



BENEMÉRITA UNIVERSIDAD AUTÓNOMA DE PUEBLA

FACULTAD DE CIENCIAS BIOLÓGICAS



**“DESARROLLO DE UNA HERRAMIENTA BIOINFORMÁTICA
PARA EL ANÁLISIS POBLACIONAL DE LA VARIACIÓN
GENÓMICA”**

T E S I S

PRESENTADA PARA OBTENER EL GRADO DE:
LICENCIATURA EN BIOTECNOLOGÍA

PRESENTA:
JUDITH BALLESTEROS VILLASCÁN

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

ASESORES DE TESIS:
DR. CARLOS ALBERTO CONTRERAS PAREDES
DRA. NORMA ANGÉLICA CABALLERO CONCHA



PUEBLA, PUE.

DICIEMBRE 2019

AGRADECIMIENTOS

Agradezco a mis padres Ramón Ballesteros Pérez y Judith Villascán Arana por su amor y apoyo en cada etapa de mi vida académica, por todos los sacrificios que hicieron para que hoy tengan este trabajo en sus manos. A mi hermano, Juan Ramón Ballesteros Villascán por hacerme más llevaderas las tardes sentada trabajando. A toda mi familia, especialmente a mis abuelitos Juan Villascán Arcos, Blanca Aurora Arana González, Ramón Ballesteros Añas y Josefina García, por su apoyo, su cariño y su motivación.

A la Dra. Liliana López Pliego por recibirme y hacerme parte del Laboratorio de Genética Molecular Microbiana junto con el Dr. Miguel Castañeda Lucio, por permitirme aprender tanto y sobre la marcha de la parte experimental de la Genética. Gracias por convertirme en mi amiga, por enseñarme las primeras cosas de Bioinformática que hice y por alentarme a seguir aprendiendo de este maravilloso mundo.

Al M.C. Israel Aguilar Ordoñez, por darme una oportunidad cuando sabía apenas nada de Bioinformática, por tu tiempo cada semana en los primeros talleres y días de discusión de artículos, por permitirnos colaborar más en los grandes proyectos, por las videollamadas a la hora que fuera necesaria si tenía algún problema con el código, por corregirme, por tu guía y por haber hecho posible este trabajo.

Al Dr. Juan Enrique Morett Sánchez, quien con el apoyo del M.C. Israel, me permitió entrar al desarrollo del Proyecto 100G-MX, por ser un ejemplo de humildad, por sus enseñanzas y su confianza.

Al equipo de INMEGEN, el Dr. Cristóbal Fresno Rodríguez por ser mi tutor ante INMEGEN y permitirme acceso al clúster para llevar a cabo muchos análisis del proyecto. A Fernando Pérez, por ser un ejemplo de que siempre hay tiempo de hacer análisis extra, por su pasión a la Ciencia y por enriquecer los jueves de discusión para el proyecto.

A Ram González Buenfil, por ser mi compañero de Laboratorio desde lo experimental hasta el salto a Bioinformática, por todos los Talleres y Cursos a los que fuimos juntos, por tu ayuda cuando me atoraba con algún comando, por las mini discusiones que hacíamos, por las veces en que trabajábamos turnándonos la silla mientras platicábamos, por convertirte en mi amigo y porque fuiste una persona clave para que ambos llegáramos a este proyecto.

A mi novio Jorge Antonio Amador Balderas, por tranquilizarme cuando me sentía abrumada, por motivarme a que siempre puedo hacerlo, por impulsarme a lograr cualquier cosa, por escucharme y tratar de entender sobre todas las cosas de código que te platico y de Genómica Humana, aunque no sea tu fuerte, por ayudarme en todos los trámites para llegar aquí, por las veces en que nuestras citas eran para trabajar juntos, por tu apoyo incondicional y por tu amor.

A mis amigos de la Licenciatura, Alejandra Fuentes, Rafael López, Luz Alondra Hernández, Eric Navarro, Jesús López, Oswaldo Sánchez y Jael Flores por su amistad, por ser estudiantes ejemplares, por las comidas y festejos que mejoraron mucho estos años de Licenciatura juntos.



Secretaría
de Educación
Gobierno de Puebla

CONCYTEP

Consejo de Ciencia
y Tecnología del Estado
de Puebla

*“Esta investigación fue realizada gracias al apoyo del Consejo de Ciencia y
Tecnología del Estado de Puebla”*

DEDICATORIA

A mis padres Judith Villascán Arana y Ramón Ballesteros Pérez que me han apoyado y hecho tantos sacrificios por mí para darme todo lo que tengo ahora.

Nunca tendré suficiente para agradecerles, los amo.

A mi hermano Juan Ramón Ballesteros Villascán, de quien espero ser un buen ejemplo a seguir y que quiero tanto.

A mis abuelitos, mis tíos y familiares que me han apoyado y motivado en mi vida estudiantil.

A mi novio Jorge Antonio Amador Balderas, mi apoyo durante el desarrollo de este trabajo y a quien admiro y quiero demasiado.

ÍNDICE GENERAL

1. RESUMEN	9
2. INTRODUCCIÓN	11
2.1 LA REVOLUCIÓN GENÓMICA	11
3. ANTECEDENTES	14
3.1 PROCESO DE DETECCIÓN DE VARIANTES.....	14
3.2 PROYECTOS POBLACIONALES A NIVEL MUNDIAL	18
3.2.1 HapMap.....	19
3.2.2 El Proyecto de los 1000 Genomas.....	19
3.2.3 ExAC.....	21
3.2.4 gnomAD	22
3.3 PROYECTOS POBLACIONALES EN MÉXICO	23
3.3.1 Estudio de ancestría nativa de México a través de microarreglos	23
3.3.2 Secuenciación de genomas completos de Nativos Americanos	24
3.3.3 Secuenciación de exomas de mexicanos con ancestría nativa americana	26
4. MARCO CONCEPTUAL	27
4.1 DBSNP.....	28
4.2 VARIANT EFFECTOR PREDICTOR (VEP)	28
4.3 BASES DE DATOS CON ASOCIACIONES CLÍNICAS DE VARIANTES	28
4.3.1 GWAS Catalog.....	29
4.3.2 ClinVar	30
5. PLANTEAMIENTO DEL PROBLEMA	30
6. HIPÓTESIS	31
7. OBJETIVOS.....	31
7.1 GENERAL.....	31
7.2 ESPECÍFICOS	31
8. MATERIAL Y MÉTODOS	32

8.1 METODOLOGÍA	32
8.2 MUESTRAS.....	33
8.3 DESCRIPCIÓN BÁSICA DE LOS DATOS	34
8.4 IDENTIFICACIÓN DE VARIANTES DE INTERÉS BIOMÉDICO.....	35
8.5 CURACIÓN DE VARIANTES DE INTERÉS BIOMÉDICO	38
8.6 <i>PIPELINES</i> AUTOMATIZADOS.....	38
9. RESULTADOS.....	40
9.1 DESCRIPCIÓN BÁSICA DE LOS DATOS	40
9.2 CATALOGACIÓN DE LOS DATOS.....	41
9.3 VARIANTES DE INTERÉS BIOMÉDICO	42
10. DISCUSIÓN	50
11. CONCLUSIONES	53
12. BIBLIOGRAFÍA	54
13. ANEXOS	69

ÍNDICE DE FIGURAS

Figura 1. ¿Qué son las variantes genéticas? Clasificación e impacto.	15
Figura 2. Abordajes de secuenciación y microarreglos de ADN.	16
Figura 3. Descripción y ejemplos de algunos formatos usados en Bioinformática.	18
Figura 4. ¿De dónde viene el rsID?	25
Figura 5. Definición de Frecuencia Alélica. Concepto clave en este trabajo.	25
Figura 6. Número de variantes por individuo de cada grupo étnico	40
Figura 7. Consecuencia anotada por VEP de acuerdo al tipo de variante.	41
Figura 8. Frecuencia de variantes biomédicamente relevantes de 100GMX, europeos y asiáticos del este, reportadas en ClinVar o GWAS Catalog	43
Figura 9. Diagrama de Venn de las variantes de interés biomédico con frecuencias alélicas diferenciales entre la población de nativos mexicanos y otras poblaciones a nivel mundial.	44

ÍNDICE DE TABLAS

Tabla 1. Poblaciones del Proyecto 1000 Genomas	20
Tabla 2. Exomas y Genomas secuenciados por población, y subpoblaciones que las conforman del proyecto gnomAD.	22
Tabla 3. Repositorio de pipelines utilizados	39
Tabla 4. Grupos de variantes separados por frecuencias, tipo y categoría.	42
Tabla 5. rsID de variantes de interés biomédico con las frecuencias diferenciales en el proyecto de 100G-MX referidos como NatMex, y otras poblaciones	45

ÍNDICE DE ESQUEMAS

Esquema 1. Diagrama de Trabajo	33
Esquema 2. Funcionamiento del pipeline para la descripción básica de los datos	34
Esquema 3. Funcionamiento de pipeline para la identificación de variantes de interés biomédico	36

1. RESUMEN

El genoma humano consta de 3 mil millones de pares de bases, sin embargo, no todos los genomas son iguales, si no que existen diferencias llamadas polimorfismos o variantes, que van de un solo nucleótido (SNV) o unas cuantas pares de bases a fragmentos de cromosomas. Estas variaciones pudieron empezar a ser estudiadas para conocer sus impactos o si guardaban relaciones clínicas gracias a la Genómica Poblacional, con proyectos que secuenciaron cientos o miles de individuos a nivel mundial. Sin embargo, algunas regiones del mundo no fueron tan estudiadas, como es el caso de México, además por ser un país con eventos de mestizaje, se iniciaron algunos proyectos con el objetivo de conocer la variación de los nativos amerindios mexicanos.

Los primeros proyectos a nivel nacional implicaban el uso de microarreglos o exomas, pero también se llevaron a cabo iniciativas que buscaban secuenciar genomas completos. Ya se han reportado algunas variantes con información de fenotipos relacionados por estos proyectos, sin embargo, la búsqueda de información sobre cientos de miles de variantes es una labor que toma mucho tiempo y recursos.

En el presente trabajo se realizó el desarrollo de una herramienta bioinformática que permite el estudio de la variación genómica en proyectos poblacionales, validándola con los datos de genoma completo del proyecto 100G-MX, el cual es un proyecto de caracterización de genomas nativos mexicanos que explora el componente nativo mediante secuenciación de genoma completo y busca inferir la historia demográfica de los antepasados nativos americanos, con el objetivo de obtener variantes biomédicamente relevantes. Se describió el conjunto de datos de las variantes, obteniendo el número total de variantes por individuo y el total, a partir de lo cual se seleccionaron aquellas que se encontraban con algún fenotipo biomédico reportado en las bases de datos ClinVar y GWAS Catalog, que además mostraran frecuencias diferenciales en comparación con otras poblaciones a nivel mundial. Se hizo una revisión bibliográfica de estas variantes catalogadas como

biomédicamente relevantes, acompañada de la descripción del impacto, gen al que pertenecía (en el caso de variantes en regiones codificantes) y un resumen breve sobre con qué rasgo se reportaba esta variante. Interesantemente, se encontraron variantes relacionadas con fenotipos previamente reportados dentro de la población mexicana, como el aumento del índice de masa corporal, diferencias en los niveles de colesterol, y otros que no han sido estudiados profundamente como hábitos de consumo de alcohol, depresión y funciones cognitivas. Es importante mencionar que no se encontraron variantes reportadas como no benignas o patogénicas por ClinVar para este conjunto de datos, y en el caso de las variantes reportadas por estudios de asociación, la mayoría era de individuos europeos, y sólo uno incluía individuos con ancestría americana, lo cual resalta la importancia y necesidad de tener más proyectos de secuenciación de poblaciones nativas y mestizas en México que contribuyan al desarrollo de la Medicina Genómica en México, y se espera que el desarrollo de esta herramienta facilite la identificación de variantes de interés biomédico en los futuros proyectos poblacionales.

2. INTRODUCCIÓN

2.1 La revolución genómica

En el año de 1900, DeVries, Correns y Tschermak de forma independiente redescubrían el trabajo de Gregor Mendel^[1,2], las llamadas leyes de la herencia, que sentarían la base para el desarrollo de la Genética. En 1953, los científicos Watson & Crick identificaron a la molécula de Ácido Desoxirribonucleico (ADN) como la base química de la herencia y dedujeron su estructura gracias a la famosa “fotografía 51” tomada por Rosalind Franklin a través de difracción de rayos X^[3]. Conociendo la estructura de esta molécula, los químicos y biólogos buscaron conocer el orden y secuencia de los nucleótidos de la molécula de ADN, es así como en 1977 Sanger y Gilbert desarrollaron métodos para secuenciar el ADN basados en el marcaje radiactivo de nucleótidos y reacciones químicas que fragmentan o polimerizan al ADN^[4,5], de forma paralela, los genetistas usaron marcadores de ADN para poder ubicar genes en regiones cromosomales, desarrollando mapas de ligación basados en patrones de herencia de alelos^[6]. Posteriormente, en la década de 1990s, se logró secuenciar el código genético completo de distintos organismos, siendo los primeros genomas secuenciados algunos virus y bacterias debido a su facilidad de estudio^[7-9]; tras el éxito de secuenciar genomas pequeños, surgió la iniciativa de conocer la secuencia completa del genoma humano, dando inicio el Proyecto del Genoma Humano en 1990^[10].

En el año 2001 se publicó el primer borrador del genoma humano, siendo esta la culminación de un gran esfuerzo colaborativo tecnológico y científico^[11,12]. Una de las primeras sorpresas del genoma humano fue que la cantidad genes que codifican para proteínas era mucho menor a lo que se esperaba, dando lugar a nuevas preguntas como: ¿qué detalles genómicos nos hacen diferentes de otras especies?, ¿qué tan diferente es el genoma de dos personas?

Gracias al avance de tecnologías masivas de secuenciación, recientemente se han completado los primeros proyectos de secuenciación completa y profunda de

genomas de la población general. A diferencia del proyecto original del genoma humano (donde se reportó un solo genoma, integrado a partir de muestras voluntarias), los proyectos de secuenciación poblacional permiten analizar las diferencias genómicas entre personas a través de la identificación de mutaciones presentes en cada individuo (propriadamente denominadas “variantes”, para evitar la asociación negativa con el término “mutación”). El estudio de las variantes particulares de los individuos y la variación compartida por grupos poblacionales forma la base de las iniciativas en medicina genómica, particularmente en el área de la farmacogenómica (relativa a cómo las variantes afectan la forma en que un individuo metaboliza los fármacos) y la medicina personalizada (el uso de la información genómica particular de cada individuo para mejorar la prevención, diagnóstico y tratamiento de enfermedades).

Aunque previamente existieron proyectos de análisis de la variación poblacional a nivel mundial basados en tecnologías de genotipificación por microarreglos^[13] (que son placas de vidrio que contienen moléculas de ADN de posiciones específicas del genoma, que sirven para detectar variantes en una muestra de ADN), dichas tecnologías permitían la comparación de una fracción pequeña del genoma humano. El proyecto pionero de secuenciación profunda poblacional, basado en tecnologías de secuenciación masiva, fue el Proyecto de los Mil Genomas, el cual consistió en la primera descripción detallada de la variación entre los genomas de distintas poblaciones del mundo^[14]. Otros proyectos más extensos le siguieron, como el proyecto gnomAD^[15] una descripción general de la variación en decenas de miles de genomas completos, y cientos de miles de exomas (conjunto de exones; son los fragmentos de ADN transcritos para dar lugar a las proteínas); o el UK10K, una búsqueda por marcadores genómicos en patologías de interés para el Reino Unido que permitan distinguir entre pacientes enfermos y sanos^[16], lo cual sentará las bases para una mejora directa de sus servicios de salud.

Por otro lado, gracias a las iniciativas de medicina personalizada puestas en marcha en distintos países^[17–23], se pronostica que en el futuro no muy lejano la mayoría de las personas tendrá acceso a su genoma completo. La generación y acumulación

de datos genéticos humanos, junto con metadatos fenotípicos (edad, estatura, índice de masa corporal, padecimientos clínicos, respuestas a tratamientos, etc.) han dado paso a una etapa denominada la Revolución Genómica, etapa en la que el conocimiento detallado de los genomas y su relación con fenotipos de interés clínico harán posible la prevención oportuna y el tratamiento personalizado de las enfermedades.

En estos inicios la Revolución Genómica se enfrenta principalmente a dos tipos de problemas: limitaciones en tecnologías de la información, y falta de conocimiento biológico profundo sobre las causas genéticas (de haberlas) en múltiples padecimientos. Por lo anterior, la bioinformática ocupa un rol central en la Revolución Genómica.

La secuenciación de genomas completos provee millones de datos crudos, es decir, datos sobre la variación de los individuos que no incluyen información clínica de interés. En México, actualmente contamos con la información de 12 genomas completos amerindios representativos de algunos pueblos originarios, junto con datos de exoma y microarreglos^[24–26]. Asimismo, actualmente se está llevando a cabo un proyecto por el Instituto Nacional de Medicina Genómica (INMEGEN), el cual cuenta con 76 genomas completos de poblaciones nativas mexicanas (NatMex). Las variantes reportadas por estos datos genómicos son importantes ya que representan hasta el 50% de las variantes que se encontrarán en la población general, dado que la población mexicana es mayoritariamente mestiza^[27,28].

En el presente proyecto analizaremos el conjunto de datos poblacionales de los 76 genomas nativos mexicanos utilizando bases de datos fenotípicos actualizadas, para identificar y caracterizar *in silico* variantes de interés clínico que sirvan como puntos de partida para futuras líneas de investigación en medicina genómica de relevancia nacional. Paralelamente, nuestro proyecto está basado en el diseño, desarrollo y validación de una herramienta bioinformática para el análisis reproducible de proyectos genómicos poblacionales; el propósito principal de la herramienta es agilizar el paso desde la detección de polimorfismos hasta el análisis de variantes de una manera amigable con los usuarios, lo cual facilitará el uso de

datos genómicos masivos y públicamente disponibles en proyectos del ámbito de la Medicina Genómica y la Genómica de Poblaciones.

3. ANTECEDENTES

3.1 Proceso de detección de variantes

El genoma humano consta de aproximadamente 3 mil millones de pares de bases y cerca de 20,000 genes que codifican proteínas^[29]. Todos los humanos son diferentes uno del otro, y la mayoría de estas diferencias tiene una base genética, es decir, diferencias en el fenotipo son reflejo de diferencias en el genotipo^[30]; algunas de estas diferencias son causadas por la tasa de mutación en humanos que va de $1.0 - 1.5 \times 10^{-8}$ pb, lo cual permite transmitir ~30 nuevas variantes en cada gameto^[31]. Si se comparan las diferencias entre dos genomas de humanos seleccionados al azar, cerca del 1.2% son diferencias que surgen por mutaciones, propiamente llamadas polimorfismos^[32,33], las cuales son responsables de la variación entre poblaciones^[34], por lo que también se les llama “variantes”.

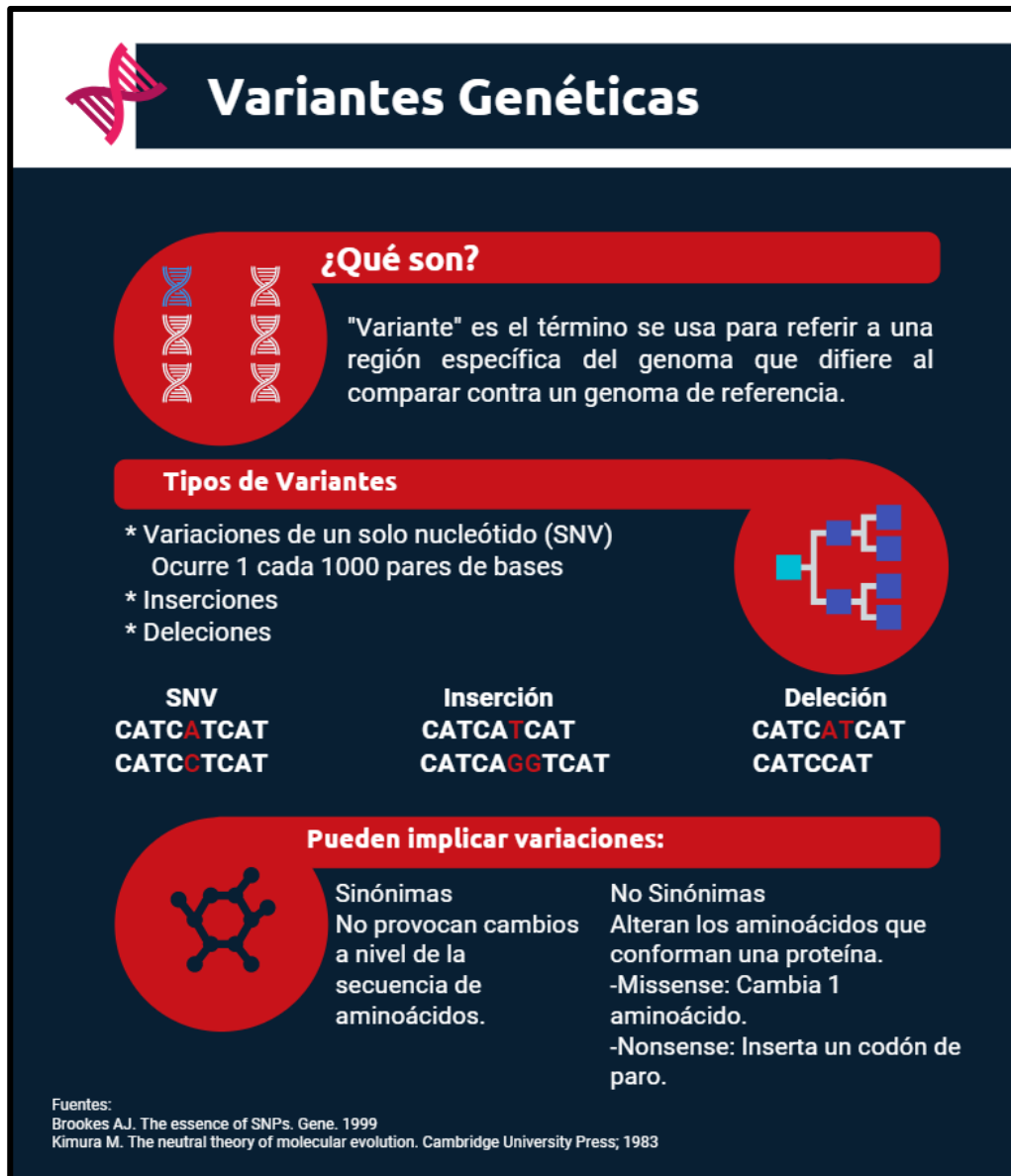


Figura 1. ¿Qué son las variantes genéticas? Clasificación e impacto.

Hoy en día las técnicas de secuenciación desarrolladas para el Proyecto del Genoma Humano han quedado en segundo plano, siendo las tecnologías de secuenciación de siguiente generación (Next Generation Sequencing, NGS) las que las han superado en cantidad de lecturas generadas, rapidez y abaratamiento de costos, incluso reemplazado en algunos aspectos a los microarreglos como la plataforma líder para la investigación del genoma^[12,35–37].

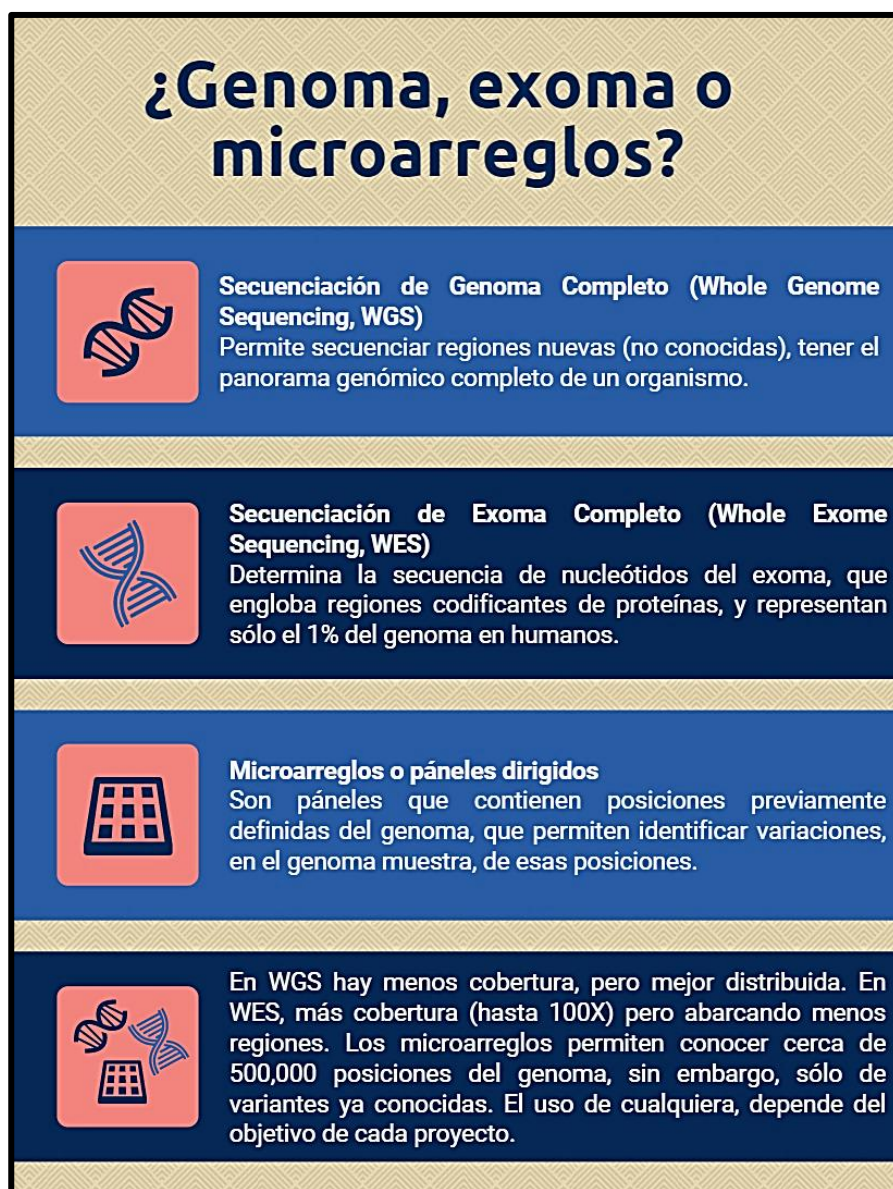


Figura 2. Abordajes de secuenciación y microarreglos de ADN.

La principal plataforma para la obtención de datos NGS es la secuenciación por síntesis, de Illumina; dicha tecnología permite digitalizar muestras biológicas, dando como resultado una enorme cantidad de datos (en el rango de Gigabytes para genomas completos; Terabytes en proyectos poblacionales compuestos por cientos de muestras individuales). Un factor importante a considerar en un proyecto de secuenciación ya sea de exoma o de genoma completo, es la “cobertura”, que se refiere al promedio de veces en que un nucleótido fue secuenciado.

Pero, los datos crudos de una secuenciación aún contienen errores y están en desorden, de tal manera que no es posible hacer análisis directamente sobre este tipo de datos; se requieren herramientas bioinformáticas que permitan integrar y analizar los datos para poder generar hipótesis biológicas de interés. Un ejemplo de estas es “Genome Analysis Tool Kit” (GATK) que es un kit de herramientas usado desde la línea de comando para análisis de datos de secuenciación genómica principalmente para el descubrimiento de variantes^[38]. Los pipelines (conjunto de transformaciones para procesar los datos crudos a algo interpretable, o flujos de trabajo) de GATK comprenden dos etapas; la del preprocesamiento de los datos crudos para producir archivos BAM (Fig. 3) alineados a un genoma de referencia y corregidos del sesgo técnico listos para el análisis; y la siguiente etapa que identifica la variación genómica con métodos de filtración, generando un archivo de salida en formato VCF^[39]. El formato VCF (Fig. 3) contiene datos para cada posición en el genoma donde se encontró un polimorfismo^[40].

Formatos en Genómica Humana

FASTQ

```
@read1
AGCTTATCCTCTGCTCACCCCGGGTAGCGCACTTGATGTATTCACAGC
+
BA1@CC7CBCCC9C8;B2@>C?B@B@B3=9?@B1:AB7B?B8B?B6B.7.
```

Archivo de texto que contiene información sobre los fragmentos de DNA y calidad de su secuenciación en formato hexadecimal.

BAM/SAM

```
SRR013667.1 99 19 8882171 60 76M = 8882214 119
NCCAGCAGCCATAAAGTGAATGGGAAATAAACACTATGTTCAAAGCAGA
#>A@BABAAAAADDEGCEFDHDEDBCFDBCDBCBDCEACB>AC@CDB@>
...
```

El archivo SAM contiene las coordenadas donde alineó cada fragmento del FastQ. El BAM es su formato binario para reducir el tamaño de los archivos generados.

VCF

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	GEN
chr1	82154	rs4477212	a	.	.	.	.	GT	0/0

Reporta cada posición variante en el genoma humano, en comparación con el genoma de referencia. Se infiere a partir de las alineaciones BAM.

Fuentes:

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011.
<http://samtools.github.io/hts-specs/SAMv1.pdf>
<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>

Figura 3. Descripción y ejemplos de algunos formatos usados en Bioinformática.

3.2 Proyectos poblacionales a nivel mundial

Después de la publicación del Proyecto del Genoma Humano, se iniciaron líneas de investigación para el estudio de la variación genómica en distintas poblaciones del mundo, principalmente aquellas poco o nada representadas, para conocer su estructura genómica y dinámica poblacional. A continuación, se sintetizan los principales proyectos de genotipificación poblacional.

3.2.1 HapMap

El proyecto HapMap^[13], lanzado en 2002 y concluido en 2010, tenía como objetivo determinar patrones de variación común o haplotipos (definidos como un conjunto específico de alelos observados en un solo cromosoma, o parte de un cromosoma) en distintas poblaciones. El HapMap provee el primer recurso público para entender la variación genómica en 4 poblaciones geográficamente diversas (África, Europa, Sur de Asia y Este de Asia)^[41,42]. Este proyecto se llevó a cabo conjuntando el uso de microarreglos para la detección de variantes, y fue útil para iniciar la búsqueda de genes que contribuyen a la enfermedad o a la respuesta de fármacos en poblaciones diversas^[41]. Al estar basado en tecnología de microarreglos, este proyecto pudo cuestionar apenas un par de millones de posiciones del genoma humano; sin embargo, HapMap 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^[43,44], la identificación y correlación de la estratificación de poblaciones y su parentesco^[45,46].

3.2.2 El Proyecto de los 1000 Genomas

El Proyecto de los 1000 Genomas dio inicio en 2009 y se concluyó en 2015, logrando caracterizar el 95% de las variantes que se encontraban en regiones genómicas secuenciadas de 2504 individuos, provenientes de 5 regiones continentales (África, Europa, Asia del sur, Asia del este y América (Tabla 1). Además de detectar variantes comunes (con frecuencias alélicas iguales o mayores al 1% en la población mundial), también se reportaron variantes raras con frecuencias alélicas menores al 0.1% en regiones codificantes del genoma.

Es importante mencionar que las variantes de baja frecuencia en su mayoría fueron variaciones potencialmente funcionales, como aquellas que modifican una proteína bajo señal de selección^[47–49]. Además, debido a su baja frecuencia tienden a ser recientes en su origen ya que exhiben un alto nivel de diferenciación entre poblaciones, por lo que son sitios de variación de interés para su estudio^[49–51]. En total, el proyecto de los 1000 Genomas reportó aproximadamente 80 millones de

variantes, el catálogo de variación genómica humana más grande y detallado hasta ese momento.

En el proyecto de los 1000 genomas se muestrearon individuos de 26 poblaciones que se agruparon por su ancestría predominante:

Tabla 1. Poblaciones del Proyecto 1000 Genomas

Superpoblación	Subpoblación	Superpoblación	Subpoblación
África	Yorubas de Ibadan Luhya en Webuye, Kenya Personas con ancestría africana en el sureste de Estados Unidos Individuos de las divisiones occidentales de Gambia Mende en Sierra Leona Caribeños africanos en Barbados Esan en Nigeria	Asia del Este	Chinos Han en Beijing Japoneses en Japón Tokio Chinos de la dinastía Han del sur de China Kinh en la ciudad de Ho Chi Minh, Vietnam Dai chino en Xishuangbanna, China
Europa	Residentes de Utah con ancestría del Norte y Este de Europa Toscani en Italia Británicos de Inglaterra y Escocia Finlandeses en Finlandia	América	Personas con ancestría mexicana viviendo en Los Ángeles, California Colombianos en Medellín, Colombia Puerto Riqueños en Puerto Rico Peruanos de Lima, Perú

	Poblaciones ibéricas en España		
Asia del Sur	Bengalíes de Bangladesh Tamil de Sri Lanka del Reino Unido Indios Telugu del Reino Unido Punjabi de Lahore, Pakistán Indios Gujarati de Houston, Texas		

3.2.3 ExAC

El Exome Aggregation Consortium (ExAC; <http://exac.broadinstitute.org>), publicado en 2016, fue un proyecto que incrementó la cantidad de variantes de alta calidad (muchos exomas secuenciados con una buena cobertura) conocidas en diferentes poblaciones a nivel mundial. El conjunto de datos final abarca 60,706 exomas humanos que pertenecen a poblaciones distintas: África, Europa, Asia del sur, Asia del este, y América (referida como población latina), además de separar a la población europea debido a un gran cuello de botella en Finlandeses y No-Finlandeses; la anotación de variantes, proceso por el cual se identifica el posible impacto funcional de los polimorfismos, se hizo con la herramienta “Variant Effect Predictor” (VEP) versión 8^[52]. El proyecto ExAC identificó 7,404,909 variantes de alta calidad, incluyendo 317,381 inserciones o deleciones (indeles). A pesar de que el exoma compone apenas el 1% del genoma humano, el 72% de las variantes encontradas por ExAC estaban ausentes de los datos encontrados por proyectos de secuenciación de genoma completo como el proyecto de los mil genomas^[14] y Exome Sequencing Project (ESP)^[53,54] este dato refleja la importancia de actualizar periódicamente los procesos de detección de variantes en el acervo público de

datos genómicos, así como el incluir más muestras, con mayor diversidad poblacional^[54].

3.2.4 gnomAD

Tabla 2. Exomas y Genomas secuenciados por población, y subpoblaciones que las conforman del proyecto gnomAD.

Tomado de: <https://macarthurlab.org/2018/10/17/gnomad-v2-1/>

Población	Descripción	Genomas	Exomas	Total
Afr	Africanos/africanos-americanos	4,359	8,128	12,487
Amr	Latinos	424	17,296	17,720
Asj	Ashkenazi judíos	145	5,040	5,185
Eas	Asiáticos del Este	(780)	(9,197)	(9,977)
Kor	Coreanos	0	1,909	1,909
Jpn	Japoneses	0	76	76
Oea	Otros asiáticos del este	780	7,212	7,992
Fin	Finlandeses	1,738	10,824	12,562
Nfe	Europeos no-finlandeses	(7,718)	(56,885)	(64,603)
Bgr	Búlgaros	0	1,335	1,335
Est	Estonios	2,297	121	2,418
Nwe	Europeos del noreste	4,299	21,111	25,410
Seu	Europeos del sur	53	5,752	5,805
Swe	Europeos del sureste	0	13,067	13,067
Onf	Otros europeos no-finlandeses	1,069	15,499	16,568
Sas	Sur asiáticos	0	15,308	15,308
Oth	Otros (población no asignada)	544	3,070	3,614
Total		15,708	125,748	141,456

Conceptualmente una extensión de ExAC, el “Genome Aggregation Database” (gnomAD), actualizado por última vez a principios de 2019, agregó las secuencias de 15,708 genomas y 125,748 exomas, siendo la base de datos

públicamente disponible con mayor número de secuencias de alta cobertura hasta la fecha. La anotación de variantes se llevó a cabo usando la herramienta VEP versión 85^[52]. Los datos son públicamente accesibles en <https://gnomad.broadinstitute.org>. Actualmente el mayor catálogo de variantes poblacionales, gnomAD reporta aproximadamente 17 millones de variantes en exoma y 261 millones en genoma^[15].

3.3 Proyectos poblacionales en México

Gracias a los proyectos anteriormente mencionados se conoce la variación genómica general a nivel mundial. Pero variantes raras, con frecuencias menores suelen ser específicas de poblaciones no incluidas en proyectos internacionales, por ejemplo, la población amerindia mexicana. Dado que el genoma promedio de la población mexicana está compuesto hasta en un 50% por la aportación amerindia, la variación presente en las poblaciones indígenas de México puede ser relevante médica y funcionalmente o representativa de nuestra genómica nacional^[55]. La población mexicana es el resultado de un complejo proceso de mestizaje^[26,56]. Sin embargo, sólo se han reportado pocas secuencias que permitan la exploración de variantes funcionales y médicamente relevantes que albergan las poblaciones indígenas^[18,55,57–59]. El estudio de variantes funcionales particulares han arrojado interesantes resultados como en el caso del haplotipo SLCA11, el cual se ha identificado como un haplotipo de riesgo, que explica cerca del ~20% del incremento de la Diabetes Tipo 2 en mexicanos, en comparación con poblaciones de ancestría europea^[60]. A continuación, se exponen los enfoques metodológicos de los proyectos en población mexicana recientes.

3.3.1 Estudio de ancestría nativa de México a través de microarreglos

Moreno Estrada et al. ^[24], examinaron patrones locales de variación utilizando microarreglos (Affymetrix 6.0 y Affymetrix 500K) con 785, 663 SNPs informativos, en 511 individuos conformados por muestras de poblaciones indígenas y de individuos mestizos provenientes de 28 poblaciones mexicanas. Además de definir un patrón general de distribución de grupos ancestrales, los autores lograron

detectar variantes debidas a la ancestría nativa posiblemente relacionadas con fenotipos de función pulmonar en la población mestiza mexicana.

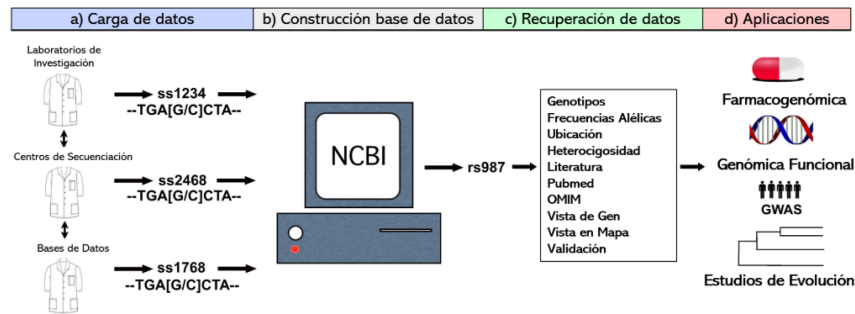
3.3.2 Secuenciación de genomas completos de Nativos Americanos

En 2017, Romero-Hidalgo et al.^[25] reportaron la secuenciación completa de 12 individuos amerindios mexicanos provenientes de 6 grupos étnicos (Tarahumara y Tepehuano del Norte; Nahuatl, Totonaca y Zapoteca del Sur; y Maya del Sur-Este) y un trío mexicano mestizo (madre, padre e hijo) los cuales fueron analizados en conjunto con datos del Panel de Diversidad Nativa Americana (NMPD) (compuesta por 362 muestras de 15 poblaciones indígenas genotipadas por la Universidad de California, San Francisco, usando Affymetrix) y 466 muestras adicionales del Instituto Nacional de Medicina Genómica (INMEGEN) genotipadas utilizando una combinación de microarreglos de Affymetrix e Illumina (las muestras incluyen a 370 mexicanos cosmopolitas de 10 estados diferentes y 96 muestras de tres poblaciones indígenas, colectadas como parte del Proyecto de Diversidad Genómica Mexicana (GDM))^[28]. Las lecturas de los genomas completos se alinearon al genoma de referencia (GRCh37) y las variantes fueron llamadas usando la versión 2.4 del Software Complete Genomics^[61]. Las consecuencias funcionales de las variantes no sinónimas fueron predichas utilizando el indicador Polyphen-2 (Polymorphism Phenotyping v2), para identificar variantes con efectos deletéreos^[62]. Tres muestras adicionales de tres Tarahumaras fueron secuenciadas por Illumina NextSeq 500 en la Unidad de Secuenciación de INMEGEN. Anotadas con el software snpEff para la predicción de impacto funcional^[63]. En cuanto a variantes de interés biomédico, solo se reportaron variantes no sinónimas con registro en la base de datos ClinVar para identificar patogenicidad potencial^[25]; cabe señalar que actualmente es posible expandir la búsqueda de variantes de interés clínico utilizando más bases de datos, tales como GWAS Catalog y ClinVar.

dbSNP

dbSNP es una base de datos de dominio público que contiene una colección de polimorfismos genéticos (SNPs). Incluye información del contexto de la secuencia, la frecuencia en una población, y métodos experimentales (si los hay). Acepta variantes de cualquier especie de cualquier parte del genoma.

Flujo de Datos en dbSNP



Como se esquematiza, una vez que se sube un nuevo SNP, se le asigna un identificador único (ss##), se alinea a la referencia y si hay varias variantes de esa posición, se le llama "grupo SNP de referencia" y se le asigna un solo ID (rs##). Se hacen las anotaciones de otra información usando recursos contenidos en NCBI, para que los usuarios puedan tener el contexto de información relacionada a la variante. El formato final descargable de dbSNP es un VCF.

Fuentes:

Sherry ST, Ward M-H, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* 2001.

Ilustración modificada de: https://en.wikipedia.org/wiki/DbSNP#/media/File:DbSNP_diagram_no_caption.jpg

Figura 4. ¿De dónde viene el rsID?

Frecuencia Alélica:

- Representa la incidencia de una posición del genoma (variante) en una población.
- Se calcula al dividir el número de veces que un alelo de interés es observado en una población (*alelos contados*, AC) entre el número de copias total de todos los alelos en un locus genético particular de la población (*número de alelos*, AN).
- Pueden ser representadas como decimal, porcentaje o fracción.

Fuente:

<https://www.nature.com/scitable/definition/allele-frequency-298/>

Figura 5. Definición de Frecuencia Alélica.

Concepto clave en este trabajo.

El proyecto de los 12 genomas identificó algunos alelos relevantes biomédicamente, entre los que destacan (rs199472885, *R312C*), mutación que portaba un individuo en el gen *KCNH2*, y que es una mutación previamente encontrada en pacientes con síndrome congénito de QT largo^[64]; rs80358606 es otra variante que ClinVar registra como clínicamente incierta; también se identificó una variante que provocaba la pérdida de función de la proteína al cambiar una Treonina por Serina en la posición 187 de la proteína codificada por el gen *SCN5A*, esta variante la portaba un individuo con síndrome de Brugada^[65]. Se compararon las frecuencias alélicas de variantes encontradas en los 12 Genomas amerindios con otras poblaciones del proyecto de los mil genomas (europeos y asiáticos), mostrando que algunas de las variantes que difieren más habían sido reportadas previamente involucradas en señales de selección natural, como rs1834640 que es una variante de pigmentación de la piel cercana al gen *SLC24A5*, la cual difiere en frecuencia de los europeos, pero tiene valores similares con los asiáticos. Además, se reportó que la variante rs765132, asociada con la respuesta a terapia por factor de necrosis tumoral alfa en la enfermedad inflamatoria intestinal^[66], se encontraba en una frecuencia alélica el 86% de los 12 Genomas, pero tan sólo en el 5% de europeos y ausente en asiáticos^[25]. Este tipo de variantes particularmente presentes en las poblaciones amerindias son cada vez más importantes para el área biomédica.

Actualmente también está en marcha el proyecto 100G-MX por el INMEGEN, el cual buscaba en un inicio conocer la variación de 100 genomas completos de nativos amerindios mexicanos, sin embargo, por cuestiones técnicas y metodológicas, el proyecto continúa con 76 genomas completos. Considerado una expansión del proyecto de los 12 Genomas, 100G-MX será el proyecto más extenso en su tipo a nivel nacional. Las herramienta bioinformática que proponemos desarrollar en el presente trabajo, servirá para la descripción y detección de variantes de interés biomédico en el proyecto 100G-MX.

3.3.3 Secuenciación de exomas de mexicanos con ancestría nativa americana

Recientemente en proceso de publicación, Ávila-Arcos et al^[26]. reportan la secuenciación de exomas proveniente de 78 individuos de cinco diferentes grupos

indígenas del Norte (Rarámuri o Tarahumara y Huichol), Centro (Nahua), Sur (Triqui) y Sureste (Maya). Se reportan 58,968 variantes de nucleótido único, con 4,181 SNPs previamente no reportados (aunque la base de datos utilizada fue dbSNP142 una versión no actualizada). El 67.5% fueron variantes no sinónimas y de baja frecuencia. Algunas de las variantes encontradas eran candidatos a evolución adaptativa, pertenecientes a genes involucrados en metabolismo, pigmentación de la piel o sistema inmune. La predicción de impacto funcional se hizo con la herramienta ANNOVAR (versión 2015Jun17)^[67], con lo cual se identificaron variantes de interés biomédico^[26].

Este estudio permitió identificar algunas señales candidatas a evolución adaptativa en ancestría nativa americana, incluyendo genes que ya habían sido identificados como blancos de selección en otras poblaciones como *SLC24A5* de pigmentación de la piel y *FADS3* involucrado en el metabolismo de lípidos^[68]. Interesantemente también se encontraron variantes en genes relacionados con la respuesta del sistema inmune como *SYT5* y *TBC1D10C* involucrados en la respuesta innata y adaptativa respectivamente. También se encontraron genes involucrados en espermatogénesis (*GMCL*) y angiogénesis (*NCKIPSD*)^[26].

4. MARCO CONCEPTUAL

Los proyectos poblacionales a nivel nacional e internacional descritos en las secciones 2 y 3 sirvieron como la base para el estudio a gran escala del genoma humano, mejorando el análisis de llamado de variantes, la identificación y correlaciones entre las poblaciones, de tal forma que mejoraran los datos de producción primaria, reduciendo los errores y mejorando el análisis en todas sus etapas, identificación filtración y mapeo. Además, ampliaron la representación de la diversidad y variación genómica humana, haciendo públicos los datos generados, desde los datos secuenciados hasta sus metadatos, sirviendo como punto de

partida para los subsecuentes estudios poblacionales o de asociaciones a fenotipos clínicos.

De manera paralela a la aparición de estos proyectos, surgieron bases de datos y herramientas que permitirán el análisis de este proyecto. A continuación, se describen para abordar correctamente el diseño metodológico.

4.1 dbSNP

dbSNP (<https://www.ncbi.nlm.nih.gov/snp/>) es la base de datos pública que recopila todas las variantes conocidas en el genoma humano. Integrada por el “National Center for Biotechnology Information” (NCBI), siendo su versión más actualizada la 152 (Abril 2019), es el principal recurso para el estudio general de la genómica de poblaciones, farmacogenómica y asociaciones fenotípicas^[69].

4.2 Variant Effector Predictor (VEP)

VEP (<https://www.ensembl.org/Tools/VEP>) es una herramienta bioinformática que permite la anotación de información biológica para cada variante, indicando si una variante afecta un gen o una región no codificante, y brindando información poblacional como la frecuencia de esa variante en distintas poblaciones (reportada por 1000G y gnomAD); capaz de usar los genomas humanos de referencia más recientes (GRCh38 y GRCh37), es la herramienta de anotación usada para la última fase del proyecto 1000g, ExAC y gnomAD. La anotación incluye posibles efectos en los cambios del transcrito o a nivel de aminoácidos contemplando las propiedades biofísicas de las proteínas, de tal forma que puede hacerse una interpretación de la variante aun cuando no está asociada a un fenotipo. VEP cuenta con un formato API flexible que nos permitirá incluir anotaciones de diseño propio (custom annotations) para integrar las bases de datos actuales GWAScatalog y ClinVar^[52].

4.3 Bases de datos con asociaciones clínicas de variantes

Las bases de datos de genotipos-fenotipos tienen un objetivo principal: proveer acceso a suficientes datos y conocimiento para permitir establecer y documentar el significado funcional y patogénico de variantes genéticas^[70]. Las

relaciones entre el genotipo de un individuo y la enfermedad (fenotipo) pueden ahora ser exploradas usando microarreglos^[71], secuenciación de exoma completo (whole-exome sequencing, WES)^[72] y secuenciación de genoma completo (whole-genome sequencing, WGS)^[73]. Muchas bases de datos genotipo-fenotipo son esfuerzos de escala pequeña que respaldan proyectos de investigación únicos, áreas específicas de enfermedad, registros específicos de pacientes o conjuntos específicos de genes. Estas bases de datos buscan servir a comunidades internacionales de diagnóstico quienes se especializan en el gen o genes en cuestión al facilitar la interpretación de variantes^[74]. Por otro lado, bases de datos más amplias de genotipos-fenotipos han sido creadas, ofreciendo ligas de las variantes a enfermedades y metadatos (se refiere a información que acompaña a otra, explicando su contexto) y que cubren estudio de asociación de genoma completo (GWASs). También, existen bases de conocimiento general que invierten un esfuerzo manual en coleccionar, armonizar e inter-relacionar grandes cantidades de información primaria para crear compilaciones grandes y poderosas de conocimiento acerca de genes relacionados a enfermedades y a tópicos clínicos o de investigación^[75].

Las siguientes bases de datos son recursos que relacionan variantes con fenotipos clínicos como la predisposición a ciertas enfermedades o fenotipos. Estos recursos forman la base técnica para la herramienta bioinformática que proponemos

4.3.1 GWAS Catalog

El catálogo de Asociaciones de Genoma Completo (GWASC por sus siglas en inglés), fue fundado en 2008 por el “National Human Genome Research Institute” (NHGRI), en respuesta al incremento de estudios GWAS publicados, lo cual permitió el acceso sistemático a las asociaciones genotipo-fenotipo reportadas a nivel mundial. GWASC es albergado por EMBL-EBI (<https://www.ebi.ac.uk/gwas/home>)^[76]. La principal ventaja de las asociaciones reportada en GWASCatalog es su respaldo estadístico. Un ejemplo de asociación reportada es rs77924615 detectada por el proyecto de 1000 genomas, la cual pertenece al gen *PDILT* que codifica para una proteína disulfuro isomerasa, y ha

sido asociada a un incremento en la tasa de filtración glomerular, e incremento en la medición de creatinina sérica en individuos con ancestría europea.

4.3.2 ClinVar

ClinVar es un conjunto de reportes de asociación clínica entre variantes y fenotipos, lanzado por primera vez en 2012. ClinVar facilita el proceso de detección de variantes con interpretación clínica. Los alelos descritos son mapeados a la versión del genoma de referencia y presenta la información de forma interactiva para los usuarios en su portal (<https://www.ncbi.nlm.nih.gov/clinvar/>) así como también puede ser parte de flujos de trabajo al poder descargarse la base de datos^[77]. Es importante señalar que ClinVar no contiene asociaciones GWAS, pero contiene información de los estudios y condiciones clínicas con las que se relaciona la variante, principalmente apoyados por la comunidad científica y médicos genetistas. Además, ClinVar contiene información de la base de datos OMIM (Online Mendelian Inheritance in Man, <https://www.omim.org/>), el cual alberga un catálogo de genes relacionados a desórdenes humanos.

5. PLANTEAMIENTO DEL PROBLEMA

Los proyectos de Genómica Humana han tenido un gran auge en los últimos años, aportando una gran cantidad de datos de diferentes grupos poblacionales y diferentes enfoques para estudiar la variación. En México se ha comenzado con la generación de datos de alta calidad, particularmente enfocados en la población indígena, puesto que a partir de ella se puede explicar un 50% de la variación encontrada en la población mestiza. Sin embargo, ninguno de los proyectos poblacionales reportados ha incluido estrategias estandarizadas para el análisis de la variación genómica masiva, mucho menos herramientas que faciliten la identificación de variantes de interés biomédico. Esto ha creado un área de oportunidad para proveer un análisis más extenso y actualizado de las variantes de interés clínico, utilizando bases de datos de relaciones genotipo-fenotipo como

Clinvar y GWAScatalog en sus versiones más actuales. Paralelamente es necesario proponer una herramienta bioinformática estandarizada para el estudio de datos genómicos poblacionales. Por lo anterior, este trabajo pretende profundizar en el estudio de variantes de interés clínico presentes en población amerindia mexicana utilizando los datos de genoma completo del proyecto en marcha de “100 Genomas mexicanos” del INMEGEN, a través de una herramienta bioinformática propiamente documentada, escalable y reproducible. Las relaciones entre variantes y fenotipos particulares que reportemos, además de la herramienta que proveeremos para su estudio, contribuirá al desarrollo de la Medicina Genómica nacional.

6. HIPÓTESIS

Existen variantes de interés biomédico en la población nativa mexicana implicadas en asociaciones a rasgos fenotípicos y predisposición de enfermedades.

7. OBJETIVOS

7.1 General

Identificar variantes asociadas a fenotipos de interés biomédico-clínico en población nativa mexicana.

7.2 Específicos

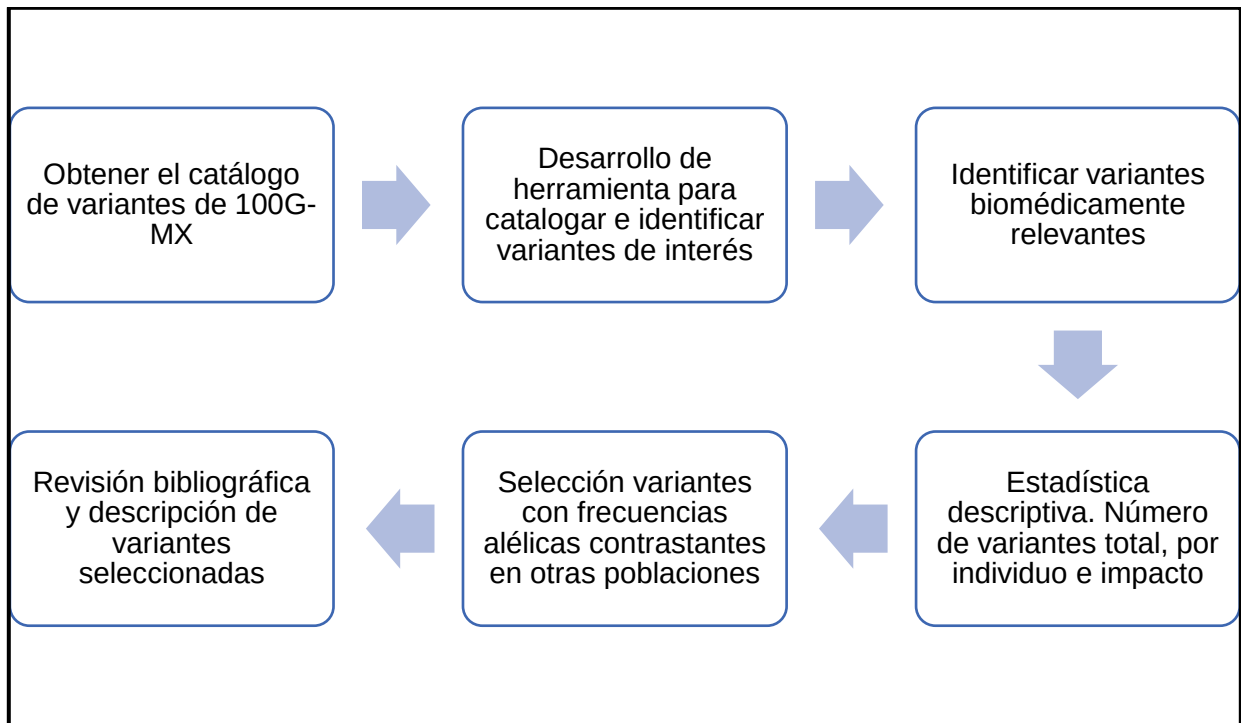
- Describir el número de variantes total, por individuo, y el tipo de variación (sinónimas y no sinónimas) del proyecto 100G-MX de genomas completos de población nativa americana.
- Catalogar variantes de interés biomédico asociadas a GWAS Catalog y ClinVar.

- Identificar variantes particulares de la población amerindia mexicana en comparación con otras superpoblaciones (Europea y Asiática del Este).
- Describir el contexto biomédico de las variantes particulares de en población nativa mexicana.

8. MATERIAL Y MÉTODOS

8.1 Metodología

De manera general, el Esquema 1 ilustra el diagrama de trabajo. El archivo de las variantes filtradas y llamadas fue el conjunto de datos entregados por parte del INMEGEN y con ayuda de la Bioinformática se llevó a cabo la manipulación de los datos como se describe a continuación.



Esquema 1. Diagrama de Trabajo

8.2 Muestras

INMEGEN compartió el conjunto de variantes poblacionales que incluye 76 individuos no emparentados, de 27 grupos indígenas. Los individuos incluidos son parte de la cohorte de “Análisis Metabólico en Muestras Indígenas”, utilizadas en el proyecto 100G-MX del INMEGEN. El estudio se hizo en conformidad con la declaración de Helsinki, y fue aprobada por el comité de ética e investigación humana del INMEGEN. Cada participante firmó una forma de consentimiento^[78]. Los individuos presentan un promedio de 97.22% de genoma nativo americano.

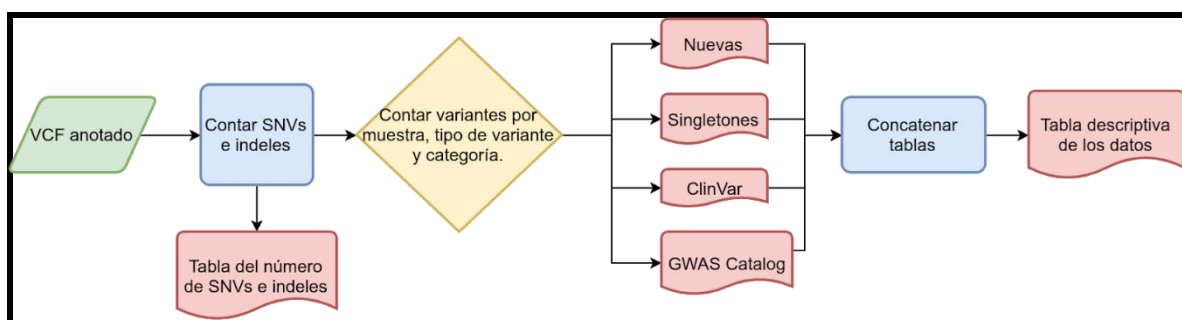
El conjunto de datos 100G-MX está reportado en la versión GRCh38 del genoma humano. Después del llamado de variantes, de mantener sólo las variantes marcadas como “PASS” por GATK, separar los sitios multialélicos y realizar la anotación funcional de las variantes con VEP, estos fueron los datos que el INMEGEN proporcionó para este trabajo de Tesis.

8.3 Descripción básica de los datos

Para continuar con los objetivos propuestos, se llevó a cabo la descripción de los datos genómicos para conocer el número de variantes total, el número de variantes por individuo y el tipo de variación (sinónimas y no sinónimas); así como la estadística descriptiva del número de variantes por categoría.

Para calcular el número de variantes por individuo, se usó el pipeline nf-100GMX-variant-summarizer (<https://github.com/laguilaror/nf-100GMX-variant-summarizer>).

El proceso que lleva a cabo el pipeline es:



Esquema 2. Funcionamiento del pipeline para la descripción básica de los datos

1. Contar el número de variantes de tipo SNV e indeles.
2. Contar el número de variantes de todas las muestras.
3. Crear una tabla por categoría, separada por tabuladores (TSV) con el número de variantes de tipo SNV o indel por muestra. Las categorías son:
 - a) Variantes nuevas.
 - b) Singletones (variantes con un solo alelo contado, AC=1) a nivel mundial.
 - c) Variantes en ClinVar.
 - d) Variantes en GWAS Catalog
4. Concatenar las tablas del conteo de cada categoría.

5. Agrupar muestras por grupo (Maya, Mixe, Nahua, Otomi, Popoluca, Seri y Zapoteco para este conjunto de datos) y generar un archivo por grupo.

El tipo de formato de archivo de entrada que usa este pipeline automatizado es un archivo comprimido VCF descrito previamente en la Figura 3.

La información técnica del funcionamiento del código del pipeline por módulo se encuentra detalladamente descrita en la documentación de este, el cual está albergado en el enlace indicado en la Tabla 3.

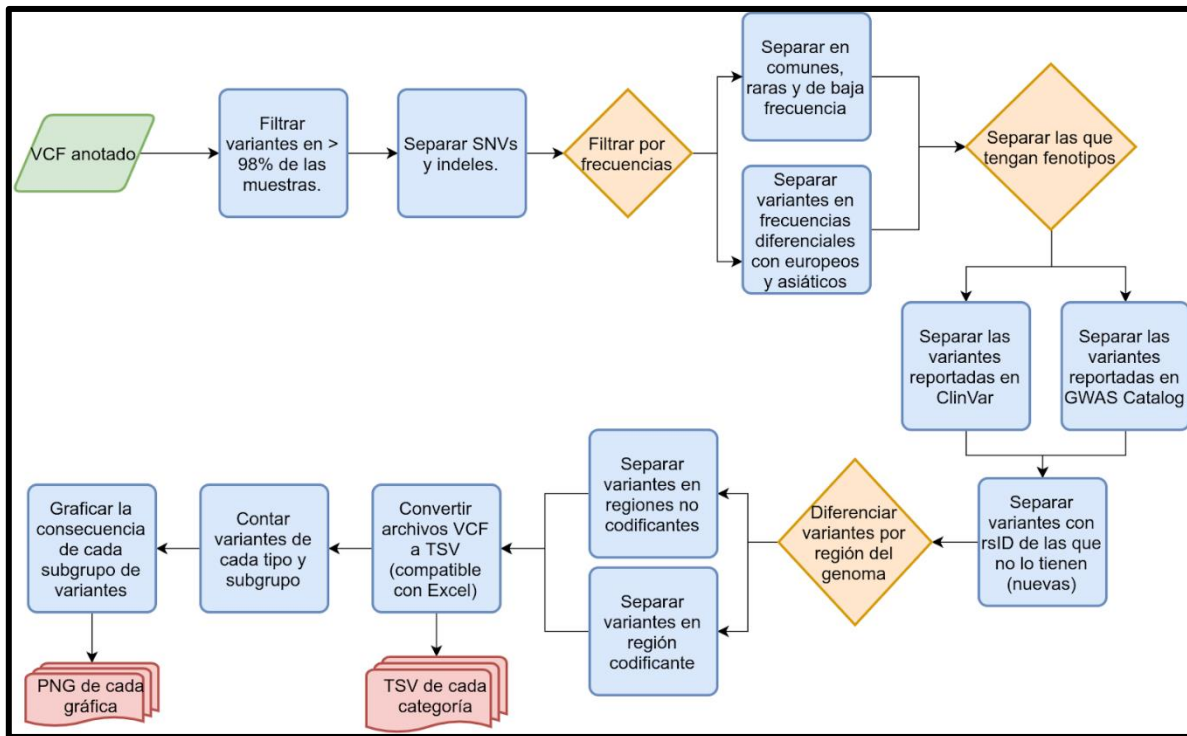
El resultado de la descripción de los datos se ilustra en la Figura 6 de la sección de resultados.

Tomando el archivo de las variantes anotadas y el número de SNPs e Indeles se hizo un gráfico de barras que representa la consecuencia anotada por VEP de cada tipo de variante (Figura 7 en sección de resultados).

8.4 Identificación de variantes de interés biomédico

Con el archivo de las variantes anotadas, se seleccionaron aquellas que tuvieran algún registro en las bases de datos GWAS Catalog y Clinvar. Sin embargo, para que fuera una cantidad de variantes humanamente manejables, se limitó a filtrar también por frecuencias alélicas de gnomAD que fueran diferenciales entre la población de nativos mexicanos y la asiática y europea.

En esta etapa, se utilizó el pipeline `nf-vcf-cataloguer` (<https://github.com/laguilaror/nf-vcf-cataloguer>), el cual cumple con las tareas a continuación enlistadas:



Esquema 3. Funcionamiento de pipeline para la identificación de variantes de interés biomédico

1. Incluir sólo las variantes en donde hubo información por lo menos 75 de los 76 individuos.
2. Conservar sólo ciertos tipos de variantes. Las de tipo SNP e indeles.
3. Separar las variantes en grupos de acuerdo con sus frecuencias alélicas. Los grupos son frecuencias raras (0-0.01), bajas (0.010001-0.05) y comunes (0.050001-1).
4. Separar las variantes con frecuencias diferenciales entre la población de nativos mexicanos y otras poblaciones del proyecto gnomAD.

Más del 5% de frecuencia alélica en la población local y menos del 5% en EAS (asiáticos del este) y NFE (europeos no finlandeses).

Más del 5% de frecuencia alélica en EAS y NFE y menos del 5% en la población local.

5. Separar las variantes anotadas que se encuentren en la base de datos ClinVar. ClinVar contiene datos de OMIM también.
6. Separar las variantes que tengan un estudio de asociación en GWAS Catalog.
7. Separar las variantes conocidas de las no conocidas. Las conocidas son aquellas que ya tienen un rsID asignado en la versión más reciente de dbSNP.
8. Separar variantes que se encuentran en regiones codificantes, es decir, aquellas que se encuentran en exones de las que pertenecen a intrones o intergénicas.
9. Separar las variantes que se encuentran en regiones UTR 5' o 3'.
10. Convertir los archivos separados de formato VCF a uno en que las anotaciones están separadas por tabulaciones, el cual por sus siglas en inglés Tab Separated Values File, se abrevia TSV, el cual es un formato compatible con Excel haciendo que los resultados del análisis sean accesibles al público sin conocimiento bioinformático avanzado o el requerimiento de un sistema operativo específico.
11. Usando un script de R^[79] que es una herramienta para contar las variantes de diferentes tipos y subgrupos (las variantes previamente separadas), generando una tabla con estos valores.
12. Hacer gráficos de la consecuencia anotada por VEP de cada categoría de variantes.

En resumen, para la identificación de variantes biomédicamente relevantes, este trabajo consideraron dos puntos:

- a) Frecuencias diferenciales con otras poblaciones, principalmente Asia del Este y Europa. Estrategia previamente en trabajos con individuos nativos americanos para encontrar variantes relevantes^[25,26].
- b) Variantes con información reportada sobre asociaciones fenotípicas y clínicas.

La información técnica del funcionamiento del código del *pipeline* por módulo se encuentra detalladamente descrita en la documentación de este, el cual está albergado en el enlace indicado en la Tabla 3.

8.5 Curación de variantes de interés biomédico

Para iniciar el proceso de descripción de variantes, se graficaron las frecuencias alélicas (Figura 8 de la sección de resultados) de las variantes alojadas en ClinVar y GWAS Catalog y aquellas con frecuencias diferenciales se representaron en un Diagrama de Venn (Figura 9) para conocer en qué base y región del genoma se encontraban.

La curación de las variantes incluye una descripción minuciosa sobre el rsID de la variante, la posición y el cambio del nucleótido, el cambio de aminoácido y la posición en que ocurre dentro de la proteína, la frecuencia alélica en que se encuentra dentro de los 76 genomas, la población africana, americana, asiática (del sur y del este), europea (no finlandesa) de gnomAD junto con la consecuencia de la variante. Además, si esta afectaba región codificante se reportó si afectaba algún dominio proteico y su función. De manera general se describió la función del gen al que pertenece la variante, y los tejidos en que el gen tiene principal expresión. En cuanto a la información de las bases de datos, si esta se encontraba en ClinVar, se reportó el fenotipo clínico al que se relacionaba y si se consideraba benigna o no; recordando que ClinVar contiene información de OMIM, por lo que, si es el caso, también se reportó el fenotipo de OMIM. En el caso de la base de datos de GWAS Catalog, se reportó información concreta con base a los artículos en que las variantes se asociaron, con el objetivo de que el análisis sirva para que puedan plantearse futuros estudios en población mexicana con fenotipos particulares.

8.6 Pipelines automatizados

GitHub (<https://github.com/>) es una plataforma de desarrollo de código de que permite almacenarlo de forma pública (o privada) y manejar proyectos utilizando el sistema de control de versiones de Git, lo cual posibilita el desarrollo de software colaborativo, la reproducibilidad y replicabilidad del código en las tareas bioinformáticas.

La replicabilidad en herramientas bioinformáticas ha sido limitada en el campo científico, por lo que se han sugerido herramientas y buenas prácticas de

programación, como: resolver tareas por módulos, re-usar código, tener unidades de prueba (test) y el diseño de la documentación^[80].

Es por esto que, en GitHub se encuentran los repositorios del código desarrollado para esta herramienta, haciéndola replicable y accesible. Los pipelines utilizados en este trabajo, que conforman el desarrollo de la herramienta bioinformática para el estudio de la variación genómica, se enlistan en la Tabla 3, indicando el link al repositorio en que se encuentra cada uno. Dentro de cada repositorio se encuentra un archivo llamado README.MD que da la descripción general del pipeline; por módulo existe un archivo readme.md que indica el funcionamiento general del módulo y un archivo (mkfile) que contiene el código para llevar a cabo las tareas, el cual tiene comentarios que explican el porqué de comandos, argumentos, o software utilizado.

Tabla 3. Repositorio de pipelines utilizados

Pipeline	Descripción	Repositorio
nf-variants-per-sample	Este pipeline automatizado, cuenta variantes por individuo (muestra), total y por tipo.	https://github.com/laquilaror/nf-100GMX-variant-summarizer
nf-vcf-cataloguer	Este pipeline automatizado, separa y genera una tabla general que describe cada categoría y subgrupo de las variantes anotadas por VEP. Además, grafica cada subconjunto de las consecuencias de las variantes.	https://github.com/laquilaror/nf-vcf-cataloguer

9. RESULTADOS

El principal resultado de este trabajo es el desarrollo y la obtención de la herramienta bioinformática que permitió llevar a cabo la identificación de variantes para su posterior catalogación y descripción. Debido a la automatización del código bioinformático, puede llevarse este análisis de obtención de variantes de interés biomédico en otras poblaciones o proyectos. A continuación, se enlistan los resultados obtenidos para el conjunto de datos del proyecto 100G-MX.

9.1 Descripción Básica de los Datos

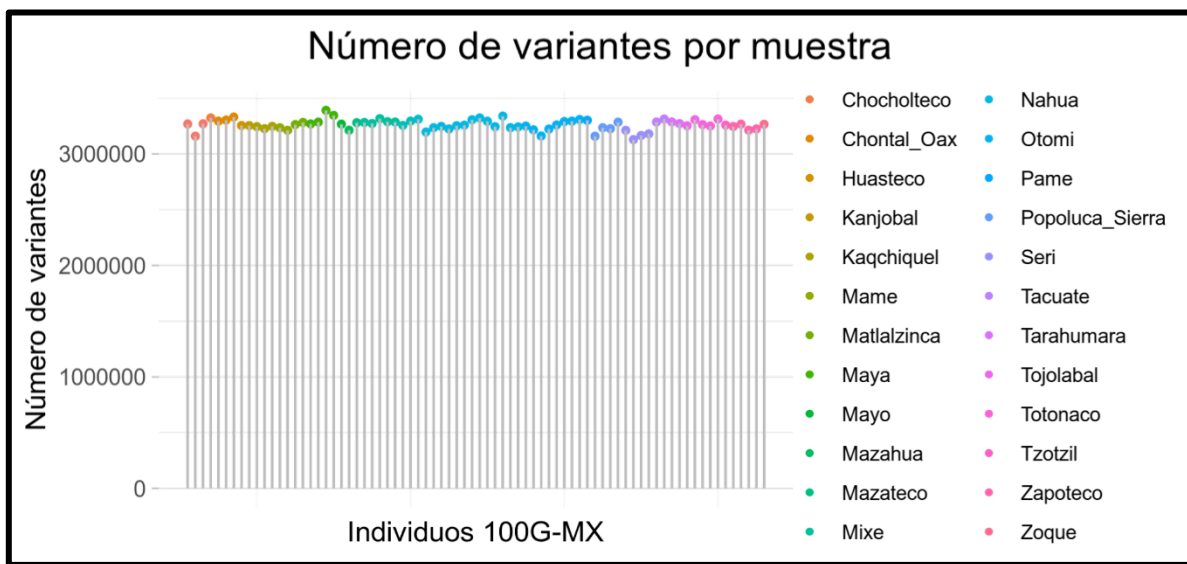


Figura 6. Número de variantes por individuo de cada grupo étnico

La Figura 6 muestra el número de variantes de cada individuo del grupo de datos de 100G-MX. En promedio, de forma individual se obtuvieron 3.26 millones de variantes y de los 76 genomas completos se obtuvieron 8,638,130 SNVs y 1,099,022 indeles.

En la Figura 7 se muestran las consecuencias anotadas por VEP de las variantes. Como puede observarse, la mayor parte de las variantes caen en regiones intergénicas o intrónicas, y también en regiones reguladoras, tanto para las variantes de tipo SNV como los indeles.

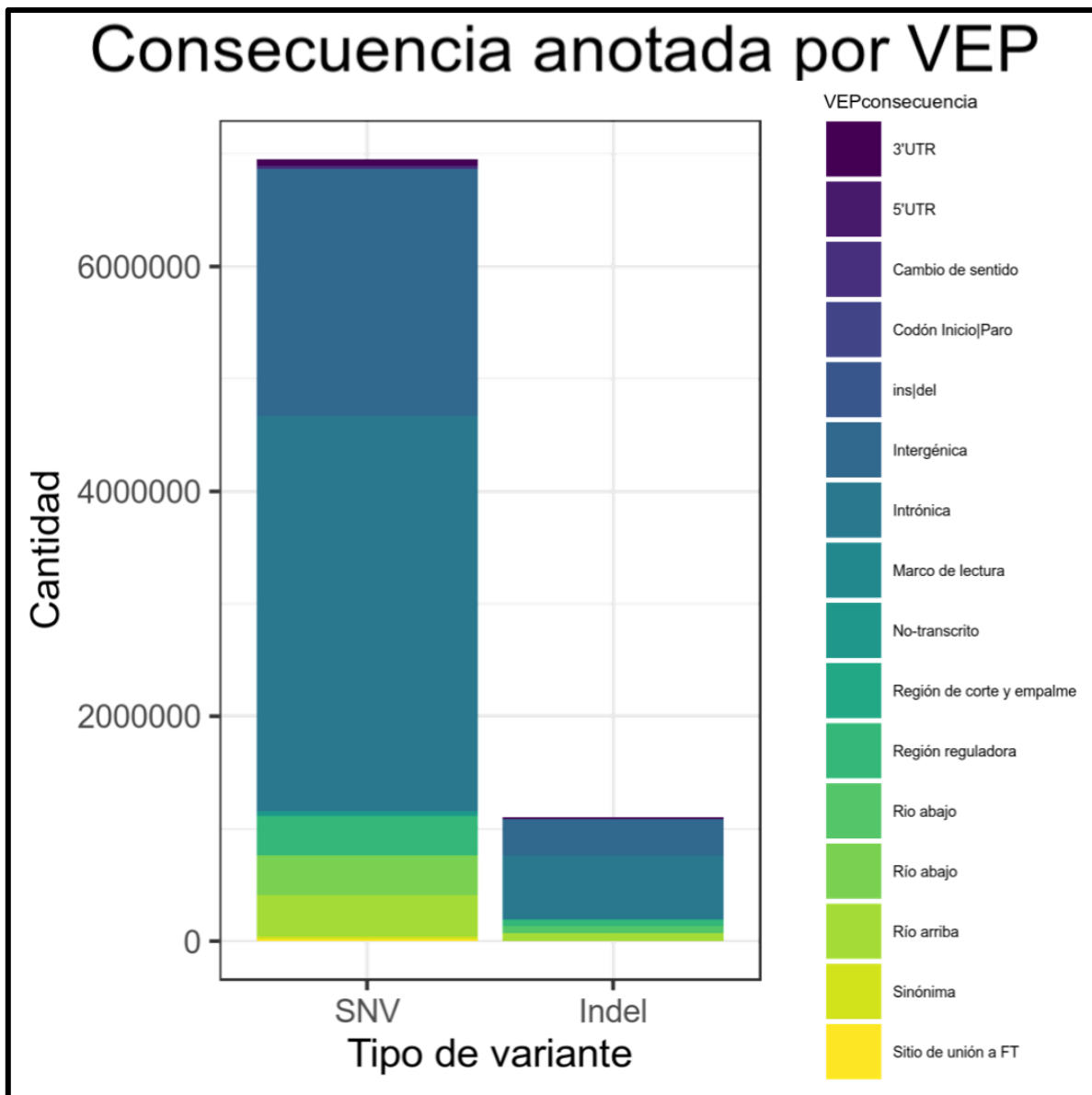


Figura 7. Consecuencia anotada por VEP de acuerdo al tipo de variante. FT refiere a Factor de Transcripción.

9.2 Catalogación de los Datos

En la Tabla 4, se resume el número de variantes por grupo de acuerdo a sus frecuencias y tipo de variante (SNV o indeles), de acuerdo a la descripción del pipeline catalogador del Esquema 3.

Tabla 4. Grupos de variantes separados por frecuencias, tipo y categoría.

FA= Frecuencia Alélica

	SNV			INDEL		
	FA Rara	FA Baja	FA Común	Fa Rara	FA Baja	FA Común
Región Codificante	25,784	9,118	28,945	843	342	592
Reportadas en dbSNP	1,918,765	1,128,329	4,930,622	225,923	211,317	592,412
Nuevas	0	34,791	314	0	7,148	1,865
ClinVar	3,566	2,182	12,932	262	269	962
GWAS Catalog	1,623	2,789	20,037	28	58	371

9.3 Variantes de Interés Biomédico

Como puede observarse en la Figura 8, hacer la descripción del total de las variantes con asociaciones clínicas y fenotípicas para conocer el impacto de la variante al gen que pertenece, las funciones relacionadas o el fenotipo al que se asocia, sería una tarea laboriosa e interminable, por lo que, como se propuso anteriormente, se hizo la descripción de las variantes que podrían ser relevantes para la población nativa mexicana, al comparar las frecuencias alélicas con las otras

poblaciones y encontrarse en cantidades con una gran diferencia.

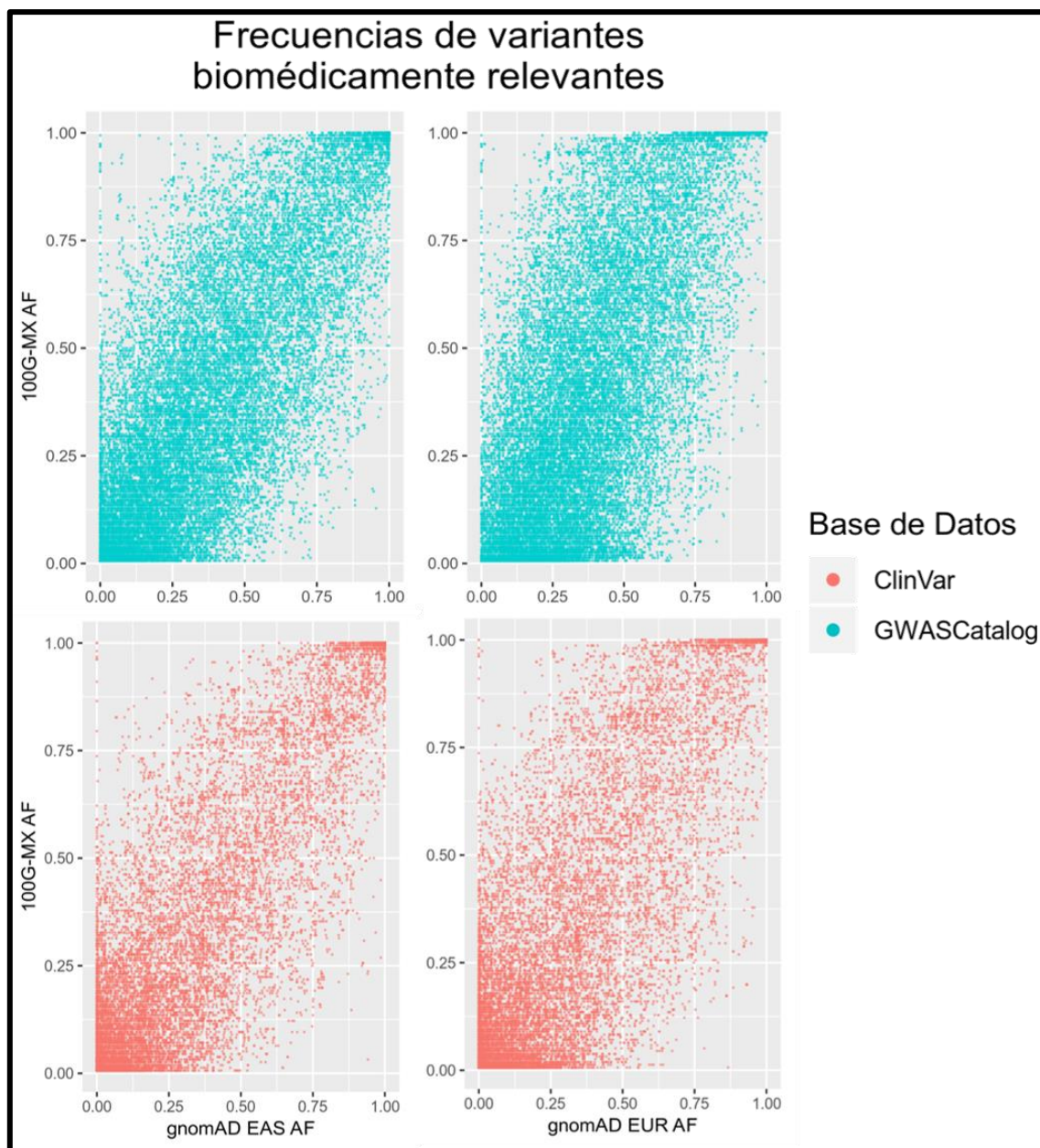


Figura 8. Frecuencia de variantes biomédicamente relevantes de 100GMX, europeos y asiáticos del este, reportadas en ClinVar o GWAS Catalog

Las variantes que se seleccionaron de acuerdo a sus frecuencias alélicas y la región del genoma a la que pertenecen se representa en el diagrama de Venn de la Figura 9.

Variantes de Interés Biomédico

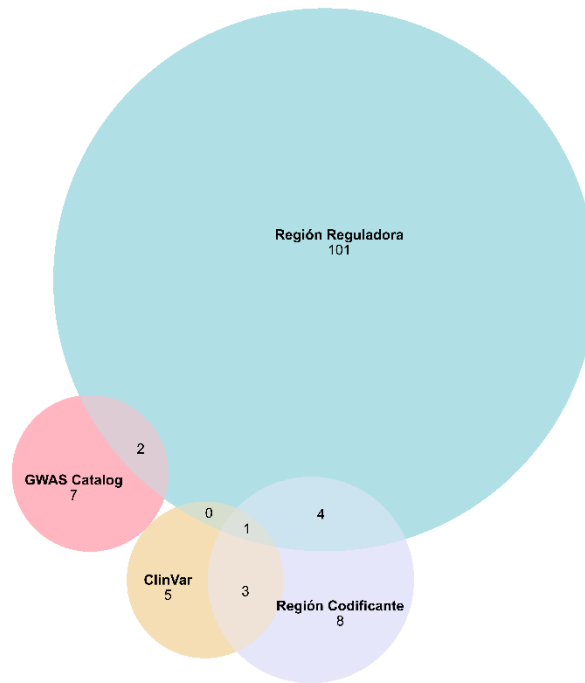


Figura 9. Diagrama de Venn de las variantes de interés biomédico con frecuencias alélicas diferenciales entre la población de nativos mexicanos y otras poblaciones a nivel mundial.

Un resumen de la información relevante de las variantes seleccionadas se describe en la Tabla 5. En color naranja se encuentran las variantes observadas en ClinVar y en rosa las de GWAS Catalog.

Tabla 5. rsID de variantes de interés biomédico con las frecuencias diferenciales en el proyecto de 100G-MX referidos como NatMex, y otras poblaciones

rsID	NatMex FA	gnomAD AMR FA	gnomAD EAS FA	Consecuencia	Fenotipo
rs231402	85.90%	57.09%	43.70%	Sinónima	Displasia fibrosa maxilar
rs2239796	79.50%	43.63%	27.80%	Intrónica	Eritrodermia ictiosiforme congénita
rs12199580	78.20%	51.42%	18.66%	Cambio de sentido	Eritrodermia ictiosiforme congénita
rs62405438	67.90%	36.97%	16.15%	3' UTR	Síndrome de Char
rs1194182	36.50%	43.13%	67.63%	Intrónica	Deficiencia de glucoproteína IV plaquetaria
rs2296872	9.60%	20.05%	67.78%	Intrónica	Sin especificar
rs11037882	87.20%	53.90%	43.26%	Intrónica	Sin especificar
rs75985315	46.80%	17.22%	7.71%	Sinónima	Microcefalia primaria autosómica recesiva
rs4790235	51.30%	19.81%	0.00%	Cambio de sentido	Síndrome miasténico congénito
rs6961069	36.50%	41.61%	66.95%	Intrónica	Conteo plaquetario
rs9895848	51.30%	32.62%	6.83%	Intrónica	Conteo de linfocitos
rs6002333	94.20%	56.26%	7.21%	Intrónica	Porcentaje de eosinófilos
rs2483058	50.60%	NA	NA	Intrónica	Niveles de colesterol
rs12485003	69.20%	2.03%	9.76%	Intrónica	Presión sanguínea
rs11940694	23.70%	40.66%	42.64%	Intrónica	Consumo de alcohol
rs2413631	92.90%	54.95%	4.12%	Intrónica	Síntomas de depresión psicológica
rs72887147	69.90%	35.50%	16.05%	Intergénica	Incremento de masa corporal
rs202629	17.30%	49.28%	94.26%	Intergénica	Dependencia a cannabis

Existe un gran número de variantes (SNVs e indels) en un individuo, lo cual representa un reto complejo para determinar con precisión las variantes causales de enfermedades^[81]. Desde el desarrollo del Proyecto del Genoma Humano, se han mejorado los datos de referencia y el *software* usados en las tecnologías de secuenciación. Pero, los parámetros y guías para priorizar variantes relevantes en la investigación de enfermedades aún siguen en su etapa inicial^[82]. Hoy en día, las tecnologías NGS ofrecen un gran potencial en el diagnóstico clínico. Usando el conocimiento de las variaciones genéticas específicas, podría ofrecerse un enfoque “personalizado” al tratamiento de pacientes con ciertos desórdenes clínicos^[83]. Sin embargo, la interpretación del significado de variantes es un proceso complejo que no termina con la determinación de patogenicidad^[84]. El conocimiento existente sobre mecanismos moleculares involucrados en una enfermedad o fenotipo

determinado puede ser útil para priorizar posibles variantes relevantes biomédicamente^[82].

A continuación, se describe la información recolectada sobre estas variantes de interés biomédico.

ClinVar

Se encontraron loci con frecuencias diferenciales en varios genes que además tenían algún fenotipo relacionado en ClinVar, como en el gen *SH3BP2* (dominio SH3 de unión a proteína 2) que funciona como una proteína adaptadora que regula positivamente la transcripción en células T, NK y basófilos, al unir el dominio SH3 con varias proteínas, principalmente tirosinas cinasas^[85], cuya expresión principal es en vejiga y piel^[86]. Variaciones en este gen, se han asociado con querubismo, la cual es una enfermedad genética fibroósea que se caracteriza por distintos grados de crecimiento anómalo progresivo de la mandíbula, el maxilar o ambos^[87]. Específicamente, la variante rs231402 que afecta este gen, da como consecuencia una variación sinónima, sin embargo, en los 76 genomas de este estudio se encuentra en una frecuencia alélica de 85.9%.

También se encontraron variantes en el gen *PNPLA1*, el cual regula la diferenciación de adipocitos^[88]. rs12199580 y rs2239796 con 78.2% y 79.5% de frecuencia alélica respectivamente en los 76 genomas, ambas denominadas como benignas por ClinVar.

Otro gen resultado del análisis contrastando frecuencias diferenciales fue *TFAP2B*, principalmente expresado en piel, glándulas suprarrenales y riñón^[86], el cual codifica para un miembro de la familia de los factores de transcripción AP-2. Se piensa que estimula la proliferación celular y suprime la diferenciación terminal en el desarrollo embrional, funcionando tanto como activador como represor^[85]. Mutaciones en este gen se han relacionado con el Síndrome de Char, condición que afecta el desarrollo de cara, corazón y extremidades^[89]. rs62405438, una variante que se encuentra en el extremo 3' UTR de este gen con frecuencia en los 76 genomas de 67.9% ha sido denominada como benigna de acuerdo con ClinVar.

SETX es una proteína que en *Saccharomyces cerevisiae* está involucrada en el splicing y terminación del tRNA, sRNA y snRNA, funcionando como RNA helicasa^[90], la cual es expresada principalmente en tejidos como tiroides, testículos, nódulos linfáticos, vejiga y apéndice en humanos^[86]. Se encontró una variante intrónica (rs2296872) en frecuencia alélica de 9.6%, catalogada como benigna por ClinVar.

Igualmente, otro gen con variantes en frecuencias diferenciales es EXT2, el cual codifica para una glicosiltransferasa implicada en la síntesis de heparina sulfato proteoglicanos, involucrados en varias funciones celulares^[91], expresado principalmente en placenta y endometrio^[86]. La variante rs11037882, establecida como benigna por ClinVar, se encuentra en una frecuencia alélica de 87.2% en los 76 genomas.

Se encontró variación con frecuencias contrastantes en *CENPJ*, el cual es un gen que codifica para una proteína centrosomal con un posible rol en la regulación del ensamble de microtúbulos^[92] expresada principalmente en testículos^[86]. La variante rs75985315 con una frecuencia alélica de 46.8% en los 76 genomas, más del doble de la frecuencia reportada para la población americana del proyecto gnomAD y que en el resto de las poblaciones se encuentra en menos del 8%, es una variante sinónima y catalogada como posiblemente benigna por ClinVar.

Interesantemente, *CHRNE* es un gen que codifica para una subunidad de los receptores de acetilcolina^[85]. Variaciones en este gen se han relacionado con un grupo de desórdenes presinápticos, sinápticos y post sinápticos^[93]. rs4790235 es una variante etiquetada como posiblemente benigna por ClinVar que afecta a este gen al provocar un cambio de glicina por valina, y se encuentra en una frecuencia alélica del 51.3% en los 76 genomas, y menos del 5% en otras poblaciones del proyecto gnomAD, exceptuando a la población americana con el 23.05%.

GWAS Catalog

Se encontraron algunas variantes con frecuencias diferenciales en comparación con otras poblaciones relacionadas con fenotipos particulares

reportados en GWAS Catalog como rs6961069, rs9895848, rs6002333, rs2483058, rs11940694, rs72887147, rs2413631, rs202629 y rs12485003.

Tres de estas variantes se reportaron por un estudio que estudiaba la variación en células sanguíneas, relacionadas con desórdenes complejos, principalmente asociaciones a variantes en baja y rara frecuencia; para este estudio se tomaron tres cohortes de ancestría europea, The UK Biobank Study, The UK BiLEVE Study y The INTERVAL Study, fenotipos que fueron seleccionados a través de mediciones físicas y cuestionarios online del estilo de vida y salud^[94].

rs6961069 es una variante intrónica que afecta al gen *CD36*, la proteína que codifica es la 4ª glicoproteína más grande plaquetaria y funciona como un receptor de trombospondina en varias líneas celulares. Además, de que une a colágeno, a fosfolípidos aniónicos y LDL oxidado, y, al unirse a ácidos grasos de cadena larga participa en su regulación y transporte^[85], y es principalmente expresado en tejido adiposo^[86]. Variaciones en este gen están relacionadas con deficiencia de glicoproteínas IV en plaquetas^[95]; sin embargo, esta variante con frecuencia alélica en los 76 genomas de 36.5% se asoció en este estudio a un incremento en el conteo de plaquetas^[94], y ClinVar la denomina como posiblemente benigna.

rs9895848 también es una variante intrónica que mapea al gen *MINK1*, el cual pertenece a la familia “germinal center kinase” (GCK), los cuales regulan varios procesos celulares al acoplar receptores de la superficie celular a vías de cinasas activadas por mitógenos (MAPK), como la vía JNK^[96]; este gen es principalmente expresado en piel y cerebro^[86]. Esta variante, con una frecuencia alélica de 51.3% en los 76 genomas, se asoció a un incremento en el conteo de linfocitos^[94].

Del mismo modo, rs6002333, variante intrónica con una frecuencia alélica de 94.2% en los 76 genomas, pertenece al gen *ZC3H7B* que es una proteína de unión a RNA que puede estar involucrada en la regulación de la traducción, y al interactuar con la proteína no estructural de rotavirus NSP3 puede tener un papel en su replicación^[97]. Esta variante se asoció con el incremento del porcentaje de leucocitos^[94].

Otro estudio de genoma completo reportado en GWAS Catalog en gemelos monocigóticos relacionada con la variabilidad de sueros lipídicos y niveles de apolipoproteínas en 1,720 pares de mujeres gemelas monocigóticas de ascendencia europea del Proyecto GenomEUtwin, midiendo las fracciones lipídicas en ambos gemelos^[98]. rs2483058 fue la variante en selección relacionada con este estudio, que pertenece al gen *SRGAP2* el cual codifica para una proteína altamente conservada que regula la migración y diferenciación neuronal^[99,100] el cual es mayormente expresado en piel, vejiga y cerebro^[86]. La variante con una frecuencia alélica de 50.6% fue asociada a un incremento en la medida del colesterol de alta densidad^[98].

Por otra parte, dos estudios relacionados con el meta-análisis de los desórdenes causados por el uso del alcohol reportaron la asociación con una variante en frecuencia diferencial. Ambos estudios se llevaron a cabo usando cohortes del UK Biobank, 23andMe, y otras más de ascendencia europea, definiendo a los fenotipos a través de los metadatos del test “Alcohol Use Disorders Identification Test” (AUDIT) y cuestionarios para conocer el número de bebidas alcohólicas por semana^[101,102]. rs11940694 en frecuencia alélica de 23.7% es una variante intrónica que pertenece al gen *KLB* el cual tiene como función la unión del receptor del factor de crecimiento de fibroblastos y se expresa principalmente en tejido adiposo^[86]. Esta variante se asoció en ambos estudios al incremento en la medición del consumo del alcohol^[101,102].

Igualmente, un meta-análisis GWAS con individuos de ascendencia africana para índice de masa corporal (BMI) y la relación de estudio-cadera reportó una variante intergénica (rs72887147) con frecuencia alélica de 69.9% en los 76 genomas que se asocia con un incremento del índice de masa corporal^[103]; esta variante afecta al pseudogene *FTH1P5*, sin embargo, no hay información respecto a éste.

Una variante en particular, rs2413631, ha sido relacionada por varios estudios GWAS sobre desórdenes psiquiátricos en individuos de cohortes como UK Biobank, 23andMe, Genetics of Personality Consortium, y otras de ascendencia europea, en donde los fenotipos se establecieron con cuestionarios dicotómicos, midiendo los

años de educación, el mayor nivel de estudios completados, de rendimiento cognitivo y otros^[104–107]. Esta variante intrónica con una frecuencia alélica de 92.9%, se asoció a desórdenes como incremento de ansiedad, decremento de medición de función cognitiva, incremento de medición de síntoma de depresión, incremento de neuroticismo e incremento en la medición de preocupación; afectando al gen *RBX1*, el cual tiene un papel en el proceso de ubiquitinación necesario para la progresión del ciclo celular^[85].

Otra asociada a un fenotipo particular es rs202629, una variante intergénica con una frecuencia alélica de 17.3% en los 76 genomas que mapea a dos genes cercanos, *PHF5A*, que es una subunidad del complejo del *spliceosoma* que cataliza el splicing del pre-RNA^[108] y *TOB2* que pertenece a la familia de proteínas antiproliferativas involucradas en la progresión del ciclo celular^[109], la cual se analizó como parte de un estudio de asociación de dependencia a cannabis en individuos descendientes de europeos de 5 cohortes, resultando en un incremento en la medición de dependencia^[110].

Finalmente, un estudio transétnico asociado a presión sanguínea realizó un metaanálisis para participantes del Million Veteran Program (MVP) y UKBK, teniendo participantes con ancestría europea, africana, latina/hispana, asiáticos y nativos americanos; a quienes se les pidió una muestra de sangre, se aplicaron cuestionarios de estilo de vida y se usó el reporte estadístico de los datos. Asociándose esta variante intrónica al incremento de 0.4076 mmHg en la medición de la presión de pulso^[111], encontrándose en 69.2% de frecuencia alélica en los 76 genomas y perteneciendo al gen *TNRC6B*, que está involucrado en el silenciamiento del gen por RNA, la regulación positiva del acortamiento de colas de poli A del mRNA en el núcleo^[85].

10. DISCUSIÓN

Hacer un escrutinio para identificar variantes de interés biomédico dentro de las 3 billones de bases que constituyen al genoma humano consume mucho tiempo,

por lo que esta herramienta busca hacer eficiente a través de métodos computacionales y paralelizables este proceso. La manera en la que este trabajo destaca y facilita esta tarea es brindando un subconjunto de variantes involucradas en una ruta metabólica o fenotipos clínicos, lo cual es el mayor reto actual en la investigación. Cabe resaltar que en comparación con otras herramientas actuales que buscan abordar esta problemática, este trabajo tiene la ventaja de incluir variantes en regiones no codificantes con asociaciones clínicas o fenotípicas^[112].

La herramienta desarrollada en este trabajo permitió ofrecer, a partir de un catálogo de variación con anotaciones funcionales, identificar las variantes de mayor interés biomédico que posiblemente se encuentran implicadas en rasgos fenotípicos de la población estudiada (nativos americanos). El dataset completo de variación nativa mexicana fue de 8,638,130 SNVs y 1,099,022 indeles cortos. Como puede observarse en la Figura 6, el número de variantes por muestra permanece constante, reflejando el buen llamado de variantes realizado previamente por el INMEGEN; sin embargo, resaltan tres individuos del grupo étnico Seri, los cuales son un grupo pequeño de Sonora y que su aislamiento reflejado en la disminución de variación concuerda con la información antropológica (Fig. 6). Por otro lado, manejar millones de variantes, para enfocarse en algunas de interés, es difícil sin el uso de recursos bioinformáticos, por lo que es clave facilitar la identificación de variantes biomédicamente relevantes. Lo anterior favorece el proceso de generación de hipótesis basadas en observaciones genómicas, al aportar información asociada a fenotipos, genes, o frecuencias poblacionales. Por lo que esta herramienta resalta diferentes conjuntos de datos de acuerdo al interés del usuario (Tabla 4), comparación de frecuencias alélicas, asociaciones fenotípicas o clínicas, tipo de consecuencia o variante o variantes nuevas, siendo las que proveen mayor posibilidad de hipótesis y validaciones experimentales, aquellas con frecuencias alélicas contrastantes y meta-datos asociados. Como ejemplo de este proceso generador de hipótesis, utilizamos variantes con frecuencias alélicas diferenciales entre la población nativa, la población Europea, y la población de Asia del Este. Consideramos variantes con frecuencias diferenciales a aquellas con alta frecuencia dentro de los nativos mexicanos, y baja en otras poblaciones, y

visceversa. Las variantes con frecuencias diferenciales son sugerentes de procesos de selección, y en ocasiones evidencian la importancia funcional de algunas posiciones del genoma^[47–49]. Algunas variantes de baja frecuencia tienden a ser recientes en su origen, y exhiben un alto nivel de diferenciación entre poblaciones, por lo que son sitios de variación de interés para su estudio^[49–51].

Las variantes de interés en el catálogo nativo mexicano descritas en este trabajo están relacionadas con genes que participan en procesos de respuesta inmune, y metabolismo. Esto coincide con trabajos previos de genómica poblacional donde se destacan variantes en frecuencias diferenciales^[24–26]. El 72.5% de la población actual en México tiene sobrepeso u obesidad^[113], por lo que llaman la atención en particular variantes asociadas en estudios de otras poblaciones con fenotipos como obesidad, aumento de niveles de colesterol o índice de masa corporal. El estudio de la estructura genómica de las poblaciones nativas mexicanas permitirá comprender mejor elementos genómicos y particularidades regionales que se comparten con la actual población cosmopolita mexicana.

En este trabajo se analizaron algunas de las variantes del set de datos poblacional de 76 genomas amerindios mexicanos para identificar y caracterizar *in silico* variantes de interés clínico, no obstante, esto sirve como punto de partida ya que se necesita la comprobación de forma experimental de si estas variantes a) se encuentran en población cosmopolita de México en frecuencias similares, b) si verdaderamente participa afectando el rasgo fenotípico con el que se ha relacionado en otras poblaciones. Este tipo de variantes interesantes por la información disponible que se tiene de ellas es la que conforma la base de la Medicina Genómica, y por los resultados obtenidos este trabajo contribuye significativamente al avance de la Medicina Genómica del país al aumentar el número de variantes relevantes a nivel poblacional y médico.

Considerando los estudios que desde 2014 se han llevado a cabo para conocer la variación del país^[24–26], se ha cubierto una parte de la variación existente, sin embargo, México cuenta con 67 grupos indígenas^[114], esto refleja la importancia de incluir más muestras, con mayor diversidad poblacional^[54], ya que todavía quedan

poblaciones sin ser estudiadas a nivel genómico que contribuyen a la composición genómica de la población mestiza mexicana, por lo que en este trabajo se llevó a cabo el diseño, desarrollo y validación con datos reales de una herramienta bioinformática para el análisis de la variación genómica, lo cual busca agilizar la detección de variantes no benignas o relacionadas a fenotipos relevantes para futuros proyectos poblacionales.

Con el avance de las tecnologías de secuenciación, el abaratamiento de costos por reacción y considerando que las regiones codificantes del genoma sólo aportan un porcentaje muy bajo de la variación humana, la secuenciación de genomas completos para llevar a cabo proyectos poblacionales en distintas regiones del mundo tomó importancia^[17-23], lo cual refleja la relevancia de los datos del proyecto 100G-MX empleados en este trabajo, que permitieron el estudio de la variación en regiones codificantes, reguladoras e intergénicas como puede observarse en la Figura 7, sin embargo, uno de los grandes retos que aún existen es determinar el impacto o función de una variante en regiones intergénicas (no reguladoras), y también conocer qué tanto afecta una sola variante a los elementos reguladores (enhancers, promotores y miRNAs); debido a esto es importante el uso de bases de datos con información relacionada a este tipo de elementos genómicos y la continuación del estudio de los múltiples elementos del genoma que enriquezcan estas bases de datos.

11. CONCLUSIONES

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

- 1.** El uso de una herramienta que estandarice el análisis de la variación genómica masiva de forma escalable y reproducible, permite la identificación de variantes de interés biomédico al facilitar el análisis a través del uso de las versiones más actuales de las bases de datos de relaciones genotipo-fenotipo.

2. Existen variantes asociadas a fenotipos de interés biomédico en población amerindia mexicana que explican características del fenotipo de la población mestiza mexicana, por lo que es importante que se continúe el estudio de estas variantes a profundidad.
3. Existen variantes de baja frecuencia en la población nativa amerindia mexicana, importantes debido a que pueden ser sitios de diferenciación reciente, que tienen asociaciones con fenotipos de interés clínico.
4. En conjunto con el Proyecto de los 12 genomas y el Proyecto no publicado de secuenciación de exoma de M. Ávila, este trabajo estudió la variación del proyecto de 100G-MX, proyecto complementa el estudio de la variación regional abarcando la diversidad genómica de 31 de los 68 grupos de NatMex, estudiados por tecnologías de secuenciación masiva.

12. BIBLIOGRAFÍA

1. Mendel G. Versuche über Pflanzen-Hybriden [Internet]. 1. Auflage. Brunn: Naturforschender Verein; 1866. Recuperado: http://www.deutschestextarchiv.de/book/show/mendel_pflanzenhybriden_1866
2. Herbert Fuller Roberts. Plant Hybridization Before Mendel. En: Princeton University Press [Internet]. 1929 [citado el 18 de septiembre de 2019] p. 374. Recuperado: https://books.google.com.mx/books/about/Plant_Hybridization_Before_Mendel.html?id=R505AAAAMAAJ&redir_esc=y
3. WATSON JD, CRICK FHC. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. Nature. 1953;171: 737–738. doi:10.1038/171737a0
4. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A. 1977;74: 5463–7.

doi:10.1073/pnas.74.12.5463

5. Maxam AM, Gilbert W. A new method for sequencing DNA. *Proc Natl Acad Sci U S A*. 1977;74: 560–4. doi:10.1073/pnas.74.2.560
6. Griffiths A, Wessler S, Carrol S, Doebley J. *A Map of Genetics*. 2015. doi:10.1002/chem.201405884
7. Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, Fiddes JC, et al. Nucleotide sequence of bacteriophage ϕ X174 DNA. *Nature*. 1977;265: 687–695. doi:10.1038/265687a0
8. Blattner FR, Plunkett G, Bloch CA, Perna NT, Burland V, Riley M, et al. The complete genome sequence of *Escherichia coli* K-12. *Science*. 1997;277: 1453–62. doi:10.1126/science.277.5331.1453
9. McGeoch DJ, Dalrymple MA, Davison AJ, Dolan A, Frame MC, McNab D, et al. The Complete DNA Sequence of the Long Unique Region in the Genome of Herpes Simplex Virus Type 1. *J Gen Virol*. 1988;69: 1531–1574. doi:10.1099/0022-1317-69-7-1531
10. Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, et al. Complementary DNA sequencing: expressed sequence tags and human genome project. *Science*. 1991;252: 1651–6. doi:10.1126/science.2047873
11. Initial sequencing and analysis of the human genome. *Nature*. 2001;409: 860–921. doi:10.1038/35057062
12. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The sequence of the human genome. *Science*. 2001;291: 1304–51. doi:10.1126/science.1058040
13. The International HapMap Consortium. A haplotype map of the human genome. *Nature*. 2005;437: 1299–320. doi:10.1038/nature04226
14. Consortium T 1000 GP. A global reference for human genetic variation. *Nature*. 2015;526: 68–74. doi:10.1038/nature15393

15. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. *bioRxiv*. 2019; 531210. doi:10.1101/531210
16. Consortium TU. The UK10K project identifies rare variants in health and disease. *Nature*. 2015;526: 82–90. doi:10.1038/nature14962
17. The Genome of the Netherlands Consortium, Francioli LC, Menelaou A, Pulit SL, Dijk F van, Palamara PF, et al. Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nat Genet*. 2014;46: 818–825. doi:10.1038/ng.3021
18. Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, et al. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature*. 2016;538: 201–206. doi:10.1038/nature18964
19. Malaspinas A-S, Westaway MC, Muller C, Sousa VC, Lao O, Alves I, et al. A genomic history of Aboriginal Australia. *Nature*. 2016;538: 207–214. doi:10.1038/nature18299
20. Maretty L, Jensen JM, Petersen B, Sibbesen JA, Liu S, Villesen P, et al. Sequencing and de novo assembly of 150 genomes from Denmark as a population reference. *Nature*. 2017;548: 87–91. doi:10.1038/nature23264
21. Nagasaki M, Yasuda J, Katsuoka F, Nariiai N, Kojima K, Kawai Y, et al. Rare variant discovery by deep whole-genome sequencing of 1,070 Japanese individuals. *Nat Commun*. 2015;6: 8018. doi:10.1038/ncomms9018
22. Zhernakova D V., Brukhin V, Malov S, Oleksyk TK, Koepfli KP, Zhuk A, et al. Genome-wide sequence analyses of ethnic populations across Russia. *Genomics*. 2019; doi:10.1016/J.YGENO.2019.03.007
23. Reuter MS, Walker S, Thiruvahindrapuram B, Whitney J, Cohn I, Sondheim N, et al. The Personal Genome Project Canada: findings from whole genome sequences of the inaugural 56 participants. *CMAJ*. 2018;190: E126–E136.

doi:10.1503/cmaj.171151

24. Moreno-Estrada A, Gignoux CR, Fernández-López JC, Zakharia F, Sikora M, Contreras A V, et al. Human genetics. The genetics of Mexico recapitulates Native American substructure and affects biomedical traits. *Science*. 2014;344: 1280–5. doi:10.1126/science.1251688
25. Romero-Hidalgo S, Ochoa-Leyva A, Garcíarrubio A, Acuña-Alonzo V, Antúnez-Argüelles E, Balcazar-Quintero M, et al. Demographic history and biologically relevant genetic variation of Native Mexicans inferred from whole-genome sequencing. *Nat Commun*. 2017;8: 1005. doi:10.1038/s41467-017-01194-z
26. Ávila-Arcos MC, McManus KF, Sandoval K, Rodríguez-Rodríguez JE, Martin AR, Luisi P, et al. Population history and gene divergence in Native Mexicans inferred from 76 human exomes. *bioRxiv*. 2019; 534818. doi:10.1101/534818
27. Price AL, Patterson N, Yu F, Cox DR, Waliszewska A, McDonald GJ, et al. A genomewide admixture map for Latino populations. *Am J Hum Genet*. 2007;80: 1024–36. doi:10.1086/518313
28. Silva-Zolezzi I, Hidalgo-Miranda A, Estrada-Gil J, Fernandez-Lopez JC, Uribe-Figueroa L, Contreras A, et al. Analysis of genomic diversity in Mexican Mestizo populations to develop genomic medicine in Mexico. *Proc Natl Acad Sci U S A*. 2009;106: 8611–6. doi:10.1073/pnas.0903045106
29. Clamp M, Fry B, Kamal M, Xie X, Cuff J, Lin MF, et al. Distinguishing protein-coding and noncoding genes in the human genome. *Proc Natl Acad Sci U S A*. 2007;104: 19428–33. doi:10.1073/pnas.0709013104
30. Jobling M., Hurles M., & Tyler-Smith C. *Human evolutionary genetics*. 2a ed. Garland Science, editor. 2014.
31. Jorde LB. *Genetic Variation and Human Evolution* [Internet]. Recuperado: <http://www.ashg.org/education/pdf/geneticvariation.pdf>
32. Shastri BS. SNP alleles in human disease and evolution. *J Hum Genet*.

2002;47: 0561–0566. doi:10.1007/s100380200086

33. Pang AW, MacDonald JR, Pinto D, Wei J, Rafiq MA, Conrad DF, et al. Towards a comprehensive structural variation map of an individual human genome. *Genome Biol.* 2010;11: R52. doi:10.1186/gb-2010-11-5-r52
34. Chanock S. Candidate Genes and Single Nucleotide Polymorphisms (SNPs) in the Study of Human Disease. *Dis Markers.* 2001;17: 89–98. doi:10.1155/2001/858760
35. Pinkel D, Se Graves R, Sudar D, Clark S, Poole I, Kowbel D, et al. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat Genet.* 1998;20: 207–211. doi:10.1038/2524
36. Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, et al. Detection of large-scale variation in the human genome. *Nat Genet.* 2004;36: 949–951. doi:10.1038/ng1416
37. Sebat J, Lakshmi B, Malhotra D, Troke J, Lese-Martin C, Walsh T, et al. Strong association of de novo copy number mutations with autism. *Science.* 2007;316: 445–9. doi:10.1126/science.1138659
38. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 2010;20: 1297–303. doi:10.1101/gr.107524.110
39. Poplin R, Ruano-Rubio V, DePristo MA, Fennell TJ, Carneiro MO, Auwera GA Van der, et al. Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv.* 2018; 201178. doi:10.1101/201178
40. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. *Bioinformatics.* 2011;27: 2156–2158. doi:10.1093/bioinformatics/btr330
41. The International HapMap Project. *Nature.* 2003;426: 789–796. doi:10.1038/nature02168

42. Integrating ethics and science in the International HapMap Project. *Nat Rev Genet.* 2004;5: 467–475. doi:10.1038/nrg1351
43. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature.* 2007;447: 661–678. doi:10.1038/nature05911
44. Rabbee N, Speed TP. A genotype calling algorithm for affymetrix SNP arrays. *Bioinformatics.* 2006;22: 7–12. doi:10.1093/bioinformatics/bti741
45. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am J Hum Genet.* 2007;81: 559–575. doi:10.1086/519795
46. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet.* 2006;38: 904–909. doi:10.1038/ng1847
47. Tennessen JA, Bigham AW, O'Connor TD, Fu W, Kenny EE, Gravel S, et al. Evolution and Functional Impact of Rare Coding Variation from Deep Sequencing of Human Exomes. *Science* (80-). 2012;337: 64–69. doi:10.1126/science.1219240
48. Marth GT, Yu F, Indap AR, Garimella K, Gravel S, Leong W, et al. The functional spectrum of low-frequency coding variation. *Genome Biol.* 2011;12: R84. doi:10.1186/gb-2011-12-9-r84
49. Nelson MR, Wegmann D, Ehm MG, Kessner D, St. Jean P, Verzilli C, et al. An Abundance of Rare Functional Variants in 202 Drug Target Genes Sequenced in 14,002 People. *Science* (80-). 2012;337: 100–104. doi:10.1126/science.1217876
50. Mathieson I, McVean G. Differential confounding of rare and common variants in spatially structured populations. *Nat Genet.* 2012;44: 243–246. doi:10.1038/ng.1074
51. Gravel S, Henn BM, Gutenkunst RN, Indap AR, Marth GT, Clark AG, et al.

Demographic history and rare allele sharing among human populations. *Proc Natl Acad Sci U S A*. 2011;108: 11983–8. doi:10.1073/pnas.1019276108

52. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant Effect Predictor. *Genome Biol*. 2016;17: 122. doi:10.1186/s13059-016-0974-4
53. Fu W, O'Connor TD, Jun G, Kang HM, Abecasis G, Leal SM, et al. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature*. 2013;493: 216–220. doi:10.1038/nature11690
54. Lek M, Karczewski KJ, Minikel E V., Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*. 2016;536: 285–291. doi:10.1038/nature19057
55. Consortium T2D. Sequence variants in SLC16A11 are a common risk factor for type 2 diabetes in Mexico. *Nature*. 2014;506: 97–101. doi:10.1038/nature12828
56. Sanchez-Albornoz N. The population of Latin America. A history [Internet]. Berkeley, CA (USA) Univ. of California Press; 1974. Recuperado: <http://agris.fao.org/agris-search/search.do?recordID=XF2015016437>
57. Raghavan M, Steinrucken M, Harris K, Schiffels S, Rasmussen S, DeGiorgio M, et al. Genomic evidence for the Pleistocene and recent population history of Native Americans. *Science* (80-). 2015;349: aab3884–aab3884. doi:10.1126/science.aab3884
58. Ribeiro-dos-Santos AM, de Souza JES, Almeida R, Alencar DO, Barbosa MS, Gusmão L, et al. High-Throughput Sequencing of a South American Amerindian. Calafell F, editor. *PLoS One*. 2013;8: e83340. doi:10.1371/journal.pone.0083340
59. Acuña-Alonzo V, Flores-Dorantes T, Kruit JK, Villarreal-Molina T, Arellano-Campos O, Hünemeier T, et al. A functional ABCA1 gene variant is associated with low HDL-cholesterol levels and shows evidence of positive selection in

Native Americans. *Hum Mol Genet.* 2010;19: 2877–2885. doi:10.1093/hmg/ddq173

60. Rusu V, Hoch E, Mercader JM, Tenen DE, Gymrek M, Hartigan CR, et al. Type 2 Diabetes Variants Disrupt Function of SLC16A11 through Two Distinct Mechanisms. *Cell.* 2017;170: 199-212.e20. doi:10.1016/j.cell.2017.06.011
61. Drmanac R, Sparks AB, Callow MJ, Halpern AL, Burns NL, Kermani BG, et al. Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science.* 2010;327: 78–81. doi:10.1126/science.1181498
62. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. *Nat Methods.* 2010;7: 248–249. doi:10.1038/nmeth0410-248
63. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly (Austin).* 2012;6: 80–92. doi:10.4161/fly.19695
64. Splawski I, Shen J, Timothy KW, Lehmann MH, Priori S, Robinson JL, et al. Spectrum of mutations in long-QT syndrome genes. KVLQT1, HERG, SCN5A, KCNE1, and KCNE2. *Circulation.* 2000;102: 1178–85. doi:10.1161/01.cir.102.10.1178
65. Makiyama T, Akao M, Tsuji K, Doi T, Ohno S, Takenaka K, et al. High Risk for Bradyarrhythmic Complications in Patients With Brugada Syndrome Caused by SCN5A Gene Mutations. *J Am Coll Cardiol.* 2005;46: 2100–2106. doi:10.1016/j.jacc.2005.08.043
66. Dubinsky MC, Mei L, Friedman M, Dhere T, Haritunians T, Hakonarson H, et al. Genome wide association (GWA) predictors of anti-TNF α therapeutic responsiveness in pediatric inflammatory bowel disease. *Inflamm Bowel Dis.* 2010;16: 1357–1366. doi:10.1002/ibd.21174
67. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* 2010;38:

e164–e164. doi:10.1093/nar/gkq603

68. Racimo F, Marnetto D, Huerta-Sánchez E. Signatures of Archaic Adaptive Introgression in Present-Day Human Populations. *Mol Biol Evol.* 2016;34: msw216. doi:10.1093/molbev/msw216
69. Sherry ST, Ward M-H, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* 2001;29: 308–311. doi:10.1093/nar/29.1.308
70. Johnston JJ, Biesecker LG. Databases of genomic variation and phenotypes: existing resources and future needs. *Hum Mol Genet.* 2013;22: R27–R31. doi:10.1093/hmg/ddt384
71. Rehm HL. Disease-targeted sequencing: a cornerstone in the clinic. *Nat Rev Genet.* 2013;14: 295–300. doi:10.1038/nrg3463
72. Yang Y, Muzny DM, Reid JG, Bainbridge MN, Willis A, Ward PA, et al. Clinical Whole-Exome Sequencing for the Diagnosis of Mendelian Disorders. *N Engl J Med.* 2013;369: 1502–1511. doi:10.1056/NEJMoa1306555
73. Saunders CJ, Miller NA, Soden SE, Dinwiddie DL, Noll A, Alnadi NA, et al. Rapid Whole-Genome Sequencing for Genetic Disease Diagnosis in Neonatal Intensive Care Units. *Sci Transl Med.* 2012;4: 154ra135-154ra135. doi:10.1126/scitranslmed.3004041
74. Horaitis O, Talbot CC, Phommavanh M, Phillips KM, Cotton RGH. A database of locus-specific databases. *Nat Genet.* 2007;39: 425–425. doi:10.1038/ng0407-425
75. Brookes AJ, Robinson PN. Human genotype–phenotype databases: aims, challenges and opportunities. *Nat Rev Genet.* 2015;16: 702–715. doi:10.1038/nrg3932
76. Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* 2014;42: D1001–D1006. doi:10.1093/nar/gkt1229

77. Landrum MJ, Lee JM, Benson M, Brown G, Chao C, Chitipiralla S, et al. ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res.* 2016;44: D862–D868. doi:10.1093/nar/gkv1222
78. Contreras-Cubas C, Sánchez-Hernández BE, García-Ortiz H, Martínez-Hernández A, Barajas-Olmos F, Cid M, et al. Heterogenous Distribution of MTHFR Gene Variants among Mestizos and Diverse Amerindian Groups from Mexico. Kibar Z, editor. *PLoS One.* 2016;11: e0163248. doi:10.1371/journal.pone.0163248
79. R Development Core Team RFFSC. *R: A language and environment for statistical computing.* 2011. p. 409.
80. Kim YM, Poline JB, Dumas G. Experimenting with reproducibility: A case study of robustness in bioinformatics. *GigaScience.* Oxford University Press; 2018. doi:10.1093/gigascience/giy077
81. Zhang L, Zhang J, Yang J, Ying D, Lau Y lung, Yang W. PriVar: a toolkit for prioritizing SNVs and indels from next-generation sequencing data. *Bioinformatics.* 2013;29: 124–125. doi:10.1093/bioinformatics/bts627
82. Jalali Sefid Dashti M, Gamiieldien J. A practical guide to filtering and prioritizing genetic variants. *Biotechniques.* 2017;62. doi:10.2144/000114492
83. Becich M, Santana-Santos L, Gullapalli R, Desai K, Kant J. Next generation sequencing in clinical medicine: Challenges and lessons for pathology and biomedical informatics. *J Pathol Inform.* 2012;3: 40. doi:10.4103/2153-3539.103013
84. Quintáns B, Ordóñez-Ugalde A, Cacheiro P, Carracedo A, Sobrido MJ. Medical genomics: The intricate path from genetic variant identification to clinical interpretation. *Applied and Translational Genomics.* Elsevier; 2014. pp. 60–67. doi:10.1016/j.atg.2014.06.001
85. O’Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic

expansion, and functional annotation. *Nucleic Acids Res.* 2016;44: D733–D745. doi:10.1093/nar/gkv1189

86. Fagerberg L, Hallström BM, Oksvold P, Kampf C, Djureinovic D, Odeberg J, et al. Analysis of the Human Tissue-specific Expression by Genome-wide Integration of Transcriptomics and Antibody-based Proteomics. *Mol Cell Proteomics.* 2014;13: 397–406. doi:10.1074/MCP.M113.035600
87. Tiziani V, Reichenberger E, Buzzo CL, Niazi S, Fukai N, Stiller M, et al. The gene for cherubism maps to chromosome 4p16. *Am J Hum Genet.* 1999;65: 158–66. doi:10.1086/302456
88. Wilson PA, Gardner SD, Lambie NM, Commans SA, Crowther DJ. Characterization of the human patatin-like phospholipase family. *J Lipid Res.* 2006;47: 1940–9. doi:10.1194/jlr.M600185-JLR200
89. Mani A, Radhakrishnan J, Farhi A, Carew KS, Warnes CA, Nelson-Williams C, et al. Syndromic patent ductus arteriosus: Evidence for haploinsufficient TFAP2B mutations and identification of a linked sleep disorder. *Proc Natl Acad Sci.* 2005;102: 2975–2979. doi:10.1073/pnas.0409852102
90. Moreira M-C, Klur S, Watanabe M, Németh AH, Ber I Le, Moniz J-C, et al. Senataxin, the ortholog of a yeast RNA helicase, is mutant in ataxia-ocular apraxia 2. *Nat Genet.* 2004;36: 225–227. doi:10.1038/ng1303
91. Farhan SMK, Wang J, Robinson JF, Prasad AN, Rupar CA, Siu VM, et al. Old gene, new phenotype: mutations in heparan sulfate synthesis enzyme, EXT2 leads to seizure and developmental disorder, no exostoses. *J Med Genet.* 2015;52: 666–75. doi:10.1136/jmedgenet-2015-103279
92. Hung LY, Tang CJ, Tang TK. Protein 4.1 R-135 interacts with a novel centrosomal protein (CPAP) which is associated with the gamma-tubulin complex. *Mol Cell Biol.* 2000;20: 7813–25. doi:10.1128/mcb.20.20.7813-7825.2000
93. Engel AG, Shen X-M, Selcen D, Sine SM. Congenital myasthenic syndromes:

- pathogenesis, diagnosis, and treatment. *Lancet Neurol.* 2015;14: 420–34. doi:10.1016/S1474-4422(14)70201-7
94. Astle WJ, Elding H, Jiang T, Allen D, Ruklisa D, Mann AL, et al. The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. *Cell.* 2016;167: 1415-1429.e19. doi:10.1016/j.cell.2016.10.042
 95. Yamamoto N, Akamatsu N, Sakuraba H, Yamazaki H, Tanoue K. Platelet glycoprotein IV (CD36) deficiency is associated with the absence (type I) or the presence (type II) of glycoprotein IV on monocytes. *Blood.* 1994;83. Recuperado: <http://www.bloodjournal.org/content/83/2/392?sso-checked=true>
 96. Hu Y, Leo C, Yu S, Huang BCB, Wang H, Shen M, et al. Identification and functional characterization of a novel human misshapen/Nck interacting kinase-related kinase, hMINK beta. *J Biol Chem.* 2004;279: 54387–97. doi:10.1074/jbc.M404497200
 97. Vitour D, Lindenbaum P, Vende P, Becker MM, Poncet D. RoXaN, a novel cellular protein containing TPR, LD, and zinc finger motifs, forms a ternary complex with eukaryotic initiation factor 4G and rotavirus NSP3. *J Virol.* 2004;78: 3851–62. doi:10.1128/jvi.78.8.3851-3862.2004
 98. Surakka I, Whitfield JB, Perola M, Visscher PM, Montgomery GW, Falchi M, et al. A Genome-Wide Association Study of Monozygotic Twin-Pairs Suggests a Locus Related to Variability of Serum High-Density Lipoprotein Cholesterol. *Twin Res Hum Genet.* 2012;15: 691–699. doi:10.1017/thg.2012.63
 99. Charrier C, Joshi K, Coutinho-Budd J, Kim J-E, Lambert N, de Marchena J, et al. Inhibition of SRGAP2 function by its human-specific paralogs induces neoteny during spine maturation. *Cell.* 2012;149: 923–35. doi:10.1016/j.cell.2012.03.034
 100. Dennis MY, Nuttle X, Sudmant PH, Antonacci F, Graves TA, Nefedov M, et al. Evolution of human-specific neural SRGAP2 genes by incomplete segmental

duplication. *Cell*. 2012;149: 912–22. doi:10.1016/j.cell.2012.03.033

101. Sanchez-Roige S, Palmer AA, Fontanillas P, Elson SL, Adams MJ, Howard DM, et al. Genome-Wide Association Study Meta-Analysis of the Alcohol Use Disorders Identification Test (AUDIT) in Two Population-Based Cohorts. *Am J Psychiatry*. 2019;176: 107–118. doi:10.1176/appi.ajp.2018.18040369
102. Liu M, Jiang Y, Wedow R, Li Y, Brazel DM, Chen F, et al. Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use. *Nat Genet*. 2019;51: 237–244. doi:10.1038/s41588-018-0307-5
103. Ng MCY, Graff M, Lu Y, Justice AE, Mudgal P, Liu CT, et al. Discovery and fine-mapping of adiposity loci using high density imputation of genome-wide association studies in individuals of African ancestry: African ancestry anthropometry genetics consortium. *PLoS Genet*. 2017;13. doi:10.1371/journal.pgen.1006719
104. Nagel M, Jansen PR, Stringer S, Watanabe K, de Leeuw CA, Bryois J, et al. Meta-analysis of genome-wide association studies for neuroticism in 449,484 individuals identifies novel genetic loci and pathways. *Nat Genet*. 2018;50: 920–927. doi:10.1038/s41588-018-0151-7
105. Lee JJ, Wedow R, Okbay A, Kong E, Maghzian O, Zacher M, et al. Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nat Genet*. 2018;50: 1112–1121. doi:10.1038/s41588-018-0147-3
106. Mbarek H, van de Weijer MP, van der Zee MD, Ip HF, Beck JJ, Abdellaoui A, et al. Biological insights into multiple birth: genetic findings from UK Biobank. *Eur J Hum Genet*. 2019;27: 970–979. doi:10.1038/s41431-019-0355-z
107. Nagel M, Watanabe K, Stringer S, Posthuma D, van der Sluis S. Item-level analyses reveal genetic heterogeneity in neuroticism. *Nat Commun*. 2018;9: 905. doi:10.1038/s41467-018-03242-8

108. Will CL, Urlaub H, Achsel T, Gentzel M, Wilm M, Lührmann R. Characterization of novel SF3b and 17S U2 snRNP proteins, including a human Prp5p homologue and an SF3b DEAD-box protein. *EMBO J.* 2002;21: 4978–88. doi:10.1093/emboj/cdf480
109. Ikematsu N, Yoshida Y, Kawamura-Tsuzuku J, Ohsugi M, Onda M, Hirai M, et al. Tob2, a novel anti-proliferative Tob/BTG1 family member, associates with a component of the CCR4 transcriptional regulatory complex capable of binding cyclin-dependent kinases. *Oncogene.* 1999;18: 7432–7441. doi:10.1038/sj.onc.1203193
110. Agrawal A, Chou Y-L, Carey CE, Baranger DAA, Zhang B, Sherva R, et al. Genome-wide association study identifies a novel locus for cannabis dependence. *Mol Psychiatry.* 2018;23: 1293–1302. doi:10.1038/mp.2017.200
111. Giri A, Hellwege JN, Keaton JM, Park J, Qiu C, Warren HR, et al. Trans-ethnic association study of blood pressure determinants in over 750,000 individuals. *Nat Genet.* 2019;51: 51–62. doi:10.1038/s41588-018-0303-9
112. Dharanipragada P, Seelam SR, Parekh N. SeqVItA: Sequence Variant Identification and Annotation Platform for Next Generation Sequencing Data. *Front Genet.* 2018;9. doi:10.3389/fgene.2018.00537
113. “Instituto Nacional de Estadística y Geografía”. Encuesta Intercensal 2015. En: Encuesta Intercensal [Internet]. 2015 [citado el 12 de noviembre de 2019]. Recuperado:
<https://www.inegi.org.mx/programas/intercensal/2015/default.html#Documentacion>
114. Sistema de Información Cultural. Pueblos indígenas en México : Sistema de Información Cultural-Secretaría de Cultura [Internet]. [citado el 12 de septiembre de 2019]. Recuperado:
https://sic.cultura.gob.mx/?table=grupo_etnico&disciplina=&estado_id=0

13. ANEXOS

-abreviaturas