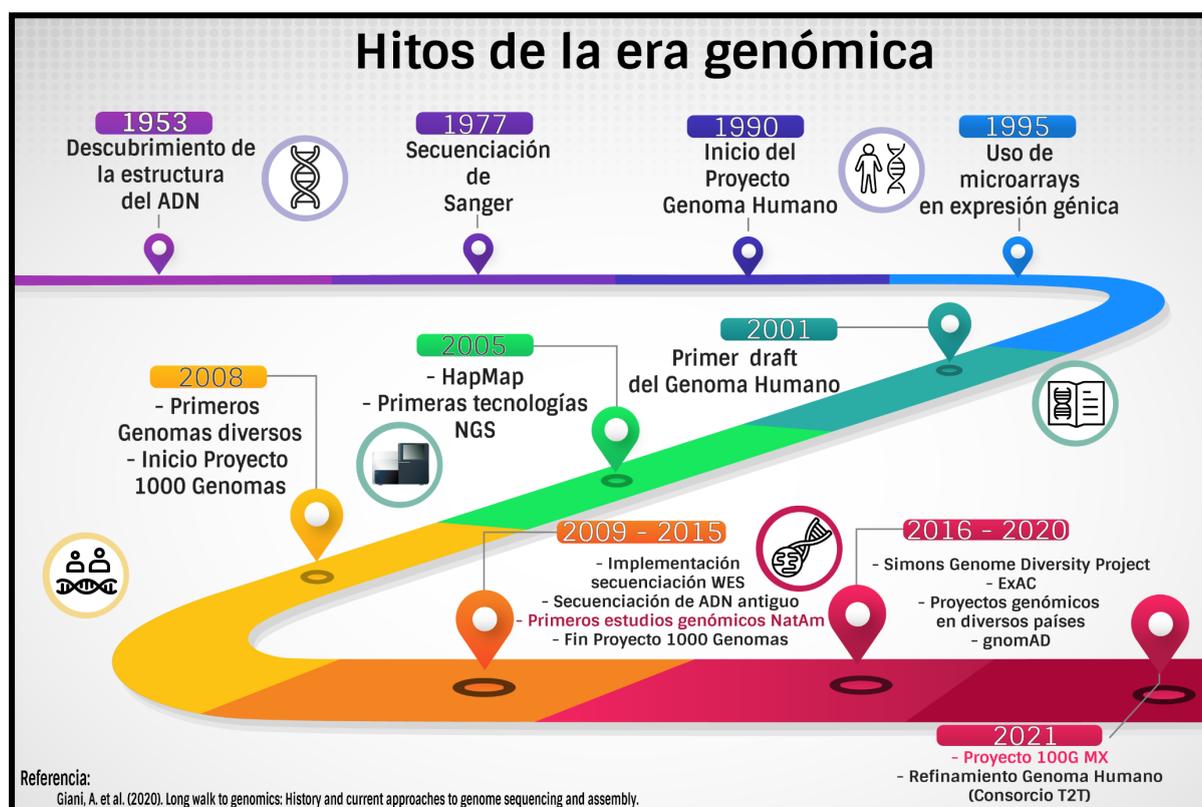
En nuestro grupo de investigación surgieron las siguientes preguntas: ¿Cuántos genomas NatAm estudiados mediante secuenciación o Arrays se han reportado en artículos de investigación?, ¿los datos generados están disponibles libremente?, es decir, ¿forman parte del acervo mundial para poder integrarlo en estudios posteriores?.

Para responder estas preguntas, en el presente proyecto se creó la primera base de datos que contiene información de genomas secuenciados o genotipados de NatAm, para contribuir al análisis de esta población continental en futuros proyectos.

## 2. ANTECEDENTES

### 2.1 El inicio de la era genómica

La coexistencia de varias tecnologías y descubrimientos científicos fueron necesarios para que la era del genoma surgiera y evolucionara, permitiendo la secuenciación y el análisis de genomas completos (Figura 1). El descubrimiento de las enzimas de restricción (Smith & Welcox, 1970), el desarrollo de vectores de clonación para la creación de bibliotecas genómicas (Borck et al., 1976), las técnicas de hibridación como el Southern blot (Southern, 1975), y el descubrimiento de la PCR (reacción en cadena de la polimerasa) para facilitar la amplificación del ADN (Mullis et al., 1986), fueron algunas técnicas sumamente importantes para dar pie a una revolución genómica, que comenzó con el Proyecto del Genoma Humano (PGH).



**Figura 1: Hitos de la era genómica.** Línea de tiempo que ilustra algunos de los principales logros del estudio de los genomas, desde el inicio de la era de la secuenciación hasta los proyectos genómicos a gran escala de la actualidad. Adaptada de: (Giani et al., 2020).

El PGH ha sido una de las grandes hazañas de exploración de la historia humana. Fue un viaje de descubrimiento de la especie dirigido por un equipo internacional de investigadores que buscaban secuenciar y mapear todos los genes de nuestra especie, *Homo sapiens* (humanos) (National Human Genome Research Institute, 2020). Comenzando el 1 de octubre de 1990 y completado en abril de 2003, el PGH nos dio la capacidad, por primera vez, de empezar a leer el libro que contiene la información biológica que define a un humano.

El primer borrador del PGH ha tenido un gran impacto en la investigación biomédica y el desarrollo tecnológico (Lander et al., 2001; Venter et al., 2001). Los grandes conjuntos de datos genómicos que se han generado (y continúan creciendo) a partir de ese gran avance son ahora recursos valiosos para guiar la investigación científica y clínica, el desarrollo de medicamentos y la práctica médica. Al día de hoy el genoma humano de referencia sigue siendo mejorado y estudiado (Nurk et al., 2021) para lograr comprender cómo la información genómica conduce al desarrollo, la función y el mal funcionamiento de las células y los organismos, y aprovechar plenamente este conocimiento para promover la salud humana y tratar las enfermedades. Esto nos permite empezar a responder cuestiones biológicas como: ¿Por qué algunas personas enferman gravemente, mientras que otras permanecen, relativamente sanas durante toda su vida?, ¿por qué una enfermedad evoluciona de forma tan diferente en cada persona?, ¿por qué un fármaco funciona de manera óptima en ciertos individuos y no en otros?, incluso nos permite responder cuestiones evolutivas y/o antropológicas como: ¿históricamente cuáles fueron los caminos de nuestros ancestros?

La respuesta a estas preguntas puede variar en cada persona del mundo. Es por eso que, si bien el genoma humano de referencia es una herramienta valiosa, no es suficiente para entender en su totalidad a toda la especie humana (Ballouz

et al., 2019), por lo tanto, se necesitan las secuencias completas de un gran número de individuos de diferentes componentes ancestrales para representar mejor la variabilidad genómica humana en el mundo. Se necesitan diversos estudios genómicos a gran escala.

## **2.2 Las tecnologías genómicas: El nuevo microscopio del siglo XXI**

Los estudios de la variación genética humana mediante la secuenciación de ADN han experimentado un gran desarrollo desde hace más de 40 años. Casi 25 años después de que la estructura del ADN fue descubierta (Watson & Crick, 1953), el primer método de secuenciación de ADN fue publicado en 1977, conocido como secuenciación de Sanger (Sanger et al., 1977). El método siguió mejorando con la introducción de la electroforesis capilar y ganó una amplia aceptación como método de "secuenciación de primera generación" en la década de 1990 para secuenciar genomas pequeños y grandes, desde bacterias y fagos, hasta su mayor aplicación en el Proyecto del Genoma Humano.

Sin embargo, la secuenciación de Sanger tuvo limitaciones que no la hicieron adecuada para aplicarla a proyectos de secuenciación poblacional. Si bien tiene una buena precisión, se limita a secuenciar un solo fragmento de ADN y a una longitud máxima de hasta 1,000 pares de bases (Petersen et al., 2017).

En vista de las dificultades técnicas y económicas, los pioneros en secuenciación poblacional comenzaron a buscar y desarrollar tecnologías más baratas y rápidas para secuenciar, lo que dio como resultado en 2005 la publicación de los dos primeros estudios que describieron las tecnologías de siguiente generación de secuenciación genómica (NGS) (Margulies et al., 2005; Shendure et al., 2005). El costo y el tiempo de la secuenciación han disminuido gracias al éxito comercial y a la fiabilidad de las tecnologías NGS, cambiando de forma radical los estudios de la diversidad genética y las enfermedades humanas en el mundo (Petersen et al., 2017). A continuación, resumiré algunas de las características de las diferentes tecnologías usadas para estudiar el genoma humano.

### 2.2.1 *Secuenciación de genoma completo (WGS)*

La WGS proporciona el orden de la secuencia de las 3 mil millones de pares de bases en el genoma humano (Figura 2), y da una imagen de todo el panorama genómico de un organismo, permitiendo descubrir nuevas regiones y variantes genéticas, ya sean comunes o raras (Manzoni et al., 2018).

Esta tecnología ha tenido un gran impacto en los estudios a gran escala de genómica poblacional, pues ha permitido explorar la diversidad genética de varias poblaciones del mundo. Ha permitido integrar biobancos que son útiles para la investigación genómica y su aplicación en la medicina (Carress et al., 2021). Es ideal que los grupos nativos americanos con menos (o nula) representación en genómica sean estudiados con esta tecnología, pues permitirá conocer totalmente su variación genética.

En 2018 se estimó que a nivel mundial el número de individuos cuyos genomas completos han sido secuenciados desde 2003 ascendía a 1.5 millones (Shendure et al., 2019) y al día de hoy la cifra sigue aumentando gracias a la reducción de los costos de la secuenciación, llegando al increíble precio de 300 dólares por individuo (Nebula Genomics, 2022).

### 2.2.2 *Secuenciación de exoma completo (WES)*

La región genómica que codifica las proteínas está fragmentada en varios trozos, llamados exones (Sastre, 2014). Los exones (que en conjunto se conocen como exoma) representan una porción pequeña del genoma, aproximadamente 1% (Pruitt et al., 2009). Por lo que la WES tiene la capacidad de “interrogar” a todas las regiones de codificación de proteínas con una gran profundidad de secuenciación y de forma más económica que WGS (Figura 2). Desde su implementación en 2009 (Ng et al., 2009), ha tenido impacto en áreas como la genómica del cáncer (Vogan, 2021) y en el conocimiento de la historia genómica de los grupos nativos mexicanos (Ávila-Arcos et al., 2020). La ventaja de que WES pueda cubrir una parte tan funcional del genoma, puede ser aprovechada por la genética de poblaciones para

diseñar estudios centrados en fenotipos específicos, o para sondear hipótesis de adaptación y selección natural (Udpa et al., 2014).

### *2.2.3 Genotipado por microarreglos (Arrays)*

Existe una tercera técnica comúnmente utilizada para el estudio genético de las poblaciones, conocida como Arrays de ADN (Figura 2). Los Arrays de ADN son colecciones de sondas ancladas a un soporte sólido que evalúan la existencia de millones de variantes genéticas específicas en la muestra analizada (Gresham et al., 2008). Si WGS ve todo el panorama del genoma, y WES mira aproximadamente el 1% del mismo, los arrays de ADN se centran en puntos todavía más específicos del panorama. Estos puntos clave suelen tener una importancia demostrada como marcadores de componentes ancestrales, asociaciones fenotipo-genotipo (GWAS) (Witte, 2010) o biomarcadores en general. Los Arrays de ADN pueden considerarse de bajo coste en comparación con WGS y WES, pero hay que tener en cuenta que los arrays sólo pueden examinar efectivamente las variantes interesantes ya conocidas. Los Arrays de ADN son “ciegos” a las variaciones nuevas y desconocidas. Para el estudio de poblaciones previamente subrepresentadas en los estudios genómicos (como los nativos americanos), el uso de la tecnología de arrays de ADN para los proyectos de genotipado podría no ser la mejor estrategia, ya que la variación importante puede ocurrir en algunas poblaciones y no en otras.



**Figura 2: Tecnologías genómicas para estudiar el genoma.** Se muestran las principales características, ventajas y desventajas de WGS, WES y Arrays. Adaptada de: Ballesteros-Villascán, J. (2020).

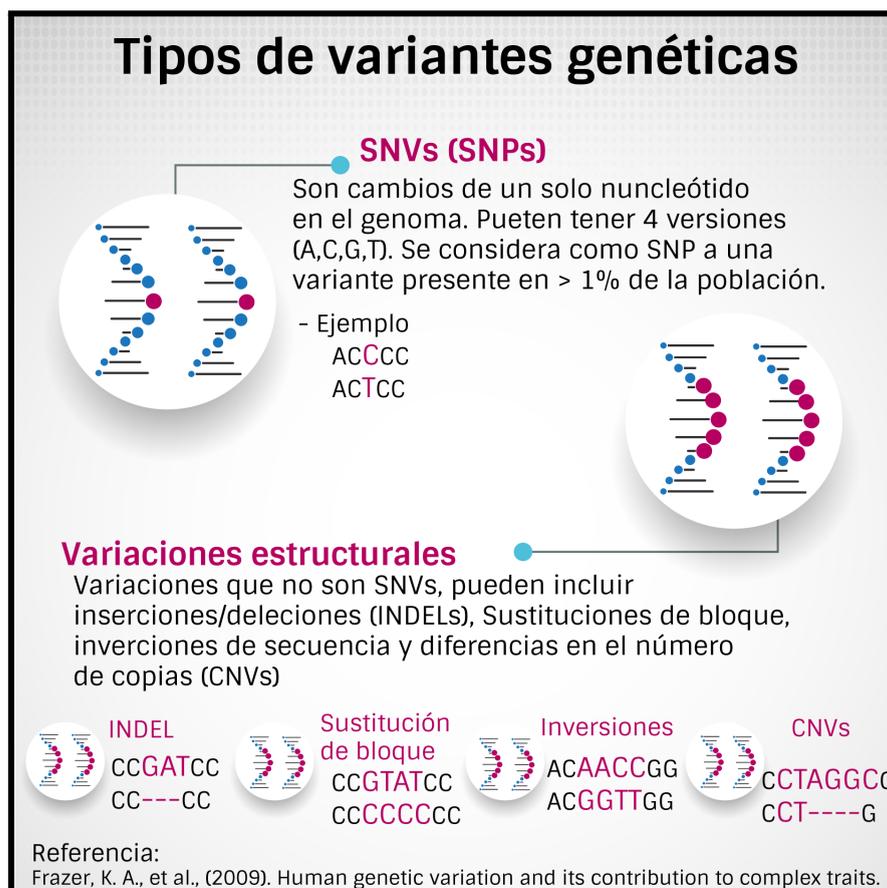
### 2.2.4 Variantes genéticas

El descubrimiento de variantes genéticas es esencial en los experimentos donde se usan tecnologías de WGS y/o WES, así como de Arrays. Una variante genética puede definirse como una diferencia en la secuencia entre dos genomas. Por ejemplo, una base en el genoma de un organismo que no es la misma que la de un genoma de referencia representa una variante (Frazer et al., 2009). Las variantes pueden ser la causa directa de una afección, dar lugar a un aumento o disminución de la predisposición al desarrollo de problemas de salud comunes, o pueden no tener efecto en absoluto. Las variantes genéticas se pueden clasificar de acuerdo al tipo de célula en el que se presenten y en el tipo de alteración que provocan en el genoma (Garvan Institute of Medical Research, 2021). A continuación se describen las variantes clasificadas por la alteración que provocan.

### 2.2.4.1 Clasificación de variantes

SNP y SNV: El Polimorfismo de Nucleótido Único (SNP por sus siglas en inglés), y la Variación de Nucleótido Único (SNV por sus siglas en inglés) hacen referencia a los cambios en un solo nucleótido en el genoma (Ranganathan et al., 2018). Pueden tener hasta cuatro versiones: una para cada nucleótido, A, C, G y T. Sin embargo, se considera SNP cuando la sustitución de un solo nucleótido está presente en más del 1% de la población (Figura 3). Un SNV, en cambio, es una variación en una posición que no ha sido bien caracterizada, por ejemplo, cuando sólo se ve en un individuo (Eichler, 2019; Frazer et al., 2009).

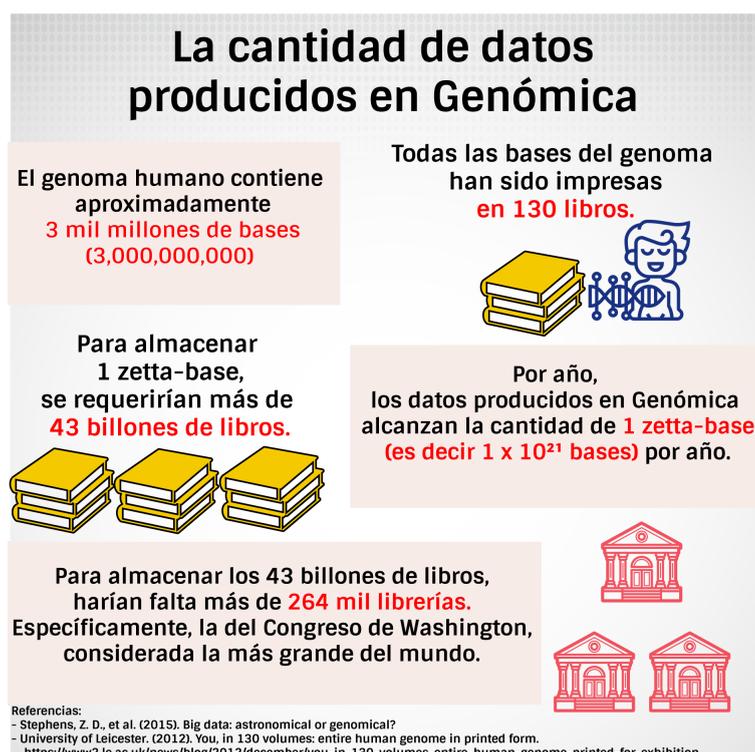
Variaciones estructurales: Se refieren a aquellas diferencias genómicas entre individuos que no son variantes de un solo nucleótido (Figura 3). Dicha variación incluye inserciones-delecciones (indels), sustituciones en bloque, inversiones de secuencias de ADN y diferencias en el número de copias (Frazer et al., 2009).



**Figura 3: Variantes genéticas.** Se describen las variantes genéticas clasificadas por su impacto biológico. Adaptada de Frazer, K. A., et al. (2009).

### 2.2.5 La importancia de la bioinformática y la ciencia de datos en genómica

La frase "Nos estamos ahogando en información, pero hambrientos de conocimiento" (Naisbitt, 1982) es muy aplicable en el mundo científico actual. En los últimos años la cantidad de datos generados por el uso de tecnologías genómicas ha sido enorme, llegando a producir ya desde la década pasada una cantidad de datos promedio de 1 zetta-base por año, es decir,  $10^{21}$  bases nucleotídicas (Stephens et al., 2015), lo que representaría más de 43 billones ( $4.3 \times 10^{13}$ ) de libros impresos con todas las bases nucleotídicas, y que podrían almacenarse en más de 264 mil bibliotecas (Figura 4). Sin embargo, no nos alcanzaría la vida para imprimir e interpretar esa cantidad de información.



**Figura 4. La cantidad de datos producidos en genómica.** Para entender la magnitud del número de datos biológicos generados por la genómica se puede entender a los genomas como libros, y la cantidad necesaria de estos para poder almacenar toda la información generada en genómica es inmensa.

Para lograr el objetivo de analizar la gran cantidad de datos producida es necesario contar con habilidades del campo de la bioinformática y la ciencia de datos. Si bien no es necesario volverse un experto en estas áreas, es esencial tener habilidades computacionales que permitan a un científico resolver tareas sencillas (como manejo y visualización de datos, uso consola de comandos, realización de scripts, entre otras), comunicarse de mejor manera con otros científicos más experimentados en el área computacional e incluso mejorar el pensamiento crítico a la hora de analizar los descubrimientos de las investigaciones (Attwood et al., 2019).

Se estima que en esta década se logren secuenciar entre cien millones y dos mil millones de genomas humanos (Stephens et al., 2015), es por eso que existe y seguirá habiendo una gran demanda de profesionales con conocimientos en bioinformática y ciencia de datos (Melendrez et al., 2021). Sin estas disciplinas no sería posible generar y analizar todos los datos genómicos.

### **2.3 ¿Cuáles son los beneficios de estudiar genomas?: La promesa de la medicina genómica**

Secuenciar genomas humanos nos permite comprender cómo la variación genómica contribuye a los estados de salud y enfermedad. Este conocimiento ha empezado a marcar el camino a la aplicación y promesa de la medicina genómica (Owen et al., 2021), la cual puede ser definida simplemente como el reconocimiento de la variación genética y/o genómica de cada persona para el tratamiento y prevención de las enfermedades (Lupski, 2016). Gracias a la reducción de costos de la secuenciación y al desarrollo de herramientas bioinformáticas en los últimos años, estudiar genomas humanos ha tenido impacto en áreas como la investigación genómica de enfermedades comunes y raras, en la salud reproductiva y prenatal, y en cáncer (Shendure et al., 2019).

Algunos países alrededor del mundo han empezado a invertir en la medicina genómica a través de la realización de proyectos genómicos poblacionales (Aguilar-Ordoñez et al., 2021; Genome of the Netherlands Consortium, 2014; Nagasaki et al., 2015; Wonkam, 2021). El objetivo es poder asegurar un futuro

donde las personas puedan acceder a la información contenida en sus genomas, al análisis de estos y un mejor sistema de salud que sea personalizado (Ashley, 2016).

## **2.4 Proyectos Genómicos Poblacionales**

Desde la publicación del borrador del genoma humano, las tecnologías de estudio del genoma han sido fundamentales para estudiar la variación genética humana, marcando el camino para implementar estudios genómicos a gran escala. Se han estudiado poblaciones de distintas partes del mundo para poder conocer su estructura genómica, así como la dinámica poblacional (Ballesteros Villascán, 2019). A continuación, se resumen los principales proyectos de genómica poblacional a nivel internacional que se han realizado.

### *2.4.1 Hapmap*

El proyecto HapMap (International HapMap Consortium, 2005), se inició en 2002 y concluyó en 2010, y tuvo como objetivo determinar patrones de variación común en 4 poblaciones pertenecientes a distintas partes del mundo (África, Europa, Sur de Asia y Este de Asia) (International HapMap Consortium, 2004, 2005). Este proyecto se llevó a cabo implementando el uso de Arrays de ADN para la detección de variantes, y fue útil para iniciar la búsqueda de genes que contribuyen a enfermedades o a la respuesta de fármacos (Ballesteros Villascán, 2019). Al estar basado en tecnología de microarreglos, el proyecto no pudo capturar el panorama completo de los genomas de los participantes del estudio. Sin embargo fue la base para la mejora de estudios de asociación de genoma completo a gran escala al proveer un contraste para la detección de variantes (Rabbee & Speed, 2006; Wellcome Trust Case Control Consortium, 2007).

### *2.4.2 1000G*

El proyecto de los 1,000 Genomas (1000 Genomes Project Consortium et al., 2015) se inició en 2009 y terminó en 2015, con el objetivo de secuenciar y detectar variantes genéticas en 2,504 humanos pertenecientes a 26 poblaciones distribuidas

en 5 regiones del mundo (África, Europa, Asia del sur, Asia del este y América). Fue el primer proyecto en crear una base de datos públicamente disponible sobre la variación humana.

#### *2.4.3 SGBP*

El "Simons Genome Diversity Project", fue publicado en 2016, y en este se secuenciaron 278 genomas que representan 142 poblaciones humanas para determinar elementos importantes en el panorama evolutivo de la humanidad, incluidas las medidas de parentesco y las diferentes tasas de mutación entre las poblaciones africanas y no africanas, así como la influencia de los primeros homínidos en las poblaciones actuales (Mallick et al., 2016). Este estudio es una respuesta al Proyecto 1,000 Genomas, que mientras analizaba un mayor número de individuos, sólo incluyó 26 poblaciones humanas geográficamente distintas. Los autores del SGBP sostienen que las poblaciones que se pasaron por alto en el Proyecto 1,000 Genomas, como la nativa de América, son importantes para comprender la variación en las poblaciones modernas de distintas poblaciones.

#### *2.4.4 ExAC*

El "Exome Aggregation Consortium" (ExAC) se publicó en 2016, reportando más de 60,000 secuencias para las partes codificantes de proteínas del genoma (exoma) de poblaciones humanas distintas (África, Europa, Asia del sur y del este, y la población latina de América), en una base de datos que ha continuado expandiéndose (Lek et al., 2016). El proyecto identificó 7,404,909 variantes de alta calidad, incluyendo 317,381 indels (Ballesteros Villascán, 2019), de las cuales la mayoría no habían sido reportadas en proyectos anteriores de secuenciación poblacional (1000 Genomes Project Consortium et al., 2015; Mallick et al., 2016).

#### *2.4.5 HGDP*

El "Human Genome Diversity Project" (HGDP) comenzó en la década de 1990 (Cavalli-Sforza, 2005), y su última publicación fue en 2020; a través de la secuenciación de 929 genomas humanos de 54 poblaciones diversas de todo el mundo, describió otra gran cantidad de variantes genéticas no descubiertas

anteriormente y proporcionó nuevos conocimientos sobre el pasado evolutivo de la humanidad, destacando la complejidad del proceso a través del cual nuestros antepasados se diversificaron, migraron y mezclaron en todo el mundo (Bergström et al., 2020). Este estudio representa el resultado de muchos años de trabajo de Luigi Cavalli-Sforza, uno de los científicos más destacados del mundo en el área de la genética para entender la migración y la evolución humana. Además, representa un gran recurso de variación genética de poblaciones humanas poco representadas como las nativas americanas.

#### 2.4.6 gnomAD

El “Genome Aggregation Database” (gnomAD) es un estudio sucesor del ExAC, e incluye la variación genética de 15,708 genomas completos y la adición de 125,748 exomas de 17 poblaciones distintas del mundo (Karczewski et al., 2020). El aumento del tamaño de las muestras secuenciadas y la inclusión de las regiones no codificantes permitieron obtener más de 240 millones de pequeñas variantes genéticas, así como variaciones estructurales. La Base de datos gnomAD es actualmente la mayor colección de variación genética poblacional parcialmente disponible a partir de datos de secuenciación, sin embargo, sigue dejando fuera de su estudio a la población NatAm.

#### 2.4.7 *La importancia de la diversidad genómica: la subrepresentación de grupos nativos americanos*

Las bases de datos de poblaciones de referencia (obtenidas a partir de estudios genómicos de gran escala) son una herramienta esencial en la interpretación de variantes genéticas en todos los genomas humanos, y apoyan al descubrimiento de nuevas relaciones entre enfermedades y genes (Gudmundsson et al., 2021). Como ya se describió en la sección anterior, los investigadores han tratado de genotipificar y secuenciar la mayor cantidad de humanos posible a través de esfuerzos de investigación internacionales, sin embargo, si se presta especial atención a las poblaciones y al número de individuos estudiados en estos proyectos, carecen de la

representación adecuada de algunas poblaciones, como lo es la asiática, africana y en mayor medida la nativa americana.

La falta de una mejor representación de poblaciones diversas en la investigación genómica humana ha sido un tema que se ha resaltado a lo largo del tiempo (Bustamante et al., 2011; Need & Goldstein, 2009; Popejoy & Fullerton, 2016; Sirugo et al., 2019). A pesar de los grandes avances en la identificación de variantes genéticas de relevancia biomédica, la mayoría de estudios genómicos a gran escala se han basado en estudiar poblaciones con un componente ancestral europeo (Guglielmi, 2019; Sirugo et al., 2019). El área donde se puede evidenciar este sesgo es en los participantes de los estudios GWAS (estudios destinados a encontrar relaciones estadísticas entre genotipo-fenotipo), en donde casi el 80 % pertenecen a la población europea, un 10 % son individuos de Asia, 6 % no tienen un componente ancestral definido, 2 % pertenecen a África, 1 % son individuos hispanos o latinos de América, y finalmente menos del 1 % son individuos nativos americanos (Guglielmi, 2019; Sirugo et al., 2019).

La mayoría de los proyectos genómicos en los países del mundo se han centrado en la descripción de la genómica de la población en países de altos ingresos económicos (Carress et al., 2021; Sirugo et al., 2019), dejando un vacío en la posible comprensión de la genómica subyacente a los procesos de salud y enfermedad del resto del mundo. La escasa cantidad de información genómica de los grupos NatAm limita los beneficios potenciales del conocimiento genómico para esta población, como son las aplicaciones biomédicas y la elaboración de políticas sanitarias, o incluso la información antropológica derivada de las secuencias genómicas de los NatAm (Bien et al., 2019; Hindorff et al., 2018).

## **2.5 Proyectos genómicos de la población nativa americana**

Gracias a los esfuerzos internacionales de investigación genómica en las poblaciones se tiene conocimiento de la variación genética que se presenta en el mundo de forma común. Sin embargo, las variantes raras pueden estar presentes

en las poblaciones no incluidas en los estudios genómicos, por ejemplo: la NatAm (Ballesteros Villascán, 2019).

Los estudios genómicos en la población NatAm son de suma importancia, pues conocer su diversidad genética representará en cierto modo saber cómo contribuyen a la diversidad genética del mundo (Adhikari et al., 2017). En total 56 proyectos se han realizado en América, en los que se han utilizado tecnologías de secuenciación y/o Arrays, y han incluido individuos NatAm representando a diversos grupos de distintos países del continente, o bien, analizan restos antiguos de individuos NatAm. Si bien la cantidad de individuos y grupos no han sido abundantes en comparación con proyectos de otros países con grandes recursos económicos, su contribución ha empezado a impactar en la investigación biomédica y antropológica (Jiménez-Kaufmann et al., 2022).

## **2.6 La disponibilidad de datos en la genómica**

Compartir los datos generados en un estudio científico es un elemento clave que es característico en la ciencia reproducible. En genómica NatAm aún falta mucho por alcanzar en este tema. Mucha información sigue siendo parcialmente disponible y/o privada, y con políticas y expectativas poco claras e inconsistentes sobre la forma en que se comparten los datos genómicos (Popkin, 2019). La ciencia en general avanza hacia una mayor apertura en términos de datos, liberación de investigación, código bioinformático y/o incluso flujos de trabajo completos. Hoy en día el avance de la tecnología ha favorecido la forma en que se comparten los datos, ya que cualquiera que tenga una conexión a Internet puede tener acceso a información científicamente relevante. Los investigadores que han compartido los datos generados en sus estudios pueden tener muchos beneficios: pueden generar nuevas colaboraciones, aumentar la confianza en los hallazgos y generar buena voluntad entre el gremio, incluso aumentar el impacto de su investigación ya que es más citada por ser más requerida (Colavizza et al., 2020).

En los estudios genómicos de poblaciones humanas las buenas prácticas para compartir datos son poco claras. Un obstáculo importante es la preocupación por la legalidad de compartir datos, especialmente cuando los sujetos de la investigación son personas (Popkin, 2019). Los investigadores deben considerar las cuestiones éticas, legales y de privacidad antes de poner a disposición los datos, por ejemplo, de grupos nativos, pues podría comprometer la privacidad o el bienestar de las personas pertenecientes a grupos ya de por sí vulnerables (Byrd et al., 2020; Claw et al., 2018). A continuación describiré las formas en que los datos genómicos generados en un estudio poblacional se pueden compartir.

### 2.6.1 Datos compartidos públicamente

La disponibilidad pública (Figura 5) ocurre cuando los datos generados de un estudio son almacenados y puestos a disposición sin restricción en alguna base de datos (en este caso, las cuestiones legales y éticas ya fueron consideradas, y se espera que el usuario esté familiarizado con ellas). Esta forma de compartir datos es la más beneficiosa para la comunidad científica, pues tiene barreras casi nulas de acceso, permite el avance de algún campo de investigación y los datos ya están estandarizados bajo una base de datos común (Byrd et al., 2020).

### 2.6.2 *Datos compartidos mediante acceso controlado*

La disponibilidad controlada de datos (Figura 5) tiene la característica de que, si bien la información es almacenada en una base de datos, el acceso al solicitante sólo será posible si se cumplen ciertos requisitos que están relacionados a cuestiones éticas y legales, tales como una revisión del protocolo del uso de los datos o un compromiso de utilizar los datos sólo para investigaciones relacionadas con ciertos temas como salud o antropología molecular. Esta forma de compartir suele suponer una pequeña barrera para los esfuerzos de reutilización, pero promueve la máxima seguridad y privacidad de los individuos (Byrd et al., 2020).

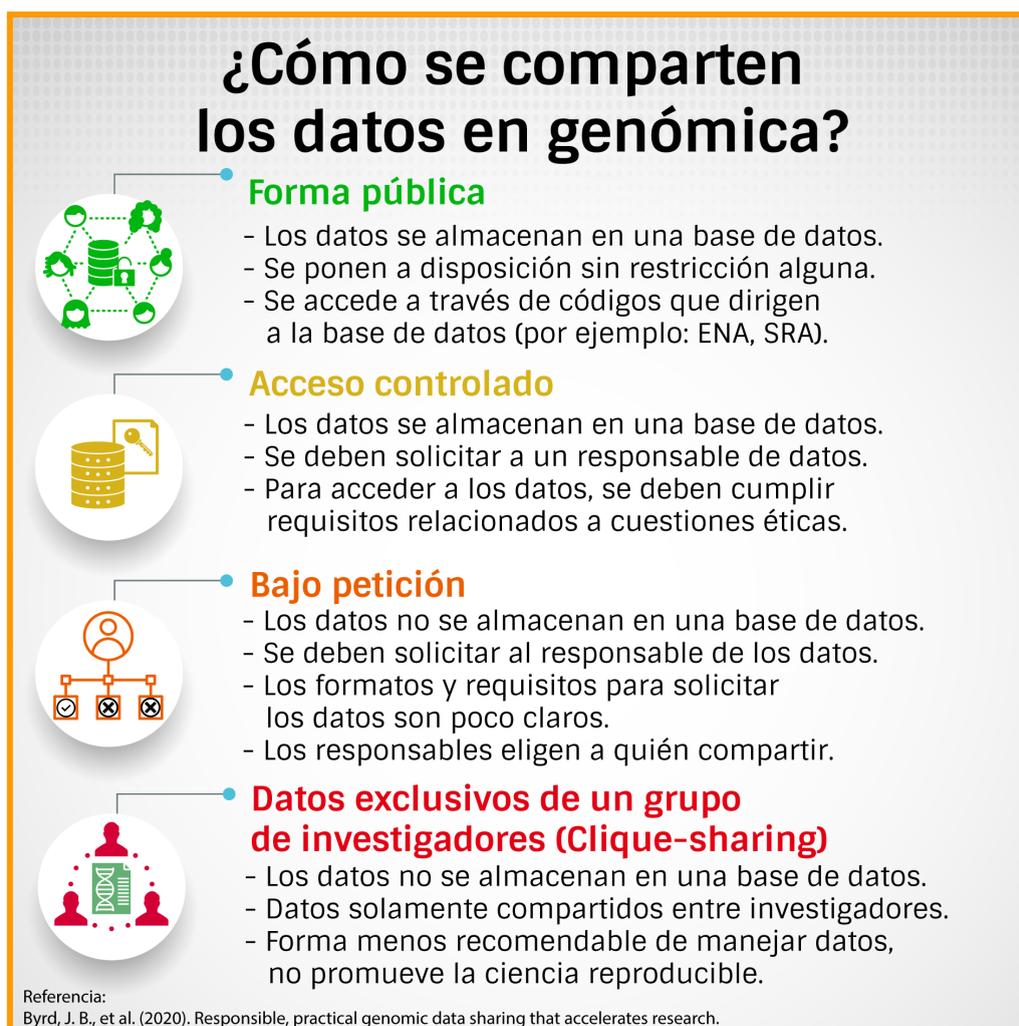
### 2.6.3 *Datos compartidos bajo petición*

Al generarse los datos de un estudio, si los investigadores no planean compartir sus datos de manera pública o controlada (Figura 5), indicarán en los artículos que el

acceso puede ser solicitado mediante una petición directa a uno de los autores que es responsable del resguardo de los datos. Sin embargo, suele suceder que la forma en que se deben solicitar los datos no es clara, complicando el acceso y el aprovechamiento de los datos. En esta forma de compartir datos, los responsables podrán seleccionar qué solicitantes pueden aprovechar los datos (Byrd et al., 2020).

#### *2.6.4 Datos compartidos exclusivamente en un grupo de investigadores*

Cuando los datos generados no son almacenados en una base de datos ni compartidos de forma pública, controlada o bajo petición, suelen estancarse en un nicho de investigadores (Figura 5), los cuales entre ellos tendrán exclusivamente el acceso a los datos (Byrd et al., 2020). Tanto en esta forma de compartir datos como en la forma de petición previa para acceder a ellos, las decisiones sobre la disponibilidad de datos recaerá exclusivamente en los científicos responsables de los datos. Son formas y conductas que ponen en desventaja a los solicitantes de datos y por consecuencia al avance en la investigación genómica.



**Figura 5: Formas de compartir datos en la genómica.** En la figura se muestran las características, ventajas y desventajas de las formas en las que se comparten los datos en genómica. Adaptado de Byrd, J. B., et al (2020).

### 3. MARCO TEÓRICO

Para comprender la disponibilidad de datos genómicos NatAm es necesario conocer el contexto geográfico de la población objetivo, así como distinguir entre los análisis de NatAm contemporáneos y antiguos. A continuación abordaremos estos puntos.

### **3.1 Nativos americanos contemporáneos, o restos antiguos de nativos americanos**

En los estudios revisados para este trabajo se encontraron dos tipos generales de análisis: los estudios de ADN contemporáneo, y los estudios del ADN de restos antiguos (también referidos como ADN antiguo).

#### *3.1.1 ADN de nativos americanos contemporáneos*

Se considera a un estudio de este tipo cuando la muestra de ADN corresponde a individuos pertenecientes a una comunidad o grupo NatAm que está presente actualmente en algún país de América. Incluye individuos que son considerados descendientes de grupos nativos pre-existentes a la conquista europea, así como a descendientes de la migración humana original a las Américas desde Siberia (Bolnick et al., 2016; Morales et al., 2018).

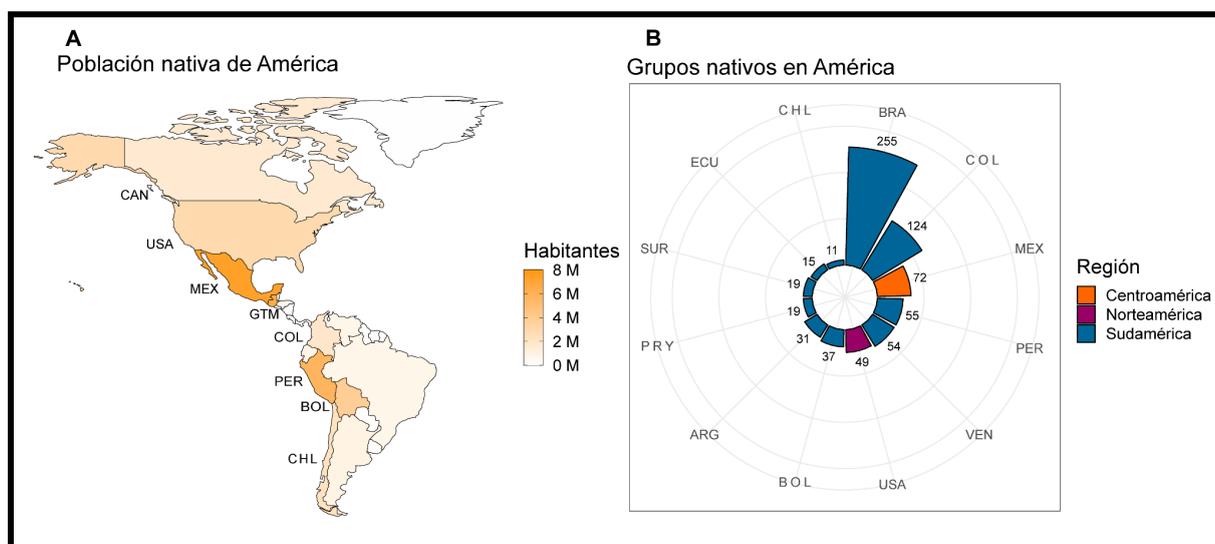
#### *3.1.2 ADN de restos de nativos americanos (o ADN antiguo)*

Se considera a un estudio de este tipo cuando la muestra de ADN fue recuperada a partir de restos antiguos de individuos que pertenecían a grupos NatAm. Usualmente dichas muestras provienen de excavaciones antropológicas. Estos restos son de gran interés en años recientes debido a que se mejoró la tecnología para obtener ADN de buena calidad a partir de muestras antiguas (Skoglund & Reich, 2016; Willerslev & Meltzer, 2021).

### **3.2 Distribución de la población nativa en América**

El continente americano puede clasificarse en dos grandes regiones: América del Norte, y América Latina y el Caribe, con 55 países distribuidos en estas regiones (United Nations et al., 2019). El continente Americano alberga una gran diversidad de grupos NatAm. La información oficial de los censos de los países en los que están presentes los grupos NatAm puede estar actualizada o no, o ser escasa, dependiendo del país. En el presente trabajo hemos recopilado información censal gubernamental y de terceros sobre los grupos contemporáneos NatAm. Hemos contabilizado un estimado de 37,036,134 individuos NatAm distribuidos en 27

países (Figura 6A, tabla suplementaria ST1: <https://github.com/laguilaror/natam-review/blob/main/Tables%20A%20tale%20of%20native%20american%20whole%20genome%20sequencing.xlsx>). Los países con la mayor población NatAm son México (~7,3 millones), Guatemala (~6,5 millones), Perú (~5,7 millones) y Bolivia (~4,1 millones). En cuanto a la diversidad de grupos, Brasil y Colombia albergan más de 100 grupos nativos diferentes cada uno (Figura 6B). Muchos de estos grupos siguen sin estar representados en las bases de datos genómicas.



**Figura 6: Países de origen y diversidad de nativos americanos. (A):** El mapa representa la cantidad de individuos nativos en los países de América. Se destaca con etiquetas a los países con más de 1 millón de individuos nativos. **(B):** El gráfico muestra el número de grupos nativos presentes en algunos países de América. Solo se muestran los países que albergan más de 10 grupos nativos.

### 3.3 ¿En dónde se almacenan los datos genómicos?

Es recomendable que a la hora de compartir datos estos se almacenen en una base de datos bien estructurada, que asegure la seguridad y privacidad de los datos y además, que tenga un mantenimiento sostenido (Byrd et al., 2020). En las ciencias ómicas existen diversas bases de datos para almacenar algún tipo específico de información (genomas, transcriptomas, etc). Para este trabajo consideraremos algunas de las bases de datos más importantes para almacenar información y datos generados por experimentos de secuenciación de genomas, exomas o Arrays.

### 3.3.1 NCBI SRA

La base de datos Sequence Read Archive (SRA) (<https://www.ncbi.nlm.nih.gov/sra/docs/>) pone a disposición pública datos de secuenciación de genoma completo y exoma completo, así como de otras técnicas como RNA-Seq y ChIP-Seq. El repositorio está destinado principalmente a apoyar las investigaciones financiadas por el Instituto Nacional de Salud (NIH) de Estados Unidos; sin embargo, se pueden cargar conjuntos de datos de menos de 1 TB sin coste alguno.

### 3.3.2 EBI ENA

European Nucleotide Archive (ENA) (<https://www.ebi.ac.uk/ena/browser/home>) es una base de datos que también soporta datos públicos de secuenciación, en general los mismos tipos de datos que SRA. Esta base de datos comparte un modelo de espejo con SRA, por lo que los datos públicos cargados se reflejan en ambos sistemas. Si algún proyecto no cuenta con financiación del NIH, ENA puede ser una gran opción como mecanismo para difundir públicamente los datos genómicos.

### 3.3.3 *The international genome sample resource*

La base de datos Internacional de Muestras Genómicas (IGSR) (<https://www.internationalgenome.org/data>) mantiene y comparte públicamente los genomas generados por los proyectos 1000 Genomas, Simons Genome Diversity Project y Human Genome Diversity Project. Es una fuente de información ejemplar, por el hecho de que demuestra que sí se pueden compartir datos genómicos de manera anonimizada, respetuosa y libre.

### 3.3.4 EBI EGA

La base de datos European Genome-phenome Archive (EGA) (<https://ega-archive.org/>) permite almacenar todo tipo de datos genéticos y fenotípicos de identificación personal resultantes de proyectos de investigación biomédica. También permite almacenar datos de secuenciación, pero a diferencia de

las bases SRA y ENA, esta es de acceso controlado. Es una de las opciones más seguras para compartir datos en genómica humana.

### 3.3.5 Otras bases de datos

Las bases de datos mencionadas anteriormente son las más conocidas y usadas a la hora de compartir datos. Se consideran el estándar para el tipo de estudios revisados en este trabajo. Pero también encontramos que existen datos NatAm depositados en bases de datos con objetivos particulares. A continuación las indicamos.

#### 3.3.5.1 Base de datos de David Reich

David Reich es un importante genetista estadounidense reconocido por su investigación sobre la historia evolutiva de las poblaciones humanas. Junto con su equipo de investigación ha creado un repositorio de bases de datos genómicos (microarreglos y secuenciación) donde se albergan principalmente datos de ADN antiguo de diversas partes del mundo (<https://reich.hms.harvard.edu/datasets>).

#### 3.3.5.2 Figshare

Figshare es un repositorio donde los usuarios pueden hacer que todos sus resultados de investigación estén disponibles de una manera citable, compartible y detectable <https://figshare.com/>. No es una base de datos especializada en datos genómicos, sin embargo, sí contiene particularmente algunas muestras de arrays NatAm.

## 4. PLANTEAMIENTO DEL PROBLEMA

Los estudios de genómica poblacional en grupos Nativos de Americanos son escasos, pero se puede aprovechar la disponibilidad pública de los datos generados para agregar la información y plantear investigaciones futuras. La acumulación de datos genómicos es esencial para describir la diversidad genética del continente. Es necesario contar con un compendio que registre todos los estudios de genómica poblacional NatAm realizados a la fecha, y que además proporcione links de descarga a los datos públicamente disponibles. En el presente trabajo nos

planteamos construir dicho compendio para crear una referencia fuerte que promueva futuras investigaciones genómicas en NatAm, con el objetivo final de disminuir la brecha de subrepresentación que existe actualmente.

## **5. PREGUNTA DE INVESTIGACIÓN**

¿Dónde están los datos genómicos de individuos nativos de América generados en estudios de genómica poblacional?

## **6. OBJETIVOS**

### **6.1 Objetivo general**

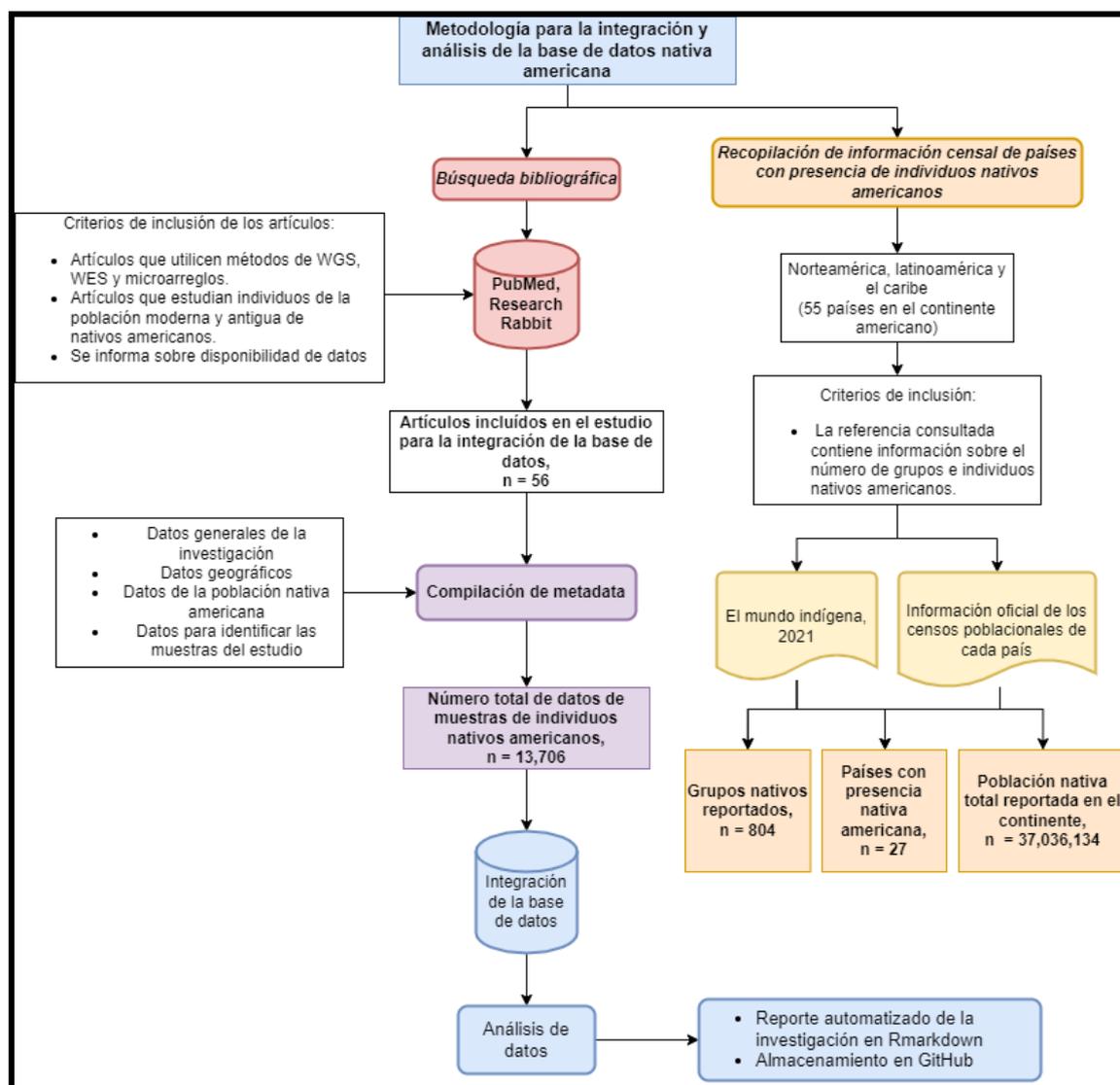
Promover el estudio de la diversidad genómica de Nativos Americanos a través de la recopilación de datos obtenidos por secuenciación de genomas y exomas, así como de Arrays públicamente disponibles.

### **6.2 Objetivos específicos**

1. Identificar los grupos NatAm previamente estudiados mediante alguna tecnología genómica.
2. Describir la disponibilidad de datos en los estudios genómicos de NatAm.
3. Construir un compendio que indique dónde están los datos públicamente disponibles en genómica NatAm.

## **7. MATERIALES Y MÉTODOS**

La metodología de este trabajo se divide en 3 partes, que son búsqueda bibliográfica, Selección de metadatos de los artículos e integración del compilado y Recopilación de información censal de NatAm en los países de América. Al final logramos integrar y analizar una base de datos y metadatos que recopila todos los estudios genómicos NatAm publicados a la fecha.



Esquema 1: Metodología general para la integración de la base de datos de nativos americanos.

### 7.1 Búsqueda bibliográfica.

Se establecieron los términos de búsqueda pertinentes para encontrar artículos que cumplieran con los objetivos del presente trabajo. Los motores de búsqueda de literatura y los criterios de selección se describen a continuación.

### *7.1.1 Criterios de inclusión de los artículos revisados en este trabajo.*

#### *7.1.1.1 Tecnología utilizada*

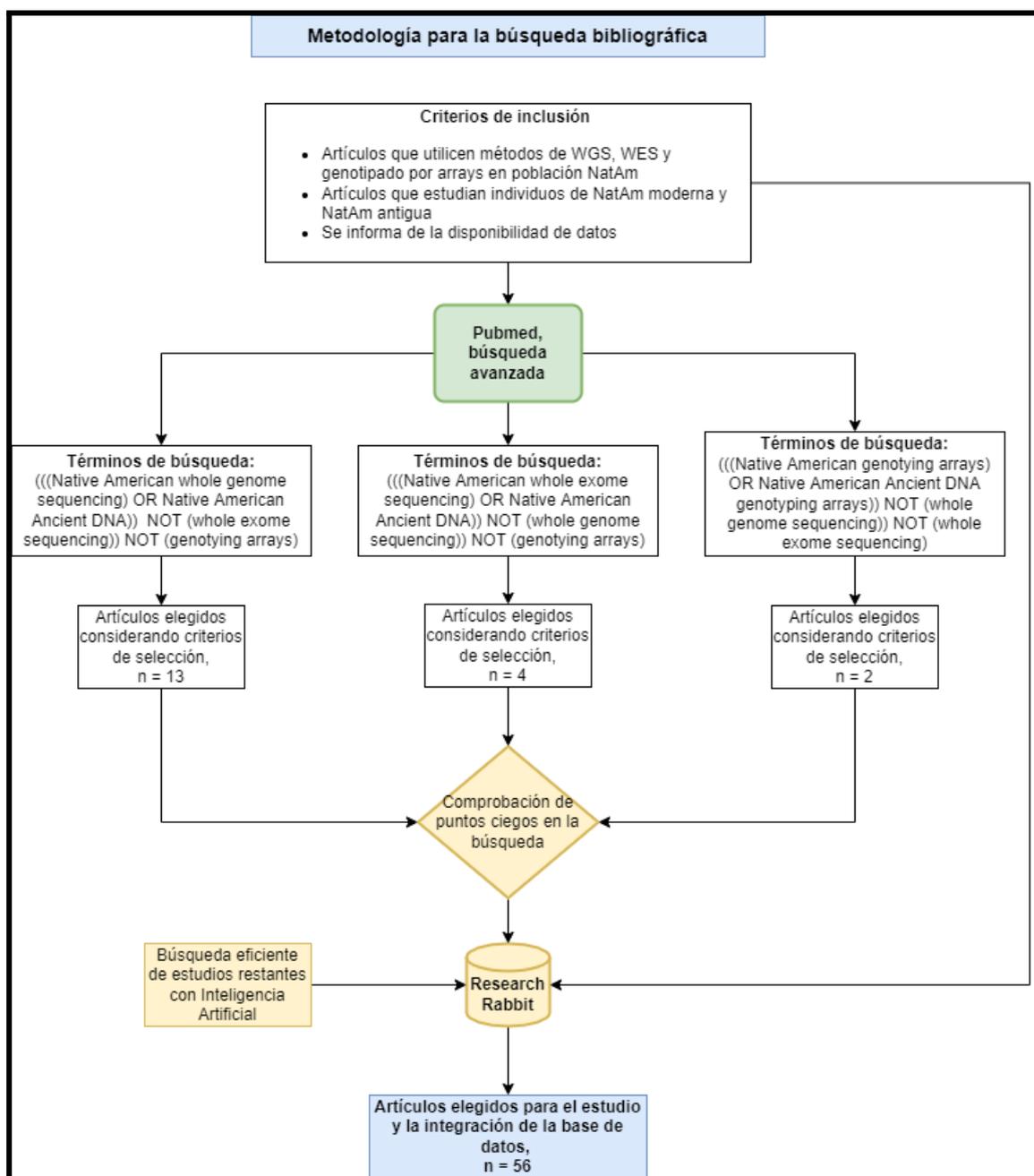
Este criterio hace referencia a que los artículos debían especificar en su metodología que utilizaron **WGS** (secuenciación de genoma completo), **WES** (secuenciación de exoma completo) y/o **Arrays** (genotipificación).

#### *7.1.1.2 Tipo de ADN estudiado*

Este criterio establece dos distinciones. **(I): ADN moderno de NatAm** – ADN de NatAm perteneciente a una comunidad nativa moderna, descendiente de grupos nativos anteriores a la conquista, o descendiente de la migración humana original a las Américas desde Siberia (Bolnick et al., 2016; Morales et al., 2018). **(II): ADN antiguo de NatAm** – ADN recuperado y analizado de restos antiguos de NatAm, que pertenecían a grupos nativos de América del Norte, Central y del Sur (Skoglund & Reich, 2016; Willerslev & Meltzer, 2021).

#### *7.1.1.3 Disponibilidad de datos:*

Este criterio establece tres distinciones: **(I): Disponibilidad pública:** Artículos que presentan una sección de disponibilidad de datos y especifican que todas las muestras generadas están disponibles en una base de datos (pueden incluir el código de acceso o el enlace a la base de datos). Al integrar la información de los artículos con este criterio en la base de datos, se clasificaron como “Sí”. **(II): Disponibilidad no pública:** Artículos que especifican en su sección de disponibilidad de datos que todas las muestras generadas no están disponibles en ninguna base de datos o no tienen una sección de disponibilidad de datos para los individuos nativos americanos estudiados. Al integrar la información de los artículos con este criterio en la base de datos, se clasificaron como “No”. **(III): Disponibilidad parcial:** Artículos que especifican en su sección de disponibilidad de datos que algunas muestras están disponibles y otras no.



Esquema 2: Metodología para la búsqueda bibliográfica de artículos.

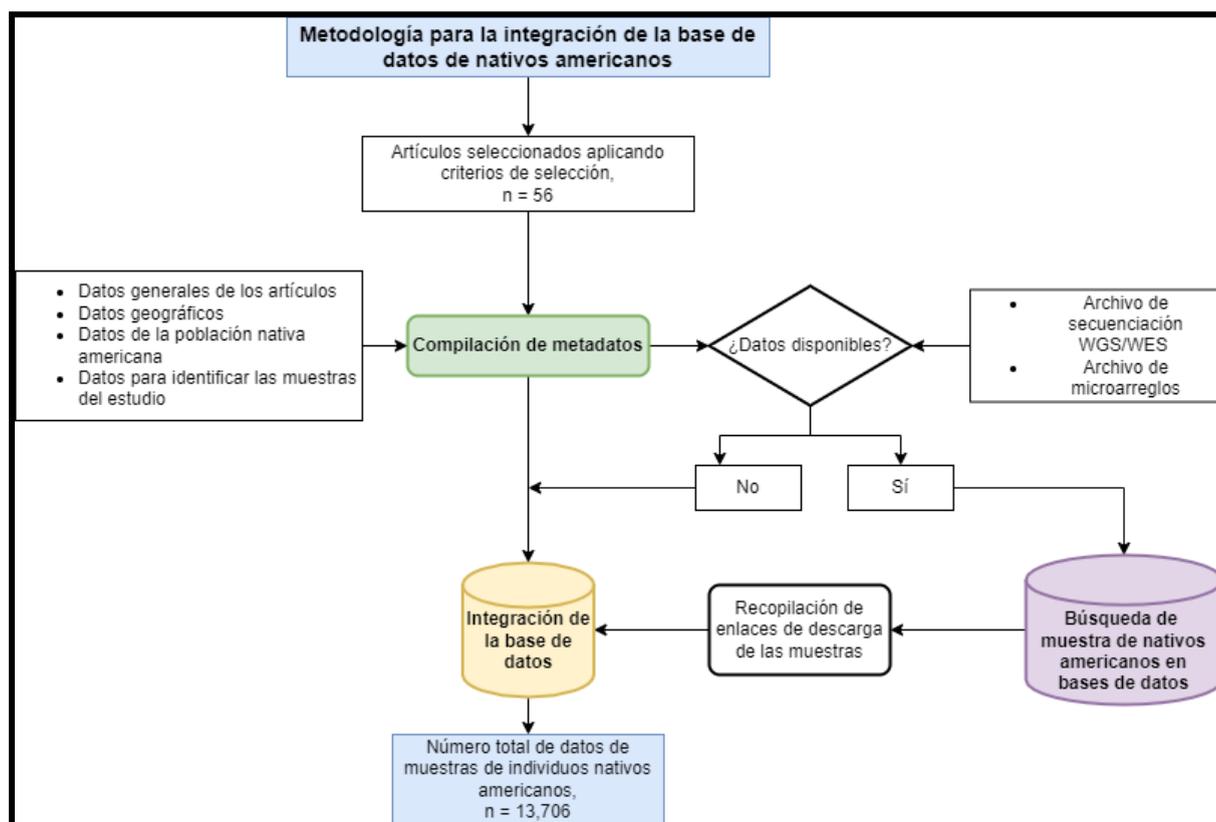
## 7.2 Selección de metadatos de los artículos e integración del compilado.

Una vez recopilados los artículos a revisar se definieron los metadatos necesarios para integrar la base de datos final de este trabajo. Estos datos se clasificaron en 4 partes: **(I): Datos generales del artículo** – Título del artículo, apellido del primer autor, año de publicación, Digital Object Identifier (DOI) del artículo, afiliaciones de los autores restantes y países de las afiliaciones. **(II) Datos geográficos** – De estar

disponibles, indican el origen aproximado de la toma de muestra, en coordenadas (lat, lon). **(III) Datos de los grupos NatAm estudiados** – Grupo NatAm estudiado, código de identificación, sexo de los individuos. **(IV) Datos para identificar las muestras del estudio y su disponibilidad** – Tipo de ADN estudiado (moderno, antiguo), tecnología genómica usada (WGS, WES, Arrays), ID original de la muestra, disponibilidad de los datos, link a base de datos, archivo disponible, link de descarga de datos genómicos.

### 7.2.1 Integración de la base de datos

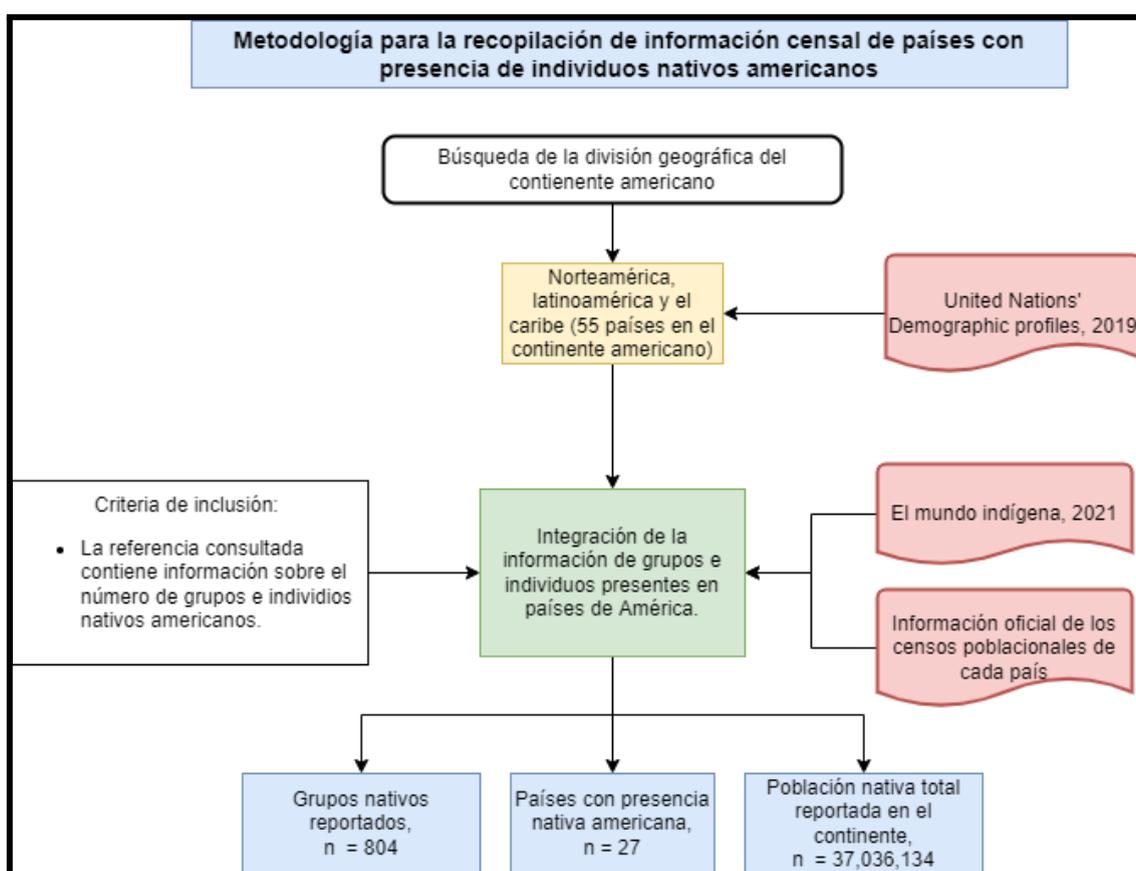
Los datos se integraron en archivos de formato tabular para su análisis usando el lenguaje R, y la interfaz gráfica RStudio, toda la información está almacenada y disponible en GitHub (<https://github.com/laguilaror/natam-review>) para garantizar su reproducibilidad.



Esquema 3: Metodología para la recopilación de metadata y la integración de la base de datos.

### 7.3 Recopilación de información censal de NatAm en los países de América.

Uno de los objetivos de este trabajo es conocer qué grupos NatAm han sido representados en estudios de secuenciación y arrays. Esto nos permitirá comparar con la información censal de los países hogar de NatAm y saber qué otros grupos no han sido representados en cada país. Para ello, recopilamos la información censal de poblaciones NatAm en cada país de América. En caso de no estar disponible la información oficial gubernamental, se buscó una referencia de terceros.



Esquema 4: Metodología para la recopilación de censos poblacionales de América.

## 8. RESULTADOS

El resultado más importante de este trabajo es la integración de una base de datos que contiene información sobre 56 estudios genómicos de nativos americanos realizados desde 2009 hasta 2021. Un resumen de los estudios se muestra en la

Tabla 1. En las siguientes secciones describiremos algunos hallazgos interesantes derivados del análisis de los metadatos, tales como la temporalidad de los estudios, la disponibilidad de los datos, así como la participación de países latinoamericanos en este tipo de estudios.

**Tabla 1. Estudios genómicos de NatAm recopilados.**

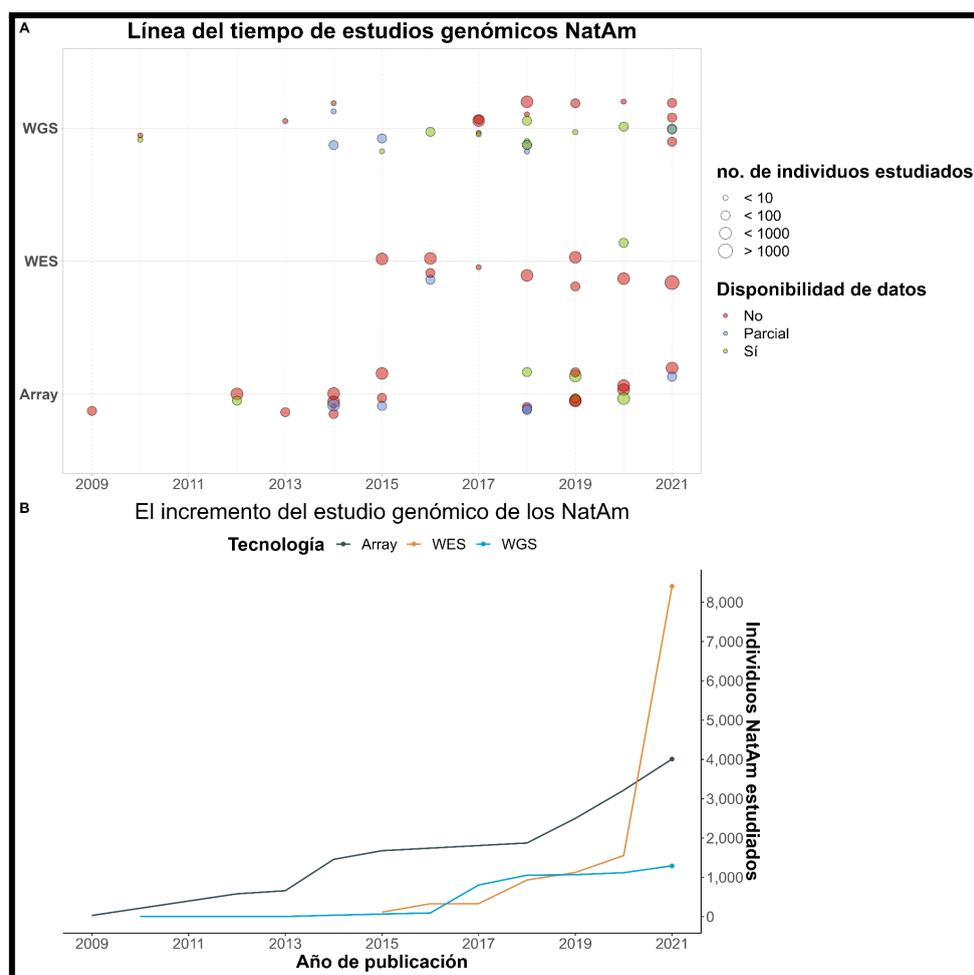
Estudio	Individuos estudiados	Referencia
Analysis of genomic diversity in Mexican Mestizo populations to develop genomic medicine in Mexico	30	(Silva-Zolezzi et al., 2009)
Genetic history of an archaic hominin group from Denisova Cave in Siberia	1	(Reich et al., 2010)
Ancient human genome sequence of an extinct Palaeo-Eskimo	1	(M. Rasmussen et al., 2010)
Reconstructing Native American population history	493	(Reich et al., 2012)
Ancient Admixture in Human History	57	(Patterson et al., 2012)
High-Throughput Sequencing of a South American Amerindian	1	(Ribeiro-dos-Santos et al., 2013)
Reconstructing the Population Genetic History of the Caribbean	79	(Moreno-Estrada et al., 2013)
Two ancient human genomes reveal Polynesian ancestry among the indigenous Botocudos of Brazil	3	(Malaspinas et al., 2014)
The genetic prehistory of the New World Arctic	29	(Raghavan et al., 2014)
The genome of a Late Pleistocene human from a Clovis burial site in western Montana	2	(M. Rasmussen et al., 2014)
Patterns of Admixture and Population Structure in Native Populations of Northwest North America	104	(Verdu et al., 2014)
The genetics of Mexico recapitulates Native American substructure and affects biomedical traits	511	(Moreno-Estrada et al., 2014)
Genome-wide Ancestry Patterns in Rapanui Suggest Pre-European Admixture with Native Americans	27	(Moreno-Mayar et al., 2014)
Ancient human genomes suggest three ancestral populations for present-day Europeans	155	(Lazaridis et al., 2014)
Increased Missense Mutation Burden of Fatty Acid Metabolism Related Genes in Nunavik Inuit Population	226	(Zhou et al., 2015)
The ancestry and affiliations of Kennewick Man	1	(M. F. Rasmussen et al., 2015)
Genomic evidence for the Pleistocene and recent population history of Native Americans	87	(Raghavan et al., 2015)
Genetic evidence for two founding populations of the Americas	48	(Skoglund et al., 2015)
A time transect of exomes from a Native American population before and after European contact	50	(Lindo et al., 2016)
Toward precision medicine: TBC1D4 disruption is common among the Inuit and leads to underdiagnosis of type 2 diabetes	114	(Manousaki et al., 2016)
Distribution and clinical impact of functional variants in 50,726 whole-exome sequences from the DiscovEHR study	51	(Dewey et al., 2016)
The Simons Genome Diversity Project: 300 genomes from 142 diverse populations	28	(Mallick et al., 2016)
Whole genome sequence association and ancestry-informed polygenic profile of	692	(Peng et al., 2017)

EEG alpha in a Native American population		
Demographic history and biologically relevant genetic variation of Native Mexicans inferred from whole-genome sequencing	15	(Romero-Hidalgo et al., 2017)
Ancient individuals from the North American Northwest Coast reveal 10,000 years of regional genetic continuity	3	(Lindo et al., 2017)
Genetic Ancestry of Rapanui before and after European Contact	5	(Fehren-Schmitz et al., 2017)
Whole-exome sequencing in maya indigenous families: variant in PPP1R3A is associated with type 2 diabetes	600	(Sánchez-Pozos et al., 2018)
Terminal Pleistocene Alaskan genome reveals first founding population of Native Americans	2	(Moreno-Mayar, Potter, et al., 2018)
Evolutionary genomic dynamics of Peruvians before, during, and after the Inca Empire	150	(Harris et al., 2018)
The genetic prehistory of the Andean highlands 7000 years BP though European contact	71	(Lindo et al., 2018)
Origins and genetic legacies of the Caribbean Taino	1	(Schroeder et al., 2018)
Genomic insights into the origin and diversification of late maritime hunter-gatherers from the Chilean Patagonia	65	(de la Fuente et al., 2018)
Ancient human parallel lineages within North America contributed to a coastal expansion	91	(Scheib et al., 2018)
Early human dispersals within the Americas	16	(Moreno-Mayar, Vinner, et al., 2018)
Reconstructing the Deep Population History of Central and South America	49	(Posth et al., 2018)
Population history and gene divergence in Native Mexicans inferred from 76 human exomes	78	(Ávila-Arcos et al., 2020)
Genetic architecture and adaptations of Nunavik Inuit	279	(Zhou et al., 2019)
Whole Genome Sequence, Variant Discovery and Annotation in Mapuche-Huilliche Native South Americans	11	(Vidal et al., 2019)
Palaeo-Eskimo genetic ancestry and the peopling of Chukotka and North America	14	(Flegontov et al., 2019)
Dissecting the Pre-Columbian Genomic Ancestry of Native Americans along the Andes–Amazonia Divide	229	(Gnecchi-Ruscone et al., 2019)
Comparing signals of natural selection between three Indigenous North American populations	150	(Reynolds et al., 2019)
The Current Genomic Landscape of Western South America: Andes, Amazonia, and Pacific Coast	70	(Barbieri et al., 2019)
Exome Sequencing of Native Populations From the Amazon Reveals Patterns on the Peopling of South America	58	(Ribeiro-dos-Santos et al., 2020)
Exome sequencing identifies a nonsense variant in DAO associated with reduced energy expenditure in American Indians	373	(Piaggi et al., 2020)
Insights into human genetic variation and population history from 929 diverse genomes	46	(Bergström et al., 2020)
Integration of ancient DNA with transdisciplinary dataset finds strong support for Inca resettlement in the south Peruvian coast	6	(Bongers et al., 2020)
Native American gene flow into Polynesia predating Easter Island settlement	227	(Ioannidis et al., 2020)
A genetic history of the pre-contact Caribbean	174	(Fernandes et al., 2021)
The genetic structure and adaptation of Andean highlanders and Amazonians are influenced by the interplay between geography and culture	316	(Borda et al., 2020)
Exome Sequencing of 21 Bardet-Biedl Syndrome (BBS) Genes to Identify Obesity Variants in 6,851 American Indians	6851	(Day et al., 2021)

Whole genome variation in 27 Mexican indigenous populations, demographic and biomedical insights	76	(Aguilar-Ordoñez et al., 2021)
Indigenous Ancestry and Admixture in the Uruguayan Population	10	(Spangenberg et al., 2021)
The genomic landscape of Mexican Indigenous populations brings insights into the peopling of the Americas	716	(García-Ortiz et al., 2021)
Ancient genomes reveal long-range influence of the pre-Columbian culture and site of Tiwanaku	17	(Popović et al., 2021)
Archaeogenomic distinctiveness of the Isthmo-Colombian area	97	(Capodiferro et al., 2021)
Imputation Performance in Latin American Populations: Improving Rare Variants Representation With the Inclusion of Native American Genomes	50	(Jiménez-Kaufmann et al., 2022)

### 8.1 Análisis general de estudios genómicos NatAm

Desde el primer proyecto publicado en 2009, el número de trabajos en genómica NatAm ha aumentado de forma constante hasta un total de 56 proyectos a principios de 2022 (Figura 7A). Este es un buen panorama para continuar cerrando la brecha de los grupos subrepresentados en la genómica moderna. De todos los proyectos que se revisaron en este trabajo, 26 han incluido individuos NatAm en estudios basados en genotipado por Arrays, mientras que 11 proyectos utilizaron WES, y 29 utilizaron WGS. Las tres tecnologías genómicas se siguen utilizando en los proyectos más recientes. Hay un número creciente de proyectos WGS que son publicados; sin embargo, las publicaciones que usan tecnologías WES y Arrays suelen implicar una mayor cantidad de individuos. En general, el aumento del número y la frecuencia de los trabajos ha venido acompañado de un aumento en el número de individuos estudiados. Hasta la fecha se han estudiado un total de 13,706 individuos NatAm con estas tecnologías genómicas: 4,009 con Arrays, 8,405 con WES y 1,292 con WGS (Figura 7B, tabla suplementaria ST3: <https://github.com/laquilaror/natam-review/blob/main/Tables%20A%20table%20of%20native%20american%20whole%20genome%20sequencing.xlsx>). Si nos centramos en los estudios WGS (la tecnología que proporciona una visión más amplia de la genómica poblacional) la marca de 1,000 genomas completos estudiados se alcanzó en 2018. Esto significa que hay aproximadamente más de 3,000,000,000,000 de nucleótidos estudiados ya publicados, pintando el cuadro de la diversidad y la historia del continente americano.

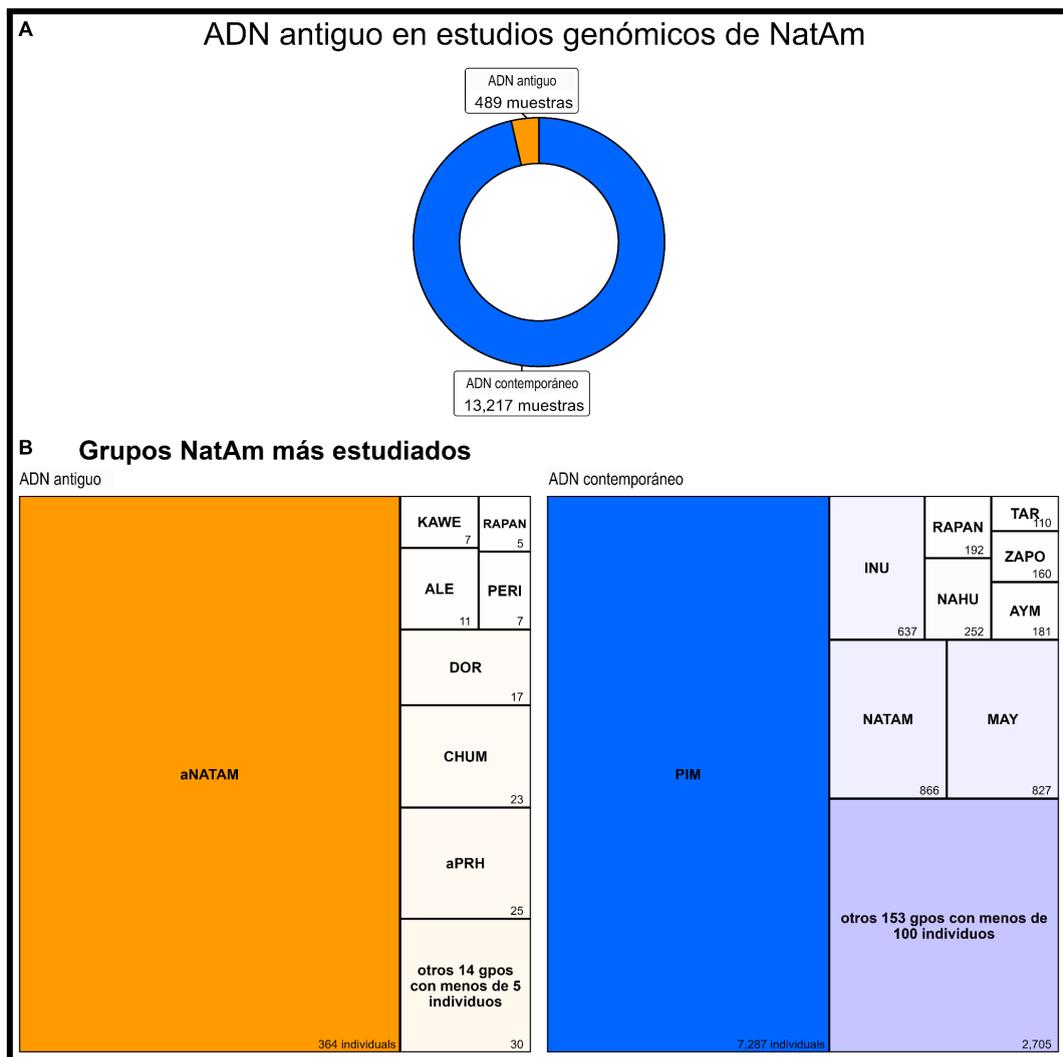


**Figura 7: Los estudios genómicos de nativos americanos a través de los años. (A):** Línea del tiempo que muestra los proyectos recopilados en este trabajo clasificados por tecnología genómica. El tamaño de los círculos indica el número de muestras generadas de individuos nativos en los estudios, (desde menos de 10 hasta más de 1,000). La variación del color de los círculos indica la disponibilidad de los datos del estudio (verde: Disponibles, Azul: Parcial, Rojo: No Disponibles). **(B):** Incremento del número de muestras de individuos nativos acumuladas en los estudios genómicos.

## 8.2 Los grupos NatAm estudiados

Hasta la fecha 489 de los individuos estudiados son muestras antiguas, y 13,217 son contemporáneas (Figura 8A). En total 178 grupos NatAm (contemporáneos y antiguos) han sido representados en estudios genómicos de poblaciones y 607 grupos NatAm siguen sin estar representados en el archivo genómico colectivo (Figura 9, tabla suplementaria ST5:

<https://github.com/laguilaror/natam-review/blob/main/Tables%20A%20tale%20of%20native%20american%20whole%20genome%20sequencing.xlsx>). Se ha utilizado la tecnología de arrays para estudiar 152 grupos NatAm, 19 grupos han sido estudiados mediante WES, y la WGS se ha implementado para analizar 70 grupos. Los grupos Pima, Maya, Inuit, Nahua, Rapa Nui y Aymara se encuentran entre las poblaciones NatAm contemporáneas más estudiadas, mientras que los Chumash, los Dorset, los Aleutianos, los Pericúes, los Rapa Nui y los Kawéskar se encuentran entre los más estudiados por análisis de ADN antiguo (Figura 8B).

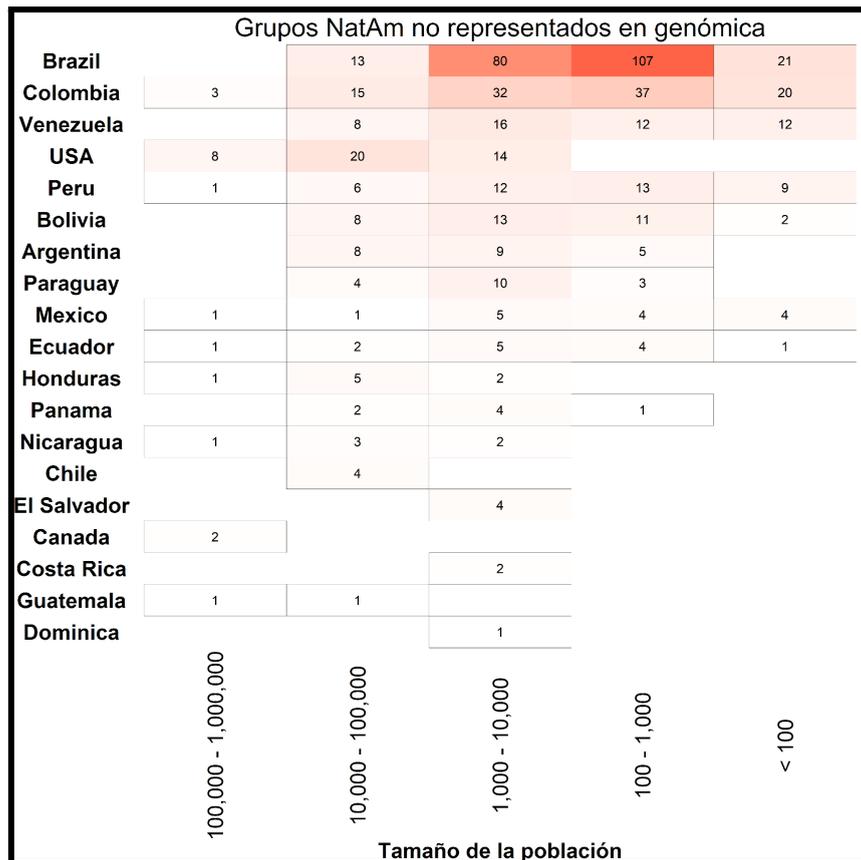


**Figura 8: Vista general de los grupos NatAm representados en genómica. (A):** Las muestras antiguas son una fracción pequeña pero importante del total de muestras genómicas de los NatAm. **(B):** Se muestran los grupos NatAm que han sido estudiados a en genómica; Pima (PIM), Maya (MAY) e Inuit (INU) son algunos de los grupos NatAm contemporáneos más estudiados, mientras que Chumash (CHUM), Dorset (DOR) y Aleutiano (ALE) se encuentran entre los más representados en

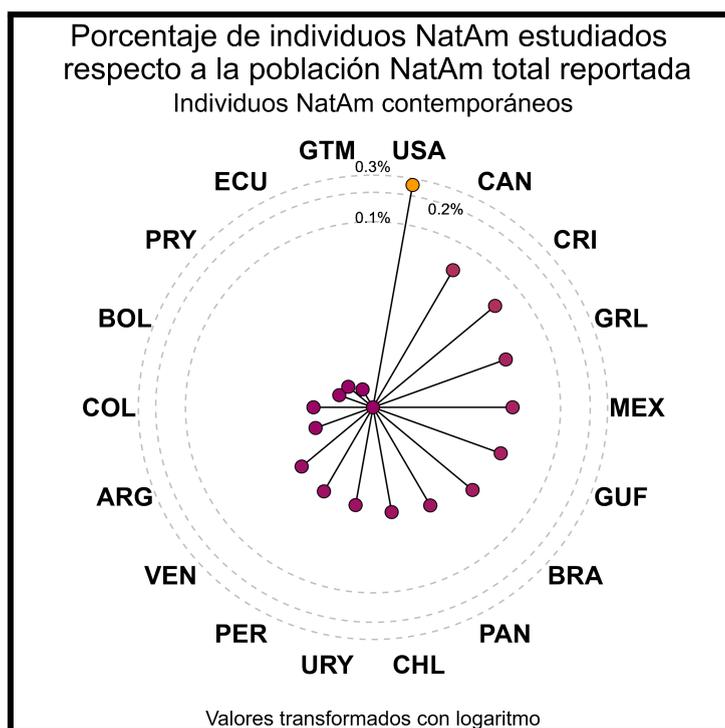
los conjuntos de datos antiguos. Los códigos homologados y el nombre completo de los grupos NatAm estudiados en genómica pueden encontrarse en la tabla S1 en la sección Material Suplementario.

### *8.2.1 Comparación de la población NatAm representada en genómica contra censos de población NatAm*

En este trabajo hemos recopilado datos sobre los censos de los diferentes países que albergan poblaciones NatAm contemporáneas (tabla suplementaria ST4: <https://github.com/laguilaror/natam-review/blob/main/Tables%20A%20tale%20of%20native%20american%20whole%20genome%20sequencing.xlsx>). En 27 países hay presencia de grupos nativos americanos, sin embargo solo en 18 países hay individuos representados en estudios genómicos. Al comparar los datos de los censos con los estudios genómicos publicados, encontramos que actualmente menos del 0.3% de los individuos NatAm en cada país americano ha sido representada (Figura 10). Como hemos insistido antes, la WGS proporciona el contexto genómico completo de la población estudiada, por lo que los futuros esfuerzos genómicos podrían beneficiarse enormemente del análisis de genomas completos en estas poblaciones NatAm que faltan.



**Figura 9: Grupos NatAm no representados en genómica.** El mapa de calor muestra el número de grupos nativos que no están representados en estudios genómicos; esta información se obtuvo al comparar datos de censos nacionales con los grupos reportados en los estudios genómicos publicados.

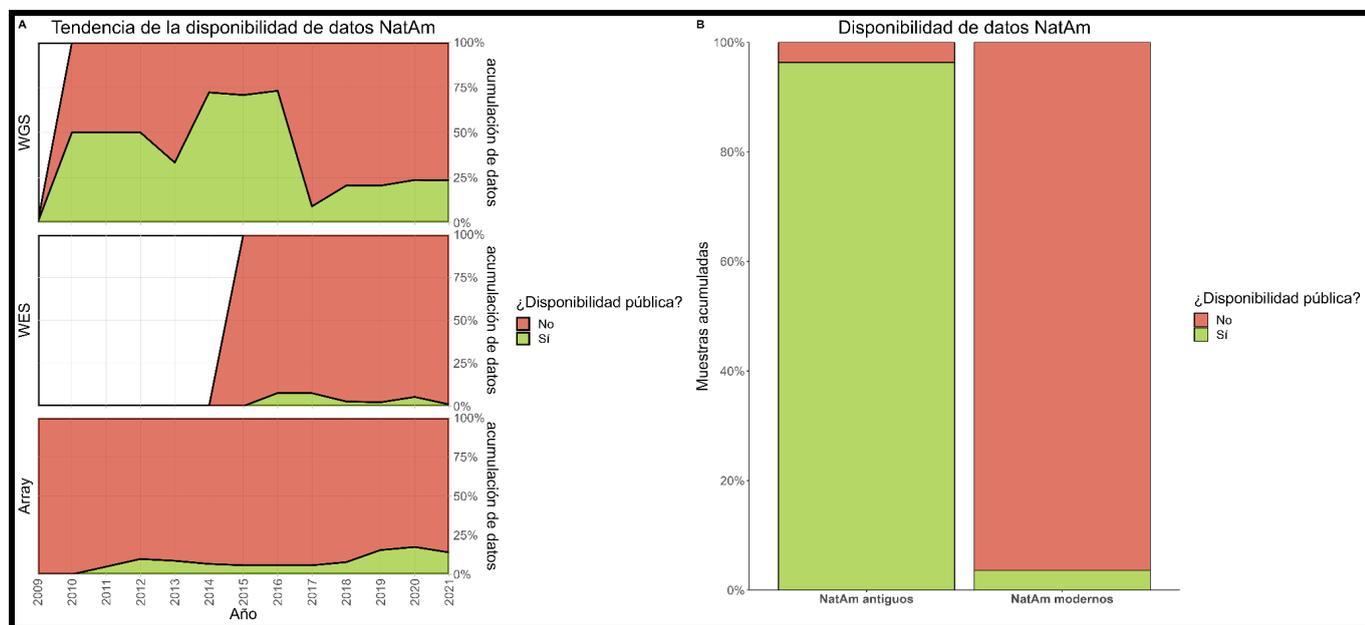


**Figura 10: Comparación de individuos NatAm representados en genómica con censos poblacionales.** La figura muestra por país, la cantidad porcentual del total de individuos nativos que han sido representados en genómica. En la mayoría de países con presencia nativa, menos del 0.1% total de individuos nativos han sido estudiados. Los valores se transformaron con logaritmo base 10 para lograr una mejor representación visual.

### 8.3 Disponibilidad de datos

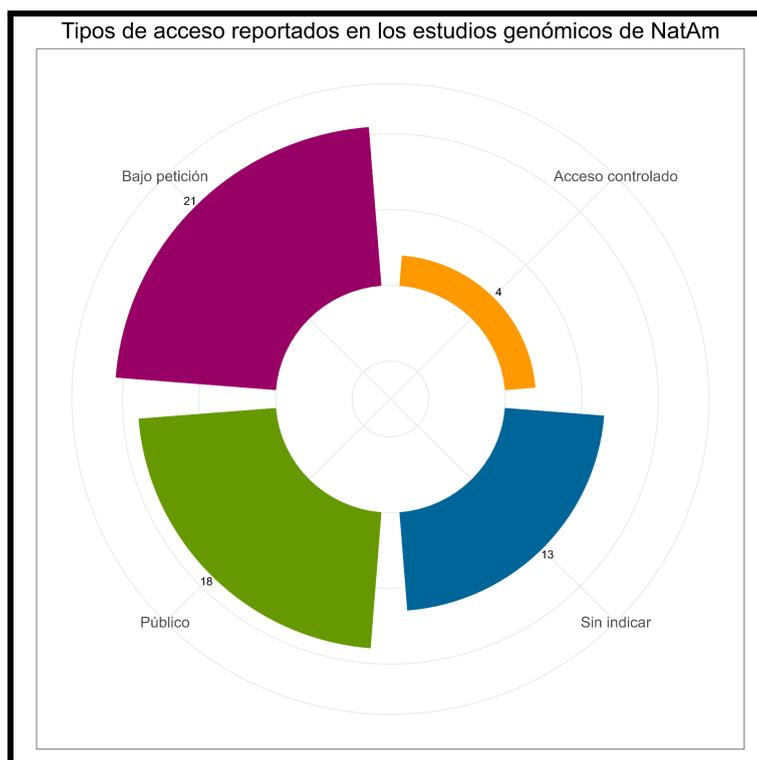
La genómica de NatAm no proporciona una fuente de datos accesible, tal como muestra nuestro análisis de la disponibilidad de datos en los 56 trabajos revisados. Recopilamos los enlaces de descarga proporcionados para la descarga de datos en bases de datos públicas a través de los artículos científicos publicados y etiquetamos cada muestra como disponible o no disponible de manera independiente. Una muestra "disponible" proporciona información directa o enlaces para la descarga de archivos de secuenciación a través de repositorios establecidos como las bases Sequence Read Archive, European Nucleotide Archive, o similares, mientras que una muestra "no disponible" no proporciona un enlace de descarga o información directa y clara sobre cómo acceder a los datos. Las muestras "disponibles bajo petición" se marcaron como "no disponibles". Analizamos la cuota acumulada de individuos estudiados y encontramos que hasta 2021 sólo el 14 % de

las muestras de Array están disponibles, para WES sólo hay un 0.9 %, y para WGS la cuota disponible es del 23.4 % (Figura 11A). Es interesante notar que el 96.3% de las muestras antiguas proporcionan datos directamente disponibles, en contraste con sólo el 3.6% de las muestras modernas (Figura 11B).



**Figura 11: La disponibilidad de datos de los estudios de NatAm. (A):** La dinámica de la disponibilidad muestra la apertura de los datos genómicos NatAm según la tecnología utilizada (Array, WES o WGS). **(B):** El gráfico muestra que el área de ADN antiguo tiene más apertura a compartir los datos que los estudios de ADN contemporáneo.

La forma en que los investigadores comparten datos en los estudios genómicos NatAm no es la más adecuada. Analizamos el tipo de acceso a los datos indicado en los artículos revisados (Figura 12), y encontramos que 21 están bajo el formato de acceso bajo petición, 18 son públicos (sin embargo 3 estudios presentan enlaces de acceso no funcionales), 4 se rigen bajo el acceso controlado y 13 estudios no especifican la disponibilidad de datos.



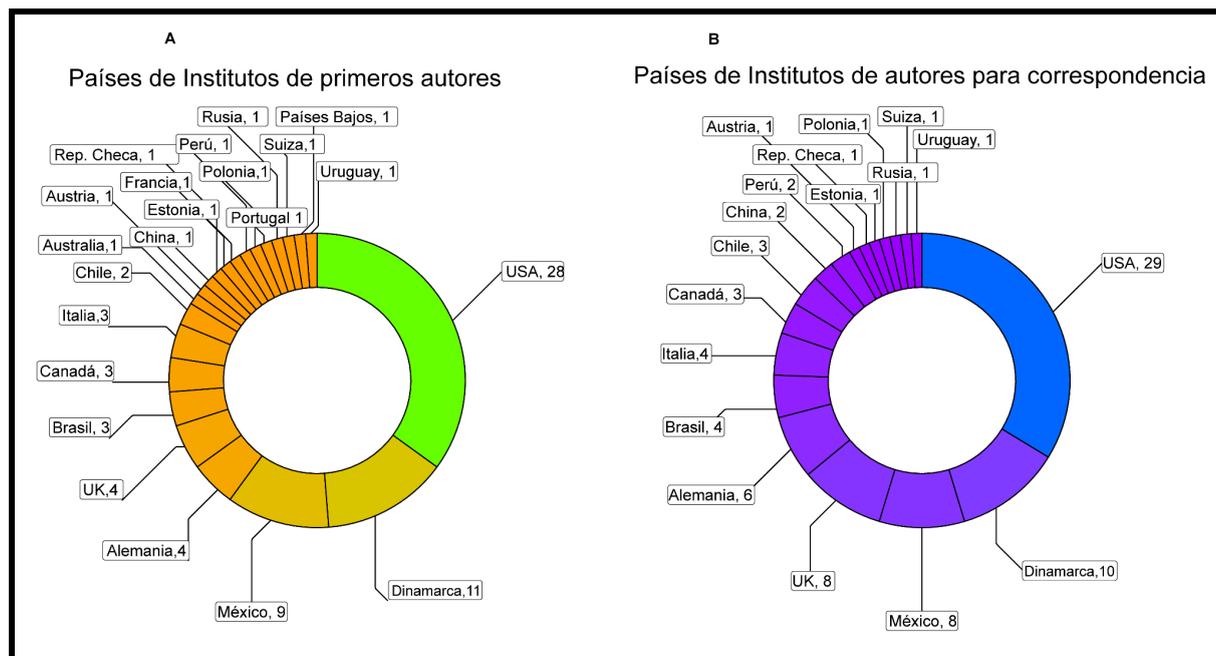
**Figura 12: Tipos de Acceso reportados en los estudios de NatAm.** El tipo de acceso que predomina en los artículos de investigación genómica de NatAm es la modalidad “Bajo petición”, la cual presenta formatos poco claros para realizar la solicitud. Solamente 18 artículos tienen un tipo de acceso público (3 presentan enlaces de acceso no funcionales), 4 artículos presentan un tipo de acceso controlado y el resto de artículos no indica ninguna forma de solicitud de acceso.

#### 8.4 El interés en la investigación genómica nativa americana

Analizamos el país de origen de cada institución de investigación que participó en los proyectos revisados, y encontramos que Estados Unidos, Dinamarca y México son los países que lideran los proyectos revisados (Figura 13, tabla suplementaria ST6:

<https://github.com/laguilaror/natam-review/blob/main/Tables%20A%20tale%20of%20native%20american%20whole%20genome%20sequencing.xlsx>). Si comparamos la

Figura 6 con la Figura 13, resulta evidente que muchos de los países de origen de los grupos NatAm que faltan también carecen de representación como líderes en los proyectos de investigación.



**Figura 13. Participación de países en la investigación genómica de la población. (A):** Se muestran los países sede para los institutos de los primeros autores. **(B):** Lo mismo pero para autores de correspondencia. En ambos gráficos se observa que hacen falta más representación de los países de origen de los grupos NatAm.

## 9. DISCUSIÓN

En los últimos años han incrementado los proyectos genómicos que incluyen individuos NatAm (Figura 7, con lo cual se busca disminuir la brecha de los grupos subrepresentados en la genómica moderna (Sirugo et al., 2019). La continuación de este tipo de estudios permitirá describir la variación genética nativa, un aspecto de gran relevancia regional. La genómica de poblaciones y los descubrimientos de grandes y diversos esfuerzos de secuenciación indican que la gran mayoría de la variación genética humana es rara y se espera que sea específica de las poblaciones regionales (Gravel et al., 2011). Por último, es importante considerar los beneficios que la investigación genómica NatAm puede aportar en diversas disciplinas científicas como lo es la medicina genómica, farmacogenómica, antropología molecular, genética de poblaciones y la biotecnología (Bolnick et al., 2016).

La integración y el análisis de la base de datos presentada en este trabajo permitió identificar el estado actual de la investigación genómica de la población NatAm. 56 estudios fueron analizados, y encontramos resultados que son preocupantes, pues hasta la fecha menos del 0.3% de la población NatAm ha sido estudiada. Aún queda muchísima información genética por conocer. Algunos de los grupos que ya han sido estudiados con tecnologías genómicas necesitan más representación, preferentemente mediante secuenciación de genoma completo para tener el panorama completo; tal es el caso de los grupos Huichol, Tojolabal, Lacandón, Seri, que son algunos de los grupos NatAm más distintivos genéticamente (Aguilar-Ordoñez et al., 2021; Moreno-Estrada et al., 2014). En general, el estudio genómico profundo de los grupos NatAm permitirá establecer nuevas hipótesis, contribuir con nuevo conocimiento genético y esclarecer más la historia evolutiva de las poblaciones NatAm (Jiménez-Kaufmann et al., 2022).

Si bien los puntos anteriores resaltan la importancia de promover la investigación genómica en NatAm, es importante reconocer las limitaciones en estos estudios. Un punto importante es analizar cómo se podría involucrar más a los países del continente que aún no hayan sido parte de estos estudios, ya que como observamos en la figura 13, son pocos los que han contribuido a la realización de estudios genómicos NatAm. La implementación de la medicina genética y genómica depende en gran medida de la capacitación exitosa de una fuerza laboral en la genómica con experiencia en la interpretación, comunicación e integración de la información genómica aplicado a entornos con impacto social, por ejemplo el entorno clínico-genómico (Sirisena & Dissanayake, 2019). Sin embargo, los países en desarrollo en América, al no mejorar su capacidad de investigación y desarrollo en este campo, han permanecido excluidos de la revolución genómica y sus beneficios. En muchos de estos países el poco avance en genómica se debe a los numerosos desafíos derivados de la falta de capacidad local para planificar y llevar a cabo la capacitación requerida para proyectos genómicos de gran escala (Hetu et al., 2019). Los factores que contribuyen a esta situación son la escasez de fondos económicos adecuados, así como el establecimiento de las instalaciones tecnológicas necesarias para la prestación de servicios genómicos en torno a los

cuales dichos programas de capacitación podrían implementarse (Sirisena & Dissanayake, 2019). Para tratar de superar estos desafíos se requiere de la acción conjunta del sector gubernamental, académico e incluso privado. Acciones puntuales que pueden promover el desarrollo de la genómica son el incremento de presupuesto económico para la investigación, el establecimiento de centros especializados en análisis genómico, el fomento a las colaboraciones internacionales, el acceso a datos y publicaciones genómicas actualizadas y de relevancia, el desarrollo de métodos experimentales (Helmy et al., 2016) y por último, la creación de programas de formación en genómica, bioinformática y ciencia de datos, pues sin las habilidades de estas áreas, es prácticamente imposible promover la investigación genómica realizada a nivel regional (Mangul et al., 2019).

Al analizar la tendencia en la disponibilidad de datos en los 56 estudios recopilados por tecnología genómica, observamos que WGS tiene el porcentaje más alto (23.4%) de disponibilidad, seguido de Arrays (14%) y WES (0.9%) (Figura 2). Además, al considerar la forma en que se comparten los datos de estos estudios, 21 se rigen bajo la modalidad de compartir datos bajo petición, y presentan formatos poco claros para solicitar el acceso a la información. Los pocos detalles que se proporcionan para elaborar la petición a los datos varían, desde contactar al responsable de los datos, hasta no presentar ningún detalle. Solo un investigador responsable ha cambiado de adscripción, por lo que, en general, hasta el momento la información genómica sigue bajo la responsabilidad y exclusividad de investigadores e institutos de investigación fijos. Como ya se mencionó, esta forma de compartir datos no es la adecuada, pues deja en desventaja a los investigadores solicitantes, ya que los que poseen los datos pueden seleccionar quién tendrá acceso, y además, los datos no se encuentran en una base de datos que mantenga su seguridad, ni control eficiente de la metadata (Byrd et al., 2020). Por lo anterior consideramos que la disponibilidad de datos es un punto muy mejorable en la genómica NatAm. Compartir los datos obtenidos en los estudios científicos es un elemento esencial e imprescindible para garantizar la transparencia y la reproducibilidad de la ciencia (Wilson et al., 2021). Los investigadores suelen solicitar y reutilizar los datos compartidos para realizar nuevos estudios y generar

nuevos datos que permitan el progreso de la genómica. Si bien reconocemos la importancia de los datos, también conocemos algunas de las razones para no compartir sus datos, por ejemplo la preocupación por el mal uso de los mismos y la posible pérdida de futuras publicaciones propias como resultado de abrir la data a otros investigadores (scooping) (Byrd et al., 2020; Popkin, 2019). Este problema puede resolverse compartiendo los datos sólo después de haber realizado una publicación. Además, se puede optar por publicar un preprint de un estudio de investigación cuando se decide compartir los datos. Además, se puede optar por sólo compartir los datos y metadatos necesarios para reproducir el estudio publicado (Wilson et al., 2021). El concepto de compartir datos tiene que ver con la colaboración científica en beneficio de la humanidad, pero parece que, ahora mismo, la genómica NatAm es una competición entre investigadores dentro de una torre de babel (Powell, 2021).

La falta de presencia de más grupos e individuos NatAm se ha hecho notar en estudios genómicos poblacionales y en bases de datos genómicas. Por ejemplo, la base de datos gnomAD (<http://gnomad.broadinstitute.org/about>) incluye información de referencia sobre variantes genómicas de más de 141,000 individuos en todo el mundo, pero carece de información explícita sobre grupos NatAm (Karczewski et al., 2020). En consecuencia, no se conoce la frecuencia poblacional de las variantes genómicas en poblaciones indígenas, lo que da lugar a resultados de diagnóstico potencialmente menos precisos en comparación con los de las poblaciones mejor representadas. El acceso desigual a las tecnologías genómicas, los determinantes socioeconómicos negativos y la falta de datos pertinentes sobre las variaciones genéticas de la población contribuyen a que la investigación genómica tenga una relevancia limitada y una eficacia reducida para los pueblos indígenas (Garrison et al., 2019). Si los investigadores planean incrementar el número de estudios, y por consecuencia, los datos genómicos de los grupos NatAm, desde la planeación de los proyectos se deben tener en consideración los aspectos éticos y legales que aseguren el bienestar de los individuos (Claw et al., 2018; Garrison et al., 2019; Hudson et al., 2020). Sin embargo, esta tarea no ha sido sencilla a través de los últimos años, por diversas razones, entre las que destacan la

incapacidad de los investigadores para hacer partícipes a las comunidades indígenas de forma ética e inclusiva, la falta de transparencia de los estudios y las malas prácticas de investigación (Bowe Katy & Davis, 2003; Dukepoo, 1998; Mello & Wolf, 2010), todo lo cual ha sembrado la desconfianza y ha justificado la falta de voluntad de algunos pueblos nativos para compartir su información genética personal.

Existen reportes y propuestas que han puesto de manifiesto diversas recomendaciones para mejorar la participación de las comunidades indígenas (Claw et al., 2018; Garrison et al., 2019; Hudson et al., 2020; Morgan et al., 2019), las cuales básicamente se pueden resumir en establecer relaciones de confianza y colaboración, escuchar, aprender y respetar las perspectivas culturales, los conocimientos y los valores de las comunidades, mejorar la transparencia de las prácticas de investigación mediante la claridad de los objetivos, procesos y la comunicación de los resultados a obtener, así como fomentar la participación de la comunidad nativa en el proceso de la investigación. Por último, no se debe dejar de lado la difusión de los resultados, con métodos entendibles, accesibles y culturalmente respetuosos.

## 10. CONCLUSIONES

Lo expuesto a lo largo de este trabajo permite llegar a las siguientes conclusiones.

1. La utilidad de la base de datos genómicos integrada de nativos americanos busca promover el estudio de la diversidad genómica de nativos americanos a través del conocimiento del estado actual de la genómica NatAm. Lo anterior permitirá plantear más y mejores investigaciones usando tecnologías genómicas para capturar el panorama genómico de los grupos nativos faltantes.
2. Se han estudiado un total de 13,706 individuos NatAm con estas tecnologías genómicas: 4,009 con Arrays, 8,405 con WES y 1,292 con WGS. 489 de los individuos estudiados son muestras antiguas, y 13.217 son contemporáneas. 178 grupos NatAm (contemporáneos y antiguos) han sido representados en

estudios genómicos de poblaciones y 607 grupos NatAm siguen sin estar representados en el archivo genómico colectivo. Se ha utilizado la tecnología de arrays para estudiar 152 grupos NatAm, 19 grupos han sido estudiados mediante WES, y WGS se ha implementado para analizar 70 grupos. Esta falta de representación de los grupos NatAm en los estudios genómicos exagera la idea de que la genómica no es globalmente diversa.

3. Al analizar la disponibilidad de datos del total de individuos estudiados hasta 2021, sólo el 14 % de las muestras de Array están disponibles, para WES sólo hay un 0,9 %, y para WGS la cuota disponible es del 23,4 %. De manera contrastante, el 96,3% de las muestras antiguas proporcionan datos directamente disponibles, en contraste con el 3,6% de las muestras modernas.
4. La forma predominante en que los investigadores comparten es bajo petición a los responsables de los datos.
5. Encontramos que muchos de los países de origen de los grupos NatAm también carecen de representación como líderes en los proyectos de investigación en genómica poblacional.

## 11. PERSPECTIVAS

A partir de los resultados y conclusiones presentados en esta Tesis, las perspectivas a futuro son las siguientes.

La base de datos creada podrá usarse para promover y desarrollar nuevas investigaciones en nativos americanos usando tecnologías de secuenciación en países donde hay poca o nula representación de estos individuos, por ejemplo, países caribeños como Santa Lucía, Dominica o San Vicente y las Granadinas, o en centroamérica, países como El Salvador, Nicaragua u Honduras.

La tendencia que existe actualmente en la disponibilidad de datos debería llamar la atención, e iniciar la conversación en busca de un cambio de paradigma, hacia uno en el que los investigadores accedan más a compartir los datos

generados en los estudios genómicos para promover la aplicación y los beneficios a la población nativa americana.

Finalmente, será de gran interés estudiar la información públicamente recopilada mediante nuevos análisis bioinformáticos en busca de expandir la investigación genómica sin la necesidad de inversión en la generación de nueva información genética.

## 12. REFERENCIAS

1. 1000 Genomes Project Consortium, Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., Korbel, J. O., Marchini, J. L., McCarthy, S., McVean, G. A., & Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74. <https://doi.org/10.1038/nature15393>
2. Adhikari, K., Chacón-Duque, J. C., Mendoza-Revilla, J., Fuentes-Guajardo, M., & Ruiz-Linares, A. (2017). The Genetic Diversity of the Americas. *Annual Review of Genomics and Human Genetics*, *18*(1), 277–296. <https://doi.org/10.1146/annurev-genom-083115-022331>
3. Aguilar-Ordoñez, I., Pérez-Villatoro, F., García-Ortiz, H., Barajas-Olmos, F., Ballesteros-Villascán, J., González-Buenfil, R., Fresno, C., Garcíarrubio, A., Fernández-López, J. C., Tovar, H., Hernández-Lemus, E., Orozco, L., Soberón, X., & Morett, E. (2021). Whole genome variation in 27 Mexican indigenous populations, demographic and biomedical insights. *PLOS ONE*, *16*(4). <https://doi.org/10.1371/journal.pone.0249773>
4. Ashley, E. A. (2016). Towards precision medicine. *Nature Reviews Genetics*, *17*(9), 507–522. <https://doi.org/10.1038/nrg.2016.86>
5. Attwood, T. K., Blackford, S., Brazas, M. D., Davies, A. C., & Schneider, M. V. (2019). A global perspective on evolving bioinformatics and data science training needs. *Briefings in Bioinformatics*, *20*(2), 398–404. <https://doi.org/10.1093/bib/bbx100>

6. Ávila-Arcos, M. C., McManus, K. F., Sandoval, K., Rodríguez-Rodríguez, J. E., Villa-Islas, V., Martin, A. R., Luisi, P., Peñaloza-Espinosa, R. I., Eng, C., Huntsman, S., Burchard, E. G., Gignoux, C. R., Bustamante, C. D., & Moreno-Estrada, A. (2020). Population History and Gene Divergence in Native Mexicans Inferred from 76 Human Exomes. *Molecular Biology and Evolution*, 37(4), 994–1006. <https://doi.org/10.1093/molbev/msz282>
7. Ballesteros Villascán, J. (2019). Desarrollo de una herramienta bioinformática para el análisis poblacional de la variación genómica. *Benemérita Universidad Autónoma de Puebla*. <https://repositorioinstitucional.buap.mx/handle/20.500.12371/4640>
8. Ballouz, S., Dobin, A., & Gillis, J. A. (2019). Is it time to change the reference genome? *Genome Biology*, 20(1), 159. <https://doi.org/10.1186/s13059-019-1774-4>
9. Barbieri, C., Barquera, R., Arias, L., Sandoval, J. R., Acosta, O., Zurita, C., Aguilar-Campos, A., Tito-Álvarez, A. M., Serrano-Osuna, R., Gray, R. D., Mafessoni, F., Paul Heggarty, Heggarty, P., Shimizu, K. K., Shimizu, K., Fujita, R., Stoneking, M., Pugach, I., & Fehren-Schmitz, L. (2019). The Current Genomic Landscape of Western South America: Andes, Amazonia, and Pacific Coast. *Molecular Biology and Evolution*, 36(12), 2698–2713. <https://doi.org/10.1093/molbev/msz174>
10. Bergström, A., McCarthy, S. A., Hui, R., Almarri, M. A., Ayub, Q., Danecek, P., Chen, Y., Felkel, S., Hallast, P., Kamm, J., Blanché, H., Deleuze, J.-F., Cann, H., Mallick, S., Reich, D., Sandhu, M. S., Skoglund, P., Scally, A., Xue, Y., ... Tyler-Smith, C. (2020). Insights into human genetic variation and population history from 929 diverse genomes. *Science (New York, N.Y.)*, 367(6484), eaay5012. <https://doi.org/10.1126/science.aay5012>
11. Bien, S. A., Wojcik, G. L., Hodonsky, C. J., Gignoux, C. R., Cheng, I., Matise, T. C., Peters, U., Kenny, E. E., & North, K. E. (2019). The Future of Genomic Studies Must Be Globally Representative: Perspectives from PAGE. *Annual Review of Genomics and Human Genetics*, 20(1), 181–200. <https://doi.org/10.1146/annurev-genom-091416-035517>
12. Bolnick, D. A., Raff, J. A., Springs, L. C., Reynolds, A. W., & Miró-Herrans, A.

- T. (2016). Native American Genomics and Population Histories. *Annual Review of Anthropology*, 45(1), 319–340. <https://doi.org/10.1146/annurev-anthro-102215-100036>
13. Bongers, J., Nakatsuka, N., O’Shea, C., Harper, T. K., Tantaleán, H., Tantaleán, H., Stanish, C., & Fehren-Schmitz, L. (2020). Integration of ancient DNA with transdisciplinary dataset finds strong support for Inca resettlement in the south Peruvian coast. *Proceedings of the National Academy of Sciences of the United States of America*, 117(31), 18359–18368. <https://doi.org/10.1073/pnas.2005965117>
14. Borck, K., Beggs, J. D., Brammar, W. J., Hopkins, A. S., & Murray, N. E. (1976). The construction in vitro of transducing derivatives of phage lambda. *Molecular and General Genetics MGG*, 146(2), 199–207. <https://doi.org/10.1007/BF00268089>
15. Borda, V., Alvim, I., Mendes, M., Silva-Carvalho, C., Soares-Souza, G. B., Leal, T. P., Furlan, V., Scliar, M. O., Marilia O Scliar, Zamudio, R., Zolini, C., Araújo, G. S., Luizon, M. R., Padilla, C., Cáceres, O., Levano, K. S., Sánchez, C. C., Sanchez, C., Trujillo, O., ... Guio, H. (2020). The genetic structure and adaptation of Andean highlanders and Amazonians are influenced by the interplay between geography and culture. *Proceedings of the National Academy of Sciences of the United States of America*, 117(51), 32557–32565. <https://doi.org/10.1073/pnas.2013773117>
16. Bowekaty, M. B., & Davis, D. S. (2003). Cultural Issues in Genetic Research with American Indian and Alaskan Native People. *IRB: Ethics & Human Research*, 25(4), 12–15. <https://doi.org/10.2307/3563819>
17. Bustamante, C. D., De La Vega, F. M., & Burchard, E. G. (2011). Genomics for the world. *Nature*, 475(7355), 163–165. <https://doi.org/10.1038/475163a>
18. Byrd, J. B., Greene, A. C., Prasad, D. V., Jiang, X., Xiaoqian Jiang, Xiaoqian Jiang, Xiaoqian Jiang, & Greene, C. S. (2020). Responsible, practical genomic data sharing that accelerates research. *Nature Reviews Genetics*, 21(10), 615–629. <https://doi.org/10.1038/s41576-020-0257-5>
19. Capodiferro, M. R., Aram, B., Raveane, A., Migliore, N. R., Colombo, G., Ongaro, L., Rivera, J., Mendizábal, T., Hernández-Mora, I., Tribaldos, M.,

- Perego, U. A., Li, H., Scheib, C. L., Christiana Lyn Scheib, Christiana L. Scheib, Modi, A., Gómez-Carballa, A., Grugni, V., Lombardo, G., ... Achilli, A. (2021). Archaeogenomic distinctiveness of the Isthmo-Colombian area. *Cell*, 184(7), 1706. <https://doi.org/10.1016/j.cell.2021.02.040>
20. Carress, H., Lawson, D. J., Lawson, D., & Elhaik, E. (2021). Population genetic considerations for using biobanks as international resources in the pandemic era and beyond. *BMC Genomics*, 22(1), 1–19. <https://doi.org/10.1186/s12864-021-07618-x>
21. Cavalli-Sforza, L. L. (2005). The Human Genome Diversity Project: Past, present and future. *Nature Reviews Genetics*, 6(4), 333–340. <https://doi.org/10.1038/nrg1596>
22. Claw, K. G., Anderson, M. Z., Begay, R. L., Tsosie, K. S., Fox, K., & Garrison, N. A. (2018). A framework for enhancing ethical genomic research with Indigenous communities. *Nature Communications*, 9(1), 2957. <https://doi.org/10.1038/s41467-018-05188-3>
23. Colavizza, G., Hrynaszkiewicz, I., Staden, I., Whitaker, K., & McGillivray, B. (2020). The citation advantage of linking publications to research data. *PLOS ONE*, 15(4), e0230416. <https://doi.org/10.1371/journal.pone.0230416>
24. Day, S. E., Muller, Y. L., Köroğlu, Ç., Kobes, S., Wiedrich, K., Mahkee, D., Kim, H. I., Van Hout, C., Gosalia, N., Ye, B., Shuldiner, A. R., Knowler, W. C., Hanson, R. L., Bogardus, C., & Baier, L. J. (2021). Exome Sequencing of 21 Bardet-Biedl Syndrome (BBS) Genes to Identify Obesity Variants in 6,851 American Indians. *Obesity*, 29(4), 748–754. <https://doi.org/10.1002/oby.23115>
25. de la Fuente, C., Ávila-Arcos, M. C., Galimany, J., Carpenter, M. L., Homburger, J. R., Blanco, A. T., Contreras, P., Dávalos, D. I. C., Diana Cruz Dávalos, Reyes, O., San Román, M., Moreno-Estrada, A., Campos, P. F., Eng, C., Huntsman, S., Burchard, E. G., Malaspinas, A.-S., Bustamante, C., Willerslev, E., ... Moraga, M. (2018). Genomic insights into the origin and diversification of late maritime hunter-gatherers from the Chilean Patagonia. *Proceedings of the National Academy of Sciences of the United States of America*, 115(17), 201715688. <https://doi.org/10.1073/pnas.1715688115>
26. Dewey, F. E., Murray, M. F., Overton, J. D., Habegger, L., Leader, J. B.,

- Fetterolf, S. N., O'Dushlaine, C., Van Hout, C. V., Jeffrey Staples, Staples, J., Gonzaga-Jauregui, C., Metpally, R., Pendergrass, S. A., Monica A. Giovanni, Giovanni, M. A., Monica A. Giovanni, Kirchner, H. L., Balasubramanian, S., Abul-Husn, N. S., ... Carey, D. J. (2016). Distribution and clinical impact of functional variants in 50,726 whole-exome sequences from the DiscovEHR study. *Science*, 354(6319). <https://doi.org/10.1126/science.aaf6814>
27. Dukepoo, F. C. (1998). Genetic Services in the New Era: Native American Perspectives. *Community Genetics*, 1(3), 130–133.
28. Eichler, E. E. (2019). Genetic Variation, Comparative Genomics, and the Diagnosis of Disease. *The New England Journal of Medicine*, 381(1), 64. <https://doi.org/10.1056/NEJMra1809315>
29. Fehren-Schmitz, L., Jarman, C. L., Harkins, K. M., Kayser, M., Popp, B. N., & Skoglund, P. (2017). Genetic Ancestry of Rapanui before and after European Contact. *Current Biology*, 27(20), 3209. <https://doi.org/10.1016/j.cub.2017.09.029>
30. Fernandes, D. M., Sirak, K. A., Ringbauer, H., Sedig, J., Rohland, N., Cheronet, O., Mah, M., Mallick, S., Olalde, I., Culleton, B. J., Adamski, N., Bernardos, R., Bravo, G., Broomandkhoshbacht, N., Callan, K., Candilio, F., Demetz, L., Carlson, K. S. D., Eccles, L., ... Reich, D. (2021). A genetic history of the pre-contact Caribbean. *Nature*, 590(7844), 103–110. <https://doi.org/10.1038/s41586-020-03053-2>
31. Flegontov, P., Altınışık, N. E., Changmai, P., Rohland, N., Mallick, S., Adamski, N., Deborah A. Bolnick, Bolnick, D. A., Broomandkhoshbacht, N., Candilio, F., Culleton, B. J., Culleton, B. J., Flegontova, O., Friesen, T. M., Jeong, C., Harper, T. K., Keating, D., Kennett, D. J., Kennett, D. J., ... Schiffels, S. (2019). Palaeo-Eskimo genetic ancestry and the peopling of Chukotka and North America. *Nature*, 570(7760), 236–240. <https://doi.org/10.1038/s41586-019-1251-y>
32. Frazer, K. A., Murray, S. S., Schork, N. J., & Topol, E. J. (2009). Human genetic variation and its contribution to complex traits. *Nature Reviews Genetics*, 10(4), 241–251. <https://doi.org/10.1038/nrg2554>
33. García-Ortiz, H., Barajas-Olmos, F., Contreras-Cubas, C., Cid-Soto, M. A.,

- Córdova, E., Centeno-Cruz, F., Mendoza-Caamal, E., Cicerón-Arellano, I., Marlen Flores-Huacuja, Baca, P., Bolnick, D. A., Snow, M., Flores-Martínez, S. E., Ortiz-Lopez, R., Austin W. Reynolds, Antonio Blanchet, Morales-Marin, M. E., Velázquez-Cruz, R., Kostić, A., ... Orozco, L. (2021). The genomic landscape of Mexican Indigenous populations brings insights into the peopling of the Americas. *Nature Communications*, 12(1), 5942. <https://doi.org/10.1038/s41467-021-26188-w>
34. Garrison, N. A., Hudson, M., Ballantyne, L. L., Garba, I., Martinez, A., Taulii, M., Arbour, L., Caron, N. R., & Rainie, S. C. (2019). Genomic Research Through an Indigenous Lens: Understanding the Expectations. *Annual Review of Genomics and Human Genetics*, 20(1), 495–517. <https://doi.org/10.1146/annurev-genom-083118-015434>
35. Garvan Institute of Medical Research. (2021). *Types of variants*. Garvan Institute of Medical Research. <https://www.garvan.org.au/research/kinghorn-centre-for-clinical-genomics/learn-about-genomics/for-gp/genetics-refresher-1/types-of-variants>
36. Genome of the Netherlands Consortium. (2014). Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nature Genetics*, 46(8), 818–825. <https://doi.org/10.1038/ng.3021>
37. Giani, A. M., Gallo, G. R., Gianfranceschi, L., & Formenti, G. (2020). Long walk to genomics: History and current approaches to genome sequencing and assembly. *Computational and Structural Biotechnology Journal*, 18, 9–19. <https://doi.org/10.1016/j.csbj.2019.11.002>
38. Gnecci-Ruscone, G. A., Stefania Sarno, Stefania Sarno, Sarno, S., De Fanti, S., Gianvincenzo, L., Giuliani, C., Boattini, A., Bortolini, E., Di Corcia, T., Mellado, C. S., Francia, T. J. D., Gentilini, D., Di Blasio, A. M., Di Cosimo, P., Cilli, E., Elisabetta Cilli, Elisabetta Cilli, González-Martín, A., ... Pettener, D. (2019). Dissecting the Pre-Columbian Genomic Ancestry of Native Americans along the Andes–Amazonia Divide. *Molecular Biology and Evolution*, 36(6), 1254–1269. <https://doi.org/10.1093/molbev/msz066>
39. Gravel, S., Henn, B. M., Gutenkunst, R. N., Indap, A. R., Marth, G. T., Clark, A. G., Yu, F., Gibbs, R. A., 1000 Genomes Project, & Bustamante, C. D.

- (2011). Demographic history and rare allele sharing among human populations. *Proceedings of the National Academy of Sciences of the United States of America*, 108(29), 11983–11988. <https://doi.org/10.1073/pnas.1019276108>
40. Gresham, D., Dunham, M. J., & Botstein, D. (2008). Comparing whole genomes using DNA microarrays. *Nature Reviews. Genetics*, 9(4), 291. <https://doi.org/10.1038/nrg2335>
41. Gudmundsson, S., Singer-Berk, M., Watts, N. A., Phu, W., Goodrich, J. K., Solomonson, M., Genome Aggregation Database Consortium, Rehm, H. L., MacArthur, D. G., & O'Donnell-Luria, A. (2021). Variant interpretation using population databases: Lessons from gnomAD. *Human Mutation*. <https://doi.org/10.1002/humu.24309>
42. Guglielmi, G. (2019). Facing up to injustice in genome science. *Nature*, 568(7752), 290–293. <https://doi.org/10.1038/d41586-019-01166-x>
43. Harris, D. N., Song, W., Shetty, A. C., Levano, K. S., Cáceres, O., Padilla, C., Borda, V., Tarazona, D., Trujillo, O., Sanchez, C., Kessler, M. D., Galarza, M., Capristano, S., Montejó, H., Flores-Villanueva, P. O., Tarazona-Santos, E., O'Connor, T. D., & Guio, H. (2018). Evolutionary genomic dynamics of Peruvians before, during, and after the Inca Empire. *Proceedings of the National Academy of Sciences*, 115(28), E6526–E6535. <https://doi.org/10.1073/pnas.1720798115>
44. Helmy, M., Awad, M., & Mosa, K. A. (2016). Limited resources of genome sequencing in developing countries: Challenges and solutions. *Applied & Translational Genomics*, 9, 15–19. <https://doi.org/10.1016/j.atg.2016.03.003>
45. Hetu, M., Koutouki, K., & Joly, Y. (2019). Genomics for All: International Open Science Genomics Projects and Capacity Building in the Developing World. *Frontiers in Genetics*, 10. <https://www.frontiersin.org/article/10.3389/fgene.2019.00095>
46. Hindorff, L. A., Bonham, V. L., Brody, L. C., Ginoza, M. E. C., Hutter, C. M., Manolio, T. A., & Green, E. D. (2018). Prioritizing diversity in human genomics research. *Nature Reviews Genetics*, 19(3), 175–185. <https://doi.org/10.1038/nrg.2017.89>

47. Hudson, M., Hudson, M., Garrison, N. A., Sterling, R., Caron, N. R., Fox, K., Yracheta, J., Yracheta, J. M., Anderson, J., Anderson, J., Wilcox, P., Arbour, L., Brown, A., Taulii, M., Kukutai, T., Haring, R. C., Aika, B. T., Baynam, G., Dearden, P. K., ... Carroll, S. R. (2020). Rights, interests and expectations: Indigenous perspectives on unrestricted access to genomic data. *Nature Reviews Genetics*, 21(6), 377–384. <https://doi.org/10.1038/s41576-020-0228-x>
48. International HapMap Consortium. (2004). Integrating ethics and science in the International HapMap Project. *Nature Reviews Genetics*, 5(6), 467–475. <https://doi.org/10.1038/nrg1351>
49. International HapMap Consortium. (2005). A haplotype map of the human genome. *Nature*, 437(7063), 1299–1320. <https://doi.org/10.1038/nature04226>
50. Ioannidis, A. G., Blanco-Portillo, J., Sandoval, K., Hagelberg, E., Miquel-Poblete, J. F., Moreno-Mayar, J. V., Rodríguez-Rodríguez, J. E., Quinto-Cortés, C. D., Auckland, K., Parks, T., Robson, K., Hill, A. V. S., Avila-Arcos, M. C., Sockell, A., Homburger, J. R., Wojcik, G. L., Barnes, K. C., Herrera, L., Berríos, S., ... Moreno-Estrada, A. (2020). Native American gene flow into Polynesia predating Easter Island settlement. *Nature*, 583(7817), 572–577. <https://doi.org/10.1038/s41586-020-2487-2>
51. Jiménez-Kaufmann, A., Chong, A. Y., Cortés, A., Quinto-Cortés, C. D., Fernandez-Valverde, S. L., Ferreyra-Reyes, L., Cruz-Hervert, L. P., Medina-Muñoz, S. G., Sohail, M., Palma-Martinez, M. J., Delgado-Sánchez, G., Mongua-Rodríguez, N., Mentzer, A. J., Hill, A. V. S., Moreno-Macías, H., Huerta-Chagoya, A., Aguilar-Salinas, C. A., Torres, M., Kim, H. L., ... Moreno-Estrada, A. (2022). Imputation Performance in Latin American Populations: Improving Rare Variants Representation With the Inclusion of Native American Genomes. *Frontiers in Genetics*, 12. <https://www.frontiersin.org/article/10.3389/fgene.2021.719791>
52. Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., Collins, R. L., Laricchia, K. M., Ganna, A., Birnbaum, D. P., Gauthier, L. D., Brand, H., Solomonson, M., Watts, N. A., Rhodes, D., Singer-Berk, M., England, E. M., Seaby, E. G., Kosmicki, J. A., ... MacArthur,

- D. G. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, 581(7809), 434–443. <https://doi.org/10.1038/s41586-020-2308-7>
53. Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., ... The Wellcome Trust: (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822), 860–921. <https://doi.org/10.1038/35057062>
54. Lazaridis, I., Patterson, N., Mitnik, A., Renaud, G., Mallick, S., Kirsanow, K., Sudmant, P. H., Schraiber, J. G., Castellano, S., Lipson, M., Berger, B., Economou, C., Bollongino, R., Bollongino, R., Fu, Q., Bos, K. I., Nordenfelt, S., Li, H., de Filippo, C., ... Krause, J. (2014). Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature*, 513(7518), 409–413. <https://doi.org/10.1038/nature13673>
55. Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., O'Donnell-Luria, A. H., Ware, J. S., Hill, A. J., Cummings, B. B., Tukiainen, T., Birnbaum, D. P., Kosmicki, J. A., Duncan, L. E., Estrada, K., Zhao, F., Zou, J., Pierce-Hoffman, E., Berghout, J., ... MacArthur, D. G. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, 536(7616), 285–291. <https://doi.org/10.1038/nature19057>
56. Lindo, J., Achilli, A., Perego, U. A., Archer, D., Archer, D. R., Valdiosera, C., Petzelt, B., Mitchell, J., Worl, R., Dixon, E. J., Fifield, T. E., Rasmussen, M., Rasmussen, M. F., Morten Fischer Rasmussen, Willerslev, E., Cybulski, J. S., Kemp, B. M., Brian M. Kemp, DeGiorgio, M., & Malhi, R. S. (2017). Ancient individuals from the North American Northwest Coast reveal 10,000 years of regional genetic continuity. *Proceedings of the National Academy of Sciences of the United States of America*, 114(16), 4093–4098. <https://doi.org/10.1073/pnas.1620410114>
57. Lindo, J., Haas, R., Hofman, C. A., Apata, M., Moraga, M., Verdugo, R. A., Watson, J. T., Llave, C. V., Witonsky, D. B., Beall, C. M., Warinner, C., Novembre, J., Aldenderfer, M., Mark Aldenderfer, & Di Rienzo, A. (2018). The genetic prehistory of the Andean highlands 7000 years BP though European

- contact. *Science Advances*, 4(11). <https://doi.org/10.1126/sciadv.aau4921>
58. Lindo, J., Huerta-Sánchez, E., Nakagome, S., Rasmussen, M., Petzelt, B., Mitchell, J., Cybulski, J. S., Willerslev, E., DeGiorgio, M., & Malhi, R. S. (2016). A time transect of exomes from a Native American population before and after European contact. *Nature Communications*, 7(1), 13175. <https://doi.org/10.1038/ncomms13175>
59. Lupski, J. R. (2016). Clinical genomics: From a truly personal genome viewpoint. *Human Genetics*, 135(6), 591–601. <https://doi.org/10.1007/s00439-016-1682-6>
60. Malaspina, A.-S., Lao, O., Schroeder, H., Rasmussen, M. F., Rasmussen, M., Rasmussen, M. A., Raghavan, M., Moltke, I., Campos, P. F., Sagredo, F. S., Rasmussen, S., Gonçalves, V. F., Albrechtsen, A., Allentoft, M. E., Johnson, P. L. F., Li, M., Reis, S., Bernardo, D. V., DeGiorgio, M., ... Willerslev, E. (2014). Two ancient human genomes reveal Polynesian ancestry among the indigenous Botocudos of Brazil. *Current Biology*, 24(21). <https://doi.org/10.1016/j.cub.2014.09.078>
61. Mallick, S., Li, H., Lipson, M., Mathieson, I., Gymrek, M., Racimo, F., Zhao, M., Chennagiri, N., Nordenfelt, S., Tandon, A., Skoglund, P., Lazaridis, I., Sankararaman, S., Fu, Q., Rohland, N., Renaud, G., Erlich, Y., Willems, T., Gallo, C., ... Reich, D. (2016). The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature*, 538(7624), 201–206. <https://doi.org/10.1038/nature18964>
62. Mangul, S., Serghei Mangul, Serghei Mangul, Martin, L. S., Langmead, B., Sanchez-Galan, J. E., Toma, I., Hormozdiari, F., Pevzner, P. A., & Eskin, E. (2019). How bioinformatics and open data can boost basic science in countries and universities with limited resources. *Nature Biotechnology*, 37(3), 324–326. <https://doi.org/10.1038/s41587-019-0053-y>
63. Manousaki, D., Kent, J. W., Haack, K., Zhou, S., Xie, P., Greenwood, C. M. T., Brassard, P., Newman, D. E., Cole, S. A., Umans, J. G., Rouleau, G. A., Comuzzie, A. G., & Richards, J. B. (2016). Toward Precision Medicine: TBC1D4 Disruption Is Common Among the Inuit and Leads to Underdiagnosis of Type 2 Diabetes. *Diabetes Care*, 39(11), 1889–1895.

<https://doi.org/10.2337/dc16-0769>

64. Manzoni, C., Kia, D. A., Vandrovцова, J., Hardy, J., Wood, N. W., Lewis, P. A., & Ferrari, R. (2018). Genome, transcriptome and proteome: The rise of omics data and their integration in biomedical sciences. *Briefings in Bioinformatics*, 19(2), 286–302. <https://doi.org/10.1093/bib/bbw114>
65. Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y.-J., Chen, Z., Dewell, S. B., Du, L., Fierro, J. M., Gomes, X. V., Godwin, B. C., He, W., Helgesen, S., Ho, C. H., Irzyk, G. P., ... Rothberg, J. M. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057), 376–380. <https://doi.org/10.1038/nature03959>
66. Melendrez, M. C., Shaw, S., S. Shaw, Brown, C. T., Brown, C. T., Goodner, B., B. W. Goodner, & Kvaal, C. (2021). Editorial: Curriculum Applications in Microbiology: Bioinformatics in the Classroom. *Frontiers in Microbiology*, 12, 705233. <https://doi.org/10.3389/fmicb.2021.705233>
67. Mello, M. M., & Wolf, L. E. (2010). The Havasupai Indian Tribe Case—Lessons for Research Involving Stored Biologic Samples. *New England Journal of Medicine*, 363(3), 204–207. <https://doi.org/10.1056/NEJMp1005203>
68. Morales, J., Welter, D., Bowler, E. H., Cerezo, M., Harris, L. W., McMahon, A. C., Hall, P., Junkins, H. A., Milano, A., Hastings, E., Malangone, C., Buniello, A., Burdett, T., Flicek, P., Parkinson, H., Cunningham, F., Hindorff, L. A., & MacArthur, J. A. L. (2018). A standardized framework for representation of ancestry data in genomics studies, with application to the NHGRI-EBI GWAS Catalog. *Genome Biology*, 19(1), 21. <https://doi.org/10.1186/s13059-018-1396-2>
69. Moreno-Estrada, A., Gignoux, C. R., Fernández-López, J. C., Zakharia, F., Sikora, M., Contreras, A. V., Acuña-Alonzo, V., Víctor Acuña-Alonzo, Sandoval, K., Eng, C., Romero-Hidalgo, S., Ortiz-Tello, P., Robles, V., Kenny, E. E., Nuño-Arana, I., Barquera-Lozano, R., Macín-Pérez, G., Julio Granados-Arriola, Granados-Arriola, J., ... Bustamante, C. (2014). The genetics of Mexico recapitulates Native American substructure and affects

- biomedical traits. *Science*, 344(6189), 1280–1285.  
<https://doi.org/10.1126/science.1251688>
70. Moreno-Estrada, A., Gravel, S., Zakharia, F., McCauley, J. L., Byrnes, J. K., Gignoux, C. R., Ortiz-Tello, P., Martinez, R., Hedges, D. J., Morris, R. W., Eng, C., Sandoval, K., Suehelay Acevedo-Acevedo, Norman, P., Layrisse, Z., Parham, P., Martínez-Cruzado, J. C., Burchard, E. G., Cuccaro, M. L., ... Bustamante, C. (2013). Reconstructing the population genetic history of the Caribbean. *PLOS Genetics*, 9(11).  
<https://doi.org/10.1371/journal.pgen.1003925>
71. Moreno-Mayar, J. V., Potter, B. A., Vinner, L., Steinrücken, M., Rasmussen, S., Terhorst, J., John A. Kamm, Kamm, J., John A. Kamm, Albrechtsen, A., Malaspinas, A.-S., Sikora, M., Reuther, J. D., Irish, J. D., Malhi, R. S., Orlando, L., Song, Y. S., Nielsen, R., Meltzer, D. J., & Willerslev, E. (2018). Terminal Pleistocene Alaskan genome reveals first founding population of Native Americans. *Nature*, 553(7687), 203–207.  
<https://doi.org/10.1038/nature25173>
72. Moreno-Mayar, J. V., Rasmussen, S., Seguin-Orlando, A., Rasmussen, M. A., Rasmussen, M., Rasmussen, M. F., Liang, M., Flåm, S. T., Lie, B. A., Gilfillan, G. D., Nielsen, R., Thorsby, E., Willerslev, E., & Malaspinas, A.-S. (2014). Genome-wide Ancestry Patterns in Rapanui Suggest Pre-European Admixture with Native Americans. *Current Biology*, 24(21), 2518–2525.  
<https://doi.org/10.1016/j.cub.2014.09.057>
73. Moreno-Mayar, J. V., Vinner, L., de Barros Damgaard, P., de la Fuente, C., Jeffrey Chan, Chan, J., Spence, J. P., Allentoft, M. E., Vimala, T., Racimo, F., Pinotti, T., Rasmussen, S., Margaryan, A., Orbegozo, M. I., Mylopotamitaki, D., Wooller, M. J., Bataille, C. P., Becerra-Valdivia, L., Chivall, D., ... Willerslev, E. (2018). Early human dispersals within the Americas. *Science*, 362(6419). <https://doi.org/10.1126/science.aav2621>
74. Morgan, J., Coe, R. R., Lesueur, R., Kenny, R., Price, R., Makela, N., & Birch, P. H. (2019). Indigenous Peoples and genomics: Starting a conversation. *Journal of Genetic Counseling*, 28(2), 407–418.  
<https://doi.org/10.1002/jgc4.1073>

75. Mullis, K., Faloona, F., Scharf, S., Saiki, R., Horn, G., & Erlich, H. (1986). Specific Enzymatic Amplification of DNA In Vitro: The Polymerase Chain Reaction. *Cold Spring Harbor Symposia on Quantitative Biology*, 51(0), 263–273. <https://doi.org/10.1101/SQB.1986.051.01.032>
76. Nagasaki, M., Yasuda, J., Katsuoka, F., Nariai, N., Kojima, K., Kawai, Y., Yamaguchi-Kabata, Y., Yokozawa, J., Danjoh, I., Saito, S., Sato, Y., Mimori, T., Tsuda, K., Saito, R., Pan, X., Nishikawa, S., Ito, S., Kuroki, Y., Tanabe, O., ... Yamamoto, M. (2015). Rare variant discovery by deep whole-genome sequencing of 1,070 Japanese individuals. *Nature Communications*, 6, 8018. <https://doi.org/10.1038/ncomms9018>
77. Naisbitt, J. (1982). *Megatrends: Ten New Directions Transforming Our Lives*. Warner Books. <https://books.google.com.mx/books?id=LzMEy5bBONcC>
78. National Human Genome Research Institute. (2020, diciembre 22). *The Human Genome Project*. Genome.Gov. <https://www.genome.gov/human-genome-project>
79. Nebula Genomics. (2022). *Whole Genome Sequencing DNA Test*. Nebula Genomics. <https://nebula.org/whole-genome-sequencing-dna-test/>
80. Need, A. C., & Goldstein, D. B. (2009). Next generation disparities in human genomics: Concerns and remedies. *Trends in Genetics*, 25(11), 489–494. <https://doi.org/10.1016/j.tig.2009.09.012>
81. Ng, S. B., Turner, E. H., Robertson, P. D., Flygare, S. D., Bigham, A. W., Lee, C., Shaffer, T., Wong, M., Bhattacharjee, A., Eichler, E. E., Bamshad, M., Nickerson, D. A., & Shendure, J. (2009). Targeted capture and massively parallel sequencing of 12 human exomes. *Nature*, 461(7261), 272–276. <https://doi.org/10.1038/nature08250>
82. Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bizkadze, A. V., Mikheenko, A., Vollger, M. R., Altemose, N., Uralsky, L., Gershman, A., Aganezov, S., Hoyt, S. J., Diekhans, M., Logsdon, G. A., Alonge, M., Antonarakis, S. E., Borchers, M., Bouffard, G. G., Brooks, S. Y., ... Phillippy, A. M. (2021). *The complete sequence of a human genome* (p. 2021.05.26.445798). bioRxiv. <https://doi.org/10.1101/2021.05.26.445798>
83. Owen, M. J., Niemi, A.-K., Dimmock, D. P., Speziale, M., Nespeca, M., Chau,

- K. K., Van Der Kraan, L., Wright, M. S., Hansen, C., Veeraraghavan, N., Ding, Y., Lenberg, J., Chowdhury, S., Hobbs, C. A., Batalov, S., Zhu, Z., Nahas, S. A., Gilmer, S., Knight, G., ... Kingsmore, S. F. (2021). Rapid Sequencing-Based Diagnosis of Thiamine Metabolism Dysfunction Syndrome. *New England Journal of Medicine*, *384*(22), 2159–2161. <https://doi.org/10.1056/NEJMc2100365>
84. Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., Genschoreck, T., Webster, T., & Reich, D. (2012). Ancient Admixture in Human History. *Genetics*, *192*(3), 1065–1093. <https://doi.org/10.1534/genetics.112.145037>
85. Peng, Q., Schork, N. J., Wilhelmsen, K. C., & Ehlers, C. L. (2017). Whole genome sequence association and ancestry-informed polygenic profile of EEG alpha in a Native American population. *American Journal of Medical Genetics*, *174*(4), 435–450. <https://doi.org/10.1002/ajmg.b.32533>
86. Petersen, B.-S., Fredrich, B., Hoepfner, M. P., Ellinghaus, D., & Franke, A. (2017). Opportunities and challenges of whole-genome and -exome sequencing. *BMC Genetics*, *18*(1), 14. <https://doi.org/10.1186/s12863-017-0479-5>
87. Piaggi, P., Köroğlu, Ç., Nair, A. K., Sutherland, J., Muller, Y. L., Kumar, P., Hsueh, W.-C., Kobes, S., Shuldiner, A. R., Kim, H. I., Gosalia, N., Van Hout, C. V., Cristopher V. Van Hout, Cristopher V. Van Hout, Jones, M., Knowler, W. C., Krakoff, J., Hanson, R. L., Bogardus, C., & Baier, L. J. (2020). Exome Sequencing Identifies A Nonsense Variant in DAO Associated With Reduced Energy Expenditure in American Indians. *The Journal of Clinical Endocrinology and Metabolism*, *105*(11). <https://doi.org/10.1210/clinem/dgaa548>
88. Popejoy, A. B., & Fullerton, S. M. (2016). Genomics is failing on diversity. *Nature*, *538*(7624), 161–164. <https://doi.org/10.1038/538161a>
89. Popkin, G. (2019). Data sharing and how it can benefit your scientific career. *Nature*, *569*(7756), 445–447. <https://doi.org/10.1038/d41586-019-01506-x>
90. Popović, D., Molak, M., Ziółkowski, M., Vranich, A., Sobczyk, M., Vidaurre, D. U., Guido Agresti, Skrzypczak, M., Ginalski, K., Lamnidis, T. C., Nakatsuka,

- N., Mallick, S., & Baca, M. (2021). Ancient genomes reveal long-range influence of the pre-Columbian culture and site of Tiwanaku. *Science Advances*, 7(39). <https://doi.org/10.1126/sciadv.abg7261>
91. Posth, C., Nakatsuka, N., Lazaridis, I., Skoglund, P., Mallick, S., Lamnidis, T. C., Rohland, N., Nägele, K., Adamski, N., Bertolini, E., Broomandkoshbacht, N., Cooper, A., Culleton, B. J., Tiago Ferraz, Ferry, M., Furtwängler, A., Haak, W., Harkins, K. M., Harper, T. K., ... Reich, D. (2018). Reconstructing the Deep Population History of Central and South America. *Cell*, 175(5), 1185–1197. <https://doi.org/10.1016/j.cell.2018.10.027>
92. Powell, K. (2021). The broken promise that undermines human genome research. *Nature*, 590(7845), 198–201. <https://doi.org/10.1038/d41586-021-00331-5>
93. Pruitt, K. D., Harrow, J., Harte, R. A., Wallin, C., Diekhans, M., Maglott, D. R., Searle, S., Farrell, C. M., Loveland, J. E., Ruff, B. J., Hart, E., Suner, M.-M., Landrum, M. J., Aken, B., Ayling, S., Baertsch, R., Fernandez-Banet, J., Cherry, J. L., Curwen, V., ... Lipman, D. (2009). The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. *Genome Research*, 19(7), 1316–1323. <https://doi.org/10.1101/gr.080531.108>
94. Rabbee, N., & Speed, T. P. (2006). A genotype calling algorithm for affymetrix SNP arrays. *Bioinformatics (Oxford, England)*, 22(1), 7–12. <https://doi.org/10.1093/bioinformatics/bti741>
95. Raghavan, M., DeGiorgio, M., Albrechtsen, A., Moltke, I., Skoglund, P., Korneliussen, T. S., Grønnow, B., Appelt, M., Gulløv, H. C., Friesen, T. M., Fitzhugh, W. W., Malmström, H., Rasmussen, S., Olsen, J. V., Melchior, L. C., Fuller, B. T., Fahrni, S., Stafford, T. W., Grimes, V., ... Willerslev, E. (2014). The genetic prehistory of the New World Arctic. *Science*, 345(6200), 1255832–1255832. <https://doi.org/10.1126/science.1255832>
96. Raghavan, M., Steinrücken, M., Harris, K., Schiffels, S., Rasmussen, S., DeGiorgio, M., Albrechtsen, A., Valdiosera, C., Ávila-Arcos, M. C., Malaspina, A.-S., Eriksson, A., Eriksson, A., Moltke, I., Metspalu, M., Homburger, J. R., Wall, J. D., Cornejo, O. E., Moreno-Mayar, J. V.,

- Korneliusson, T. S., ... Willerslev, E. (2015). Genomic evidence for the Pleistocene and recent population history of Native Americans. *Science*, 349(6250), 3384–3396. <https://doi.org/10.1126/science.aab3884>
97. Ranganathan, S., Nakai, K., & Schonbach, C. (2018). *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics* (1st ed.). Elsevier Science Publishers B. V.
98. Rasmussen, M., Anzick, S. L., Waters, M. R., Skoglund, P., DeGiorgio, M., Stafford, T. W., Rasmussen, S., Moltke, I., Albrechtsen, A., Doyle, S. M., Poznik, G. D., Gudmundsdottir, V., Yadav, R., Malaspina, A.-S., White, V. S. S., Allentoft, M. E., Cornejo, O. E., Tambets, K., Eriksson, A., ... Willerslev, E. (2014). The genome of a Late Pleistocene human from a Clovis burial site in western Montana. *Nature*, 506(7487), 225–229. <https://doi.org/10.1038/nature13025>
99. Rasmussen, M. F., Rasmussen, M., Sikora, M., Korneliusson, T. S., Poznik, G. D., Zollikofer, C. P. E., Allentoft, M. E., Moltke, I., Jónsson, H., Malhi, R. S., Orlando, L., Stafford, T. W., Meltzer, D. J., Willerslev, E., Rasmussen, M., Albrechtsen, A., Moreno-Mayar, J. V., León, M. S. P. de, Valdiosera, C., ... Nielsen, R. (2015). The ancestry and affiliations of Kennewick Man. *Nature*. <https://doi.org/10.1038/nature14625>
100. Rasmussen, M., Li, Y., Yingrui Li, Li, Y., Li, Y., Lindgreen, S., Pedersen, J. S., Albrechtsen, A., Moltke, I., Metspalu, M., Ene Metspalu, Metspalu, E., Kivisild, T., Gupta, R., Bertalan, M., Nielsen, K., Gilbert, M. T. P., Wang, Y., Yong Wang, ... Willerslev, E. (2010). Ancient human genome sequence of an extinct Palaeo-Eskimo. *Nature*, 463(7282), 757–762. <https://doi.org/10.1038/nature08835>
101. Reich, D., Green, R. E., Kircher, M., Krause, J., Patterson, N., Durand, E. Y., Durand, E., Eric Durand, Viola, B., Briggs, A. W., Stenzel, U., Johnson, P. L. F., Maricic, T., Good, J. M., Marques-Bonet, T., Alkan, C., Fu, Q., Mallick, S., Li, H., ... Pääbo, S. (2010). Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature*, 468(7327), 1053–1060. <https://doi.org/10.1038/nature09710>
102. Reich, D., Patterson, N., Campbell, D., Tandon, A., Mazières, S.,

- Mazières, S., Ray, N., Parra, M. V., Winston Rojas, W., Duque, C., Mesa, N., García, L. F., Triana, O., Blair, S., Maestre, A., Dib, J. C., Dib, J. C., ... Ruiz-Linares, A. (2012). Reconstructing Native American population history. *Nature*, *488*(7411), 370–374. <https://doi.org/10.1038/nature11258>
103. Reynolds, A. W., Austin W. Reynolds, Mata-Míguez, J., Miró-Herrans, A. T., Briggs-Cloud, M., Sylestine, A., Barajas-Olmos, F., García-Ortiz, H., Rzhetskaya, M., Orozco, L., Raff, J., M. Geoffrey Hayes, Hayes, M. G., Bolnick, D. A., & Deborah A. Bolnick. (2019). Comparing signals of natural selection between three Indigenous North American populations. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(19), 9312–9317. <https://doi.org/10.1073/pnas.1819467116>
104. Ribeiro-dos-Santos, A. M., de Souza, J. E. S., Almeida, R., Alencar, D. O., Barbosa, M. S., Gusmão, L., Silva, W. A., de Souza, S. J., Silva, A., Ribeiro-dos-Santos, Â., Darnet, S., & Santos, S. (2013). High-Throughput Sequencing of a South American Amerindian. *PLOS ONE*, *8*(12). <https://doi.org/10.1371/journal.pone.0083340>
105. Ribeiro-dos-Santos, A. M., Vidal, A. F., Vinasco-Sandoval, T., Guerreiro, J. F., Santos, S., Ribeiro-dos-Santos, Â., de Souza, S. J., & Sandro J. de Souza. (2020). Exome Sequencing of Native Populations From the Amazon Reveals Patterns on the Peopling of South America. *Frontiers in Genetics*, *11*, 548507–548507. <https://doi.org/10.3389/fgene.2020.548507>
106. Romero-Hidalgo, S., Ochoa-Leyva, A., Garcíarrubio, A., Acuña-Alonzo, V., Antúnez-Argüelles, E., Balcazar-Quintero, M., Barquera-Lozano, R., Carnevale, A., Cornejo-Granados, F., Fernández-López, J. C., García-Herrera, R., García-Ortiz, H., Granados-Silvestre, Á., Granados, J., Guerrero-Romero, F., Hernández-Lemus, E., León-Mimila, P., Macín-Pérez, G., Martínez-Hernández, A., ... Soberón, X. (2017). Demographic history and biologically relevant genetic variation of Native Mexicans inferred from whole-genome sequencing. *Nature Communications*, *8*(1), 1005. <https://doi.org/10.1038/s41467-017-01194-z>
107. Sánchez-Pozos, K., Ortiz-López, M. G., Peña-Espinoza, B. I., de los

- Ángeles Granados-Silvestre, M., Jiménez-Jacinto, V., Verleyen, J., Tekola-Ayele, F., Sanchez-Flores, A., & Menjivar, M. (2018). Whole-exome sequencing in maya indigenous families: Variant in PPP1R3A is associated with type 2 diabetes. *Molecular Genetics and Genomics*, 293(5), 1205–1216. <https://doi.org/10.1007/s00438-018-1453-2>
108. Sanger, Nicklen, & Coulson. (1977, diciembre 1). *DNA sequencing with chain-terminating inhibitors*. PNAS. <https://www.pnas.org/doi/abs/10.1073/pnas.74.12.5463>
109. Sastre, L. (2014). Exome sequencing: What clinicians need to know. *Advances in Genomics and Genetics*, 4, 15–27. <https://doi.org/10.2147/AGG.S39108>
110. Scheib, C. L., Li, H., Desai, T., Link, V., Kendall, C., Dewar, G., Griffith, P. W., Mörseburg, A., Johnson, J. R., Potter, A., Kerr, S. L., Endicott, P., Lindo, J., Haber, M., Xue, Y., Tyler-Smith, C., Sandhu, M. S., Lorenz, J. G., Randall, T. D., ... Kivisild, T. (2018). Ancient human parallel lineages within North America contributed to a coastal expansion. *Science (New York, N.Y.)*, 360(6392), 1024–1027. <https://doi.org/10.1126/science.aar6851>
111. Schroeder, H., Sikora, M., Gopalakrishnan, S., Cassidy, L. M., Delser, P. M., Velasco, M. S., Schraiber, J. G., Rasmussen, S., Homburger, J. R., Ávila-Arcos, M. C., Allentoft, M. E., Moreno-Mayar, J. V., Renaud, G., Gabriel Renaud, Gómez-Carballa, A., Laffoon, J. E., Hopkins, R. J. A., Higham, T., Carr, R. S., ... Willerslev, E. (2018). Origins and genetic legacies of the Caribbean Taino. *Proceedings of the National Academy of Sciences of the United States of America*, 115(10), 2341–2346. <https://doi.org/10.1073/pnas.1716839115>
112. Shendure, J., Findlay, G. M., & Snyder, M. W. (2019). Genomic Medicine—Progress, Pitfalls, and Promise. *Cell*, 177(1), 45–57. <https://doi.org/10.1016/j.cell.2019.02.003>
113. Shendure, J., Porreca, G. J., Reppas, N. B., Lin, X., McCutcheon, J. P., Rosenbaum, A. M., Wang, M. D., Zhang, K., Mitra, R. D., & Church, G. M. (2005). Accurate Multiplex Polony Sequencing of an Evolved Bacterial Genome. *Science*, 309(5741), 1728–1732.

- <https://doi.org/10.1126/science.1117389>
114. Silva-Zolezzi, I., Hidalgo-Miranda, A., Estrada-Gil, J. K., Fernández-López, J. C., Uribe-Figueroa, L., Contreras, A. V., Balam-Ortiz, E., del Bosque-Plata, L., Velázquez-Fernández, D., Velázquez-Fernández, D., Lara, C., Goya, R., Hernández-Lemus, E., Carlos Davila, Carlos Davila, Dávila, C., Davila, C., Barrientos, E., March, S., & Jimenez-Sanchez, G. (2009). Analysis of genomic diversity in Mexican Mestizo populations to develop genomic medicine in Mexico. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(21), 8611–8616. <https://doi.org/10.1073/pnas.0903045106>
115. Sirisena, N. D., & Dissanayake, V. H. W. (2019). Strategies for Genomic Medicine Education in Low- and Middle-Income Countries. *Frontiers in Genetics*, *10*. <https://www.frontiersin.org/article/10.3389/fgene.2019.00944>
116. Sirugo, G., Williams, S. M., & Tishkoff, S. A. (2019). The Missing Diversity in Human Genetic Studies. *Cell*, *177*(1), 26–31. <https://doi.org/10.1016/j.cell.2019.02.048>
117. Skoglund, P., Mallick, S., Bortolini, M. C., Chennagiri, N., Hünemeier, T., Petzl-Erler, M. L., Salzano, F. M., Patterson, N., & Reich, D. (2015). Genetic evidence for two founding populations of the Americas. *Nature*, *525*(7567), 104–108. <https://doi.org/10.1038/nature14895>
118. Skoglund, P., & Reich, D. (2016). A genomic view of the peopling of the Americas. *Current Opinion in Genetics & Development*, *41*, 27–35. <https://doi.org/10.1016/j.gde.2016.06.016>
119. Smith, H. O., & Welcox, K. W. (1970). A Restriction enzyme from *Hemophilus influenzae*: I. Purification and general properties. *Journal of Molecular Biology*, *51*(2), 379–391. [https://doi.org/10.1016/0022-2836\(70\)90149-X](https://doi.org/10.1016/0022-2836(70)90149-X)
120. Southern, E. M. (1975). Detection of specific sequences among DNA fragments separated by gel electrophoresis. *Journal of Molecular Biology*, *98*(3), 503–517. [https://doi.org/10.1016/S0022-2836\(75\)80083-0](https://doi.org/10.1016/S0022-2836(75)80083-0)
121. Spangenberg, L., Fariello, M. I., Arce, D., Illanes, G., Greif, G., Shin, J.-Y., Yoo, S.-K., Seo, J.-S., Robello, C., Kim, C., Novembre, J., Sans, M., & Naya,

- H. (2021). Indigenous Ancestry and Admixture in the Uruguayan Population. *Frontiers in Genetics*, 12. <https://www.frontiersin.org/article/10.3389/fgene.2021.733195>
122. Stephens, Z. D., Lee, S. Y., Faghri, F., Faghri, F., Campbell, R. H., Zhai, C., Efron, M., Iyer, R. K., Schatz, M. C., Sinha, S., & Robinson, G. E. (2015). Big data: Astronomical or genetical? *PLOS Biology*, 13(7), 1002195. <https://doi.org/10.1371/journal.pbio.1002195>
123. Udpa, N., Ronen, R., Zhou, D., Liang, J., Stobdan, T., Appenzeller, O., Yin, Y., Du, Y., Guo, L., Cao, R., Wang, Y., Jin, X., Huang, C., Jia, W., Cao, D., Guo, G., Claydon, V. E., Hainsworth, R., Gamboa, J. L., ... Haddad, G. G. (2014). Whole genome sequencing of Ethiopian highlanders reveals conserved hypoxia tolerance genes. *Genome Biology*, 15(2), R36. <https://doi.org/10.1186/gb-2014-15-2-r36>
124. United Nations, Department of Economic and Social Affairs, & Population Division. (2019). *World population prospects 2019 Volume II: Demographic profiles*.
125. Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., ... Zhu, X. (2001). The Sequence of the Human Genome. *Science*, 291(5507), 1304–1351. <https://doi.org/10.1126/science.1058040>
126. Verdu, P., Pemberton, T. J., Laurent, R., Kemp, B. M., González-Oliver, A., Gorodezky, C., Hughes, C. E., Shattuck, M. R., Petzelt, B., Mitchell, J., Harry, H., William, T., Worl, R., Cybulski, J. S., Rosenberg, N. A., & Malhi, R. S. (2014). Patterns of Admixture and Population Structure in Native Populations of Northwest North America. *PLOS Genetics*, 10(8). <https://doi.org/10.1371/journal.pgen.1004530>
127. Vidal, E. A., Moyano, T. C., Bernabé I. Bustos, Bernabé I. Bustos, Bustos, B. I., Pérez-Palma, E., Moraga, C., Riveras, E., Montecinos, A. M., Azocar, L., Soto, D. C., Mabel Vidal, Vidal, M., Alex Di Genova, Di Genova, A., Di Genova, A., Puschel, K., Nürnberg, P., Nürnberg, P., ... Gutiérrez, R. A.

- (2019). Whole Genome Sequence, Variant Discovery and Annotation in Mapuche-Huilliche Native South Americans. *Scientific Reports*, 9(1), 2132–2132. <https://doi.org/10.1038/s41598-019-39391-z>
128. Vogan, K. (2021). Exploring whole exomes. *Nature Research*. <https://doi.org/10.1038/d42859-020-00109-1>
129. Watson, J. D., & Crick, F. H. C. (1953). Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature*, 171(4356), 737–738. <https://doi.org/10.1038/171737a0>
130. Wellcome Trust Case Control Consortium. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145), 661–678. <https://doi.org/10.1038/nature05911>
131. Willerslev, E., & Meltzer, D. J. (2021). Peopling of the Americas as inferred from ancient genomics. *Nature*, 594(7863), 356–364. <https://doi.org/10.1038/s41586-021-03499-y>
132. Wilson, S. L., Way, G. P., Bittremieux, W., Armache, J.-P., Haendel, M. A., & Hoffman, M. M. (2021). Sharing biological data: Why, when, and how. *FEBS Letters*, 595(7), 847–863. <https://doi.org/10.1002/1873-3468.14067>
133. Witte, J. S. (2010). Genome-Wide Association Studies and Beyond. *Annual Review of Public Health*, 31, 9. <https://doi.org/10.1146/annurev.publhealth.012809.103723>
134. Wonkam, A. (2021). Sequence three million genomes across Africa. *Nature*, 590(7845), 209–211. <https://doi.org/10.1038/d41586-021-00313-7>
135. Zhou, S., Xie, P., Quibion, A., Ambalavanan, A., Dionne-Laporte, A., Spiegelman, D., Bourassa, C. V., Xiong, L., Dion, P. A., & Rouleau, G. A. (2019). Genetic architecture and adaptations of Nunavik Inuit. *Proceedings of the National Academy of Sciences of the United States of America*, 116(32), 16012–16017. <https://doi.org/10.1073/pnas.1810388116>
136. Zhou, S., Xiong, L., Xie, P., Ambalavanan, A., Bourassa, C. V., Dionne-Laporte, A., Spiegelman, D., Maude Turcotte Gauthier, Maude Turcotte Gauthier, Maude Turcotte Gauthier, Gauthier, M., Henrion, E., Diallo, O., Dion, P. A., & Rouleau, G. A. (2015). Increased Missense Mutation Burden

of Fatty Acid Metabolism Related Genes in Nunavik Inuit Population. *PLOS ONE*, 10(5). <https://doi.org/10.1371/journal.pone.0128255>

### 13. MATERIAL SUPLEMENTARIO

<b>Tabla S1: Códigos homologados y el nombre completo de los grupos NatAm estudiados en genómica</b>	
<b>Nombre del grupo NatAm</b>	<b>Código de nombre</b>
Pima	PIM
Karitiana	KARIT
Maya	MAY
Surui	SURU
Chane	CHAN
Piapoco	PIA
Mixe	MIX
Mixteco	MIXT
Zapoteco	ZAPO
Quechua	QUE
Aymara	AYM
Tarahumara	TAR
Tepehuano	TEPE
Nahua	NAHU
Totonaco	TOT
Mapuche-Huilliche	MAPU
Mayo	MAYO
Seri	SER
Chinanteco	CHIN
Chocholteco	CHOC
Chontal de Oaxaca	CHONT
Huasteco	HUAS
Matlatzinca	MATL
Mazahua	MAZ
Mazateco	MAZA
Otomi	OTO
Pame	PAM
Tacuate	TACU
Q'anjob'al	QAN
Kaqchikel	KAQ
Mam	MAM
Mocho	MOCH

Popoluca_Sierra	POPO
Tojolabal	TOJ
Tzotzil	TSOT
Zoque	ZOQ
Dakelh	DAK
Inuit	INU
Aleutian_Islander	ALE
Tsimshian	TSIM
Huichol	HUI
Yukpa	YUKP
Chopccas	CHOP
Matzes	MATZ
Moches	MOC
Uros	URO
Charrúa	CHAR
Chipewyan	CHIP
Cree	CRE
Native_American	NATAM
Assurini Kwatinemo	ASSU
Arara	ARAR
Araweté	ARAWE
Awa Guajá	AWAG
Kayapó	KAY
Xikrin do Odjá	XIKO
Xikrin do Katete	XIKK
Zoé (Puturujará)	ZOE
Wajãpi	WAJA
Triqui	TRIQ
Huambisa	HUA
Cashibo	CAS
Shipibo	SHI
Ashaninka	ASH
Yanesha	YANE
Wichi	WIC
Guarani	GUAR
Algonquin	ALG
Ojibwa	OJI

Chilote	CHIL
Inga	ING
Ticuna	TIC
Wayuu	WAYU
Cabecar	CAB
Cochimi	COCH
Cucupa	CUC
Haida	HAID
Kumiai	KUMI
Athabaskan	ATH
Tlingit	TLI
Matsigenka	MATSI
Q'eros	QER
Huilliche-Pehuenche	HUIL
Pehuenche	PEH
Kawéskar	Kawe
Yámana	YAMA
Rapa Nui	RAPAN
Nisga'a	NIS
Stswecem'c	STS
Splatsin	SPL
Purepecha	PUR
Lacandon	LAC
Bari	BARI
Warao	WAR
Apalai	APAL
Urubu-Kaapor	URUB
Xavante	XAV
Zoro	ZOR
Arhuaco	ARH
Bribri	BRI
Chono	CHON
Chorotega	CHORO
Diaguita	DIA
Embera	EMB
Guahibo	GUA
Guaymi	GUAYM
Huetar	HUE

Jamamadi	JAM
Kaingang	KAIN
Kogi	KOG
Maleku	MAL
Palikur	PALI
Parakana	PARA
Teribe	TERI
Toba	TOB
Waunana	WAUN
Yaghan	YAG
Yaqui	YAQ
Kiliwa	KIL
Kuahl	KUA
Paipai	PAIP
Papago	PAP
Guarijio	GUARI
Mexicanero	MEXI
Cora	COR
Tepehua	TEP
Amuzgo	AMU
Chatino	CHAT
Chichimeca	CHI
Cuicateco	CUIC
Ixcateco	IXC
Popoloca	POP
Tlahuica	TLA
Tlapaneco	TLAP
Huave	HUAV
Awakateko	AWAK
Chol	CHO
Chuj	CHUJ
Ixil	IXI
Popticalteco	POPT
Kanjobal	KANJ
Kekchi	KEK
Tzeltal	TSE
Iñupiat	INUP
Kichwa	KICH

Kamentsa	KAME
Cofan	COF
Cocama	COC
Chamicuro	CHAM
Shawi	SHAW
Muniche	MUNI
Wayku	WAYK
Tallanes	TAL
Puno (Aimaran)	PUN
Jaqarus	JAQ
Candoshi	CAND
Awajun	AWAJ
Chachapoyas (Quechuan)	CHA
Lamas (Quechuan)	LAM
Naso	NASO
Ngabe	NGAB
Guna	GUN
Ancient_NatAm_Related	aNATAM
Pre-Dorset	PRED
Saqqaq	SAQ
Dorset	DOR
Thule	THU
Norton	NOR
Clovis	CLO
Pericues	PERI
Pre-Columbian-Mummies	PRECM
Selknam	SEL
Enoque	ENO
Chinchorro	CHINC
Chumash	CHUM
Botocudo	BOT
Lucayan	LUC
PRH_Ancient	aPRH