



**UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO**

**Programa de Maestría y Doctorado en Ciencias Bioquímicas**

**ESTUDIO GENÉTICO DE LAS DISLIPIDEMIAS EN UNA COHORTE DE LA  
CIUDAD DE MÉXICO - PROYECTO TLALPAN 2020.**

**TESIS**

**QUE PARA OPTAR POR EL GRADO DE:  
MAESTRO EN CIENCIAS BIOQUÍMICAS**

**PRESENTA:**

**TOMÁS EDUARDO TEXIS VALENCIA**

**TUTOR PRINCIPAL**

**DR. ENRIQUE HERNÁNDEZ LEMUS  
INSTITUTO NACIONAL DE MEDICINA GENÓMICA**

**MIEMBROS DEL COMITÉ TUTOR**

**DRA. MIRELLE VANESSA GONZÁLEZ COVARRUBIAS  
INSTITUTO NACIONAL DE MEDICINA GENÓMICA**

**DR. SAMUEL CANIZALES QUINTEROS**

**FACULTAD DE QUÍMICA (UNIDAD PERIFÉRICA INMEGEN)**

Ciudad de México, julio 2020



Universidad Nacional  
Autónoma de México

Dirección General de Bibliotecas de la UNAM

**Biblioteca Central**



**UNAM – Dirección General de Bibliotecas**  
**Tesis Digitales**  
**Restricciones de uso**

**DERECHOS RESERVADOS ©**  
**PROHIBIDA SU REPRODUCCIÓN TOTAL O PARCIAL**

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

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

Este trabajo de investigación fue realizado gracias a los financiamientos del Consejo Nacional de Ciencia y Tecnología (CONACYT) número 247,237 y del Consejo Nacional de Ciencia y Tecnología – Fondo Sectorial de Investigación en Salud y Seguridad Social (CONACYT – FOSSIS) número 272,795.

## Agradecimientos

Realicé este trabajo con mucha dedicación y amor por la investigación, pero quiero comenzar agradeciendo a aquellas personas que me ayudaron a encontrar esa pasión por la ciencia y que además me guiaron por este proceso.

Desde hace más de 4 años, la Dra. Vanessa González, del laboratorio de farmacogenómica del INMEGEN, me abrió las puertas y me ayudó a encontrar mi vocación. Ella ha sido un elemento clave en mi formación, desde nivel licenciatura y siempre le estaré agradecido. Gracias por compartirme su conocimiento y experiencia en estos años.

Durante la estancia de maestría conocí a dos grandes investigadores que tiene el país y que me ayudaron en la integración del conocimiento para realizar este trabajo, el Dr. Enrique Hernández Lemus, mi tutor, y el Dr. Samuel Canizales. Les agradezco mucho su tiempo, tolerancia, dedicación y humildad, son excelentes tutores.

Quiero reconocer que el INMEGEN merece una gran mención en este apartado, ya que es un instituto de alta especialidad, que alberga recursos tanto materiales como humanos, con los cuales logré finalizar satisfactoriamente mi estancia. En este sentido, quiero agradecer igualmente a la UNAM, la máxima casa de estudios, que me aportó, en gran medida, con el cuerpo de profesores de tan excelente calidad, quienes de igual forma me ayudaron durante este proceso e indirectamente en la formación de este trabajo, quiero agradecer específicamente a la Dra. Blanca Taboada, al Dr. Lorenzo Segovia y al Dr. Alejandro Garciarrubio.

Agradezco al Instituto Nacional de Cardiología, por permitirme colaborar en este proyecto tan ambicioso y enriquecedor como lo es, Tlalpan 2020, en específico a la Dra. Susana Rivera y a la Dra. Maite Vallejo, quienes coordinaron, junto con su equipo de trabajo, la recepción y manejo de base de datos de todas las muestras.

De igual forma, quiero reconocer el apoyo que me brindaron mis compañeros de laboratorio: Elba, Ari, Cintia, Karla, Martin, Karina y David, quienes siempre me han brindado su amistad, pero en los momentos complicados me han apoyado, dándome consejos o ayudándome.

También quiero agradecer a mis “pokeamigos” con quienes he pasado grandes momentos y siempre me han brindado la mano o han estado ahí para escucharme, aun cuando no les interesa el tema o no del todo. Les tengo mucho aprecio y admiración. Agradezco a Luis Tule, Pamas, Nestor y Manuel.

Por último, pero no por eso los menos importantes sino todo lo contrario, agradezco a mis padres, ya que sin su perseverancia y apoyo, yo no hubiera siquiera entrado al posgrado y aún más finalizado. Gracias por estar conmigo tantos años y escucharme, más durante estos 4 meses de pandemia que han sido bastante complicados.

## TABLA DE CONTENIDO

<b>ÍNDICE DE TABLAS</b> .....	<b>6</b>
<b>ÍNDICE DE FIGURAS</b> .....	<b>7</b>
<b>ÍNDICE ANEXO</b> .....	<b>8</b>
<b>RESUMEN</b> .....	<b>9</b>
<b>ANTECEDENTES</b> .....	<b>10</b>
Estudio Tlalpan 2020 .....	10
<b>INTRODUCCIÓN</b> .....	<b>11</b>
Metabolismo de lípidos.....	11
Dislipidemias .....	13
Etiología de las dislipidemias .....	14
Prevalencia mundial de las dislipidemias .....	14
Prevalencia en México de las dislipidemias.....	15
Índices Aterogénicos .....	16
Marcadores genéticos de dislipidemia y riesgo aterogenico: estudios de asociación. ....	18
Estudios genéticos de las dislipidemias .....	19
<b>OBJETIVOS</b> .....	<b>21</b>
General .....	21
Específicos .....	21
<b>JUSTIFICACIÓN</b> .....	<b>21</b>
<b>HIPÓTESIS</b> .....	<b>21</b>
<b>METODOLOGÍA</b> .....	<b>22</b>
Muestras Biológicas.....	22
COLECCIÓN DE MUESTRAS SANGUÍNEAS .....	22
EXTRACCIÓN DE ADN .....	23
CONTROL DE DALIDAD DEL ADN: CUANTIFICACIÓN, EVALUACION DE PUREZA E INTEGRIDAD ...	24
Genotipado .....	25
GENOTIPIFICACIÓN POR MICROARREGLOS .....	25
Análisis bioinformático.....	26
CONTROL DE CALIDAD DE DATOS GENÓMICOS .....	26
Análisis de mestizaje .....	30
ESTRATEGIA DE ENRIQUECIMIENTO DE GEN CANDIDATO .....	31

Estadística descriptiva .....	32
CARACTERIZACIÓN DE LA POBLACIÓN .....	32
SELECCIÓN DE COVARIANTES .....	33
Análisis de asociación .....	34
MODELOS DE ASOCIACIÓN DE GEN CANDIDATO (PRIMER ETAPA).....	34
MODELOS DE ASOCIACIÓN CONDICIONAL DE GENES RELACIONADOS (REGRESIÓN DE COMPONENTES PRINCIPALES: SEGUNDA ETAPA) .....	35
<b>RESULTADOS.....</b>	<b>38</b>
Estadística descriptiva .....	38
Selección de covariantes clínicas.....	39
Control de calidad de muestras .....	42
Control de calidad de datos genómicos para muestras control.....	44
Control de calidad de datos genómicos para estudios GWAS .....	45
EFICIENCIA EN EL GENOTIPADO POR VARIANTE Y MUESTRA.....	45
VERIFICACIÓN DE SEXO.....	45
HARDY-WEINBERG .....	45
Análisis de mestizaje (proporción del componente ancestral).....	46
PROPORCION ANCESTRAL DE LA POBLACION DE ESTUDIO: ANÁLISIS DE COMPONENTES PRINCIPALES (PCA).....	47
PROPORCIÓN DEL COMPONENTE ANCESTRAL .....	48
Estrategia de enriquecimiento de gen candidato.....	49
Desarrollo de modelos de asociación genotipo-fenotipo.....	51
Desarrollo de modelos de regresión condicional .....	59
<b>DISCUSIÓN.....</b>	<b>62</b>
Variables clínicas de la población de estudio.....	62
Estrategia gen candidato.....	62
Análisis de asociación multivariable.....	63
Análisis de asociación condicional .....	64
Variantes encontradas que forman haplotipo.....	65
Genes relacionados con el metabolismo de lípidos: <i>HNF4A, ANGPTL4, FAIM2, LINGO2, MTUS1</i> .....	69
Genes relacionados con el síndrome metabólico: <i>MUC16, YPEL2, TMEM132C, FAM19A2</i> .....	72
Otros genes: <i>CDH26 y SHISA6</i> .....	74
<b>CONCLUSIONES.....</b>	<b>78</b>
<b>PERSPECTIVAS.....</b>	<b>79</b>

<b>REFERENCIAS.....</b>	<b>80</b>
<b>ANEXO. Geles de agarosa .....</b>	<b>87</b>

## ABREVIATURAS

ADN	Ácido Desoxirribonucleico
C	Colesterol
CT	Colesterol Total
CAD	Enfermedad Arterial Coronaria
CHB	Chinos Han de Beijing
CEU	Europeos del Norte de Utah
dL	Decilitros
EDTA	Ácido etilendiaminotetraacético
EUR	Superpoblación europea
GWAS	Estudios de Asociación del Genoma Completo
h	Horas
C-HDL	Lipoproteínas de Alta Densidad
HTA	Hipertensión arterial
IDL	Lipoproteínas de Densidad Intermedia
kb	Kilobases
C-LDL	Lipoproteínas de Baja Densidad
LH	Lipasa Hepática
LPL	Lipasa Lipoproteica
MAF	Frecuencia del Alelo Menor
MXL	Mexicanos de Los Ángeles
min	Minutos
mg	Miligramos
mL	Mililitros
mmol	Milimol
Nat	Nativos mexicanos
pb	Pares de bases

PC	Componentes principales
PCA	Análisis de componentes principales
QM	Quilomicrones
rpm	Revoluciones por minuto
SNP	Polimorfismo de Nucleótido Sencillo
TG	Triglicérido
μL	Microlitro
C-VLDL	Lipoproteínas de Muy Baja Densidad
YRI	Yoruba en Ibadan de Nigeria

## ÍNDICE DE TABLAS

Tabla 1. Partículas lipoproteicas con sus respectivas apolipoproteinas asociadas, así como la composición de colesterol y proteína. ....	12
Tabla 2. Clasificación de las dislipidemias según La Sociedad Mexicana de Endocrinología .....	14
Tabla 3. Porcentaje de adultos con hipercolesterolemia por diagnóstico previo. ....	15
Tabla 4. Límites de riesgo para los índices aterogénicos estratificados por género.....	18
Tabla 5. Características demográficas y clínicas de la cohorte Tlalpan 2020, comparando entre hombres y mujeres. ....	39
Tabla 6. Número de muestras y/o variantes que pasaban en cada filtro del control de calidad para estudios de asociación genotipo-fenotipo.....	46
Tabla 7. Número de SNPs (genes) asociados significativamente ( $P < 0.05$ ) con los diferentes fenotipos usando regresiones lineales y logísticas múltiples.....	52
Tabla 8. SNPs asociados significativamente con hipercolesterolemia, usando regresiones logísticas múltiples, comparando entre voluntarios sanos y dislipidémicos ( $P < 9E-4$ ). ....	52
Tabla 9. SNPs asociados significativamente con hipertrigliceridemia, usando regresiones logísticas múltiples, comparando entre voluntarios sanos (controles) y dislipidémicos ( $P < 9E-4$ ). ....	53
Tabla 10. SNPs asociados significativamente con hipoalfalipoproteinemia, usando regresiones logísticas múltiples, comparando entre voluntarios sanos y dislipidémicos ( $P < 9E-4$ ). ....	53
Tabla 11. SNPs asociados significativamente con dislipidemia mixta, usando regresiones logísticas múltiples, comparando entre voluntarios sanos y dislipidémicos ( $P < 9E-4$ ). ....	54
Tabla 12. SNPs asociados significativamente con colesterol total de voluntarios sanos y dislipidémicos usando regresiones lineales múltiples ( $P < 9E-4$ ). ....	54
Tabla 13. SNPs asociados significativamente con triglicéridos de voluntarios sanos y dislipidémicos usando regresiones lineales múltiples ( $P < 9E-4$ ). ....	55
Tabla 14. SNPs asociados significativamente con C-LDL de voluntarios sanos y dislipidémicos usando regresiones lineales múltiples ( $P < 9E-4$ ). ....	55
Tabla 15. SNPs asociados significativamente con C-HDL de voluntarios sanos y dislipidémicos usando regresiones lineales múltiples ( $P < 9E-4$ ). ....	55
Tabla 16. SNPs asociados significativamente con índice aterogénico de plasma de voluntarios sanos y dislipidémicos usando regresiones lineales múltiples ( $P < 9E-4$ ). ....	56
Tabla 17. SNPs asociados significativamente con índice de Castelli II de voluntarios sanos y dislipidémicos usando regresiones lineales múltiples ( $P < 9E-4$ ). ....	57
Tabla 18. Polimorfismos de efecto independiente asociados con el Índice Aterogénico de Plasma (AIP).....	60
Tabla 19. Polimorfismos de efecto independiente asociados al Índice de Castelli II (IC II) .....	61

## ÍNDICE DE FIGURAS

Figura 1. Rutas metabólicas asociadas a lípidos. ....	13
Figura 2. Porcentaje de la prevalencia mundial de altas concentraciones de colesterol (hipercolesterolemia) en adultos (>25 años) de ambos sexos en el 2008. ....	15
Figura 3. Riesgos aterogénicos asociados a las lipoproteínas. ....	17
Figura 4. Estrategia propuesta para abordar los objetivos propuestos.....	22
Figura 5. Diagrama de flujo del control de calidad de datos genómicos.....	26
Figura 6. Representación de los estratos para la comparación de la frecuencia alélica. Se dividió por género (hombres y mujeres), por tipo de dislipidemia (hipercolesterolemia, hipertrigliceridemia, hipoalfalipoproteinemia y dislipidemia mixta) y por genero-tipo de dislipidemia.....	31
Figura 7. Agrupación jerárquica de las covariantes clínicas.....	40
Figura 8. Dispersión de las muestras en los primeros dos componentes principales que explican el 51% de la variabilidad.....	41
Figura 9. Gel de agarosa al 2%. Carril 1º, escalera de peso molecular 1 Kb, seguido del identificador para cada muestra. ....	43
Figura 10. Dispersión de las muestras en base a las diferentes estructuras poblacionales en los primeros componentes principales que explican el 74% del total de la variación del panel de variantes para evaluar ancestría.....	47
Figura 11. Proporción de las cuatro estructuras ancestrales (CHB, CEU, YRI, Nativos) para los mestizos de este proyecto, los nativos mexicanos y las muestras de 1000 genomas para CHB, CEU e YRI.....	48
Figura 12. Número de genes que resultaron de la comparación de las frecuencias alélicas entre casos y controles para cada grupo.....	50
Figura 13. Estrategias utilizadas para la identificación de variantes asociadas a dislipidemias a través de modelos de asociación tipo gen candidato .....	58
Figura 14. Desequilibrio de ligamiento en D' entre las variantes de DSCAM rs57292936, rs2837379, rs760168 y CHODL rs77923858, rs77779037 que fueron asociadas a índices aterogénicos.....	65

## ÍNDICE ANEXO

Figura I. Gel de agarosa al 2%. Carril 1º, escalera de peso molecular 1 Kb, seguido del identificador para cada muestra. ....	87
Figura II. Gel de agarosa al 2%. Carril 1º, escalera de peso molecular 1 Kb, seguido del identificador para cada muestra. ....	87
Figura III. Gel de agarosa al 2%. Carril 1º, escalera de peso molecular 1 Kb, seguido del identificador para cada muestra. ....	88
Figura IV. Gel de agarosa al 2%. Carril 1º, escalera de peso molecular 1 Kb, seguido del identificador para cada muestra. ....	88
Figura V. Gel de agarosa al 2%. Carril 1º, escalera de peso molecular 1 Kb, seguido del identificador para cada muestra. ....	89
Figura VI. Gel de agarosa al 2%. Carril 1º, escalera de peso molecular 1 Kb, seguido del identificador para cada muestra. ....	89
Figura VII. Gel de agarosa al 2%. Carril 1º, escalera de peso molecular 1 Kb, seguido del identificador para cada muestra. ....	90
Figura VIII. Gel de agarosa al 2%. Carril 1º, escalera de peso molecular 1 Kb, seguido del identificador para cada muestra. ....	90
Figura IX. Gel de agarosa al 2%. Carril 1º, escalera de peso molecular 1 Kb, seguido del identificador para cada muestra. ....	91
Figura X. Gel de agarosa al 2%. Carril 1º, escalera de peso molecular 1 Kb, seguido del identificador para cada muestra. ....	91

## RESUMEN

### *Justificación y antecedentes*

Las dislipidemias son un conjunto de enfermedades asintomáticas causadas por concentraciones anormales de lípidos en sangre <sup>1</sup>. Se estima que variaciones genéticas pueden explicar alrededor del 40-50% de la variabilidad en los niveles de lípidos. Existen más de 100 loci asociados a niveles de triglicéridos, colesterol, LDL-C y HDL-C <sup>2,3</sup>. Recientemente, la relación entre diferentes lípidos ha dado origen a diversos índices aterogénicos como el AIP ( $\log \frac{TG}{HDL-C}$ ) y el Castelli II ( $\frac{LDL-C}{HDL-C}$ ), los cuales son más precisos para predecir morbilidades cardiovasculares. Por lo anterior, este proyecto busca identificar variantes genéticas asociadas a índices de riesgo aterogénico, en individuos de la cohorte Tlalpan 2020.

### *Metodología*

La población de estudio (n=440) representa una submuestra de iguales proporciones a la muestra original 70-30% casos-contróles. La genotipificación se realizó utilizando el microarreglo GSA-Illumina. Los datos genómicos se sometieron a análisis de calidad <sup>4</sup>, bioinformáticos y estadísticos. Para el desarrollo de modelos de asociación, se seleccionó una estrategia de gen candidato posterior a una exhaustiva revisión bibliográfica <sup>3,5-7</sup>. Para los análisis estadísticos y de asociación se incluyeron 426 muestras y 373 genes con 10,990 variantes considerando como covariantes clínicas, ácido úrico, peso, cintura, sexo y mestizaje y como variables de respuesta los índices aterogénicos AIP y Castelli II. El desarrollo de modelos de asociación incluyeron dos pasos a) un método de regresión lineal multivariable y b) modelo multivariable condicional usando componentes principales, para identificar variantes genéticas de efecto independiente entre ellas, se utilizó la estrategia FDR (false discovery rate) para corregir por comparaciones múltiples/ falsos positivos.

### *Resultados.*

Se observaron diferencias entre casos (dislipidemias) y controles. Por ejemplo, individuos con dislipidemia mostraron valores más altos de IMC, edad, glucosa, ácido úrico, colesterol total (TC), triglicéridos (TG), y lipoproteínas de baja densidad (LDL-C). La regresión lineal multivariable identificó 680 variantes (en 180 genes) asociadas con los índices aterogénicos: AIP y Castelli II ( $9E-7 > P < 0.05$ ). Por ejemplo, *APOC3* rs5128 ( $P = 9E-7$ ), variante previamente asociada con el metabolismo triglicéridos y colesterol <sup>8</sup>. Y posteriormente, a través de la regresión condicional identificamos 32 variantes en 16 genes asociadas a los índices aterogénicos. Estos genes están involucrados con el metabolismo lípidos, síndrome metabólico y el desarrollo neuronal embrionario.

### *Conclusiones/Perspectivas*

Actualmente se procesa de validación del modelo, análisis exploratorios identificaron dos haplotipos asociados a ambos índices aterogénicos, uno conformado por las variantes *DSCAM* rs2837379 y rs760168, el otro por *NDUFB9* rs3829038 y rs77538674. El haplotipo *DSCAM* tiene un tamaño del efecto de -0.23, por lo que el alelo C, está asociado a menores niveles de Castelli II. La frecuencia de este haplotipo es hasta dos veces mayor en la población europea con respecto a la mexicana ( $AF_{EUR} = 0.38$  contra  $AF_{T2020} = 0.14$ ,  $P = 2.2E-3$ ). El haplotipo de *NDUFB9* tiene un tamaño del efecto 0.21, por lo que el alelo A, está asociado a mayores niveles de AIP y es hasta dos veces más frecuente en la población mexicana con respecto a la europea ( $AF_{EUR} = 0.03$  contra  $AF_{T2020} = 0.06$ ,  $P = 4.6E-3$ ).

## **ANTECEDENTES**

### **Estudio Tlalpan 2020**

En 2014 el Instituto Nacional de Cardiología formó la iniciativa del proyecto “Tlalpan 2020”, cuyo objetivo inicial fue la determinación de la incidencia de hipertensión arterial sistémica en mestizos mexicanos con residencia en la Ciudad de México (CDMX). Este estudio epidemiológico longitudinal con duración de 10 años, busca integrar diversos factores biológicos, genéticos, sociodemográficos y de estilo de vida, para asociarlos al desarrollo de la hipertensión.

Se invitó a la población mexicana de la CDMX a participar de forma voluntaria, incluyendo a hombres y mujeres clínicamente sanos, entre 20 y 50 años de edad para evaluarlos durante 10 años e identificar quienes desarrollan hipertensión y que factores pudieran estar involucrados en el desarrollo.

Se reclutaron 1330 participantes hasta mayo del 2014, sin embargo, en diciembre del 2015, cuando ya se comenzaron a revisar resultados de estudios clínicos, se detectó una incidencia del 70% en anomalías en el metabolismo de lípidos (dislipidemia), y aproximadamente la mitad de ellos (33% del total) tenían dislipidemia mixta, por lo que se supuso existe una susceptibilidad en la población mexicana en el desarrollo a dislipidemias. Estas observaciones condujeron a proponer un nuevo planteamiento para investigar las posibles causas de estas anomalías en los niveles de lípidos en sangre. Una posibilidad es que las dislipidemias puedan ser explicadas, en parte, por factores genéticos y dado que habrá un seguimiento de 10 años es posible asociarlas en un futuro al desarrollo de la hipertensión y complicaciones cardiovasculares. Este estudio se enfoca a encontrar las posibles variantes genéticas que influyen en la aparición de dislipidemia, por lo tanto se presentan como antecedentes la etiología de la dislipidemia, así como estudios de asociación genética previamente realizados con este padecimiento.

## INTRODUCCIÓN

### **Metabolismo de lípidos.**

Los lípidos más abundantes en los humanos son los triglicéridos (TG) y el colesterol total (CT), los cuales tienen como precursor a los ácidos grasos de las vías exógena (dieta) y endógena <sup>9</sup>.

Los TG y el colesterol de la ingesta son transportados en emulsiones estables con ayuda de ácidos biliares para llegar al lumen del intestino, donde la lipasa pancreática los hidroliza en monoacilgliceroles y ácidos grasos. Estos lípidos son absorbidos por los enterocitos, son resintetizados y empaquetados con apolipoproteínas para formar quilomicrones (QM) las cuales son expulsados al sistema linfático. Los TG contenidos dentro de los QM, son hidrolizados por la lipasa lipoproteica (LPL) expresada en las superficies lumbales de las células endoteliales de tejidos vasculares, musculares, adiposos y en el hígado por la lipasa hepática (LH) promoviendo la absorción de los lípidos. Los remanentes de QM son retirados de la circulación por acción del hígado, a través de receptores para lipoproteínas de baja densidad (C-LDL) y en menor proporción por un sistema de receptores denominados LRP-1, que actúan en conjunto con el proteinglicano de superficie celular. Casi todos los TG que son transportados por los QM son utilizados en los tejidos extrahepáticos, mientras que el colesterol restante de los remanentes de QM regresa al hígado <sup>10</sup>.

Por otro lado, la vía endógena comienza en el hígado donde se sintetizan lipoproteínas de muy baja densidad (C-VLDL) que transportan TG y colesterol a los tejidos periféricos, ocasionado por la ingesta de ácidos grasos y/o carbohidratos. Los remanentes que resultan de la hidrólisis de las C-VLDL son capturados por el hígado para su reciclado y una parte cae en la cascada lipolítica de las lipoproteínas VLDL-IDL-LDL.

Las lipoproteínas de densidad intermedia (IDL) son reconocidas e internalizadas por el hígado y la porción restante de estas lipoproteínas sigue su curso en el plasma resultando en las lipoproteínas de baja densidad (C-LDL). Las C-LDL

constituyen los principales transportadores plásmaticos de colesterol y la presencia de la apolipoproteína apo B-100, con sus respectivos receptores de reconocimiento, son fundamentales en el metabolismo del colesterol <sup>10</sup>.

Por último, se encuentra la ruta de retorno, para el transporte del colesterol desde los tejidos periféricos al hígado, mediada por la proteína apo A-I que se empaqueta en las Lipoproteínas de Alta Densidad (C-HDL) e interconecta a las rutas endógena y exógena. La ruta comienza cuando apo A-I recolecta colesterol y fosfolípidos provenientes de tejidos integrándolos en una estructura discoidal C-HDL, la cual es enriquecida con ésteres de colesterol para formar esferas nacientes de C-HDL. Éstas tienen la función principal del intercambio de ésteres de colesterol con las C-LDL que serán recapturadas <sup>11</sup>.

Las partículas lipoproteicas descritas previamente tienen asociadas a un conjunto de apolipoproteínas heterogéneas que se encuentran en distintas proporciones, descritas en la tabla 1. Variaciones en estas proporciones pueden causar el desarrollo de enfermedades cardiovasculares y alteraciones metabólicas como dislipidemia.

Tabla 1. Partículas lipoproteicas con sus respectivas apolipoproteínas asociadas, así como la composición de colesterol y proteína.

Lipoproteína	Apolipoproteína	Diámetro, nm	Colesterol %	% Proteína
C-HDL	A(-I,-II,-IV,-V), D, H	5-15	30	33
C-LDL	B-100, H	18-28	50	25
IDL	E	25-50	29	18
VC-LDL	A-V, B-100, C-II, C-III, H	30-80	22	10
Quilomicrones	B-48, C-I, C-II, C-III, H	100-1000	8	<2

Recuperado de Mclaughlin, C. (2015).

Las rutas metabólicas previamente descritas forman una red inter-dependiente de enzimas asociadas que conectan a las diferentes vías (endógenas, exógenas y de retorno) que se describen en la figura 1.

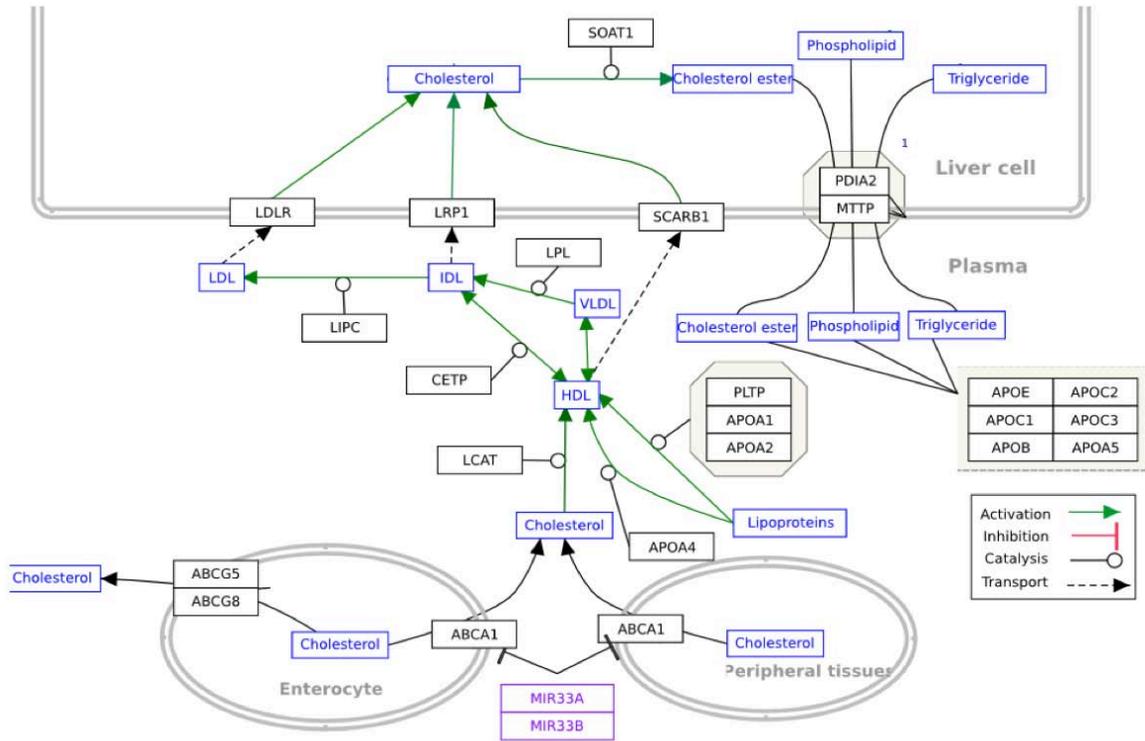


Figura 1. Rutas metabólicas asociadas a lípidos.

En azul esta representado los metabolitos intermediarios, en morado miRNAs que regulan la expresión, en rojo un ejemplo de fármaco usado para el control de niveles lípidos en plasma y en negro las enzimas encargadas de la catálisis, activación o transporte.

Como se observa en la figura 1, que describe parte de las rutas del metabolismo de lípidos, existe un grupo de proteínas implicadas en el transporte y metabolismo de lípidos, para las cuales se ha reportado variación genética importante relacionadas con anomalías a las rutas y con el desarrollo de dislipidemias.

### Dislipidemias

Las dislipidemias se caracterizan por concentraciones anormales de lípidos sanguíneos<sup>1</sup>. Se pueden clasificar según lo propuesto por La Sociedad Mexicana de Endocrinología<sup>12</sup> de acuerdo a las concentraciones plasmáticas de referencia (Tabla 2).

Tabla 2. Clasificación de las dislipidemias según La Sociedad Mexicana de Endocrinología

Dislipidemia	Lípidos (mg/dL)		C-HDL	C-LDL
	CT	TG		
Hipercolesterolemia Aislada	> 200	< 150	-	-
Hipertrigliceridemia Aislada	< 200	> 150	< 150	> 150
Hipoalfalipoproteinemia	-	-	< 50 para mujeres < 40 para hombres	< 50 para mujeres < 40 para hombres
Dislipidemia mixta	-	-	< 130	< 130

El diagnóstico clínico de las dislipidemias se basa en los niveles séricos de las lipoproteínas y lípidos o el depósito de ellos en la piel y tendones, tomándo como referencia la clasificación propuesta por La Sociedad Mexicana de Endocrinología.

### **Etiología de las dislipidemias**

Las causas en el desarrollo de las dislipidemias son variables, teniendo dos clasificaciones: las primarias asociadas a la carga genética y las secundarias al estilo de vida o comorbilidades. Es de relevancia mencionar que la dislipidemia sólo se expresa en presencia de causas secundarias significativas, teniendo que el genotipo determina entre el 40-50% del aporte en la variación total del desarrollo de la dislipidemia<sup>2</sup>.

### **Prevalencia mundial de las dislipidemias**

La dislipidemia más común es la hipercolesterolemia (HTC), caracterizada por un aumento en la concentración plásmatica del colesterol, mayor a 200 mg/dL o 5mmol/L. A nivel mundial, se estima que la HTC esta asociada a un tercio de las cardiopatías isquémicas, además de 2.6 millones de muertes (4.5% del total) <sup>13</sup>.

La prevalencia global de HTC para el 2008 en adultos (mayores a 25 años) fue del 39%, con un foco elevado en Europa (54%) seguido de América (48%). Entre las regiones con porcentajes bajos se encuentran África (22.6%) y Asia Sudoriental (29%). Específicamente en México la prevalencia fue de 50.7% que es mayor al promedio de América y tan sólo 4 puntos por debajo de Europa como se observa en la figura 3.

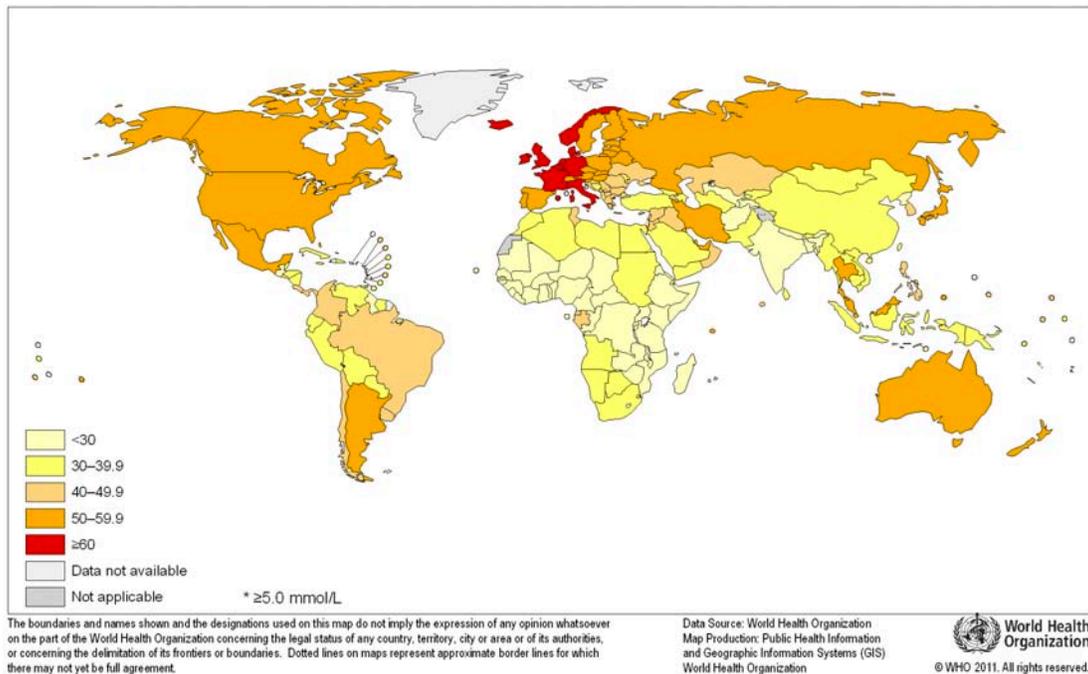


Figura 2. Porcentaje de la prevalencia mundial de altas concentraciones de colesterol (hipercolesterolemia) en adultos (>25 años) de ambos sexos en el 2008.

### Prevalencia en México de las dislipidemias

En la Encuesta Nacional de Salud de Medio Camino del 2016, se reportó una prevalencia de hipercolesterolemia en la población general de 44.7%. De estos, 28% informaron haber recibido un diagnóstico médico previo de hipercolesterolemia. La prevalencia más elevada por diagnóstico médico se encuentra en el grupo poblacional de 50 a 79 años de edad (Tabla 3).

Tabla 3. Porcentaje de adultos con hipercolesterolemia por diagnóstico previo.

Edad (años)	Hombres		Mujeres		Total	
	N (miles)	%	N (miles)	%	N (miles)	%
20 a 29	260	11	475.2	16.3	735.2	13.9
30 a 39	926.8	25.2	868.2	25	1795	25.1
40 a 49	926.6	36	912.8	24.1	1839.4	28.9
50 a 59	687.4	35.4	1216.4	40.1	1903.8	38.3
60 a 69	571.4	31.9	891.2	39.5	1462.5	36.1
70 a 79	223.9	32.5	320.6	33.7	544.5	33.2
80 y más	55.3	18.5	88.1	23.3	143.5	21.2
Total	3651.3	27.4	4772.6	28.4	8423.9	28

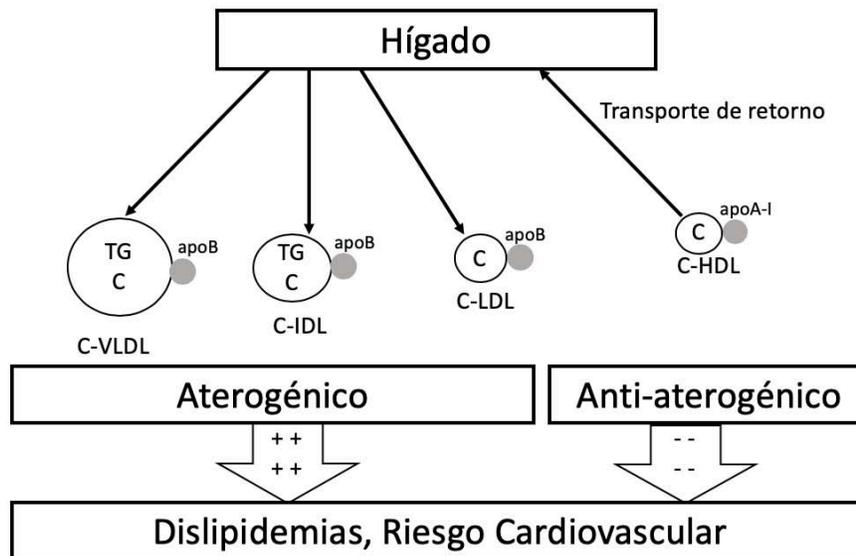
## Índices Aterogénicos

Para optimizar la capacidad predictiva del perfil lipídico, se han desarrollado algunos índices aterogénicos o lipoproteícos, que proveen información sobre factores de riesgo a dislipidemia y arteriosclerosis, que aunque son difíciles de definir y diagnosticar por análisis de rutina pueden ser usados como reflejo de las interacciones metabólicas-clínicas entre las fracciones plasmáticas de lípidos y definen el riesgo cardiovascular y de alteración metabólica: dislipidemia.

- Índice de Castelli I =  $CT/HDL$
- Índice de Castelli II =  $C - LDL/C - HDL$
- Índice Aterogénico de Plasma =  $\log(TG/C - HDL)$

Los índices de Castelli I y II son muy similares entre sí pues aproximadamente dos terceras partes de colesterol en plasma se encuentran como C-LDL, consecuentemente, el colesterol total y el C-LDL están relacionados cuantitativamente. En este sentido, cuando un paciente tiene un incremento de CT, específicamente C-LDL, es un marcador aterógeno; mientras que, el reducir la concentración plasmática de C-HDL se correlaciona con factores de riesgo incluyendo dislipidemias o enfermedad arterial coronaria (Figura 3).

La aplicabilidad clínica de estos índices se aprecia al comparar valores de CT o C-HDL, contra los índices aterogénicos. Sólo los índices son capaces de distinguir entre las personas aparentemente sanas y sobrevivientes de infartos al miocardio.



Modificado de Millán, J. et al (2009).

Figura 3. Riesgos aterogénicos asociados a las lipoproteínas.

En la figura 3 se observa que la apolipoproteína B (apoB) se encuentra en las lipoproteínas de muy baja densidad (C-VLDL), densidad intermedia (C-IDL) y lipoproteínas de baja densidad (C-LDL), las cuales son de riesgo aterogénico, mientras que la apolipoproteína A-I (apo A-I) se asocia a la lipoproteína de alta densidad (C-HDL), responsable de iniciar el transporte inverso del colesterol y por tanto se considera antiaterogénico. El equilibrio entre apoB y apoA-I es indicativo de riesgo cardiovascular: cuanto mayor es la proporción, mayor es el riesgo (ídem).

Se ha observado que el predominio de partículas pequeñas de C-LDL por sobre partículas grandes de C-HDL está relacionada con un aumento en el riesgo aterogénico; por lo que, medir la distribución del tamaño de partícula de C-LDL o C-HDL mediante electroforesis en gel, ha sido necesario para obtener una estimación más precisa del riesgo aterogénico (ídem).

Por otro lado, Dobiášová et al. encontraron una correlación positiva ( $r = 0.803$ ) entre el índice aterogénico de plasma (AIP) y la rapidez de esterificación de

colesterol en plasma sin lipoproteína apoB ( $FER_{HDL}$ ), la cual a su vez refleja la distribución que existe entre en C-HDL y C-LDL, estableciendo la relación entre la distribución de C-LDL y C-HDL y AIP <sup>15</sup>.

Finalmente, cada índice tiene valores límite que apoyan a la determinación del riesgo aterogénico como la dislipidemia (Tabla 4).

Tabla 4. Límites de riesgo para los índices aterogénicos estratificados por género.

Índice	Riesgo aterogénico (paciente dislipidémico)	
	Hombres	Mujeres
Índice de Castelli I	> 4.0	> 3.5
Índice de Castelli II	> 3.0	> 2.5
AIP	>0.5 para hombres o mujeres	

Modificado de Millán, J. et al (2009).

### **Marcadores genéticos de dislipidemia y riesgo aterogénico: estudios de asociación.**

De manera general, se pueden clasificar a los estudios de asociación genotipo-fenotipo en dos tipos: aquellos que asocian el genoma completo (GWAS) y los de gen candidato. Los primeros no tienen una suposición previa, por lo que no delimitan el estudio a cierto número de variantes genéticas y requieren un análisis computacional exhaustivo que permita localizar regiones cromosomales asociadas al fenotipo <sup>16</sup>, la única limitante es el número de marcadores analizados de acuerdo a la tecnología utilizada.

La aproximación por gen candidato racionaliza al problema de estudio, estableciendo que la mayor variación genética resulta de cambios en un conjunto de genes. Este estudio define al gen como una función biológica conocida que participa directa o indirectamente regulando la aparición del fenotipo. Además, establece que dicho efecto puede ser evaluado usando sus variantes en análisis de asociación.

Cabe mencionar que las asociaciones genotipo-fenotipo encontradas por un estudio GWAS o gen candidato, como las de este proyecto, no pueden ser

trasladadas de manera directa a la clínica, ya que las variantes genéticas de un locus no son directamente informativas con respecto al mecanismo por el cual esta asociado con un fenotipo, es decir se identifica una asociación entre una variante genética y un fenotipo determinado lo cual no implica causalidad.

### **Estudios genéticos de las dislipidemias**

Existen varios factores asociados al desarrollo de la dislipidemia, se estima que la genética puede determinar hasta el 40-50% de los niveles de lípidos en plasma. Por lo tanto, estudios de genoma completo, *GWAS*, se han dedicado a identificar variantes en genes que contribuyan a explicar y predecir la variación inter-individual en la concentración plasmática de lípidos <sup>17</sup>. A continuación algunos estudios relevantes:

Las investigaciones de Teslovich et al., que marcan un punto de referencia importante, identifican 95 loci genéticos asociados a concentraciones de lípidos en plasma en personas con enfermedad arterial coronaria y se comparan con controles en diversas poblaciones del mundo <sup>3</sup>. El conjunto de variantes reportado es una pauta para el desarrollo de estudios de replicación o validación en poblaciones específicas. Por ejemplo, el grupo de Willer et al. logró refinar al conjunto de los primeros 95 loci, encontrando 62 nuevos asociados a los 4 principales lípidos plasmáticos (CT, TG, HDL y LDL) en 188 mil europeos en un estudio tipo *GWAS* <sup>18</sup>.

Consecuentemente, se generaron estudios para replicar dichas variantes asociadas en europeos con niveles de lípidos, pero en la población mexicana. En 2013 un grupo de investigadores en colaboración con el Instituto Nacional de Nutrición en México, fue capaz de replicar las variantes *APOA5* rs964184, *GCKR* rs1260326 y *LPL* rs12678919 asociadas a hipertrigliceridemia en mexicanos <sup>6</sup>. Es lógico esperar que no todos los 95 loci fueron replicados y que líderes en la materia insistan en que urge realizar estudios tipo GWAs para identificar que variantes en mexicanos pueden predecir las diferentes dislipidemias <sup>19</sup>.

Por lo tanto, es evidente que los factores genéticos que determinan los niveles de lípidos en plasma aun no han sido identificados del todo para la población mexicana. Este estudio busca contribuir al listado de genes y sus variantes que sirvan de marcadores genéticos asociados a: niveles de lípidos, tipos de dislipidemia e índices aterogénicos. Es importante mencionar que estudios tipo *GWAS* a nivel mundial son realizados en grupos de cientos y miles de individuos y aquí se cuenta con 426 sujetos de estudio, por lo que se recurrió a una estrategia de gen candidato para responder a la pregunta de investigación.

Dado el éxito de los *GWAS* para identificar marcadores genéticos, este proyecto tomó las estrategias metodológicas de dichos estudios incluyendo: rigurosos controles de calidad en cuanto a genotipos y fenotipos publicados como guía de análisis previamente <sup>4</sup>.

Finalmente, se asociaron variantes seleccionadas de genes candidato, pero esta lista fue enriquecida tomando en cuenta las variantes de una lista de 669672 cuya frecuencia alélica difirió significativamente entre casos (dislipidémicos) y controles (sanos) utilizando una prueba Chi-cuadrada. El análisis de gen candidato con la lista ya enriquecida tomo en cuanta además, variables fisiológicas, bioquímicas y el componente ancestral individual.

## **OBJETIVOS**

### **General**

- Identificar variantes genéticas asociadas a marcadores de riesgo aterogénico (dislipidemias) de una cohorte de mexicanos mestizos (Tlalpan 2020) utilizando microarreglos.

### **Específicos**

- Investigar y contribuir al listado de variantes genéticas asociadas a niveles de lípidos, tipo de dislipidemia e índices aterogénicos en población mexicana.
- Desarrollar modelos de regresión condicionales que permitan encontrar variantes genéticas con efecto independiente asociadas a índices aterogénicos.

## **JUSTIFICACIÓN**

La dislipidemia es una enfermedad causada por múltiples factores y se caracteriza por el metabolismo anormal de lípidos. Existen cuatro tipos de dislipidemia, la más prevalente es la hipercolesterolemia seguida de la dislipidemia mixta, la hipoalfalipoproteinemia y la hipertrigliceridemia aislada. La hipercolesterolemia es responsable de un tercio de las cardiopatías y de 2.6 millones de muertes a nivel mundial. Se ha demostrado que el componente genético puede explicar hasta el 50% de la variación multifactorial en los niveles de lípidos circulantes y que la población mexicana, así como otras con descendencia amerindia, tienen una predisposición significativa en comparación con otras poblaciones, por lo que es relevante conocer que variantes genéticas pueden explicar dicha susceptibilidad.

## **HIPÓTESIS**

$H_0$ = Las variaciones en índices aterogénicos en la cohorte Tlalpan 2020 no se asocian a la lista de variantes contenidas en los genes candidato.

$H_1$ = Las variaciones en índices aterogénicos en la cohorte Tlalpan 2020 se asocian a la lista de variantes contenidas en los genes candidato.

# METODOLOGÍA

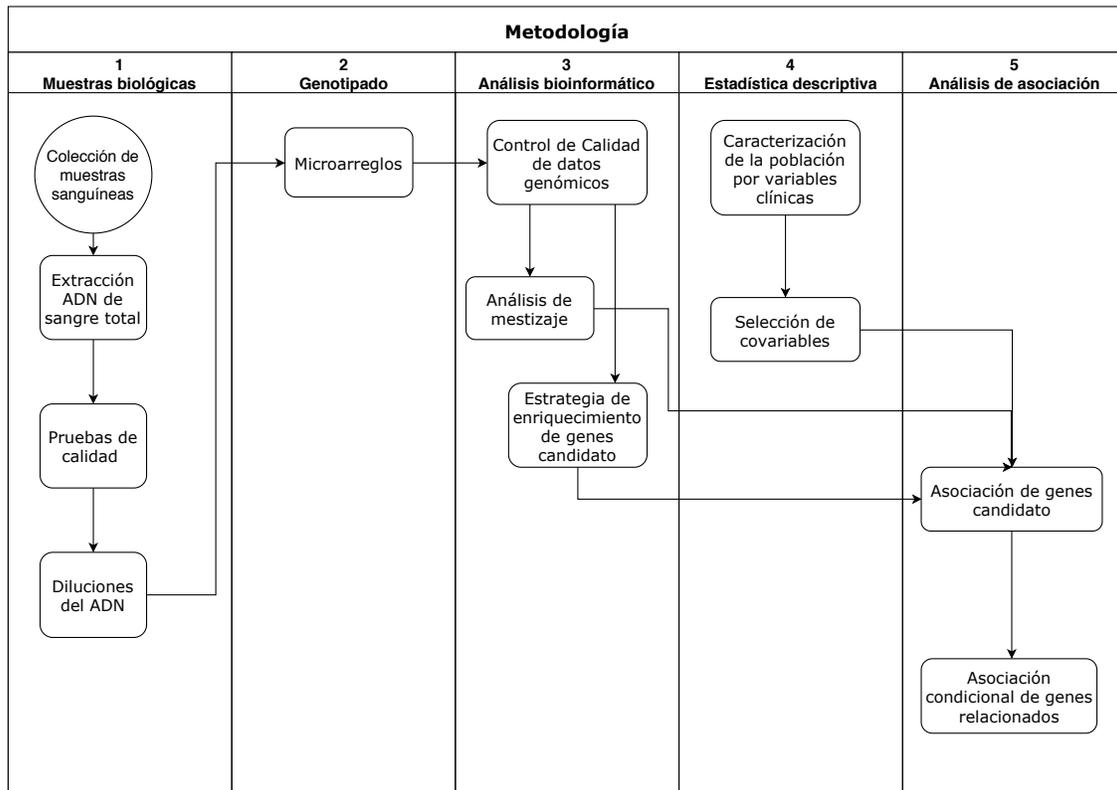


Figura 4. Estrategia propuesta para abordar los objetivos propuestos

## Muestras Biológicas

### COLECCIÓN DE MUESTRAS SANGUÍNEAS

Se recolectó una submuestra del Proyecto Tlalpan 2020 del Instituto Nacional de Cardiología, conformada por voluntarios sanos y con algún tipo de dislipidemia, provenientes de la CDMX entre 2013 y 2014. Este proyecto incluye 124 controles y 316 individuos con dislipidemia, manteniendo las proporciones originales de la cohorte de 30% individuos sanos (controles) y 70% individuos con dislipidemia (casos). Los voluntarios acudieron semestralmente al Instituto Nacional de Cardiología para mediciones de presión arterial, peso, cintura y química sanguínea. El protocolo fue aprobado por los comités de ética e investigación del INMEGEN y del Instituto Nacional de Cardiología.

Criterios de inclusión de la muestra.

- Voluntarios que aceptaron donar sangre por un consentimiento firmado.
- Controles: Adultos entre 20 y 50 años, de ambos sexos, sin hipertensión (HTA) y/o diabetes.
- Casos: Adultos entre 20 y 50 años, de ambos sexos, diagnosticados con algún tipo de dislipidemia por el Instituto Nacional de Cardiología, tomando en cuenta los niveles lipídicos en sangre propuestos por La Sociedad Mexicana de Endocrinología descritos en la Tabla 2.

Criterios de exclusión de la muestra.

- Pacientes incapaces de donar una muestra sanguínea.

#### **EXTRACCIÓN DE ADN**

Durante una de las visitas al Instituto Nacional de Cardiología los voluntarios donaron una muestra de sangre en tubos vacutainer-EDTA, los cuales fueron centrifugados 3000 rpm por 10min, se removio el plasma y el paquete celular fue congelado a  $-70^{\circ}\text{C}$  y enviado al INMEGEN, en las siguientes 4 semanas posteriores a su extracción. El ADN fue extraído usando el kit comercial PureGen Blood Kit (Qiagen, Valencia, Estados Unidos) de acuerdo a las instrucciones del proveedor con algunas adaptaciones para eficientar el procesamiento de muestra. El protocolo modificado se describe a continuación:

- El paquete celular (globulos rojos y blancos) se transfirió a un tubo cónico de 15 mL y se agregaron 10 mL buffer de EL (buffer de lisis de eritrocitos).
- Agitación en vortex a velocidad media por 1 min hasta observar homogeneidad en la mezcla.
- Incubación por 10 min a temperatura ambiente.
- Centrifugación a temperatura ambiente ( $25^{\circ}\text{C}$ ) por 15 min a 3500 rpm y se descartó el sobrenadante.
- Resuspensión de la pastilla o sedimento (pellet) en buffer de lisis celular y se incubó por 1 h a  $37^{\circ}\text{C}$ .

- Al lisado celular se adicionó 1 mL de solución de precipitación de proteínas y se homogeneizó en vortex a velocidad media por 1 min e inmediatamente se centrifugó a temperatura ambiente por 15 min a 4000 rpm.
- Sin eliminar el sobrenadante, se agregaron 800  $\mu$ L de solución de precipitación de proteínas.
- Disgregación del pellet previo a una segunda centrifugación usando las condiciones anteriores.
- El sobrenadante se transfirió a un nuevo tubo falcón, donde se adicionaron 5 mL de isopropanol frío (4°C) y se homogeneizó por inmersión.
- Incubación 15 min a 4°C y se centrifugó a temperatura ambiente por 15 min a 4000 rpm.
- Se descartó el sobrenadante y se realizó un lavado con etanol al 70% (v/v) a 4°C, homogenizando completamente la mezcla en un vortex a velocidad media.
- Centrifugación a temperatura ambiente por 15 min a 3500 rpm.
- El sobrenadante se descartó y el pellet se dejó secar a temperatura ambiente hasta evaporarse el etanol.
- Resuspensión del pellet en 200  $\mu$ L de agua libre de nucleasas y se almacenó a -20°C.

#### **CONTROL DE DALIDAD DEL ADN: CUANTIFICACIÓN, EVALUACION DE PUREZA E INTEGRIDAD**

La calidad del ADN se evaluó con el fin de asegurar el correcto genotipado de las mismas. Por lo anterior, se llevó acabo el siguiente protocolo:

- El ADN fue cuantificado y se determinó la calidad por espectroscopia UV usando un NanoDrop One (Thermo Scientific, Maddison, Estados Unidos).
- La integridad del ADN se evaluó por electroforesis en geles de agarosa al 2% con buffer *TBE* (p/v), verificando el corrimiento de las bandas en un fotodocumentador (BioRad, California, Estados Unidos)

Aquellas muestras que tuvieran un índice de absorbancia 260/280 nm de al menos 1.8 y 230/260 nm de al menos 2.2, fueron consideradas para ser genotipadas.

## **Genotipado**

### **GENOTIPIFICACIÓN POR MICROARREGLOS**

El ADN se diluyó en agua libre de nucleasas a una concentración final de 50 ng/μL en un volumen total de 50 μL, para su posterior genotipificación por microarreglos.

Se usó el microarreglo “Custom” *Infinium Global Screening Array-24 v1.0* de la compañía Illumina, adaptado con variantes elegidas por varios grupos de investigación del INMEGEN. Este microarreglo permite el análisis de 642,824 variantes representativas de múltiples etnias, informativas de ancestría, para pruebas farmacogenómicas, relacionadas a enfermedades cardiovasculares,<sup>20</sup>. El arreglo comercial incluye un total de 642,824 variantes (Illumina, n.d.) adicionales a las 26,848 variantes seleccionadas por el INMEGEN. El microarreglo personalizado tiene propiedad intelectual INMEGEN-Código 46, compañía que procesó los microarreglos.

## Análisis bioinformático

### CONTROL DE CALIDAD DE DATOS GENÓMICOS

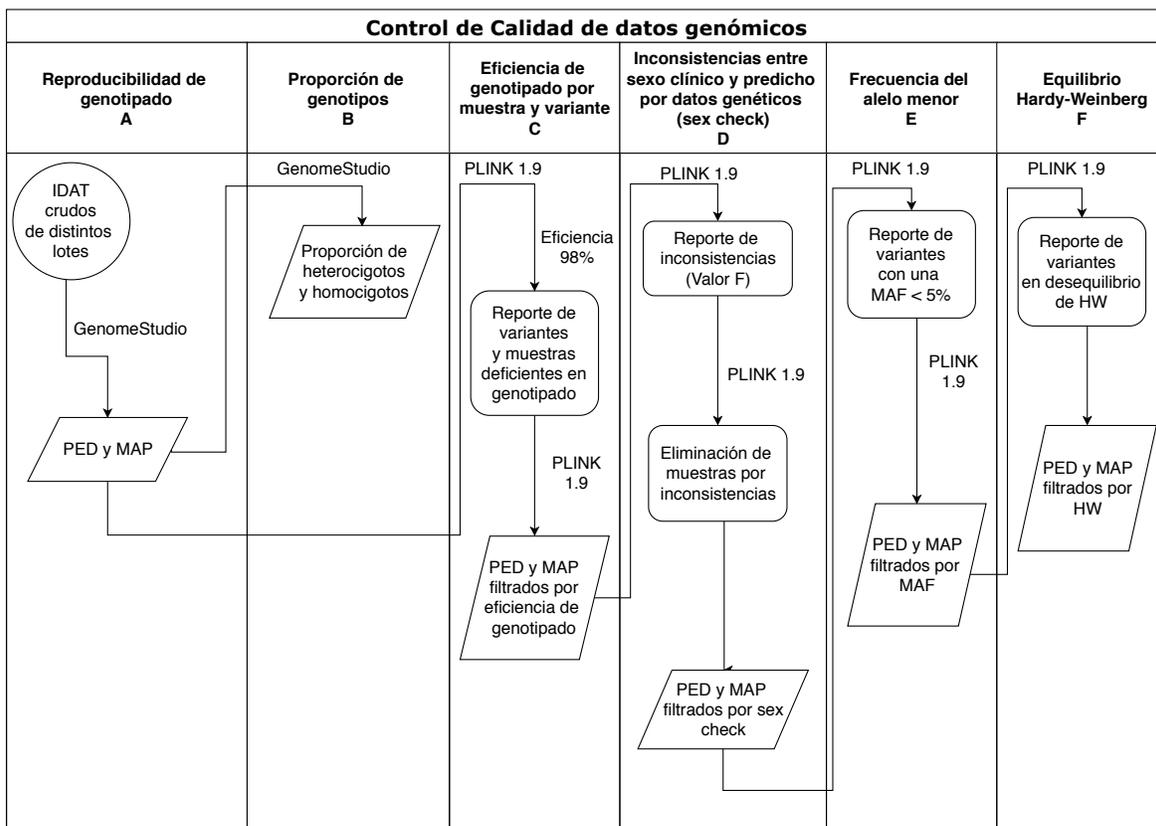


Figura 5. Diagrama de flujo del control de calidad de datos genómicos.

Este estudio fue uno de los primeros en utilizar el microarreglo personalizado, por lo tanto, era necesario validar que el llamado de variantes fuera consistente y homogéneo para lo cual se corrieron los incisos A y B (figura 5) para muestras control. Los incisos siguientes, del C al F, forman parte del control de calidad previo a cualquier asociación genotipo-fenotipo (GWAS).

#### **CONTROL DE REPRODUCIBILIDAD DE GENOTIPADO POR MICROARREGLOS (A)**

El microarreglo personalizado fue validado con muestras intra e inter-lotes empleando el módulo *Genotyping* de *GenomeStudio* <sup>22</sup>, usando el formato *samplesheet* (consta de los identificadores de las muestras), el Manifiesto (contiene el orden específico de las sondas de cada variante dentro del microarreglo) y los datos crudos de fluorescencia en formato .IDAT. De acuerdo al

protocolo de Illumina, se utilizó el algoritmo *GenCall* para obtener los genotipos procesados a partir de los datos crudos. Una vez hecha la genotipificación, se agruparon los SNPs, al seleccionar la opción *ClusterAllSNPs* dentro de *GenomeStudio*, usando un umbral de 0.15 para el llamado de variantes.

Los datos de genotipificación por fluorescencia se traducen a archivos de texto plano PED y MAP, mediante el *plugin* de *PLINK 1.9*. Input Report versión 2.1.4 (*GenomeStudio*, Illumina, California, Estados Unidos). Estos archivos de texto se compararon con los generados en el laboratorio y los entregados por la empresa Código 46, evaluando la reproducibilidad de los datos al no encontrar diferencias significativas entre el genotipado.

Los datos genotipados en formatos crudos PED son archivos de texto plano separado por tabuladores, que tienen la descripción de cada muestra en seis columnas. La primera identifica al individuo asignando un número que no debe repetirse, la siguiente identifica a la familia por lo que si existen hermanos se asigna un mismo número para el conjunto de individuos relacionados y las siguientes dos columnas identifican a los padres del individuo. Las últimas dos columnas asignan sexo (1 para hombre, 2 para mujer) y fenotipo, que puede ser casos (2), controles (1) o valores continuos. Las columnas siguientes contienen los genotipos de los individuos, ordenados según la matriz transpuesta de los archivos MAP.

Los archivos MAP de texto plano separado por tabuladores, identifican a las variantes genéticas en 4 columnas. La primera asigna el cromosoma, la siguiente el identificador rs de la variante, a continuación la distancia genética entre variantes en centimorgans y por último la posición en el cromosoma en pares de base.

#### ***REPRODUCIBILIDAD DE GENOTIPADO ENTRE REPLICAS INTRA E INTER-LOTES (A)***

La reproducibilidad del análisis de microarreglos fue evaluada incorporando replicas intra e inter-lotes, calculando la proporción de diferencias de genotipos entre replicas usando una prueba de *t-Student*. Es decir, se compararon los

genotipos entre muestras intra e inter-lotes con sus respectivas replicas usando los archivos PED generados con *GenomeStudio*.

#### ***PROPORCIÓN DE GENOTIPOS: HOMOCIGOTOS Y HETEROCIGOTOS (B)***

Como método adicional, en *GenomeStudio* se analizaron las proporciones de genotipos: homocigotos y heterocigotos en las muestras tomadas de diferentes lotes, ya que microarreglos de poca calidad se caracterizan por distribuciones aberrantes de intensidad, exceso de variantes no caracterizadas, perdidas o heterocigotas, o ambas <sup>23</sup>.

#### ***EFICIENCIA DE GENOTIPADO POR MUESTRA (C)***

Fallas en la eficiencia del llamado de variantes (*call rate*) puede ser indicativo de una calidad pobre en la muestra de DNA o fallas técnicas durante el proceso de genotipificación. Esta mala eficiencia puede llevar a un genotipado aberrante y por lo tanto muestras con posibles fallas deben ser descartadas. Es recomendable usar un límite entre el 98-99% en la eficiencia del genotipado por muestra <sup>4</sup>.

#### ***EFICIENCIA DE GENOTIPADO POR VARIANTE (SNP) (C)***

Al igual que la anterior, el llamado de variantes por muestra, se recomienda un límite del 98-99% para el caso de las variantes. Para llevar acabo el *call rate* de muestras y variantes se utilizó el comando *--missing* de *PLINK 1.9*, el cual genera un archivo que lista la proporción de variantes (SNPs) faltantes de genotipo para cada individuo y la proporción de individuos faltantes para cada variante. Aquellas muestras y/o variantes que no cumplieron con este criterio fueron descartadas.

#### ***FILTRADO DE MUESTRAS POR INCONSISTENCIAS ENTRE SEXO CLÍNICO Y EL PREDICHO POR LOS DATOS GENÉTICOS (D)***

Una vez validada la confiabilidad de genotipificación con *GenomeStudio*, además de descartar muestras y variantes por la eficiencia del genotipado, se identificaron problemas de identidad por mal manejo de muestreo. Uno de los métodos más sencillos para corroborar que no existe error por mezclado de muestras, es mediante el chequeo del sexo reportado por la base de datos clínica contra el sexo predicho por los datos genéticos.

El software de libre acceso *PLINK 1.9*, se basa en las proporciones de heterocigosidad del cromosoma X (valor F) para determinar de manera empírica el sexo y reporta los casos en los que el sexo no coincide, al estimar un valor de F para cada muestra (mujeres  $F < 0.2$  y hombres  $F > 0.8$ ) el cual se corrobora con el sexo biológico reportado por el medico.

#### ***FILTRADO DE VARIANTES SEGÚN LA FRECUENCIA DEL ALELO MENOR (MAF) (E)***

Es importante filtrar variantes genotipadas en baja proporción, basándose en la frecuencia del alelo menor (*MAF*) ya que presentan un poder estadístico muy bajo en aquellas variantes extremadamente raras (<1% de frecuencia) al tener una muestra pequeña. El límite depende del tamaño de la muestra y el efecto del tamaño esperado, en este estudio se eliminaron SNPs con una *MAF* < 5%.

La *MAF* se obtuvo para cada SNP usando la opción *--freq* de *PLINK 1.9*. Posteriormente, los SNPs se removieron del archivo PED usando la opción *--maf* seguida del límite (5%). Es importante recalcar que, en estudios con una muestra muy grande se recomienda usar una *MAF* < 1%.

#### ***EQUILIBRIO DE HARDY-WEINBERG (HW) (F)***

Verificar el equilibrio de Hardy-Weinberg (HW) es uno de los pasos finales en el control de calidad de las variantes genéticas previo a un análisis de asociación. El equilibrio de Hardy-Weinberg, supone que las frecuencias de alelos y genotipos se mantienen constantes entre generaciones. Cuando una variante está fuera del equilibrio puede ser indicativo de errores de genotipificación, estratificación de la población o incluso asociación real con el rasgo en estudio.

Previo a las pruebas de asociación con el fenotipo, antes de eliminar variantes que no cumplan con la prueba de HW, se debe verificar si este desequilibrio es verdadero o es un error de genotipado utilizando el comando *--hardy* ( $P < 1E-5$ ) de *PLINK 1.9*, el cual determina el grado de desequilibrio de HW para cada SNP. Un valor de  $P < 1E-5$  asociado a la prueba de HW indica que el desequilibrio puede no ser verdadero y por tanto la variante no es considerada en los análisis posteriores.

## **Análisis de mestizaje**

La proporción del componente ancestral es comúnmente usada en la corrección de análisis de asociación genética al ser usada como covariable, ya que las muestras de estudio pueden pertenecer a individuos relacionados y por tanto resultar en falsos positivos que ocultan las asociaciones reales.

- Para determinar el mestizaje se usaron los archivos MAP y PED con la herramienta *PLINK 1.9* usando la intersección de datos genómicos de voluntarios de Tlalpan 2020, Nativos Mexicanos de:
  - Durango
  - Michoacán
  - Oaxaca

y de 1000 genomas para poblaciones:

- Chinos Han de Beijing (CHB)
  - Europeos del Norte de Utah (CEU)
  - Yoruba en Ibadan de Nigeria (YRI)
- 
- Se filtraron las variantes genéticas tomando aquellas que no estuvieran en desequilibrio de ligamiento (LD) usando una  $R=0.5$  y aplicando los filtros de calidad previamente descritos usando *PLINK 1.9*.
  - Se aseguró que las muestras no estuvieran relacionadas al menos en primer grado usando el factor de  $P_{\hat{}}$  (coeficiente de relación parental) con el comando *--genotype* de *PLINK 1.9*.
  - Los archivos PED y MAP filtrados se cargaron en *PLINK 1.9* y se generaron los componentes principales (PC) con el comando *--pca*.
  - Los archivos PED y MAP en formato binario se cargaron en *ADMIXTURE* para estimar las fracciones ancestrales.
  - Se realizó una validación cruzada usando un intervalo del número probable de poblaciones ancestrales con el comando *--cv*, para encontrar el menor error asociado.

- *ADMIXTURE* da un archivo con las fracciones ancestrales para cada muestra.
- Las fracciones ancestrales para cada muestra serán usadas en los modelos de asociación como covariantes.

### ESTRATEGIA DE ENRIQUECIMIENTO DE GEN CANDIDATO

Previo a la asociación entre dislipidemias y variantes genéticas se realizó una exhaustiva investigación bibliográfica generando una lista de 100 genes candidato los cuales se eligieron tomando como base otros de estudios que incluyeran cualquier la asociación genética con niveles de lipidos y dislipidemias <sup>3,5-7</sup>.

Esta lista, se enriqueció considerando variantes/genes de todo el microarreglo (669 456) que mostraron diferencias significativas en frecuencias alélica entre los casos y controles. Las comparaciones se realizaron entre los grupos de diferentes dislipidemias (casos) y estratificando por género, como se muestra en la figura 6.



Figura 6. Representación de los estratos para la comparación de la frecuencia alélica. Se dividió por género (hombres y mujeres), por tipo de dislipidemia (hipercolesterolemia, hipertrigliceridemia, hipoalfalipoproteinemia y dislipidemia mixta) y por genero-tipo de dislipidemia.

Las variantes significativas fueron determinadas utilizando PLINK basado en el análisis de tablas de contingencia y la prueba de Chi-cuadrada ( $X^2$ ) El punto de corte de significancia fue de  $P < 1E-6$  <sup>24</sup>.

$$X^2 = \sum \frac{(\text{observados} - \text{esperados})^2}{\text{esperados}}$$

## Estadística descriptiva

### CARACTERIZACIÓN DE LA POBLACIÓN

Los datos clínicos y demográficos fueron provistos por el equipo de trabajo del Proyecto Tlalpan 2020 del Instituto Nacional de Cardiología, los cuales incluyen:

- Edad, años
- Peso, kg
- Estatura, m
- IMC, kg/m<sup>2</sup>
- Cintura, cm
- Glucosa, mg/dL
- Ácido úrico, mg/dL
- Creatinina, mg/dL
- Colesterol total (TC), mg/dL
- Lipoproteínas de alta (C-HDL) y baja densidad (C-LDL), mg/dL
- Triglicéridos (TG), mg/dL

Además, como variables de respuesta que reflejan dislipidemia se calculó:

- Índice de Castelli II (IC II)

$$IC II = (LDL/HDL)$$

- Índice Aterogénico de Plasma (AIP)

$$AIP = \left( \log(TG/HDL) \right)$$

## SELECCIÓN DE COVARIANTES

En modelos de asociación genómica se consideran fenotipos confusores (covariantes) de tal forma que, el modelo tome en cuenta el mayor número de variantes clínicas y genéticas e identifique la asociación genotipo-fenotipo más precisa a la situación real del paciente. Es importante prevenir efectos de multicolinealidad al incluir en el modelo covariantes correlacionadas que no son independientes como es el caso del peso y el IMC <sup>25</sup>.

- La colinealidad entre covariantes se determinó a través del coeficiente de correlación de Pearson, comparando por pares todas las variables para determinar el grado de independencia usando la paquetería de Python *seaborn*.
- A partir de un coeficiente de Pearson de 0.3 se consideraron correlacionadas no independientes y solo una de ellas es incluida en el modelo.
- Para decidir que covariante se iba a seleccionar del conjunto de variables colineales, se analizaron por componentes principales. Este método disminuye la complejidad de datos multidimensionales, mediante la transformación de las covariantes en vectores propios (eigenvector) con el comando *prcomp*.
- Se tomaron las covariantes con mayor variabilidad, ya que son las que más información aportan a los modelos de asociación, usando los valores absolutos de los scores de cada covariante en el primer componente principal.
- Las covariantes elegidas formaron parte de la corrección de los modelos de asociación posteriores.

## **Análisis de asociación**

Nuestra estrategia consta de dos etapas

1. Regresión multivariable lineal (variables continuas) y logística (variable discretas o categóricas).
2. Regresión condicional de componentes principales utilizando variantes asociadas significativamente de la etapa uno.

### **MODELOS DE ASOCIACIÓN DE GEN CANDIDATO (PRIMER ETAPA)**

Como primer paso se identificaron variantes genéticas asociadas a diferentes variables de respuesta, para lo cual se desarrollaron regresiones lineales y logísticas múltiples corregidas por las siguientes covariantes:

- Ácido úrico
- Peso
- Cintura
- Sexo
- Fracción ancestral: CEU y YRI

para los siguientes fenotipos:

- Lineal (variable continua)
  - Índices Aterogénicos (AIP e IC II)
  - Niveles lipídicos (TC, TG, C-HDL, C-LDL)
- Logístico (casos vs controles)
  - Tipo de dislipidemia (HTC, HTG, DM, HA vs controles)

De manera genérica los modelos de regresión se aplicaban para cada variante genética (SNP) como se presenta a continuación:

$$\begin{aligned} \text{Fenotipo} = & \beta_0 + \beta_1 \text{SNP}_{\text{prueba}} + \beta_1 \text{Covariable}_1 \\ & + \beta_2 \text{Covariable}_2 + \dots + \beta_n \text{Covariable}_n \end{aligned}$$

Donde

$\beta_0$  es el valor de la intersección lineal

$\beta_1$  es el efecto o cofactor del SNP de prueba

$\beta_2, \beta_3, \dots, \beta_n$  son el tamaño del efecto de cada covariante clínica o genética

Es importante mencionar que este método no considera que los tamaños del efecto ( $\beta_2, \beta_3, \dots, \beta_n$ ) de cada SNP sean independientes respecto a otros SNPs cercanos o en LD.

- En el caso de los análisis tipo casos (1) y controles (0), se realizó una regresión logística para cada tipo de dislipidemia (hipercolesterolemia, hipertrigliceridemia, hipoalfalipoproteinemia y mixta) como variable de respuesta usando *PLINK 1.9* con el comando *--logistic*.
- Los archivos de datos genómicos MAP y PED generados con la lista de genes candidato, además de los archivos *.txt* con las covariantes y fenotipos, usando los comandos *--covar* y *--pheno*.
- Se filtraron los SNPs con punto de corte  $P < 0.05$ .
- También se desarrollaron regresiones lineales donde la variable de respuesta son los valores lipídicos de plasma (TC, TG, C-HDL y C-LDL) e índices aterogénicos (IC II y AIP) y las covariantes seleccionadas previamente con el comando *--linear* en *PLINK 1.9*.
- Se compararon los SNPs asociados y los valores de P obtenidos para cada regresión y se reportaron aquellos significativos para cada fenotipo ( $P < 1E-4$ ).

#### **MODELOS DE ASOCIACIÓN CONDICIONAL DE GENES RELACIONADOS (REGRESIÓN DE COMPONENTES PRINCIPALES: SEGUNDA ETAPA)**

Las asociaciones condicionales buscan eliminar variantes de la etapa anterior que no sean independientes. Para lo cual, los SNPs obtenidos como significativos de las regresiones lineales ( $P < 0.05$ ) se sometieron a un análisis adicional en el cual se analizaba que el tamaño del efecto no dependiera de las demás variantes

genéticas y esto permite evaluar el efecto de cada SNP independientemente de las variantes que tenga asociadas por LD o cercanas <sup>26</sup>.

*Justificación de la segunda etapa de regresión condicional.*

SNPs significativos cercanos entre sí pueden estar en LD lo cual impide discernir entre aquellos asociados con el fenotipo, que tienen una posible relación causal, y aquellos que estén en LD que muestran asociación. Además, los SNPs en LD tienen efectos de multicolinealidad generando inestabilidad en los coeficientes ( $\beta$ ), por lo que se eligió utilizar una regresión condicional de componentes principales las cuales se han utilizado en estudios previos tomando fenotipos como paratuberculosis en bovinos <sup>27</sup> y Alzheimer <sup>28</sup> usando una cohorte de hasta 300 individuos.

La ecuación genérica de la regresión condicional de componentes principales es la siguiente:

$$\begin{aligned} \text{Fenotipo} = & \beta_0 + \beta_1 \text{SNP}_{\text{prueba}} + \beta_1 \text{SNP}_1 + \beta_1 \text{Covariable}_1 + \dots + \beta_n \text{SNP}_n \\ & + \beta_n \text{Covariable}_n \end{aligned}$$

Donde

$\beta_0$  es la intersección con el eje de las ordenadas

$\beta_1$  es el efecto o cofactor del SNP de prueba

$\beta_1, \beta_2, \dots, \beta_n$  son los efectos o cofactores de las covariantes

$\text{SNP}_{\text{prueba}}$  al cual se evalúa el efecto independiente

$\text{SNP}_n$  en el mismo cromosoma y que sirven como covariantes

- Como variable de respuesta se usaron los índices aterogénicos en asociaciones independientes, pues tienen un nivel de predicción mayor a dislipidemias en comparación con los niveles lipídicos en sangre. Además, las variantes asociadas en la primera etapa tenían un nivel de significancia

mayor hasta por dos logaritmos con respecto a los demás fenotipos evaluados en comparación con los índices aterogénicos.

- El filtrado de variantes (valor de  $P < 0.05$ ) de la primera etapa da un nuevo archivo de datos genómicos, PED y MAP, que hace referencia a la matriz  $G$  de genotipos dónde el individuo  $i$ -ésimo tiene un genotipo  $g_{ij}$  para el  $j$ -ésimo SNP.
- Se transformó la matriz  $G$  a una matriz de relación para calcular los coeficientes (scores) de los componentes principales (PCs) por el método de descomposición en valores singulares para cada SNP dentro del cromosoma  $\theta$  usando el comando `--pca` en *PLINK 1.9*.
- *PLINK 1.9* determina las varianzas  $\lambda_k$  (valores eigen) de  $k$  PCs y se tomaron los necesarios para obtener el 90% de la varianza total de la matriz original de genotipos  $G$ .
- El software genera el  $a_k$  vector eigen, asociado a  $\lambda_k$  valor eigen, que consiste en los coeficientes para los genotipos dentro de la matriz  $G$  para el PC  $k$ -ésimo y calcula los valores para el individuo  $i$ -ésimo de dicho PC cómo el producto de  $(g_{i1}, g_{i2}, \dots, g_{i(n-1)}) * a_k$
- Los PCs calculados por *PLINK 1.9* fueron usados como covariantes generando la ecuación.

$$\begin{aligned} \text{Fenotipo} = & \beta_0 + \beta_1 \text{SNP}_{prueba} + \beta_1 \text{PC}_1 + \beta_n \text{PC}_n + \dots + \beta_1 \text{Covariante}_1 \\ & + \beta_k \text{Covariante}_k \end{aligned}$$

- El proceso se repite por SNP y por cromosoma, para evaluar el efecto independiente de todos los SNPs previamente asociados.
- El análisis de asociación genera una lista de variantes con el respectivo valor de significancia. Se consideró una significancia a partir de valor de  $P$  menor a 0.05 corregido por FDR.
- En este estudio se discutió la posible causalidad o relación de estos SNPs con los fenotipos clínicos.

## RESULTADOS

### Estadística descriptiva

Se analizaron las características clínicas de los 440 voluntarios, 124 (28.2%) controles y 316 (71.8%) casos, colectados por el Instituto Nacional de Cardiología. La proporción entre hombres y mujeres fue del 40% y 60% respectivamente; además, la edad promedio fue de  $38 \pm 9.8$  años. Al comparar las medias de todas las variantes clínicas, se observó que en el IMC, edad (años), glucosa (mg/dL) y ácido úrico (mg/dL) en sangre fueron estadísticamente mayores en casos (individuos con dislipidemia) en comparación con los controles (individuos considerados sanos)  $P < 0.05$ . También, se encontraron diferencias relevantes, entre controles y casos, en cuanto a los niveles plasmáticos de:

- TC, mg/dL (mayor en casos)
- C-LDL, mg/dL (mayor en casos)
- TG, mg/dL (mayor en casos)
- C-HDL, mg/dL (menor en casos)

Posteriormente, se realizaron comparaciones de todas las variables clínicas entre hombres y mujeres, resultando que existen diferencias significativas para todos los niveles lipídicos, y en todas las variables clínicas con excepción de edad e IMC, como se observa en la tabla 5.

Tabla 5. Características demográficas y clínicas de la cohorte Tlalpan 2020, comparando entre hombres y mujeres.

Variable	Hombres	Mujeres	P*
Edad, años	36.6 ± 9.7	37.8 ± 9.9	NS**
Dislipidemicos [n (%)]	125 (71%)	191 (72%)	NS
Peso, kg	79.2 ± 14.3	66.1 ± 13.2	5.53E-20
Altura, m	1.70 ± 0.07	1.57 ± 0.06	2.12E-57
IMC	27.4 ± 4.3	26.9 ± 5.1	NS
Cintura, cm	94.0 ± 11.7	86.8 ± 12.9	3.13E-09
Glucosa, mg/dL	94.7 ± 10.5	92.4 ± 12.3	3.39E-02
Ácido úrico, mg/dL	6.54 ± 1.14	4.65 ± 0.95	3.91E-51
Creatinina, mg/dL	0.95 ± 0.12	0.70 ± 0.10	1.12E-70
Colesterol Total, mmol/L	4.73 ± 0.89	4.49 ± 0.74	3.77E-03
C-HDL, mmol/L	1.12 ± 0.24	1.24 ± 0.28	2.06E-06
C-LDL, mmol/L	3.15 ± 0.82	2.96 ± 0.71	1.38E-02
Triglicéridos, mmol/L	1.90 ± 1.6	1.36 ± 0.68	4.17E-05
Índice de Castelli II	2.90 ± 0.84	2.49 ± 0.79	6.18E-07
Índice Aterogénico de Plasma	0.51 ± 0.04	0.36 ± 0.25	1.99E-07

\*Valor P para una prueba de *t* de Student de dos colas o *U* de Mann-Whitney según corresponda.

\*\*Diferencia no significativa, considerando una  $P < 0.05$ .

### Selección de covariantes clínicas

Después de que se caracterizó la muestra de estudio, se evaluaron las variantes clínicas para elegir aquellas que serían usadas en los modelos de asociación genotipo-fenotipo. Se seleccionaron variantes clínicas independientes unas de otras, es decir que no fueran colineales. Las variables clínicas fueron seleccionadas según su aproximación matemática (figura 7). Aquellas variantes que empataban y eran colineales se agrupaban y de éstas solo se eligió una usando el método de componentes principales.

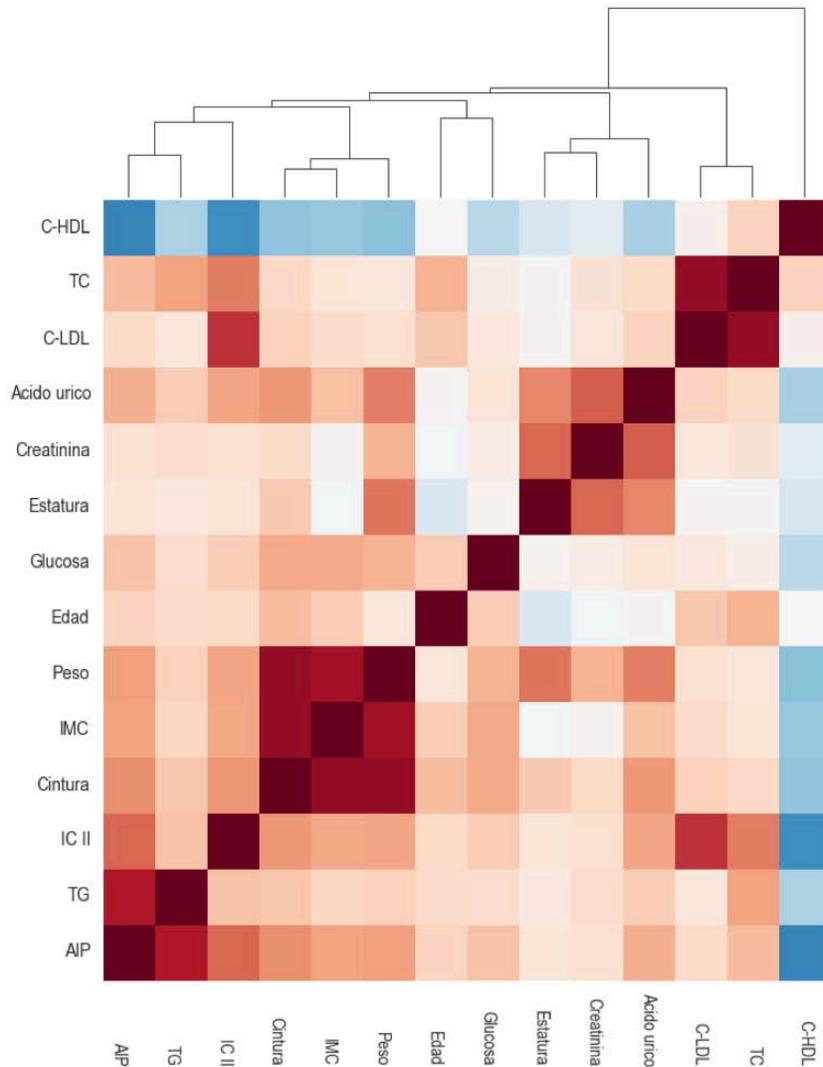


Figura 7. Agrupación jerárquica de las covariantes clínicas.

La proximidad matemática entre las variables clínicas (figura 7) se obtuvo mediante los valores de correlación de Pearson, donde el color se intensifica representando una relación mayor (0.3-1). Usando dichos valores, se formaron los siguientes grupos de las variables clínicas correlacionadas matemáticamente:

- TC y C-LDL
- Ácido úrico, creatinina y estatura
- Glucosa, edad, peso, IMC, cintura e IC II
- TG y AIP
- C-HDL, peso, IMC, cintura, IC II, TG, TC y AIP

Con el propósito de identificar la(s) covariable(s) clínica con la mayor aportación en cuanto a varianza en cada grupo de estudio, se realizó un análisis de componentes principales (PCA) y el coeficiente de Pearson. Este análisis transforma los valores continuos de las variables clínicas en vectores (componentes principales) que permiten encontrar la covariable clínica que explica la mayor variación, asignando valores a cada una de las covariables (Scores) según su aporte en la suma total de la variación. En la figura 8, se observa en un gráfico de dispersión los primeros dos componentes.

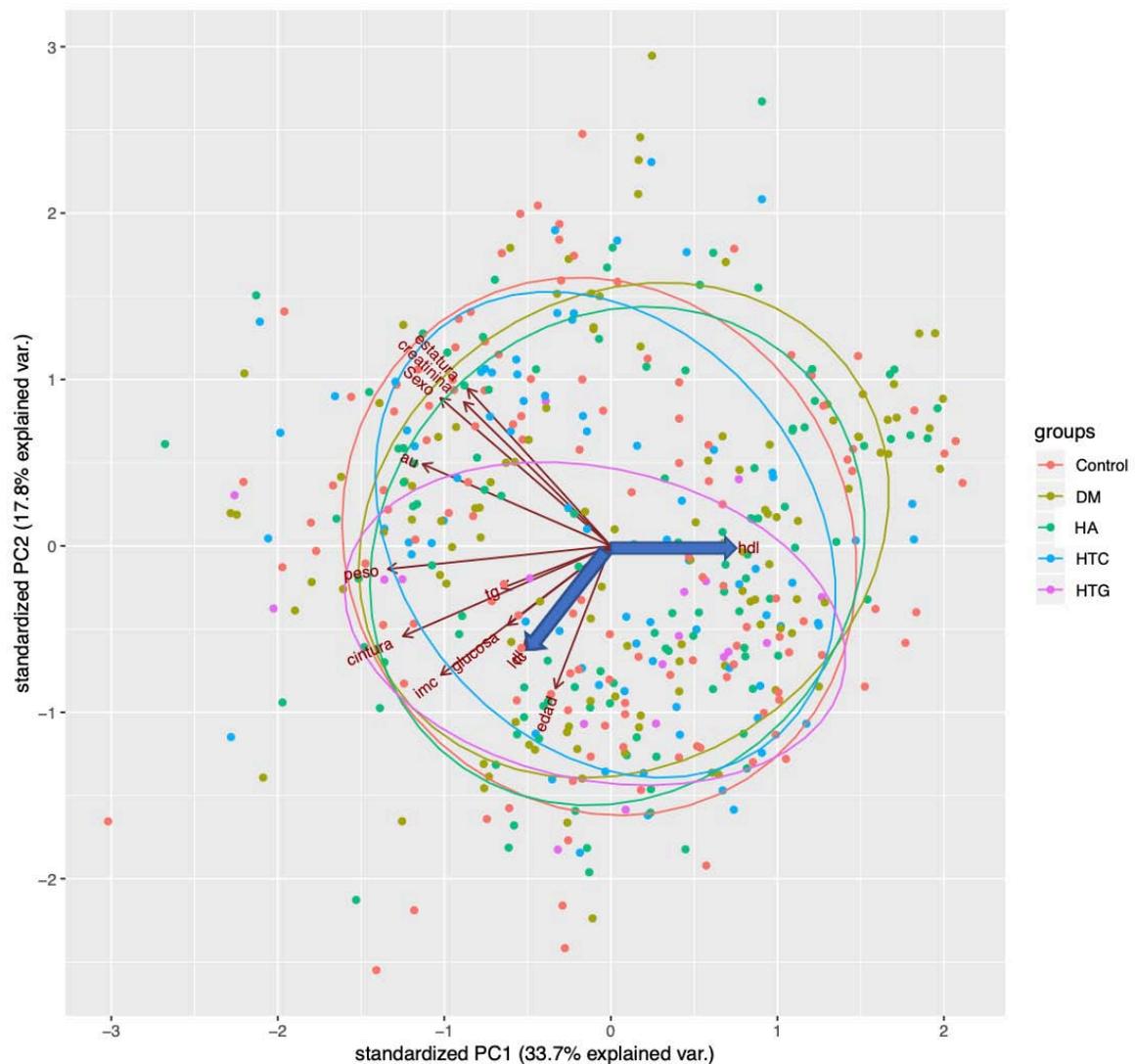


Figura 8. Dispersión de las muestras en los primeros dos componentes principales que explican el 51% de la variabilidad.

Las muestras se encuentran dispersas por tipo de dislipidemia (DM: dislipidemia mixta, HDL: hipoalfalipoproteinemia, HTC: hipercolesterolemia y HTG: hipertrigliceridemia) y controles. Además, se aprecian los vectores para cada covariable clínica, se puede observar que el único factor antiaterogénico es el C-HDL ya que es el único vector cuya dirección es completamente opuesta a las demás variables clínicas. Así mismo, se observa el sobrelape de las variables clínicas que son colineales como el C-LDL y el CT, ya que el C-LDL representa el 70% del CT.

Las siguientes covariables clínicas fueron incluidas en el análisis de asociación, pues mostraron los valores de PCA (scores) más altos y no eran colineales:

- Ácido úrico (mg/dL)
- Peso (kg)
- Cintura (cm)
- Sexo (1 Hombre, 2 Mujer)

### **Control de calidad de muestras**

Las muestras de ADN (35-50 ng/ $\mu$ L) fueron diluidas en agua libre de nucleasas y evaluadas en cuanto a pureza e integridad mediante electroforesis en geles de agarosa como se observa en la figura 9, donde todas las muestras tienen una sola banda correspondiente al ADN genómico y no hay corrimiento de bandas a lo largo del carril, por lo que se puede decir que el ADN está íntegro para cada muestra. Cabe mencionar que las demás muestras evaluadas por geles de agarosa se encuentran en el apéndice A.

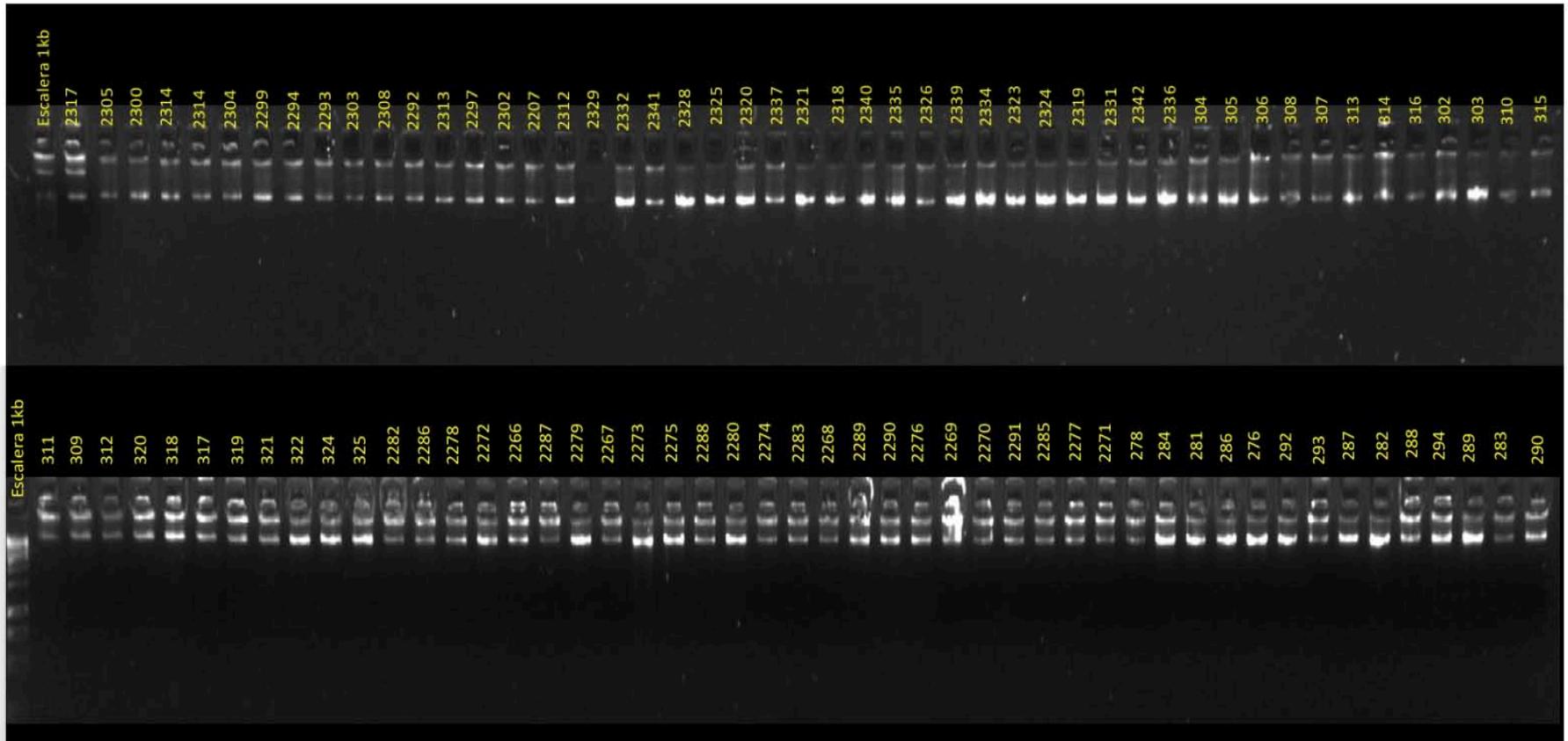


Figura 9. Gel de agarosa al 2%. Carril 1º, escalera de peso molecular 1 Kb, seguido del identificador para cada muestra.

### **Control de calidad de datos genómicos para muestras control**

Las muestras de ADN se genotiparon utilizando 40  $\mu\text{L}$  de ADN a una concentración de 40  $\text{ng}/\mu\text{L}$  en el microarreglo de *Infinium GSA* personalizado por el INMEGEN (669,672 variantes) y debido a que era la primera vez que se usaba este diseño, se evaluó la calidad del genotipado para 51 muestras control, que incluían 8 réplicas intra e inter-lotes tomadas al azar siguiendo los apartados A y B del diagrama de flujo de la figura 5 en el capítulo de metodología.

Los datos crudos de fluorescencia del microarreglo se analizaron para su transformación en datos con información de genotipos en archivos de texto plano (PED/MAP) y se compararon con los mismos archivos de genotipos (PED/MAP) generados por el servicio de genotipificación, con el fin de corroborar que la asignación de los genotipos era correcta.

Para lo anterior, se determinó el grado de similitud de los genotipos obtenidos en el laboratorio contra los entregados por el servicio de genotipificación usando una prueba *t-Student*, la cual comparó la diferencia entre las medias de las proporciones sobre los genotipos. Se obtuvo que la diferencia entre las muestras entregadas, contra las generadas en el laboratorio era menor al 99% con un nivel de confianza de  $\alpha = 0.01$ , lo que aseguraba la reproducibilidad del genotipado con este microarreglo.

También se evaluó la proporción de homocigotos y heterocigotos de las 51 muestras control, apartado B de la figura 5, debido a que fluctuaciones en la variación de las proporciones indican que el microarreglo es de mala calidad por tener variantes que durante el proceso no fueron debidamente genotipadas por mala hibridación de la muestra con la sonda del microarreglo.

Tras determinar las proporciones para homocigotos, heterocigotos y las variantes no genotipadas de las 51 muestras control, se determinó que todas las proporciones son uniformes entre las muestras, confirmando que las sondas

hibridaron correctamente y que por tanto las muestras fueron debidamente genotipadas.

## **Control de calidad de datos genómicos para estudios GWAS**

### **EFICIENCIA EN EL GENOTIPADO POR VARIANTE Y MUESTRA**

Después de validar el microarreglo usando las 51 muestras control, se procedió con el análisis de control de calidad para las 440 muestras de este proyecto, como se muestra en el diagrama de flujo de la figura 5 en el apartado C, tomando una eficiencia de genotipado del 98% por muestra y por variante. Se encontró que 33,981 variantes, que representaban el 5% del total, no pasaron el filtro de eficiencia de genotipado al 98%. Por otra parte, la eficiencia de genotipado en 3 de las 440 muestras presentaron un valor cercano al corte (97.9%), por lo que no fueron descartadas.

### **VERIFICACIÓN DE SEXO**

Se validó que el sexo de la base clínica coincidiera con el predicho usando datos genómicos, tomando como base el valor F (homocigocidad del cromosoma X) para determinar el sexo masculino ( $F > 0.8$ ) o el sexo femenino ( $F < 0.2$ ), resultando en 12 muestras dispares que fueron descartadas. El valor de F se basa en las proporciones de heterocigosidad del cromosoma X para determinar de manera empírica el sexo, reportando aquellos casos en los que el sexo no coincide.

### **HARDY-WEINBERG**

Se determinó la frecuencia del alelo menor y el equilibrio de Hardy-Weinberg (apartados E y F de la figura 5), y se seleccionaron las variantes con una frecuencia de al menos el 5% dentro de la población de estudio y que no estuvieran fuera del equilibrio HW considerando un valor de  $P < 1E-5$ . Cuando un marcador está fuera del equilibrio puede ser indicativo de errores de genotificación, estratificación de la población o incluso asociación real con el rasgo en estudio.

Al final de estos filtros de calidad se eliminaron 329,627 variantes y 12 muestras, por lo que pasaron los controles de calidad 296,064 SNPs y 428 muestras las cuales fueron consideradas para los análisis de mestizaje y asociaciones genotipo-fenotipo candidato. El número de variantes y muestras que pasaron o fueron descartados durante cada etapa del control de calidad se encuentran descritos en la tabla 6.

Tabla 6. Número de muestras y/o variantes que pasaban en cada filtro del control de calidad para estudios de asociación genotipo-fenotipo.

Control de calidad	Pasan		Descartan	
	Muestras	Variante	Muestras	Variante
Eficiencia de genotipado por variante y muestra (C)	440	635,691	0	33,981
Inconsistencias entre sexo clínico y predicho (D)	428	NA	12	NA
Frecuencia del alelo menor (E)	NA	308,463	NA	327,228
Equilibrio de Hardy-Weinberg (F)	NA	296,064	NA	12,399

\*NA = El control de calidad no aplica sobre ese parámetro

### **Análisis de mestizaje (proporción del componente ancestral)**

Los 296,064 SNPs y 428 muestras de este proyecto que pasaron los controles de calidad para estudios *GWAS*, se unieron con la intersección de los datos genómicos provenientes de 1000 genomas para las poblaciones ancestrales de Chinos Han de Beijing (CHB), Europeos del Norte de Utah (CEU) y Yoruba en Ibadan de Nigeria (YRI), además de los datos genómicos de 20 nativos mexicanos provenientes de Campeche, Durango y Oaxaca que previamente fueron genotipados usando el mismo microarreglo personalizado de Illumina.

Una vez conformada la base de datos genómicos para evaluar el mestizaje de las muestras, se aplicaron nuevamente los controles de calidad para estudios *GWAS* y se agregó el filtro de desequilibrio de ligamiento entre los marcadores, además del parentesco entre las muestras. Finalmente, se obtuvo un panel de variantes para determinar la ancestría con 56,028 variantes y 756 muestras.

Es importante mencionar que, tras realizar el análisis de parentesco entre las 428 muestras de este proyecto, se encontró con un par de gemelos, por lo que se descartó una de estas muestras. Adicionalmente se etiquetaron debidamente otras muestras relacionadas en primer nivel como primos, hermanos, padres e hijos para el análisis posterior. Finalmente, se obtuvo un total de 426 muestras y 296,064 variantes que fueron consideradas para los siguientes análisis.

### PROPORCIÓN ANCESTRAL DE LA POBLACION DE ESTUDIO: ANÁLISIS DE COMPONENTES PRINCIPALES (PCA)

En este paso se buscó que las muestras se agruparan de acuerdo a la estructura y proporción poblacional. La varianza que existe entre las frecuencias alélicas del panel de variantes para evaluar ancestría es indicativa de que dichas variantes segregan en conjunto y en diferente frecuencia, lo que permite diferenciar a los grupos como se observa en la figura 10.

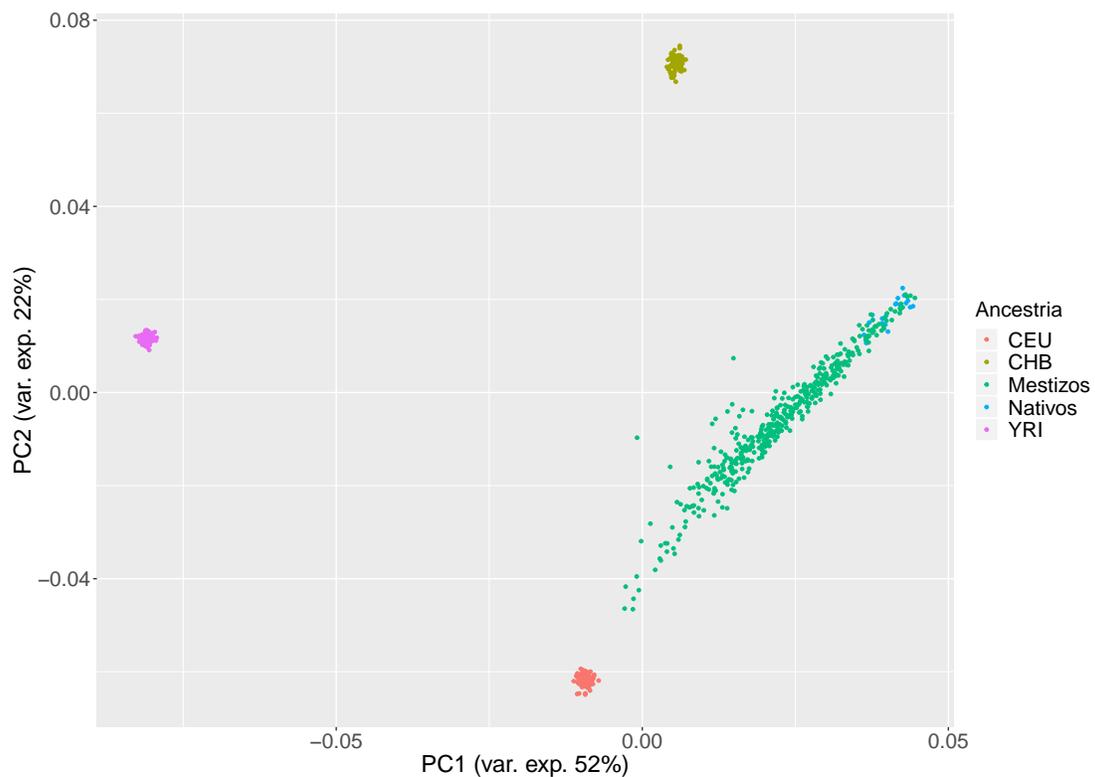


Figura 10. Dispersión de las muestras en base a las diferentes estructuras poblacionales en los primeros componentes principales que explican el 74% del total de la variación del panel de variantes para evaluar ancestría.

### PROPORCIÓN DEL COMPONENTE ANCESTRAL

Para la proporción del componente ancestral, se eligieron 4 estructuras ancestrales (K = 4) usando el método de validación cruzada del software *ADMIXTURE*. Esta estrategia, confirmó lo observado en el análisis de PCA, figura 10, donde se encuentran cuatro centroides correspondientes a las 4 estructuras ancestrales: CEU, CHB, Nativos y YRI.

Adicionalmente, se calcularon las proporciones ancestrales para las 426 muestras, (Figura 11), de estas se eligió la proporción del componente nativo y africano como covariables entre ellas adicionales para el modelo de asociación posterior, ya que no eran colineales.

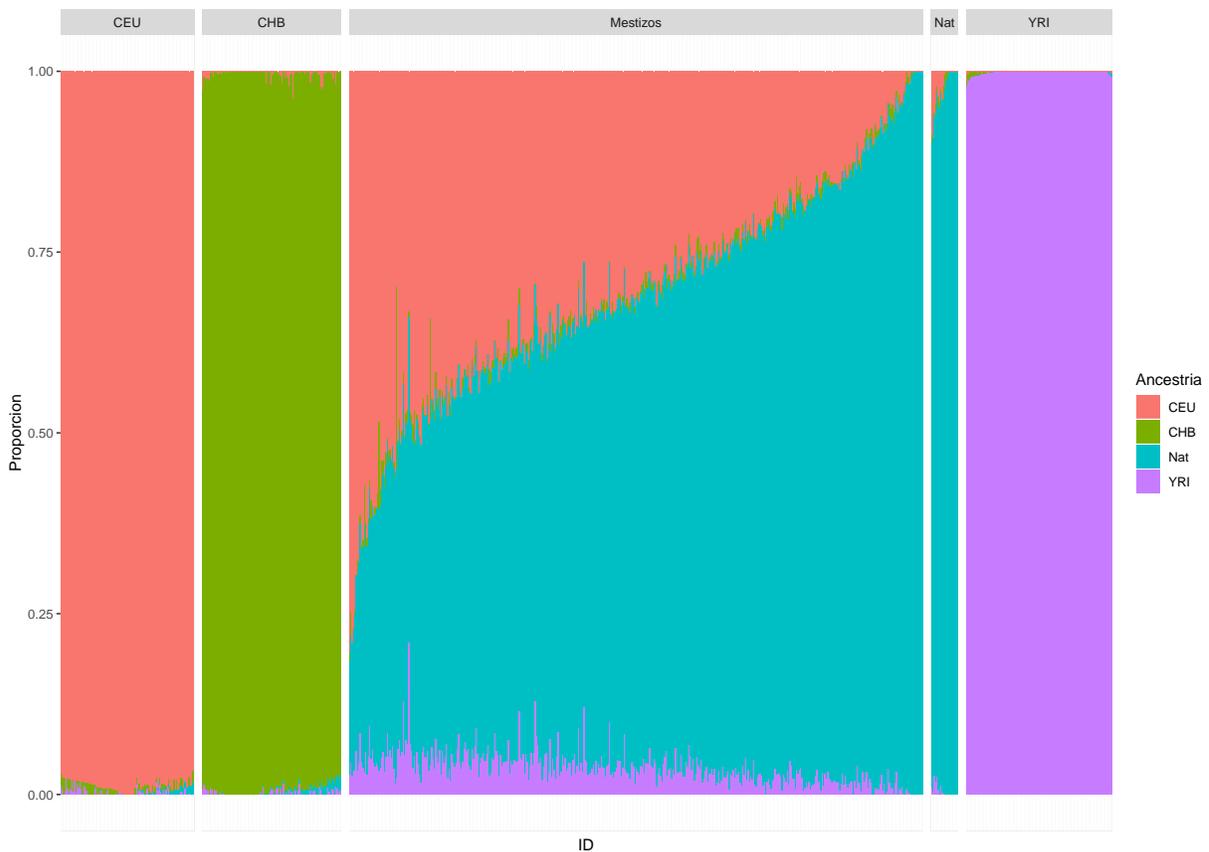


Figura 11. Proporción de las cuatro estructuras ancestrales (CHB, CEU, YRI, Nativos) para los mestizos de este proyecto, los nativos mexicanos y las muestras de 1000 genomas para CHB, CEU e YRI.

Para este punto ya se han definido las muestras que cumplen con el control de calidad (N=426, variantes =296,064) y las covariables clínicas (ácido úrico, peso, cintura, sexo y componente ancestral) que fueron consideradas en los modelos de asociación genotipo-fenotipo. Se eligió una estrategia de asociación tipo gen candidato, la cual consideró una lista de genes candidato que fue seleccionada a través de una revisión de literatura. Dichos genes han sido reportados en estudios previos de asociación con dislipidemias o metabolismo de lípidos para diferentes poblaciones <sup>3,5,6</sup>. De esta recopilación bibliográfica elegimos 150 genes y 183 variantes para realizar el presente estudio de asociación.

### **Estrategia de enriquecimiento de gen candidato**

Debido a que existen variantes que posiblemente no esten consideradas en la lista de gen candidato por la falta de estudios en poblaciones latinas y mexicanas o que no fueron encontradas, se propuso identificar variantes con diferencias significativas entre frecuencia alélica sobre casos y controles, para incorporar los genes/variantes a nuestra lista de gen candidato a partir de un punto de corte de  $P_{\text{Bonferroni}} < 0.05$ .

Estas comparaciones entre casos y controles se realizaron estratificando por género y tipo de dislipidemia:

- Género
  - Hombres casos y controles
  - Mujeres casos y controles
- Dislipidemia
  - Hipercolesterolemia casos y controles
  - Hipertrigliceridemia casos y controles
  - Hipoalfalipoproteinemia casos y controles
  - Dislipidemia mixta casos y controles
- Género y dislipidemia
  - Hombres
    - Hipercolesterolemia casos y controles
    - Hipertrigliceridemia casos y controles

- Hipoalfalipoproteinemia casos y controles
- Dislipidemia mixta casos y controles
- Mujeres
  - Hipercolesterolemia casos y controles
  - Hipertrigliceridemia casos y controles
  - Hipoalfalipoproteinemia casos y controles
  - Dislipidemia mixta casos y controles

De cada grupo se comparó la frecuencia alélica entre casos y controles y se determinó su significancia (426 muestras y 296,064 SNPs). De estas comparaciones se seleccionaron 373 genes (10,990 variantes), los cuales fueron incluidos en la lista de genes (variantes) candidato. Cabe resaltar que en esta nueva lista se encontraban el 90% de los 150 genes ya considerados a partir de la revisión de la literatura. En la figura 12 se enlista el número de genes que se encontraron por grupo.

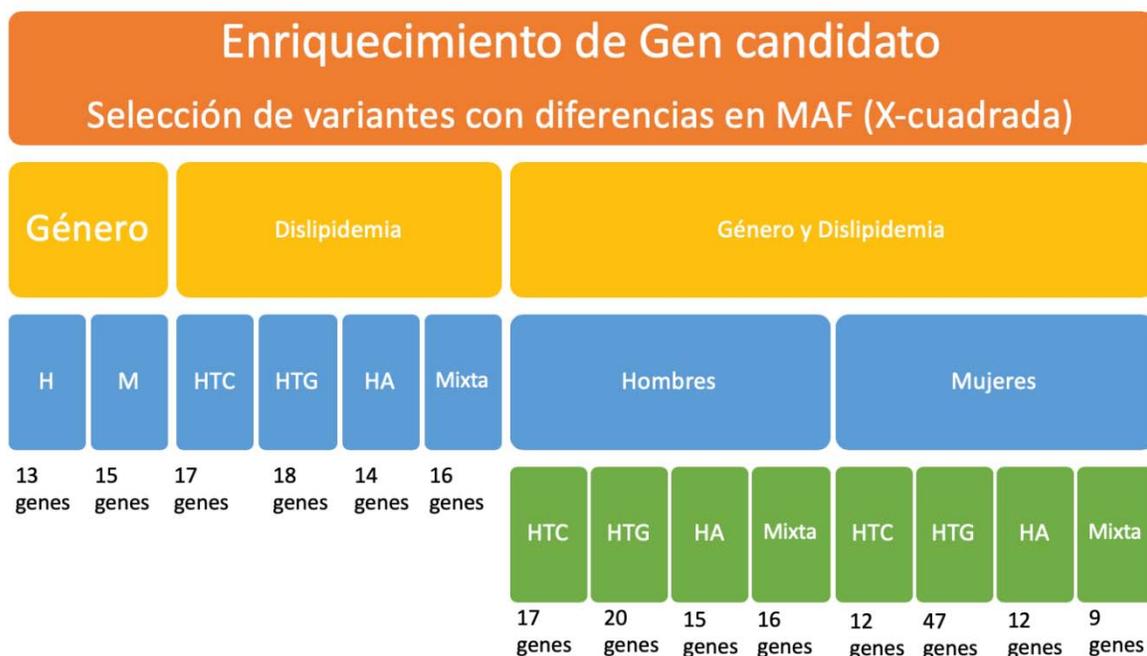


Figura 12. Número de genes que resultaron de la comparación de las frecuencias alélicas entre casos y controles para cada grupo.

Se obtuvo una lista con 373 genes y 10990 variantes que, en conjunto con las covariables clínicas previamente definidas y el fenotipo, se utilizaron para crear los modelos de asociación genotipo-fenotipo por una estrategia de gen candidato.

### **Desarrollo de modelos de asociación genotipo-fenotipo.**

Se desarrollaron modelos de regresión múltiple considerando como variable independiente los siguientes fenotipos:

- Lineal (Valores continuos)
  - Perfiles lipídicos (TC, TG, C-LDL y C-HDL)
  - Índices aterogénicos (AIP e IC II)
- Logístico (Casos contra controles)
  - Tipos de dislipidemia
    - Hipercolesterolemia
    - Hipertrigliceridemia
    - Hipoalfalipoproteinemia
    - Dislipidemia mixta

Las regresiones consideraron las siguientes covariables seleccionadas de tal forma que no existió colinealidad entre ellas:

- Ácido úrico, mg/dL
- Peso, kg
- Cintura, cm
- Sexo (1 Hombre, 2 Mujer)
- Fracciones del componente ancestral (CEU y YRI)

El resultado de las regresiones incluyó variantes asociadas a cada fenotipo, pero sólo aquellos con valores de  $P < 0.05$  (sin corregir) fueron seleccionados para la siguiente etapa del análisis. En la tabla 7 se encuentra el número de SNPs con valores de  $P < 0.05$  y el número de genes cubiertos por dichos SNPs, además de los rangos de valores de  $P$  entre los SNPs asociados con cada fenotipo. Puesto que existían más de 500 SNPs sugestivamente ( $P < 0.05$ ) asociados con cada fenotipo, se decidió detallar sólo aquellos que tuvieran un valor de significancia de

$P < 9E-4$  en las tablas 8 a 17, dividiendo por cada fenotipo analizado en el primer análisis de regresión.

Tabla 7. Número de SNPs (genes) asociados significativamente ( $P < 0.05$ ) con los diferentes fenotipos usando regresiones lineales y logísticas múltiples.

Fenotipo*	No. de SNPs asociados $P < 0.05$ (No. genes)	Rango de valores de $P$ entre los SNPs asociados
Hipercolesterolemia (Casos vs Controles)	686 (181)	$9E-5 \geq P \leq 1$
Hipertrigliceridemia (Casos vs Controles)	653 (177)	$1E-4 \geq P \leq 1$
Hipoalfalipoproteinemia (Casos vs Controles)	858 (206)	$5E-6 \geq P \leq 1$
Dislipidemia mixta (Casos vs Controles)	677 (173)	$2E-6 \geq P \leq 1$
Colesterol Total (Variable continua)	650 (180)	$4E-5 \geq P \leq 1$
Triglicéridos (Variable continua)	499 (160)	$2E-6 \geq P \leq 1$
C-HDL (Variable continua)	687 (151)	$2E-6 \geq P \leq 1$
C-LDL (Variable continua)	617 (183)	$1E-5 \geq P \leq 1$
Índice de Castelli II (Variable continua)	629 (181)	$7E-8 \geq P \leq 1$
Índice Aterogénico de Plasma (Variable continua)	687 (173)	$7E-8 \geq P \leq 1$

\*Los fenotipos están mezclados entre valores continuos y categóricos, donde los únicos valores categóricos son los tipos de dislipidemia (1 para casos y 2 para controles). Dependiendo de la clase de fenotipo fue el tipo de regresión, lineal para continuos y logística para categóricos.

Tabla 8. SNPs asociados significativamente con hipercolesterolemia, usando regresiones logísticas múltiples, comparando entre voluntarios sanos y dislipidémicos ( $P < 9E-4$ ).

SNP	Gen	Chr	Beta	P	MAF
					EUR / MXL / T2020
rs10770114	<i>AMPD3</i>	11	-3.62	2.96E-04	0.43 / 0.23 / 0.15
rs2213565	<i>HLA-DQA2</i>	6	-3.48	5.06E-04	0.79 / 0.16 / 0.15
rs7997299	<i>MYO16</i>	13	-3.46	5.32E-04	0.32 / 0.28 / 0.26
rs2213573	<i>HLA-DQA2</i>	6	-3.39	6.88E-04	0.79 / 0.16 / 0.15
rs2213572	<i>HLA-DQA2</i>	6	-3.39	6.98E-04	0.61 / 0.26 / 0.22
rs2051597	<i>HLA-DQA2</i>	6	-3.39	7.04E-04	- / 0.18 / 0.14
rs3827434	<i>GPR50</i>	X	-3.36	7.75E-04	0.79 / 0.47 / 0.42

rs72657320	<i>DNAH11</i>	7	-3.36	7.79E-04	0 / 0.06 / 0.08
rs7738021	<i>MARCH3</i>	5	-3.35	8.23E-04	0.7 / 0.27 / 0.24
rs954811	<i>WWOX</i>	16	-3.34	8.33E-04	0.64 / 0.47 / 0.42
rs13028762	<i>CTNNA2</i>	2	-3.33	8.80E-04	0.34 / 0.48 / 0.48
rs7755597	<i>HLA-DQA2</i>	6	-3.31	9.46E-04	0.79 / 0.16 / 0.15
rs7773068	<i>HLA-DQA2</i>	6	-3.31	9.46E-04	0.79 / 0.16 / 0.15
rs7773694	<i>HLA-DQA2</i>	6	-3.31	9.46E-04	0.79 / 0.16 / 0.15
rs9276408	<i>HLA-DQA2</i>	6	-3.31	9.48E-04	0.77 / 0.21 / 0.17

Chr indica el cromosoma; Beta, efecto asociado a la variante; P, valor de p; MAF, Frecuencia del alelo menor en población europea (EUR), mexicanos de Los Ángeles (MXL) y los mexicanos de este proyecto (T2020).

Tabla 9. SNPs asociados significativamente con hipertrigliceridemia, usando regresiones logísticas múltiples, comparando entre voluntarios sanos (controles) y dislipidémicos ( $P < 9E-4$ ).

SNP	Gen	Chr	Beta	P	MAF
					EUR / MXL / T2020
rs12093524	<i>H6PD</i>	1	-3.37	7.44E-04	0.36 / 0.45 / 0.46

Chr indica el cromosoma; Beta, efecto asociado a la variante; P, valor de p; MAF, Frecuencia del alelo menor en población europea (EUR), mexicanos de Los Ángeles (MXL) y los mexicanos de este proyecto (T2020).

Tabla 10. SNPs asociados significativamente con hipoalfalipoproteinemia, usando regresiones logísticas múltiples, comparando entre voluntarios sanos y dislipidémicos ( $P < 9E-4$ ).

SNP	Gen	Chr	Beta	P	MAF
					EUR / MXL / T2020
rs4766308	<i>KCNA1</i>	12	-4.07	4.66E-05	0.76 / 0.22 / 0.22
rs62458216	<i>HECW1</i>	7	-3.86	1.12E-04	0.02 / 0.09 / 0.14
rs9666182	<i>IFITM1</i>	11	-3.82	1.33E-04	0.73 / 0.13 / 0.14
rs71337756	<i>CTNNA2</i>	2	-3.81	1.39E-04	0.05 / 0.03 / 0.08
rs113489099	<i>TBC1D16</i>	17	-3.77	1.63E-04	0.09 / 0.09 / 0.09
rs7894429	<i>STK32C</i>	10	-3.76	1.73E-04	0.28 / 0.2 / 0.17
rs2967846	<i>COTL1</i>	16	-3.68	2.32E-04	0.33 / 0.2 / 0.13
rs72657320	<i>DNAH11</i>	7	-3.66	2.52E-04	0 / 0.06 / 0.08
rs28526765	<i>SLCO3A1</i>	15	-3.59	3.33E-04	0.19 / 0.16 / 0.14
rs174600	<i>FADS2</i>	11	-3.56	3.68E-04	0.33 / 0.32 / 0.28
rs710192	<i>MIEF1</i>	22	-3.56	3.76E-04	0.32 / 0.28 / 0.28
rs2294360	<i>MIEF1</i>	22	-3.49	4.90E-04	0.39 / 0.36 / 0.32
rs17644210	<i>SLCO3A1</i>	15	-3.45	5.56E-04	0.15 / 0.2 / 0.17
rs6982017	<i>EBF2</i>	8	-3.44	5.82E-04	0.39 / 0.46 / 0.41

rs7001120	<i>EBF2</i>	8	-3.44	5.82E-04	0.39 / 0.46 / 0.41
rs1544028	<i>SPOCK1</i>	5	-3.43	6.13E-04	0.65 / 0.31 / 0.33
rs11857019	<i>LIPC</i>	15	-3.43	6.16E-04	0.05 / 0.2 / 0.24
rs174448	<i>FADS2</i>	11	-3.41	6.53E-04	0.62 / 0.32 / 0.31
rs10503781	<i>EBF2</i>	8	-3.34	8.35E-04	0.08 / 0.06 / 0.09
rs7124405	<i>SHANK2</i>	11	-3.33	8.85E-04	0.39 / 0.32 / 0.33
rs174619	<i>FADS2</i>	11	-3.29	9.98E-04	0.54 / 0.29 / 0.26

Chr indica el cromosoma; Beta, efecto asociado a la variante; P, valor de p; MAF, Frecuencia del alelo menor en población europea (EUR), mexicanos de Los Ángeles (MXL) y los mexicanos de este proyecto (T2020).

Tabla 11. SNPs asociados significativamente con dislipidemia mixta, usando regresiones logísticas múltiples, comparando entre voluntarios sanos y dislipidémicos ( $P < 9E-4$ ).

SNP	Gen	Chr	Beta	P	MAF
					EUR / MXL / T2020
rs76234113	<i>PSMB3</i>	17	-4.06	4.99E-05	0.06 / 0.24 / 0.29
rs897173	<i>GABRG3</i>	15	-3.49	4.77E-04	0.18 / 0.41 / 0.35
rs664936	<i>FAM63B</i>	15	-3.47	5.14E-04	- / 0.46 / 0.5
rs1528961	<i>SPOCK1</i>	5	-3.44	5.74E-04	0.43 / 0.4 / 0.39
rs3827434	<i>GPR50</i>	X	-3.44	5.89E-04	0.79 / 0.47 / 0.42
rs4243105	<i>GABRG3</i>	15	-3.37	7.65E-04	0.09 / 0.31 / 0.3
rs4766308	<i>KCNA1</i>	12	-3.33	8.83E-04	0.76 / 0.22 / 0.22

Chr indica el cromosoma; Beta, efecto asociado a la variante; P, valor de p; MAF, Frecuencia del alelo menor en población europea (EUR), mexicanos de Los Ángeles (MXL) y los mexicanos de este proyecto (T2020).

Tabla 12. SNPs asociados significativamente con colesterol total de voluntarios sanos y dislipidémicos usando regresiones lineales múltiples ( $P < 9E-4$ ).

SNP	Gen	Chr	Beta	P	MAF
					EUR / MXL / T2020
rs9282541	<i>ABCA1</i>	9	-4.17	3.67E-05	0 / 0.07 / 0.1
rs732350	<i>BCAS3</i>	17	-4.08	5.42E-05	0.65 / 0.31 / 0.26
rs1886684	<i>EVI5</i>	1	-3.75	2.03E-04	- / 0.26 / 0.19
rs8077052	<i>SKA2</i>	17	-3.69	2.51E-04	0.29 / 0.14 / 0.18
rs2278426	<i>C19orf80</i>	19	-3.61	3.44E-04	0.04 / 0.28 / 0.31
rs34818467	<i>TEX14</i>	17	-3.6	3.55E-04	0.18 / 0.1 / 0.11
rs2120659	<i>DNAH17</i>	17	-3.47	5.67E-04	- / 0.39 / 0.3
rs737337	<i>C19orf80</i>	19	-3.39	7.56E-04	0.07 / 0.29 / 0.32
rs3754751	<i>RAPGEF4</i>	2	-3.38	8.02E-04	0.81 / 0.2 / 0.15
rs12971537	<i>DOCK6</i>	19	-3.37	8.11E-04	0.09 / 0.29 / 0.3

rs4367982	<i>LRP1</i>	12	-3.35	8.83E-04	0.32 / 0.45 / 0.5
rs9896398	<i>DNAH17</i>	17	-3.33	9.54E-04	0.6 / 0.4 / 0.3

Chr indica el cromosoma; Beta, efecto asociado a la variante; P, valor de p; MAF, Frecuencia del alelo menor en población europea (EUR), mexicanos de Los Ángeles (MXL) y los mexicanos de este proyecto (T2020).

Tabla 13. SNPs asociados significativamente con triglicéridos de voluntarios sanos y dislipidémicos usando regresiones lineales múltiples ( $P < 9E-4$ ).

SNP	Gen	Chr	Beta	P	MAF
					EUR / MXL / T2020
rs3741298	<i>ZNF259</i>	11	-3.67	2.80E-04	0.77 / 0.44 / 0.45
rs4520	<i>APOC3</i>	11	-3.49	5.33E-04	0.7 / 0.41 / 0.49

Chr indica el cromosoma; Beta, efecto asociado a la variante; P, valor de p; MAF, Frecuencia del alelo menor en población europea (EUR), mexicanos de Los Ángeles (MXL) y los mexicanos de este proyecto (T2020).

Tabla 14. SNPs asociados significativamente con C-LDL de voluntarios sanos y dislipidémicos usando regresiones lineales múltiples ( $P < 9E-4$ ).

SNP	Gen	Chr	Beta	P	MAF
					EUR / MXL / T2020
rs1886684	<i>EVI5</i>	1	-4.33	1.85E-05	- / 0.26 / 0.19
rs2339324	<i>DCC</i>	18	-3.54	4.41E-04	0.12 / 0.29 / 0.24

Chr indica el cromosoma; Beta, efecto asociado a la variante; P, valor de p; MAF, Frecuencia del alelo menor en población europea (EUR), mexicanos de Los Ángeles (MXL) y los mexicanos de este proyecto (T2020).

Tabla 15. SNPs asociados significativamente con C-HDL de voluntarios sanos y dislipidémicos usando regresiones lineales múltiples ( $P < 9E-4$ ).

SNP	Gen	Chr	Beta	P	MAF
					EUR / MXL / T2020
rs737337	<i>C19orf80</i>	19	-4.42	1.27E-05	0.07 / 0.29 / 0.32
rs9282541	<i>ABCA1</i>	9	-4.39	1.44E-05	0 / 0.07 / 0.1
rs2278426	<i>C19orf80</i>	19	-4.23	2.85E-05	0.04 / 0.28 / 0.31
rs17699089	<i>DOCK6</i>	19	-4	7.65E-05	0.09 / 0.3 / 0.32
rs4774302	<i>LIPC</i>	15	-3.82	1.53E-04	0.6 / 0.33 / 0.34
rs72829648	<i>TMEM87B</i>	2	-3.66	2.85E-04	0.08 / 0.03 / 0.06
rs483431	<i>MAML2</i>	11	-3.56	4.18E-04	0.85 / 0.17 / 0.13
rs7203984	<i>CETP</i>	16	-3.54	4.51E-04	0.22 / 0.28 / 0.33
rs11021499	<i>MAML2</i>	11	-3.51	4.98E-04	0.38 / 0.29 / 0.29
rs11021504	<i>MAML2</i>	11	-3.43	6.68E-04	0.37 / 0.29 / 0.28
rs12971537	<i>DOCK6</i>	19	-3.42	6.89E-04	0.09 / 0.29 / 0.3
rs1978763	<i>MAML2</i>	11	-3.33	9.47E-04	- / 0.28 / 0.28

rs73021218	<i>CDON</i>	11	-3.33	9.54E-04	0.09 / 0.12 / 0.16
rs1077834	<i>LIPC</i>	15	-3.31	9.99E-04	0.21 / 0.48 / 0.4
rs2070895	<i>LIPC</i>	15	-3.31	9.99E-04	0.21 / 0.48 / 0.4

Chr indica el cromosoma; Beta, efecto asociado a la variante; P, valor de p; MAF, Frecuencia del alelo menor en población europea (EUR), mexicanos de Los Ángeles (MXL) y los mexicanos de este proyecto (T2020).

Tabla 16. SNPs asociados significativamente con índice aterogénico de plasma de voluntarios sanos y dislipidémicos usando regresiones lineales múltiples ( $P < 9E-4$ ).

SNP	Gen	Chr	Beta	P	MAF
					EUR / MXL / T2020
<b>rs5128</b>	<b><i>APOC3</i></b>	<b>11</b>	<b>0.21</b>	<b>7.50E-08</b>	<b>0.88 / 0.19 / 0.23</b>
<b>rs5072</b>	<b><i>APOC3</i></b>	<b>11</b>	<b>0.2</b>	<b>2.79E-07</b>	<b>0.89 / 0.18 / 0.22</b>
<b>rs1436261</b>	<b><i>ITGBL1</i></b>	<b>13</b>	<b>0.18</b>	<b>1.47E-06</b>	<b>0.78 / 0.41 / 0.43</b>
<b>rs2854117</b>	<b><i>APOC3</i></b>	<b>11</b>	<b>0.18</b>	<b>2.36E-06</b>	<b>- / 0.34 / 0.32</b>
<b>rs734104</b>	<b><i>APOC3</i></b>	<b>11</b>	<b>0.18</b>	<b>2.40E-05</b>	<b>0.83 / 0.2 / 0.25</b>
rs10750097	<i>APOA5</i>	11	0.18	3.70E-05	0.76 / 0.41 / 0.46
rs2854116	<i>APOC3</i>	11	0.17	5.66E-05	0.61 / 0.42 / 0.37
rs5092	<i>APOA4</i>	11	0.17	1.00E-04	- / 0.24 / 0.27
rs17072775	<i>KCNIP1</i>	5	0.17	1.21E-04	0.21 / 0.07 / 0.07
rs483431	<i>MAML2</i>	11	0.16	1.32E-04	0.85 / 0.17 / 0.13
rs2076188	<i>ADTRP</i>	6	-0.16	2.12E-04	0.75 / 0.3 / 0.24
rs9518471	<i>ITGBL1</i>	13	0.16	2.20E-04	0.26 / 0.38 / 0.37
rs2266788	<i>ZNF259</i>	11	0.16	2.46E-04	0.91 / 0.17 / 0.21
rs11021499	<i>MAML2</i>	11	0.16	2.67E-04	0.38 / 0.29 / 0.29
rs12545377	<i>CSMD1</i>	8	-0.16	3.07E-04	0.15 / 0.07 / 0.08
rs2542052	<i>APOC3</i>	11	0.15	3.31E-04	0.61 / 0.41 / 0.37
rs79849883	<i>ELK3</i>	12	0.15	3.85E-04	0.02 / 0.09 / 0.1
rs6724963	<i>COBLL1</i>	2	-0.15	4.29E-04	0.19 / 0.27 / 0.21
rs9557710	<i>ITGBL1</i>	13	-0.15	4.75E-04	0.7 / 0.45 / 0.46
rs4520	<i>APOC3</i>	11	-0.15	5.15E-04	0.7 / 0.41 / 0.49
rs412000	<i>TEX14</i>	17	-0.15	5.29E-04	- / 0.32 / 0.29
rs11069450	<i>ITGBL1</i>	13	0.15	5.45E-04	0.21 / 0.36 / 0.38
rs2072560	<i>ZNF259</i>	11	0.15	6.20E-04	0.92 / 0.17 / 0.21
rs11021504	<i>MAML2</i>	11	0.14	6.38E-04	0.37 / 0.29 / 0.28
rs1978763	<i>MAML2</i>	11	0.14	6.54E-04	- / 0.28 / 0.28
rs2542051	<i>APOC3</i>	11	0.14	6.81E-04	0.61 / 0.42 / 0.37
rs4774302	<i>LIPC</i>	15	0.15	6.86E-04	0.6 / 0.33 / 0.34
rs411988	<i>TEX14</i>	17	-0.15	7.56E-04	0.56 / 0.32 / 0.28
rs10831500	<i>MAML2</i>	11	0.14	7.82E-04	0.38 / 0.3 / 0.32

rs441460	<i>LRRC16A</i>	6	0.15	8.49E-04	0.44 / 0.42 / 0.47
rs1263177	<i>APOA4</i>	11	0.14	9.47E-04	0.62 / 0.38 / 0.37
rs2849176	<i>APOC3</i>	11	0.14	9.71E-04	- / 0.41 / 0.37

Chr indica el cromosoma; Beta, efecto asociado a la variante; P, valor de p; MAF, Frecuencia del alelo menor en población europea (EUR), mexicanos de Los Ángeles (MXL) y los mexicanos de este proyecto (T2020).

Tabla 17. SNPs asociados significativamente con índice de Castelli II de voluntarios sanos y dislipidemicos usando regresiones lineales múltiples (P < 9E-4).

SNP	Gen	Chr	Beta	P	MAF
					EUR / MXL / T2020
<b>rs261342</b>	<b><i>LIPC</i></b>	<b>15</b>	<b>0.19</b>	<b>7.77E-08</b>	<b>- / 0.48 / 0.41</b>
<b>rs17648647</b>	<b><i>WVVOX</i></b>	<b>16</b>	<b>-0.17</b>	<b>5.00E-07</b>	<b>0.48 / 0.45 / 0.47</b>
<b>rs1512123</b>	<b><i>CNTN5</i></b>	<b>11</b>	<b>0.17</b>	<b>9.58E-07</b>	<b>- / 0.1 / 0.09</b>
<b>rs716583</b>	<b><i>ST18</i></b>	<b>8</b>	<b>-0.16</b>	<b>3.07E-06</b>	<b>0.71 / 0.39 / 0.37</b>
rs812941	<i>CSMD1</i>	8	-0.15	3.84E-05	0.69 / 0.23 / 0.17
rs2120659	<i>DNAH17</i>	17	-0.15	7.19E-04	- / 0.39 / 0.3
rs1160536	<i>IL21R</i>	16	0.14	8.14E-04	0.02 / 0.03 / 0.09
rs7935705	<i>CNTN5</i>	11	0.14	8.19E-04	0.18 / 0.17 / 0.13
rs628419	<i>IMPA2</i>	18	0.15	8.73E-04	- / - / 0.29
rs1886684	<i>EVI5</i>	1	-0.15	8.97E-04	- / 0.26 / 0.19
rs4875319	<i>CSMD1</i>	8	-0.14	9.64E-04	0.15 / 0.09 / 0.1

Chr indica el cromosoma; Beta, efecto asociado a la variante; P, valor de p; MAF, Frecuencia del alelo menor en población europea (EUR), mexicanos de Los Ángeles (MXL) y los mexicanos de este proyecto (T2020).

En la figura 13 , se presenta un resumen de los procesos de filtración/eliminación de genes y por cada etapa del análisis de asociación.

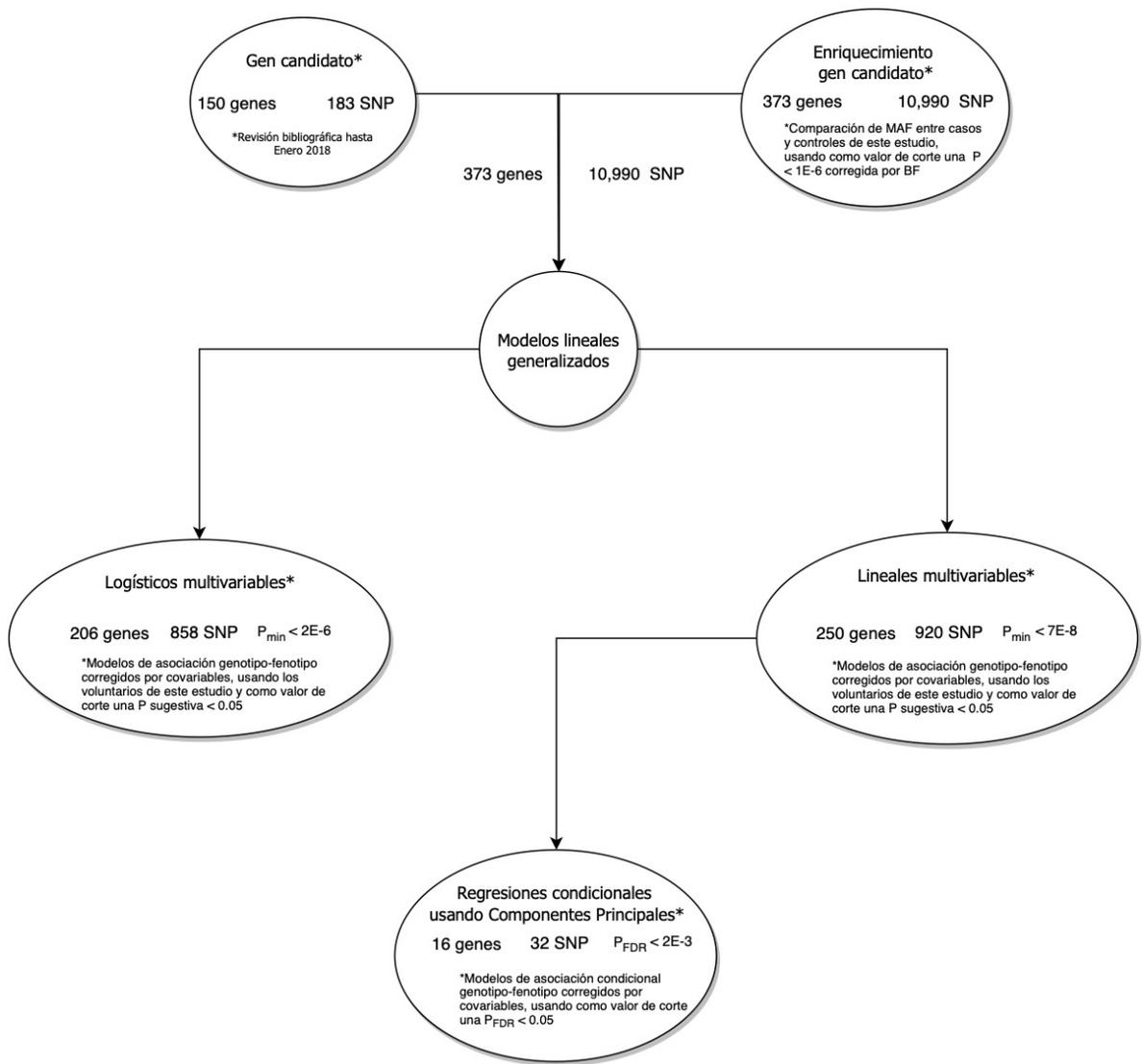


Figura 13. Estrategias utilizadas para la identificación de variantes asociadas a dislipidemias a través de modelos de asociación tipo gen candidato

Cada óvalo representa un método diferente de asociación y/o análisis bioestadístico de comparación de frecuencias, se especifican los criterios para el filtrado de genes/variantes y aquellas que al pasar el filtro resultan significativamente asociadas a los índices aterogénicos.

### **Desarrollo de modelos de regresión condicional**

En la primera etapa de regresiones/asociaciones genotipo-fenotipo, se encontraron 690 variantes en 180 genes significativamente asociadas a los dos índices aterogénicos, ( $P < 1E-6$ ). Durante esta segunda etapa de modelos de asociación se buscó seleccionar variantes de la primera etapa, las cuales estaban significativamente asociadas a los índices aterogénicos, pero que tuvieran un efecto independiente en comparación con las otras. En este segundo análisis se consideró una significancia de  $P_{FDR} < 0.05$ . Los modelos de asociación condicional utilizaron como covariantes, las mismas variables clínicas y las variantes genéticas significativas de la primer regresión. Se desarrollaron modelos condicionales utilizando 629 variantes asociadas significativamente al IC II y los 687 SNPs al AIP (Tablas 16 y 17). Esta estrategia involucra identificar si el efecto de cada una de las variantes, asociadas de la primera etapa es independiente de las variantes genéticas dentro del mismo cromosoma y en caso de que no tuvieran un efecto independiente el algoritmo las eliminó al obtener un valor de significancia  $P_{FDR} > 0.05$ .

Se identificó que, al menos una de las variantes genéticas cercanas entre ellas a menos de 500kb, asociadas a los índices aterogénicos, se encontraban en desequilibrio de ligamiento, por lo que son colineales/redundantes matemática o clínicamente, lo que puede provocar inconsistencias en el modelo, como cambios de sentido en el efecto (beta) de cada variante, por lo que se usaron transformaciones en componentes principales.

Los modelos condicionales arrojaron 133 variantes significativas, las cuales ya habían sido previamente asociados a AIP y 116 a IC II ( $P < 0.05$ ) consolidando su asociación con estos fenotipos y su independencia matemática. Posterior a la corrección estadística por el método de FDR se obtuvieron 15 variantes asociadas a AIP y 17 variantes a IC II, las cuales tienen un efecto independiente con respecto a los demás variantes asociados en la primer etapa (Tablas 18 y 19).

Tabla 18. Polimorfismos de efecto independiente asociados con el Índice Aterogénico de Plasma (AIP)

<b>SNP</b>	<b>Gen</b>	<b>Chr</b>	<b>Beta</b>	<b>P (corregido por FDR)</b>	<b>MAF EUR/MXL/T2020</b>
<b>rs57292936</b>	<b><i>DSCAM</i></b>	<b>21</b>	<b>-0.15</b>	<b>2.53E-03</b>	<b>0.05 / 0.16 / 0.16</b>
<b>rs3829038</b>	<b><i>NDUFB9</i></b>	<b>8</b>	<b>0.21</b>	<b>4.62E-03</b>	<b>0.03 / 0.06 / 0.06</b>
<b>rs77538674</b>	<b><i>NDUFB9</i></b>	<b>8</b>	<b>0.21</b>	<b>4.62E-03</b>	<b>0.03 / 0.06 / 0.06</b>
<b>rs6992199</b>	<b><i>MTSS1</i></b>	<b>8</b>	<b>-0.15</b>	<b>4.82E-03</b>	<b>0.13 / 0.06 / 0.06</b>
rs7855149	<i>LINGO2</i>	9	-0.14	2.09E-02	0.21 / 0.27 / 0.3
rs62118272	<i>MUC16</i>	19	0.13	2.18E-02	0.05 / 0.05 / 0.09
rs6027217	<i>CDH26</i>	20	-0.12	2.24E-02	0.6 / 0.3 / 0.33
rs72996742	<i>ANGPTL4</i>	19	-0.11	2.89E-02	0.27 / 0.27 / 0.27
rs10968632	<i>LINGO2</i>	9	0.12	3.75E-02	0.19 / 0.22 / 0.25
rs12545377	<i>CSMD1</i>	8	-0.12	4.49E-02	0.15 / 0.07 / 0.08
rs209568	<i>MTUS1</i>	8	0.11	4.49E-02	0.22 / 0.21 / 0.23
rs76091103	<i>CSMD1</i>	8	0.11	4.49E-02	0.07 / 0.06 / 0.06
rs17836106	<i>CHODL</i>	21	0.10	4.67E-02	0.07 / 0.13 / 0.09
rs3859701	<i>DSCAM</i>	21	-0.10	4.67E-02	0.12 / 0.09 / 0.08

Chr indica el cromosoma; Beta, efecto asociado a la variante; P, valor de p; MAF, Frecuencia del alelo menor en población europea (EUR), mexicanos de Los Ángeles (MXL) y los mexicanos de este proyecto (T2020).

Tabla 19. Polimorfismos de efecto independiente asociados al Índice de Castelli II (IC II)

SNP	Gen	Chr	Beta	P (corregido por FDR)	MAF EUR/MXL/T2020
<b>rs2837379</b>	<b>DSCAM</b>	<b>21</b>	<b>-0.23</b>	<b>2.18E-03</b>	<b>0.38 / 0.13 / 0.14</b>
<b>rs760168</b>	<b>DSCAM</b>	<b>21</b>	<b>-0.23</b>	<b>2.18E-03</b>	<b>0.38 / 0.13 / 0.14</b>
<b>rs77923858</b>	<b>CHODL</b>	<b>21</b>	<b>0.17</b>	<b>2.18E-03</b>	<b>0.02 / 0.13 / 0.17</b>
<b>rs11786902</b>	<b>CSMD1</b>	<b>8</b>	<b>0.17</b>	<b>4.34E-03</b>	<b>0.36 / 0.19 / 0.15</b>
<b>rs77779037</b>	<b>CHODL</b>	<b>21</b>	<b>0.16</b>	<b>5.69E-03</b>	<b>0.03 / 0.13 / 0.18</b>
rs2824700	CHODL	21	0.11	1.68E-02	0.35 / 0.49 / 0.44
rs1002762	CDH26	20	0.11	2.61E-02	0.19 / 0.26 / 0.22
rs3212183	HNF4A	20	-0.11	2.61E-02	0.56 / 0.37 / 0.34
rs6103716	HNF4A	20	0.10	2.61E-02	0.36 / 0.33 / 0.3
rs11651038	SHISA6	17	0.13	2.91E-02	0.36 / 0.25 / 0.27
rs1027711	FAIM2	12	0.12	3.22E-02	0.14 / 0.05 / 0.06
rs349885	FAM19A2	12	-0.15	3.22E-02	0.12 / 0.08 / 0.06
rs6486713	TMEM132C	12	0.12	3.22E-02	0.84 / 0.13 / 0.08
rs1983474	CSMD1	8	0.14	3.91E-02	0.73 / 0.16 / 0.13
rs6027217	CDH26	20	-0.09	4.46E-02	0.6 / 0.3 / 0.33
rs6503916	YPEL2	17	-0.15	4.78E-02	0.34 / 0.21 / 0.21

Chr indica el cromosoma; Beta, efecto asociado a la variante; P, valor de p; MAF, Frecuencia del alelo menor en población europea (EUR), mexicanos de Los Ángeles (MXL) y los mexicanos de este proyecto (T2020).

## **DISCUSIÓN**

La dislipidemia es una patología caracterizada por un conjunto de alteraciones en el metabolismo de lípidos con etiología multifactorial, donde el componente genético puede explicar el 50% de la variación multifactorial en el desarrollo de dislipidemia <sup>17</sup>. Además, en México y poblaciones con ancestría amerindia, existe una predisposición genética de más del 40% en el desarrollo a dislipidemias <sup>19</sup>.

En este estudio se identificaron variantes genéticas asociadas con fenotipos dislipidémicos e índices aterogénicos, que proveen información sobre el estado de dislipidemia de un individuo y factores de riesgo cardiovascular <sup>14</sup>, a partir de una submuestra de la cohorte Tlalpan 2020, 316 (71.8%) casos y 124 (28.2%) controles. Esta clasificación en casos y controles se realizó de acuerdo con los lineamientos de La Sociedad Mexicana de Endocrinología consolidando una base de datos clínica y genómica para el desarrollo de modelos de asociación genotipo-fenotipo.

### **Variables clínicas de la población de estudio**

En cuanto a diferencias de género se observaron variaciones significativas para todos los lípidos plasmáticos entre hombres y mujeres, confirmando así reportes previos de las diferencias en el metabolismo de lípidos entre sexos. Existen varias hipótesis que intentan explicarlas, una de ellas es que las diferentes hormonas entre hombres y mujeres influye de manera diferencial al metabolismo de lípidos. Por otro lado, la influencia de algunos genes que escapan de la inactivación del cromosoma X podrían permitir la expansión de los adipocitos lo cual se ha observado en ratones hembra, aumentando la susceptibilidad del sexo femenino en el metabolismo anormal de lípidos <sup>29,30</sup>.

### **Estrategia gen candidato**

Se eligió una estrategia de enriquecimiento de genes y variantes tomando variantes con una diferencia significativa en cuanto a frecuencia alélica entre casos y a partir de un valor de significancia de  $P < 0.05$  corregido por Bonferroni. La lista consolidada de genes candidato potencialmente asociados a dislipidemia

representa una herramienta de referencia que incluye todos los genes identificados hasta ahora que se relacionan con el fenotipo de estudio <sup>3,5,6</sup>. El presente análisis de frecuencias alélicas confirmó la presencia de al menos el 90% de los genes reportados en la literatura, lo cual es indicativo de que la población mexicana comparte en gran medida el componente genómico de las dislipidemias con otras poblaciones del mundo.

Esta lista gen candidato incluyó 373 genes, 10990 variantes, a partir de los cuales se generaron modelos de asociación genotipo-fenotipo primero a través de un modelo de regresión lineal múltiple de los índices aterogénicos considerando covariables clínicas y genéticas así como el componente ancestral.

### **Analisis de asociación multivariable**

Se identificaron 3830 variantes asociadas con los diferentes fenotipos (índices aterogénicos, tipos de dislipidemia y concentración de lípidos en sangre), el 35% (1227) estaban relacionadas con los índices de riesgo aterogénico. De las cuales, más de la mitad (58%) son compartidas con las asociadas a la concentración de lípidos (TG, TC, C-LDL y C-HDL), pero con valores de P más significativos ( $< 1E-6$ ). Estos resultados confirman que los índices aterogénicos son capaces de predecir la dislipidemia con mayor precisión, en comparación con las concentraciones de lípidos en plasma como triglicéridos o colesterol total <sup>14,31,32</sup>. Por lo anterior, los índices aterogénicos junto con las 1227 variantes asociadas fueron usados para una segunda etapa de regresiones con el propósito de evaluar el efecto independiente de cada una de las variantes previamente asociadas ( $P < 0.05$ ) al fenotipo.

La asociación de dislipidemia con los genes *APOA5*, *APOA4*, *APOC3* y *LIPC* ya había sido previamente reportada, estos codifican para lipoproteínas y la lipasa pancreática, y también se han relacionado con dislipidemia y enfermedades coronarias en poblaciones asiáticas, europeas y mexicanas <sup>3,6,33,34</sup>. La variante más significativa para AIP, *APOC3* rs5128 ( $P = 7.5 E-8$ ) se sabe esta asociada con niveles altos de TG, TC y C-LDL para diferentes poblaciones, incluyendo caucásicas y asiáticas <sup>8</sup>. En esta etapa se identificaron 1227 variantes en 354

genes relacionados con los niveles de lípidos y los índices aterogénicos de las cuales 80 genes y 1100 variantes (90%) no se habían reportado anteriormente.

### **Análisis de asociación condicional**

Con el propósito de seleccionar variantes de la primera etapa que fueran independientes entre sí y que estuvieran asociadas a los índices aterogénicos, se realizó un segundo análisis usando un modelo de regresión condicional.

La regresión condicional identificó 29 variantes en 16 genes asociadas significativamente a AIP (16 variantes y 14 genes) e IC II (14 variantes y 13 genes;  $P_{FDR} < 0.05$ ) con un efecto independiente con otras variantes, solamente la variante *CDH26* rs6027217 se compartió entre los índices.

Al comparar las variantes asociadas en la primera etapa con las de la segunda, se observa que el 99% de la primera regresión “se pierden”, es decir disminuye su significancia debido a que el tamaño del efecto (beta) no era independiente respecto a otras también asociadas significativamente. La variante *CSMD1* rs12545377 (C>G>A) identificada en la primera regresión, es la única que se mantiene en el mismo rango de asociación significativa al AIP ( $P = 1E-2$  corregido por FDR ) conservando también el tamaño del efecto. Esta observación es indicativa de que esta variante en particular es independiente genéticamente de las demás asociadas en el mismo gen/cromosoma, lo que confirma que la significancia estadística a través del valor de P no está relacionada con la independencia en el efecto <sup>35</sup>.

Los genes significativos de la regresión condicional incluyen 29 variantes en 16 genes, *ANGPTL4*, *CDH26*, *CHODL*, *CSMD1*, *DSCAM*, *FAIM2*, *FAM19A2*, *HNF4A*, *LINGO2*, *MTSS1*, *MTUS1*, *MUC16*, *NDUFB9*, *SHISA6*, *TMEM132C* e *YPEL2*. Estos participan en el inicio de diversas rutas metabólicas que culminan en diferentes desórdenes asociados a síndrome metabólico, dislipidemias, obesidad y diabetes <sup>3,36,45–51,37–44</sup> (tablas 18 y 19). A continuación, se discute el posible papel

de dichas variantes en el desarrollo o asociación con dislipidemias y/o índice aterogénico.

### Variantes encontradas que forman haplotipo

Variantes en el gen *DSCAM* y *CHODL* del cromosoma 21 (rs57292936 (T>C), rs2837379 (C>A>G), rs760168 (T>C), rs77923858 (G>A) y rs77779037 (T>C)) se agrupan en un haplotipo, es decir están en desequilibrio de ligamiento rs77923858 con rs77779037 ( $D' = 99$ ,  $R = 0.96$ , valor de  $P < 1E-5$ ), y rs760168 con rs2837379 ( $D' = 1$ ,  $R = 1$ , valor de  $P < 1E-5$ ) como se observa en la figura 14.

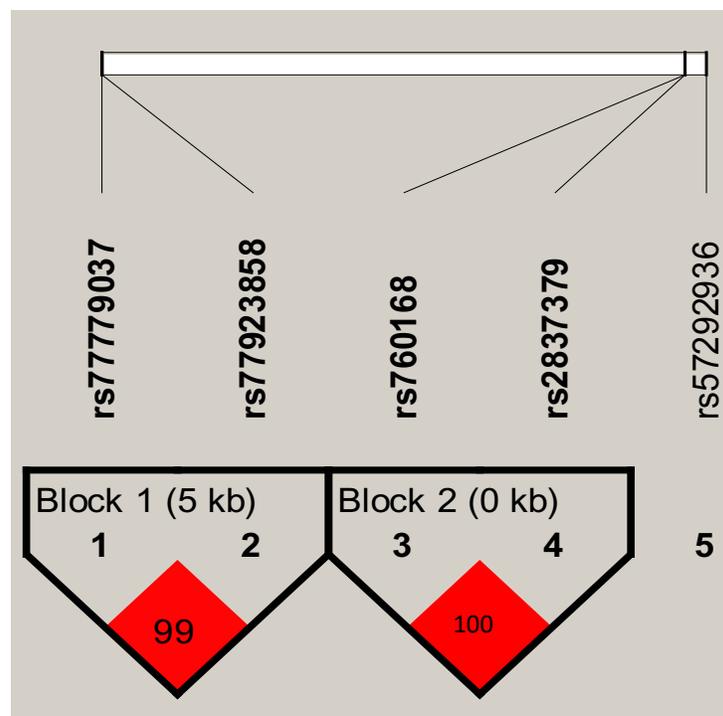


Figura 14. Desequilibrio de ligamiento en  $D'$  entre las variantes de *DSCAM* rs57292936, rs2837379, rs760168 y *CHODL* rs77923858, rs77779037 que fueron asociadas a índices aterogénicos.

Curiosamente las variantes intrónicas *DSCAM* rs57292936, rs2837379 y rs760168 tienen un efecto negativo (beta con signo negativo) es decir los alelos A, G y G respectivamente, están asociados a un mayor índice AIP y Castelli II, por lo que pueden ser consideradas como variantes de protección, mientras que los alelos A y C de las variantes intrónicas *CHODL* rs77923858 y rs77779037 tienen un efecto

positivo (beta con signo positivo) que puede ser considerado como factor de riesgo a ambos índices.

El gen *CHODL* codifica a la proteína chondrolectina, un péptido monomérico transmembranal con dominio CTLD, el cual pertenece a la superfamilia CLEC. La chondrolectina está implicada en el desarrollo neuronal, específicamente en el proceso por el cual las neuronas proyectan los axones <sup>37</sup>. Previamente, se demostró que *CHODL* interacciona con la subunidad beta de la rab geranilgeranil transferasa (*Rabggtb, idem*), la que a su vez cataliza la prenilación de las Rab GTPasas, involucradas en procesos de tráfico vesicular y de estructuras especializadas como autofagosomas y cuerpos lipídicos <sup>52</sup>. Aquí, se encontró la variante *CHODL* rs77923858, la cual no ha sido reportada en relación a síndrome metabólico o dislipidemias, la única variante reportada es *CHODL* rs77600076 relacionada a bulimia nerviosa en una cohorte de 2564 gemelas de Australia, <sup>53</sup>, y no está en desequilibrio de ligamiento con ninguna de las reportadas en este estudio. Por lo que serán necesarios estudios futuros para definir el impacto funcional de esta variante y su potencial implicación en dislipidemias.

El gen *DSCAM* codifica para la adhesina DSCAM “molécula de adhesión celular de Síndrome de Down”, una proteína transmembranal miembro de la superfamilia de inmunoglobulinas de adhesión (Ig-CAMs) involucrada en el desarrollo del Síndrome de Down, <sup>54</sup>. Un estudio reciente de Thomas Hummel, encontró que la proteína Dscam se asocia con esfingolípidos para la segregación y formación de los axones en *Drosophila* durante las etapas tempranas del desarrollo neuronal gracias a la comunicación entre la serina palmitoil transferasa (SPT) y Dscam. Mutaciones de la SPT ocasionan isoformas de Dscam que conducen a malformaciones en el desarrollo neuronal <sup>55</sup>. En humanos se ha reportado que alteraciones metabólicas de lípidos, como la dislipidemia, pueden originar cambios en el desarrollo neuronal <sup>38</sup>. Un ejemplo de esto es la neuropatía sensitiva autonómica hereditaria tipo 1 (HSAN-1) que es causada por variaciones en el gen *SPTLC1* que codifica para la subunidad de la SPT en humanos, es probable que estas mutaciones causen inespecificidad en la SPT, produciendo metabolitos que

modifiquen la segregación neuronal <sup>55</sup>. Lo anterior supone una relación entre el desarrollo neuronal y anomalías en el metabolismo de lípidos, ambas originadas por variaciones genéticas. Para *DSCAM* también, se han relacionado cambios en la organización e inervación axónica con el metabolismo de lípidos en tejido adiposo en donde la inervación axónica está involucrada en el almacenamiento y uso de energía, pues la noradrenalina procedente de las neuronas llega a los receptores  $\beta$ -adrenérgicos de los adipocitos, promoviendo la lipólisis <sup>56</sup>. Mueller et al. encontraron en 1906 que en las piernas de pacientes hemipléjicos, existe una mayor acumulación de adipocitos blancos. Al extraer médula espinal de perros, notó que su ausencia provoca una acumulación de los lípidos en extremidades, lo que denota la necesidad de la inervación axónica para el adecuado el metabolismo de lípidos <sup>57</sup>.

Las variantes *NDUFB9* rs3829038 (G>A) y rs77538674 (T>C) fueron asociadas con un tamaño de efecto positivo al índice aterogénico de plasma e índice de Castelli II (IC II) con una magnitud del efecto idéntico, misma frecuencia alélica y significancia estadística, por lo que se calculó el valor de desequilibrio de ligamiento, y se apreció que ambas están en perfecto desequilibrio (D=1, R=1, P<1E-5) es decir, estas variantes se heredan en conjunto, lo que explica en parte que compartan el mismo tamaño y sentido del efecto. Las variantes *CSMD1* rs11786902 y *MTSS1* rs6992199 están en el mismo cromosoma, pero no se encuentran en LD (Figura 15).

*NDUFB9* codifica para la subunidad NADH dehidrogenasa 1,3 subcomplejo 9 del sistema mitocondrial de fosforilación oxidativa (OXPHOS), necesario para la estabilidad y función del mismo <sup>36</sup>. Variaciones epigenéticas en, *Ndufb9* se han relacionado con una dieta alta en grasas (HFD) en ratones obesos. *Ndfub9* se encuentra silenciado a través de la hipermetilación lo cual podría desencadenar un desbalance en el metabolismo energético. Variantes de este gen también, se han identificado en pacientes obesos, con silenciamiento en genes que involucran a OXPHOS <sup>58</sup>. En biopsias de músculo se ha demostrado el silenciamiento de *NDUFB9* y su relación con desórdenes del almacenamiento de lípidos, lo cual

contribuye al desarrollo de dislipidemias <sup>59</sup>. Este gen es necesario para un correcto funcionamiento metabólico y energético a nivel de mitocondria <sup>60</sup>.

Otras variantes significativamente asociadas al índice de Castelli II y AIP, respectivamente, *CSMD1* rs11786902 y *MTSS1* rs6992199 ambas en el cromosoma 8 y no están en LD ( $P < 1E-3$ ). El gen *CSMD1* codifica para una proteína transmembranal, con 15 dominios de control complemental proteico (CCP) y con capacidad de reconocimiento a patógenos <sup>61</sup>. Variantes en este gen han sido asociadas a cáncer y enfermedades mentales, sin embargo existen reportes relacionando las variantes *CSMD1* rs1997137 y rs2930355 con el IMC, y *CSMD1* rs1997137 y rs7013078 con el síndrome metabólico <sup>62</sup>. Un estudio independiente encontró que existe una relación entre la variante *CSMD1* rs995322 y niveles altos de presión arterial en población coreana <sup>63</sup>. Las variantes encontradas en este estudio no han sido asociadas previamente al metabolismo de lípidos, sin embargo *CSMD1* rs11786902 relacionado aquí con el IC II, fue reportado en 2016 a un menor riesgo a Alzheimer en hispanos del Caribe <sup>64</sup>.

La proteína que codifica *MTSS1* pertenece a la familia de IMD [IRSp53 and MIM (missing in metastasis domain), con una función de plataforma para actina, implicada en carcinogénesis y metástasis. *MTSS1* se ha asociado a la organización del citoesqueleto promoviendo el ensamblado de filamentos de actina, y movilidad celular al favorecer la sobreexpresión de *RAC1-GTP* <sup>65</sup>. La expresión de *RAC1-GTP* se ha relacionado con la formación de placas ateroscleróticas por su efecto en la producción de citosinas proinflamatorias.

Además, *RAC1-GTP* está asociado con el tráfico, almacenamiento y metabolismo de colesterol en macrófagos, lo que puede implicar un desorden en el almacenamiento de grasas y conducir a dislipidemia <sup>66</sup>. Un estudio GWAS de cohorte europea que buscaba variantes genéticas relacionadas con factores de riesgos cardiovasculares, encontró una asociación entre la variante *MTSS1* rs891541 y el diámetro de la cintura <sup>39</sup>. Esta variante no está en LD con la aquí reportada. En este estudio *MTSS1* rs6992199 (G>A), se identificó asociada al

índice aterogénico de plasma con un tamaño de efecto protector de  $\beta=0.15$ . MTSS1 interactúa con NDUFB9 a nivel de membrana mitocondrial y este último al ser una proteína que participa en la cadena de electrones se le conoce como gen energético ya identificado en desordenes metabólicos <sup>67</sup>.

El resto de las 21 variantes significativamente asociadas a los índices aterogénicos se encuentran en genes que pueden estar relacionados con el metabolismo de lípidos, síndrome metabólico y otros. Aquí se discuten éstos en su conjunto:

**Genes relacionados con el metabolismo de lípidos: *HNF4A*, *ANGPTL4*, *FAIM2*, *LINGO2*, *MTUS1***

El gen *HNF4A* codifica para el factor hepático nuclear 4 alfa, es un factor de transcripción que se une al ADN en forma de homodímero y controla la expresión de genes como *HNF1A* que a su vez controla la expresión de GLUT1 y GLUT2 por lo que se ha asociado a diabetes. La variante *HNF4A* rs1885088 esta significativamente asociada con los niveles de LDL mostrado en un estudio piloto en pacientes del noreste de México <sup>40</sup> y esta variante se encuentra en desequilibrio de ligamiento (LD) con la reportada en este estudio (rs6103716), con un tamaño de efecto positivo de  $\beta = 0.1$  asociado al índice de Castelli II (LDL/HDL), lo que confirma el efecto del alelo C de la variante rs6103716 al riesgo de aumento en LDL en plasma, reportado por Gallardo-Blanco et al.

*ANGPTL4* codifica para angiopoyetina 4, proteína glicosilada secretada a la matriz extracelular, capaz de modular la actividad de la lipoproteinlipasa <sup>68</sup>, provocando el aumento en la concentración plasmática de estos lípidos y por tanto su distribución en el cuerpo, potencialmente generando un metabolismo anormal de lípidos (dislipidemia).

Se han encontrado algunas variantes que conducen a la pérdida de función de *ANGPTL4* por ejemplo, el rs116843064 que aumenta la sensibilidad a la insulina y el metabolismo de la glucosa, además modula positivamente la actividad de LPL,

lo que conduce a la disminución de los triglicéridos plasmáticos y C-LDL <sup>69</sup>. Adicionalmente, se encontró en estudios previos de asociación al gen *ANGPTL4* con un riesgo reducido de enfermedad coronaria <sup>41</sup>, diabetes tipo 2 <sup>70</sup> y trastornos cardiometabólicos en general <sup>69</sup>. Sin embargo, ninguna de estas variantes se encuentran en LD con la que se reportó aquí, *ANGPTL4* rs72996742 (G>A), una variante intergénica asociada con un tamaño de efecto de 0.11 puntos negativos sobre AIP. Esta es la primera vez que se informa de una variante (rs72996742) en *ANGPTL4* en mexicanos asociada al índice aterogénico de plasma.

*FAIM2* se expresa predominantemente en el sistema nervioso central, específicamente en las membranas celulares y está relacionado con una variedad de enfermedades como la isquemia, meningitis bacteriana y neuroblastoma. También se ha relacionado a *FAIM2* con patologías en el metabolismo de lípidos, por ejemplo, la variante *FAIM2* rs7138803 tiene un impacto en el aumento al riesgo de obesidad, presión arterial diastólica, frecuencia cardíaca, en población mexicana <sup>71</sup>. Sin embargo, esta variante no está en LD con *FAIM2* rs1027711 (A>G) que se encontró en este estudio asociada de manera positiva con 0.12 puntos para el índice de Castelli II, aumentando el riesgo al desarrollo de dislipidemias.

Se ha determinado que la expresión de *FAIM2* depende de la vía de fosfatidilinositol 3 cinasa – Akt (PI3K-Akt) <sup>72</sup>, la cual a su vez es conocida por su relevancia en la síntesis y secreción de TG <sup>73</sup>, y se encuentra en balsas lipídicas de membrana <sup>74</sup>. Es posible que *FAIM2* sea una proteína que participa río abajo en la vía PI3K-Akt que puede influir en la estabilidad lipídica y en el metabolismo de TG.

*LINGO2* codifica para la proteína homónima, la cual se expresa con mayor frecuencia en cerebro y se ha propuesto como regulador de la saciedad <sup>75</sup>. La variante *LINGO2* rs824248 fue asociada con diabetes tipo 2 en una cohorte de 3,848 casos y 4,366 controles en latinoamericanos y las variantes *LINGO2*

rs10968576 y rs1412239 con obesidad e IMC en individuos provenientes de Europa <sup>76</sup>.

Se han identificado diversas variantes en *LINGO2* que afectan la expresión de genes como *LIPF* que esta involucrado en el metabolismo de ácidos grasos libres, y que cuya variante ocasionada por un cambio de lectura por la delección de 19 pb (*LINGO2* rs587777699) se ha relacionado con dislipidemia <sup>44</sup>.

De forma interesante se demostró que existe una relación entre variantes del gen *LINGO2*, en específico un bloque intergénico localizado río arriba del gen, que ha sido asociado con obesidad y desarrollo neuronal incluyendo a *LINGO2* rs2150864, y la abundancia de *Blautia* en la microbiota intestinal de personas con obesidad, lo que sugiere una interacción directa o indirecta con *LINGO2* y algunas especies del género *Blautia*. Se ha propuesto que debido a que *LINGO2* se expresa con mayor frecuencia en el cerebro, variantes de este gen pueden afectar el comportamiento y la saciedad, lo que puede modificar no sólo la microbiota intestinal sino promover cambios en el metabolismo como la dislipidemia <sup>77</sup>.

Cabe mencionar que ninguna de las variantes previamente mencionadas se encuentran en LD con las variantes encontradas *LINGO2* rs7855149 (G>A) y rs10968632 (T>C), ambas intrónicas, asociadas con el índice aterogénico de plasma, pero con sentidos opuestos, ya que la primera tiene un tamaño de efecto negativo de 0.14 puntos y la segunda un efecto positivo de 0.12. Además las variantes aquí reportadas no han sido relacionadas previamente con poblaciones latinas o mexicanas asociadas con los índices aterogénicos o el desarrollo de dislipidemias.

Por otro lado, se asoció a *MTUS1* rs209568 (T>C), que es una variante de tipo cambio de sentido que provoca un cambio de Glutamina por Aspartato. Ésta se encontró relacionada con el índice aterogénico de plasma con un efecto positivo de 0.11 puntos, aumentando en el riesgo al desarrollo de dislipidemias.

*MTUS1* es un gen que se expresa durante la proliferación celular, variantes en este gen se han asociado con la inhibición del crecimiento tumoral en líneas

cancerosas, por lo que se le conoce como un supresor tumoral <sup>78</sup>. Sin embargo, *MTUS1* también juega un papel clave en la diferenciación de adipocitos cafés <sup>49</sup>, los cuales podrían combatir la obesidad y desordenes como la dislipidemia, al disminuir el almacenamiento de grasas, aumentando la energía celular por desacoplamiento de la cadena de electrones en mitocondria <sup>50</sup>.

Previamente, se observó que la isoforma 3 de *MTUS1*, es inducida en adipocitos cafés cuando se exponen a bajas temperaturas (19°C) en comparación con aquellas de temperaturas neutras (30°C) y en pruebas de *LOF* (Loss-of-function) se determinó que esta proteína es necesaria para la diferenciación celular de dichos adipocitos. Además, se encontró correlación entre los niveles del transcrito de *Mtus1* contra marcadores de la diferenciación de adipocitos café como son *Prmd16*, *Ppargc1a* y *Cidea* <sup>49 50</sup>.

### **Genes relacionados con el síndrome metabólico: MUC16, YPEL2, TMEM132C, FAM19A2**

Los genes *MUC16*, *YPEL2*, *TMEM132C* y *FAM19A2* fueron asociados a índices aterogénicos y se encontró que están involucrados en desordenes del metabolismo de carbohidratos ocasionados por diferentes vías que involucran mecanismos como apoptosis de células  $\beta$  pancreáticas, originados por exceso de glucosa o incluso por estrés auditivo <sup>43 49 50</sup>.

*MUC16* codifica para la isoforma CA-125 (antígeno 125 de cáncer) que pertenece a la familia de las mucinas glicoprotéicas y se expresa en células epiteliales de bronquios, endometrio, cornea y ovario <sup>79</sup>. Además de ser usado como biomarcador, se encontró una correlación entre los niveles plasmáticos de esta proteína con la edad, histerectomía y obesidad en pacientes post-menopáusicas sin cáncer de ovario. Aunado a esto, se determinó una correlación negativa de CA-125 en 13,000 mujeres coreanas con síndrome metabólico, niveles bajos de C-HDL y niveles altos de TG en plasma, lo que concuerda con lo esperado de pacientes dislipidémicos (Joo et al., 2011) <sup>43 49 50</sup>. En este estudio se identificó la

variante rs62118272 en este gen, sin embargo no se ha reportado ninguna variante que está asociada a síndrome metabólico o dislipidemia.

Al momento no se ha encontrado ningún reporte que relacione variantes de *MUC16* con dislipidemia o con los índices aterogénicos, sin embargo en este estudio se identificó la variante *MUC16* rs62118272 (C>T) asociada al índice aterogénico de plasma (AIP) que relaciona los niveles de TG y C-HDL, lo que concuerda con lo observado en el estudio de Joo et al., quienes también relacionaron la proteína expresada por *MUC16* y niveles de TG y C-HDL. La variante aquí reportada tiene tamaño de efecto positivo de 0.13 puntos, por lo que es una variante de riesgo por aumentar el AIP y por tanto incrementar la probabilidad del desarrollo a dislipidemia.

*YPEL2*, cuya proteína se expresa en el núcleo celular <sup>80</sup>, cuenta con sitios de unión a metales, es capaz de interactuar con fosfatasa y tiene influencia en la señalización por vesículas de calcio liberadas durante la repolarización celular <sup>81</sup>. *YPEL2* controla la expresión de *DUSP26*, que a su vez regula a *p53* evitando la ruta apoptótica celular.

En condiciones de alta glucosa (diabetes) y estrés auditivo se ha visto que se induce la transcripción de miR-200, el cual inhibe a *Ypel2*, provocando el silenciamiento de *DUSP26*, provocando así apoptosis de células beta pancreáticas y por tanto diabetes <sup>42,43</sup>. Además, se ha asociado el aumento en la expresión de *Ypel2* en ratones con hiperlipidemia, diabetes y obesidad <sup>82</sup>. En este estudio se identificó la variante rs6503916 (T>C) en este gen asociada con un tamaño de efecto protector de menos 0.15 puntos al índice de Castelli II. Esta variante no se ha reportado con antelación asociada a desordenes metabólicos como dislipidemia, pero se puede suponer que podría afectar al desarrollo de diabetes que a su vez promueve un ambiente de estrés oxidativo que puede desencadenar en dislipidemia.

*TMEM132C*, cuya variante intrónica rs6486713 fue asociada al índice de Castelli II, es un gen que ha sido pobremente estudiado. En el 2018 se reportó que esta

adhesina neuronal, tiene la función de conectar al medio extracelular con el citoesqueleto de actina intracelular. En estudios en ratas diabéticas se identificó una sobreexpresión de 3 veces en 138 genes, entre ellos el gen *Tmem132c*, por lo que se encuentra asociado a desordenes metabólicos, incluido la dislipidemia <sup>47</sup> (Joo et al., 2011) <sup>43 49 50</sup>.

*TFA2/FAM19A2* codifica a la proteína TFA2 que pertenece a la familia TFA, compuesta por 5 genes expresados predominantemente en el cerebro altamente homólogos que codifican pequeñas proteínas que son secretadas, y se postula que funcionan como quimiocinas o neurocininas específicas que actúan como reguladores de las células inmunes y nerviosas <sup>46</sup>. En modelos murinos *FAM19A2* estimula el comportamiento del apetito y aumenta el gasto de energía, lo que sugiere un posible papel en las funciones metabólicas como el catabolismo de lípidos <sup>46 47</sup> (Joo et al., 2011) <sup>43 49 50</sup>.

En un reporte de asociación del genoma completo, Walford et al. encontraron que la variante *FAM19A2* rs10506418 estaba relacionada con la sensibilidad a la insulina en pacientes diabéticos caucásicos de Estados Unidos con un tamaño de efecto positivo, lo que significa que es un factor de riesgo en el desarrollo de la diabetes, el cual forma parte del síndrome metabólico, que a su vez propicia un ambiente de estrés oxidativo que puede propiciar el desarrollo de otras alteraciones metabólicas como la dislipidemia <sup>45 46 47</sup> (Joo et al., 2011) <sup>43 49 50</sup>.

Aunado a las variantes previamente reportadas, aquí se encontró que la variante *FAM19A2* rs349885 estuvo asociada a una disminución de 0.15 puntos en el índice de Castelli II, es decir puede ser considerada posiblemente una variante de protección contra el desarrollo de la dislipidemia al modificar el apetito del paciente. Esta variante no estuvo en desequilibrio de ligamiento con ninguna otra relacionada con alteraciones o síndromes metabólicos.

**Otros genes: CDH26 y SHISA6**

Por último, se clasificaron los genes *CDH26* y *SHISA6* ya que no existe una ruta metabólica reportada por la cuál puedan participar en el desarrollo a dislipidemias.

Previamente, se descubrió que la proteína *CDH26* regula la estructura y la polaridad del citoesqueleto de las células epiteliales de las vías respiratorias al localizarse en la membrana apical y promover la formación de actina cortical <sup>83</sup>.

Previamente, se ha observado que la proteína codificada esta altamente expresada en personas con desordenes alérgicos del tejido gastrointestinal (gastroenteritis eosinofílica) y se caracteriza por concentraciones elevadas de citocinas y linfocitos, los cuales son activados a su vez por células del tracto gastrointestinal que promueven la inflamación y que esta indirectamente relacionada con el síndrome metabólico <sup>51</sup>. En este sentido, se ha determinado que personas en condiciones de estrés oxidativo, como son los síndromes metabólicos (dislipidemia, diabetes tipo 2 y obesidad), tienen inflamación en diferentes tejidos, pero aun no se elucida la ruta de esta asociación <sup>84</sup>.

Se encontró que la cadherina *CDH26* podría ser parte del mecanismo por el cual se produce la resolución de la inflamación para promover el retorno del tejido a la homeostasis <sup>51</sup>, lo cual puede ayudar en la protección contra el estrés oxidativo causado por la dislipidemia.

En este estudio se encontró a la variante intrónica *CDH26* rs1002762 (A>G) asociada al índice de Castelli II con un tamaño de efecto positivo de 0.11 y a la variante rs6027217 (A>G) asociada con un tamaño de efecto negativo a los dos índices aterogénicos con valores de 0.09 y 0.12 para el de Castelli II y el de plasma respectivamente, que otorgan un efecto protector de esta variante sobre el desarrollo en la dislipidemia.

La variante intrónica *SHISA6* rs11651038 (C>T) fue asociada a un tamaño de efecto positivo de 0.13 con el índice de Castelli II, que por tanto incrementa la probabilidad al desarrollo de dislipidemia. La proteína codificada por *SHISA6*

forma parte de los complejos AMPAR en el cerebro y estos tienen la función de mediar la transmisión sináptica, sin embargo, existe muy poca información acerca de la función de este gen <sup>85</sup>.

Recientemente se encontró que los perfiles de expresión de ciertos miRNAs podrían ser usados como blancos terapéuticos en personas con enfermedad coronaria y al estudiar dichos perfiles se encontró que uno de ellos se expresaba diferencialmente y estaba dirigido a *SHISA6* y otros dos a tres genes de la familia *TMEM*, que están involucrados en el metabolismo de lípidos <sup>86</sup>. Además, previamente se ha estudiado la relación causal entre la dislipidemia y el desarrollo de la enfermedad coronaria, por lo que puede existir una relación entre *SHISA6*, dislipidemia y enfermedad coronaria <sup>87</sup>.

#### Análisis Poblacional

Finalmente, se compararon algunas de las frecuencias del alelo menor (MAF) de las variantes más significativas ( $P < 1E-3$ ) entre las poblaciones EUR, MXL y T2020. De manera importante, se encontró que el haplotipo conformado por las variantes *DSCAM* rs2837379 y rs760168 ( $D'=1$ ), que tienen un valor de protección aterogénico está también en LD en poblaciones europeas, sin embargo este haplotipo tiene una menor frecuencia (MAF EUR= 0.38) , de hasta el 50%, en las poblaciones mexicanas de Los Ángeles (MAF MXL= 0.13, MAF T2020= 0.14) y de los voluntarios analizados en este proyecto (T2020).

En contraste, el haplotipo conformado por las variantes *NDUFB9* rs3829038 y rs77538674 ( $D'=1$ ), con un tamaño del efecto de 0.21, el cual muestra una MAF de 0.03 en población europeas, lo cual representa el doble en comparación con las poblaciones mexicanas (MAF MXL= 0.06 Y MAF T2020= 0.06).

Estas diferencias apoyan la noción de la variabilidad en el desarrollo de dislipidemias entre poblaciones y diferentes ambientes geográficos <sup>3,88</sup>, por lo que es importante que cada población consolide y ordene de forma jerárquica las

variantes genéticas relacionadas con la enfermedad, en este caso dislipidemias, de acuerdo a las frecuencias dictadas por la geografía.

Este estudio provee un listado de estas variantes, que podrá ser utilizado como referencia para la validación del impacto funcional de genes y sus variantes en el desarrollo de dislipidemias y sus comorbilidades.

## CONCLUSIONES

- Entre las 29 variantes, se encontraron dos haplotipos asociados a los índices aterogénicos con los valores de P más significativos, uno conformado por las *DSCAM* rs2837379 y rs760168, el otro por *NDUFB9* rs3829038 y rs77538674.
- El haplotipo en *DSCAM* está asociado estadísticamente con 0.23 puntos de protección contra el índice de Castelli II y es hasta dos veces más frecuente en la población europea con respecto a la mexicana.
- El otro haplotipo en *NDUFB9* está asociado estadísticamente con 0.21 puntos de riesgo con el índice de aterogénico de plasma y es hasta dos veces más frecuente en la población mexicana con respecto a la europea.
- De estas variantes, sólo *HNF4A* rs6103716 se encontró en desequilibrio de ligamiento con *HNF4A* rs1885088 que se asoció al aumento de C-LDL en mexicanos del norte.
- Se discutieron las inferencias y el posible papel de 16 genes asociados a índices aterogénicos es decir de riesgo a dislipidemia, los cuales se clasificaron según su participación en desordenes del metabolismo de lípidos (*NDUFB9*, *HNF4A*, *ANGPTL4*, *FAIM2*, *LINGO2*, *MTUS1*, *MTSS1*), síndrome metabólico (*CSMD1*, *MUC16*, *YPEL2*, *TMEM132C*, *FAM19A2*), desarrollo neuronal (*DSCAM*, *CHODL*) y otros (*SHISA6*, *CDH26*) que en conjunto explican parte de la variación interindividual que existe en la aparición de dislipidemia.
- Se obtuvo una base de datos genómica filtrada por diversos controles de calidad para estudios GWAS con 294,000 variantes para 426 voluntarios mestizos de la CDMX que contribuye a próximos estudios sobre la población mexicana y a nivel mundial para comprender las diferencias interpoblación.

## **PERSPECTIVAS**

El presente trabajo logró identificar 29 variantes que previamente no habían sido reportadas con asociación a índices aterogénicos o dislipidemia para ninguna población, lo que permite consolidar bases para próximos estudios encaminados a conocer la dislipidemia en la población mexicana. Como siguiente paso, se propone validar estas 29 variantes, por lo que se plantea continuar la colaboración con el Instituto Nacional de Cardiología y en específico el proyecto Tlalpan 2020, para obtener una nueva cohorte de mexicanos mestizos de la CDMX con la cual se validen las variantes encontradas aquí.

Así mismo tras haber validado estas variantes, se pueden realizar estudios de ganancia o pérdida de la función usando técnicas *in vitro*, que permitan determinar los mecanismos moléculares o las rutas bioquímicas afectadas por estas variantes/genes con el propósito de elucidar el mecanismo que conlleva a la dislipidemia.

## REFERENCIAS

1. Canalizo-miranda, E., Salas-anaya, J. A., Jara-espino, R. & Viniegra-osorio, A. Guía de práctica clínica Diagnóstico y tratamiento de las dislipidemias. **51**, (2013).
2. Goldberg, A. Dislipidemia - Trastornos endocrinos y metabólicos - Manual MSD versión para profesionales. Available at: <https://www.msmanuals.com/es/professional/trastornos-endocrinos-y-metabolicos/trastornos-de-los-lipidos/dislipidemia>. (Accessed: 3rd December 2018)
3. Teslovich, T. M. *et al.* Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466**, 707–713 (2010).
4. Turner, S. *et al.* Quality control procedures for genome wide association studies. *Curr. Proc. Hum. Genet.* **68**, 1–24 (2011).
5. Speliotes, E. K. *et al.* Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat. Genet.* **42**, 937–948 (2010).
6. Weissglas-Volkov, D. *et al.* Genomic study in Mexicans identifies a new locus for triglycerides and refines European lipid loci. *J. Med. Genet.* **50**, 298–308 (2013).
7. Ou, H. J. *et al.* Relationship of the APOA5/A4/C3/A1 gene cluster and APOB gene polymorphisms with dyslipidemia. *Genet. Mol. Res.* **14**, 9277–9290 (2015).
8. Song, Y. *et al.* Associations of the APOC3 rs5128 polymorphism with plasma APOC3 and lipid levels: a meta-analysis. (2015). doi:10.1186/s12944-015-0027-0
9. Alves-Bezerra, M. & Cohen, D. E. Triglyceride metabolism in the liver. *Compr. Physiol.* **8**, 1–22 (2018).
10. Mclaughlin, C. L. Development of novel therapeutic approaches for the reduction of apolipoprotein B expression. *Res. Gate* (2015). doi:10.13140/RG.2.1.3104.7840
11. Oram, J. F. ATP-Binding Cassette Transporter A1: A Cell Cholesterol Exporter That Protects Against Cardiovascular Disease. *Physiol. Rev.* **85**, 1343–1372 (2005).
12. Aguilar, C. A. *et al.* Diagnóstico y tratamiento de las dislipidemias: posición de la Sociedad Mexicana de Nutrición y Endocrinología. *Rev. Endocrinol. y Nutr.* **12**, 7–41 (2004).
13. WHO. Global\_BloodCholesterolPrevalence\_BothSexes\_2008.png (2027×1358). (2008). Available at: [http://gamapserver.who.int/mapLibrary/Files/Maps/Global\\_BloodCholesterol](http://gamapserver.who.int/mapLibrary/Files/Maps/Global_BloodCholesterol)

Prevalence\_BothSexes\_2008.png. (Accessed: 3rd December 2018)

14. Millán, J. *et al.* Lipoprotein ratios: Physiological significance and clinical usefulness in cardiovascular prevention. *Vasc. Health Risk Manag.* **5**, 757–765 (2009).
15. Dobiášová, M. & Frohlich, J. The plasma parameter log (TG/HDL-C) as an atherogenic index: Correlation with lipoprotein particle size and esterification rate in apob-lipoprotein-depleted plasma (FERHDL). *Clin. Biochem.* **34**, 583–588 (2001).
16. Zhu, M. & Zhao, S. *Candidate Gene Identification Approach: Progress and Challenges. International Journal of Biological Sciences* **3**, (2007).
17. Willer, C. J. *et al.* Newly identified loci that influence lipid concentrations and risk of coronary artery disease. *Nat. Genet.* **40**, 161–169 (2008).
18. Willer, C. J. *et al.* Discovery and Refinement of Loci Associated with Lipid Levels. *Nat. Genet.* **45**, 1274–1285 (2013).
19. Aguilar-Salinas, C. A. *et al.* Hypoalphalipoproteinemia in populations of Native American ancestry: An opportunity to assess the interaction of genes and the environment. *Curr. Opin. Lipidol.* **20**, 92–97 (2009).
20. Illumina. Infinium Global Screening Array-24 v1 . 0. 1–7 (2017).
21. Illumina. Infinium® Global Screening Array-24 v1.0.
22. Illumina. GenomeStudio Genotyping Module v2.0 User Guide. (2016).
23. Scharpf, R. B. *et al.* A multilevel model to address batch effects in copy number estimation using SNP arrays. *Biostatistics* **12**, 33–50 (2011).
24. Jafari, M. & Ansari-Pour, N. Why, when and how to adjust your P values? *Cell J.* **20**, 604–607 (2019).
25. Altman, D. G. Covariate Imbalance, Adjustment for. in *Wiley StatsRef: Statistics Reference Online* (John Wiley & Sons, Ltd, 2014). doi:10.1002/9781118445112.stat04929
26. Yang, J. *et al.* Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits Genetic Investigation of ANthropometric Traits (GIANT) Consortium 4 , DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium NIH Public Access Author Manuscript. *Nat Genet* **44**, 369–372 (2013).
27. Pant, S. D. *et al.* A principal component regression based genome wide analysis approach reveals the presence of a novel QTL on BTA7 for MAP resistance in holstein cattle. (2010). doi:10.1016/j.ygeno.2010.01.001
28. Hibar, D. P. *et al.* Principal components regression: Multivariate, gene-based

- tests in imaging genomics. in *Proceedings - International Symposium on Biomedical Imaging* 289–293 (2011). doi:10.1109/ISBI.2011.5872408
29. Palmisano, B. T., Zhu, L., Eckel, R. H. & Stafford, J. M. Sex differences in lipid and lipoprotein metabolism. (2018). doi:10.1016/j.molmet.2018.05.008
  30. Wang, X., Magkos, F. & Mittendorfer, B. Sex differences in lipid and lipoprotein metabolism: It's not just about sex hormones. *J. Clin. Endocrinol. Metab.* **96**, 885–893 (2011).
  31. Juárez-Pérez, C. A. *et al.* Increased cardiovascular risk using atherogenic index measurement among healthcare workers. *Arch. Med. Res.* **46**, 233–239 (2015).
  32. Wu, T. T., Gao, Y., Zheng, Y. Y., Ma, Y. T. & Xie, X. Atherogenic index of plasma (AIP): A novel predictive indicator for the coronary artery disease in postmenopausal women. *Lipids Health Dis.* **17**, 1–7 (2018).
  33. Fu, Q. *et al.* Effects of Polymorphisms in APOA4-APOA5-ZNF259-BUD13 Gene Cluster on Plasma Levels of Triglycerides and Risk of Coronary Heart Disease in a Chinese Han Population. *PLoS One* **10**, e0138652 (2015).
  34. Rocco, C. A., Mecikovsky, D., Aulicino, P., Bologna, R. & Sen, L. Hypercholesterolemia Is Associated with the Apolipoprotein C-III (APOC3) Genotype in Children Receiving HAART: An Eight-Year Retrospective Study. *PLoS One* **7**, 39678 (2012).
  35. Wang, K. & Abbott, D. A principal components regression approach to multilocus genetic association studies. *Genet. Epidemiol.* **32**, 108–118 (2008).
  36. Yoon, A., Tammen, S. A., Park, S., Han, S. N. & Choi, S.-W. Genome-wide hepatic DNA methylation changes in high-fat diet-induced obese mice. *Nutr. Res. Pract.* **11**, 105 (2017).
  37. Sleight, J. N. *et al.* Chondrolectin affects cell survival and neuronal outgrowth in in vitro and in vivo models of spinal muscular atrophy. *Hum. Mol. Genet.* **23**, 855–869 (2014).
  38. Lamari, F., Mochel, F., Sedel, F. & Saudubray, J. M. Disorders of phospholipids, sphingolipids and fatty acids biosynthesis: toward a new category of inherited metabolic diseases. *J. Inherit. Metab. Dis.* **36**, 411–425 (2013).
  39. Smith, E. N., Chen, W., Kähönen, M., Kettunen, J. & Lehtimäki, T. Longitudinal Genome-Wide Association of Cardiovascular Disease Risk Factors in the Bogalusa Heart Study. *PLoS Genet* **6**, 1001094 (2010).
  40. Gallardo-Blanco, H. L. *et al.* Genetic variants in KCNJ11, TCF7L2 and HNF4A are associated with type 2 diabetes, BMI and dyslipidemia in families

- of northeastern Mexico: A pilot study. *Exp. Ther. Med.* **13**, 523–529 (2017).
41. Dewey, F. E. *et al.* Inactivating Variants in ANGPTL4 and Risk of Coronary Artery Disease. *N. Engl. J. Med.* **374**, 1123–1133 (2016).
  42. Hashimoto, N. & Tanaka, T. Role of miRNAs in the pathogenesis and susceptibility of diabetes mellitus. *J. Hum. Genet.* **62**, 141–150 (2017).
  43. Kung, C.-P. & Murphy, M. E. The role of the p53 tumor suppressor in metabolism and diabetes. *J Endocrinol.* **231**, R61–R75 (2016).
  44. Ellis, K. L. *et al.* Common variants associated with changes in levels of circulating free fatty acids after administration of glucose–insulin–potassium (GIK) therapy in the IMMEDIATE trial. *Pharmacogenomics J.* **17**, 76–83 (2017).
  45. Walford, G. A. *et al.* Genome-Wide Association Study of Modified Stumvoll Insulin Sensivity Index Identifies BCL2 and FAM19A2 as Novel Insulin Sensitivity Loci. *Gemma Rojo-Martínez* **65**, 3200–3211 (2016).
  46. Okada, J. *et al.* Analysis of FAM19A2/TAFA-2 function. *Physiol. Behav.* **208**, (2019).
  47. Sárközy, M. *et al.* Transcriptomic alterations in the heart of non-obese type 2 diabetic Goto-Kakizaki rats. *Cardiovasc Diabetol* **15**, 110 (2016).
  48. Joo, N.-S., Kim, K.-N. & Kim, K. S. Serum CA125 Concentration has Inverse Correlation with Metabolic Syndrome. *J Korean Med Sci* **26**, 1328–1332 (2011).
  49. Shinoda, K. *et al.* Genetic and functional characterization of clonally derived adult human brown adipocytes HHS Public Access. *Nat Med* **21**, 389–394 (2015).
  50. Nascimento, E. B. M. *et al.* Genetic Markers of Brown Adipose Tissue Identity and In Vitro Brown Adipose Tissue Activity in Humans. *Obesity* **26**, 135–140 (2018).
  51. Caldwell, J. M. *et al.* Cadherin 26 is an alpha integrin-binding epithelial receptor regulated during allergic inflammation HHS Public Access. *Mucosal Immunol* **10**, 1190–1201 (2017).
  52. Zhen, Y. & Stenmark, H. Cellular functions of Rab GTPases at a glance. *J. Cell Sci.* **00**, 1–6 (2015).
  53. Wade, T. D. *et al.* Genetic variants associated with disordered eating. *Int. J. Eat. Disord.* **46**, 594–608 (2013).
  54. Agarwala, K. L., Nakamura, S., Tsutsumi, Y. & Yamakawa, K. Down syndrome cell adhesion molecule DSCAM mediates homophilic intercellular

- adhesion. *Brain Res. Mol. Brain Res.* **79**, 118–26 (2000).
55. Goyal, G. *et al.* Sphingolipid-dependent Dscam sorting regulates axon segregation. *Nat. Commun.* **10**, (2019).
  56. Caron, A., Lee, S., Elmquist, J. K. & Gautron, L. Leptin and brain-adipose crosstalks. *Nature Reviews Neuroscience* **19**, 153–165 (2018).
  57. Bartness, T. & Bamshad, M. Innervation of mammalian white adipose tissue: implications for the regulation of total body fat. *Am. J. Physiol* (1998).
  58. Dahlman, I. *et al.* Downregulation of Electron Transport Chain Genes in Visceral Adipose Tissue in Type 2 Diabetes Independent of Obesity and Possibly Involving Tumor Necrosis Factor-. *Diabetes* **55**, 1792–1799 (2006).
  59. Debashree, B. *et al.* Mitochondrial dysfunction in human skeletal muscle biopsies of lipid storage disorder. *J. Neurochem.* **145**, 323–341 (2018).
  60. Jung, U. J., Cho, Y.-Y. & Choi, M.-S. Apigenin Ameliorates Dyslipidemia, Hepatic Steatosis and Insulin Resistance by Modulating Metabolic and Transcriptional Profiles in the Liver of High-Fat Diet-Induced Obese Mice. (2016). doi:10.3390/nu8050305
  61. Athanasiu, L. *et al.* A genetic association study of CSMD1 and CSMD2 with cognitive function. *Brain. Behav. Immun.* **61**, 209–216 (2017).
  62. Nock, N. L. *et al.* BMC Proceedings Proceedings Defining genetic determinants of the Metabolic Syndrome in the Framingham Heart Study using association and structural equation modeling methods. *BMC Proc.* **3**, 50 (2009).
  63. Hong, K.-W. *et al.* Genetic variations in ATP2B1, CSK, ARSG and CSMD1 loci are related to blood pressure and/or hypertension in two Korean cohorts. *J. Hum. Hypertens.* **24**, 367–372 (2010).
  64. Escudero-Esparza, A. *et al.* Complement inhibitor CSMD1 acts as tumor suppressor in human breast cancer. *Oncotarget* **7**, 76920–76933 (2016).
  65. Dawson, J. C., Bruche, S., Spence, H. J., Braga, V. M. M. & Machesky, L. M. Mtss1 Promotes Cell-Cell Junction Assembly and Stability through the Small GTPase Rac1. *PLoS One* **7**, 31141 (2012).
  66. Adorni, M. P., Ronda, N., Bernini, F. & Favari, E. *Rac1 and Cholesterol Metabolism in Macrophage.* (2013).
  67. Seabury, C. M. *et al.* Genome-wide association study for feed efficiency and growth traits in U.S. beef cattle. *BMC Genomics* **18**, 386 (2017).
  68. Dijk, W. *et al.* Regulation of angiopoietin-like 4 and lipoprotein lipase in human adipose tissue. *J. Clin. Lipidol.* **12**, 773–783 (2018).

69. Aryal, B., Price, N. L., Suarez, Y. & Fernández-Hernando, C. ANGPTL4 in Metabolic and Cardiovascular Disease. *Trends Mol. Med.* **25**, 723–734 (2019).
70. Bailetti, D. *et al.* ANGPTL4 gene E40K variation protects against obesity-associated dyslipidemia in participants with obesity. *Obes. Sci. Pract.* **5**, 83–90 (2019).
71. Corella, D. (Dolores) *et al.* NOVEL ASSOCIATION OF THE OBESITY RISK-ALLELE NEAR FAS APOPTOTIC INHIBITORY MOLECULE 2 (FAIM2) GENE WITH HEART RATE AND STUDY OF ITS EFFECTS ON MYOCARDIAL INFARCTION IN DIABETIC PARTICIPANTS OF THE PREDIMED TRIAL. (BioMed Central, 2002).
72. Beier, C. P. *et al.* FasL (CD95L/APO-1L) Resistance of Neurons Mediated by Phosphatidylinositol 3-Kinase-Akt/Protein Kinase B-Dependent Expression of Lifeguard/Neuronal Membrane Protein 35. *J. Neurosci.* **25**, 6765–6774 (2005).
73. Liu, D. D. *et al.* Effects of inhibiting PI3K-Akt-mTOR pathway on lipid metabolism homeostasis in goose primary hepatocytes. *animal* **10**, 1319–1327 (2016).
74. Planells-Ferrer, L. *et al.* Fas apoptosis inhibitory molecules: more than death-receptor antagonists in the nervous system. *J. Neurochem.* **139**, 11–21 (2016).
75. Speakman, J. & Speakman, J. R. Functional Analysis of Seven Genes Linked to Body Mass Index and Adiposity by Genome-Wide Association Studies: A Review Graded effects of calorie restriction View project Dietary whey proteins and energy deficit View project Functional Analysis of Seven Genes Linked to Body Mass Index and Adiposity by Genome-Wide Association Studies: A Review. *Hum Hered* **75**, 57–79 (2013).
76. Su, T. *et al.* A genetic variant in LINGO2 contributes to the risk of gestational diabetes mellitus in a Chinese population. *J. Cell. Physiol.* **234**, 7012–7018 (2019).
77. Hall, A. B., Tolonen, A. C. & Xavier, R. J. Human genetic variation and the gut microbiome in disease. *Genome Med.* **10**, (2018).
78. Di Benedetto, M. *et al.* Structural organization and expression of human MTUS1, a candidate 8p22 tumor suppressor gene encoding a family of angiotensin II AT2 receptor-interacting proteins, ATIP. *Gene* **380**, 127–136 (2006).
79. Felder, M. *et al.* MUC16 (CA125): tumor biomarker to cancer therapy, a work in progress. (2014).
80. Hosono, K., Sasaki, T., Minoshima, S. & Shimizu, N. Identification and

characterization of a novel gene family YPEL in a wide spectrum of eukaryotic species. *Gene* **340**, 31–43 (2004).

81. Münzel, T. *et al.* The Adverse Effects of Environmental Noise Exposure on Oxidative Stress and Cardiovascular Risk. *Antioxid. Redox Signal* **28**, 873–908 (2018).
82. Stewart, T. P., Kim, H. Y., Saxton, A. M. & Kim, J. H. *Genetic and genomic analysis of hyperlipidemia, obesity and diabetes using (C57BL/6J × TALLYHO/JngJ) F2 mice.* (2010). doi:10.1186/1471-2164-11-713
83. Lachowicz-Scroggins, M. E. *et al.* Cell Discovery Cadherin-26 (CDH26) regulates airway epithelial cell cytoskeletal structure and polarity. *Cell Discov.* **4**, 7 (2018).
84. Marques, P. *et al.* Systemic Inflammation in Metabolic Syndrome: Increased Platelet and Leukocyte Activation, and Key Role of CX3CL1/CX3CR1 and CCL2/CCR2 Axes in Arterial Platelet-Proinflammatory Monocyte Adhesion. *J. Clin. Med.* **8**, 708 (2019).
85. Klaassen, R. V *et al.* Shisa6 traps AMPA receptors at postsynaptic sites and prevents their desensitization during synaptic activity. *Nat. Commun.* (2016). doi:10.1038/ncomms10682
86. Zhou, Z. *et al.* Profiling of plasma circulating miRNA in coronary heart disease patients detected by next-generation small RNA sequencing. *Int J Clin Exp Med* **10**, (2017).
87. T.KUO, P. Dyslipidemia and coronary artery disease. *Clin. Cardiol.* **17**, 519–527 (1994).
88. Li, C. *et al.* Common rs7138803 variant of FAIM2 and obesity in Han Chinese. (2013). doi:10.1186/1471-2261-13-56

## ANEXO. Geles de agarosa

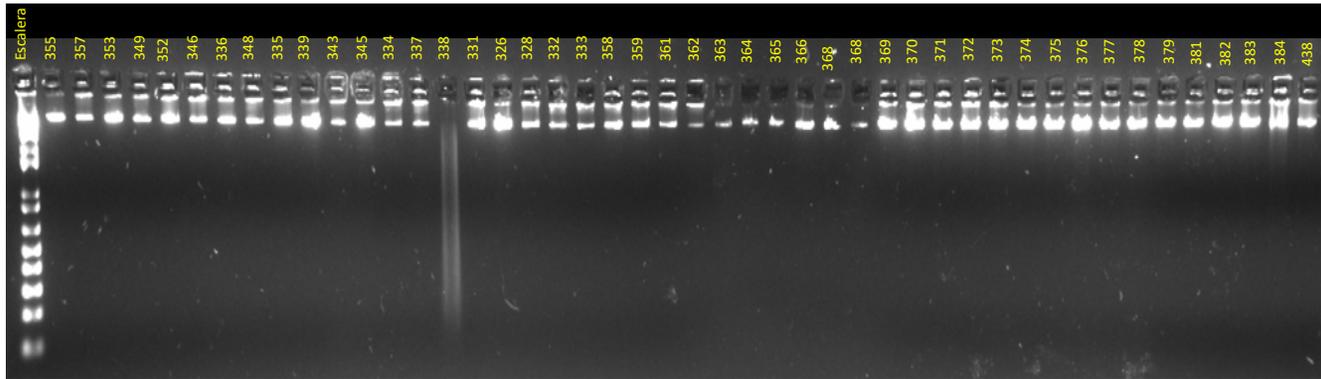


Figura I. Gel de agarosa al 2%. Carril 1º, escalera de peso molecular 1 Kb, seguido del identificador para cada muestra.

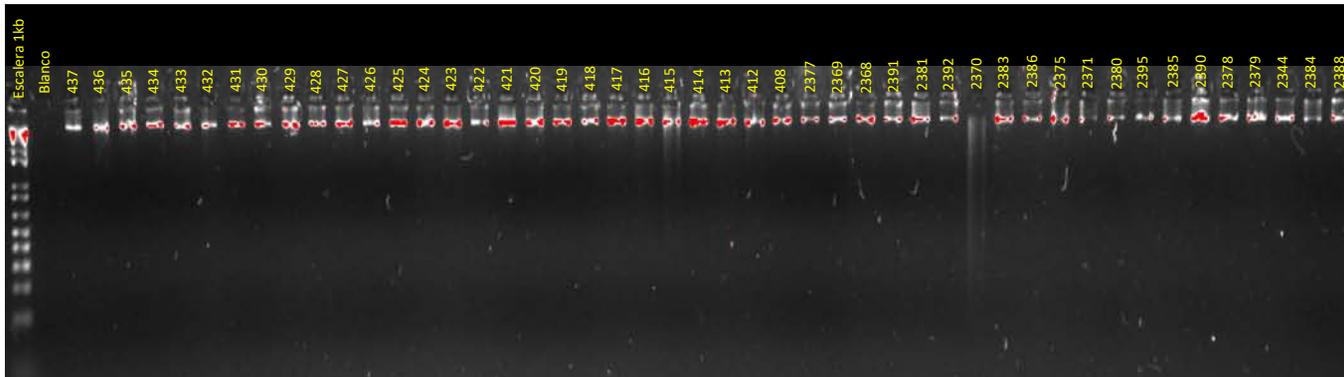


Figura II. Gel de agarosa al 2%. Carril 1º, escalera de peso molecular 1 Kb, seguido del identificador para cada muestra.

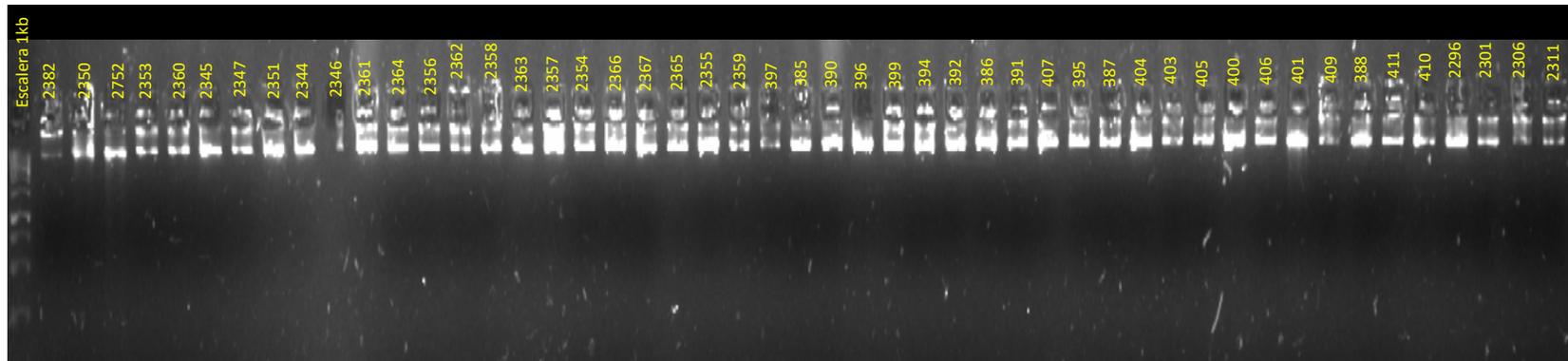


Figura III. Gel de agarosa al 2%. Carril 1º, escalera de peso molecular 1 Kb, seguido del identificador para cada muestra.

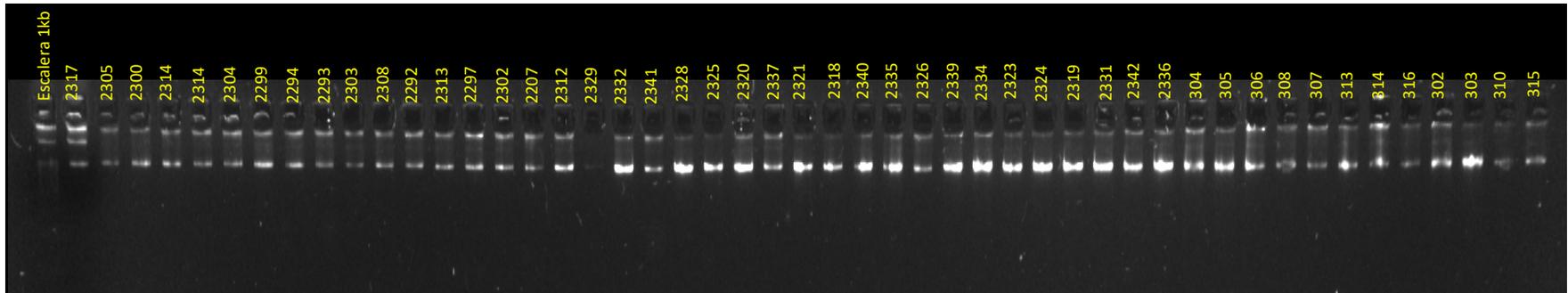


Figura IV. Gel de agarosa al 2%. Carril 1º, escalera de peso molecular 1 Kb, seguido del identificador para cada muestra.

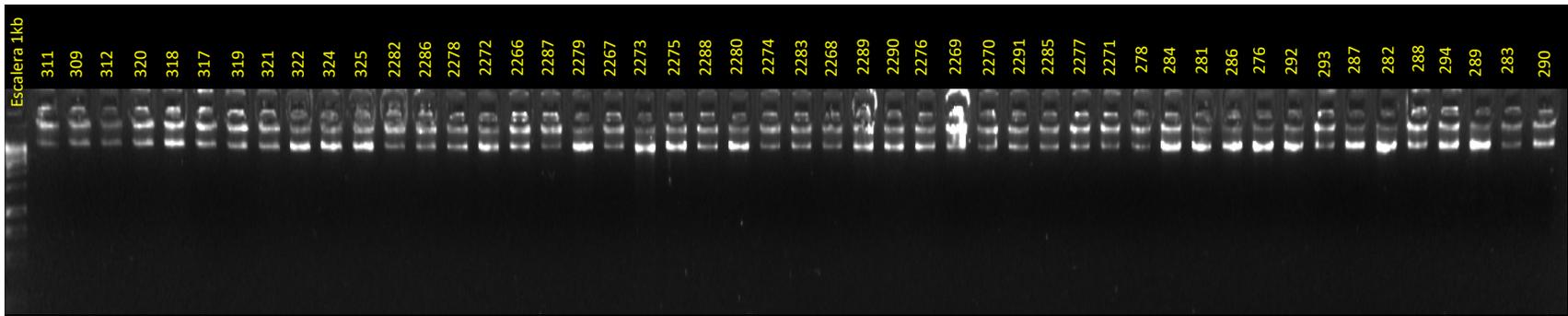


Figura V. Gel de agarosa al 2%. Carril 1º, escalera de peso molecular 1 Kb, seguido del identificador para cada muestra.

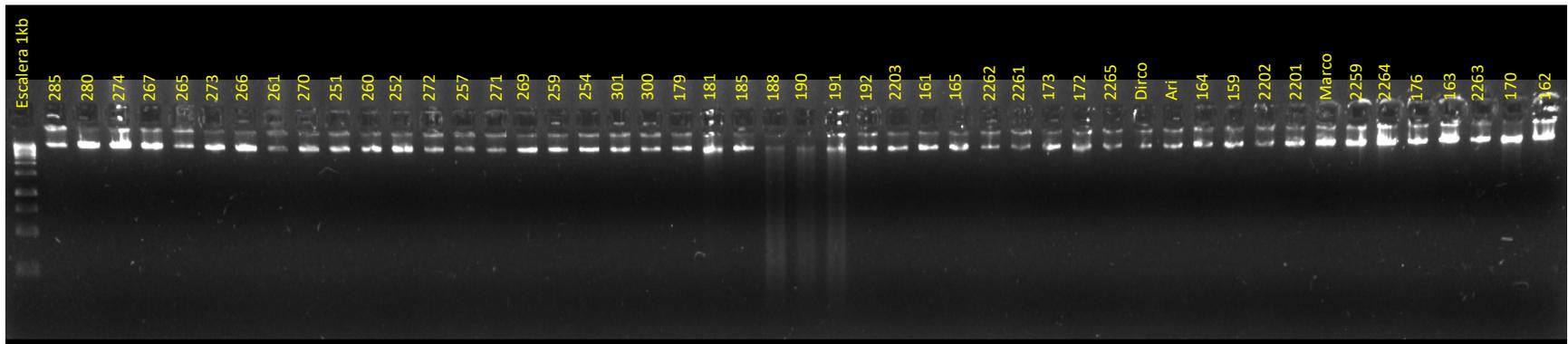


Figura VI. Gel de agarosa al 2%. Carril 1º, escalera de peso molecular 1 Kb, seguido del identificador para cada muestra.

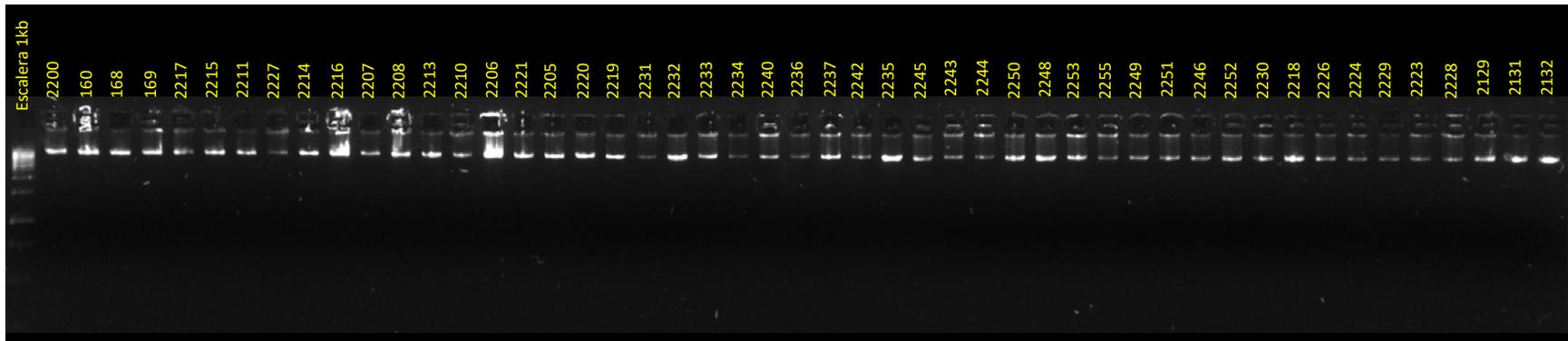


Figura VII. Gel de agarosa al 2%. Carril 1º, escalera de peso molecular 1 Kb, seguido del identificador para cada muestra.

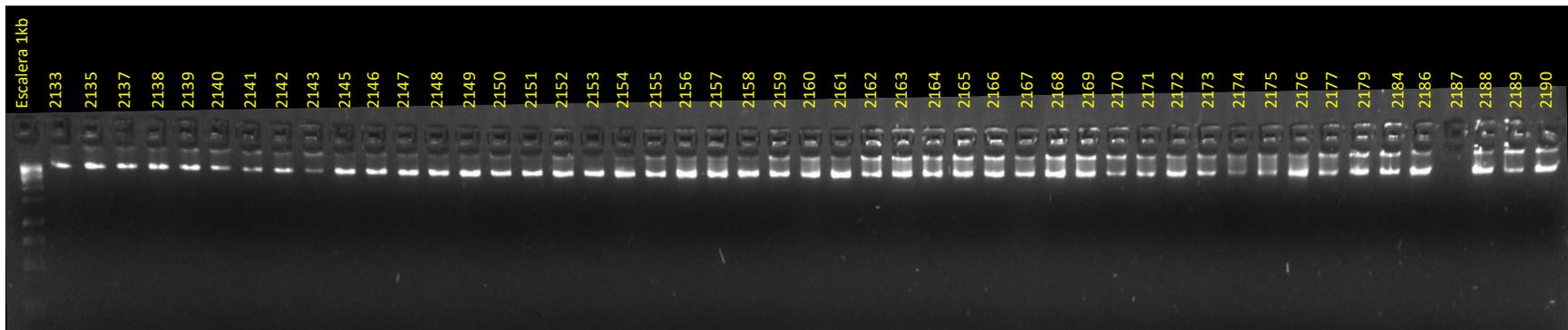


Figura VIII. Gel de agarosa al 2%. Carril 1º, escalera de peso molecular 1 Kb, seguido del identificador para cada muestra.

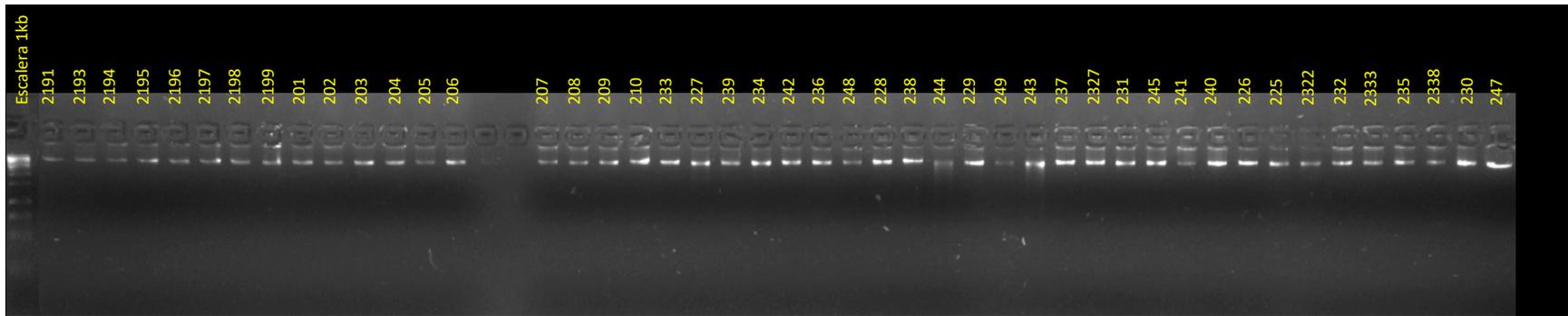


Figura IX. Gel de agarosa al 2%. Carril 1º, escalera de peso molecular 1 Kb, seguido del identificador para cada muestra.

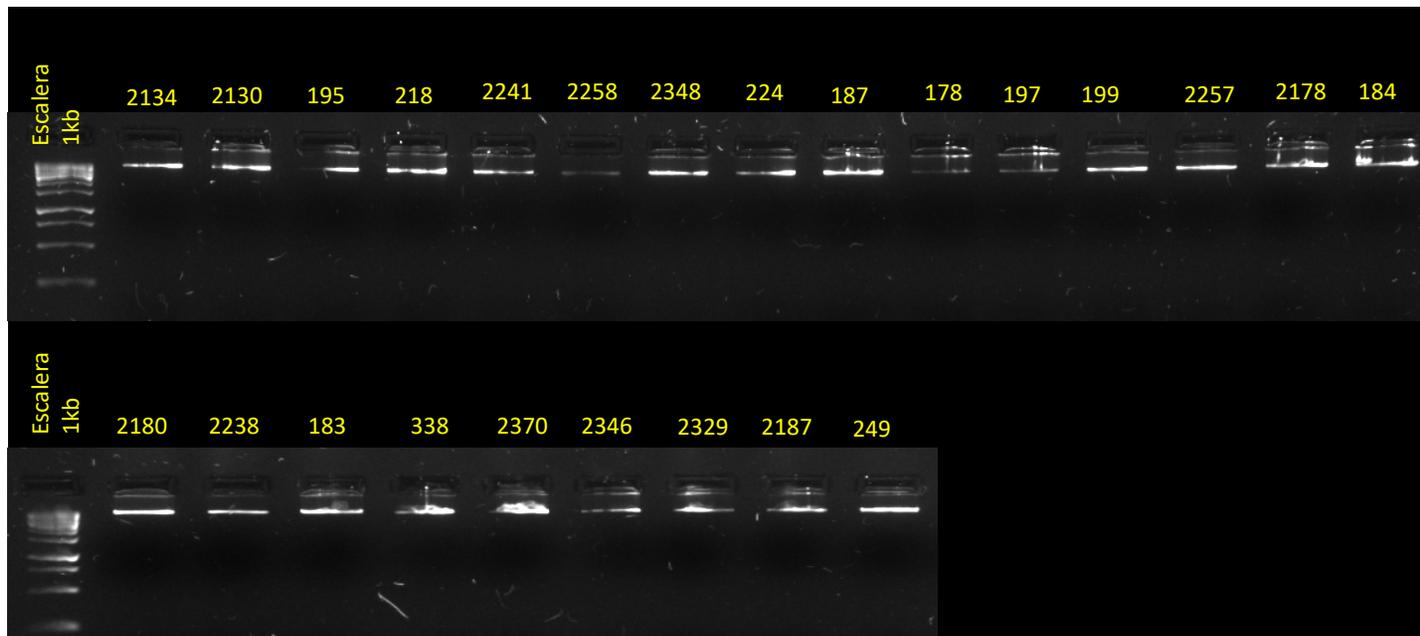


Figura X. Gel de agarosa al 2%. Carril 1º, escalera de peso molecular 1 Kb, seguido del identificador para cada muestra.