BENEMÉRITA UNIVERSIDAD AUTÓNOMA DE PUEBLA

FACULTAD DE CIENCIAS BIOLÓGICAS

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

TESIS

PRESENTADA PARA OBTENER EL GRADO DE

LICENCIADO EN BIOTECNOLOGÍA

PRESENTA:

RAM GONZÁLEZ BUENFIL

DIRECTORA DE TESIS: DRA. LILIANA LÓPEZ PLIEGO

CODIRECTOR DE TESIS: M.C. ISRAEL AGUILAR ORDÓÑEZ

ASESORES DE TESIS:

DR. CARLOS ALBERTO CONTRERAS PAREDES

M.C. ELDA CARREÓN MORENO

PUEBLA, PUE. AGOSTO, 2020

Agradecimientos

Gracias a mi familia, por acompañarme los años de la carrera aún estando tan lejos. Por llamarme y escribirme. Por el apoyo de mis padres y mi hermana para que pudiera cumplir mis sueños y metas. Por su cariño y amor incondicional.

A mis tutores de tesis. Lili, por enseñarme desde el primer día en el laboratorio, por sus pláticas y consejos, su guía en mis primeros experimentos, su constancia y compromiso para revisar mis avances en este proyecto. Israel, por creer en mí y ofrecerme grandes oportunidades, por la constancia a tus alumnos y el tiempo, esfuerzo y dedicación, por darme consejos de vida y académicos. A ambos, por ofrecerme su amistad.

A Judith, mi amiga y colega con quien he crecido todos estos años como científico. Por las tardes en el laboratorio compartiendo una silla para hacer bioinformática, por las pláticas en los viajes a cursos, congresos e institutos. Por las reflexiones, los paseos y las risas compartidas.

A mis queridos amigos, Jael, Jesús, Eric, Rigo, Oswaldo, Alejandra, Sam, Mariana y José Luis. Por acompañarme durante toda la carrera, por el tiempo en las clases y fuera de ellas. Las sesiones de estudio en la biblioteca, las pláticas reflexivas, las tardes de juegos, las fiestas y las desveladas estudiando para algún examen.

Sin ustedes no sería lo que soy ahora.

Este trabajo fue posible gracias al apoyo económico otorgado por el Consejo Nacional de Ciencia y Tecnología como Ayudante de Investigador Nacional SNI III o Emérito del Dr. Juan Enrique Morett Sánchez, durante el periodo septiembre 2018 – agosto 2020, con el expediente de ayudante número 16508. Mismo trabajo fue realizado en su laboratorio en el Instituto de Biotecnología de la UNAM. También fue posible gracias a la estancia académica (No. de matrícula 1805508) durante el periodo de octubre 2018 – enero 2020 en el Instituto Nacional de Medicina Genómica, bajo la tutoría del Dr. Cristóbal Fresno Rodríguez. De igual forma, la Subdirección de Bioinformática, a través del departamento de Supercómputo, puso a disposición del proyecto del que fue parte este trabajo, la infraestructura y el ambiente necesario para llevar a cabo el análisis masivo de datos.

Contenido

RESUMEN 5
1. Introducción 6
1.2 Variación Genética a Nivel Mundial 8
1.3 Variación Genética en México 10
1.3.1 El Proyecto 100 G-MX, Variación Nativa Mexicana en Genoma Completo 11
1.4 Variación Genética en Elementos Reguladores Enhancer 12
2. Marco Teórico 14
2.1 Almacenamiento y Repositorios de Variantes Genéticas 14
2.1.1 La Base de Datos dbSNP y el Número de Referencia de SNP (rsID) 14
2.1.2 La Base de Datos de Agregación de Genoma (gnomAD) 15
2.2 Proyectos y Consorcios Enfocados en Estudiar Regiones Reguladoras del Genoma Humano 16
2.3 La Base de datos GeneHancer 18
3. Justificación 20
4. Hipótesis 21
5. Objetivos 21
6. Materiales y Métodos 22
6.1 Descripción de los Datos Usados para el Análisis 23
6.1.1 El Catálogo de Variantes del Proyecto 100 G-MX. 23
6.1.2 Variantes con Señal de Selección 25
6.2 Metodología 26
7. Resultados 28
7.2 Variación Genética en Regiones Enhancer de 76 Genomas de población Nativa Mexicana 28
7.3 Variantes por Elemento Enhancer 30
7.4 Genes con Variantes Localizadas en Elementos Reguladores Enhancer 32
7.5 Densidad de Variantes en Elementos Reguladores por Gen 34
7.6 Variantes en Enhancers de Interés Clínico 35
8. Discusión 36
9. Conclusiones 41
10. Perspectivas 41
11. Referencias 43

12. Material Suplementario 63
12.1 Formato de Llamado de Variante (VCF) 63
12.2 Genome Analysis Toolkit (GATK) Best Practices 64
12.3 Base de datos de GeneHancer 65
12.3.1 Extracción de Enhancers 65
12.3.2 Integración de los elementos en GeneHancer 66
12.3.3 Determinación del Valor de Confianza de los Enhancers 67
12.3.4 Asociaciones Enhancer - Gen 68
12.4 Resultados 70
12.4.3 Variación Genética en Regiones Enhancer de 76 Genomas de población Nativa Mexicana 70
12.4.4 Variantes por Elemento Enhancer 78
12.4.5 Genes con Variantes Genéticas Localizadas en Elementos Reguladores Enhancer 78
12.4.6 Densidad de Variantes en Elementos Reguladores por Gen 80
12.4.7 Variantes en Enhancers de PPARG, SYN2, TIMP4 y TSEN2 81

RESUMEN

Los estudios basados en secuenciación genómica a nivel regional permiten entender la historia de una población desde el punto de vista biológico, ayudando a revelar procesos evolutivos y a detectar posibles cambios en el genoma (Jorde & Wooding, 2004; Mallick et al., 2016; The 1000 Genomes Project Consortium, 2015), que estén implicados en procesos de salud o enfermedad. Estos cambios en el genoma se conocen como variantes, y en conjunto conforman la variación genética entre individuos. En un contexto genético, México es un país con una diversidad y subestructura poblacional muy prevalente. Es decir, dentro de la población mexicana, existen grupos cuya variación genética es compartida por individuos dentro del mismo grupo, y diferente en cierto grado, entre grupos (Ávila-Arcos et al., 2019; Moreno-Estrada et al., 2014; Romero-Hidalgo et al., 2017). Una combinación originada por los eventos de poblamiento del territorio, migraciones, el periodo de conquista y mestizaje, han dado como consecuencia que los principales aportes genéticos a nuestra población actual provengan de población nativa mexicana y europea. A pesar de que la variación en población nativo mexicana permanece poco explorada, recientemente han surgido proyectos de gran interés (Moreno-Estrada et al., 2014; Romero-Hidalgo et al., 2017), los cuales han contribuido a comprender mejor la diversidad de variación genómica en la población nativo mexicana y sus posibles implicaciones biomédicas. Sin embargo, el estudio de la variación en regiones no codificantes, específicamente aquellas en regiones reguladoras, permanece inexplorada.

En el presente trabajo se analizaron variantes genéticas detectadas en regiones reguladoras de la expresión génica tipo enhancer. Utilizando datos del proyecto 100 G-MX (Aguilar-Ordoñez et al., en proceso de publicación), donde se secuenciaron 76 genomas completos de población nativa mexicana pertenecientes a distintos grupos indígenas en México. Con base en el catálogo de variantes obtenidas por el grupo del Dr. Morett en IBT-UNAM, se detectaron 108 polimorfismos de un solo nucleótido (SNPs) que ocurren en elementos identificados como enhancers de acuerdo con la base de datos GeneHancer (una base de datos de enhancers humanos y sus genes blanco inferidos) (Fishilevich et al., 2017). Dos de estas variantes están presentes en la base de datos GWAS Catalog y han sido asociadas a fenotipos de interés clínico. PPARgamma, un gen implicado en el metabolismo de lípidos con participación en patologías como la diabetes tipo 2, obesidad y retención de insulina en la población mexicana (Evans et al., 2004; Stryjecki et al., 2016), destaca por ser el gen con el mayor número de variantes en región reguladora, distribuidas en 3 elementos enhancer. De manera interesante, las variantes detectadas en enhancers reguladores de PPARgamma, SYN2, TIMP4 y TSEN2 parecen

pertenecer a un haplotipo presente en población africana y nativo mexicana, pero no en población europea ni asiática. Hasta el momento, este es el primer trabajo enfocado en explorar el panorama genético en elementos enhancer de genomas de la población nativa mexicana.

1. Introducción

Explorar la variación genética en las poblaciones humanas ha sido uno de los principales esfuerzos científicos en el siglo XXI. Esta variación se refleja a través cambios en las secuencias genéticas de un individuo respecto a otro, originados por eventos de mutación, a los que se les refiere comúnmente como polimorfismos genéticos. Un polimorfismo genético se define como la ocurrencia de múltiples alelos en un locus, donde al menos dos alelos ocurren en una frecuencia mayor al 1% (Forrester et al., 2016).

Uno de los tipos más comunes de variación genética es la sustitución de un solo nucleótido por otro en un sitio específico. Llamado polimorfismo de un solo nucleótido (SNP; del inglés Single Nucleotide Polymorphism (Shastry, 2002). La figura 1 muestra un ejemplo de SNP, junto con otros dos tipos comunes de polimorfismos genéticos: Inserciones/deleciones cortas (INDEL), y variaciones estructurales complejas como las duplicaciones.

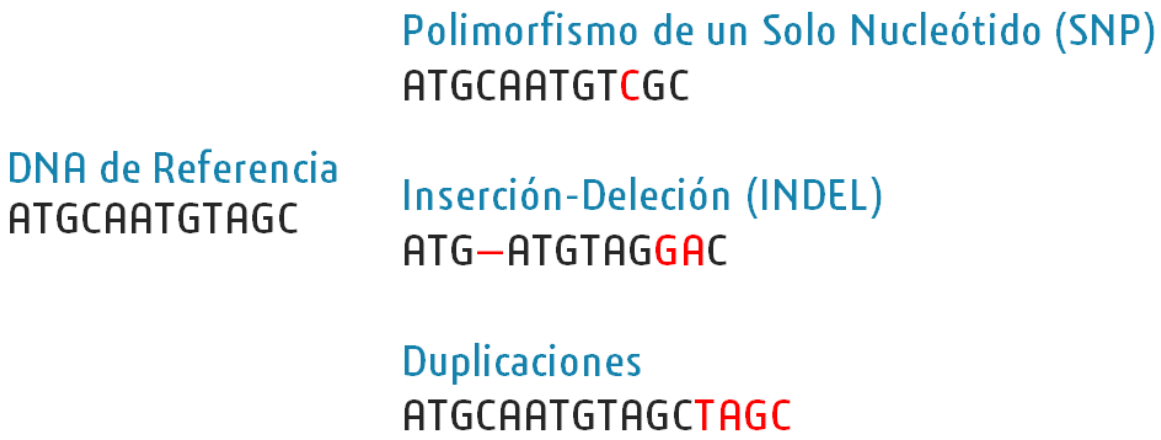


Figura 1. Algunos tipos comunes de polimorfismos genéticos. Los polimorfismos son detectados al comparar con una secuencia de DNA de referencia. Un SNP representa el cambio de una base nucleotídica por otra, un INDEL es la adición o eliminación de algunas bases nucleotídicas y una duplicación es la repetición exacta de una sección de la secuencia.

Al analizar la variación genética en diversos grupos poblacionales alrededor del mundo, se ha observado que existen polimorfismos que prevalecen sólo en ciertas poblaciones, es decir, que algunos alelos se encuentran con mayor frecuencia en una población que otra. Es claro que esto

es un resultado del agrupamiento de distintos pueblos por su historia ancestral y distribución geográfica (The 1000 Genomes Project Consortium, 2015).

El término Frecuencia Alélica será recurrente en este trabajo, y se utiliza para describir la proporción que representa un alelo particular del número total de alelos en una población. Para el caso de los humanos, la ecuación que describe el cálculo de la frecuencia alélica es la siguiente:

$$AF_Z = AC_Z / 2N$$

Donde AC_Z es el número total de alelos Z observados en una población y N el número de individuos que componen a la población, que se multiplica por 2 dado que el ser humano posee un genoma diploide. En términos simples, la frecuencia alélica mide qué fracción del total de alelos de una población corresponde a un alelo en particular. La suma de todas las frecuencias alélicas para un locus específico siempre debe ser 1 (Stephenson, 2016). Un ejemplo del cálculo de la frecuencia alélica en una población puede consultarse en la figura 2.

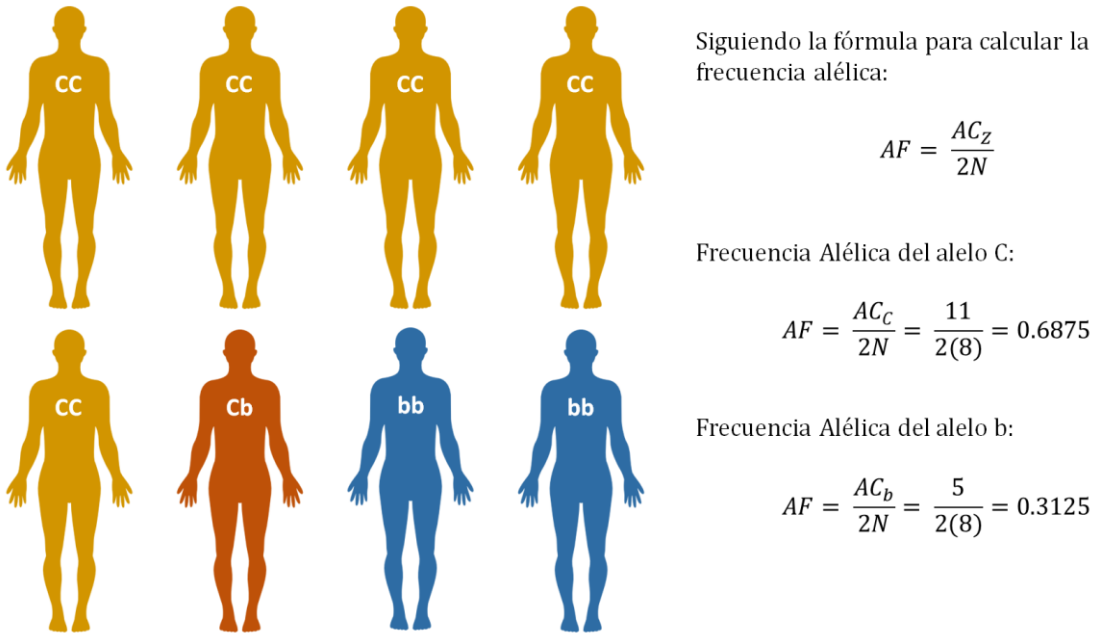


Figura 2. **Ejemplo del cálculo de frecuencias alélicas**. En una población de 8 individuos con 3 genotipos diferentes, indicados por los colores amarillo, naranja y azul, se calcula la frecuencia alélica para los alelos *C* y *b*.

Además de reflejar eventos demográficos y de adaptación, los polimorfismos genéticos pueden asociarse con distintos rasgos de interés clínico que podrían presentar un riesgo a la salud de la población estudiada. Por lo que estudiar la variación genética de una región particular

permitiría comprender mejor aspectos genotípicos que pudieran explicar rasgos fenotípicos de interés.

1.2 Variación Genética a Nivel Mundial

La publicación del primer borrador de la secuencia del genoma humano en 2001 (International Human Genome Sequencing Consortium, 2001), marcó un paso importante para la genómica humana. La idea de estudiar el genoma para comprender enfermedades complejas y explorar la diversidad de nuestra especie parecía cada vez más latente. Es por ello que en los últimos años han surgido grandes proyectos, producto de la colaboración internacional, que buscan estudiar la variación genética de distintas poblaciones con la intención de elucidar el papel que juega en la biología humana. Existen diversos proyectos cuya finalidad ha sido comprender mejor la diversidad genética a través de la genómica de poblaciones. Tres de ellos, el proyecto HapMap, el proyecto de los Mil Genomas y el "Simons Genome Diversity Project", se consideran relevantes para entender el contexto de la genómica de poblaciones, por haber incluido un amplio número de poblaciones diferentes.

El Proyecto HapMap (The International HapMap Consortium, 2005), tuvo como objetivo determinar los patrones en los que la variación genética era heredada a través del análisis de haplotipos, definidos como bloques específicos de alelos físicamente cercanos, observados en regiones particulares del genoma, que son heredados a un individuo por sus padres como una sola unidad (Crawford & Nickerson, 2005), conforme pasan las generaciones se espera que un haplotipo heredado empiece a perder longitud a causa de eventos de recombinación (Figura 3). Utilizando principalmente tecnología de microarreglos para la genotipificación, este proyecto incluyó poblaciones de 6 países: Japón, Canadá, Estados Unidos, Reino Unido, China y Nigeria, con la meta de que los resultados pudieran ser útiles para encontrar alelos que contribuyen a enfermedades o en la respuesta a fármacos en esas poblaciones. El proyecto HapMap terminó en 2005, y reportó alrededor de un millón de SNPs, construyendo el camino para futuros estudios de asociación genotipo-fenotipo.

En 2015, la versión final del proyecto de los mil genomas (The 1000 Genomes Project Consortium, 2015) reportó la secuenciación por genoma completo de 2,504 individuos pertenecientes a 26 diferentes poblaciones, entre ellas individuos catalogados como mexicanos de Los Ángeles (entre los cuales no se incluyeron individuos con alta ancestría nativa mexicana). El resultado fue una descripción de la variación genética a nivel mundial al

catalogar 88 millones de variantes, de las cuales 84.7 millones corresponden a SNPs. Además de ser uno de los primeros proyectos en analizar genomas completos de múltiples poblaciones, los resultados de este trabajo sirvieron para presentar una visión general de la variación genética humana, útil para investigación médica en estudios de asociación y medicina genómica, que hasta el momento había sido realizada principalmente en poblaciones europeas y que tenían sesgos al capturar variaciones funcionalmente importantes en otras poblaciones.

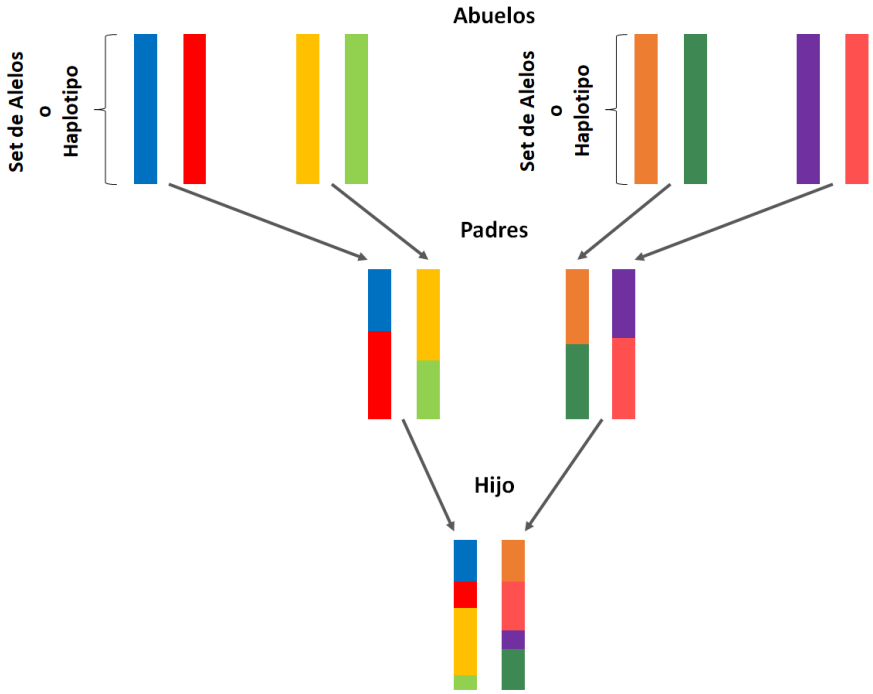


Figura 3. Bloques de haplotipo. Un set de alelos físicamente cercanos, llamado haplotipo, es heredado a un individuo por sus padres. Conforme avanzan las generaciones, la longitud de un haplotipo empieza a reducirse como resultado de procesos de recombinación.

Con una meta similar, en 2016, el Simons Genome Diversity Project (Mallick et al., 2016) reportó 300 genomas pertenecientes a 142 poblaciones distribuidas a lo largo del mundo. Aportando una visión más amplia del panorama de la diversidad genética, con 5.8 millones de pares de bases no incluidas en el genoma humano de referencia (una secuencia genómica cuyo objetivo es representar fielmente a toda la humanidad. (Schneider et al., 2016) y nuevos conocimientos de índole demográfica acerca del tiempo de separación de algunas poblaciones, acumulación de mutaciones, y el aporte de homínidos arcaicos (neandertales y denisova) hacia las poblaciones humanas modernas. Otro aspecto muy relevante de este proyecto es que incluyó los genomas de 13 individuos nativo mexicanos pertenecientes a 6 diferentes grupos indígenas.

La Medicina Genómica se nutre del descubrimiento de variantes y a través de proyectos como los antes mencionados, sin embargo, su alcance se ve limitado cuando los estudios existentes carecen de la representación de algunos grupos poblacionales. Un análisis reciente del GWAS Catalog (MacArthur et al., 2017), un repositorio de variantes genómicas cuyo efecto, presencia en diversas fuentes de literatura, y otra información relevante (es decir, variantes curadas), que han sido asociadas a enfermedades humanas y creada en colaboración del National Human Genome Research Institute (NHGRI) y el European Bioinformatics Institute (EMBL-EBI), reveló una gran falta de diversidad y baja representación de individuos con ancestría africana, latina, del pacífico, de medio oriente, mestiza y nativa frente a individuos con ancestría europea y asiática en sus datos. Dicho sesgo en la información genera disparidad respecto a las poblaciones menos representadas. Esto podría causar un impacto en la "generalización" de asociación de variantes a rasgos fenotípicos, ya que las variantes podrían no tener el mismo efecto en poblaciones con mayor representación de datos que en poblaciones menos estudiadas, afectando la manera en la que las últimas reciben servicios de salud por ser estas variantes menos relevantes o de distinto impacto para ellas (Popejoy & Fullerton, 2016).

1.3 Variación Genética en México

El contexto genético de la población mexicana ha sido descrito por una subestructura y diversidad poblacional particular. Es decir, dentro de la población mexicana, existen grupos cuya variación genética es compartida por individuos dentro del mismo grupo, y diferente entre grupos (Moreno-Estrada et al., 2014; Romero-Hidalgo et al., 2017). Una mezcla originada por los eventos de poblamiento del territorio, migraciones, el periodo de conquista y mestizaje han dado como consecuencia que los principales aportes genéticos a nuestra población provengan de población nativo mexicana y europea (Moreno-Estrada et al., 2014; Romero-Hidalgo et al., 2017). El resultado ha sido una gran riqueza genética donde los individuos varían en grados de ancestría mestiza, europea y nativo mexicana; este último grupo compuesto por 68 agrupaciones lingüísticas en el territorio mexicano (Instituto Nacional de Lenguas Indígenas (Mexico), 2009).

En México, a pesar de que la variación en población nativa mexicana permanece poco explorada, recientemente han surgido proyectos poblacionales de gran interés, como los que se describen a continuación.

En 2014, Moreno-Estrada y colaboradores exploraron los patrones de variación genómica regional en individuos de 20 grupos nativo-mexicanos, evaluaron alrededor de 1 millón de SNPs utilizando genotipificación por microarreglos. Este estudio introdujo la perspectiva de una subestructura poblacional al establecer evidencia de una diferenciación genética entre diferentes grupos nativos en el país, principalmente organizada por la región geográfica y la historia de cada pueblo. El estudio logró asociar genotipos de poblaciones en ciertas regiones del país con una diferencia en el valor del volumen espiratorio en un segundo (FEV1), una medida estándar para evaluar la función pulmonar, estimando un cambio en esta medida dependiendo de la ancestría de cada individuo, remarcando el impacto que los estudios poblacionales pueden tener en el ámbito clínico de la medicina en México.

Así mismo, en 2017 Romero-Hidalgo y colaboradores exploraron la diversidad nativa mexicana con una mayor profundidad. Utilizando datos de genoma completo provenientes de 12 individuos indígenas, reportando alrededor de 3.2 millones de SNPs (por individuo). Los resultados de este estudio ampliaron el panorama de la estructura poblacional mexicana, sugiriendo, a través del análisis de enriquecimiento de vías (una metodología que identifica las vías biológicas alteradas por variantes. Reimand et al., 2019), que la resistencia física de la etnia tarahumara podría ser un reflejo de procesos adaptativos. Asimismo, fueron detectadas variantes con altas frecuencias alélicas previamente asociadas con rasgos fenotípicos en regiones codificantes y promotoras. Es importante señalar que, en este proyecto, los promotores se definieron como regiones de 7.5 kilobases en la región río arriba del sitio de inicio de transcripción 5’ de los genes. Actualmente se cuenta con evidencia más directa sobre la localización de elementos reguladores; particularmente en la base de datos GeneHancer, que se utilizará en el presente proyecto.

1.3.1 El Proyecto 100 G-MX, Variación Nativa Mexicana en Genoma Completo

El proyecto 100 G-MX es un esfuerzo del Instituto Nacional de Medicina Genómica (INMEGEN) en conjunto con el Instituto de Biotecnología de la Universidad Nacional Autónoma de México (IBt UNAM) por secuenciar el genoma completo de 76 individuos no emparentados, pertenecientes a 27 grupos indígenas de México, con más del 97% de ancestría nativo mexicana en promedio; cuyo objetivo principal es describir niveles de variación genética

que hasta el momento no han sido detectados en población nativa. De este proyecto se ha generado un catálogo de aproximadamente 9 millones de variantes, de las cuales una fracción considerable son variantes *novel*, es decir, variantes que no han sido detectadas con anterioridad en ninguna otra población.

El presente proyecto de tesis utilizó datos generados por 100 G-MX para explorar y analizar la variación genética nativo-mexicana detectada en elementos reguladores tipo enhancer.

Adicionalmente, y como parte del reciente trabajo de tesis de Judith Ballesteros Villascán, se ha explorado la variación en regiones codificantes de genomas secuenciados por el proyecto, detectando variantes de interés clínico en estas regiones.

1.4 Variación Genética en Elementos Reguladores Enhancer

A pesar de que muchos estudios se concentran en analizar variantes en regiones codificantes, dado que estas impactarían directamente a la proteína codificada y en consecuencia al fenotipo, recientemente se ha encontrado que el 88% de las variantes asociadas a alguna enfermedad se encuentran en las regiones no codificantes (MacArthur et al., 2017; Edwards et al., 2013; Buniello et al., 2019; *GWAS Catalog*, n.d.), tal como se ha observado con variantes que contribuyen al desarrollo de enfermedades y que ocurren en regiones reguladoras (Rojano et al., 2018), impactando en la expresión génica.

Un potenciador transcripcional o enhancer es una secuencia de nucleótidos que interactúa con factores de transcripción y regiones promotoras para favorecer la expresión génica. Comúnmente actúa sobre sobre regiones en el cromosoma en el que se encuentra en *cis*, pero en algunos casos también puede actuar sobre otros cromosomas en *trans* (Hans Peter Müller & Walter Schaffner, 1990; Müller, H.-P. et al., 1989; Sasaki-Iwaoka et al., 1999), por lo que la distancia de esta interacción es variable. La activación transcripcional mediada por enhancers opera cuando estas secuencias interactúan con regiones promotoras a través de factores transcripcionales formando un bucle en la cromatina permitiendo el reclutamiento de la maquinaria transcripcional (Marsman & Horsfield, 2012. Figura 4). La asociación enhancer-promotor no está restringida a un solo elemento enhancer, tal como se ha demostrado a partir de un análisis de interacciones enhancer-promotor a lo largo del genoma humano, en donde

alrededor del 25% de los promotores identificados estaban asociados con dos o más elementos enhancer (Chepelev et al., 2012).

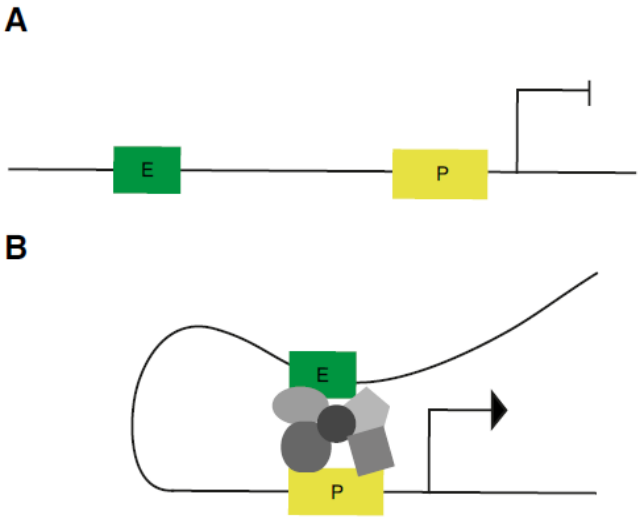


Figura 4. Interacción Enhancer-Promotor. **(A)** Esquema de un gen sin transcribir que río arriba contiene un elemento enhancer inactivo (E, verde) y un elemento promotor (P, amarillo). **(B)** El enhancer es traído en proximidad con el promotor y activa la transcripción al reclutar factores de transcripción (FT) y coactivadores (figuras grises). El gen activo se muestra con una línea que termina en flecha y el gen inactivo como una línea que termina en T. Tomado de Marsman & Horsfield, 2012.

Existen evidencias de la importancia de detectar y analizar variación en elementos reguladores enhancer, las cuales pueden jugar un papel muy importante en el desarrollo de fenotipos de interés clínico y que, en algunos casos, han demostrado ser más prevalentes en ciertas poblaciones que en otras. Los siguientes son algunos ejemplos que evidencian la importancia de buscar variantes clínicamente relevantes en regiones reguladoras, ya que se tratan de variantes que afectan la regulación de sus genes blanco y por lo tanto es posible que causen algún tipo de enfermedad:

En 2013, French y colaboradores lograron identificar 3 variantes en enhancers que regulan al gen CCND1, estas variantes se encuentran asociadas al desarrollo de cáncer de mama y fueron detectadas con mayor frecuencia en población europea que en poblaciones asiáticas.

Los islotes pancreáticos de las células β son las unidades fisiológicas donde se coordina la síntesis de insulina, fundamental para el metabolismo de la glucosa (Pokrywczynska et al., 2019); y su estudio ha sido clave en la comprensión del desarrollo de la diabetes (Ashcroft & Rorsman, 2012). En 2014, se detectaron conjuntos de SNPs asociados a diabetes tipo 2 localizados en distintos grupos de enhancers que activan la expresión de genes en islotes pancreáticos, prevalentes principalmente en población asiática del este. Entre los genes

afectados se encontraban SLC2A2, CDKN2A, C2CD4A y SLC30A8, entre otros (Pasquali et al., 2014).

La enfermedad coronaria es un padecimiento del corazón que se manifiesta principalmente como dolores en el pecho, ataques al corazón y muerte repentina (Wong, 2014). En 2018, Gong y colaboradores buscaron estudiar la posible asociación de esta enfermedad con la presencia de SNPs en diversos grupos de enhancers. Realizando un estudio de asociación de genoma completo, junto con datos de patrones de modificación de histonas, pudieron encontrar 366 SNPs en grupos de enhancers asociados con esta enfermedad. Un análisis de enriquecimiento de vías reveló que estas variantes estaban enriquecidas en procesos de señalización y regulación bien estudiados, tales como la vía de señalización del cAMP y la vía de señalización de ErbB; dos vías que juegan un papel importante en el metabolismo de la enfermedad coronaria.

2. Marco Teórico

2.1 Almacenamiento y Repositorios de Variantes Genéticas

Dentro del campo de la genómica de poblaciones, existen diversas bases de datos que sirven como un medio para guardar y organizar la información generada por distintos proyectos o laboratorios alrededor del mundo. Estas bases de datos contienen, principalmente, información acerca de las variantes detectadas, como el tipo de variante genética, la población donde ha sido detectada y su frecuencia alélica. A continuación, se describen dos bases de datos importantes en el contexto de este trabajo.

2.1.1 La Base de Datos dbSNP y el Número de Referencia de SNP (rsID)

Ya que los SNPs son una forma muy común de variación genética (Shastry, 2002) ocurriendo aproximadamente una vez por cada mil pares de bases a lo largo del genoma (Brookes, 1999), y que pueden ayudar a explicar en gran parte fenotipos de interés, ha habido un aumento reciente en la atención científica por detectarlos y descubrirlos. Es por eso que en 1998 el National Human Genome Research Institute (NHGRI) y el National Center for Biotechnology Information (NCBI) crearon la base de datos dbSNP (https://www.ncbi.nlm.nih.gov/snp/) para servir como un repositorio de variación molecular nutrido por los datos generados por grupos

de investigación alrededor del mundo. Almacenando principalmente polimorfismos de un solo nucleótido pero también otros tipo de variantes genéticas, dbSNP organiza las variantes detectadas en etiquetas llamadas "Número de Referencia de SNP (*rs*)" (Sherry et al., 1999), referidos en este trabajo como *rsID*, los cuales son únicos para cada variante (por ejemplo, *rs22458*)

Los números de referencia de SNPs facilitan los estudios en campos de medicina genómica, genómica de poblaciones, genómica evolutiva, farmacogenómica, estudios de asociación, ya que proveen una notación estable para reportar nuevas variantes, realizar análisis de polimorfismos y minería e integración de datos. Esta notación es constante para todas las variantes almacenadas sin importar la versión del genoma de referencia que se utilizó para su descubrimiento.

2.1.2 La Base de Datos de Agregación de Genoma (gnomAD)

El repositorio más extenso de datos de secuenciación poblacional, la Base de Datos de Agregación de Genoma (en inglés Genome Aggregation Database; gnomAD. Karczewski et al., 2019), es el resultado de una coalición de investigadores alrededor del mundo que busca recopilar y agregar datos de secuenciación de exoma (la proporción del genoma que codifica para proteínas. Haworth et al., 2016) y de genoma completo, provenientes de proyectos poblacionales de gran escala.

La versión 2.1 de este repositorio incluye 16 millones de variantes de un solo nucleótido (SNV por Single Nucleotide Variant), un término utilizado para referirse a cambios de un solo nucleótido independientemente de su frecuencia alélica (si todas ocurrieran en una frecuencia alélica mayor al 1%, el término correcto para utilizar sería SNP (Arias et al., 1991); provenientes de 125,748 exomas y 229 millones de SNVs provenientes de 15,708 genomas. Organizados en archivos VCF (ver material suplementario 1.1) por cromosoma (1 al 22 y cromosoma X). Todos los VCFs incluyen anotada la frecuencia alélica con la que ocurre cada variante en 7 principales poblaciones continentales (tabla 1).

Estos datos se encuentran públicamente disponibles para su descarga en https://gnomad.broadinstitute.org/downloads.

Tabla 1. Poblaciones en el repositorio de gnomAD. Se muestran las 7 principales poblaciones continentales (más la población no asignada) y el número de exomas y genomas de cada una que tiene el repositorio.

Población	Descripción	Genomas	Exomas	Total
afr	Africana/ Afroamericana	4,359	8,128	12,487
amr	Latina/Americana Mezclada	424	17,296	17,720
asj	Judía Ashkenazi	145	5,040	5,185
eas	Asiática del Este	780	9,197	9,977
fin	Finlandesa	1,738	10,824	12,562
nfe	Europea no Finlandesa	7,718	56,885	64,603
sas	Asiática del Sur	0	1,335	1,355
oth	Otra (Población no asignada)	544	3,070	3,614
Total		**15,708**	**125,748**	**141,456**

Esta versión incluye también subpoblaciones continentales que, por la poca relevancia con este trabajo, no han sido mostradas. La información contenida en esta tabla fue obtenida de la documentación de la versión 2.1 de gnomAD, disponible en *https://macarthurlab.org/2018/10/17/gnomad-v2-1/*.

GnomAD provee el panorama de variación a nivel global contra el cual podemos contrastar frecuencias alélicas en proyectos regionales, como 100G-MX.

2.2 Proyectos y Consorcios Enfocados en Estudiar Regiones Reguladoras del Genoma Humano

La gran cantidad de datos e información obtenida a partir de datos de secuenciación de genoma completo ha dado lugar al planteamiento de nuevas preguntas científicas. Una de ellas ha sido definir, dentro de las zonas no codificantes del genoma, qué regiones tienen un papel en la regulación de la expresión génica, cuáles son las características de estos elementos reguladores y cuál es la región genómica que las define. Existen diversos proyectos que se han enfocado en

el estudio de elementos reguladores a lo largo de todo el genoma, los cuales se describen a continuación:

ENCODE: El proyecto *ENcyclopedia of DNA Elements* fue un consorcio de colaboración internacional fundado por el National Human Genome Research Institute cuyo objetivo fue construir una lista comprensiva de todos los elementos funcionales en el genoma humano. Los elementos mapeados, y las metodologías usadas incluyen: regiones transcritas de RNA (obtenidas por RNA-seq, CAGE, RNA-PET y anotación manual), regiones codificantes de proteínas (obtenidas por espectrometría de masa), sitios de unión a factores de transcripción (obtenidas por ChIP-seq y DNase-seq), estructura de la cromatina (obtenida por DNase-seq, FAIRE-seq, Histone ChIP-seq y MNase-seq) y sitios de metilación de DNA (por ensayo RRBS), resultados que fueron obtenidos gracias al apoyo de métodos bioinformáticos y curación manual (The ENCODE Project Consortium, 2012).

FANTOM5: *Functional ANnotation Of the Mammalian genome*, es un proyecto internacional con el objetivo de crear un atlas de expresión génica en humanos partiendo de proyectos de secuenciación de DNA complementario (cDNA) para identificar sitios de inicio de la transcripción (The FANTOM Consortium and the RIKEN PMI and CLST (DGT), 2014). A través de la tecnología cap analysis of gene expression (CAGE), una metodología que captura las orillas de RNAs mensajeros para después ser secuenciados, este proyecto ha podido determinar genes, sitios de inicio de la transcripción y reconocer la actividad de elementos regulatorios como promotores y enhancers, entre otros (The FANTOM Consortium et al., 2014). FANTOM5 ha identificado más de 180,000 promotores y casi 44,000 secuencias candidatos a enhancers, ambos en genomas de células primarias de humano y ratón (The FANTOM Consortium and the RIKEN PMI and CLST (DGT), 2014).

RoadMap Epigenomics: El objetivo de este proyecto fue crear un mapa epigenético del humano usando secuenciación de nueva generación para analizar marcas genómicas como metilación del DNA, modificación de histonas y transcritos de RNA en diferentes tejidos y tipos celulares (Bernstein et al., 2010). Con esta información es posible anotar variantes en elementos reguladores con expresión diferencial por tipo de tejido o tipo celular (Zhou et al., 2015).

GTEx: El proyecto *Genotype-Tissue Expression* está formado por diversos grupos de investigación cuyo objetivo principal es el de establecer una relación entre variantes genéticas con rasgos o enfermedades analizando cambios en la expresión génica. Ayudando a determinar qué vías son las afectadas en alguna enfermedad. Entre sus datos se encuentra un atlas de expresión para la identificación putativa de regiones reguladoras, así como también para conocer los niveles de loci de expresión de caracteres cuantitativos (del inglés *expression Quantitative Trait Loci*, eQTL) asociados a enfermedades (Lonsdale et al., 2013).

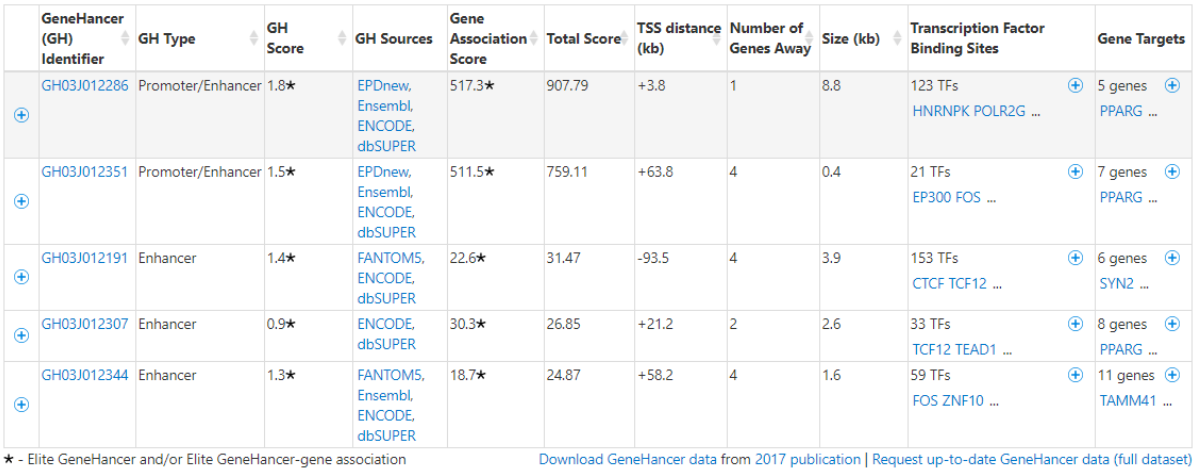
2.3 La Base de datos GeneHancer

GeneHancer es una base de datos desarrollada por el Weizmann Institute of Sciences y LifeMap Sciences Inc. de elementos enhancer detectados en humanos y sus genes blanco inferidos, integrados en el marco de GeneCards (Stelzer et al., 2016).

La base de datos contiene 284,834 elementos enhancers definidos, integrados a partir de 4 fuentes de datos genómicas: Encyclopedia of DNA Elements (ENCODE), Ensembl Regulatory Build, Functional Annotation of the Mammalian Genome (FANTOM) project, y el VISTA Enhancer Browser. De estos enhancers, 93,970 han sido derivados de más de una de las bases de datos antes mencionadas y se le ha asignado a cada elemento enhancer un puntaje derivado de su anotación; dicho puntaje permite distinguir elementos enhancer con mayor o menor evidencia de soporte.

Los elementos en GeneHancer tienen un identificador único e informativo llamado Identificador GeneHancer (GHid). Todos los identificadores inician con GH, seguidos del número de cromosoma donde se encuentra el elemento, una letra relacionada con la versión de GeneHancer (‘J’ para la versión 4.8) y una coordenada aproximada de la kilobase donde se encuentra.

En la base de datos, para cada elemento en GeneHancer se incluye la siguiente información: GHid, el tipo de elemento (enhancer, promotor/enhancer), los proyectos que reportaron al enhancer, tamaño del elemento, valor de confianza del elemento, factores de transcripción con sitios de unión en el elemento, patrones en tejidos de expresión y, de ser el caso, asociaciones enfermedad-enhancer con la base de datos DiseaseEnhancer (Zhang et al., 2018). Un ejemplo de la estructura de los datos de GeneHancer en GeneCards puede consultarse en la figura 5.



GeneHancer (GH) Identifier	GH Type	GH Score	GH Sources	Gene Association Score	Total Score	TSS distance (kb)	Number of Genes Away	Size (kb)	Transcription Factor Binding Sites	Gene Targets
GH03J012286	Promoter/Enhancer	1.8★	EPDnew, Ensembl, ENCODE, dbSUPER	517.3★	907.79	+3.8	1	8.8	123 TFs HNRNPK POLR2G ...	5 genes PPARG ...
GH03J012351	Promoter/Enhancer	1.5★	EPDnew, Ensembl, ENCODE, dbSUPER	511.5★	759.11	+63.8	4	0.4	21 TFs EP300 FOS ...	7 genes PPARG ...
GH03J012191	Enhancer	1.4★	FANTOM5, ENCODE, dbSUPER	22.6★	31.47	-93.5	4	3.9	153 TFs CTCF TCF12 ...	6 genes SYN2 ...
GH03J012307	Enhancer	0.9★	ENCODE, dbSUPER	30.3★	26.85	+21.2	2	2.6	33 TFs TCF12 TEAD1 ...	8 genes PPARG ...
GH03J012344	Enhancer	1.3★	FANTOM5, Ensembl, ENCODE, dbSUPER	18.7★	24.87	+58.2	4	1.6	59 TFs FOS ZNF10 ...	11 genes TAMM41 ...

★ - Elite GeneHancer and/or Elite GeneHancer-gene association

Download GeneHancer data from 2017 publication | Request up-to-date GeneHancer data (full dataset)

Figura 5. Estructura de los datos de GeneHancer dentro de GeneCards. Cada elemento en GeneHancer presenta información relevante, como el tipo de elemento, el puntaje asignado, los proyectos que lo reportaron, un puntaje de asociación con genes, la distancia con el sitio de inicio de transcripción, el tamaño, factores de transcripción que se unen al elemento y los genes a los que regula. Esta imagen fue capturada el día 26 de marzo de 2019 de https://www.genecards.org/.

La metodología del proyecto GeneHancer para identificar elementos enhancer y los genes a los que están asociados puede resumirse en los siguientes pasos:

1. **Minería de enhancers:** Obtención de los enhancers a partir de distintas bases de datos.
2. **Estimación del valor de confianza de los enhancers:** A los enhancers obtenidos se les asignó un valor que refleja la certeza de que, en efecto, se trata de un elemento regulador tipo enhancer.
3. **Estimación de una asociación enhancer-gen.** A través de distintos métodos, se infirió la relación reguladora que tiene cada enhancer con uno o varios genes.

Como un indicador de la calidad de los enhancers y asociaciones enhancer-gen contenidas en su base de datos, GeneHancer establece tres categorías *elite*:

- **Elite enhancer:** Elementos enhancer reportados por dos o más fuentes.
- **Asociaciones elite:** Asociaciones enhancer-gen apoyadas por dos o más métodos.
- **Elementos doble elite:** Enhancers elite con asociaciones elite.

El proceso de identificación de elementos en GeneHancer se describe con mayor detalle en el material suplementario 1.3 y un esquema general de la metodología del proyecto GeneHancer puede consultarse en la figura 6.

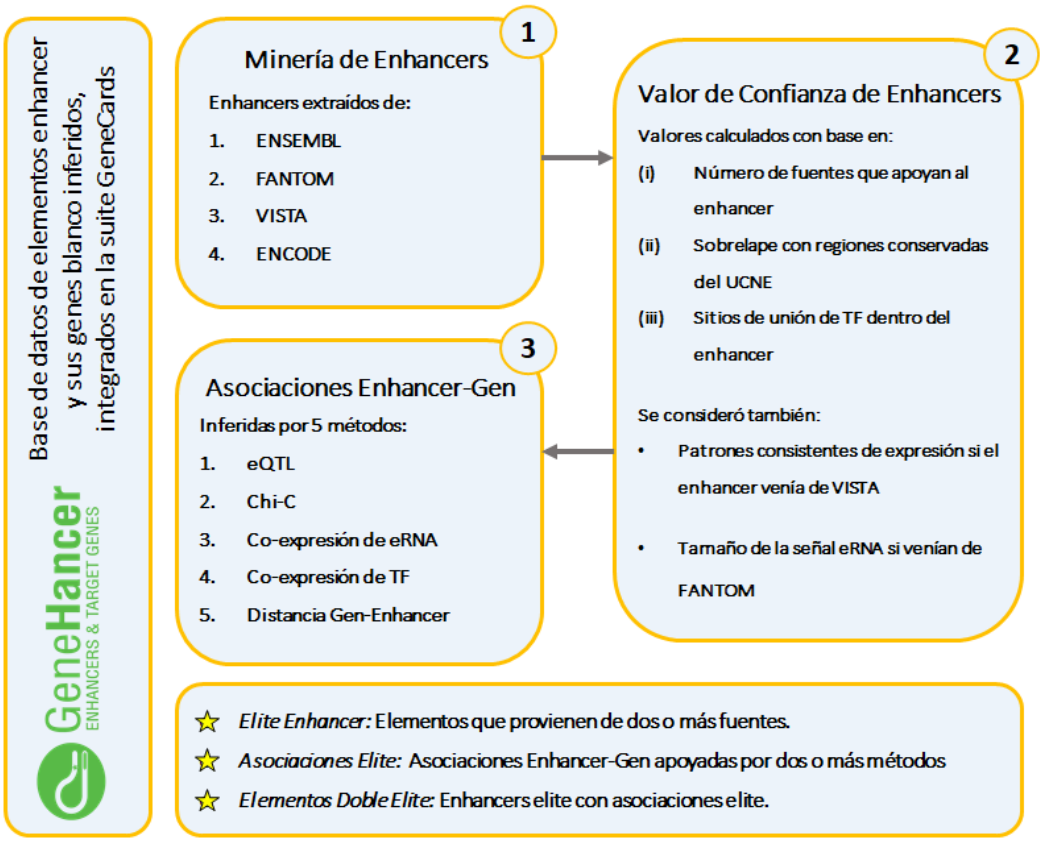


Figura 6. Metodología de GeneHancer. Se muestran los 3 pasos principales de la metodología seguida por el proyecto, desde la obtención de posibles enhancers hasta la asociación de elementos enhancer validados con posibles genes blanco. También, se destacan las tres tipos de categorías *élite* (estrellas) de los elementos contenidos en la base de datos.

En total, los elementos recopilados en GeneHancer cubren un 12.4% del genoma humano. Siendo que las regiones codificantes del genoma abarcan solamente alrededor del 1%, para exones, y alrededor del 3% para genes codificantes de proteínas (The ENCODE Project Consortium, 2012).

Es de esperar que exista gran cantidad de variantes genéticas en esta fracción del genoma. GeneHancer ofrece una buena oportunidad de estudiar la variación en región reguladora, utilizando elementos enhancer y promotores con fuerte evidencia que soporte su función

3. Justificación

Aunque son pocos los estudios genéticos específicamente enfocados en población mexicana (mestiza y/o nativa), en la última década se han identificado variantes clínicamente relevantes en población mestiza (Balam-Ortiz et al., 2011; Gamboa-Melendez et al., 2012; León-Mimila et al., 2013), dos de los estudios más notables han sido la identificación de variantes en el

haplotipo SLC16A11 como un factor de riesgo para la diabetes tipo 2 (The SIGMA Type 2 Diabetes Consortium et al., 2013) y la asociación de una variante en el gen PPARgamma con la resistencia a insulina en niños mexicanos (Stryjecki et al., 2016); así como variantes con relevancia clínica relacionadas con presión evolutiva en población nativa mexicana (Ávila-Arcos et al., 2019; Moreno-Estrada et al., 2014; Romero-Hidalgo et al., 2017). Sin embargo, además del trabajo de 12 genomas nativo mexicanos de (Romero-Hidalgo et al., 2017) donde se detectaron variantes en región promotora, ninguna otra investigación ha abordado el estudio de variantes en región reguladora en población nativa, especialmente en las regiones enhancer.

Este trabajo aprovecha los recientes datos de genoma completo del proyecto 100 G-MX, permitiendo estudiar los elementos reguladores tipo enhancer y aportando un análisis de las variantes contenidas en ellos. Comparando, además, la frecuencia en la que ocurren estas variantes en población nativa con la frecuencia en la que ocurren en otras poblaciones globales, permitiendo inferir posibles variantes de interés clínico.

4. Hipótesis

Existen variantes genéticas de interés clínico en regiones reguladoras tipo enhancer en la población nativa mexicana.

5. Objetivos

General: Estudiar variantes en regiones reguladoras en población Nativa Mexicana

Específico:

Caracterizar *in silico* posibles variantes de interés clínico para la población nativo-mexicana localizadas en elementos enhancer, definidas como aquellas con frecuencias alélicas diferenciables respecto a otras poblaciones

6. Materiales y Métodos

En este trabajo, se analizaron y describieron variantes genéticas con señal de selección (ver sección 6.1.2) en elementos reguladores de tipo enhancer detectados en población nativo-mexicana por el proyecto 100 G-MX. La figura 7 presenta un esquema general de la metodología seguida, la cual se describe de manera más extensa a continuación.

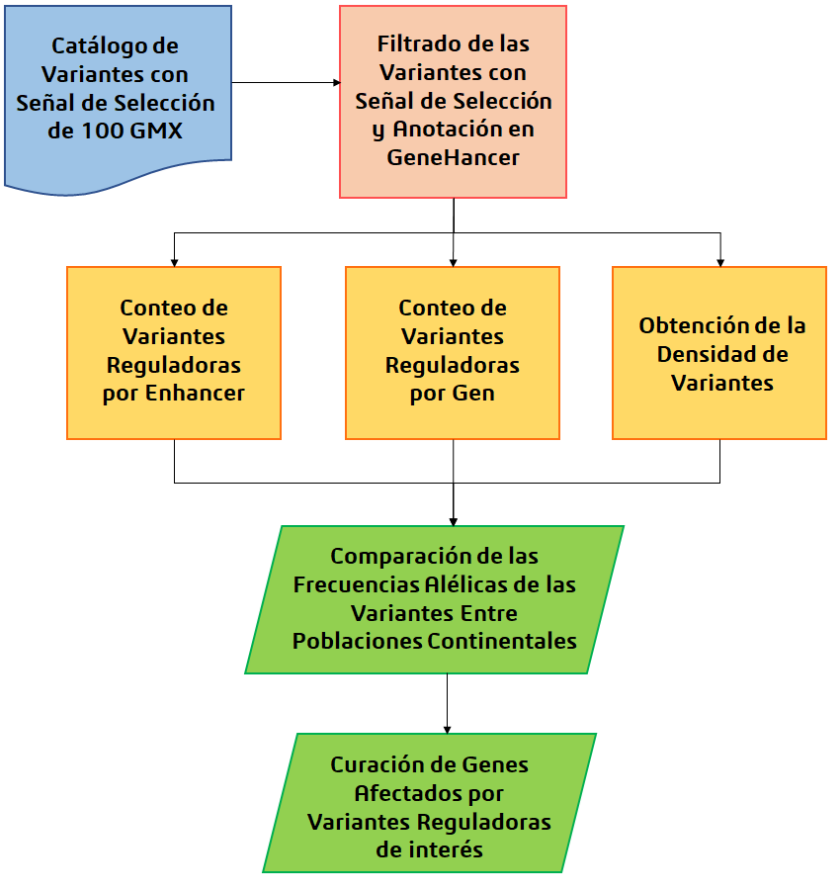


Figura 7. Metodología seguida para analizar las variantes detectadas en enhancers por el proyecto 100 G-MX.

El filtrado, operación o conteo de datos fue realizado a través de la línea de comandos de la terminal de un sistema Linux Ubuntu 16.04 o utilizando código escrito en el lenguaje de programación R.

6.1 Descripción de los Datos Usados para el Análisis

6.1.1 El Catálogo de Variantes del Proyecto 100 G-MX.

El principal resultado del proyecto 100 G-MX fue un catálogo de variantes, que contiene 9, 737 152 variantes genéticas, de las cuales 8,638,130 son polimorfismos de un solo nucleótido y 1, 099, 022 son indeles (inserciones-deleciones). Obtenidas a partir del llamado de variantes con el protocolo de GATK Best Practices (Ver Material suplementario 1.2), alineadas a la versión del genoma humano de referencia GRCh38.

Estas variantes fueron anotadas con información clínicamente relevante utilizando la herramienta nf-VEPextended (descargable en https://github.com/Iaguilaror/nf-VEPextended). Las bases de datos clínicamente relevantes fueron: 1) GeneHancer (enero 2019, v4.8. Fishilevich et al., 2017), utilizando solo los elementos doble elite, descargada del UCSC Table Browser (Karolchik, 2004), 2) PharmGKB, una base de datos con información sobre el impacto de la variación genética humana en la respuesta a fármacos (var_drug_ann.tsv descargada el 14/06/2019. Whirl-Carrillo et al., 2012), 3) GWAS Catalog, una base de datos de asociaciones SNP-fenotipo (v1.0.2. Buniello et al., 2019), 4) ClinVar, una base de datos de variación genética y su relación con la salud humana (clinvar_20190403. Landrum et al., 2016) y 5) gnomAD 2.1.1 (Karczewski et al., 2019. Sección 5.1.4), esta última es clave para comparar la frecuencia alélica de las variantes detectadas en población nativa mexicana con la de otras poblaciones.

La comparación de la frecuencia alélica de las variantes en nativos mexicanos con las de otras poblaciones continentales permite empezar a inferir a las variantes de interés, las cuales, bajo esta estrategia, tendrán una frecuencia alélica diferente a la de otras poblaciones, por razones que en este punto son aún desconocidas. Al estudiarlas, es posible detectar que se encuentran en alguna región que esté asociada a algún rasgo fenotípico de interés, por ejemplo, el desarrollo de alguna enfermedad.

De este catálogo, la fracción de datos analizados en este proyecto de tesis fue la de variantes de un solo nucleótido con señal de selección (ver sección 6.1.2) del catálogo que tuvieran anotación de GeneHancer. Buscando principalmente cuántos y qué genes son regulados por los enhancers donde fueron detectadas, los elementos enhancers involucrados y buscar variantes de interés particular al comparar la frecuencia alélica de estas variantes en población nativa con la frecuencia alélica de otras poblaciones continentales.

La tabla 2 muestra esta fracción en el catálogo obtenido por el proyecto 100 G-MX, un total de 1810 variantes, todas reportadas previamente en la base de datos dbSNP, representando la distribución de sus anotaciones en las bases de datos antes mencionadas.

Tabla 2. Catálogo de variantes de un solo nucleótido con señal de selección del proyecto 100 G-MX, con anotaciones en distintas bases de datos.

Base de datos	SNVs con Señal de Selección
dbSNP	1810
Clinvar	9
Codificantes	16
GeneHancer	108
GWAS Catalog	9
PharmGKB	0

La anotación de las variantes con las diferentes bases de datos no es mutuamente excluyente. Por ejemplo, una variante que tiene una anotación en GWAS Catalog también podría tener anotación GeneHancer (Figura 8).

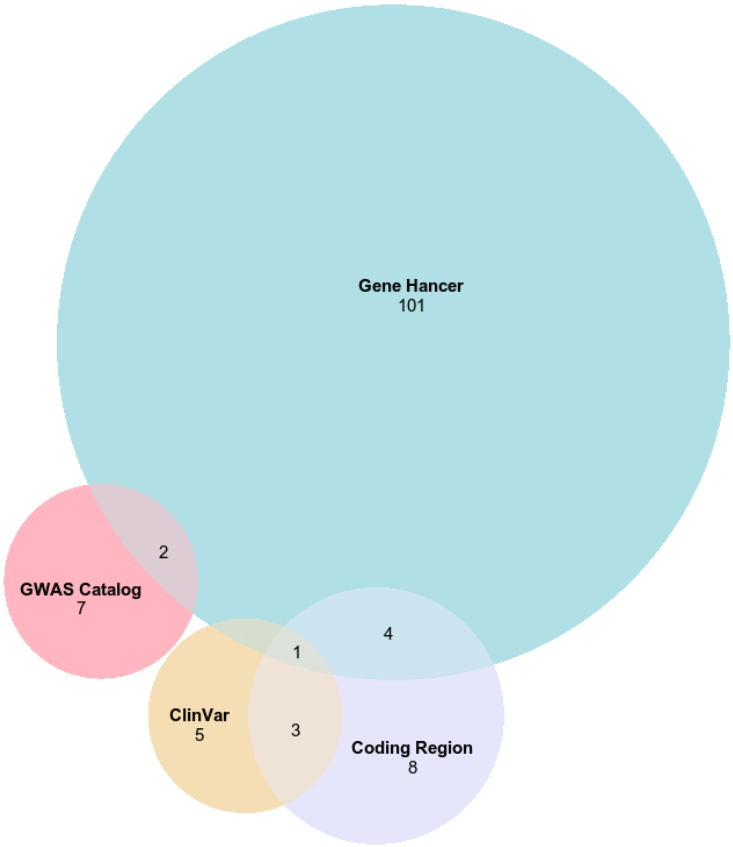


Figura 8. Diagrama de Venn de la distribución de variantes en distintas bases de datos

6.1.2 Variantes con Señal de Selección

Las poblaciones humanas se adaptan al ambiente donde residen. Dicha adaptación se puede estudiar a través de la detección de "señales de selección", las cuales son definidas como regiones en el genoma que han incrementado su frecuencia de manera preferente en una población (Pérez O'Brien et al., 2014), esto puede ser debido a su importancia funcional en procesos específicos, o a procesos de deriva génica. Las señales de selección pueden ayudar a entender procesos de presión evolutiva experimentados por poblaciones a lo largo de la historia. Y, anteriormente, se ha demostrado que estas señales pueden estar asociadas con regiones del genoma involucradas en diferenciación cultural y enfermedades complejas en humanos (Laland et al., 2010; Lappalainen et al., 2010).

Existen diferentes métodos y pruebas estadísticas utilizadas para encontrar señales de selección en el genoma; basados en frecuencias alélicas, en diferenciación dentro y entre poblaciones, contraste de diferentes tipos de mutaciones, basados en desequilibrio de ligamiento, entre otros, revisados con mayor detalle por Harris & Meyer, 2006, y Scheinfeldt & Tishkoff, 2013.

En el proyecto 100 G-MX, para identificar SNPs con señal de selección se utilizaron dos métodos: Population Branch Statistics; PBS (Yi et al., 2010), un método para identificar variantes genéticas con frecuencia alélica diferencial, y el Valor de Haplotipo Integrado (del inglés Integrated Haplotype Score; iHS (Voight et al., 2006), un método para detectar haplotipos más largos de lo esperado, los cuales indican la presencia de uno o más alelos que se encuentran bajo selección.

El análisis de PBS fue realizado utilizando el conjunto de datos de nativos mexicanos del proyecto, un conjunto de datos de población peruana con alta ancestría nativa y un conjunto de datos de población Han de China. Alternativamente, el análisis de iHS permitió detectar tramos extendidos de homocigosidad que incluyen sitios de interés como una señal de selección en el conjunto de datos de nativos mexicanos. El 1% de los resultados más significativos de ambos métodos fueron intersectados para determinar al conjunto de SNPs bajo selección en población nativa mexicana.

6.2 Metodología

Filtrado de Variantes con Señal de Selección y Anotación de GeneHancer

Se utilizó el conjunto de 1,810 variantes de un solo nucleótido con señal de selección del proyecto 100 G-MX. De este conjunto, se filtraron todas las variantes que contenían anotación de GeneHancer. Extrayendo también, para cada una su *rsID* y frecuencia alélica en las poblaciones nativa mexicana (obtenida por 100 G-MX), latina, asiática del este, europea no finlandesa y africana (obtenidas de la anotación con gnomAD 2.1.1).

A partir de la información contenida en la columna de anotación de GeneHancer, se obtuvo para cada variante el GHid, tipo de elemento regulador (enhancer o promotor/enhancer [1]) y el gen o genes que regula.

En total, se extrajeron 108 variantes con señal de selección y anotación de GeneHancer con las características antes discutidas. Esta información fue integrada en una tabla para los análisis posteriores.

[1] Es importante explicar la etiqueta “promotor/enhancer” en algunos elementos de GeneHancer. La base de datos argumenta que este término refleja la idea de que los elementos reguladores de tipo enhancer y promotor están íntimamente relacionados, teniendo en común algunos aspectos moleculares como hipersensibilidad a la DNAsa I,

modificaciones y actividades transcripcionales. Compartir estas características dificulta la diferenciación y obtención de uno sobre el otro con técnicas moleculares, especialmente cuando ambos tipos de elementos se encuentran genómicamente cercanos. Puntualmente, la etiqueta "promotor/enhancer" es usada por GeneHancer para identificar a los elementos enhancer recopilados del proyecto ENCODE, el cual etiqueta a algunos de sus elementos como "enhancers próximos" (Fishilevich et al., 2017). Utilizar datos que conllevan cierta ambigüedad respecto a la verdadera identidad del elemento podría crear falsos positivos; sin embargo, el proyecto GeneHancer considera importante incluirlos con la idea de que los elementos en su base de datos funcionen como puntos de partida para futuras investigaciones y análisis de caracterización en elementos reguladores.
Por lo anterior, en este trabajo se refiere a todos los elementos reguladores de GeneHancer involucrados como Elementos Enhancer. Especificando la etiqueta de cada elemento en la tabla S1 del material suplementario 2.

Conteo de Variantes por Elemento Enhancer y por Gen

A partir de las 108 variantes extraídas, se contó el número de veces que se repetía un GHid, obteniendo el número de variantes asociadas a cada elemento enhancer. De la misma manera, se realizó un conteo del número de instancias en las que se encontraba a cada gen, obteniendo el número de variantes por gen.

Los resultados de ambas operaciones fueron guardados en tablas separadas.

Obtención de la Densidad de Variantes

Dado que el tamaño de un enhancer podría explicar la cantidad de variantes detectadas en él, se normalizó la ocurrencia de variantes de acuerdo con la longitud de cada elemento. Para cada gen involucrado, se tomó el número de variantes que ocurren en sus enhancers reguladores y se dividió entre la sumatoria de la longitud en pares de base de estos. La siguiente ecuación describe este proceso:

$$\frac{Número\ de\ variantes\ en\ enhancers}{\Sigma\ longitud\ de\ elementos\ enhancers}\ x\ 1000$$

Obteniendo, respecto a cada gen involucrado, la densidad de variantes en elementos enhancer por kilobase.

Comparación de las Frecuencias Alélicas de las Variantes entre Poblaciones Continentales

De la tabla resultante del filtrado de variantes, para cada variante se extrajo el *rsID* y la frecuencia alélica que tiene en cada población. Creando una tabla que facilitara la comparación de la frecuencia con la que ocurre cada variante en población nativa mexicana con la frecuencia que ocurre en otras poblaciones continentales. Esta comparación permite, de manera somera, identificar rápidamente variantes de interés por la diferencia o el patrón en la que ocurre su frecuencia alélica en distintas poblaciones.

Curación de Genes Afectados por Variantes en sus Enhancers Reguladores

Finalmente, se realizó una curación manual de los 5 genes con mayor número de variantes en su elemento (o elementos) regulador enhancer. Buscando información relevante en la literatura científica sobre su función, su principal tejido de expresión, y, de estar disponible, información sobre efectos de mutaciones en ese gen y las poblaciones en las que se ha estudiado. De manera general, recopilar esta información permite establecer una idea del contexto genético de las variantes: Si están asociados a alguna enfermedad, si existe algún patrón en los tejidos de expresión, si son especialmente relevantes para una población, etc. Estableciendo una primera línea de información para guiar futuros trabajos de investigación.

7. Resultados

7.2 Variación Genética en Regiones Enhancer de 76 Genomas de población Nativa Mexicana

Después de filtrar las variantes con señal de selección y con anotación de GeneHancer en catálogo de variantes del proyecto 100 G-MX, fueron obtenidos y analizados un total de 108 SNPs. La figura 9 muestra una distribución de la frecuencia alélica de estas variantes en la población nativa mexicana comparada con la frecuencia que tienen en otras poblaciones continentales. Resultan interesantes aquellas variantes que tienen alta frecuencia en población nativa y baja en todas las demás poblaciones (excepto la población latina), ya que podrían tratarse de variantes propias de la población nativa de México, o de variantes que por alguna razón han aumentado su frecuencia en esta población. Una tabla que acompaña a la figura 9 puede consultarse en el material suplementario 2, tabla S1.

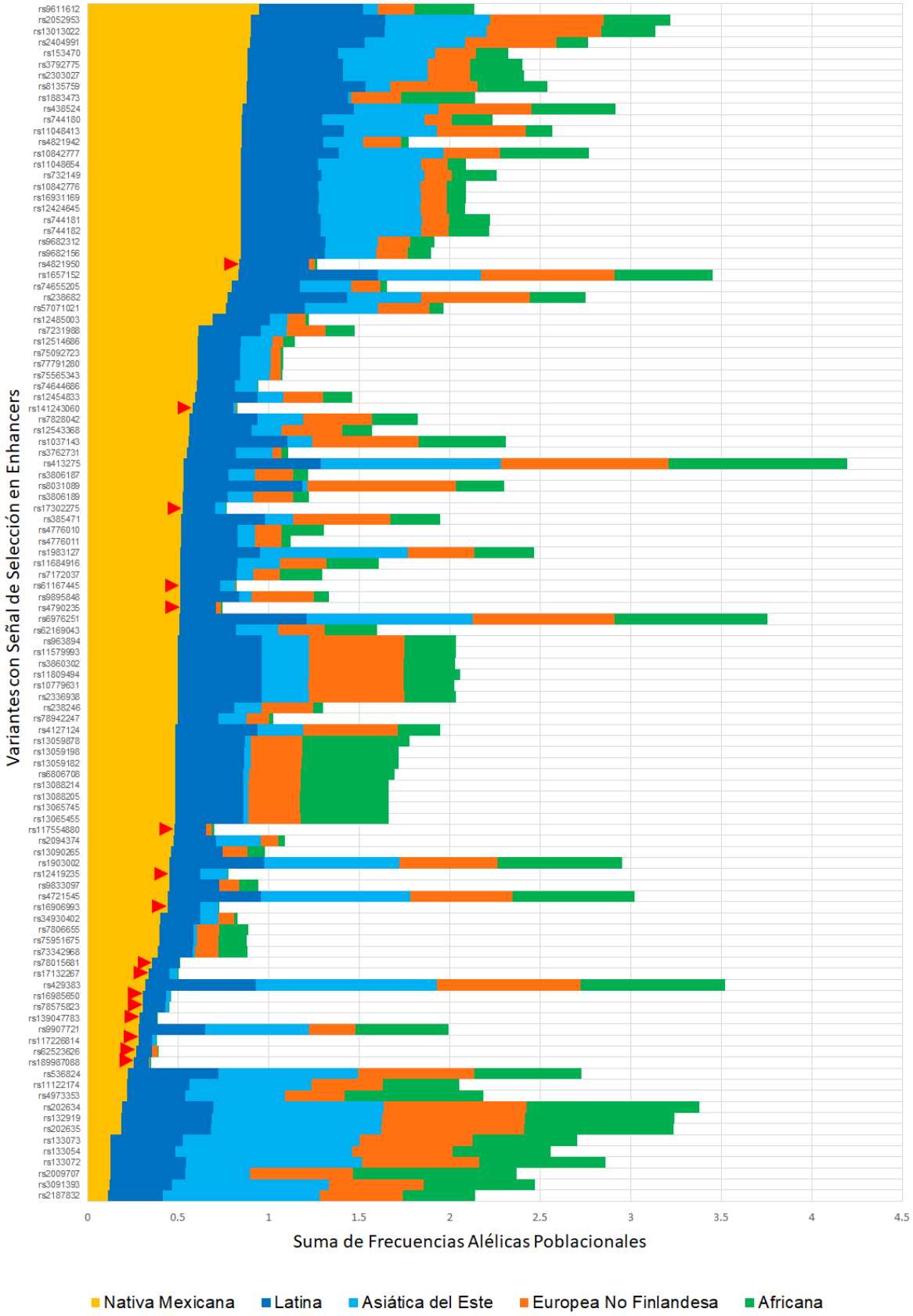


Figura 9. Suma de frecuencias alélicas de las variantes en poblaciones continentales. Se muestra la suma de las frecuencias alélicas en diferentes poblaciones para cada variante, ordenadas por la frecuencia alélica de la población nativa mexicana, en cada población, y representadas por bloques apilados. Aquellas variantes que tienen una frecuencia alélica alta en nativos mexicanos, pero baja en cualquier otra población (excepto la latina) resultan de especial interés y están señaladas por triángulos rojos.

7.3 Variantes por Elemento Enhancer

Se encontró que los 108 SNPs obtenidos se encuentran distribuidos en 74 elementos reguladores enhancer. Sólo en el 10.81% de ellos se detectaron 3 o más variantes, 13.52% contienen 2 variantes y el 75.67% restante solamente cuenta con una variante.

La tabla 3 muestra al grupo de enhancers que tuvieron más de tres variantes. Destaca el enhancer con el identificador *GH03J012439*, un regulador del gen PPARG y SYN2, con seis variantes.

Tabla 3. **Elementos Enhancer con Tres o Más Variantes**

GeneHancerID	Variantes	Coordenadas Genómicas (GRCh38/hg38)	Genes Blanco	Función del Gen Blanco
GH03J012439	6	chr3:12439068-12447058	PPARG SYN2	**PPARG:** Involucrado en la señalización de nutrientes y regulación del metabolismo de lípidos y carbohidratos (Pérez et al., 2007). **SYN2:** involucrada en la plasticidad y transmisión sináptica (Cruceanu et al., 2012)
GH01J206448	5	chr1:206448621-206452230	IKBKE	Quinasa no canónica que actúa como modulador positivo en el proceso de autofagia; también, ha sido descrito como un oncogen (Leonardi et al., 2019).
GH12J026690	5	chr12:26688001-26693327	ITPR2	Segundo mensajero de los canales liberadores de calcio.(Wiel et al.,

				2014, p. 2)
GH12J026703	4	chr12:26700401-26706799	ITPR2	Segundo mensajero de los canales liberadores de calcio (Wiel et al., 2014, p. 2).
GH03J012458	3	chr3:12456601-12465199	PPARG TSEN2	**PPARG:** Involucrado en la señalización de nutrientes y regulación del metabolismo de lípidos y carbohidratos (Pérez et al., 2007). **TSEN2:** Subunidad del complejo de endonucleasa de empalme de tRNA (Bierhals et al., 2013, p. 2)
GH05J096699	3	chr5: 96720801-96722400	CAST ERAP1 ENSG00000248734 ENSG00000249180	**CAST:** Inhibidor de cisteína proteasas dependientes de calcio (Ai et al., 2013, p. 2). **ERAP1:** Aminopeptidasa involucrada en el corte de péptidos para ser cargados en el complejo mayor de histocompatibilidad de clase I (Padula et al., 2019). **ENSG00000248734:** RNA largo no codificante con función desconocida. **ENSG00000249180:** RNA largo no codificante con función desconocida.

GH15J052178	3	chr15:52178400-52181227	CERNA1 GNB5	**CERNA1:** RNA largo no codificante, involucrado en la inhibición la apoptosis de células endoteliales (Lu et al., 2016). **GNB5:** Regula negativamente la señalización de proteínas G en el sistema nervioso central (Poke et al., 2019, p. 5).
GH22J041409	3	chr22:41409898-41416053	EP300	Transfiere grupos acetilo a las histonas para promover la transcripción génica (Kalkhoven, 2004).

Una tabla con el número de variantes para todos los elementos enhancers puede consultarse en el material suplementario 2, tabla S2.

7.4 Genes con Variantes Localizadas en Elementos Reguladores Enhancer

Para cada elemento enhancer involucrado, se identificó al gen o genes que regula, de acuerdo con GeneHancer; así como el número de variantes localizadas en cada uno. La tabla 5 reporta a los cinco genes con mayor número de variantes en sus elementos enhancer. Se encontró que las 108 variantes en enhancers están asociadas a 100 genes; cabe aclarar, que una variante puede estar relacionada a 2 o más genes debido al multi-control que pueden ejercer los elementos enhancer. De estos, se curó información relevante sobre los cinco genes con más variantes en su enhancer o enhancers reguladores. Se reporta principalmente la función del gen, efectos asociados a mutaciones y principales tejidos de expresión.

Tabla 5. Cinco Genes con el Mayor Número de Variantes en sus Elementos Reguladores Enhancer

Gen	Variantes en Enhancers

PPARG	10
ITPR2	9
IKBKE	7
MKL1	6
SYN2	6

PPARG: La proteína codificada por este gen, Receptor de Peroxisoma-Proliferador-Activado Gamma, es un regulador de la diferenciación de adipocitos. La función de los Receptores Activados por Proliferadores de Peroxisomas (PPARs), como PPARG, es unirse a químicos que inducen la proliferación de peroxisomas, que contribuyen a la oxidación de ácidos grasos. Este tipo de receptores puede activarse por ácidos grasos provenientes de la dieta o derivados de procesos metabólicos en el cuerpo. Sirven como sensores de lípidos que al activarse pueden redirigir el metabolismo. El principal tejido de expresión de este gen es en el tejido adiposo (Evans et al., 2004).

ITPR2: La proteína Receptor Tipo 2 de Inositol 1,4,5-trifosfato, perteneciente a la familia de los receptores inositol 1,4,5-trifosfato cuyos miembros son segundos mensajeros de los canales intracelulares liberadores de calcio. Estas proteínas son mediadores del aumento en la cantidad de calcio citoplasmático en respuesta a la producción de inositol trifosfato por la activación del receptor. Esta señalización mediada por receptor de inositol trifosfato está involucrada en procesos de migración celular, división celular, contracción del músculo liso, y señalización neuronal. Mutaciones en este gen se han asociado a anhidrosis (dificultad para sudar), sugiriendo que la liberación de calcio intracelular es requerida para la la producción de sudor en glándulas ecrinas (Shi et al., 2015; Wiel et al., 2014). Los principales tejidos de expresión en este gen son hígado y riñón (O'Leary et al., 2016).

IKBKE: Inhibidor del Factor Nuclear Kappa B Quinasa Subunidad Epsilon, es una quinasa I-kappa-B no canónica (IKK) que actúa como modulador positivo en el proceso de autofagia (Leonardi et al., 2019). También ha sido identificado como un oncogén de cáncer de mama, al ser sobrexpresado en más del 30% de los carcinomas mamarios y líneas celulares de cáncer de mama (Hutti et al., 2009).

MKL1 (MRTFA): Leucemia Megacarioblástica 1 ó Factor de Transcripción A Relacionado a Miocardina, la proteína codificada por este gen interactúa con el factor de transcripción miocardina, un regulador en la diferenciación del tejido muscular liso (Alajbegovic et al., 2017). Ha sido involucrado en un evento de translocación específica, creando una fusión con la proteína RBM15 (RNA binding motif protein 15). Esta translocación ha sido asociada con leucemia megacarioblástica aguda. El gen también ha sido asociado con esquizofrenia en la población Han de China (Wang et al., 2016). Su principal tejido de expresión es la médula ósea (O'Leary et al., 2016).

SYN2: Synapsina II es un miembro de la familia de genes sinapsinas, las cuales codifican fosfoproteínas neuronales que se asocian con la superficie de vesículas sinápticas. Algunos polimorfismos en este gen se han asociado con funciones presinápticas anormales y algunos desórdenes neuronales como el autismo, epilepsia, desorden bipolar y esquizofrenia (Corradi et al., 2014; Cruceanu et al., 2016; Saviouk et al., 2007). El principal tejido de expresión de este gen es en el cerebro (O'Leary et al., 2016).

Una tabla con el número de variantes en cada uno de los genes de este estudio puede consultarse en el material suplementario 2, tabla S3.

7.5 Densidad de Variantes en Elementos Reguladores por Gen

Se encontró que los genes con mayor densidad de variantes en sus elementos reguladores (tomando en cuenta la longitud de cada elemento) difieren de aquellos genes donde se detectaron más variantes en sus enhancers (sección 7.4), confirmando que el tamaño de un enhancer explica la cantidad de variantes detectadas en él. La tabla 6 presenta a los cinco genes con la mayor densidad de variantes, calculada como el número de variantes en enhancers por kilobase.

Tabla 6. Cinco Genes con la Mayor Densidad de Variantes en sus Elementos Reguladores Enhancer.

Gen	Número de Variantes por Kilobase de Región Reguladora	Total de Variantes Encontradas	Longitud Total de Elementos Enhancer en pares de base	Número de Elementos Enhancer

LOC101928386	1.11731844	2	1790	2
CERNA1	1.07526882	3	2790	1
NEK4P2	1.00806452	1	992	1
SLURP1	0.75075075	2	2664	1
DUSP23	0.6097561	2	3280	1

Tres genes tienen valores similares que conforman el máximo número de variantes por kilobase de longitud de sus enhancers reguladores. LOC101928386, un gen que al transcribirse crea un RNA largo no codificante sin función conocida (GeneID de NCBI: 101928386), es el que presenta la mayor densidad de variantes, aproximadamente 1.1 variantes por kilobase. Seguido de CERNA1, un gen que codifica para un RNA largo no codificante cuyos blancos son *miR-4707-5p* y *miR-4767*, y que juega un papel en inhibir la apoptosis de células endoteliales, la cual está relacionada con enfermedades cardiovasculares (Lu et al., 2016), y NEK4P2, un pseudogen de la quinasa 4 relacionada a NIMA (Never in Mitosis gene A) con función desconocida (GeneID de NCBI: 100421734), los cuales tienen una densidad de variantes casi igual a la máxima.

El resto de los genes presenta menos de una variante por kilobase. Con SLURP1, que codifica para una proteína involucrada en la transducción de señales transmembranales y en el desarrollo del mal de Meleda (Bergqvist et al., 2018), y DUSP23, la cual codifica para una fosfatasa de doble especificidad de sustrato (Wu et al., 2004).

Una tabla con la densidad de variantes por cada uno de los 100 genes implicados en este estudio puede consultarse en la sección del material suplementario 2, tabla S4.

7.6 Variantes en Enhancers de Interés Clínico

Durante el proceso de filtrado de las variantes se capturaron dos que además de tener anotación de GeneHancer tuvieron anotación en la base de datos GWAS Catalog (Buniello et al., 2019),

un repositorio de variantes genómicas curadas asociadas a enfermedades, convirtiendo a estas en variantes de interés clínico.
Estas variantes fueron exploradas más detalladamente, enfocándose en el fenotipo asociado y en los estudios poblacionales en los que habían sido descritas:

1) **rs12485003**. Localizada en la posición 40,239,272 del cromosoma 22, la cual se encuentra dentro de una región enhancer que regula al gen TNRC6B reportada por Giri y colaboradores en 2019. La asociación se hizo con 365,998 individuos de ancestría europea, 63,490 individuos de ancestría africana, 22,802 individuos de ancestría Hispana, 4,792 individuos de ancestría asiática y 2,695 individuos de ancestría Nativo Americana provenientes de los proyectos poblacionales UK Biobank y Million Veteran Program (donde cerca de la mitad estaban bajo tratamientos hipertensivos y un cuarto padecía diabetes de algún tipo). Se observó, con un P-value de 1x10^-16, que sugiere que el alelo A en rs12485003 produce un aumento de 0.4076 mmHg en la medida de la presión de pulso sanguíneo.

2) **rs9895848**. Localizada en la posición 4,835,663 del cromosoma 17, la cual se encuentra dentro de una región reportada por GeneHancer con la etiqueta Promotor/Enhancer, que regula al gen MINK1. La asociación fue identificada por (Astle et al., 2016), donde se estudiaron 171,643 individuos de ancestría Europea provenientes de los proyectos poblacionales INTERVAL, UK Biobank y UK BeLIEVE. En este estudio, se observó, con un P-value de 1x10^-14 que el alelo alterno T en rs9895848 produce una reducción de 0.03045967 unidades en la cuenta de linfocitos.

8. Discusión

Variantes con Señal de Selección en Enhancers de Población Nativa Mexicana

Previamente, estudios de variantes con señal de selección han detectado o confirmado diferentes regiones importantes relacionadas con eventos de adaptación humana. Algunos ejemplos incluyen la persistencia de la lactasa (Bersaglieri et al., 2004), pigmentación de la piel (The International HapMap Consortium et al., 2007; Voight et al., 2006; Williamson et al., 2007) y morfología del cabello (Fujimoto et al., 2008; J. L. Kelley, 2006). No obstante, hasta

ese momento la diversidad de las poblaciones usadas para estos estudios ha sido muy baja, por lo que en 2009 (Pickrell et al., 2009) buscaron señales de selección en 53 poblaciones diversas (incluyendo muestras de la población maya) utilizando metodologías basadas en el análisis de haplotipos. Sin embargo, a pesar de los esfuerzos por caracterizar variantes con señal de selección, ninguno de los proyectos antes mencionados ha detectado variantes con señal de selección en elementos enhancer. Esto debido al contexto tecnológico del periodo en el que se llevaron a cabo los estudios, ya que la anotación de elementos reguladores era limitada (Joanna L. Kelley & Swanson, 2008), además de que no se disponía con suficiente información de genomas completos. Hoy, proyectos como 100 G-MX y fuentes de anotación como la base de datos de GeneHancer hacen posible esta tarea.

Describimos 108 SNPs con señal de selección presentes en 100 elementos enhancer, obtenidas por el proyecto 100G-MX. Esta cantidad de SNPs no ha podido ser comparada por falta de otras investigaciones sobre señales de selección en enhancers. Sin embargo, en estudios anteriores (Voight et al., 2006; Williamson et al., 2007), los cuales han analizado principalmente población africana, europea y asiática, se ha encontrado que algunas regiones no codificantes incluyen fuertes señales de selección.

Los tejidos de expresión de los cinco genes con mayor número de variantes en su elemento regulador no demostraron algún patrón aparente entre ellos, encontrando expresión en diversos tejidos que van desde tejido adiposo, hasta riñón y médula ósea. Por otro lado, es interesante mencionar que polimorfismos en los genes con enhancers asociados, IKBKE y MKL1(MRTFA) se han asociado al desarrollo de ciertos tipos de cáncer (cáncer de mama y leucemia megacarioblástica aguda, respectivamente); mientras que polimorfismos en MKL1(MRTFA) y SYN2 han demostrado estar asociados a desórdenes neuronales como la esquizofrenia, autismo, epilepsia y desorden bipolar. Ya que se trata de variantes con señal de selección, estos polimorfismos podrían estar asociados a rasgos biomédicamente relevantes en la población nativo-mexicana; no obstante, el poco estudio que se les ha dedicado a los elementos enhancers de estos genes presenta el obstáculo principal para definir, con certeza, el rol que juegan en el desarrollo de enfermedades a las que sus genes regulados están asociadas.

Adicionalmente, para profundizar más la investigación de los elementos reguladores, todos los enhancers donde se detectaron variantes con señal de selección fueron intersectados con aquellos presentes en la base de datos DiseaseEnhancer (Zhang et al., 2018), buscando enhancers que previamente hayan sido asociados a enfermedades en nuestro conjunto de datos.

Sin embargo, no se encontró ningún sobrelape de regiones entre los enhancers analizados y los de DiseaseEnhancer.

Variantes en Elementos Enhancer de PPARG

PPARG es el gen en el que se detectaron más variantes en elementos enhancer que lo regulan, con 10 variantes distribuidas en tres elementos enhancer: GH03J012436, GH03J012439 y GH03J012458, ubicados río abajo del gen (Figura 10).

Este gen pertenece a la familia de receptores nucleares PPARs (Receptor Activado por Proliferadores de Peroxisomas), junto con PPARdelta y PPARalpha, PPARgamma está involucrado en la señalización de nutrientes y regulación del metabolismo de lípidos y carbohidratos, principalmente en procesos de asimilación celular de los lípidos por vías anabólicas, en la diferenciación de adipocitos y en la síntesis de triglicéridos (Pérez et al., 2007).

La importancia clínica de PPARG ha sido altamente resaltada ya que mutaciones en este gen pueden llevar al Síndrome de Resistencia al Ligando de PPARgamma (PLRS). Un síndrome cuyos síntomas incluyen hipertensión, diabetes mellitus, resistencia a la insulina, dislipidemia, lipodistrofia, síndrome del ovario poliquístico y enfermedad del hígado graso (Semple, 2006). Las principales mutaciones que se han observado dentro de este gen, y asociado a PLRS, son: P467L, V290M, R425C y F388L (Agarwal & Garg, 2002; Barroso et al., 1999; Hegele et al., 2002)

Este resultó ser un hallazgo especialmente interesante, ya que variantes en PPARG se han asociado previamente a la diabetes tipo 2 (Altshuler et al., 2000; Chan et al., 2013; Kilpeläinen et al., 2008), una enfermedad prevalente en la población mexicana (Hernández-Ávila et al., 2013), y a la resistencia a insulina en niños mexicanos (Stryjecki et al., 2016).

PPARG Comparte Enhancers con SYN2, TIMP4 y TSEN2.

Adicionalmente, se encontró que estos elementos enhancer regulan no solo al gen PPARG, sino que también comparten otros genes blanco (Figura 10). El enhancer GH03J012439 tiene además como gen blanco a SYN2, que codifica para una fosfoproteína neuronal involucrada en la plasticidad y transmisión sináptica (Cruceanu et al., 2012); GH03J012436 regula también a TIMP4, cuya proteína pertenece a la familia de inhibidores tisulares de metaloproteinasa, con actividades en la diferenciación celular, proliferación y apoptosis (Melendez-Zajgla et al., 2008); y GH03J012458 a TSEN2, un gen que codifica para una subunidad del complejo de

endonucleasa de empalme de tRNA, el cual ayuda a remover intrones durante la maduración de las moléculas de tRNA (Bierhals et al., 2013, p. 2). A pesar de esto, PPARG es el gen que es regulado por todos estos elementos enhancer, sugiriendo que estas variantes están más íntimamente relacionadas con él que con los otros genes.

Una tabla con las variantes detectadas en enhancers de estos genes, y su distribución de frecuencias alélicas puede consultarse en el material suplementario 2, tabla S5.

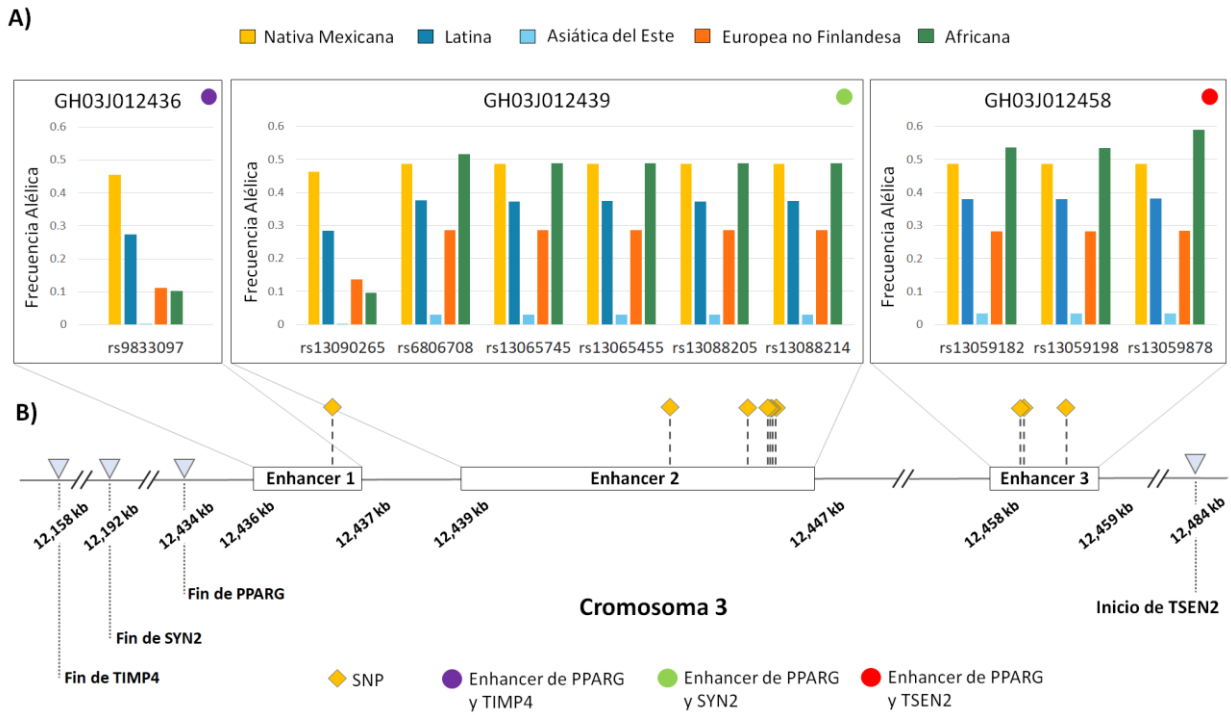


Figura 10. Variantes detectadas en enhancers reguladores del gen PPARG, SYN2, TIMP4 y TSEN2. (A) Comparación de la frecuencia alélica de las variantes detectadas en población nativo-mexicana con 4 poblaciones continentales. **(B)** Localización genómica de las variantes, de los elementos enhancer que las contienen y de los genes regulados.

Posible Haplotipo de Variantes en Enhancers reguladores del Gen PPARG, SYN2, TIMP4 y TSEN2.

Las variantes detectadas en elementos enhancer de los genes PPARgamma, SYN2, TIMP4 y TSEN2 abren la puerta a nuevas y emocionantes hipótesis sobre variantes en enhancers en población nativa mexicana. Particularmente, al comparar la frecuencia con la que ocurren estas variantes en población nativa respecto a su ocurrencia en otras poblaciones, se distingue un patrón en el que las frecuencias alélicas decaen en las poblaciones asiática del este y europea no finlandesa, mientras que se mantienen en proporciones similares en población africana y nativa mexicana (Figura 11-A); sugiriendo que estas variantes podrían pertenecer a un haplotipo y que, bajo el modelo del poblamiento de la tierra “Fuera de África” (en inglés Out

of Africa model (Rightmire, 2009; Tattersall, 2009)), podrían haberse originado en África, por alguna razón redujeron su frecuencia alélica en humanos que migraron y habitaban las regiones de Asia del este; y que no aumentó su frecuencia sino quizás hasta después de las migraciones a través del estrecho de Bering y el asentamiento del hombre en lo que hoy es el territorio mexicano (Figura 11-B). Notablemente, este haplotipo se extiende al menos unas cientos de pares de bases más en población nativa mexicana que en población africana en dirección río arriba hacia el final de PPARG, SYN2 y TIMP2, demostrado por el repentino decremento de la frecuencia alélica de las variantes *rs9833097* y *rs13090265* en población africana (Figura 11-A).

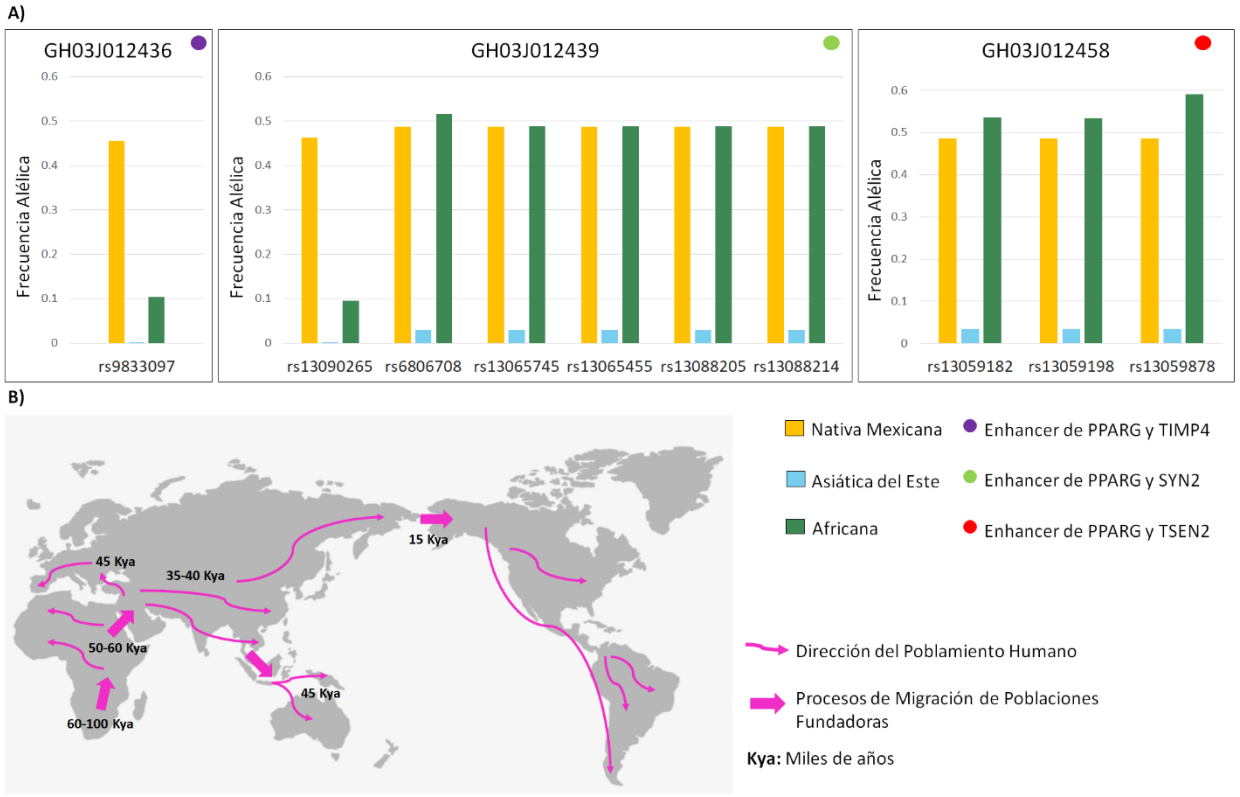


Figura 11. Posible haplotipo de variantes en enhancers del gen PPARG, SYN2, TIMP4 y TSEN2. (A) Comparación de la frecuencia alélica de las variantes en enhancers detectadas en población nativa con la frecuencia alélica que tienen en las poblaciones asiática y africana**. (B)** Patrones de migración putativos de los humanos modernos durante los últimos cien mil años, empezando con una población inicial en África y terminando con el poblamiento de Sudamérica hace aproximadamente 12-14 mil años. Este mapa fue adaptado a partir del trabajo de Henn et al., 2012.

9. Conclusiones

Este trabajo reporta el análisis de 108 SNPs con señal de selección en regiones enhancer, detectadas en 27 de los 68 grupos indígenas en México por el proyecto 100 G-MX, describiendo aspectos de interés como el número de variantes por enhancer y el número de variantes por gen regulado, entre otros. El hallazgo más importante fueron las variantes en regiones reguladoras del gen PPARG, un gen asociado al metabolismo de lípidos y con implicaciones patológicas previamente reportadas en población mestiza, el cual ha presentado el mayor número de variantes en sus enhancers reguladores. Estas variantes en enhancers, los cuales también regulan a los genes SYN2, TIMP4 y TSEN2, parecen pertenecer a un haplotipo compartido por la población nativo-mexicana y la población africana, las cuales deben explorarse con más detalle. Así mismo, se encontraron dos variantes previamente analizadas en estudios de asociación y que son de interés clínico, involucradas en un aumento en la presión del pulso sanguíneo y en un cambio en la cuenta de linfocitos.

Con estos hallazgos, se remarca la importancia de los estudios genómicos en poblaciones poco representadas, como es el caso de la población nativa-mexicana, y la necesidad de estudiar y validar el papel que juegan las variantes genéticas cuando ocurren en regiones reguladoras tipo enhancer. Los resultados de este trabajo servirán como puntos de partida para líneas de investigación sobre la variación en regiones reguladoras y su posible impacto clínico, bajo la directriz de impulsar la medicina genómica en el país.

10. Perspectivas

Habiendo identificado las variantes de interés discutidas en este proyecto, es importante realizar la validación experimental para poder esclarecer el efecto que tienen en la interacción promotor-enhancer y, consecuentemente, en la regulación de la expresión génica. Considerando especialmente a las variantes en enhancers de PPARgamma por ser el gen cuyas mutaciones han tenido un claro efecto en la población mexicana, estando asociadas a padecimientos como la diabetes tipo 2 (Altshuler et al., 2000; Chan et al., 2013; Kilpeläinen et al., 2008) y la resistencia a insulina en niños mexicanos (Stryjecki et al., 2016).

Respecto al posible haplotipo, es necesario explorar la frecuencia alélica de variantes genómicamente cercanas, y ver cómo se compara su ocurrencia en población nativo-mexicana con la de poblaciones africanas para poder definir su longitud.

A partir de los hallazgos de este trabajo, sería posible detectar a las poblaciones fundadoras de los grupos nativos en el continente americano, explorando el genoma de grupos nativos de Asia, particularmente aquellos asentados cerca de Beringia.

Finalmente, bases de datos como GeneHancer han facilitado la definición de elementos reguladores tipo enhancer, ayudando enormemente al estudio de los elementos reguladores. No obstante, es deseable que futuros proyectos de variación en el genoma humano empiecen a integrar estas anotaciones a sus estudios, enriqueciendo el poco conocimiento que se tiene de ellos en este momento.

11. Referencias

1. Agarwal, A. K., & Garg, A. (2002). A Novel Heterozygous Mutation in Peroxisome Proliferator-Activated Receptor-γ Gene in a Patient with Familial Partial Lipodystrophy. *The Journal of Clinical Endocrinology & Metabolism*, *87*(1), 408–408. https://doi.org/10.1210/jcem.87.1.8290
2. Alajbegovic, A., Turczyńska, K. M., Hien, T. T., Cidad, P., Swärd, K., Hellstrand, P., Della Corte, A., Forte, A., & Albinsson, S. (2017). Regulation of microRNA expression in vascular smooth muscle by MRTF-A and actin polymerization. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research*, *1864*(6), 1088–1098. https://doi.org/10.1016/j.bbamcr.2016.12.005
3. Altshuler, D., Hirschhorn, J. N., Klannemark, M., Lindgren, C. M., Vohl, M.-C., Nemesh, J., Lane, C. R., Schaffner, S. F., Bolk, S., Brewer, C., Tuomi, T., Gaudet, D., Hudson, T. J., Daly, M., Groop, L., & Lander, E. S. (2000). The common PPARγ Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes. *Nature Genetics*, *26*(1), 76–80. https://doi.org/10.1038/79216
4. Arias, T. D., Jorge, L. F., & Barrantes, R. (1991). Uses and misuses of definitions of genetic polymorphism. A perspective from population pharmacogenetics. *British Journal of Clinical Pharmacology*, *31*(1), 117–120.
5. Ashcroft, F. M., & Rorsman, P. (2012). Diabetes Mellitus and the β Cell: The Last Ten Years. *Cell*, *148*(6), 1160–1171. https://doi.org/10.1016/j.cell.2012.02.010
6. Astle, W. J., Elding, H., Jiang, T., Allen, D., Ruklisa, D., Mann, A. L., Mead, D., Bouman, H., Riveros-Mckay, F., Kostadima, M. A., Lambourne, J. J., Sivapalaratnam, S., Downes, K., Kundu, K., Bomba, L., Berentsen, K., Bradley, J. R., Daugherty, L. C., Delaneau, O., … Soranzo, N. (2016). The Allelic Landscape

of Human Blood Cell Trait Variation and Links to Common Complex Disease. *Cell*, *167*(5), 1415-1429.e19. https://doi.org/10.1016/j.cell.2016.10.042

7. Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., Banks, E., Garimella, K. V., Altshuler, D., Gabriel, S., & DePristo, M. A. (2013). From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. *Current Protocols in Bioinformatics*, *43*(1). https://doi.org/10.1002/0471250953.bi1110s43
8. Ávila-Arcos, M. C., McManus, K. F., Sandoval, K., Rodríguez-Rodríguez, J. E., Villa-Islas, V., Martin, A. R., Luisi, P., Peñaloza-Espinosa, R. I., Eng, C., Huntsman, S., Burchard, E. G., Gignoux, C. R., Bustamante, C. D., & Moreno-Estrada, A. (2019). Population History and Gene Divergence in Native Mexicans Inferred from 76 Human Exomes. *Molecular Biology and Evolution*, *msz282*. https://doi.org/10.1093/molbev/msz282
9. Balam-Ortiz, E., Alfaro-Ruiz, L., Carrillo, K., Jimenez-Sanchez, G., Esquivel-Villarreal, A., Elizalde, A., Gil, T., Urushihara, M., & Kobori, H. (2011). Variants and Haplotypes in Angiotensinogen Gene Are Associated With Plasmatic Angiotensinogen Level in Mexican Population. *The American Journal of the Medical Sciences*, *342*(3), 205–211. https://doi.org/10.1097/MAJ.0b013e3182121020
10. Barroso, I., Gurnell, M., Crowley, V. E. F., Agostini, M., Schwabe, J. W., Lewis, H., Schafer, A. J., Chatterjee, V. K. K., & O'Rahilly, S. (1999). *Dominant negative mutations in human PPARg associated with severe insulin resistance, diabetes mellitus and hypertension*. *402*, 4.

11. Ben-Ari, G., & Lavi, U. (2012). Marker-assisted selection in plant breeding. In *Plant Biotechnology and Agriculture* (pp. 163–184). Elsevier. https://doi.org/10.1016/B978-0-12-381466-1.00011-0

12. Bergqvist, C., Kadara, H., Hamie, L., Nemer, G., Safi, R., Karouni, M., Marrouche, N., Abbas, O., Hasbani, D. J., Kibbi, A. G., Nassar, D., Shimomura, Y., & Kurban, M. (2018). SLURP-1 is mutated in Mal de Meleda, a potential molecular signature for melanoma and a putative squamous lineage tumor suppressor gene. *International Journal of Dermatology*, *57*(2), 162–170. https://doi.org/10.1111/ijd.13850

13. Bernstein, B. E., Stamatoyannopoulos, J. A., Costello, J. F., Ren, B., Milosavljevic, A., Meissner, A., Kellis, M., Marra, M. A., Beaudet, A. L., Ecker, J. R., Farnham, P. J., Hirst, M., Lander, E. S., Mikkelsen, T. S., & Thomson, J. A. (2010). The NIH Roadmap Epigenomics Mapping Consortium. *Nature Biotechnology*, *28*(10), 1045–1048. https://doi.org/10.1038/nbt1010-1045

14. Bersaglieri, T., Sabeti, P. C., Patterson, N., Vanderploeg, T., Schaffner, S. F., Drake, J. A., Rhodes, M., Reich, D. E., & Hirschhorn, J. N. (2004). Genetic Signatures of Strong Recent Positive Selection at the Lactase Gene. *The American Journal of Human Genetics*, *74*(6), 1111–1120. https://doi.org/10.1086/421051

15. Bierhals, T., Korenke, G. C., Uyanik, G., & Kutsche, K. (2013). Pontocerebellar hypoplasia type 2 and TSEN2: Review of the literature and two novel mutations. *European Journal of Medical Genetics*, *56*(6), 325–330. https://doi.org/10.1016/j.ejmg.2013.03.009

16. Bose, D. A., & Berger, S. L. (2017). ERNA binding produces tailored CBP activity profiles to regulate gene expression. *RNA Biology*, *14*(12), 1655–1659. https://doi.org/10.1080/15476286.2017.1353862

17. Brookes, A. J. (1999). The essence of SNPs. *Gene*, *234*(2), 177–186. https://doi.org/10.1016/S0378-1119(99)00219-X

18. Buniello, A., MacArthur, J. A. L., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., Suveges, D., Vrousgou, O., Whetzel, P. L., Amode, R., Guillen, J. A., Riat, H. S., Trevanion, S. J., Hall, P., Junkins, H., … Parkinson, H. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research*, *47*(D1), D1005–D1012. https://doi.org/10.1093/nar/gky1120

19. Chan, K. H. K., Niu, T., Ma, Y., You, N. Y., Song, Y., Sobel, E. M., Hsu, Y.-H., Balasubramanian, R., Qiao, Y., Tinker, L., & Liu, S. (2013). Common Genetic Variants in Peroxisome Proliferator-Activated Receptor-γ (*PPARG*) and Type 2 Diabetes Risk Among Women's Health Initiative Postmenopausal Women. *The Journal of Clinical Endocrinology & Metabolism*, *98*(3), E600–E604. https://doi.org/10.1210/jc.2012-3644

20. Chepelev, I., Wei, G., Wangsa, D., Tang, Q., & Zhao, K. (2012). Characterization of genome-wide enhancer-promoter interactions reveals co-expression of interacting genes and modes of higher order chromatin organization. *Cell Research*, *22*(3), 490–503. https://doi.org/10.1038/cr.2012.15

21. Corradi, A., Fadda, M., Piton, A., Patry, L., Marte, A., Rossi, P., Cadieux-Dion, M., Gauthier, J., Lapointe, L., Mottron, L., Valtorta, F., Rouleau, G. A., Fassio, A., Benfenati, F., & Cossette, P. (2014). SYN2 is an autism predisposing gene: Loss-of-function mutations alter synaptic vesicle cycling and axon outgrowth. *Human Molecular Genetics*, *23*(1), 90–103. https://doi.org/10.1093/hmg/ddt401

22. Crawford, D. C., & Nickerson, D. A. (2005). Definition and Clinical Importance of Haplotypes. *Annual Review of Medicine*, *56*(1), 303–320. https://doi.org/10.1146/annurev.med.56.082103.104540

23. Cruceanu, C., Alda, M., Grof, P., Rouleau, G. A., & Turecki, G. (2012). Synapsin II Is Involved in the Molecular Pathway of Lithium Treatment in Bipolar Disorder. *PLoS ONE*, *7*(2), e32680. https://doi.org/10.1371/journal.pone.0032680

24. Cruceanu, C., Kutsarova, E., Chen, E. S., Checknita, D. R., Nagy, C., Lopez, J. P., Alda, M., Rouleau, G. A., & Turecki, G. (2016). DNA hypomethylation of Synapsin II CpG islands associates with increased gene expression in bipolar disorder and major depression. *BMC Psychiatry*, *16*(1), 286. https://doi.org/10.1186/s12888-016-0989-0

25. Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., Durbin, R., & 1000 Genomes Project Analysis Group. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

26. De Santa, F., Barozzi, I., Mietton, F., Ghisletti, S., Polletti, S., Tusi, B. K., Muller, H., Ragoussis, J., Wei, C.-L., & Natoli, G. (2010). A Large Fraction of Extragenic RNA Pol II Transcription Sites Overlap Enhancers. *PLoS Biology*, *8*(5), e1000384. https://doi.org/10.1371/journal.pbio.1000384

27. Dimitrieva, S., & Bucher, P. (2013). UCNEbase—A database of ultraconserved non-coding elements and genomic regulatory blocks. *Nucleic Acids Research*, *41*(D1), D101–D109. https://doi.org/10.1093/nar/gks1092

28. Ding, M., Liu, Y., Liao, X., Zhan, H., Liu, Y., & Huang, W. (2018). Enhancer RNAs (eRNAs): New Insights into Gene Transcription and Disease Treatment. *Journal of Cancer*, *9*(13), 2334–2340. https://doi.org/10.7150/jca.25829

29. Edwards, S. L., Beesley, J., French, J. D., & Dunning, A. M. (2013). Beyond GWASs: Illuminating the Dark Road from Association to Function. *The American Journal of Human Genetics*, *93*(5), 779–797. https://doi.org/10.1016/j.ajhg.2013.10.012

30. Evans, R. M., Barish, G. D., & Wang, Y.-X. (2004). PPARs and the complex journey to obesity. *Nature Medicine*, *10*(4), 355–361. https://doi.org/10.1038/nm1025

31. Fishilevich, S., Nudel, R., Rappaport, N., Hadar, R., Plaschkes, I., Iny Stein, T., Rosen, N., Kohn, A., Twik, M., Safran, M., Lancet, D., & Cohen, D. (2017). GeneHancer: Genome-wide integration of enhancers and target genes in GeneCards. *Database*, *2017*. https://doi.org/10.1093/database/bax028

32. Forrester, J. V., Dick, A. D., McMenamin, P. G., Roberts, F., & Pearlman, E. (2016). Genetics. In *The Eye* (pp. 130-156.e1). Elsevier. https://doi.org/10.1016/B978-0-7020-5554-6.00003-4

33. French, J. D., Ghoussaini, M., Edwards, S. L., Meyer, K. B., Michailidou, K., Ahmed, S., Khan, S., Maranian, M. J., O'Reilly, M., Hillman, K. M., Betts, J. A., Carroll, T., Bailey, P. J., Dicks, E., Beesley, J., Tyrer, J., Maia, A.-T., Beck, A., Knoblauch, N. W., … Dunning, A. M. (2013). Functional Variants at the 11q13 Risk Locus for Breast Cancer Regulate Cyclin D1 Expression through Long-Range Enhancers. *The American Journal of Human Genetics*, *92*(4), 489–503. https://doi.org/10.1016/j.ajhg.2013.01.002

34. Fujimoto, A., Kimura, R., Ohashi, J., Omi, K., Yuliwulandari, R., Batubara, L., Mustofa, M. S., Samakkarn, U., Settheetham-Ishida, W., Ishida, T., Morishita, Y., Furusawa, T., Nakazawa, M., Ohtsuka, R., & Tokunaga, K. (2008). A scan for genetic determinants of human hair morphology: EDAR is associated with Asian hair thickness. *Human Molecular Genetics*, *17*(6), 835–843. https://doi.org/10.1093/hmg/ddm355

35. Gamboa-Melendez, M. A., Huerta-Chagoya, A., Moreno-Macias, H., Vazquez-Cardenas, P., Ordonez-Sanchez, M. L., Rodriguez-Guillen, R., Riba, L., Rodriguez-Torres, M., Guerra-Garcia, M. T., Guillen-Pineda, L. E., Choudhry, S., del Bosque-Plata, L., Canizales-Quinteros, S., Perez-Ortiz, G., Escobedo-Aguirre, F., Parra, A., Lerman-Garber, I., Aguilar-Salinas, C. A., & Tusie-Luna, M. T. (2012). Contribution of Common Genetic Variation to the Risk of Type 2 Diabetes in the Mexican Mestizo Population. *Diabetes*, *61*(12), 3314–3321. https://doi.org/10.2337/db11-0550

36. Gong, J., Qiu, C., Huang, D., Zhang, Y., Yu, S., & Zeng, C. (2018). Integrative functional analysis of super enhancer SNPs for coronary artery disease. *Journal of Human Genetics*, *63*(5), 627–638. https://doi.org/10.1038/s10038-018-0422-2

37. *GWAS Catalog*. (n.d.). Retrieved April 24, 2020, from https://www.ebi.ac.uk/gwas/

38. Hans Peter Müller, & Walter Schaffner. (1990). Transcriptional enhancers can act in trans. *Trends in Genetics*, *6*, 300–304. https://doi.org/10.1016/0168-9525(90)90236-Y

39. Harris, E. E., & Meyer, D. (2006). The molecular signature of selection underlying human adaptations. *American Journal of Physical Anthropology*, *131*(S43), 89–130. https://doi.org/10.1002/ajpa.20518

40. Haworth, A., Savage, H., & Lench, N. (2016). Diagnostic Genomics and Clinical Bioinformatics. In *Medical and Health Genomics* (pp. 37–50). Elsevier. https://doi.org/10.1016/B978-0-12-420196-5.00004-6

41. Hegele, R. A., Cao, H., Frankowski, C., Mathews, S. T., & Leff, T. (2002). PPARG F388L, a Transactivation-Deficient Mutant, in Familial Partial Lipodystrophy. *Diabetes*, *51*(12), 3586–3590. https://doi.org/10.2337/diabetes.51.12.3586

42. Henn, B. M., Cavalli-Sforza, L. L., & Feldman, M. W. (2012). The great human expansion. *Proceedings of the National Academy of Sciences*, *109*(44), 17758–17764. https://doi.org/10.1073/pnas.1212380109

43. Hernández-Ávila, M., Pablo Gutiérrez, J., & Reynoso-Noverón, N. (2013). Diabetes mellitus en México. El estado de la epidemia. *Salud Pública de México*, *55*(Supl.2), 129. https://doi.org/10.21149/spm.v55s2.5108

44. Hutti, J. E., Shen, R. R., Abbott, D. W., Zhou, A. Y., Sprott, K. M., Asara, J. M., Hahn, W. C., & Cantley, L. C. (2009). Phosphorylation of the Tumor Suppressor CYLD by the Breast Cancer Oncogene IKKε Promotes Cell Transformation. *Molecular Cell*, *34*(4), 461–472. https://doi.org/10.1016/j.molcel.2009.04.031

45. Instituto Nacional de Lenguas Indígenas (Mexico) (Ed.). (2009). *Catálogo de las lenguas indígenas nacionales: Variantes lingüísticas de México con sus autodenominaciones y referencias geoestadísticas* (1. ed). Instituto Nacional de Lenguas Indígenas.

46. International Human Genome Sequencing Consortium. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*(6822), 860–921. https://doi.org/10.1038/35057062

47. Jorde, L. B., & Wooding, S. P. (2004). Genetic variation, classification and "race." *Nature Genetics*, *36*(S11), S28–S33. https://doi.org/10.1038/ng1435

48. Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., Collins, R. L., Laricchia, K. M., Ganna, A., Birnbaum, D. P., Gauthier, L. D., Brand, H., Solomonson, M., Watts, N. A., Rhodes, D., Singer-Berk, M., Seaby, E. G., Kosmicki, J. A., Walters, R. K., … MacArthur, D. G. (2019). *Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes* [Preprint]. Genomics. https://doi.org/10.1101/531210

49. Karolchik, D. (2004). The UCSC Table Browser data retrieval tool. *Nucleic Acids Research*, *32*(90001), 493D – 496. https://doi.org/10.1093/nar/gkh103

50. Kelley, J. L. (2006). Genomic signatures of positive selection in humans and the limits of outlier approaches. *Genome Research*, *16*(8), 980–989. https://doi.org/10.1101/gr.5157306

51. Kelley, Joanna L., & Swanson, W. J. (2008). Positive Selection in the Human Genome: From Genome Scans to Biological Significance. *Annual Review of Genomics and Human Genetics*, *9*(1), 143–160. https://doi.org/10.1146/annurev.genom.9.081307.164411

52. Kilpeläinen, T. O., Lakka, T. A., Laaksonen, D. E., Lindström, J., Eriksson, J. G., Valle, T. T., Hämäläinen, H., Ilanne-Parikka, P., Keinänen-Kiukaanniemi, S., Lindi, V., Tuomilehto, J., Uusitupa, M., & Laakso, M. (2008). SNPs in PPARG Associate with Type 2 Diabetes and Interact with Physical Activity: *Medicine & Science in Sports & Exercise*, *40*(1), 25–33. https://doi.org/10.1249/mss.0b013e318159d1cd

53. Kim, T.-K., Hemberg, M., Gray, J. M., Costa, A. M., Bear, D. M., Wu, J., Harmin, D. A., Laptewicz, M., Barbara-Haley, K., Kuersten, S., Markenscoff-Papadimitriou, E., Kuhl, D., Bito, H., Worley, P. F., Kreiman, G., & Greenberg, M. E. (2010).

Widespread transcription at neuronal activity-regulated enhancers. *Nature*, *465*(7295), 182–187. https://doi.org/10.1038/nature09033

54. Laland, K. N., Odling-Smee, J., & Myles, S. (2010). How culture shaped the human genome: Bringing genetics and the human sciences together. *Nature Reviews Genetics*, *11*(2), 137–148. https://doi.org/10.1038/nrg2734

55. Landrum, M. J., Lee, J. M., Benson, M., Brown, G., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Hoover, J., Jang, W., Katz, K., Ovetsky, M., Riley, G., Sethi, A., Tully, R., Villamarin-Salomon, R., Rubinstein, W., & Maglott, D. R. (2016). ClinVar: Public archive of interpretations of clinically relevant variants. *Nucleic Acids Research*, *44*(D1), D862–D868. https://doi.org/10.1093/nar/gkv1222

56. Lappalainen, T., Salmela, E., Andersen, P. M., Dahlman-Wright, K., Sistonen, P., Savontaus, M.-L., Schreiber, S., Lahermo, P., & Kere, J. (2010). Genomic landscape of positive natural selection in Northern European populations. *European Journal of Human Genetics*, *18*(4), 471–478. https://doi.org/10.1038/ejhg.2009.184

57. Leonardi, M., Perna, E., Tronnolone, S., Colecchia, D., & Chiariello, M. (2019). Activated kinase screening identifies the IKBKE oncogene as a positive regulator of autophagy. *Autophagy*, *15*(2), 312–326. https://doi.org/10.1080/15548627.2018.1517855

58. León-Mimila, P., Villamil-Ramírez, H., Villalobos-Comparán, M., Villarreal-Molina, T., Romero-Hidalgo, S., López-Contreras, B., Gutiérrez-Vidal, R., Vega-Badillo, J., Jacobo-Albavera, L., Posadas-Romeros, C., Canizalez-Román, A., Río-Navarro, B. D., Campos-Pérez, F., Acuña-Alonzo, V., Aguilar-Salinas, C., & Canizales-Quinteros, S. (2013). Contribution of Common Genetic Variants to Obesity and Obesity-Related Traits in Mexican Children and Adults. *PLoS ONE*, *8*(8), e70640. https://doi.org/10.1371/journal.pone.0070640

59. Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., Foster, B., Moser, M., Karasik, E., Gillard, B., Ramsey, K., Sullivan, S., Bridge, J., Magazine, H., Syron, J., … Moore, H. F. (2013). The Genotype-Tissue Expression (GTEx) project. *Nature Genetics*, *45*(6), 580–585. https://doi.org/10.1038/ng.2653

60. Lu, W., Huang, S. Y., Su, L., Zhao, B. X., & Miao, J. Y. (2016). Long Noncoding RNA LOC100129973 Suppresses Apoptosis by Targeting miR-4707-5p and miR-4767 in Vascular Endothelial Cells. *Scientific Reports*, *6*(1), 21620. https://doi.org/10.1038/srep21620

61. MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., Junkins, H., McMahon, A., Milano, A., Morales, J., Pendlington, Z. M., Welter, D., Burdett, T., Hindorff, L., Flicek, P., Cunningham, F., & Parkinson, H. (2017). The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Research*, *45*(D1), D896–D901. https://doi.org/10.1093/nar/gkw1133

62. Mallick, S., Li, H., Lipson, M., Mathieson, I., Gymrek, M., Racimo, F., Zhao, M., Chennagiri, N., Nordenfelt, S., Tandon, A., Skoglund, P., Lazaridis, I., Sankararaman, S., Fu, Q., Rohland, N., Renaud, G., Erlich, Y., Willems, T., Gallo, C., … Reich, D. (2016). The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature*, *538*(7624), 201–206. https://doi.org/10.1038/nature18964

63. Marsman, J., & Horsfield, J. A. (2012). Long distance relationships: Enhancer–promoter communication and dynamic gene transcription. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, *1819*(11–12), 1217–1227. https://doi.org/10.1016/j.bbagrm.2012.10.008

64. Melendez-Zajgla, J., Pozo, L., Ceballos, G., & Maldonado, V. (2008). Tissue Inhibitor of Metalloproteinases-4. The road less traveled. *Molecular Cancer*, *7*(1), 85. https://doi.org/10.1186/1476-4598-7-85

65. Mifsud, B., Tavares-Cadete, F., Young, A. N., Sugar, R., Schoenfelder, S., Ferreira, L., Wingett, S. W., Andrews, S., Grey, W., Ewels, P. A., Herman, B., Happe, S., Higgs, A., LeProust, E., Follows, G. A., Fraser, P., Luscombe, N. M., & Osborne, C. S. (2015). Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nature Genetics*, *47*(6), 598–606. https://doi.org/10.1038/ng.3286

66. Moreno-Estrada, A., Gignoux, C. R., Fernandez-Lopez, J. C., Zakharia, F., Sikora, M., Contreras, A. V., Acuna-Alonzo, V., Sandoval, K., Eng, C., Romero-Hidalgo, S., Ortiz-Tello, P., Robles, V., Kenny, E. E., Nuno-Arana, I., Barquera-Lozano, R., Macin-Perez, G., Granados-Arriola, J., Huntsman, S., Galanter, J. M., … Bustamante, C. D. (2014). The genetics of Mexico recapitulates Native American substructure and affects biomedical traits. *Science*, *344*(6189), 1280–1285. https://doi.org/10.1126/science.1251688

67. Müller, H.-P., Sogo, J., & Schaffner, W. (1989). An enhancer stimulates transcription in Trans when attached to the promoter via a protein bridge. *Cell*, *4*(58), 767–777. https://doi.org/10.1016/0092-8674(89)90110-4

68. Nica, A. C., & Dermitzakis, E. T. (2013). Expression quantitative trait loci: Present and future. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *368*(1620), 20120362. https://doi.org/10.1098/rstb.2012.0362

69. O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva,

O., … Pruitt, K. D. (2016). Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, *44*(D1), D733–D745. https://doi.org/10.1093/nar/gkv1189

70. Pasquali, L., Gaulton, K. J., Rodríguez-Seguí, S. A., Mularoni, L., Miguel-Escalada, I., Akerman, İ., Tena, J. J., Morán, I., Gómez-Marín, C., van de Bunt, M., Ponsa-Cobas, J., Castro, N., Nammo, T., Cebola, I., García-Hurtado, J., Maestro, M. A., Pattou, F., Piemonti, L., Berney, T., … Ferrer, J. (2014). Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants. *Nature Genetics*, *46*(2), 136–143. https://doi.org/10.1038/ng.2870

71. Pérez, J. C. R., Esparragón, F. R., & Mogollón, J. N. (2007). *PPAR-gamma, hipertensión arterial y riñón*. *27*(1), 6.

72. Pérez O'Brien, A. M., Utsunomiya, Y. T., Mészáros, G., Bickhart, D. M., Liu, G. E., Van Tassell, C. P., Sonstegard, T. S., Da Silva, M. V., Garcia, J., & Sölkner, J. (2014). Assessing signatures of selection through variation in linkage disequilibrium between taurine and indicine cattle. *Genetics Selection Evolution*, *46*(1), 19. https://doi.org/10.1186/1297-9686-46-19

73. Petretto, E., Mangion, J., Dickens, N. J., Cook, S. A., Kumaran, M. K., Lu, H., Fischer, J., Maatz, H., Kren, V., Pravenec, M., Hubner, N., & Aitman, T. J. (2006). Heritability and Tissue Specificity of Expression Quantitative Trait Loci. *PLoS Genetics*, *2*(10), e172. https://doi.org/10.1371/journal.pgen.0020172

74. Pickrell, J. K., Coop, G., Novembre, J., Kudaravalli, S., Li, J. Z., Absher, D., Srinivasan, B. S., Barsh, G. S., Myers, R. M., Feldman, M. W., & Pritchard, J. K. (2009). Signals of recent positive selection in a worldwide sample of human populations. *Genome Research*, *19*(5), 826–837. https://doi.org/10.1101/gr.087577.108

75. Pokrywczynska, M., Lanzoni, G., & Ricordi, C. (2019). From Adult Pancreatic Islets to Stem Cells. In *Principles of Regenerative Medicine* (pp. 335–349). Elsevier. https://doi.org/10.1016/B978-0-12-809880-6.00021-7

76. Popejoy, A. B., & Fullerton, S. M. (2016). Genomics is failing on diversity. *Nature*, *538*(7624), 161–164. https://doi.org/10.1038/538161a

77. Pulakanti, K., Pinello, L., Stelloh, C., Blinka, S., Allred, J., Milanovich, S., Kiblawi, S., Peterson, J., Wang, A., Yuan, G.-C., & Rao, S. (2013). Enhancer transcribed RNAs arise from hypomethylated, Tet-occupied genomic regions. *Epigenetics*, *8*(12), 1303–1320. https://doi.org/10.4161/epi.26597

78. Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*(6), 841–842. https://doi.org/10.1093/bioinformatics/btq033

79. Reimand, J., Isserlin, R., Voisin, V., Kucera, M., Tannus-Lopes, C., Rostamianfar, A., Wadi, L., Meyer, M., Wong, J., Xu, C., Merico, D., & Bader, G. D. (2019). Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap. *Nature Protocols*, *14*(2), 482–517. https://doi.org/10.1038/s41596-018-0103-9

80. Rightmire, G. P. (2009). Middle and later Pleistocene hominins in Africa and Southwest Asia. *Proceedings of the National Academy of Sciences*, *106*(38), 16046–16050. https://doi.org/10.1073/pnas.0903930106

81. Rojano, E., Seoane, P., Ranea, J. A. G., & Perkins, J. R. (2018). Regulatory variants: From detection to predicting impact. *Briefings in Bioinformatics*. https://doi.org/10.1093/bib/bby039

82. Romero-Hidalgo, S., Ochoa-Leyva, A., Garcíarrubio, A., Acuña-Alonzo, V., Antúnez-Argüelles, E., Balcazar-Quintero, M., Barquera-Lozano, R., Carnevale, A.,

Cornejo-Granados, F., Fernández-López, J. C., García-Herrera, R., García-Ortíz, H., Granados-Silvestre, Á., Granados, J., Guerrero-Romero, F., Hernández-Lemus, E., León-Mimila, P., Macín-Pérez, G., Martínez-Hernández, A., … Soberón, X. (2017). Demographic history and biologically relevant genetic variation of Native Mexicans inferred from whole-genome sequencing. *Nature Communications*, *8*(1), 1005. https://doi.org/10.1038/s41467-017-01194-z

83. Sasaki-Iwaoka, H., Maruyama, K., Endoh, H., Komori, T., Kato, S., & Kawashima, H. (1999). A Trans-Acting Enhancer Modulates Estrogen-Mediated Transcription of Reporter Genes in Osteoblasts. *Journal of Bone and Mineral Research*, *14*(2), 248–255. https://doi.org/10.1359/jbmr.1999.14.2.248
84. Saviouk, V., Moreau, M., Tereshchenko, I., & Brzustowicz, L. (2007). Association of synapsin 2 with schizophrenia in families of Northern European ancestry. *Schizophrenia Research*, *96*(1–3), 100–111. https://doi.org/10.1016/j.schres.2007.07.031
85. Schaukowitch, K., Joo, J.-Y., Liu, X., Watts, J. K., Martinez, C., & Kim, T.-K. (2014). Enhancer RNA Facilitates NELF Release from Immediate Early Genes. *Molecular Cell*, *56*(1), 29–42. https://doi.org/10.1016/j.molcel.2014.08.023
86. Scheinfeldt, L. B., & Tishkoff, S. A. (2013). Recent human adaptation: Genomic approaches, interpretation and insights. *Nature Reviews Genetics*, *14*(10), 692–702. https://doi.org/10.1038/nrg3604
87. Schneider, V. A., Graves-Lindsay, T., Howe, K., Bouk, N., Chen, H.-C., Kitts, P. A., Murphy, T. D., Pruitt, K. D., Thibaud-Nissen, F., Albracht, D., Fulton, R. S., Kremitzki, M., Magrini, V., Markovic, C., McGrath, S., Steinberg, K. M., Auger, K., Chow, W., Collins, J., … Church, D. M. (2016). *Evaluation of GRCh38 and de*

novo haploid genome assemblies demonstrates the enduring quality of the reference assembly [Preprint]. Genomics. https://doi.org/10.1101/072116

88. Semple, R. K. (2006). PPAR and human metabolic disease. *Journal of Clinical Investigation*, *116*(3), 581–589. https://doi.org/10.1172/JCI28003

89. Shastry, B. S. (2002). SNP alleles in human disease and evolution. *Journal of Human Genetics*, *47*(11), 0561–0566. https://doi.org/10.1007/s100380200086

90. Sherry, S. T., Ward, M., & Sirotkin, K. (1999). *DbSNP—Database for Single Nucleotide Polymorphisms and Other Classes of Minor Genetic Variation*. 4.

91. Shi, J., Fu, L., & Wang, W. (2015). High expression of inositol 1,4,5-trisphosphate receptor, type 2 (ITPR2) as a novel biomarker for worse prognosis in cytogenetically normal acute myeloid leukemia. *Oncotarget*, *6*(7). https://doi.org/10.18632/oncotarget.3024

92. Stelzer, G., Rosen, N., Plaschkes, I., Zimmerman, S., Twik, M., Fishilevich, S., Stein, T. I., Nudel, R., Lieder, I., Mazor, Y., Kaplan, S., Dahary, D., Warshawsky, D., Guan-Golan, Y., Kohn, A., Rappaport, N., Safran, M., & Lancet, D. (2016). The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses. *Current Protocols in Bioinformatics*, *54*(1). https://doi.org/10.1002/cpbi.5

93. Stephenson, F. H. (2016). Forensics and Paternity. In *Calculations for Molecular Biology and Biotechnology* (pp. 439–463). Elsevier. https://doi.org/10.1016/B978-0-12-802211-5.00013-8

94. Stryjecki, C., Peralta-Romero, J., Alyass, A., Karam-Araujo, R., Suarez, F., Gomez-Zamudio, J., Burguete-Garcia, A., Cruz, M., & Meyre, D. (2016). Association between PPAR-γ2 Pro12Ala genotype and insulin resistance is modified by circulating lipids in Mexican children. *Scientific Reports*, *6*(1), 24472. https://doi.org/10.1038/srep24472

95. Tattersall, I. (2009). Human origins: Out of Africa. *Proceedings of the National Academy of Sciences, 106*(38), 16018–16021. https://doi.org/10.1073/pnas.0903207106

96. The 1000 Genomes Project Consortium. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74. https://doi.org/10.1038/nature15393

97. The ENCODE Project Consortium. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, *489*(7414), 57–74. https://doi.org/10.1038/nature11247

98. The FANTOM Consortium and the RIKEN PMI and CLST (DGT). (2014). A promoter-level mammalian expression atlas. *Nature*, *507*(7493), 462–470. https://doi.org/10.1038/nature13182

99. The FANTOM Consortium, Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T., Ntini, E., Arner, E., Valen, E., Li, K., Schwarzfischer, L., Glatz, D., Raithel, J., Lilje, B., … Sandelin, A. (2014). An atlas of active enhancers across human cell types and tissues. *Nature*, *507*(7493), 455–461. https://doi.org/10.1038/nature12787

100. The International HapMap Consortium. (2005). A haplotype map of the human genome. *Nature*, *437*(7063), 1299–1320. https://doi.org/10.1038/nature04226

101. The International HapMap Consortium, Sabeti, P. C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., Xie, X., Byrne, E. H., McCarroll, S. A., Gaudet, R., Schaffner, S. F., & Lander, E. S. (2007). Genome-wide detection and characterization of positive selection in human populations. *Nature*, *449*(7164), 913–918. https://doi.org/10.1038/nature06250

102. The SIGMA Type 2 Diabetes Consortium, Williams, A. L., Jacobs, S. B. R., Moreno-Macías, H., Huerta-Chagoya, A., Churchhouse, C., Márquez-Luna, C.,

García-Ortíz, H., José Gómez-Vázquez, M., Burtt, N. P., Aguilar-Salinas, C. A., González-Villalpando, C., Florez, J. C., Orozco, L., Haiman, C. A., Tusié-Luna, T., Altshuler, D., Williams, A. L., Márquez-Luna, C., … Tusié-Luna, T. (2013). Sequence variants in SLC16A11 are a common risk factor for type 2 diabetes in Mexico. *Nature*, *506*, 97.

103. Visel, A., Minovitsky, S., Dubchak, I., & Pennacchio, L. A. (2007). VISTA Enhancer Browser—A database of tissue-specific human enhancers. *Nucleic Acids Research*, *35*(Database), D88–D92. https://doi.org/10.1093/nar/gkl822
104. Voight, B. F., Kudaravalli, S., Wen, X., & Pritchard, J. K. (2006). A Map of Recent Positive Selection in the Human Genome. *PLoS Biology*, *4*(3), e72. https://doi.org/10.1371/journal.pbio.0040072
105. Wang, D., Cheng, Z., Zhao, X., Li, Z., Wang, J., Dong, A., Zhou, Z., & Zhang, F. (2016). Association between MKL1 rs6001946 and schizophrenia in a Han Chinese population. *Neuroscience Letters*, *631*, 36–39. https://doi.org/10.1016/j.neulet.2016.08.008
106. Wang, D., Garcia-Bassets, I., Benner, C., Li, W., Su, X., Zhou, Y., Qiu, J., Liu, W., Kaikkonen, M. U., Ohgi, K. A., Glass, C. K., Rosenfeld, M. G., & Fu, X.-D. (2011). Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA. *Nature*, *474*(7351), 390–394. https://doi.org/10.1038/nature10006
107. Whirl-Carrillo, M., McDonagh, E. M., Hebert, J. M., Gong, L., Sangkuhl, K., Thorn, C. F., Altman, R. B., & Klein, T. E. (2012). Pharmacogenomics Knowledge for Personalized Medicine. *Clinical Pharmacology & Therapeutics*, *92*(4), 414–417. https://doi.org/10.1038/clpt.2012.96
108. Wiel, C., Lallet-Daher, H., Gitenay, D., Gras, B., Le Calvé, B., Augert, A., Ferrand, M., Prevarskaya, N., Simonnet, H., Vindrieux, D., & Bernard, D. (2014).

Endoplasmic reticulum calcium release through ITPR2 channels leads to mitochondrial calcium accumulation and senescence. *Nature Communications*, *5*(1), 3792. https://doi.org/10.1038/ncomms4792

109. Williamson, S. H., Hubisz, M. J., Clark, A. G., Payseur, B. A., Bustamante, C. D., & Nielsen, R. (2007). Localizing Recent Adaptive Evolution in the Human Genome. *PLoS Genetics*, *3*(6), 15.

110. Wong, N. D. (2014). Epidemiological studies of CHD and the evolution of preventive cardiology. *Nature Reviews Cardiology*, *11*(5), 276–289. https://doi.org/10.1038/nrcardio.2014.26

111. Wu, Q., Li, Y., Gu, S., Li, N., Zheng, D., Li, D., Zheng, Z., Ji, C., Xie, Y., & Mao, Y. (2004). Molecular cloning and characterization of a novel dual-specificity phosphatase 23 gene from human fetal brain. *The International Journal of Biochemistry & Cell Biology*, *36*(8), 1542–1553. https://doi.org/10.1016/j.biocel.2003.12.014

112. Yi, X., Liang, Y., Huerta-Sanchez, E., Jin, X., Cuo, Z. X. P., Pool, J. E., Xu, X., Jiang, H., Vinckenbosch, N., Korneliussen, T. S., Zheng, H., Liu, T., He, W., Li, K., Luo, R., Nie, X., Wu, H., Zhao, M., Cao, H., … Wang, J. (2010). Sequencing of 50 Human Exomes Reveals Adaptation to High Altitude. *Science*, *329*(5987), 75–78. https://doi.org/10.1126/science.1190371

113. Zhang, G., Shi, J., Zhu, S., Lan, Y., Xu, L., Yuan, H., Liao, G., Liu, X., Zhang, Y., Xiao, Y., & Li, X. (2018). DiseaseEnhancer: A resource of human disease-associated enhancer catalog. *Nucleic Acids Research*, *46*(D1), D78–D84. https://doi.org/10.1093/nar/gkx920

114. Zhao, Y., Wang, L., Ren, S., Wang, L., Blackburn, P. R., McNulty, M. S., Gao, X., Qiao, M., Vessella, R. L., Kohli, M., Zhang, J., Karnes, R. J., Tindall, D. J., Kim,

Y., MacLeod, R., Ekker, S. C., Kang, T., Sun, Y., & Huang, H. (2016). Activation of P-TEFb by Androgen Receptor-Regulated Enhancer RNAs in Castration-Resistant Prostate Cancer. *Cell Reports*, *15*(3), 599–610. https://doi.org/10.1016/j.celrep.2016.03.038

115. Zhou, X., Li, D., Zhang, B., Lowdon, R. F., Rockweiler, N. B., Sears, R. L., Madden, P. A. F., Smirnov, I., Costello, J. F., & Wang, T. (2015). Epigenomic annotation of genetic variants using the Roadmap Epigenome Browser. *Nature Biotechnology*, *33*(4), 345–346. https://doi.org/10.1038/nbt.3158

12. Material Suplementario

El material suplementario de este proyecto se describe a continuación y también está disponible a través de las siguientes ligas:

- Material Suplementario 1: https://docs.google.com/document/d/1aRc1nr9xke2t8Lz5UwQYcvzotm8HiLceqSRANO4IJCQ/edit?usp=sharing
- Material Suplementario 2: https://docs.google.com/spreadsheets/d/1huRz187zjA10BQt9TjLkoQ4uFVs_zMQLcZ824SzKb9U/edit?usp=sharing

12.1 Formato de Llamado de Variante (VCF)

El formato de llamado de variante (del inglés Variant Call Format; VCF) (Danecek et al., 2011) es un formato de texto estandarizado que guarda a los principales tipos de variación genética, incluyendo a los SNPs, junto con "anotaciones" o información acerca de ellas (por ejemplo, información sobre el genotipo o de las muestras).

De manera general, un archivo VCF contiene una sección de encabezado y una sección de datos. La sección de encabezado contiene líneas de metainformación, las cuales describen las etiquetas y anotaciones usadas en la sección de datos. También informa sobre el software utilizado para la creación del archivo, la versión de la secuencia o genoma de referencia utilizada y la fecha de creación.

La sección de datos contiene 8 principales columnas:

1. CHROM: Cromosoma donde se encuentra la variante
2. POS: Posición del inicio de la variante
3. ID: Identificador de la variante (comúnmente un rsID)
4. REF: Alelo de referencia (i.e. el alelo encontrado en el genoma de referencia)
5. ALT: El alelo (o alelos) alterno presentes en las muestras, en esa posición.
6. QUAL: Un valor que indica la calidad del llamado de la variante.
7. FILTER: Indica si la variante pasó o no los filtros aplicados a los datos.
8. INFO: Información adicional sobre la variante.

12.2 Genome Analysis Toolkit (GATK) Best Practices

El descubrimiento de nuevas variantes genéticas es un proceso estandarizado al que se le refiere como *llamado de variantes*. La herramienta Genome Analysis Toolkit (GATK) Best Practices (Auwera et al., 2013), desarrollada en el Broad Institute, es el programa más utilizado genómica humana para el llamado de variantes, a partir de datos de secuenciación de DNA (figura 12). De manera resumida, las operaciones que realiza GATK están divididas en tres fases:

1. **Preprocesamiento:** Esta fase inicia con los datos crudos provenientes de la secuenciación, generalmente en formato FASTQ (este formato contiene la secuencia generada más un puntaje asignado del llamado de bases), y produce archivos en formato BAM (un formato binario para guardar datos de secuenciación) que seguirán el flujo del análisis. Esta fase incluye el alineamiento de los datos con un genoma de referencia, así como otras operaciones de limpieza para corregir posibles sesgos técnicos en la secuenciación con el propósito de hacer aptos a los datos para análisis posteriores.
2. **Llamado de variantes:** Inicia con archivos BAM (formato binario para guardar datos de secuenciación) pre-procesados y produce un archivo que contiene un conjunto de llamados de vatiantes en formato VCF. En esta fase se identifican los sitios donde las muestras exhiben posible variación genética. Y se aplican métodos de filtrado apropiados al diseño experimental.
3. **Refinamiento del set de llamados:** Inicia y termina con un set de llamados en formato VCF. En esta fase se utilizan metadatos para evaluar y mejorar la precisión del proceso de genotipificación, se agrega información adicional y se evalúa la calidad del set de llamados.

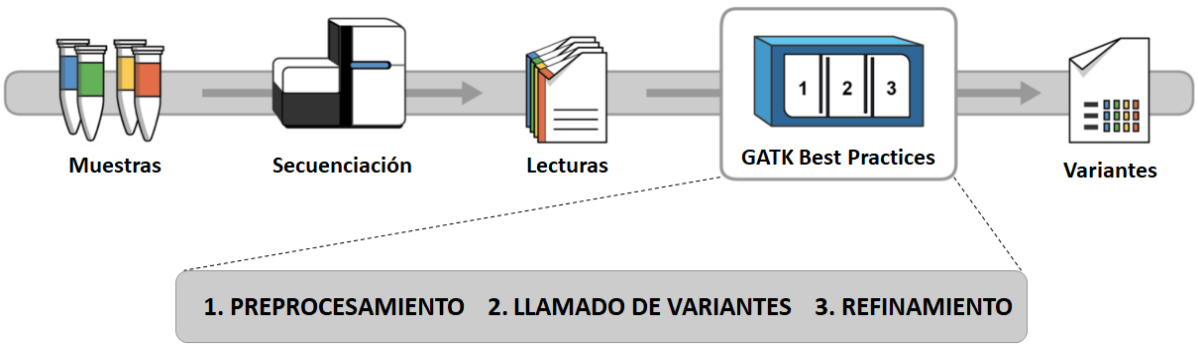


Figura 1. Flujo de trabajo para el llamado de variantes con GATK. El análisis se divide en tres fases principales, y culmina con uno o varios sets de llamado de variantes en formato VCF. Esta figura fue obtenida de la documentación de GATK disponible en *https://gatk.broadinstitute.org/* y ha sido traducida al español

12.3 Base de datos de GeneHancer

Esta sección corresponde a lo realizado en el proyecto GeneHancer. Está incluido para mejorar el entendimiento técnico de este trabajo de tesis, complementando a la información contenida en la sección 5.3.

12.3.1 Extracción de Enhancers

Los elementos enhancer fueron extraídos de las siguientes fuentes de datos:

1. Flancos de enhancers y promotores del ENSEMBL Regulatory Build versión 82, el cual se basa en sets de datos del proyecto ENCODE y el Roadmap Epigenomics.

2. El set de datos 'Enhancers Permisivos' de FANTOM5 del Transcribed Enhancer Atlas.

3. Enhancers humanos del VISTA Enhancer Browser. Este set de datos incluye elementos que presentan patrones de expresión a lo largo de varios tejidos, aumentando la certeza de estos.

4. Regiones enhancer próximas y distantes de ENCODE. Estos datos fueron proveídos a ENCODE por el Zhiping Weng Lab de la Universidad de Massachussets. La predicción de enhancers en este set de datos se basó en la identificación de regiones hipersensibles a la DNAsa y señales de acetilación de la Histona H3K27. http://zlab-annotations.umassmed.edu/enhancers/methods

La cantidad de elementos enhancer, resultado de la minería de las bases de datos antes mencionadas, fue de 284,834; estos elementos definirían al grupo de *enhancers candidatos* (Figura 13).

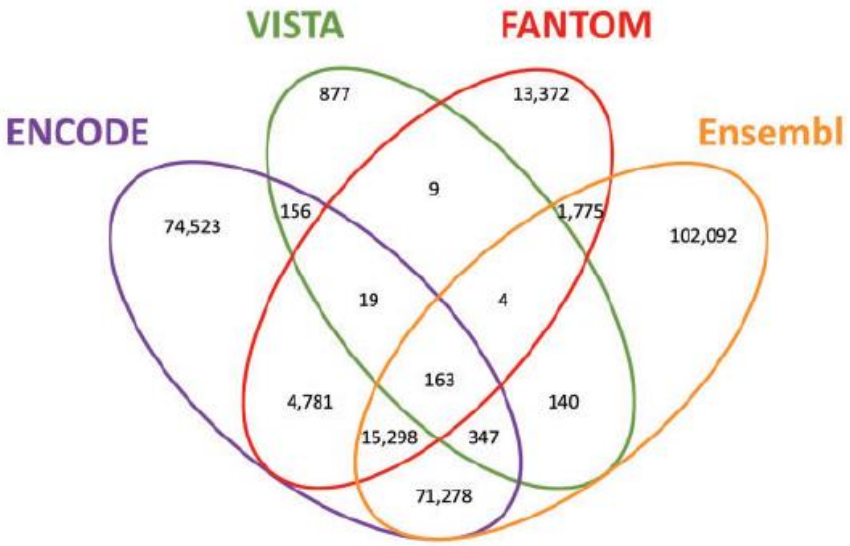


Figura 2. Enhancers candidatos de GeneHancer. Diagrama de Venn de los 284,834 enhancers candidatos, delimitados por la fuente de donde fueron obtenidos. De estos, 93,970 han sido reportados por más de una fuente. Esta figura pertenece al trabajo de *(Fishilevich et al., 2017)*.

12.3.2 Integración de los elementos en GeneHancer

Para el proceso de integración, todos los elementos enhancer recuperados de las fuentes antes mencionadas fueron sometidas a un análisis de sobrelape de regiones utilizando BEDtools (Quinlan & Hall, 2010) (un software para manejo de archivos en formato *bed*) para definir el grupo de enhancers candidatos, generando agrupamientos de elementos enhancer que compartieran posiciones similares en el genoma. Por cada uno de estos agrupamientos, se buscó generar un solo elemento enhancer candidato que representara la extensión de todos los enhancers en el grupo. Definiendo la longitud de cada enhancer candidato por la posición más pequeña de inicio y más alta del final de los enhancers en el grupo (Figura 14).

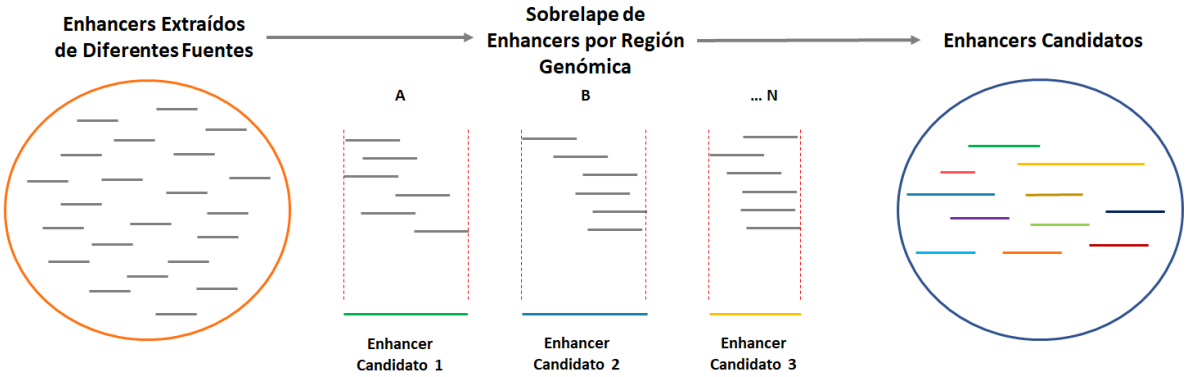


Figura 3. Integración de enhancers candidatos. Los elementos enhancers extraídos de distintas fuentes fueron sobrelapados y agrupados por región genómica (grupos A, B …N) para formar un enhancer candidato, definiendo su longitud por la posición

más pequeña de inicio y la más alta del final (líneas punteadas) de los enhancers en el grupo. Integrando así al grupo de enhancers candidatos.

Adicionalmente, se probó que los enhancers agrupados coincidieran con regiones de la base de datos de elementos no codificantes ultra conservados (UCNE) (Dimitrieva & Bucher, 2013).

12.3.3 Determinación del Valor de Confianza de los Enhancers

Después de obtener un set definido de candidatos enhancer, se asignó a cada elemento un valor de confianza, una medida de la certeza que se tiene de que realmente se trata de un elemento regulador tipo enhancer. Este valor se calcula basándose en la combinación de los siguientes componentes:

i. El número de fuentes que apoyan al enhancer dividido por 4, que es el número máximo de fuentes: $S_{fuentes}$

ii. El sobrelape con regiones conservadas del UCNE: $S_{conservado}$. Con valor de 1 para enhancers que sobrelapan con una región conservada, de otra forma este valor es 0.

iii. El valor de sitios de unión a factores de transcripción (TFBS): S_{TFBS}. Para cada enhancer, es el número de factores de transcripción únicos que tienen un sitio de unión dentro del enhancer. Los TFBS utilizados para este cálculo fueron obtenidos de los sets de datos de ChIP-Seq de ENCODE.

Para los candidatos a enhancer que tenían elementos provenientes del proyecto VISTA, se les dio un puntaje adicional: S_{VISTA}. Un valor constante de 0.25 asignado a aquellos que tenían reportados patrones de expresión consistentes a lo largo de distintos tejidos.
Por último, a los enhancers candidatos que tenían información de RNAs de enhancer (eRNAs) (ver Caja 4) de FANTOM se les añadió un valor S_{FANTOM} basado en el tamaño del cluster de la etiqueta CAGE de cada elemento de FANTOM (The FANTOM Consortium et al., 2014).

Por lo tanto, el valor de confianza de los elementos GeneHancer: S_E, se definió como:

$$S_E = S_{fuentes} + S_{conservado} + S_{TFBS} + S_{VISTA} + S_{FANTOM}$$

12.3.4 Asociaciones Enhancer - Gen

Una vez obtenido el grupo de enhancers candidatos con su respectivo valor de confianza, GeneHancer buscó encontrar a los posibles genes blanco de cada elemento enhancer. Para inferir una relación entre enhancers candidatos y genes se utilizaron 5 diferentes métodos:

1. **Loci de expresión de caracteres cuantitativos (eQTLs*).***

Esta validación consistió en identificar SNPs dentro de enhancers candidatos y después buscar correlaciones entre estos y la expresión de un potencial gen blanco al que regula el elemento enhancer, como se documentó en la base de datos GTEX (Lonsdale et al., 2013).

La base de esta validación se estableció aplicando un análisis similar de eQTL (Ben-Ari & Lavi, 2012; Nica & Dermitzakis, 2013; Petretto et al., 2006) a elementos promotor que regulaban únicamente a un gen correspondiente. Se encontró que en el 33% de los casos, el SNP que se encontraba en un promotor tenía la mejor señal de conexión eQTL hacia la expresión de su gen correspondiente, dando un estimado de verdaderos positivos en este análisis.

Adicionalmente, encontraron que los enhancers tenían una densidad de eQTL de 0.96 por kb mientras que otras regiones genómicas tuvieron una densidad mucho menor (0.60 por kb). Y que a pesar de que los enhancers abarcaban 0.184 de todos los eQTL, también sólo contenían 0.147 de todos los SNPs no asociados a la expresión génica; demostrando que los SNPs con conexiones a la expresión de un gen son más propensos a encontrarse en un elemento enhancer que los SNPs sin conexión a la expresión de un gen.

2. Captura Hi-C (Chi-C)

Un Chi-C es un ensayo genómico capaz de mapear interacciones reguladoras a escala genómica (Mifsud et al., 2015). Se utilizó un set de 22,000 promotores para este análisis, buscando fragmentos que interaccionen con los promotores que sobrelapan el set de enhancers candidatos. Se identificaron interacciones de largo alcance entre

92,621 enhancers candidatos y 19,461 promotores, resultando en 2.29 +- 1.87 genes por elemento enhancer.

3. **Co-expresión de eRNA**

   Este análisis se realizó con datos del Atlas de Enhancers Humanos generados por el proyecto FANTOM5 (The FANTOM Consortium et al., 2014). Se registraron casos de co-expresión entre eRNAs (Bose & Berger, 2017; De Santa et al., 2010; Ding et al., 2018; Kim et al., 2010; Pulakanti et al., 2013; Schaukowitch et al., 2014; Wang et al., 2011; Zhao et al., 2016) transcritos de regiones candidatas a enhancer y la expresión de mRNA de potenciales genes blanco. Esto resultó en la identificación de asociaciones entre 21,957 enhancers candidatos y 11,527 genes, siendo 2.17+-1.88 genes por enhancer candidato.

4. **Co-expresión de factores de transcripción (TF)**

   Este método se basó en analizar TFs que tienen sitios de unión, determinados por análisis de ChIP-seq, dentro de un enhancer candidato. Se buscaron correlaciones estadísticamente significativas entre la expresión de TFs y la expresión de potenciales genes blanco para los enhancers candidatos. Este método validó asociaciones entre 24,519 enhancers candidatos y 10,40 genes, en promedio 4.71 +- 5.33 genes por enhancer candidato.

5. **Distancia Gen - Enhancer.**

   La interacción gen-enhancer tiene a ocurrir en una distancia genómica considerable entre ellos (Marsman & Horsfield, 2012), sin embargo se cree que la probabilidad de que exista un evento de regulación disminuye cuando se trata de distancias muy grandes entre un gen y su enhancer asociado. Para apoyar esta aseveración, Fishilevich y colaboradores (Fishilevich et al., 2017) realizaron un análisis del comportamiento de todos los pares de gen-enhancer candidato (obtenidos por los métodos antes mencionados) en función de la distancia entre ellos: Añadiendo una medida basada en la distancia para definir un par gen-enhancer candidato, reflejando la dependencia que existe entre la distancia de estos dos elementos.

   En base a la convencionalidad de concentrarse en elementos inmediatamente adyacentes establecida por el proyecto VISTA (Visel et al., 2007), se tomó a los 2 genes

flanqueantes más próximos al elemento enhancer en una distancia no mayor a 1Mb. Esta adición generó 542,896 nuevas conexiones gen-enhancer.

Como resultado de la búsqueda de asociaciones por los 5 métodos antes descritos, fueron establecidos 1, 019, 746 pares de asociación enhancer-gen (Figura 15) a través de 284, 821 enhancers candidatos y 101, 337 genes. Encontrando también 3.58 +- 2.91 genes por elemento enhancer y 10.06 +- 11.8 elementos enhancer por gen.

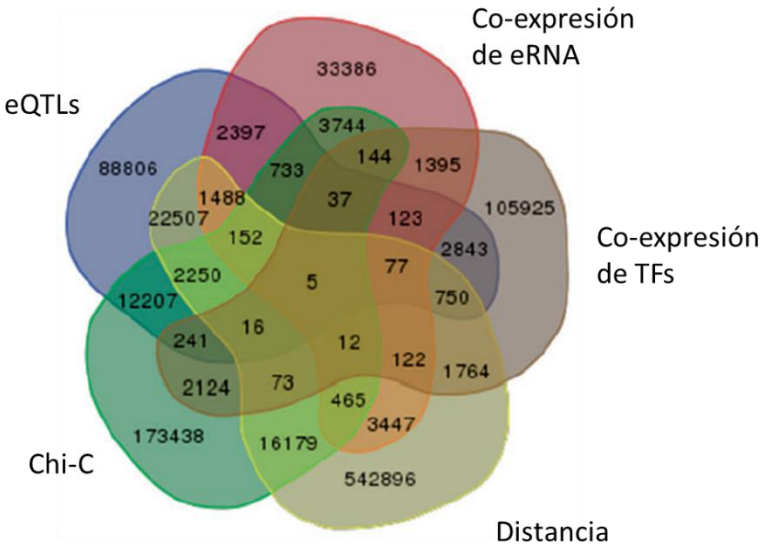


Figura 4. Asociaciones enhancer-gen de GeneHancer. Diagrama de Venn de los 1, 019, 746 pares enhancer-gen, agrupados por los 5 distintos métodos de asociación. Esta figura pertenece al trabajo de Fishilevich y colaboradores (Fishilevich et al., 2017) y ha sido traducida al español.

12.4 Resultados

12.4.3 Variación Genética en Regiones Enhancer de 76 Genomas de población Nativa Mexicana

La tabla S1 muestra, para cada variante, su rsID, el identificador del elemento enhancer y el gen regulado de acuerdo con GeneHancer, el tipo de elemento (Enhancer o Promotor/Enhancer), y la frecuencia alélica de la variante en la población de 100 G-MX y en las otras 4 poblaciones continentales consideradas.

Tabla S1. Variantes localizadas en elementos GeneHancer

rsID	GeneHancer ID	Tipo	Gen Regulado	100-G-MX	Latina	Asiática del Este	Europea no Finlandesa	Africana
rs1037143	GH01J159596	Enhancer	OR10J6P	0.558	0.54717	0.134961	0.587666	0.485274
rs10779631	GH01J206448	Enhancer	IKBKE	0.5	0.461	0.26	0.5252	0.2814
rs10842776	GH12J026690	Enhancer	ITPR2	0.846	0.429078	0.56491	0.144981	0.10432
rs10842777	GH12J026690	Enhancer	ITPR2	0.846	0.543839	0.574839	0.314798	0.487777
rs11048413	GH12J026119	Promoter /Enhancer	SSPN	0.853	0.561828	0.516513	0.487835	0.147724
rs11048413	GH12J026119	Promoter /Enhancer	ENSG00000256894	0.853	0.561828	0.516513	0.487835	0.147724
rs11048654	GH12J026690	Enhancer	ITPR2	0.846	0.429245	0.567742	0.144998	0.10108
rs11122174	GH01J230976	Promoter /Enhancer	TTC13	0.218	0.346429	0.672704	0.39412	0.422197
rs11122174	GH01J230976	Promoter /Enhancer	ARV1	0.218	0.346429	0.672704	0.39412	0.422197
rs11579993	GH01J206444	Enhancer	IKBKE	0.5	0.4645	0.2613	0.527	0.2815
rs11684916	GH02J143900	Enhancer	LOC101928386	0.513	0.316785	0.231912	0.25858	0.288868
rs117226814	GH03J150598	Promoter /Enhancer	EIF2A	0.282	0.076651	0.019231	0.000259	0.001378
rs117226814	GH03J150598	Promoter /Enhancer	SELENOT	0.282	0.076651	0.019231	0.000259	0.001378
rs117554880	GH07J006530	Enhancer	GRID2IP	0.481	0.174528	0	0.03374	0.013541
rs11809494	GH01J206448	Enhancer	IKBKE	0.5	0.4619	0.26	0.5257	0.3125
rs12419235	GH11J118633	Enhancer	PHLDB1	0.455	0.169811	0.148267	0.001751	0.003098
rs12419235	GH11J118633	Enhancer	TREH	0.455	0.169811	0.148267	0.001751	0.003098
rs12424645	GH12J026690	Enhancer	ITPR2	0.846	0.42891	0.563472	0.145241	0.104694
rs12514686	GH05J096733	Enhancer	CAST	0.609	0.238717	0.174419	0.05848	0.067394
rs12543368	GH08J142737	Enhancer	SLURP1	0.564	0.344869	0.165806	0.334097	0.165701
rs12543368	GH08J142737	Enhancer	LY6K	0.564	0.344869	0.165806	0.334097	0.165701

rs13013022	GH02J127540	Enhancer	MYO7B	0.904	0.735154	0.566452	0.630833	0.301472
rs13059182	GH03J012458	Enhancer	PPARG	0.487	0.379717	0.034018	0.283357	0.535714
rs13059182	GH03J012458	Enhancer	TSEN2	0.487	0.379717	0.034018	0.283357	0.535714
rs13059198	GH03J012458	Enhancer	PPARG	0.487	0.380615	0.034018	0.283262	0.53453
rs13059198	GH03J012458	Enhancer	TSEN2	0.487	0.380615	0.034018	0.283262	0.53453
rs13059878	GH03J012458	Enhancer	PPARG	0.487	0.382701	0.034018	0.284128	0.590471
rs13059878	GH03J012458	Enhancer	TSEN2	0.487	0.382701	0.034018	0.284128	0.590471
rs13065455	GH03J012439	Enhancer	PPARG	0.487	0.373821	0.028883	0.286225	0.489293
rs13065455	GH03J012439	Enhancer	SYN2	0.487	0.373821	0.028883	0.286225	0.489293
rs13065745	GH03J012439	Enhancer	PPARG	0.487	0.37234	0.028883	0.286225	0.48849
rs13065745	GH03J012439	Enhancer	SYN2	0.487	0.37234	0.028883	0.286225	0.48849
rs13088205	GH03J012439	Enhancer	PPARG	0.487	0.37234	0.029487	0.28628	0.48896
rs13088205	GH03J012439	Enhancer	SYN2	0.487	0.37234	0.029487	0.28628	0.48896
rs13088214	GH03J012439	Enhancer	PPARG	0.487	0.373522	0.029487	0.286048	0.48896
rs13088214	GH03J012439	Enhancer	SYN2	0.487	0.373522	0.029487	0.286048	0.48896
rs13090265	GH03J012439	Enhancer	PPARG	0.462	0.284198	0.000641	0.136328	0.095769
rs13090265	GH03J012439	Enhancer	SYN2	0.462	0.284198	0.000641	0.136328	0.095769
rs133054	GH22J040650	Enhancer	MKL1	0.128	0.359005	0.975547	0.553947	0.543338
rs133054	GH22J040650	Enhancer	XPNPEP3	0.128	0.359005	0.975547	0.553947	0.543338
rs133072	GH22J040679	Promoter /Enhancer	XPNPEP3	0.128	0.415094	0.975515	0.643712	0.698756
rs133073	GH22J040679	Promoter /Enhancer	XPNPEP3	0.128	0.398585	0.975578	0.624415	0.577552
rs139047783	GH10J103675	Enhancer	SH3PXD2A	0.288	0.096698	0	6.50E-05	0.000919
rs141243060	GH02J143479	Enhancer	ARHGAP15	0.583	0.224586	0.011539	0.003307	0.00402
rs153470	GH05J151208	Promoter /Enhancer	ANXA6	0.885	0.498818	0.536223	0.226565	0.179705
rs153470	GH05J151208	Promoter /Enhancer	SLC36A1	0.885	0.498818	0.536223	0.226565	0.179705

rs153470	GH05J151208	Promoter /Enhancer	GM2A	0.885	0.498818	0.536223	0.226565	0.179705
rs153470	GH05J151208	Promoter /Enhancer	CCDC69	0.885	0.498818	0.536223	0.226565	0.179705
rs16906993	GH11J009751	Promoter /Enhancer	TMEM9B-AS1	0.442	0.181604	0.100128	0.000454	0.004367
rs16906993	GH11J009751	Promoter /Enhancer	RPL21P97	0.442	0.181604	0.100128	0.000454	0.004367
rs16906993	GH11J009751	Promoter /Enhancer	SWAP70	0.442	0.181604	0.100128	0.000454	0.004367
rs16906993	GH11J009751	Promoter /Enhancer	SBF2-AS1	0.442	0.181604	0.100128	0.000454	0.004367
rs16906993	GH11J009751	Promoter /Enhancer	SBF2	0.442	0.181604	0.100128	0.000454	0.004367
rs16931169	GH12J026690	Enhancer	ITPR2	0.846	0.431442	0.564516	0.145314	0.104651
rs17132267	GH05J131994	Promoter /Enhancer	FNIP1	0.34	0.113208	0.044231	0.000584	0.001724
rs17132267	GH05J131994	Promoter /Enhancer	MEIKIN	0.34	0.113208	0.044231	0.000584	0.001724
rs1883473	GH22J040540	Enhancer	MKL1	0.878	0.563095	0.018662	0.274516	0.407595
rs1903002	GH04J088818	Promoter /Enhancer	FAM13A	0.455	0.520095	0.750643	0.538366	0.687155
rs1983127	GH10J069227	Enhancer	HKDC1	0.513	0.441038	0.813385	0.366931	0.33437
rs2009707	GH02J127633	Enhancer	GPR17	0.128	0.413915	0.354463	0.568797	0.903152
rs2052953	GH02J127540	Enhancer	MYO7B	0.904	0.743499	0.574194	0.630935	0.364139
rs2094374	GH09J124808	Enhancer	OLFML2A	0.474	0.237589	0.247748	0.093604	0.038329
rs2094374	GH09J124808	Enhancer	GOLGA1	0.474	0.237589	0.247748	0.093604	0.038329
rs2187832	GH22J040602	Enhancer	MKL1	0.115	0.303783	0.863402	0.460151	0.399977
rs2303027	GH05J151122	Enhancer	TNIP1	0.885	0.528302	0.468468	0.232424	0.298314
rs2303027	GH05J151122	Enhancer	ANXA6	0.885	0.528302	0.468468	0.232424	0.298314
rs2336938	GH01J206444	Enhancer	IKBKE	0.5	0.4645	0.259	0.5268	0.2834
rs2404991	GH02J127562	Enhancer	MYO7B	0.897	0.634752	0.554627	0.504613	0.175558
rs3091393	GH22J040507	Enhancer	MKL1	0.122	0.345154	0.865212	0.525591	0.613579
rs34930402	GH08J133044	Enhancer	SLA	0.404	0.219194	0.102564	0.0858	0.01654
rs3762731	GH03J101572	Promoter /Enhancer	PCNP	0.551	0.267689	0.202196	0.051214	0.035041

rs3792775	GH05J151122	Enhancer	TNIP1	0.885	0.527251	0.468468	0.231444	0.288594
rs3792775	GH05J151122	Enhancer	ANXA6	0.885	0.527251	0.468468	0.231444	0.288594
rs3806187	GH01J159778	Promoter /Enhancer	DUSP23	0.532	0.247642	0.145513	0.211169	0.082318
rs3806189	GH01J159778	Promoter /Enhancer	DUSP23	0.526	0.247045	0.145513	0.216457	0.088966
rs3860302	GH01J206448	Enhancer	IKBKE	0.5	0.4621	0.26	0.5254	0.2815
rs4127124	GH01J206448	Enhancer	IKBKE	0.487	0.4506	0.2552	0.5206	0.2358
rs413275	GH02J117813	Promoter /Enhancer	CCDC93	0.532	0.754717	0.998718	0.926024	0.983471
rs413275	GH02J117813	Promoter /Enhancer	DDX18	0.532	0.754717	0.998718	0.926024	0.983471
rs429383	GH01J006700	Promoter /Enhancer	THAP3	0.321	0.609005	0.998074	0.794221	0.80076
rs438524	GH04J002809	Promoter /Enhancer	SH3BP2	0.859	0.612293	0.467908	0.516018	0.461477
rs4721545	GH07J016846	Enhancer	AGR2	0.442	0.516667	0.823907	0.565476	0.673807
rs4721545	GH07J016846	Enhancer	LOC100287613	0.442	0.516667	0.823907	0.565476	0.673807
rs4776010	GH15J052178	Promoter /Enhancer	CERNA1	0.519	0.3125	0.095635	0.147204	0.232052
rs4776010	GH15J052178	Promoter /Enhancer	GNB5	0.519	0.3125	0.095635	0.147204	0.232052
rs4776011	GH15J052178	Promoter /Enhancer	CERNA1	0.519	0.310427	0.095513	0.146841	0.051217
rs4776011	GH15J052178	Promoter /Enhancer	GNB5	0.519	0.310427	0.095513	0.146841	0.051217
rs8031089	GH15J052231	Enhancer	GNB5	0.532	0.653302	0.031491	0.820867	0.263678
rs8031089	GH15J052231	Enhancer	MYO5C	0.532	0.653302	0.031491	0.820867	0.263678
rs17302275	GH15J059270	Enhancer	MYO1E	0.526	0.180851	0.0589744	0.00038895	0.00252467
rs1657152	GH16J001825	Promoter /Enhancer	HAGH	0.833	0.770784	0.570694	0.737256	0.541167
rs1657152	GH16J001825	Promoter /Enhancer	FAHD1	0.833	0.770784	0.570694	0.737256	0.541167
rs238682	GH16J001825	Promoter /Enhancer	HAGH	0.776	0.657518	0.407672	0.602526	0.307179
rs238682	GH16J001825	Promoter /Enhancer	FAHD1	0.776	0.657518	0.407672	0.602526	0.307179
rs61167445	GH16J048555	Enhancer	N4BP1	0.513	0.219194	0.0848329	0.00265716	0.00264246
rs61167445	GH16J048555	Enhancer	ENSG00000261267	0.513	0.219194	0.0848329	0.00265716	0.00264246

rs9895848	GH17J004832	Promoter /Enhancer	MINK1	0.513	0.326241	0.068299	0.343368	0.0812644
rs4790235	GH17J004898	Promoter /Enhancer	CHRNE	0.513	0.198113	0	0.0266096	0.0093232
rs4790235	GH17J004898	Promoter /Enhancer	C17orf107	0.513	0.198113	0	0.0266096	0.0093232
rs238246	GH17J004937	Promoter /Enhancer	RNF167	0.5	0.310142	0.153205	0.283258	0.0554153
rs238246	GH17J004937	Promoter /Enhancer	ENO3	0.5	0.310142	0.153205	0.283258	0.0554153
rs385471	GH17J004965	Promoter /Enhancer	ZNF232	0.519	0.460993	0.157216	0.534165	0.276322
rs385471	GH17J004965	Promoter /Enhancer	KIF1C	0.519	0.460993	0.157216	0.534165	0.276322
rs385471	GH17J004965	Promoter /Enhancer	INCA1	0.519	0.460993	0.157216	0.534165	0.276322
rs385471	GH17J004965	Promoter /Enhancer	CAMTA2	0.519	0.460993	0.157216	0.534165	0.276322
rs385471	GH17J004965	Promoter /Enhancer	SPAG7	0.519	0.460993	0.157216	0.534165	0.276322
rs4821950	GH22J040490	Enhancer	MKL1	0.84	0.382701	0.003205	0.031639	0.011366
rs4821950	GH22J040490	Enhancer	LOC101927257	0.84	0.382701	0.003205	0.031639	0.011366
rs4973353	GH02J230943	Enhancer	GPR55	0.218	0.323877	0.550258	0.330929	0.762075
rs536824	GH06J052325	Enhancer	PAQR8	0.224	0.50119	0.770914	0.641262	0.591851
rs62169043	GH02J143900	Enhancer	LOC101928386	0.506	0.314421	0.23166	0.258303	0.288523
rs62523626	GH08J141088	Enhancer	DENND3	0.269	0.089623	0	0.028518	0.006893
rs6806708	GH03J012439	Enhancer	PPARG	0.487	0.375297	0.029525	0.285668	0.516713
rs6806708	GH03J012439	Enhancer	SYN2	0.487	0.375297	0.029525	0.285668	0.516713
rs6976251	GH07J006007	Promoter /Enhancer	PMS2	0.506	0.705189	0.917632	0.785041	0.84258
rs7172037	GH15J052178	Promoter /Enhancer	CERNA1	0.513	0.311611	0.091784	0.146879	0.232373
rs7172037	GH15J052178	Promoter /Enhancer	GNB5	0.513	0.311611	0.091784	0.146879	0.232373
rs732149	GH12J026703	Enhancer	ITPR2	0.846	0.445755	0.567358	0.151311	0.250288
rs73342968	GH07J006446	Promoter /Enhancer	KDELR2	0.391	0.189737	0.016731	0.128566	0.159538
rs73342968	GH07J006446	Promoter /Enhancer	DAGLB	0.391	0.189737	0.016731	0.128566	0.159538
rs744180	GH12J026703	Enhancer	ITPR2	0.853	0.444575	0.563389	0.151317	0.226322

rs744181	GH12J026703	Enhancer	ITPR2	0.846	0.441943	0.558901	0.150817	0.224885
rs744182	GH12J026703	Enhancer	ITPR2	0.846	0.441805	0.556937	0.150811	0.22496
rs189987088	GH13J030879	Promoter /Enhancer	TEX26-AS1	0.256	0.0813679	0.00576923	0.00233342	0.0005736
rs189987088	GH13J030879	Promoter /Enhancer	MEDAG	0.256	0.0813679	0.00576923	0.00233342	0.0005736
rs189987088	GH13J030879	Promoter /Enhancer	LINC00545	0.256	0.0813679	0.00576923	0.00233342	0.0005736
rs78015681	GH14J104084	Promoter /Enhancer	ASPG	0.359	0.147406	0.00321337	6.50E-05	0.0005760
rs74644686	GH05J132272	Enhancer	P4HA2	0.603	0.214623	0.119537	0.001168	0.003908
rs74655205	GH03J010955	Enhancer	SLC6A1	0.795	0.376179	0.287645	0.157319	0.039903
rs75092723	GH05J096699	Promoter /Enhancer	CAST	0.609	0.234043	0.167746	0.056952	0.012969
rs75092723	GH05J096699	Promoter /Enhancer	ERAP1	0.609	0.234043	0.167746	0.056952	0.012969
rs75092723	GH05J096699	Promoter /Enhancer	ENSG00000248734	0.609	0.234043	0.167746	0.056952	0.012969
rs75092723	GH05J096699	Promoter /Enhancer	ENSG00000249180	0.609	0.234043	0.167746	0.056952	0.012969
rs75565343	GH05J096699	Promoter /Enhancer	CAST	0.609	0.234043	0.166881	0.056573	0.012747
rs75565343	GH05J096699	Promoter /Enhancer	ERAP1	0.609	0.234043	0.166881	0.056573	0.012747
rs75565343	GH05J096699	Promoter /Enhancer	ENSG00000248734	0.609	0.234043	0.166881	0.056573	0.012747
rs75565343	GH05J096699	Promoter /Enhancer	ENSG00000249180	0.609	0.234043	0.166881	0.056573	0.012747
rs75951675	GH07J006423	Enhancer	RAC1	0.397	0.188836	0.017308	0.120808	0.156649
rs75951675	GH07J006423	Enhancer	DAGLB	0.397	0.188836	0.017308	0.120808	0.156649
rs77791280	GH05J096699	Promoter /Enhancer	CAST	0.609	0.234043	0.167738	0.056595	0.012741
rs77791280	GH05J096699	Promoter /Enhancer	ERAP1	0.609	0.234043	0.167738	0.056595	0.012741
rs77791280	GH05J096699	Promoter /Enhancer	ENSG00000248734	0.609	0.234043	0.167738	0.056595	0.012741
rs77791280	GH05J096699	Promoter /Enhancer	ENSG00000249180	0.609	0.234043	0.167738	0.056595	0.012741
rs7806655	GH07J006433	Enhancer	DAGLB	0.397	0.189125	0.017375	0.125359	0.158149
rs7828042	GH08J142737	Enhancer	SLURP1	0.564	0.373522	0.256145	0.380352	0.249596
rs7828042	GH08J142737	Enhancer	LY6K	0.564	0.373522	0.256145	0.380352	0.249596

rs78575823	GH11J047399	Promoter /Enhancer	MADD	0.308	0.121103	0.021263	0.000457	0.002312
rs78575823	GH11J047399	Promoter /Enhancer	SPI1	0.308	0.121103	0.021263	0.000457	0.002312
rs78575823	GH11J047399	Promoter /Enhancer	ENSG00000255197	0.308	0.121103	0.021263	0.000457	0.002312
rs8135759	GH22J040460	Promoter /Enhancer	MKL1	0.878	0.658019	0.135604	0.482208	0.386112
rs963894	GH01J206448	Enhancer	IKBKE	0.5	0.4634	0.2619	0.5269	0.2815
rs9682156	GH03J101511	Promoter /Enhancer	SENP7	0.846	0.46274	0.287599	0.174045	0.128423
rs9682312	GH03J101511	Promoter /Enhancer	SENP7	0.846	0.47066	0.288742	0.177419	0.131269
rs9833097	GH03J012436	Enhancer	PPARG	0.455	0.27305	0.000642	0.111768	0.103019
rs9833097	GH03J012436	Enhancer	TIMP4	0.455	0.27305	0.000642	0.111768	0.103019
rs9907721	GH17J016666	Promoter /Enhancer	NEK4P2	0.282	0.370283	0.573739	0.255945	0.513237
rs7231988	GH18J049645	Enhancer	LIPG	0.615	0.343972	0.140565	0.217589	0.160304
rs12454833	GH18J049645	Enhancer	LIPG	0.596	0.343602	0.139923	0.221875	0.160377
rs78942247	GH18J059904	Enhancer	PMAIP1	0.5	0.225236	0.15276	0.124465	0.0259531
rs57071021	GH19J051748	Promoter /Enhancer	FPR1	0.763	0.437352	0.405161	0.285081	0.0783299
rs57071021	GH19J051748	Promoter /Enhancer	FPR2	0.763	0.437352	0.405161	0.285081	0.0783299
rs16985650	GH20J041307	Enhancer	LPIN3	0.308	0.125	0.0243902	0.0022035	0.00114811
rs12485003	GH22J040238	Enhancer	TNRC6B	0.692	0.315166	0.097561	0.100221	0.0203168
rs4821942	GH22J040321	Enhancer	ADSL	0.853	0.448113	0.223008	0.208323	0.0438032
rs202634	GH22J041409	Enhancer	EP300	0.192	0.505938	0.939689	0.789143	0.954524
rs202635	GH22J041409	Enhancer	EP300	0.186	0.497619	0.939845	0.788608	0.825751
rs132919	GH22J041409	Enhancer	EP300	0.186	0.501185	0.940568	0.789585	0.826092
rs9611612	GH22J041559	Promoter /Enhancer	CSDC2	0.949	0.571605	0.0827633	0.202514	0.332474

12.4.4 Variantes por Elemento Enhancer

Tabla S2. Número de Variantes en cada Elemento Enhancer

GeneHancerID	Variantes	GeneHancerID	Variantes	GeneHancerID	Variantes
GH03J012439	6	GH02J143479	1	GH03J010955	1
GH01J206448	5	GH02J230943	1	GH03J012436	1
GH12J026690	5	GH07J006423	1	GH03J101572	1
GH12J026703	4	GH07J006433	1	GH03J150598	1
GH03J012458	3	GH07J006446	1	GH04J002809	1
GH05J096699	3	GH07J006530	1	GH04J088818	1
GH15J052178	3	GH07J016846	1	GH05J096733	1
GH22J041409	3	GH08J133044	1	GH05J131994	1
GH01J159778	2	GH08J141088	1	GH05J132272	1
GH01J206444	2	GH09J124808	1	GH05J151208	1
GH02J127540	2	GH10J069227	1	GH06J052325	1
GH02J143900	2	GH10J103675	1	GH17J016666	1
GH03J101511	2	GH11J009751	1	GH18J059904	1
GH05J151122	2	GH11J047399	1	GH19J051748	1
GH08J142737	2	GH11J118633	1	GH20J041307	1
GH16J001825	2	GH12J026119	1	GH22J040238	1
GH18J049645	2	GH13J030879	1	GH22J040321	1
GH22J040679	2	GH14J104084	1	GH22J040460	1
GH01J006700	1	GH15J052231	1	GH22J040490	1
GH01J159596	1	GH15J059270	1	GH22J040507	1
GH01J230976	1	GH16J048555	1	GH22J040540	1
GH02J117813	1	GH17J004832	1	GH22J040602	1
GH02J127562	1	GH17J004898	1	GH22J040650	1
GH02J127633	1	GH17J004937	1	GH22J041559	1
GH07J006007	1	GH17J004965	1		

12.4.5 Genes con Variantes Genéticas Localizadas en Elementos Reguladores Enhancer

Tabla S3. Número de variantes en enhancers por gen.

Gen	Variantes en Enhancers	Gen	Variantes en Enhancers	Gen	Variantes en Enhancers
PPARG	10	ARHGAP15	1	ADSL	1
ITPR2	9	ARV1	1	AGR2	1
IKBKE	7	ASPG	1	INCA1	1
MKL1	6	C17orf107	1	KDELR2	1
SYN2	6	CAMTA2	1	KIF1C	1

CAST	4	CCDC69	1	LINC00545	1
GNB5	4	CCDC93	1	LOC100287613	1
ANXA6	3	CHRNE	1	LOC101927257	1
CERNA1	3	CSDC2	1	LPIN3	1
DAGLB	3	DDX18	1	MADD	1
ENSG00000248734	3	DENND3	1	MEDAG	1
ENSG00000249180	3	EIF2A	1	MEIKIN	1
EP300	3	ENO3	1	MINK1	1
ERAP1	3	ENSG00000255197	1	MYO1E	1
MYO7B	3	ENSG00000256894	1	MYO5C	1
TSEN2	3	ENSG00000261267	1	N4BP1	1
XPNPEP3	3	FAM13A	1	NEK4P2	1
DUSP23	2	FNIP1	1	OLFML2A	1
FAHD1	2	FPR1	1	OR10J6P	1
HAGH	2	FPR2	1	P4HA2	1
LIPG	2	GM2A	1	PAQR8	1
LOC101928386	2	GOLGA1	1	PCNP	1
LY6K	2	GPR17	1	PHLDB1	1
SENP7	2	GPR55	1	PMAIP1	1
SLURP1	2	GRID2IP	1	PMS2	1
TNIP1	2	HKDC1	1	RAC1	1
RNF167	1	SH3BP2	1	SSPN	1
RPL21P97	1	SH3PXD2A	1	SWAP70	1
SBF2	1	SLA	1	TEX26-AS1	1
SBF2-AS1	1	SLC36A1	1	THAP3	1
SELENOT	1	SLC6A1	1	TIMP4	1
TREH	1	SPAG7	1	TMEM9B-AS1	1
TTC13	1	SPI1	1	TNRC6B	1

12.4.6 Densidad de Variantes en Elementos Reguladores por Gen

Tabla S4. Densidad de Variantes en Enhancers por Gen

Gen	Número de Variantes por Kilobase	Gen	Número de Variantes por Kilobase
LOC101928386	1.11731844	FNIP1	0.07920792
CERNA1	1.07526882	DDX18	0.07469934
NEK4P2	1.00806452	CAST	0.07364448
SLURP1	0.75075075	CCDC93	0.07162811
DUSP23	0.6097561	RNF167	0.06922811
OR10J6P	0.36832413	FPR2	0.06721785
PCNP	0.29171529	CHRNE	0.06468305
ARV1	0.2772387	ENSG00000255197	0.06427148
ITPR2	0.26693558	GOLGA1	0.06258997
LINC00545	0.26034887	PMAIP1	0.06023371
SLC6A1	0.25265285	AGR2	0.05767013
LOC100287613	0.24740228	SPI1	0.05576623
FAHD1	0.23326335	ERAP1	0.05540678
HAGH	0.23326335	C17orf107	0.05239992
GRID2IP	0.2247191	LOC101927257	0.04792256
PPARG	0.21105061	ANXA6	0.04719578
LIPG	0.20498104	MEDAG	0.04616166
ASPG	0.1998002	MINK1	0.04592212
TEX26-AS1	0.17746229	TNIP1	0.04578545
TTC13	0.17289073	ARHGAP15	0.04463688
GNB5	0.15716475	TMEM9B-AS1	0.04456129
LY6K	0.15270673	N4BP1	0.04451171
DAGLB	0.15015015	EIF2A	0.04428306
SYN2	0.14979777	FAM13A	0.04377709
ENSG00000261267	0.14907573	HKDC1	0.04284674
EP300	0.14800928	FPR1	0.04084634
MYO7B	0.14648438	PHLDB1	0.03958045
IKBKE	0.1461683	RAC1	0.03785871
MEIKIN	0.13875399	THAP3	0.03615068
GPR55	0.12906557	ADSL	0.03586801
ENSG00000249180	0.12759985	SBF2-AS1	0.03499685
PMS2	0.12656626	ENO3	0.03431356
TSEN2	0.12572818	SH3BP2	0.03376895
CSDC2	0.12106538	P4HA2	0.0335762
KDELR2	0.11414222	RPL21P97	0.03281701
SELENOT	0.11316057	GM2A	0.03089376
ENSG00000248734	0.11112346	SSPN	0.03054648
SENP7	0.11090778	TREH	0.02968592

XPNPEP3	0.10547781	SLC36A1	0.02901242
INCA1	0.10467916	CCDC69	0.02850058
KIF1C	0.10467916	PAQR8	0.02772541
CAMTA2	0.10115315	MADD	0.02715842
ZNF232	0.10013017	LPIN3	0.02395324
ENSG00000256894	0.09843489	DENND3	0.02351669
SPAG7	0.09410879	TNRC6B	0.02275313
OLFML2A	0.09211496	SLA	0.02020161
MYO5C	0.09013069	SWAP70	0.01331416
MKL1	0.08919546	SBF2	0.01202342
TIMP4	0.08887309	MYO1E	0.01177662
GPR17	0.08537522	SH3PXD2A	0.01032919

12.4.7 Variantes en Enhancers de PPARG, SYN2, TIMP4 y TSEN2

Tabla S5. Variantes detectadas en enhancers de PPARG, SYN2, TIMP4 y TSEN2

rsID	GHid y Gen Regulado	100 G-MX	Latina	Asiática del Este	Europea no Finlandesa	Africana
rs9833097	GH03J012436 PPARG, TIMP4	0.455	0.273	0.0006418	0.1118	0.103
rs13090265	GH03J012439 PPARG, SYN2	0.462	0.2842	0.000641	0.1363	0.09577
rs6806708	GH03J012439 PPARG, SYN2	0.487	0.3753	0.02953	0.2857	0.5167
rs13065745	GH03J012439 PPARG, SYN2	0.487	0.3723	0.02888	0.2862	0.4885
rs13065455	GH03J012439 PPARG, SYN2	0.487	0.3738	0.02888	0.2862	0.4893
rs13088205	GH03J012439 PPARG, SYN2	0.487	0.3723	0.02949	0.2863	0.489
rs13088214	GH03J012439 PPARG, SYN2	0.487	0.3735	0.02949	0.286	0.489
rs13059182	GH03J012458 PPARG, TSEN2	0.487	0.3797	0.03402	0.2834	0.5357
rs13059198	GH03J012458 PPARG, TSEN2	0.487	0.3806	0.03402	0.2833	0.5345
rs13059878	GH03J012458 PPARG, TSEN2	0.487	0.3827	0.03402	0.2841	0.5905