UNIVERSIDAD AUTONOMA DE BAJA CALIFORNIA

INSTITUTO DE INGENIERIA

“ESTUDIO DE ASOCIACIÓN DEL GENOMA MITOCONDRIAL CON DIABETES MELLITUS TIPO 2”

TESIS

PARA OBTENER EL GRADO DE:

DOCTOR EN CIENCIAS

PRESENTA

JULIO ALEJANDRO VALDEZ GONZALEZ

DIRECTOR

DR. RAFAEL VILLA ANGULO

CO-DIRECTOR

DR. PEDRO MAYORGA ORTIZ

Mexicali, Baja California Enero, 2024

Agradecimientos

A gradezco a mi familia por su apoyo incondicional y por impulsarme en mi desarrollo profesional, haciéndome notar la importancia de la educación y la superación académica para lograr mis objetivos, gracias por estar siempre para mí y a todos aquellos que me animaron a avanzar y a crecer como persona.

Gracias a mis padres por ser los principales promotores de mis sueños, gracias a ellos por cada día confiar y creer en mí, por siempre desear y anhelar lo mejor para mi vida, gracias a mi madre por todas sus atenciones mientras yo me dedicaba a estudiar; gracias a mi padre por cada consejo y por cada una de sus palabras que me guían durante mi vida.

Gracias a mi esposa por apoyarme en continuar con mis estudios, por la paciencia que ha tenido mientras yo estaba estudiando y trabajando en la tesis, ya que tuvimos sacrificar varios de nuestros momentos juntos porque tenía algo en que trabajar para avanzar en el doctorado.

Gracias a Dios por la vida de mis padres, también porque cada día bendice mi vida con la hermosa oportunidad de estar y disfrutar al lado de las personas que sé que más me aman, y a las que yo sé que más amo en mi vida, gracias a Dios por permitirme amar a mis padres.

Me gustaría agradecer sinceramente a mi asesor de Tesis, Dr. Rafael Villa Angulo, por todo su esfuerzo y dedicación que me ha brindado a lo largo del desarrollo de mi tesis. Sus conocimientos, sus orientaciones, su manera de trabajar, su persistencia, su paciencia y su motivación han sido fundamentales para mi formación como investigador. Le agradezco la confianza brindada y la oportunidad para realizar este trabajo, así como su tiempo, enseñanza y paciencia. A su manera, ha sido capaz de ganarse mi lealtad y admiración, así como sentirme en deuda con él por todo lo recibido durante el periodo de tiempo que ha durado esta Tesis.

Así mismo expreso mi gratitud a la Universidad Autónoma de Baja California (UABC) y al programa de becas CONACYT por hacer posible el realizar este doctorado, al haberme apoyado económicamente, dándome la oportunidad de cumplir este sueño.

"Si he logrado ver más lejos, ha sido porque he subido sobre los hombros de 2 gigantes"

Resumen

La diabetes tipo 2 (DT2) es un trastorno complejo y multifactorial. Actualmente está catalogada como una de las epidemias más significativas del siglo XXI por la magnitud con la que crece, así como por los efectos cardiovasculares que provoca. Este trabajo presenta un análisis completo del genoma para la asociación de posiciones de pares de bases individuales del genoma mitocondrial humano con DT2. Los datos consistieron en 510 genomas mitocondriales completos, de los cuales 437 eran de pacientes con diabetes tipo 2 y 73 eran de pacientes de control. Primero, se realizó un alineamiento múltiple que permitió la visualización y selección de una región con variación alélica. Luego, se utilizó un análisis de componentes principales para visualizar la estructura de los datos. A continuación, se realizó un análisis de regresión logística para descubrir 3 posiciones de pares de bases asociadas con la diabetes tipo 2. Un análisis de Odds Ratio reveló tres posiciones asociadas como factores de riesgo para la diabetes tipo 2. Una inspección de la anotación del genoma mitocondrial identificó tres genes CYTB, TRNP y TRNT asociados con la diabetes tipo 2. El gen TRNT se informó anteriormente como asociado, mientras que los genes CYTB y TRNP se identificaron en este trabajo como asociados con la diabetes tipo 2 por primera vez. Este estudio proporciona nueva evidencia de asociación con dos genes nuevos, que satisfacen los criterios para ser factores de riesgo de diabetes tipo 2.

INDICE

Lista de Figuras

Lista de Tablas

CAPÍTULO I INTRODUCCIÓN

En este capítulo se presenta una descripción básica sobre la enfermedad de Diabetes Mellitus tipo II. Se describe en que consiste, cual es la genética que la caracteriza y se presentan algunos datos genéticos sobre la enfermedad en la sociedad mexicana. En adición, se presenta el planteamiento del problema y los objetivos y metas trazados para la realización de este trabajo de tesis.

1.1 Diabetes Tipo 2

La Diabetes Mellitus Tipo 2 (DM2) es un trastorno complejo y multifactorial caracterizado por hiperglucemia crónica debido a la interacción de múltiples variantes genéticas y varios factores ambientales. Como resultado del envejecimiento de la población y la creciente prevalencia de obesidad e inactividad física, el número de pacientes con diabetes tipo 2 ha aumentado notablemente en todo el mundo [1]. Es catalogada como una de las epidemias del siglo XXI, tanto por su creciente magnitud como por su impacto negativo en la enfermedad cardiovascular [2].

En el año 2000 fueron afectados 171 millones de personas en el mundo y durante 2011 hubo 366 millones; actualmente existen más de 340 millones de afectados, en el 90% se manifiesta como una DM2. Se proyecta que para el año 2030 exista un aumento de 552 millones de diabéticos, con un incremento de la prevalencia en países en desarrollo [2].

La DM2 es una enfermedad heterogénea, de etiología multifactorial, en la que se combinan la resistencia a la insulina y la inadecuada secreción de insulina compensatoria por células beta del páncreas; se manifiesta como una hiperglucemia crónica, acompañada por trastornos del metabolismo de carbohidratos, grasas y proteínas. La susceptibilidad de esta enfermedad está determinada por el efecto combinado de factores genéticos y ambientales [2].

El ambiente se refiere a todos los factores no genéticos que modulan el fenotipo y puede incluir tanto factores del ambiente aleatorio (climáticos, geográficos, demográficos y socioeconómicos) como el denominado estilo de vida (dieta, tabaquismo, alcoholismo y actividad física), que el individuo puede modificar [2].

La interacción biológica se define como el efecto de dos factores que actúan unidos en una reacción física o química directa y la co-participación de dos o más de ellos en el mecanismo causal de la

enfermedad, mientras que la interacción genoma-ambiente significa alguna clase de acción de influencia recíproca entre los factores genéticos y ambientales [2].

Esta enfermedad cobra mayor importancia por su morbi-mortalidad, fenómeno que podría estar relacionado al envejecimiento de la población mundial, a su crecimiento especialmente en grupos étnicos con una mayor susceptibilidad a la enfermedad, al incremento de personas obesas como consecuencia de estilos de vida cada vez más sedentarios y a un mayor consumo de comidas con un alto contenido energético, así como a los cambios sociales y factores de riesgo asociados [2].

La diabetes es un trastorno metabólico crónico que afecta negativamente la capacidad del cuerpo para fabricar y utilizar insulina, una hormona necesaria para la conversión de los alimentos en energía. La enfermedad aumenta en gran medida el riesgo de ceguera, enfermedad cardíaca, insuficiencia renal, enfermedad neurológica. La DM2 es una enfermedad en la que la insulina se secreta de forma anormal o no actúa correctamente, lo que conduce a un aumento de la glucosa en sangre [3]. Con el tiempo, los niveles elevados de glucosa pueden provocar daños en múltiples órganos. La diabetes es la principal causa de insuficiencia renal crónica, ceguera en adultos y amputación de extremidades, y es un factor de riesgo importante de enfermedad cardíaca, accidente cerebrovascular y defectos de nacimiento [4]. Se cree que la DM2 es una enfermedad multifactorial, es decir, está influenciado por factores genéticos y ambientales. Las personas con antecedentes familiares de la enfermedad tienen un mayor riesgo de desarrollarla, ya que comparten antecedentes genéticos y probablemente comparten entornos similares. Se ha estimado que el 30% -70% del riesgo de diabetes tipo 2 puede atribuirse a la genética, con múltiples genes involucrados y diferentes combinaciones de genes que desempeñan funciones en diferentes subconjuntos de individuos [4]. Aún no se sabe cuántos genes están involucrados o cuánto control ejerce cada uno sobre el desarrollo de la enfermedad, pero investigaciones recientes han identificado una serie de candidatos prometedores [4].

La DM2 es consecuencia de una compleja interacción entre múltiples genes y diversos factores ambientales, aún no completamente entendidos, y se caracteriza por defectos en la secreción y la acción de la insulina que conducen a la hiperglucemia.

1.2 Genética de la Diabetes Tipo II

La enfermedad se considera un trastorno poligénico, en el que cada variante genética confiere un efecto parcial y aditivo. Sólo del 5-10% de los casos de DM2 se deben a defectos de un solo gen; estos incluyen la diabetes de inicio en la madurez de los jóvenes, los síndromes de resistencia a la insulina, la diabetes mitocondrial y la diabetes neonatal [5]. El examen de los genes de susceptibilidad a la DM2 puede ser útil para la predicción, la prevención y el tratamiento temprano de la enfermedad.

Después de estudios previos de asociación de genoma completo (GWAS, por sus siglas en ingles), el número de variantes genéticas comunes replicadas asociadas con la DM2 ha aumentado rápidamente [6-12]. Además, se han identificado más de 40 loci genéticos asociados a la DM2; sin embargo, estos loci se han revelado principalmente sobre la base de investigaciones de individuos europeos [13]. Los genomas identificados solo explican una pequeña proporción de la heredabilidad estimada de la enfermedad, lo que sugiere que quedan por identificar factores genéticos adicionales. Una limitación de los GWAS es la gran cantidad de hipótesis y el alto costo económico de las investigaciones [14]. Varios estudios han abordado la viabilidad y eficacia de los GWAS basados en agrupaciones, con ahorros considerables en tiempo y costo [14-16]. Además, la secuenciación del genoma completo en múltiples muestras en una población brinda una oportunidad sin precedentes para caracterizar de manera integral las variantes polimórficas en las poblaciones [17].

Aunque la contribución genética a la diabetes tipo 2 es bien reconocida, ahora hay al menos 19 loci que contienen genes, que se sabe que aumentan el riesgo de la enfermedad, incluidos PPARG, KCNJ11, KCNQ1, CDKAL1, CDKN2A-2B, CDC123-CAMK1D, MTNR1B, TCF7L2, TCF2 (HNF1B), HHEX-KIF11-IDE, JAZF1, IGF2BP2, SLC30A8, THADA, ADAMTS9, WFS1, FTO, NOTCH2 y TSPAN8 [5]. Hasta la fecha, un conjunto actual de 66 loci de susceptibilidad establecidos, identificados principalmente a través de GWAS a gran escala [5, 11, 18-25], engloba, como máximo, el 10% de la agregación familiar de la enfermedad. De los loci de susceptibilidad establecidos actualmente, nueve están contenidos en 19 genes.

Como ya se ha comentado la diabetes tipo 2 generalmente es causada por una combinación de factores genéticos y ambientales. El componente genético es más fuerte que en la diabetes tipo 1: es casi seguro que el gemelo de un paciente afectado por diabetes tipo 2 desarrollará la enfermedad.

Otros factores determinantes son la dieta y el ejercicio físico. Cuando los alimentos escasean, por ejemplo, la incidencia de diabetes tipo 2 es muy baja [26].

Los indios Pima proporcionan un buen ejemplo de la complementariedad entre el origen genético y el estilo de vida. Los que viven en México tienen una tasa de incidencia de diabetes de alrededor del 8%, mientras que los que emigraron a los Estados Unidos, donde la vida es más sedentaria y tienen acceso a alimentos ricos en energía (es decir, grasas), tienen una incidencia de diabetes de hasta el 50% [26].

El principal factor de riesgo de la DM2 es la obesidad. Los estudios epidemiológicos han demostrado que, en comparación con las personas delgadas, los hombres y mujeres obesos (con un índice de masa corporal > 35) tienen, respectivamente, 60 y 90 veces más probabilidades de desarrollar DM2. En términos genéticos, la DM2 es una enfermedad multifactorial, ya que su aparición no se puede atribuir a un solo gen [26].

A diferencia de los pacientes con DM2, los pacientes prediabéticos (que padecen resistencia a la insulina) no desarrollan hiperglucemia en ayunas. Sin embargo, cuando se somete a una prueba de tolerancia a la glucosa (oGTT), que implica la ingestión de 75 g de glucosa, estos pacientes se caracterizan por tener niveles muy elevados de glucosa en sangre [26].

Durante un breve período de tiempo, las células β pancreáticas producen una gran cantidad de insulina para contrarrestar la resistencia a la insulina. Ésta es la razón por la que muchos pacientes prediabéticos presentan niveles elevados de insulina en el plasma. Sin embargo, en la mayoría de los casos, el porcentaje de muerte de las células β anula el de la regeneración de nuevas células, lo que da como resultado una disminución de las células β productoras de insulina. Cuando la productividad de la insulina en el páncreas es incapaz de satisfacer el mayor requerimiento de la hormona debido a la resistencia a la insulina, el paciente prediabético desarrolla DM22 [26].

Tres factores principales contribuyen a la hiperglucemia [26]:

1. Resistencia a la insulina en el músculo, que provoca una disminución de la captación de glucosa del torrente sanguíneo;
2. Secreción defectuosa de insulina por el páncreas;
3. Aumento de la producción de glucosa en el hígado debido a la resistencia a la insulina hepática.

Esta enfermedad ha sido considerada durante muchos años la pesadilla de los especialistas en genética, recientemente se han producido considerables avances en el conocimiento de la genética de la enfermedad, con la disponibilidad de datos procedentes de los estudios GWAS, reforzados por el desarrollo de plataformas de genotipificación de alta resolución, la profusión de SNPs (Polimorfismos de Nucleótido Simple) en bases de datos públicas, los análisis de numerosas cohortes de pacientes y la generación de herramientas de análisis muy sofisticadas [2].

La abundante evidencia que apoya la base genética de la enfermedad procede de estudios de población, de familiares y de hermanos gemelos. Su prevalencia varía considerablemente entre grupos étnicos que comparten el mismo ambiente (en los Estados Unidos es de dos a seis veces más prevalente en afroamericanos, indios Pima e hispanos que en la población de raza blanca).

Estudios realizados plantean que la presencia de familiares afectados de primer grado se debe a la predisposición genética, es decir, a la probabilidad de que sus genotipos sean más parecidos.

Los hijos de un progenitor diabético tienen un 40% de riesgo de desarrollar diabetes mellitus tipo 2 (el que es mayor si el progenitor afectado es la madre en vez del padre), frente al riesgo existente en la población de un 7% y, si ambos padres son diabéticos, el riesgo aumenta a un 70%.10 Según otros autores existe un riesgo del 38% a los 80 años si uno de los padres está afectado y del 60% a los 60 años si ambos padres están afectados [2].

Los hallazgos de la agregación preferencial de la DM2 por vía materna pudieran estar relacionados con estudios realizados que asocian varias mutaciones en el genoma mitocondrial con la diabetes y la pérdida neurosensorial de la audición. La mutación más común en el gen del ARNt mitocondrial (Leu-UUR), A3243G, asocia la diabetes mellitus con la herencia materna. Muchas otras mutaciones en el genoma mitocondrial han sido descritas en familias con diabetes mellitus tipo 2.14 Otros autores relacionan estos hallazgos con posibles mecanismos epigenéticos a través de la expresión diferencial de esta enfermedad según el sexo de los progenitores [2].

En el trabajo “PPP-Botnia and Framingham Offspring Studies” se encontró que el riesgo de presentar DM2 cuando se tiene un familiar de primer grado afectado es 1,62 veces mayor.16 En otros estudios el riesgo de desarrollar esta enfermedad es de 16,7 por cada 1 000 habitantes cuando se tiene una historia familiar positiva comparado con 8,8 por cada 1 000 habitantes con una historia familiar negativa [2].

Estudios en gemelos plantean una heredabilidad entre 25-40%, lo que es sugestivo de que ambos factores, genéticos y ambientales, contribuyen sustancialmente al riesgo individual de presentar DM2.

Los loci de riesgo hasta ahora detectados explican solo una pequeña parte (10%) del riesgo genético de la enfermedad, lo que pudiera estar en correspondencia con errores al estimar la varianza genotípica, debido al desconocimiento de variantes genéticas raras que expliquen mejor su génesis y, por otro lado, también puede influir la sobreestimación de la varianza fenotípica al ignorar la contribución potencial de los efectos intrauterinos y de la interacción gen-gen y gen-ambiente [2].

Los primeros estudios encaminados a identificar genes de susceptibilidad en la DM2 fueron estudios de ligamiento, realizados en familias, y estudios de genes candidatos.

Una de las estrategias empleadas en la búsqueda de genes candidatos son los estudios epidemiológicos. En este sentido existen diferentes ejes como factores de predisposición: la resistencia a la insulina y la disfunción de células beta, así como defectos del sistema incretínico. Estos estudios permitieron una mejor comprensión de la fisiopatología de la enfermedad, pero no permitieron identificar variantes genéticas asociadas con un elevado riesgo de padecerla; ha sido la introducción de los GWAS los que han logrado un considerable avance en el conocimiento de las bases etiopatogénicas y genéticas de dicha afección [2].

Hasta el año 2007 solo se habían asociado tres genes, de modo consistente, con la diabetes mellitus tipo 2: PPARG, KNCJ11 y TCF7L2.

La primera variante genética implicada, el Pro12Ala del gen del PPARG, codifica un receptor nuclear PPARγ, expresado de modo preferente en el tejido adiposo, en el que regula la transcripción de genes implicados en la adipogénesis; los individuos homocigotos para el alelo de la prolina son más insulinorresistentes y tienen un 20% más de riesgo de desarrollar DM tipo 2.

El KCNJ11, que codifica los canales de potasio de las células beta, funcionalmente relacionado con el receptor SUR1 de las sulfonilureas (codificado por el ABCC8), se asoció con la diabetes mellitus tipo 2 en un metaanálisis inicial y se confirmó en estudios posteriores.

El gen transcription factor 7-like 2 (TCF7L2), que codifica proteínas implicadas en la secreción de insulina es, hasta la fecha, el gen más fuertemente asociado con la DM2; cuatro polimorfismos en este gen se han asociado con la enfermedad en diferentes estudios multiétnicos publicados.

La utilización de los GWAS ha permitido identificar otros genes de riesgo implicados mayormente en la función de la célula beta y, en menor medida, en la acción de la insulina y en el desarrollo de obesidad [2].

En el estudio realizado por Brunetti y colaboradores en el año 2014 se concluyó que el gen HMGA1 muestra gran asociación respecto a la DM2 y su variante más frecuente, rs146052672, confiere el riesgo más elevado para el ser humano. Las variantes genéticas identificadas en este gen pueden representar un marcador predictivo para la temprana detección de esta enfermedad, especialmente en individuos que presentan una historia familiar. Las variantes de este gen pudieran inducir un curso clínico diferente en la diabetes comparado con pacientes diabéticos sin la variante además de predecir la respuesta para la terapia y permitir identificar, a priori, a los pacientes que podrían sacar mejor provecho de un tratamiento farmacológico específico [2].

Los loci de riesgo detectados hasta el momento explican solo una pequeña parte del riesgo genético y representan únicamente la punta del iceberg. Se necesita una mejor comprensión de las bases genéticas de la enfermedad, que permita diseñar mejores estrategias para su prevención y su tratamiento [2], Y a pesar de que los genes que predisponen a la diabetes se consideran factores esenciales en el desarrollo de la enfermedad, la activación de una predisposición genética requiere de la presencia de factores ambientales y de comportamiento, especialmente los relacionados con el estilo de vida. Los factores más importantes son la dieta, la inactividad física, el exceso de peso y la obesidad abdominal [2].

Estas son condiciones comunes en los altos estándares de vida de la población occidental, de modo que el avance de este estilo de vida en países en desarrollo también explica la explosión epidémica de la enfermedad, pues los datos epidemiológicos muestran que la distribución temporal y espacial de la DM2 en las áreas geográficas examinadas es comparable con la tendencia al sobrepeso y a la obesidad [2].

Está bien documentado el efecto de la actividad física en la prevención de esta enfermedad. Se ha expuesto que la probabilidad de desarrollarla es usualmente menor en personas que practican

ejercicios físicos versus aquellos que llevan una vida sedentaria. Un ejemplo es que mujeres que practican actividad física vigorosa tienen 16% menos riesgo de presentar DM tipo 2 [2].

La actividad física sistemática aporta beneficios en la salud, fundamentalmente en pacientes diabéticos, debido a que aumenta el contenido mitocondrial del músculo esquelético, ofrece un mecanismo adicional que mejora la sensibilidad a la insulina producida por el ejercicio, además de un control más adecuado de las cifras de glucemia, y favorece la reducción de las complicaciones de índole cardiovascular; por lo que la American Heart Association y la American Diabetes Association recomiendan realizar al menos 150 minutos de ejercicios aeróbicos moderados o 90 minutos de ejercicios aeróbicos intensos a la semana [2].

En otra de sus aristas se ha demostrado que el ejercicio no solo disminuye los niveles de marcadores inflamatorios como CRP, IL-6 y TFN-α, sino que simultáneamente eleva las concentraciones de citoquinas anti-inflamatorias (IL-4, IL-10, TGF-β), que suprimen la producción de citoquinas proinflamatorias relacionadas con desordenes metabólicos (IL-1, IL-2 y TNF-α) [2].

En el caso del exceso de peso se ha demostrado que causa resistencia a la insulina y representa el primer paso en la historia natural de la enfermedad. Inicialmente, en individuos destinados a ser diabéticos, las células beta pancreáticas compensan la resistencia a la insulina al secretar niveles aumentados de esta, así asegura la euglucemia postprandial. La hiperglucemia en los pacientes resistentes a la insulina se desarrolla más tarde, cuando las células beta pancreáticas fallan en la compensación. Como han demostrado numerosos estudios ambos defectos son el resultado de una interacción complicada entre factores genéticos y medioambientales, incluidos los agentes químicos (iones de calcio y de cinc) y las sustancias orgánicas contaminantes, que se sospecha juegan un papel importante en la formación de fibras amiloideas en las células betas pancreáticas [2].

En varios países del mundo la obesidad es prevalente en pacientes con diabetes mellitus tipo 2: en Inglaterra el 86% de los pacientes son sobrepeso u obesos, mientras que en Australia el 53% de los diabéticos son obesos y el 32% son sobrepeso.

Al analizar el papel de la obesidad en relación con la DM2 se pensaba, anteriormente, que el tejido adiposo funcionaba únicamente como depósito de grasa; sin embargo, en la actualidad se sabe que

los adipocitos desempeñan una función autocrina, paracrina y endocrina, es decir, liberan sustancias llamadas adipocinas, que tienen un efecto sobre los mismos adipocitos, las células adyacentes y las células distantes en otros órganos y tejidos. Los adipocitos liberan adipocinas pro-inflamatorias como la resistina, la leptina, el factor de necrosis tumoral (FNT-α) y la interleucina-6, entre otros. Cuando existe un exceso de grasa, y principalmente un aumento en la región visceral, se induce la liberación de sustancias pro-inflamatorias, las que están relacionadas con ciertos desórdenes metabólicos. Es por eso que la obesidad representa un factor de riesgo independiente para enfermedades metabólicas y cardiovasculares a nivel poblacional [2].

Otro factor a tener en cuenta, que influye en la génesis de la DM2, es el sexo. En varias publicaciones se ha observado una mayor prevalencia de mujeres entre los pacientes que sufren esta enfermedad. Un ejemplo es una publicación hecha en el anuario estadístico cubano de 2014, que reportó 66,3 mujeres afectadas por cada 1 000 habitantes y solo 45,1 hombres. Resultados similares fueron hallados en el Municipio Jaruco, de la Provincia de Mayabeque, en el área de salud Puentes Grandes, de Ciudad de La Habana y en la Provincia de Camagüey [2].

Estos hallazgos podrían explicarse por las diferencias de los rasgos anatómicos, fisiológicos y conductuales entre mujeres y hombres. No solo los cromosomas sexuales contribuyen a estas diferencias, pues estudios sugieren que la variación natural dentro de los cromosomas autosómicos también afecta, de manera diferente, estos rasgos. En este contexto el sexo puede considerarse una variable "ambiental" que incluye diferencias celulares, metabólicas, fisiológicas, anatómicas y conductuales entre hombres y mujeres. Por consiguiente, este pudiera interactuar con el genotipo de manera similar a otros factores ambientales. Evidencias similares se observaron en enfermedades comunes como el asma bronquial, la hipertensión arterial y la esquizofrenia [2].

El envejecimiento es otro de los factores de riesgo a tener presente. Es referido por autores que, unido a la presencia de hábitos inadecuados del estilo de vida, pudiera asociarse a que esta enfermedad aparezca con más frecuencia en las poblaciones. Según las proyecciones de la Organización de las Naciones Unidas a mediados del presente siglo el número de personas ancianas en el mundo superará al número de jóvenes y se espera un incremento alarmante en el número de diabéticos [2].

Gil Velázquez plantea que el aumento de la prevalencia de la diabetes mellitus tipo 2 está influenciado, principalmente, por el envejecimiento de la población, pues se presenta con más

frecuencia en adultos mayores; sin embargo, debido a diferencias demográficas, en países ricos predomina en los mayores de 60 años y en los países en desarrollo la edad está entre los 40 y 60 años [2].

La prevalencia de la DM entre la población adulta de los Estados Unidos (≥65 años) está en un rango entre el 22-33%, en dependencia de los criterios diagnósticos usados. Esta alta prevalencia también ha sido confirmada en estudios prospectivos de base poblacional realizados en Holanda, que muestran que los pacientes mayores de 70 años constituyen alrededor del 50% de la población diabética.

En Cuba son varias las investigaciones realizadas sobre la interacción genoma-ambiente en la aparición de la DM2. Estos estudios han demostrado que existe una relación estadísticamente significativa entre la presencia de la enfermedad con factores genéticos como la existencia de familiares de primer grado de parentesco afectados o cuando están presentes factores de riesgos ambientales como es el caso de la obesidad; pero quizás el mayor aporte de estos estudios ha sido demostrar que existe un mayor riesgo de padecer DM2 cuando ambos factores (genéticos y ambientales) coinciden en un mismo individuo, elevándose el riesgo de padecer la enfermedad de manera exponencial en la mayoría de los casos [2].

La DM2, con mucho la forma más frecuente de la enfermedad (90%), es consecuencia de una compleja interacción entre múltiples genes y diversos factores ambientales aún no completamente entendidos, y se caracteriza por defectos en la secreción y en la acción de la insulina que conducen a la hiperglucemia [27].

Considerada durante muchos años la pesadilla de los genetistas, recientemente se han producido considerables avances en el conocimiento de la genética de la enfermedad, con la disponibilidad de datos procedentes de los GWAS, reforzados por el desarrollo de plataformas de genotipificación de alta resolución, la profusión de SNP en bases de datos públicas, los análisis de numerosas cohortes de pacientes y la generación de herramientas de análisis muy sofisticadas. Se han podido identificar hasta 28 genes asociados con DM2 que, sin embargo, sólo explican un 10% de la susceptibilidad genética a presentar la enfermedad [27].

La abundante evidencia que apoya la base genética de la DM2 procede de estudios de población, de familiares y de hermanos gemelos. Su prevalencia varía considerablemente entre grupos étnicos

que comparten el mismo ambiente (en los Estados Unidos es de dos a seis veces más prevalente en afroamericanos, indios Pima e hispanos que en la población de raza blanca). Los hijos de un progenitor diabético tienen un 40% de riesgo de desarrollar DM2, frente al riesgo existente en la población, de un 7% y, si ambos padres son diabéticos, el riesgo aumenta a un 70%. El riesgo relativo para un hermano está en torno a tres. En gemelos homocigóticos si uno de los hermanos presenta DM2, en un 90% de los casos el otro hermano presentará diabetes [27].

Los primeros estudios encaminados a identificar genes de susceptibilidad a la DM2 fueron estudios de ligamiento, realizados en familias, y estudios de genes candidatos. Aunque estos últimos permitieron una mejor comprensión de la fisiopatología de la DM2, no permitieron identificar variantes genéticas asociadas con un elevado riesgo de padecer la enfermedad; ha sido la introducción de los GWAS lo que ha permitido un considerable avance en el conocimiento de las bases etiopatogénicas y genéticas de la enfermedad. Hasta el año 2007 sólo se habían asociado 3 genes de modo consistente con la DM2: PPARG, KNCJ11 y TCF7L2.

La primera variante genética implicada en la DM2 fue el Pro12Ala del gen del PPARG, que codifica un receptor nuclear PPARγ y que se expresa de modo preferente en el tejido adiposo, donde regula la transcripción de genes implicados en la adipogénesis; los individuos homocigotos para el alelo de la prolina son más insulinorresistentes y tienen un 20% más de riesgo de desarrollar DM233 [27].

El KCNJ11 que codifica los canales de potasio de las células beta, funcionalmente relacionado con el receptor SUR1 de las sulfonilureas (codificado por el ABCC8), se asoció con DM2 en un metanálisis inicial y se confirmó en estudios posteriores.

El gen transcription factor 7-like 2 (TCF7L2), que codifica proteínas implicadas en la secreción de insulina, es, hasta la fecha, el gen más fuertemente asociado con la DM2; cuatro polimorfismos en dicho gen se han asociado con la enfermedad en diferentes estudios multiétnicos [27].

La utilización de los GWAS ha permitido identificar otros genes de riesgo, implicados mayormente en la función de la célula beta (HNF1B [17cen-q21.3], WFS1 [4p16], GCK [7p15.3-p15.1], CDKN2A/2B [9p21], CDKAL1 [6p22.3], SLC30A8 [8q24.11], IGF2BP2 [3q27.2], THADA [2p21], NOTCH2 [1p13-p11], CDC123 [10p13], CAMK1D, HHEX [10q23], IDE, TSPAN8 [12q14.1-q21.1], JAZF1 [7p15.2-p15.1], KCNQ1 [11p15.5], MTNR1B [11q21-q22],

ADCY5 [3q13.2-q21], PROX1 [1q32.2-q32.3], DGKB [7p21.2]) y, en menor medida, en la acción de la insulina (ADAMTS9 [3p14.3-p14.2], IRS1 [2q36], GCKR [2p23]) y en el desarrollo de obesidad (FTO [16q12.2]). Debido a que los alelos de riesgo de DM2 son frecuentes y confieren pequeños incrementos de riesgo, muchos individuos portadores de dichos alelos no desarrollan DM2. En estudios longitudinales de población, el valor predictivo de un modelo con información genética de al menos 18 alelos de riesgo no es superior al de un modelo basado sólo en factores de riesgo convencionales, como presión arterial, triglicéridos, colesterol-HDL, índice e masa corporal (IMC) e historia familiar de diabetes36. No obstante, un estudio reciente que contempló variantes confirmadas y aún no confirmadas de susceptibilidad a la enfermedad en todo el genoma demostró que la estimación de riesgo tenía un alto poder predictivo37. Por lo tanto, cabe esperar que en un futuro el uso clínico de todas las variantes de riesgo de DM2 ya validadas, junto con las mutaciones poco frecuentes, aún por descubrir, podría identificar individuos con alto riesgo de desarrollar DM2, especialmente si tienen familiares afectados [27].

A modo de corolario final conviene subrayar que los loci de riesgo hasta ahora detectados explican sólo una pequeña parte (10%) del riesgo genético de DM2 y representan únicamente la punta del iceberg. Necesitamos una mucho mejor comprensión de la arquitectura genética de la enfermedad que nos permita diseñar mejores estrategias para su prevención y tratamiento. Lo que se ha denominado «herencia perdida» u «oscura materia» podría explicarse por la presencia de variantes poco comunes (entre 1-5%) o muy poco frecuentes (<1%), que no son bien detectados por los GWAS. La hipótesis es que múltiples y poco frecuentes variantes funcionales se acumulan en la misma dirección sobre un haplotipo y que difieren entre individuos38. El desafío actual para la comunidad genética consiste en validar esta hipótesis e identificar las variantes poco comunes responsables de los fuertes efectos genéticos, cuantificar su número e identificar su localización entre los loci específicos de DM2 [27].

Los genes asociados con riesgo a la diabetes tipo 2, incluyen:

TCF7L2, que afecta la secreción de insulina y la producción de glucosa.

ABCC8, que ayuda a regular la insulina.

CAPN10, que está asociado con el riesgo de diabetes tipo 2 en mexicoamericanos.

GLUT2, que ayuda a trasladar la glucosa al páncreas.

GCGR, una hormona de glucagón involucrada en la regulación de la glucosa.

Algunos de los genes que se han relacionado con la predisposición a la diabetes mellitus tipo 2 (DM2) [28].

Desarrollo pancreático

- Factor de transcripción TCF7L2. Implicado en el desarrollo de la célula beta. Riesgo atribuible de DM2 hasta el 70%.
- Gen KCNQ1. Codifica la subunidad alfa de una proteína que forma parte de los canales de potasio de la célula beta. Susceptibilidad a diabetes gestacional y neuropatía diabética.
- Gen WFS1. Implicado en la supervivencia de la célula beta. Asociación con el síndrome de Wolfram y el déficit de glucagon-like peptide-1 (GLP-1).

Secreción de insulina

- Factor de transcripción TCF7L2. Los polimorfismos rs12255372 y rs7903146 se asocian a alteración de la respuesta de la célula beta al efecto incretínico del GLP-1.
- Gen del receptor alfaadrenérgico ADRA2A. Menor secreción de insulina en respuesta a la glucosa incluso reversible con antagonistas alfaadrenérgicos.

Acción de la insulina

- Gen del receptor de la insulina. Mutaciones en asociación con acantosis nigricans e hiperandrogenismos.
- Genes de sustrato del receptor de insulina IRS-2. Estado de insulinorresistencia hepático.
- Gen del receptor beta-3-adrenérgico Baja tasa de metabolismo con riesgo de obesidad y DM2
- Gen del receptor PPARgamma. Condiciona una diferente sensibilidad a la insulina y el índice de masa corporal.

Tabla 1. Genes de susceptibilidad para DT2 [29].

Gen	Cromosoma	Razón de momios	FAR	Tipo de Estudio	Función y probable mecanismo
ADAMTS9	3	1.09-1.05	0.68-0.81	MA	Acción de la insulina y metaloproteinasa
ADCY5	3	1.12	0.78	MA	Acción de la insulina/ adenilciclasa

ANK1	8	1.09	0.76	MA, CC	Estabilidad celular/ función célula β
ANKRD55	5	1.08	0.7	MA, CC	Acción de la insulina
ANKS1A	6	1.11	0.91	GWAS	Regulador de vía/ desconocido
ARAP1	11	1.08-1.14	0.81-0.88	GWAS, MA	Modulador del citoesqueleto de actina/ función de célula β
BCAR1	16	1.12	0.89	MA, CC	Proteína de acoplamiento/ función de célula β
BCL2	18	1.09	0.64	GWAS	Regulador de muerte celular/ función de célula β
BCL11A	2	1.08-1.09	0.46	MA	Dedo de Zinc / función de célula β
CAMK1D	10	1.07-1.11	0.18	LA. MA	Proteincinasa/ función de célula β
CDC123					Proteína mitótica/ función de célula β
CAPN10	2	1.09-1.18	0.73-0.96	MA	Proteasa Calpaina cisteína/ acción de la insulina
CDKAL1	6	1.10-1.20	0.27-0.31	GWAS, MA	función de célula β
CDKN2A	9	1.19-1.20	0.82-0.83	GWAS	Inhibidor de cinasa ciclina-dependiente/ función de célula β
CDKN2B					
CENTD2	11	1.08-1.13	0.81-0.88	GWAS	función de célula β
CHCHD9	9	1.11-1.20	0.93	MA	Desconocido
TLE4					
CILP2	19	1.13	0.08	MA, CC	Desconocido
DGKB	7	1.04-1.06	0.47-0.54	MA	Cinasa diacilglicerol/ acción de la insulina
DUSP9	X	1.09-1.27	0.12-0.77	MA	Fosfatasa
FOLH1	11	1.10	0.09	GWAS	Glucoproteína transmembrana/ desconocido
FTO	16	1.06-1.27	0.38-0.41	GWAS, MA	Regulador metabólico/ acción de la insulina
GATAD2A	19	1.12	0.08	GWAS	Represor transcripcional/ desconocido
GCK	7	1.07	0.20	MA	Glucocinasa/ acción de la insulina
GCKR	2	1.06-1.09	0.59-0.62	MA	Regulador de glucocinasa/ acción de la insulina
GIPR	19	1.10	0.27	GWAS	Receptor acoplado a proteína-G/ desconocido
GRB14	2	1.07	0.60	MA, GCS	Adaptador de proteína/ acción de la insulina
HFE	6	1.12	0.29	MA	Proteína de membrana/ desconocido
HHEX	10	1.12-1.13	0.53-0.60	AL, MA	Represor transcripcional/ degradación intracelular de insulina/ proteína motor.
IDE					
KIF11					
HMG20A	15	1.08	0.68	MA, GCS	Proteína asociada a cromatina/ desconocido
HMGA1	6	1.34-15.8	0.10	GCS	Regulador transcripcional/ acción de la insulina
HMGA2	12	1.10-1.20	0.09-0.10	MA	Regulador transcripcional
HNF1A	12	1.07-1.14	0.77-0.85	MA	Activador transcripcional hepático y pancreático

HNF1B	17	1.08-1.17	0.47-0.51	GCS, MA	Factor de transcripción/ función de célula β
IGF2BP2	3	1.14	0.29-0.32	GWAS, MA	Proteína de unión/ función de célula β
IRS1	2	1.09-1.12	0.64-0.67	GCS, MA	Elemento de señalización de insulina/ acción de la insulina
JAZF1	7	1.10	0.52	MA	Dedo de zinc/ función de célula β
KCNJ11	11	1.09-1.14	0.37-0.47	GCS, MA	Canales de potasio/ función de célula β
KCNQ1	11	1.08-1.23	0.44	GWAS	Canales de potasio/ función de célula β
KLF14	7	1.07-1.10	0.55	MA	Factor de transcripción/ acción de la insulina
KLHDC5	12	1.10	0.80	MA, CC	Progresión mitótica y citocinesis/ desconocido
LAMA1	18	1.13	0.38	GWAS	Mediador de migración celular/ acción de la insulina
MC4R	18	1.08	0.27	MA, CC	Receptor acoplado a proteína-G/ desconocido
MTNR1B	11	1.05-1.08	0.28-0.30	GWAS, MA	Receptor de melatonina/ función de célula β
NOTCH2	1	1.06-1.13	0.10-0.11	MA	Receptor de membrana
PPARG	3	1.11-1.17	0.85-0.88	GCS, MA	Receptor nuclear/ acción de la insulina
PRC1	15	1.07-1.10	0.22	MA	Regulador de citocinesis
PROX1	1	1.07	0.50	MA	Factor de transcripción Homeobox/ acción de la insulina
PTPRD	9	1.57	0.10	GWAS	Proteína tirosina fosfatasa
RBMS1	2	1.11-1.08	0.79-0.83	MA	Modulador DNA/ acción de insulina
SLC2A2	3	1.06	0.74	GWAS	Sensor de glucosa/ función de célula β
SLC30A8	8	1.11-1.18	0.65-0.70	GWAS, MA	función de célula β
SREBF1	17	1.07	0.38	GWAS	Regulador transcripcional de lípidos/ desconocido
SRR	17	1.28	0.69	GWAS	Serina racemase
TCF7L2	10	1.31-1.71	0.26-0.30	LA, MA, GWAS	Participante en las vías de señalización/ función de célula β
THADA	2	1.15	0.90	MA	Proteína asociada a adenomga tiroidea/ función de célula β
TH/INS	11	1.14	0.39	GWAS	Síntesis de catecolamina/ desconocido
TLE1	9	1.07	0.57	MA, CC	Corepresor transcripcional/ desconocido
TP53INP1	8	1.06-1.11	0.48	MA	Proteína proapoptótica/ desconocido
TSPAN8	12	1.06-1.09	0.27-0.71	MA	Glicoproteína de superficie celular/ función de célula β
LGR5					Receptor acoplado a proteína-G/ función de célula β
WFS1	4	1.10-1.13	0.60-0.73	GCS	Proteína transmembrana/ función de célula β
ZBED3	5	1.08-1.16	0.26	MA	Dedo de zinc/ función de célula β
ZFAND6	15	1.01-1.11	0.60-0.72	MA	Dedo de zinc/ función de célula β
ZMIZ1	10	1.08	0.52	MA, CC	Regulador transcripcional/ desconocido
Haplogroup B	mtDNA	1.52	0.25	GCS	
OriB	mtDNA	1.10	0.30	MA	

GWAS: Escaneo genómico completo, MA: Meta-análisis, GCS: Estudio de Genes candidatos, LA: Análisis de ligamiento, FAR: Frecuencias de los alelos de riesgo

1.3 Datos Genéticos de la Diabetes Tipo II en la Sociedad Mexicana

Los estilos de vida poco saludables son altamente prevalentes entre niños, adolescentes y adultos mexicanos, propiciando un aumento importante de la obesidad y sobrepeso, principal factor de riesgo modificable de la diabetes. Así, la prevalencia de la diabetes en esta población ha incrementado sustancialmente en las últimas décadas: en 1993 la prevalencia de los diabéticos con diagnóstico conocido en población mayor de 20 años fue de 4.0%, mientras que en 2000 y 2007 se describió una prevalencia del 5.8 y 7%, respectivamente. Por otro lado, de acuerdo con las encuestas nacionales de esos mismos años, se ha demostrado la alta prevalencia de condiciones comórbidas en la población diabética y problemas en la calidad de la atención, lo cual contribuye de manera importante a la mayor incidencia de complicaciones macro y microvasculares. Las estrategias de prevención implementadas a escala poblacional en países con elevado riesgo que logren modificar estilos de vida en particular en la dieta, actividad física y tabaquismo pueden ser altamente costo efectivas al reducir la aparición de la diabetes y retrasar la progresión de la misma. México tiene condiciones de alto riesgo, por lo que recientemente se han impulsado políticas intersectoriales relacionadas con la salud alimentaria y con ello combatir uno de los más importantes factores de riesgo, la obesidad. Al mismo tiempo se han diseñado, ya desde hace más de una década, estrategias como PREVENIMSS, PREVENISSSTE, grupos de autoayuda, Unidades de Especialidades Médicas para Enfermedades Crónicas, entre otras al interior de las principales instituciones de salud con el propósito de mejorar la atención que se otorga a los pacientes que ya padecen la enfermedad. Sin embargo, el estado actual de los diabéticos mexicanos se conoce sólo parcialmente, información que es necesaria para cimentar y fortalecer los esfuerzos que se requieren en prevención a todos los niveles a fin de contender una de las más grandes y emergentes amenazas de la viabilidad de los sistemas de salud, la diabetes [30].

Revisiones recientes sobre los factores genéticos de la DM2 en México están ayudando a mejorar el entendimiento de la enfermedad en las distintas subpoblaciones. En una revisión hecha por en 2017 por García-chapa et. Al [31], se analizaron 19 estudios de casos y controles sobre la posible asociación de polimorfismos genéticos con DM2 en mestizos mexicanos residentes en el país. En

total, se evaluaron 68 polimorfismos de 41 genes (tabla 1). De ellos, 25 se asociaron con un mayor riesgo de DM2 y se localizaron en 20 genes, a saber, *ABCA1, ADRB3, CAPN10, CDC123 / CAMK1D, CDKN2A / 2B, CRP, ELMO1, FTO, HHEX, IGF2BP2, IRS1, JAZF1, KCNQ1, LOC387761, LTA, NXPH1, SIRT1, SLC30A8, TCF7L2* y *TNF-α*. Entre las variantes que mostraron asociación hubo 4/20 sustituciones de aminoácidos, 13/30 sitios intrónicos, 6/10 de la región promotora o región flanqueante 5' o cadena arriba del gen, 1/2 regiones intergénicas y 2/6 de 3 Región 'no traducida o flanqueante 3' del gen. Por otro lado, 12 polimorfismos fueron analizados por diferentes autores, y se observó concordancia en la mayoría de ellos, a excepción de rs3842570 *(CAPN10)*, rs13266634*(SLC30A4)*, rs7903146 *(TCF7L2)* y rs1800629 *(TNF-α)*. Once de estos polimorfismos se combinaron y analizaron como se muestra en Tabla 2. Tenga en cuenta que rs4994 *(ADRB3)* se descartó de este análisis (sospecha de superposición de). Del mismo modo, los datos para rs7903146 y rs12255372 de *TCF7L2* no se consideraron. Por lo tanto, el alelo 3R de rs3842570, que se asoció con T2D en una muestra pequeña, aparentemente no confería susceptibilidad a la enfermedad; por el contrario, el alelo C de rs7754840 *(CDKAL1)*, que no evidenció riesgo en estudios independientes, mostró asociación con T2D. Incluyendo este alelo, un total de 26 polimorfismos y 21 genes se asociaron con DT2 en mestizos mexicanos [31].

Tabla 2. Genes analizados en estudios sobre diabetes tipo 2 realizados en mestizos mexicanos [31].

Gene	**Crom**	**dbSNP loc**	**Cambio**	**Efecto**	***n*[a]; *n*[b]**	**OR (IC del 95%)**	***p***
ABCA1	9q31	rs9282541	C / **T**	R / C	244; 202	**2,50 (1,48–4,24)**	**0,001**
		rs2000069	C / **T**	Intrónico	244; 202	1,08 (0,82–1,42) [c]	0,58
		rs2230806	G / **A**	R / K	244; 202	1,17 (0,89–1,55) [c]	0,27
		rs2487037	C / **T**	Intrónico	244; 202	1,06 (0,79–1,43) [c]	0,71
		rs3818689	G / **C**	Intrónico	244; 202	0,94 (0,52–1,68) [c]	0,82
ADAMTS9	3p14	rs4607103	C / **T**	Intrónico	1027; 990	1,05 (0,91–1,20) [d]	0.521
ADRB1	10q25	rs1801253	C / **G**	R / G	501; 552	0,79 (0,61–1,02) [c]	0,07
ADRB3	8p11	rs4994	C / **T**	W / R	519; 547	**1,69 (1,37–2,09)** [c]	**0,0001**

		rs4994	C / **T**	W / R	501; 552	**1,34 (1,10–1,64)** [c]	**0,004**
ARHGEF11	1q21	rs945508	G / **A**	RH	868; 504	0,91 (0,76–1,09) [e]	0,319
CAPN10	2q37	rs3792267	G / **A**	Intrónico	132; 112	0,97 (0,66–1,42) [c]	0,86
		rs3792267	G / **A**	Intrónico	719; 746	1,11 (0,95–1,29) [c, f]	0,20
		rs3792267	G / **A**	Intrónico	211; 152	0,91 (0,66-1,26)	0,56
		rs3842570	2R / **3R**	Intrónico	132; 112	0,97 (0,68–1,40) [c]	0,89
		rs3842570	2R / **3R**	Intrónico	43; 64	**1,81 (1,03–3,18) [c]**	**0,038**
		rs3842570	2R / **3R**	Intrónico	211; 152	0,75 (0,55–1,02)	0,06
		rs5030952	C / **T**	Intrónico	132; 113	0,85 (0,56–1,29) [c]	0,45
		rs5030952	C / **T**	Intrónico	211; 152	1,35 (0,89–2,06)	0,16
		rs2975760	T / **C**	Intrónico	134; 113	**2,72 (1,16–6,35)**	**0,017**
CAPN10	2q37	rs7607759	A / **G**	EJÉRCITO DE RESERVA	127; 110	2,27 (0,98–5,25) [c]	0.051
CDC123 / CAMK1D	10p13	rs12779790	A / **G**	Intergénico	1027; 990	**1,24 (1,05–1,47)** [d]	**0.013**
CDKAL1	6p22	rs10946398	A / **C**	Intrónico	519; 547	1,09 (0,91–1,32) [c]	0.337
		rs9465871	**C** / T	Intrónico	519; 547	1,04 (0,85–1,26) [c]	0,718
		rs7754840	**C** / G	Intrónico	519; 547	1,08 (0,89–1,29) [c]	0.438
		rs7754840	**C** / G	Intrónico	1027; 990	1,13 (0,98-1,30) [d, g]	0.081
CDKN2A / 2B	9p21	rs10811661	C / **T**	Río arriba	1027; 990	**1,42 (1,15-1,75)** [d]	**0,001**
CRP	1q23	rs1130864	C / **T**	3′-UTR	166; 130	**1,59 (1,15-2,22)** [c, h, i]	**0,005**
		rs1205	G / **A**	3′-UTR	166; 130	0,82 (0,59–1,14) [c, h, i]	0,24
		rs2794521	A / **G**	5′-flanqueando	166; 130	**1,97 (1,15-3,38)** [c, h, i]	**0,012**

		rs3093062	G / **A**	Promotor	166; 130	**3,49 (0,98-12,4)** [c, h, i]	**0.039**
ELMO1	7p14	rs1345365	A / **G**	Intrónico	148; 269	**1,37 (1,02–1,84)** [c, h, i]	**0,035**
ENPP1	6q23	rs1044498	A / **C**	K / Q	519; 547	0,94 (0,76–1,16) [c]	0.577
EXT2	11p11	rs3740878	**A** / G	Intrónico	455; 234	0,83 (0,65–1,05)	0.054
FTO	16q12	rs8050136	**A** / C	Intrónico	868; 504	0,90 (0,74–1,09) [e]	0,278
		rs9939609	**A** / T	Intrónico	519; 547	**1,25 (1,02-1,54)** [c]	**0,027**
HHEX	10q23	rs5015480	**C** / T	Río arriba	519; 547	0,96 (0,80–1,14) [c]	0,631
		rs1111875	**C** / T	3′-flanqueando	1027; 990	1,01 (0,89–1,16) [d]	0,859
		rs1111875	**C** / T	3′-flanqueando	455; 234	1,12 (0,88–1,44)	0,27
HHEX	10q23	rs7923837	A / **G**	3′-flanqueando	868; 504	**1,21 (1,02–1,44)** [k]	**0,025**
HMOX1	22q12	rs2071749	A / G	Promotor	614; 956	0,98 (0,84–1,14) [c]	0,76
IGF2BP2	3q27	rs4402960	G / **T**	Intrónico	868; 504	**1,24 (1,01–1,53)** [j]	**0,042**
IRS1	2q36	rs1801278	G / **A**	GRAMO	719; 746	**2,04 (1,41–2,96)** [c, f]	**<0,001**
		rs1801278	G / **A**	GRAMO	444; 444	**3,22 (1,99–5,20)**	**0,001**
		rs1801276	C / **G**	PENSILVANIA	444; 444	0,98 (0,72–1,32)	0,83
		rs3731594	G / **A**	DAKOTA DEL NORTE	444; 444	0,83 (0,42–1,66)	0,47
		rs1801108	**G** / C	R / P	444; 444	1,07 (0,85–1,34)	0,40
JAZF1	7p15	rs864745	**T** / C	Intrónico	868; 504	**1,24 (1,04–1,47)** [k]	**0,015**
KCNJ11	11p15	rs5215	C / **T**	V / I	519; 547	1,03 (0,87–1,23) [c]	0,729
		rs5210	**A** / G	3′-UTR	519; 547	1,03 (0,86–1,23) [c]	0,764
		rs5219	C / **T**	E / K	1027; 990	1,10 (0,96–1,26) [d]	0,154

KCNQ1	11p15	rs2237892	**C** / T	Intrónico	868; 504	**1,36 (1,13-1,64)** [k]	**0,001**
LEPR	1p31	rs1137100	**A** / G	K / R	519; 547	1,00 (0,84–1,21) [c]	0,92
LOC387761	11p12	rs7480010	A / **G**	Intrónico	455; 234	**1,43 (1,05–1,94)**	**0,006**
LTA	6p21	rs909253	**A** / G	Intrónico	51; 48	**1,98 (1,02–3,8)** [c]	**0.041**
MGEA5	10q24	MGEA5-14	A / **T**	Intrónico	271; 244	1,60 (0,52–4,86)	0.404
NOTCH2	1p11	rs10923931	G / **T**	Intrónico	1027; 990	1,04 (0,82–1,32) [d]	0,731
NQO1	16q22	rs1800566	C / **T**	PD	623; 993	0,98 (0,85–1,13) [c]	0,76
NRF2	2q31	rs2364723	C / **G**	Intrónico	625; 992	0,91 (0,79–1,05) [c]	0,18
NRF2	2q31	rs6721961	C / **A**	Promotor	623; 989	0,89 (0,74–1,06) [c]	0,18
NXPH1	7p22	rs757705	A / **G**	Intrónico	868; 504	**1,25 (1,05–1,48)** [k]	**0,01**
PPARG	3p25	rs1801282	**C** / G	PENSILVANIA	719; 746	1,00 (0,81–1,24) [c, f]	1,00
		rs1801282	**C** / G	PENSILVANIA	1027; 990	1,10 (0,90–1,34) [d]	0.342
		rs17793693	**A** / C	Intrónico	519; 547	1,09 (0,91–1,31) [c]	0.329
RALGPS2	1q25	rs2773080	**A** / G	Intrónico	868; 504	0,90 (0,74–1,10) [e]	0.315
RORA	15q22	rs7164773	C / **T**	Intrónico	868; 504	1,08 (0,91–1,28) [e]	0.357
SIRT1	10q21	rs3758391	C / **T**	Río arriba	519; 547	**1,29 (1,08-1,54)** [c]	**0,004**
SLC30A4	8q24	rs13266634	**C** / T	R / W	455; 234	1,01 (0,76–1,33)	0,92
		rs13266634	**C** / T	R / W	1027; 990	**1,22 (1,05–1,41)** [d]	**0,009**
TCF7L2	10q25	rs7903146	C / **T**	Intrónico	868; 504	1,04 (0,84–1,28) [e, l]	0,735
		rs7903146	C / **T**	Intrónico	200; 200	**1,84 (1,05-3,20)** [c, m]	**0,04**
		rs7903146	C / **T**	Intrónico	519; 547	**1,48 (1,18–1,86)** [c]	**0,0007**
		rs7903146	C / **T**	Intrónico	283; 271	1,25 (0,92–1,70)	0,16

		rs12255372	G / **T**	Intrónico	200; 200	**1,83 (1,21-2,76)** [c, m]	**0,006**
		rs12255372	G / **T**	Intrónico	281; 268	**1,78 (1,11–2,88)**	**0,017**
		rs12255372	G / **T**	Intrónico	519; 547	**1,37 (1,06–1,76)** [c]	**0,014**
		DG10S478	STR CACA	Intrónico	282; 274	**1,62 (1,02-2,57)**	**0.041**
TLR2	4q32	rs5743708	G / **A**	R / Q	321; 538	0,41 (0,04–3,7)	0,40
TLR4	9q33	rs4986790	A / **G**	D / G	321; 538	1,39 (0,42–4,56)	0,58
		rs4986791	C / **T**	T / I	321; 538	1,01 (0,32–3,18)	0,98
TNF-α	6p21	rs1800629	-308G / **A**	Río arriba	51; 48	0,76 (0,31–1,85) [c, n]	0,55
		rs1800629	-308G / **A**	Río arriba	95; 87	**4,66 (1,73-12,5)** [c]	**0,001**
		rs1800629	-308G /**A**	Río arriba	259; 645	1,25 (0,83–1,87)	0,29
		rs361525	-238G / **A**	Río arriba	259; 645	**1,57** (**1,07–2,29**)	**0,018**
TSPAN8 / LGR5	12q14 – q21	rs7961581	C / **T**	Intergénico	868; 504	0,93 (0,73–1,17) [e]	0.516
TXNIP	1q21	rs7211	C / **T**	3 ′ UTR	623; 969	0,97 (0,82–1,14)	0,67
UBQLNL	11p15	rs979752	**C** / T	Río arriba	868; 504	1,04 (0,84–1,30) [e]	0,70

Cromo: cromosoma. Los alelos de riesgo están marcados en negrita.n a ; n b . Muestra para casos y controles, respectivamente. c El OR convencional (no ajustado) fue evaluado por nosotros a partir de las frecuencias de alelos o genotipos informadas. d Se registró la n más grande . e Se consideró la prueba sin corrección de ascendencia. f Se registraron conjuntos de datos combinados. g El riesgo solo se observó en pacientes con diabetes tipo 2 no obesos (OR = 1,25; p = 0,009). h En nuestro análisis solo se utilizaron genotipos de pacientes con diabetes tipo 2 y controles sanos. IEvaluación derivada de la suma de pacientes con DM2 (obesos y no obesos).j Los autores informaron un efecto protector para el alelo A (OR = 0,65; p <0,001), pero en nuestra estimación tomamos como referencia el alelo A, ya que es el más común. k Se tomó un análisis significativo con corrección de ascendencia. l Solo se encontró asociación en la DM2 de inicio temprano (OR = 1,39; p = 0,024). m Solo se registró la población de Guerrero debido a la posible superposición de los pacientes de la Ciudad de México con [10]. n Los autores evaluaron el alelo G como riesgo; pero en nuestro análisis tomamos el alelo A, el mismo que en estudios anteriores.

Tabla 3. Análisis de los SNP estudiados por dos o más grupos en mestizos mexicanos [31].

Gene	**dbSNP loc**	**Casos (alelo)**	**Controles (alelo)**	**Frecuencia de alelos de riesgo (%)**		**OR (IC del 95%)**	***p* * valor**
				Casos	**Control S**		

CAPN10	rs3792267	1086	928	29,5	28,7	1,06 (0,87–1,29)	0,56
CAPN10	rs3842570	772	656	61,7	62,5	0,96 (0,78-1,20)	0,74
CAPN10	rs5030952	686	530	19,0	18,3	1,04 (0,78–1,40)	0,78
CDKAL1	rs7754840	3092	3074	31,4	29,1	**1,12 (1,00-1,25)**	**0.044**
HHEX	rs1111875	2974	2448	62,8	61,7	1,05 (0,94–1,17)	0,43
IRS1	rs1801278	2326	2380	6.6	2.8	**2,45 (1,83–3,28)**	**<0,0001**
PPARG	rs1801282	3492	3472	87,3	86,7	1,05 (0,91–1,21)	0,51
SLC30A4	rs13266634	2964	2448	76,3	73,0	**1,19 (1,05–1,35)**	**0,005**
TCF7L2	rs7903146	3740	3042	17,8	14,1	**1,32 (1,16-1,50)**	**<0,0001**
TCF7L2	rs12255372	2476	2586	18,3	11,6	**1,40 (1,19-1,65)**	**<0,0001**
TNF-α	rs1800629	810	1560	11,5	6.2	**1,96 (1,45-2,64)**	**<0,0001**
∗Prueba de corrección de chi-cuadrado de Yates							

Esta revisión sobre la genética de la DM2 en sujetos mestizos mexicanos muestra que 26 polimorfismos distribuidos en 21 genes se asocian a esta enfermedad, por lo que la DM2 tiene una alta heterogeneidad en nuestra población, al igual que en otros grupos étnicos. Por lo tanto, en algunos individuos están involucrados los alelos de ciertos genes, mientras que en otros sujetos están implicados variantes de diferentes genes. La conclusión anterior de que la DM2 en los mestizos mexicanos es genéticamente homogénea está basada en un análisis de solo tres marcadores genéticos, por lo que no es concluyente. Aunque la población mestiza mexicana tiene una ascendencia genética europea cercana al 30%, no todos los alelos que confieren riesgo de diabetes en los europeos están asociados con DM2 en nuestra población. Estas variaciones podrían estar relacionadas con antecedentes genéticos, diferencias en las clasificaciones clínicas, tamaño de la muestra, criterios de selección y análisis y factores ambientales como la obesidad, el estilo de vida y la dieta. Por otro lado, las investigaciones en varios grupos étnicos han demostrado la asociación de DM2 con genes aún no analizados en la población mexicana. Sería importante realizar el análisis de dichos genes para determinar si estas variantes también se asocian a la DM2

en pacientes mexicanos y aumentar el conocimiento sobre la epidemiología genética de este trastorno en nuestro país [31].

En cuanto a los estudios mexicanos, se detectó un mayor riesgo cuando se realizó el análisis ajustando covariables. Por ejemplo, en [32] observaron un efecto aditivo en el riesgo de DM2 cuando se consideraron variables como la edad, la educación, el sexo, el índice de masa corporal y la ascendencia. Mientras que en [33] informaron asociación con DM2 para los polimorfismos rs7923837 (HHEX), rs4402960 (IGF2BP2) y rs2237892 (KCNQ1) sólo cuando se ajustó la ascendencia. Para los polimorfismos rs864745 (JAZF1) y rs757705 (NXPH1), el análisis estratificado por ascendencia no mostró diferencias significativas, mientras que se observó una asociación en la comparación sin tal ajuste. Además, encontraron asociación para rs7903146 (TCF7L2) y rs7754840 (CDKAL1) solo en la DM2 de inicio temprano [OR = 1,39 (1,04–1,85), p = 0,024] y en pacientes con DM2 no obesos [OR = 1,25 (1,06–1,49), p = 0,009], respectivamente. Otro estudio encontró un OR más bajo cuando el análisis se ajustó por sexo, índice de masa corporal e historia familiar de DM2 para tres polimorfismos de IRS1 en un modelo dominante [31].

La asociación informada de rs3842570 (CAPN10), rs909253 (LTA) y rs1800629 (TNF-α) con T2D debe interpretarse con precaución dado el pequeño tamaño de la muestra y el escaso poder estadístico. Con respecto al polimorfismo rs1345365 (ELMO1), los autores informaron un efecto protector para el alelo A [OR = 0,65 (0,55-0,78), p <0,001]. Pero en nuestro análisis tomamos como referencia el alelo A, ya que es el más común; por tanto, el alelo G mostró asociación con T2D [OR = 1,37 (1,02 a 1,84), p = 0,035] [31].

Dado que la DM2 es un trastorno complejo y varios genes están implicados en su etiología y evolución, la identificación de alelos de riesgo podría ser útil, ya que, si se conocen los genes implicados y su función, es más probable lograr prevención, tratamiento, pronóstico y/o control de la enfermedad. Las complicaciones también podrían prevenirse o tratarse mejor. Sin embargo, los estudios publicados demuestran que el cribado genético para la predicción de la diabetes tipo 2 en sujetos de alto riesgo tiene actualmente poco valor en la práctica clínica. En realidad, los riesgos genéticos son difíciles de calcular porque varios alelos podrían contribuir a un efecto aditivo en la susceptibilidad a la DM2, sin mencionar los diversos factores ambientales involucrados. Aunque algunos de estos genes están implicados en el metabolismo de la glucosa y

las grasas, la función de las células β y la sensibilidad y secreción de insulina, aún no se ha esclarecido cómo algunas de sus variantes aumentan el riesgo de DM2. De todos modos, es fundamental analizar la epidemiología genética de esta enfermedad en cada población debido a las diferencias subyacentes en los antecedentes genéticos y el estilo de vida entre los grupos étnicos. Por tanto, es posible que los polimorfismos asociados con la DM2 en algunas razas no muestren asociación en otras. Los estudios de asociación de todo el genoma precisarán en última instancia el panorama genético [31].

1.4 Métodos

A continuación, se describen algunos de los métodos utilizados en este trabajo de tesis. Empezaremos con el software Análisis de genética evolutiva molecular (MEGA, por sus siglas en inglés, Molecular Evolutionary Genetics Analysis) es un software de computadora para realizar análisis estadístico de evolución molecular y para construir árboles filogenéticos. Incluye muchos métodos y herramientas sofisticados para filogenómica y filomedicina.

Este en un principio fue utilizado para visualizar las secuencias del genoma y obtener la longitud máxima y mínima de todas las secuencias.

Posteriormente se utilizó el software de MAFFT para realizar la alineación de los datos debido a que eran demasiados.

MAFFT ofrece varias estrategias de alineación múltiple. Se clasifican en tres tipos, (a) el método progresivo, (b) el método de refinamiento iterativo con el puntaje WSP, y (c) el método de refinamiento iterativo utilizando tanto el WSP como los puntajes de consistencia. En general, existe una compensación entre velocidad y precisión. El orden de velocidad es a > b > c, mientras que el orden de precisión es a < b < c. Los siguientes son los procedimientos detallados para las principales opciones de MAFFT [34-36] .

(a) FFT-NS-1, FFT-NS-2 - Métodos progresivos

Estos son métodos progresivos simples como ClustalW. Mediante el uso de varias técnicas nuevas que se describen a continuación, estas opciones pueden alinear una gran cantidad de secuencias (hasta ~5.000) en una computadora de escritorio estándar [34-36].

FFT-NS-1

Es la opción más sencilla progresiva en MAFFT y uno de los métodos más rápidos disponibles en la actualidad. El procedimiento es: (1) hacer una matriz de distancias aproximada contando el número de tuplas de 6 compartidas (ver más abajo) entre cada par de secuencias, (2) construir un árbol guía y (3) alinear las secuencias de acuerdo con el orden de ramificación.

FFT-NS-2

La matriz de distancia utilizado en FFT-NS-1 es muy aproximada y poco fiable. En FFT-NS-2, (4) el árbol guía se vuelve a calcular a partir de la alineación FFT-NS-1, y (5) se lleva a cabo la segunda alineación progresiva.

Las siguientes técnicas se utilizan para mejorar el rendimiento.

k -mer contando. Para acelerar el cálculo inicial de la matriz de distancia, que requiere un tiempo de CPU de $O(N^2)$ pasos, se adopta un método aproximado similar a la opción 'quicktree' de ClustalW, en el que el número de k -mers compartidos por un par de secuencias se cuenta y se considera una aproximación del grado de similitud. MAFFT utiliza el método muy rápido propuesto por Jones et al. (1992) con una modificación menor (Katoh et al. 2002): (1) Los 20 aminoácidos se comprimen en 6 alfabetos, según Dayhoff et al. (1978) y (2) MAFFT realiza la segunda alineación progresiva (FFT-NS-2) para mejorar la precisión.

UPGMA modificado. Se utiliza una versión modificada de UPGMA para construir un árbol guía, que funciona bien para manejar secuencias de fragmentos.

La segunda alineación progresiva. La precisión de la segunda alineación progresiva (FFT-NS-2) es ligeramente superior a la de la primera alineación progresiva (FFT-NS-1) según la prueba BAliBASE, pero la cantidad de tiempo de CPU que requiere FFT-NS-2 es aproximadamente dos veces más largo que el de FFT-NS-1.

(b) FFT-NS-i, NW-NS-i - Método de refinamiento iterativo

La precisión de la alineación progresiva se puede mejorar mediante el método de refinamiento iterativo. Se implementa una versión simplificada de PRRN como la opción FFT-NS-i de MAFFT. En FFT-NS-i, una alineación inicial de FFT-NS-2 se somete a un proceso de refinamiento iterativo [34-36].

FFT-NS-i (máx. 1.000 ciclos)

El refinamiento iterativo se repite hasta que no más mejoría en la puntuación WSP se hace o el número de ciclos alcanza 1.000.

FFT-NS-i (máx. 2 ciclos)

Como la mayoría de la calidad de la mejora se obtiene en la primera etapa de la iteración, esta opción también es útil.

Función objetiva. Se utiliza la puntuación ponderada de suma de pares (WSP) propuesta por Gotoh.

Efecto de FFT. Para probar el efecto de la aproximación FFT, también implementamos las opciones NW-NS-x, en las que la aproximación FFT está deshabilitada, pero los otros procedimientos son los mismos que los del FFT-NS-x correspondiente. No hubo una reducción significativa en la precisión al introducir la aproximación FFT.

(c) L-INS-i, E-INS-i, G-INS-i: métodos de refinamiento iterativo que utilizan WSP y puntuaciones de coherencia

Para obtener alineaciones más precisas en casos extremadamente difíciles, se han agregado tres nuevas opciones, L-INS-i, G-INS-i y E-INS-i, a las versiones recientes (v.≥5) de MAFFT. Estas opciones utilizan una nueva función objetivo que combina la puntuación WSP explicada anteriormente y la puntuación similar a COFFEE, Que evalúa la consistencia entre una alineación múltiple y alineaciones por pares [34, 35].

Para la alineación por pares, se implementan tres tipos diferentes de algoritmos, alineación global (Needleman-Wunsch), alineación local (Smith-Waterman) con costos de brecha afines (Gotoh) y alineación local con costos de brecha afines generalizados (Altschul). Las diferencias en los valores de precisión entre estos métodos son pequeñas para los puntos de referencia disponibles actualmente, como se muestra aquí. Sin embargo, cada uno de ellos tiene características diferentes, según el algoritmo en la etapa de alineación por pares:

E-INS-i

Es adecuado para las alineaciones de esta manera:

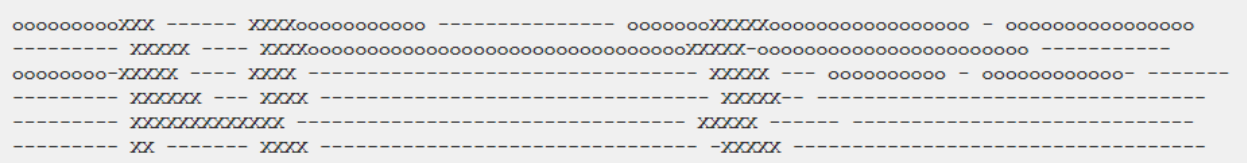


Figura 1. Tipo de Secuencias adecuadas para la alineación con E-INS-i.

donde ' X ' indica residuos alineables, ' o ' indica residuos no alineables y ' - ' indica espacios. Los residuos no alineables se dejan sin alinear en la etapa de alineación por pares, debido al uso del costo de brecha afín generalizado. Por tanto, E-INS-i es aplicable a un problema difícil como la ARN polimerasa, que tiene varios motivos conservados incrustados en regiones largas no alineables. Como E-INS-i tiene el supuesto mínimo de los tres métodos, esto se recomienda si no está clara la naturaleza de las secuencias a alinear. Tenga en cuenta que E-INS-i asume que la disposición de los motivos conservados es compartida por todas las secuencias [34-36].

Los parámetros para E-INS-i funcionan mejor para alinear un conjunto de secuencias largas y secuencias cortas que están estrechamente relacionadas entre sí. Para deshabilitar este cambio, agregue la opción --oldgenafpair.

Con los nuevos parámetros, E-INS-i puede alinear múltiples ADNc y múltiples secuencias genómicas de un gen de especies estrechamente relacionadas. Sin embargo, consume mucho espacio de memoria cuando las secuencias son largas.

L-INS-i

Es adecuado para:

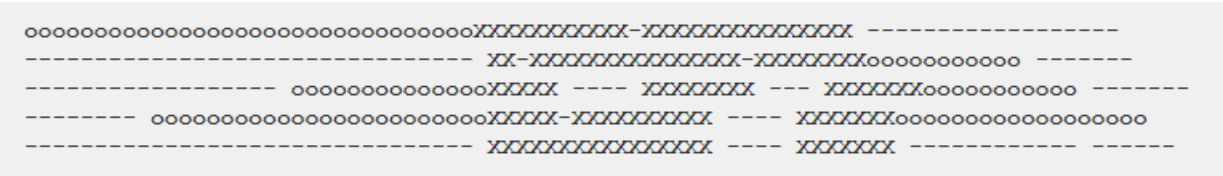


Figura 2. Tipo de Secuencias adecuadas para la alineación con L-INS-i.

L-INS-i puede alinear un conjunto de secuencias que contienen secuencias que flanquean alrededor de un dominio alineable. El algoritmo de Smith-Waterman ignora las secuencias flanqueantes en la alineación por pares. Tenga en cuenta que se supone que las secuencias de entrada tienen solo un dominio alineable. En las pruebas de referencia, la ref4 de BAliBASE corresponde a esto. Las otras categorías de BAliBASE también corresponden a situaciones similares, porque tienen secuencias flanqueantes. L-INS-i también muestra valores de precisión

más altos para una parte de SABmark y HOMSTRAD que G-INS-i, pero no hemos identificado la razón de esto [34-36].

G-INS-i

Es adecuado para:

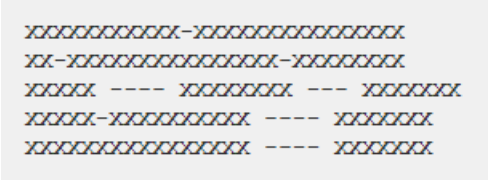


Figura 3. Tipo de Secuencias adecuadas para la alineación con G-INS-i.

G-INS-i asume que toda la región puede alinearse e intenta alinearla globalmente usando el algoritmo Needleman-Wunsch; es decir, se debe extraer un conjunto de secuencias de un dominio truncando las secuencias flanqueantes. En las pruebas de referencia, SABmark y HOMSTRAD corresponden a esto [34-36].

1.4.1 Análisis de Componentes Principales (PCA)

En este trabajo se ha aplicado Análisis de Componentes Principales (de sus siglas en ingles *Principal Component Analysis* (PCA)) de manera directa sobre la secuencia de los nucleótidos. Para esto los datos de secuencia deben transferirse a numeral. Aquí se reemplazan como un vector booleano que se presenta con 0 y 1. Por ejemplo, una base de ADN se registra mediante un conjunto de cinco dígitos (A, T, G, C y -); esta transformación del formato de grabación se conoce como 'codificación de información one-hot' en arquitectura de computadora. Esta transformación tiene el mérito de que no se pierde información y por tanto es completamente reversible; también, es aplicable tanto a secuencias de nucleótidos como de aminoácidos. Las diferencias entre dos muestras se definen mediante sustracción, la distancia se calcula utilizando la longitud euclidiana y la secuencia media se calcula como la media aritmética de las muestras. Algunas de las opciones para ajustes de ponderaciones, sustituciones paralelas, convergentes o hacia atrás pueden ser posibles, como se discutirá más adelante. Luego, la matriz de secuencia cuantificada se rota usando la descomposición de valor singular (SVD), para identificar los componentes principales (PC) [37].

La idea central del análisis de componentes principales (PCA, de sus siglas en inglés) es reducir la dimensionalidad de un conjunto de datos, que consta de un gran número de variables

interrelacionadas, conservando al mismo tiempo la mayor cantidad posible de la variación presente en un conjunto de datos. Esto se logra transformando un nuevo conjunto de variables, las componentes principales (PC's), que no están correlacionadas y se ordenan de tal manera que las primeras conservan la mayor variación presente en todas las variables originales [38].

Formalmente, PCA se define como una transformación lineal ortogonal, que transforma los datos en un nuevo sistema de coordenadas, de modo que la varianza más alta para cualquier proyección de datos se encuentra en la primera coordenada (llamada el primer componente principal), la segunda varianza más grande en la segunda coordenada. PCA es teóricamente la transformación óptima para un conjunto de datos dado, en términos de mínimos cuadrados. El procedimiento para obtener los componentes principales se puede resumir de la siguiente manera: Dado un vector X^T de n dimensiones, $X = [x_1, x_2, \ldots, x_n]^T$, cuyos vectores medios, M, y covarianza, C, están descritos por: $M=E(X)= [m_1, m_2, \ldots, m_n]^T$ y $C = E [(X - M) (X - M)^T]$. Calcule los valores propios $\lambda_1, \lambda_2, \ldots, \lambda_n$ y los vectores propios $P_1, P_2, \ldots, P_n$; y ordenarlos según su magnitud $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$. Seleccione d vectores propios para representar las n variables, $d < n$. Entonces $P_1, P_2, \ldots, P_d$ se denominan componentes principales [38].

Es un método estadístico que permite simplificar la complejidad de espacios muéstrales con muchas dimensiones a la vez que conserva su información. Supóngase que existe una muestra con n individuos cada uno con p variables ($X_1, X_2, \ldots, X_p$), es decir, el espacio muestral tiene p dimensiones. *PCA* permite encontrar un número de factores subyacentes ($z<p$)($z<p$) que explican aproximadamente lo mismo que las p variables originales. Donde antes se necesitaban p valores para caracterizar a cada individuo, ahora bastan z valores. Cada una de estas z nuevas variables recibe el nombre de componente principal [39].

Análisis de Componentes Principales pertenece a la familia de técnicas conocida como *unsupervised learning*. Los métodos de *supervised learning* tienen el objetivo de predecir una variable respuesta Y a partir de una serie de predictores. Para ello, se dispone de *p* características ($X_1, X_2, \ldots, X_p$) y de la variable respuesta Y medidas en *n* observaciones. En el caso de *unsupervised learning*, la variable respuesta Y no se tiene en cuenta ya que el objetivo no es predecir Y sino extraer información empleando los predictores, por ejemplo, para identificar subgrupos. El principal problema al que se enfrentan los métodos de *unsupervised learning* es la

dificultad para validar los resultados dado que no se dispone de una variable respuesta que permita contrastarlos [39].

El método de *PCA* permite por lo tanto "condensar" la información aportada por múltiples variables en solo unas pocas componentes. Esto lo convierte en un método muy útil de aplicar previa utilización de otras técnicas estadísticas tales como regresión, *clusterin* [39].

1.4.2 Analisis de Entropía

La teoría de la entropía de Shannon, desarrollada inicialmente por Claude E. Shannon, se aplica para medir el contraste entre criterios y esta información se utiliza para tomar decisiones. En este análisis se indica que para todo pi dentro de una distribución de probabilidad P, existe una medida H, que satisface las siguientes propiedades[40]:

1. H es una función positiva continua,

2. Si todos los p_i son igual y $p_i = 1/n$, entonces H debería ser una función monótona creciente de n; y,

3. Para todos, $n \geq 2$,

$$H(p_1, p_2, \dots, p_n) = h(p_1 + p_2, p_3 \dots, p_n) + (p_1 + p_2)H\left(\frac{p_1}{p_1 + p_2}, \frac{p_2}{p_1 + p_2}\right)$$

Shannon demostró que la única función que satisface estas condiciones es:

$$\boldsymbol{H_{Shannon}} = -\sum_i^n \boldsymbol{p_i}\ \mathbf{log}(\boldsymbol{p_i}) \qquad (1)$$

Donde: $0 \leq p_i \leq 1;\ \sum_{i=1}^{n} p_i = 1$

La entropía (H) es una medida de incertidumbre para una variable aleatoria discreta y es análoga a la variación en datos continuos. Tradicionalmente, la base del logaritmo para la entropía se calcula con bits unitarios (b=2), nats (b=e) o dits (b=10). Alternativamente, las estimaciones de entropía se pueden normalizar a una escala común donde 0<=H<=1 estableciendo b=n, el número de estados posibles. Para secuencias de ADN (n=4 nucleótidos) o proteínas (n=20 aminoácidos), la entropía normalizada H=0 indica un sitio invariable mientras que H=1 representa un sitio donde todos los estados ocurren con la misma probabilidad. Atchley et al 1999 clasificaron los aminoácidos según sus atributos fisicoquímicos para formar (n=8) grupos funcionales. Junto con la entropía AA, el valor de entropía GroupAA puede proporcionar información sobre las diferencias en la variación funcional y filogenética. Grupos AA: ácido = DE, alifático = AGILMV, amínico = NQ, aromático = FWY, básico = HKR, cisteína = C, hidroxilado = ST, prolina = P. Las

brechas se ignoran sitio por sitio, por lo que los valores de entropía pueden tener un número diferente de observaciones entre sitios. Las secuencias deben tener la misma longitud [40].

La fórmula de la entropía es $H = -\sum p*\log(p)$, donde:

- p= frecuencia de la base para calcular la entropía.
- Counts = matriz de números enteros que cuentan la presencia de cada carácter (DNA, AA o GroupAA) en cada sitio.
- Freq = matriz de frecuencias de caracteres (DNA, AA o GroupAA). Estos son simplemente recuentos de caracteres divididos por el número total de caracteres (sin espacios) en cada sitio.
- H = vector de valores de entropía para cada sitio.

1.4.2.1 Índice de diversidad de Shannon

Determinar la diversidad de la población es una parte importante de las estadísticas de un ecosistema. El índice de diversidad de Shannon es uno de varios índices estadísticos utilizados para medir la riqueza de especies. Sin embargo, existen varios otros índices de diversidad, como el índice de Simpson y el índice de Shannon Weiner [41].

El índice de Shannon es una herramienta matemática para calcular la abundancia proporcional de especies en un lugar determinado. Este tipo de abundancia de especies proporciona una indicación de la diversidad biológica en esa área. Se relaciona con la entropía de la información en que la fórmula de los dos conceptos es idéntica. La diferencia es que en la entropía de la información el valor p es la probabilidad de un arreglo particular, pero con el índice de Shannon p = n/N donde "n" es el número de individuos de una especie dada y N es el número total de individuos [41].

1.4.3 Modelos de Regresión

El objetivo de un modelo de regresión lineal es intentar explicar la relación entre una variable dependiente (variable de respuesta) y un conjunto de variables independientes (variables explicativas) $X_1,..., X_n$. En un modelo de regresión lineal simple, intentamos explicar la relación entre la variable de respuesta (Y) y una única variable explicativa (X). Utilizando las técnicas de regresión de una variable Y sobre una variable X, buscamos una función que sea una buena aproximación de una nube de puntos (x_i, y_i), mediante una curva [42].

La dependencia de la variable puede ser una regresión univariada o multivariada. La regresión univariante identifica la dependencia entre una sola variable como se representa en la ecuación (2) [43].

$$\boldsymbol{Y = \alpha + \beta X + \varepsilon} \quad (2)$$

Donde y es una variable dependiente, x es una variable independiente con coeficiente β (es la pendiente de la recta e indica cómo cambia Y cuando X aumenta en una unidad) y α es una constante (es la ordenada en el origen, la valor que toma Y cuando X es 0), y ε una variable que incluye un gran conjunto de factores, cada uno de los cuales influye en la respuesta sólo en una pequeña magnitud, a la que llamaremos error. X e Y son variables aleatorias, por lo que no se puede establecer una relación lineal exacta entre ellas [42]. Si bien la regresión multivariada consiste en identificar la dependencia entre varias variables simultáneamente, se representa en la ecuación (3) [43].

$$\boldsymbol{Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon} \quad (3)$$

Donde ε es el término de error, β_0 es la intersección, β_1-β_k son coeficientes de regresión parcial, por ejemplo, β_i cuando $1 \leq i \leq k$ representa el cambio en la respuesta media correspondiente a un cambio unitario en x_i cuando las otras variables permanecen constantes.

Los modelos de regresión predicen el resultado de las variables dependientes a partir de las variables independientes. Se considera la importancia en el análisis de regresión para manejar problemas más complicados [43]. El objetivo de la regresión lineal múltiple es resolver el conjunto de coeficientes $\Theta = \{\beta_0, \beta_1, \dots, \beta_k\}$ dadas las observaciones X y los objetivos Y [44].

1.4.3.1 Regresión Lineal

La regresión lineal es el modelo predictivo más común para identificar la relación entre variables. Puede ser una regresión lineal simple o lineal múltiple. La regresión lineal se describe en la ecuación (4) [43].

$$\boldsymbol{y = x\beta + \varepsilon} \quad \textbf{(4)}$$

En la ecuación (4) y es la variable independiente y puede ser un valor continuo o categórico; x es una variable dependiente que siempre es un valor continuo. Analiza una distribución de

probabilidad y se centra principalmente en la distribución de probabilidad condicional con análisis multivariado [43].

1.4.3.2 Regresión Lineal Simple

El proceso de regresión lineal simple que se muestra en la Figura 4 es un análisis de regresión que utiliza una única variable independiente y se describe en la ecuación (2). De manera similar a cómo la correlación expande la relación entre dos variables, la regresión lineal simple distingue entre variables dependientes e independientes; sin embargo, la correlación no lo hace [43].

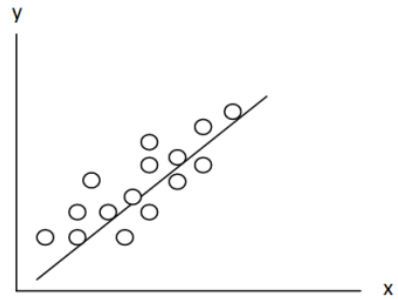


Figura 4. Regresión Lineal Simple [43].

1.4.3.3 Regresión Lineal Multiple

La regresión lineal múltiple o multivariada (MLR) representada en la Figura 5 es el proceso de predicción con más de una variable predictiva o independiente que es similar al análisis multivariado como se describe en la ecuación (3) [43].

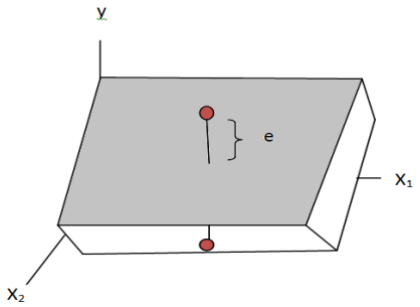


Figura 5. Regresión Lineal Multiple [43].

Una técnica estadística conocida como regresión lineal múltiple utiliza muchas variables explicativas para predecir el resultado de una variable de respuesta. El objetivo de la regresión lineal múltiple es modelar la relación entre las variables explicativas y de respuesta. El siguiente modelo es un modelo de regresión lineal múltiple con k variables predictoras., $x_1, \ldots, x_k$ [44].

El problema de MLR frecuentemente se resuelve usando mínimos cuadrados. Si cada variable predictiva x_1, x_2, . . . , x_k tiene n observaciones. Entonces x_{ij} representa la i-ésima observación de la j-ésima variable predictora x_j. Tal como, x_{31} representa el primer valor de la tercera observación. Específicamente, la ecuación (3) lo anterior se puede expresar como [44]:

$$\boldsymbol{y_j = \beta_0 + \beta_1 X_{j1} + \beta_2 X j_2 + \cdots + \beta_k X_{jk} + \varepsilon_j} \quad (5)$$

Donde $1 \leq j \leq n$, y_j es la j-ésima valor objetivo. El sistema de n ecuaciones se puede representar como una matriz de diseño como se muestra en la ecuación (2), y describe los niveles de las variables predictoras adquiridas en cada observación. Todos los coeficientes de regresión están contenidos en el vector β. Las estimaciones de mínimos cuadrados, que se indican a continuación, se utilizan para estimar y crear el modelo de regresión β [44].

$$\boldsymbol{\hat{\beta} = (X^T X)^{-1} X^T y} \quad (6)$$

Entonces el valor estimado de y se puede calcular de la siguiente manera después de obtener $\hat{\beta}$ [44].

$$\begin{aligned} \boldsymbol{\hat{y}} &= \boldsymbol{X\hat{\beta}} \\ \boldsymbol{\epsilon} &= \boldsymbol{y - \hat{y}} \end{aligned} \quad (7)$$

El propósito de utilizar datos de regresión fue buscar SNP estadísticamente asociados con la DM2.

1.4.4 Comparaciones Múltiples: Corrección del Valor P

El termino comparación hace referencia a la comparación de dos grupos mediante un contraste de hipótesis (t-test, U-test, …). El valor p devuelto por el test refleja la probabilidad de obtener una diferencia igual o más extrema que la observada, siendo cierta la hipótesis nula. Considerar la diferencia observada como significativa depende de si el valor p obtenido está por debajo de un límite establecido por el investigador al que se conoce como nivel de significancia α [45].

Las comparaciones múltiples surgen cuando un estudio estadístico conlleva la realización de varias comparaciones con el objetivo de identificar aquellos grupos para los que las diferencias son más significativas. La siguiente tabla recoge los posibles resultados obtenidos al contrastar m hipótesis nulas [45].

Tabla 4. Resultados de contrastar m hipótesis nulas.

	Test considerado significativo	Test considerado no significativo	Total
Hipótesis nula verdadera (H0)	F	$m_0 - F$	m_0
Hipótesis alternativa verdadera (HA)	T	$m_1 - T$	m_1
Total	S	$m - S$	m

- m: número total de hipótesis contrastadas.
- m_0: número de hipótesis nulas verdaderas, en la práctica este parámetro se desconoce.
- m_1: número de hipótesis alternativas verdaderas (hipótesis nulas falsas), en la práctica este parámetro se desconoce.
- F: número de falsos positivos (error tipo I, *false discoveries*).
- T: número de verdaderos positivos, hipótesis alternativas correctamente detectadas (*true discoveries*).
- m_0−F: número de verdaderos negativos.
- m_1−T: número de falsos negativos (error tipo II).
- S: número de hipótesis nulas rechazadas (test significativo), independientemente de que sean ciertas o falsas.
- m−S: número de hipótesis alternativas rechazadas, independientemente de que sean ciertas o falsas.

La estimación del false positive rate, probabilidad de rechazar la hipótesis nula (test significativo) siendo esta cierta, se obtiene como [45]:

$$F/m_0$$

La estimación del false discovery rate, probabilidad de que la hipótesis nula sea cierta a pesar de haber sido rechazada (test significativo), se obtiene como [45]:

$$F/S$$

1.4.5 False Discovery Rate (FDR)

Los métodos descritos anteriormente se centran en corregir la inflación del error de tipo I (false positive rate), es decir, la probabilidad de rechazar la hipótesis nula siendo esta cierta. Esta aproximación es útil cuando se emplea un número limitado de comparaciones. Para escenarios de large-scale multiple testing como los estudios genómicos, en los que se realizan miles de test de

forma simultánea, el resultado de estos métodos es demasiado conservativo e impide que se detecten diferencias reales. Una alternativa es controlar el false discovery rate [45].

El false discovery rate (FDR) se define como: (todas las definiciones son equivalentes)

- La proporción esperada de test en los que la hipótesis nula es cierta, de entre todos los test que se han considerado significativos.
- FDR es la probabilidad de que una hipótesis nula sea cierta habiendo sido rechazada por el test estadístico.
- De entre todos los test considerados significativos, el FDR es la proporción esperada de esos test para los que la hipótesis nula es verdadera.
- Es la proporción de test significativos que realmente no lo son.
- La proporción esperada de falsos positivos de entre todos los test considerados como significativos.

El objetivo de controlar el false discovery rate es establecer un límite de significancia para un conjunto de test tal que, de entre todos los test considerados como significativos, la proporción de hipótesis nulas verdaderas (falsos positivos) no supere un determinado valor. Otra ventaja añadida es su fácil interpretación, por ejemplo, si un estudio publica resultados estadísticamente significativos para un FDR del 10%, el lector tiene la seguridad de que, como máximo, un 10% de los resultados considerados como significativos son realmente falsos positivos [45].

Cuando un investigador emplea un nivel de significancia α, por ejemplo, de 0.05, suele esperar cierta seguridad de que solo una pequeña fracción de los test significativos se correspondan con hipótesis nulas verdaderas (falsos positivos). Sin embargo, esto no tiene por qué ser así. La razón por la que un false positive rate bajo no tiene por qué traducirse en una probabilidad baja de hipótesis nulas verdaderas entre los test significativos (false discovery rate) se debe a que esta última depende de la frecuencia con la que la hipótesis nula contrastada es realmente verdadera. Un caso extremo sería el planteado en el ejemplo 1, en el que todas las hipótesis nulas son realmente ciertas y por lo tanto el 100% de los test que resultan significativos son falsos positivos. Así pues, la proporción de falsos positivos (false discovery rate) depende de la cantidad de hipótesis nulas que sean ciertas de entre todos los contrastes [45].

Los análisis de tipo exploratorio en los que el investigador trata de identificar resultados significativos sin apenas conocimiento previo se caracterizan por una proporción alta de hipótesis nulas falsas. Los análisis que se hacen para confirmar hipótesis, en los que el diseño se ha orientado en base a un conocimiento previo, suelen tener una proporción de hipótesis nulas verdaderas alta. Idealmente, si se conociera de antemano la proporción de hipótesis nulas verdaderas de entre todos los contrastes se podría ajustar con precisión el límite significancia adecuado a cada escenario, sin embargo, esto no ocurre en la realidad [45].

1.4.6 Factor de Probabilidades (Factor Risk)

Una razón de probabilidades (OR, de sus siglas en inglés odds ratio) es una medida de asociación entre una exposición y un resultado. El OR muestra la probabilidad de que ocurra una ocurrencia dada una exposición específica en comparación con la probabilidad del resultado en ausencia de esa exposición. Los estudios de casos y controles son las aplicaciones más frecuentes de los odds ratios [46].

El odds ratio se utiliza para comparar la probabilidad de un resultado (como una enfermedad o trastorno) debido a la exposición a una variable particular (p. ej., característica de salud, elemento del historial médico). El odds ratio también se puede utilizar para evaluar si una exposición específica representa un riesgo para un resultado específico y para evaluar la importancia relativa de varias variables de riesgo para ese resultado [46].

- OR=1 Las probabilidades de resultado no se ven afectadas por la exposición.
- OR>1 La exposición está vinculada a mayores probabilidades de éxito.
- OR<1 La exposición está vinculada a una probabilidad reducida de éxito.

Se calcula utilizando el intervalo de confianza (IC) del 95% para determinar la precisión del OR. Una precisión OR alta se indica con un IC pequeño, mientras que una precisión OR baja se muestra con un IC grande. Es importante señalar que el IC del 95% no proporciona información sobre la significancia estadística de una medida, a diferencia del valor p. En realidad, si el IC del 95% no se superpone al valor nulo (por ejemplo, OR=1), con frecuencia se considera un marcador de significación estadística. Por lo tanto, sería incorrecto interpretar un IC del 95% OR que incluye el valor nulo como muestra de que la exposición y el resultado no están relacionados [46].

Para definir los factores de riesgo, se inspeccionaron las posiciones de cada par de bases que se encontraron significativas en el análisis de asociación (análisis de regresión). Se aplicaron los criterios de cálculo del Odds Ratio (OR) y la definición de factor de riesgo, como se describe en [46]. Con base en los hallazgos se examinaron la significación estadística, el valor de OR y el rango de confianza del 95% para cada variable. Luego, cada posición de par de bases que cumplía los requisitos posteriores se declaró como factor de riesgo:

1. Si la significación estadística de la posición del par de bases (valor p) fue inferior a 0.05;
2. El odds ratio (OR) no era igual a 1; y y
3. El rango de confianza del 95% para el odds ratio no contenía 1.

Por lo tanto, si una posición de par de bases cumple con estos tres criterios y su OR > 1, se declara como un factor de riesgo asociado con una mayor probabilidad de diabetes. De la misma manera, si la variable cumplió las tres condiciones, y su OR < 1, se declara como factor de riesgo asociado a menor probabilidad de padecer diabetes.

1.5 Planteamiento del Problema

La Diabetes tipo 2 se considera un trastorno poligénico, en el que cada variante genética confiere un efecto parcial y aditivo. Sólo del 5-10% de los casos de DM2 se deben a defectos de un solo gen; estos incluyen la diabetes de inicio en la madurez de los jóvenes, los síndromes de resistencia a la insulina, la diabetes mitocondrial y la diabetes neonatal [5]. El examen de los genes de susceptibilidad a la DM2 puede ser útil para la predicción, la prevención y el tratamiento temprano de la enfermedad. Mediante la implementación de estudios de asociación de genoma completo (GWAS), el número de variantes genéticas comunes asociadas con la DM2 ha aumentado rápidamente. Los primeros estudios GWAS para DM2 [6, 7, 9, 47, 48] y glucosa en ayunas [49] identificaron con éxito más de 40 loci genéticos asociados a la DM2. Estudios realizados principalmente en poblaciones europeas [13].

Recientemente, mediante un estudio de Metanálisis de GWAS de DM2 [50] y rasgos cuantitativos glucémicos [51] se aumentó drásticamente el número de loci asociados con DM2 significativos en todo el genoma en poblaciones europeas; la mayoría de estas variantes actúan a través de defectos

en la función de las células beta en lugar de la acción de la insulina. Juntas, las variantes que se sabe que están asociadas con DM2 explican ~10 % de la variación genética [50, 52], lo que indica que es probable que loci adicionales y señales independientes en loci establecidos contribuyan al riesgo de enfermedad. Este metanálisis a gran escala centrado en genes de 39 estudios multiétnicos de asociación de DM2 identificó tres loci de riesgo de DM2 europeos (GATAD2A/CILP2/PBX4, previamente conocido por tener efectos protectores sobre los lípidos; TH/INS, previamente conocido por tener efectos protectores sobre T1D y SREBF1), un locus de riesgo afroamericano de DM2 (HMGA2) y un locus de riesgo multiétnico (BCL2) y confirmó que una puntuación genética de alelos de riesgo de DM2 influye en el riesgo de DM2 en poblaciones multiétnicas que incluyen afroamericanos, hispanos y asiáticos. Por lo tanto, los GWAS multiétnicos y bien potenciados de DM2 deberían conducir al descubrimiento de genes adicionales asociados con la diabetes relevante para múltiples grupos étnicos [13].

Las regiones genéticas identificadas solo explican una pequeña proporción de la heredabilidad estimada de la DM2, lo que sugiere que quedan por identificar factores genéticos adicionales. Una limitación de GWAS es la gran cantidad de hipótesis y el alto costo económico de estas investigaciones [14]. Varios estudios han abordado la viabilidad y eficacia de los GWAS basados en agrupaciones, con ahorros considerables en tiempo y costo [14-16]. Además, la secuenciación del genoma completo en múltiples muestras en una población brinda una oportunidad sin precedentes para caracterizar de manera integral las variantes polimórficas en la población [17].

Los factores genéticos que contribuyen a la DM2 se comprenden menos en las poblaciones no europeas. Se identificó un locus nuevo (KCNQ1 [MIM 607542]) sobre la base de un GWAS en una población japonesa [53, 54] y posteriormente se demostró que alberga alelos independientes en individuos de ascendencia europea [50]. Más recientemente, los GWAS en poblaciones chinas [5, 23], japonesas [55], y del sur de Asia [56] describen loci adicionales de DM2 que superan la importancia de todo el genoma. Hasta la fecha, los GWAS DM2 en afroamericanos han tenido poca potencia para detectar nuevos loci [57].

Es innegable el avance en el entendimiento de la genética asociada a la DM2, sin embargo, para el desarrollo de medicina basada en el genoma de la diabetes, es necesario la caracterización no solo del genoma nuclear, sino del genoma mitocondrial, que está demostrado tiene genes asociados a la enfermedad. Es necesario realizar más estudios con el propósito de identificar nuevas

posiciones genómicas de base-par (bp) estadísticamente asociadas a DM2. Esto implica precisar técnicas para identificar regiones genómicas con variación estadísticamente significantes y medir su asociación con la DM2. Igualmente es necesario identificar métodos eficaces para probar si las variaciones asociadas son factores de riesgo reales para el desarrollo de la enfermedad.

1.6 Objetivo y Metas

El objetivo de este trabajo es realizar un estudio de asociación en el genoma mitocondrial para identificar nuevas posiciones genómicas base-par (bp) estadísticamente asociadas a DM2.

Las metas propuestas son las siguientes:

1. Obtener y organizar, de la base de datos de nucleótidos del Centro Nacional de Información Biotecnológica (NCBI) (https://www.ncbi.nlm.nih.gov/nucleotide), una base de información de genomas mitocondriales completos de humanos, que contenga individuos con Diabetes Mellitus tipo 2, he individuos sanos.
2. Realizar un análisis de alineación de los genomas para visualizar y seleccionar la o las regiones genómicas con variabilidad alélica.
3. Implementar un método estadístico para probar la significancia estadística de las regiones variables.
4. Realizar una minería de datos en las regiones genómicas variables para visualizar su complejidad.
5. Implementar un análisis de asociación de las posiciones variables con la DM2, mediante técnicas de regresión estadística.
6. Implementar una técnica para probar si las posiciones base-par variantes asociadas con DM2, son factores de riesgo para desarrollar la enfermedad.

CAPITULO II DESCRIPCIÓN DE LOS DATOS

En este capítulo, se describe la fuente de donde fueron recabados los datos (genomas mitocondriales completos), los archivos que fueron descargados y el formato de la información.

2.1 Obtención de los Datos

Para este trabajo de tesis se obtuvieron secuencias de genoma completo mitocondrial de 510 individuos. Los genomas fueron descargados de la base de datos del portal de información sobre diabetes tipo 2: *http://www.type2diabetesgenetics.org/*. De estos genomas, 437 correspondieron a individuos que padecen diabetes Mellitus Tipo 2, y 73 correspondieron a individuos sanos de Diabetes. El Anexo A presenta la lista de los identificadores (ID) de los genomas obtenidos para este trabajo.

2.2 Base Datos

La base de datos está separada en dos archivos con formato *fasta* (*fasta* es un formato de fichero informático basado en texto, utilizado para representar secuencias de ácidos nucleicos, de péptido, y en el que los pares de bases o los aminoácidos se representan usando códigos de una única letra). El primer archivo, llamado: *sequence (1),* tiene 437 secuencias de genoma mitocondrial completo de pacientes humanos con DM2 y el segundo archivo, llamado: *sequence (2),* tiene 73 de individuos sanos (cada secuencia es del cromosoma mitocondrial completo más común o dominante en cada individuo, con 16,569 bases-par de longitud, aunque puede variar algunos nucleótidos entre individuos diferentes). La secuencia más corta es de 16,554 bases-par.

Ambos archivos fueron combinados en un solo archivo el cual fue llamado combinación, en este nuevo archivo primero se colocaron las personas enfermas y después las sanas. Para tener un total de 510 secuencias de genoma mitocondrial completo. Las secuencias fueron alineadas con el software MEGA (https://www.megasoftware.net/). La alineación resultó con una longitud de 16,609 nucleótidos. Después de esto se realizó inspección visual de la alineación buscando regiones que presentaran mayor variabilidad. La región visualmente más variable fue de la posición 16, 170 a la 16,410 con un total de 241 posiciones.

Los nucleótidos fueron transformados a formato numérico de la siguiente manera: A = 1, C = 2, G = 3 y T = 4, GAP(-) = 5. Algunas posiciones de las secuencias, en vez de tener uno de los A, G C,

o T, tenían otras letras, tales como: R, Y, W, N. estas letras fueron cambiadas por el número 9. Estas últimas letras de acuerdo a la nomenclatura IUPAC corresponden a:

R = GA (purine - purina)

Y = TC (pyrimidine - pirimidina)

W = AT (weak bonds - lazos débiles)

N = AGCT (any - alguna)

La Figura 6 muestra la nomenclatura del código de ácidos nucleicos.

IUPAC nucleotide code	Base
A	Adenine
C	Cytosine
G	Guanine
T (or U)	Thymine (or Uracil)
R	A or G
Y	C or T
S	G or C
W	A or T
K	G or T
M	A or C
B	C or G or T
D	A or G or T
H	A or C or T
V	A or C or G
N	any base
. or -	gap

Figura 6. Código de Ácidos Nucleicos-base.

Además, se buscó información extra sobre las secuencias como por ejemplo a que etnias pertenecían las personas. De esta investigación se encontró la siguiente información. Las secuencias que comienzan con KC como, por ejemplo, KC252407.1 son personas de Taiwán. Las secuencias que comienzan con HQ como, por ejemplo, HQ610991.1 son pacientes con diabetes tipo 2 con nefropatía, por los autores y la revista parecen ser indios. Las secuencias que comienzan con JF como, por ejemplo, JF717361.1 son personas de Italia. Las secuencias que comienzan con KF como, por ejemplo, KF898145.1 por los artículos y la revista parecen ser de china. Las secuencias que comienzan con AP como, por ejemplo, AP008824.1 son personas de Japón

(japoneses1). Específicamente Paciente diabético japonés de Aichi. *No se encontró la lista de la base de datos*. Las secuencias que comienzan con AB como, por ejemplo, AB055387.1 son personas de Japón (japoneses2). Específicamente paciente japonés cardiomiopático. *No se encontró la lista de la base de datos*.

A continuación, se menciona el número de personas de cada etnia en la base de datos. Se colocan las dos primeras letras de la base de datos y a que etnia pertenecen entre paréntesis el número que ocupan en el archivo combinación, posteriormente el total de estas personas, y por ultimo un color con el cual se identifican en algunas gráficas.

KC = taiwaneses (80-245, 438-510), total=166(enfermos)+73(sanos),

color=purpura-enfermos, azul-sanos.

HQ = indios (1-19, 26-68), total=19+43=62, color=rojo.

JF = italianos (20-25), total=6, color=verde.

KF = chinos (69-79), total=11, color=amarillo.

AP = japoneses1 (246-436), total=191, color=naranja.

AB = japoneses2 (437), total=1, color=negro.

CAPÍTULO III ANÁLISIS DE LOS DATOS

En este capítulo se presentan los distintos análisis realizados a la base de datos, iniciando con la alineación de los genomas, posteriormente la detección de cúmulos a través de la técnica de Análisis de Componentes Principales (PCA), también se presentan las Proporciones Polimórficas de los genomas y los análisis de Regresión Simple y Múltiple Aplicados a los datos para encontrar las posiciones base-par asociadas con DM2.

3.1 Alineación

El análisis de alineación de los 510 genomas de personas enfermas y sanas se realizó con el software MAFFT (https://mafft.cbrc.jp/alignment/server/). La alineación resultante fue generada en un archivo de texto, en el cual las secuencias se presentaron en el mismo orden en que estaban en el archivo original. Posteriormente, las secuencias en la alineación resultante fueron codificadas de nucleótidos en formato A, C, G y T, a valores numéricos para poder ser analizados. La codificación fue la siguiente:

Tabla 5. Configuración de nucleótidos a valores numéricos.

Nucleótidos	Valor Numérico
A	1
C	2
G	3
T	4
GAP (-)	5
R, Y, W, N	9

Posteriormente, por medio de una inspección visual se localizó la región con mayor variabilidad. Esta comprende de la posición 16,170 a la 16,410.

En la Figura 7 se puede apreciar la alineación de todas las secuencias de nucleótidos, pero en la región de interés que es la de mayor variabilidad.

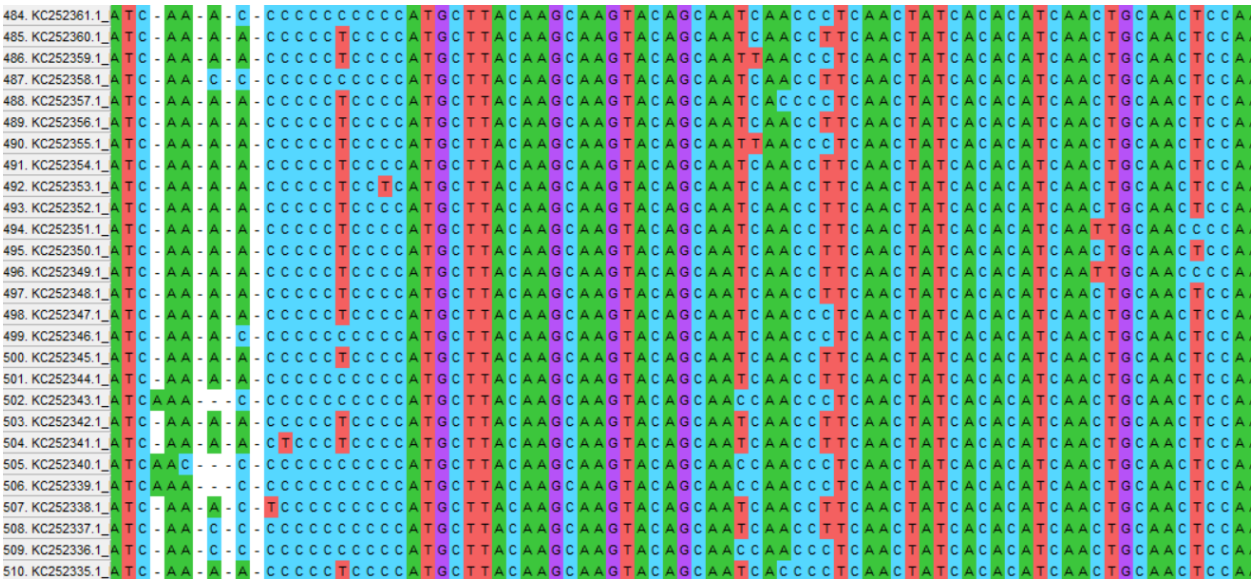


Figura 7. Alineación de las secuencias en el área de mayor variabilidad.

3.2 Prueba de Significancia Estadística de la Región Variable

En una primera etapa se calculó la entropía de una secuencia de ADN, utilizando un programa propio en el lenguaje R.

El resultado de la entropía de la región de mayor variabilidad es **16.83817**.

También se calculó la entropía para las secuencias sin tomar en cuenta la región variante, para esto se secciono la secuencia completa en fragmentos de 241 posiciones, debido a que este es el tamaño que tiene la región de mayor variabilidad.

El resultado promedio de la entropía de todas las secuencias sin contar la región variante es **1.6469847**. Como podemos comprobar hay una gran diferencia entre las entropías.

Región Variante	Resto de la Secuencia
16.83817	**1.6469847**

Después de realizar los cálculos de la entropía, se realizó una prueba de hipótesis con la prueba *t student*, para demostrar que ambos valores de entropía son estadísticamente diferentes, donde la hipótesis nula es $H_0 = |t| > t_{n-1,1-\frac{\alpha}{2}}$, mientras que la hipótesis alternativa es $H_1 = |t| \leq t_{n-1,1-\frac{\alpha}{2}}$. Por lo tanto, se calculó el valor de t como sigue:

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

Donde $\bar{x} =$ media de la distribución de los datos,

$\mu_0 =$ media de la población,

s= error estándar de la muestra,

n= tamaño de la muestra.

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{1.64 - 16.83}{\frac{1.08}{\sqrt{68}}} = \mathbf{-155.54}$$

Posteriormente en R se encontró el percentil de la prueba t para comprobar las hipótesis.

qt(p=0.975, df=67, lower.tail=TRUE)

Lo cual da como resultado **1.996008**, con estos resultados H_0 se rechaza, y por lo tanto, la región que fue seleccionada como variante es estadísticamente la región de mayor variabilidad en toda la secuencia.

Un logotipo de secuencia es una representación gráfica de la conservación de la secuencia de nucleótidos (en una hebra de ADN/ARN) o de aminoácidos (en secuencias de proteínas). Un logotipo de secuencia se crea a partir de una colección de secuencias alineadas y representa la secuencia de consenso y la diversidad de las secuencias. Los logotipos de secuencia se utilizan con frecuencia para representar características de secuencia, como sitios de unión a proteínas en el ADN o unidades funcionales en proteínas [58].

Los "bits" aquí están relacionados con los utilizados en la entropía de Shannon. Este es un equivalente teórico de la información a la entropía de Boltzmann de la termodinámica. Esta es una medida de qué tan "desordenada" está la posición, o más precisamente qué tan específica es la distribución [59].

La fórmula general para la entropía total (expresada en base a probabilidades/fracciones) es

$$S = -k \sum_i p_i * \log(p_i)$$

Dónde i varía a través de todos los diversos estados (por ejemplo, cada nucleótido).

La diferencia entre la entropía de Shannon y la entropía de Boltzmann (aparte de los contextos en los que se usa) es el valor de la constante (una para Shannon, la constante de Boltzmann para la entropía de Boltzmann) y qué logaritmo se usó (el logaritmo natural para la entropía de Boltzmann, y típicamente base- 2 para la entropía de Shannon.) Como utiliza logaritmos de base 2, la entropía de Shannon se suele medir en unidades denominadas "bits" [59].

El problema es que la entropía de una distribución no es realmente lo que desea mostrar en un logotipo de secuencia. En cambio, desea algo que se haga más grande (no cero) cuando la secuencia se vuelve más definida. Como tal, los valores que se muestran no son la entropía en sí, sino la "pérdida de entropía" de una distribución completamente aleatoria (entropía máxima) [59].

Así es como obtienes el número total de bits en una posición (la altura total). La altura de las letras individuales se obtiene tomando la altura total y multiplicándola por la probabilidad de cada estado. Es posible que esto no tenga una justificación teórica rígida (no se puede atribuir la pérdida de entropía a nucleótidos individuales como ese), pero cumple con los propósitos de visualización de hacer que las identidades de nucleótidos más prevalentes sean más grandes en tamaño de visualización [59]. Por ejemplo, si tiene una posición con 70% G y 30% C, entonces tiene:

$$\left(4 * -0.25 * log_2(0.25)\right) - \left(-0.7 * log_2(0.7) + -0.3 * log_2(0.3)\right) = 1.12$$

bits de altura total. G obtiene una altura de $1.12 * 0.7 = 0.78$ bits, mientras que C obtiene$1.12 * 0.3 = 0.34$ bits.

3.3 Minería de Datos

El proceso de aplicar PCA fue realizado a través de lenguaje R. En este caso las secuencias primero fueron alineadas y después seccionadas para quedarnos con la región de mayor variabilidad. Se puede observar una mejor visualización de nuestros componentes principales en la Figura 8 para ver la aportación de cada uno de ellos.

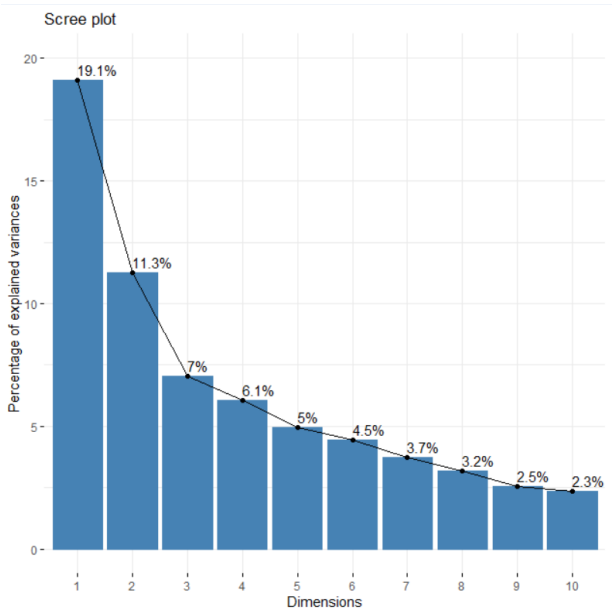


Figura 8. Primeros 10 componentes principales.

Como podemos ver de la figura anterior la primera y segunda componente son los que más varianza tienen. Ahora estos dos componentes los visualizaremos en la Figura 9 para observar si existen cúmulos o clusters donde podamos apreciar que existen grupos marcados donde la mayor cantidad de datos proviene de personas sanas o con diabetes.

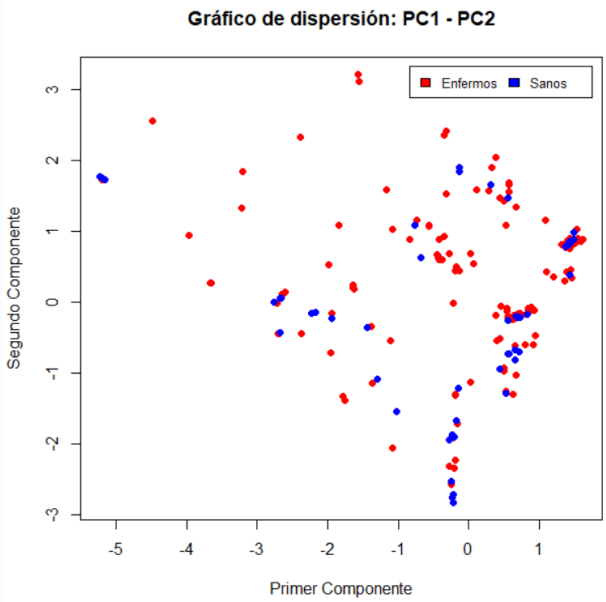


Figura 9. Gráfico de dispersión de PC1-PC2 de Sanos vs Enfermos.

También se realizó un análisis de PCA distinguiendo entre las diferentes etnias de las que se conforma la base de datos obteniendo los resultados mostrados en la Figura 10.

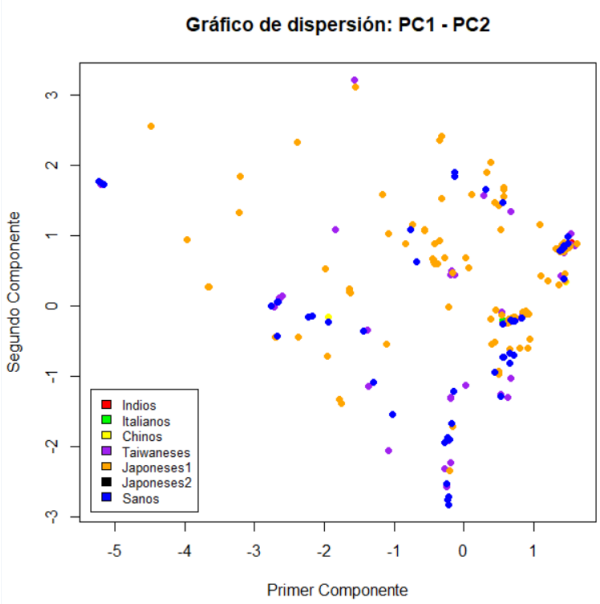


Figura 10. Gráfico de dispersión de PC1-PC2 de todas las etnias.

3.4 Análisis de Asociación

A continuación, se mostrarán los análisis de regresión simple y múltiple que fueron aplicados sobre los vectores de la región de mayor variabilidad (región de interés), esta región cuenta con 242 valores, pero la primera columna no la tomaremos en cuenta porque esa será usada para identificar a las personas enfermas de las sanas.

3.4.1 Regresión Simple

Se aplicó un análisis de regresión simple, en donde se buscó aquellas secuencias que tuvieran valores de p<0.05. Los cálculos p fueron realizados de manera individual, es decir, uno a la vez, debido a que hubo algunos problemas para automatizar este proceso. En total eran 242 posiciones de la región de mayor variabilidad. De los 242 valores aquellos que resultaron significantes o muy cercanos al valor p se pueden observar en la Tabla 6 y la

Tabla 7. De estas tablas podemos observar que 8 valores de posiciones logran superar el p<0.05 y otros 5 valores se quedaron muy cerca de ser lo.

Tabla 6. Valores p cercanos a 0.05.

Posición	Valor p
X[, 4]	0.0599
X[, 52]	0.078
X[, 61]	0.0965
X[, 90]	0.0998
X[, 167]	0.0815

Tabla 7. Valores p menores a 0.05.

Posición	Valor p
X[, 16]	0.00383
X[, 54]	0.0384
X[, 89]	0.0289
X[, 95]	0.0415
X[, 114]	0.00334
X[, 121]	0.0426
X[, 176]	0.00383
X[, 183]	0.0438

3.4.2 Regresión Múltiple

Posteriormente, se aplicó una regresión múltiple con las posiciones que obtuvieron los valores p deseados o cercanos al deseado con la comparación de uno a uno en la regresión simple.

Z. ($p \leq 0.1$) = X[,4]+ X[,52]+ X[,61]+ X[,90]+ X[,167]

Z* ($p \leq 0.05$) = X[,16]+ X[,54]+ X[,89]+ X[,95]+ X[,114]+ X[,121]+ X[,176]+ X[,183]

En un primer ensayo se aplicó la regresión múltiple tomando en cuenta solo los valores con significancia de Z*, es decir, aquellos que cumplieron con el valor $p<0.05$. Las posiciones que cumplen con el valor p requerido son:

Z* ($p \leq 0.05$) = X[,176]

Z** ($p \leq 0.01$) = X[,16]+ X[,114]

En un segundo experimento se utilizaron los valores Z. y Z* para la regresión múltiple, obteniendo los siguientes resultados:

Z* ($p \leq 0.05$) = X[,167]+ X[,176]

Z** ($p \leq 0.01$) = X[,16]+ X[,114]

A estos 4 valores que cumplían con el valor de significancia necesario se les aplico nuevamente regresión múltiple. Los resultados de esta segunda regresión, indican que 3 vectores resultaron con significancias menores al 0.05, mientras que un vector se quedó cerca de 0.05.

Z. ($p \leq 0.1$) = X[,167]

Z** ($p \leq 0.01$) = X[,16]+ X[,114]+ X[,176]

También realizamos una regresión múltiple utilizando todos los vectores, un total de 242 vectores al mismo tiempo. Los resultados de esta regresión son los siguientes:

Z. ($p \leq 0.1$) = X[,13]+ X[,144]+ X[,234]

Z* ($p \leq 0.05$) = X[,129]+ X[,167]+ X[,169]+ X[,183]

Con los valores que resultaron como significantes y los cercanos al valor significante volvimos a aplicar regresión múltiple a estos 7 vectores. Los resultados nos indican que los vectores que cumplen con el valor de significancia son:

Z. ($p \leq 0.1$) = X[,167]

Z* ($p \leq 0.05$) = X[,183]

Si volvemos a aplicar regresión múltiple a estos dos valores con significancia vuelven a aparecer los mismos valores con la misma significancia.

Todos los resultados anteriores están resumidos a continuación:

1. Usando los valores de Z* de la regresión simple
   Z*-- X[,176] y Z**--X[,16]+ X[,114]
2. Usando los valores de Z. y Z* de la regresión simple
   Z* -- X[,167]+ X[,176] y Z** -- X[,16]+ X[,114]
3. Usando los valores de Z. y Z* de la regresión simple (segunda regresión)
   Z. -- X[,167] y Z** -- X[,16]+ X[,114]+ X[,176]
4. Usando todos los valores al mismo tiempo
   Z. -- X[,13]+ X[,144]+ X[,234] y Z* -- X[,129]+ X[,167]+ X[,169]+ X[,183]
5. Usando todos los valores al mismo tiempo (segunda regresión)
   Z. -- X[,167] y Z* -- X[,183]
6. Usando todos los valores al mismo tiempo (tercera regresión)
   Z. -- X[,167] y Z* -- X[,183]

De todos estos resultados tomaremos el resultado número 2, en el cual salieron beneficiados los vectores 16, 114, 167 y 176, estos vectores son solo de la región de variabilidad, pero si los trasladamos a la secuencia completa de ADN estos vectores representarían las posiciones **16185, 16283, 16336 y 16345**.

3.5 Prueba de Factores de Riesgo

3.5.1 Razón de Probabilidades (OR)

Para realizar la prueba de razón de probabilidades se utilizaron los valores que aparecen en la Tabla 7 que son todos los vectores que superaron el valor p menor a 0.05. Los resultados los podemos encontrar en la **¡Error! No se encuentra el origen de la referencia.**. La primera columna presenta l a variable (componente de composición corporal). La segunda columna presenta la razón de probabilidad y su IC del 95%. La tercera columna presenta el valor p, y la cuarta columna presenta la declaración del Factor de Riesgo. Además, en la Figura 11 podemos encontrar una gráfica con los resultados de la razón de probabilidades.

Tabla 8. Definición de factores de riesgo de Diabetes. Una variable fue declarada como Factor de Riesgo si su valor de p era 0.05, su OR era diferente a l y su OR IC 95% no incluía 1.

Variable	Razón de Probabilidad (95% IC)	Valor p	Factor de Riesgo
X[,16]	4.301 (1.759 – 12.355)	0.00215	Si, asociado con mayores probabilidades de Diabetes.
X[,54]	1.189 (0.485 – 2.370)	0.65925	No
X[,89]	0.651 (0.375 – 1.074)	0.10374	No
X[,95]	0.933 (0.662 – 1.333)	0.69370	No
X[,114]	7.385 (2.014 – 41.343)	0.00641	Si, asociado con mayores probabilidades de Diabetes.
X[,121]	1.560 (0.593 – 7.135)	0.44478	No
X[,176]	0.580 (0.371 – 0.906)	0.01598	Si, asociado con mayores probabilidades de Diabetes.
X[,183]	0.796 (0.567 – 1.141)	0.19839	No

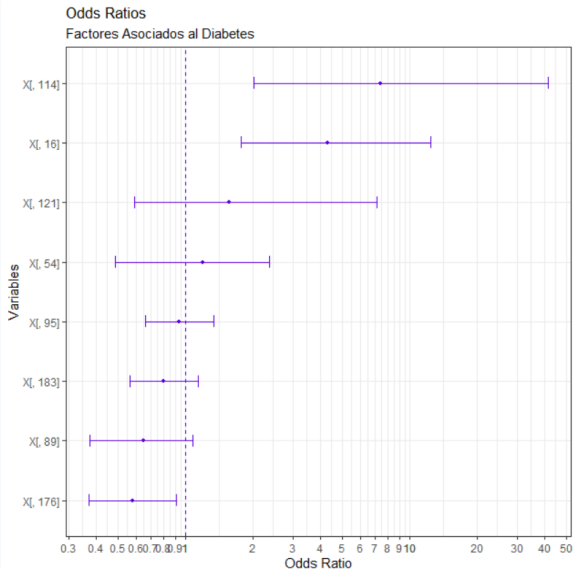


Figura 11. Grafica de Odds Ratio.

Posteriormente se realizó una corrección del valor p, utilizando varios métodos como el método de Bonferroni, FDR, Hochberg y BH. A continuación, se muestran los resultados para la regresión simple y múltiple.

Regresión Simple

> valores_p <- c(.00383, .0384, .0289, .0415, .00334, .0426, .00383, .0438)

> p.adjust(valores_p,method="bonferroni")

0.03064 0.30720 0.23120 0.33200 0.02672 0.34080 0.03064 0.35040

> p.adjust(valores_p,method="fdr")

0.01021333 0.04380000 0.04380000 0.04380000 0.01021333 0.04380000 0.01021333 0.04380000

> p.adjust(valores_p,method="hochberg")

[0.02298 0.04380 0.04380 0.04380 0.02298 0.04380 0.02298 0.04380

> p.adjust(valores_p,method="BH")

0.01021333 0.04380000 0.04380000 0.04380000 0.01021333 0.04380000 0.01021333 0.04380000

Los valores p del siguiente vector son después de la regresión múltiple.

valores_p <- c(.00215, .65925, .10374, .69370, .00641, .44478, .01598, .19839)

p.adjust(valores_p,method="bonferroni")

0.01720 1.00000 0.82992 1.00000 0.05128 1.00000 0.12784 1.00000

p.adjust(valores_p,method="hochberg")

0.01720 0.69370 0.51870 0.69370 0.04487 0.69370 0.09588 0.69370

p.adjust(valores_p,method="BH")

0.01720000 0.69370000 0.20748000 0.69370000 0.02564000 0.59304000 0.04261333 0.31742400

p.adjust(valores_p,method="fdr")

0.01720000 0.69370000 0.20748000 0.69370000 0.02564000 0.59304000 0.04261333 0.31742400

3.6 Proporciones Polimórficas

Analizar las proporciones polimórficas permite encontrar la variabilidad en cada posición de las secuencias, al igual que visualizar regiones con mayor o menor variabilidad. Obtener las frecuencias alélicas para cada posición es el primer paso para este análisis, y las gráficas de proporciones de obtienen usando la frecuencia de alelo menor (MAF). A continuación, se presentan las gráficas resultantes de las proporciones polimórficas los genomas completos y de la región variante.

3.6.1 Proporciones polimórficas de los genomas completos.

La Tabla 9 presenta en las columnas, la cantidad de posiciones base-par que tuvieron una frecuencia de alelo menor en cada uno de 5 rangos. Para la columna 2, por ejemplo, al analizar los genomas completos, incluyendo todos los individuos, la cantidad de 11,773 posiciones base-par tuvieron MAF en el rango [0 – 0.1). Considerando sólo los individuos enfermos, resultaron 12,307 posiciones base-par con una MAF en el mismo rango anterior. Y considerando sólo los individuos sanos, resultaron 11,148 bases-par en el mismo rango.

Tabla 9. Resultados de Proporciones Polimórficas de los genomas completos.

Frecuencia del MAF	**[0-0.1)**	**[0.1-0.2)**	**[0.2-0.3)**	**[0.3-0.4)**	**[0.4-0.5)**
Todos	11773	3903	645	48	14
Enfermos	12307	3441	574	48	14
Sanos	11148	3044	1599	464	129

La Figura 12 presenta una gráfica de las proporciones polimórficas de los genomas completos. En la gráfica podemos observar una diferencia clara de frecuencia MAF entre sanos y enfermos para los rangos [0.2 – 0.3) y [0.3 – 0.4). En este caso en particular, existen más posiciones base-par, en los rangos anteriores, para los individuos sanos. Esto podría ser un indicativo de que la diabetes Mellitus tipo 2 provoca una disminución en la variabilidad alélica en el genoma completo de poblaciones que la padecen.

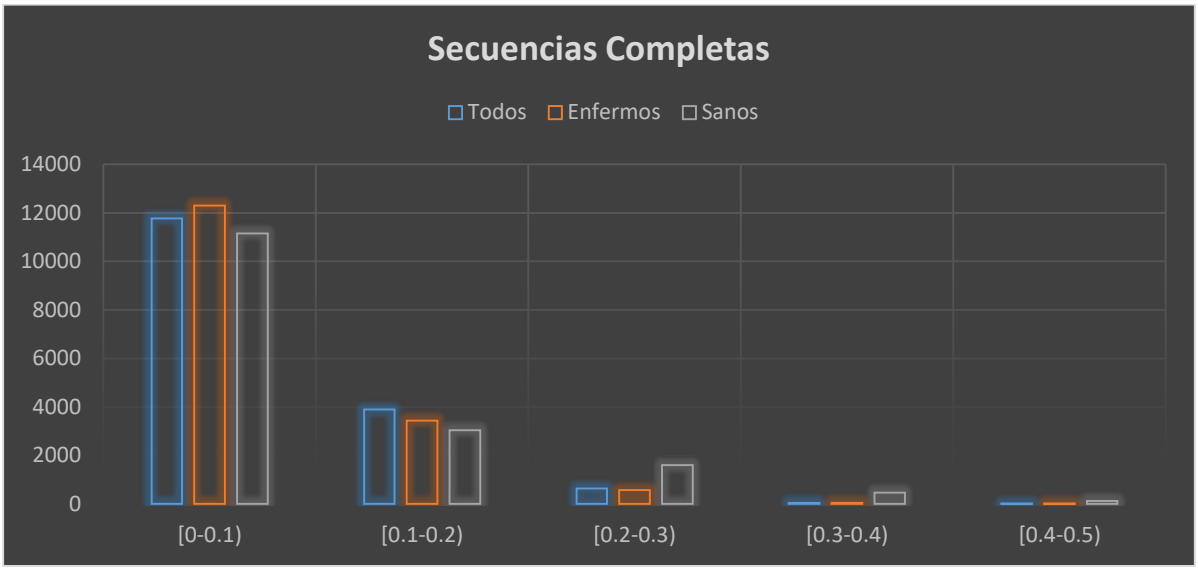


Figura 12. Grafica de proporciones polimórficas de los genomas completas.

3.6.2 Proporciones polimórficas para la región variante.

La Tabla 10 presenta en las columnas, la cantidad de posiciones base-par que tuvieron una frecuencia de alelo menor en cada uno de 5 rangos. Para la columna 2, por ejemplo, al analizar sólo la región que fue seleccionada como variante, incluyendo todos los individuos, la cantidad de 233 posiciones base-par tuvieron MAF en el rango [0 – 0.1). Considerando sólo los individuos enfermos, resultaron 234 posiciones base-par con una MAF en el mismo rango anterior. Y considerando sólo los individuos sanos, resultaron 233 bases-par en el mismo rango.

Tabla 10. Resultados de Proporciones Polimórficas de la región variante.

Grupos	[0-0.1)	[0.1-0.2)	[0.2-0.3)	[0.3-0.4)	[0.4-0.5)
Todos	233	4	1	1	2
Enfermos	234	4	1	0	2
Sanos	233	1	3	2	2

La Figura 13 presenta una gráfica de las proporciones polimórficas de la región variante.

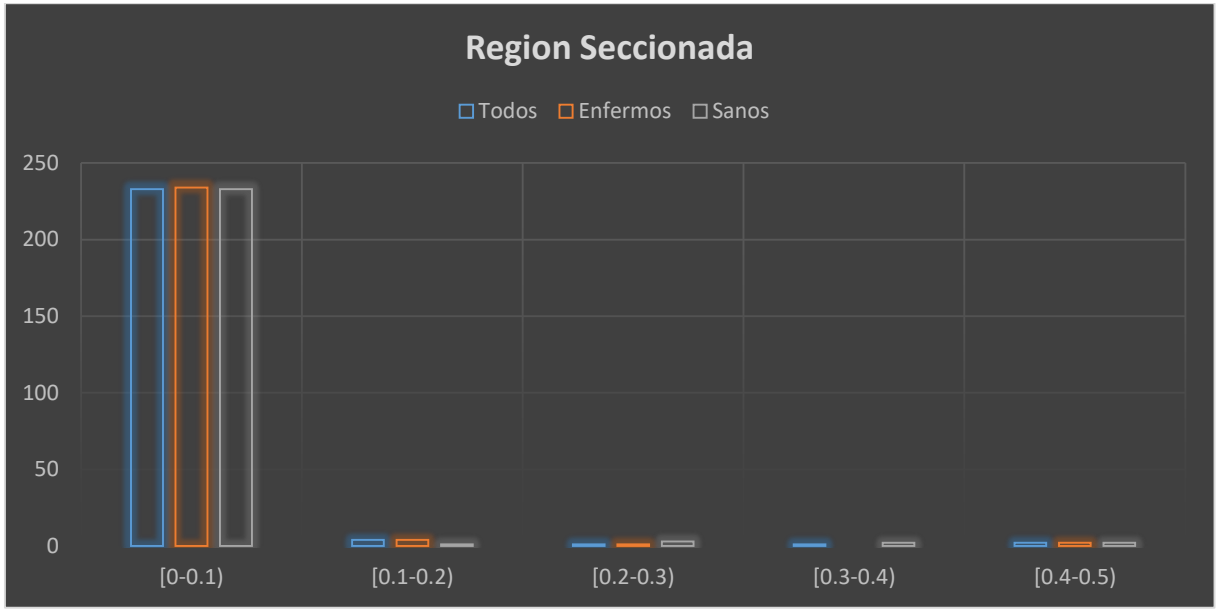


Figura 13. Grafica de proporciones polimórficas de la Región Seccionada.

3.7 Grafica de Manhattan

Un diagrama de Manhattan es un tipo específico de diagrama de dispersión ampliamente utilizado en genómica para estudiar los resultados de GWAS (Genome Wide Association Study). Cada punto representa una variante genética. El eje X muestra su posición en un cromosoma, el eje Y indica cuánto está asociado con un rasgo. A partir de los resultados de la regresión simple y múltiple se realizó la Figura 14 que es un gráfico de Manhattan.

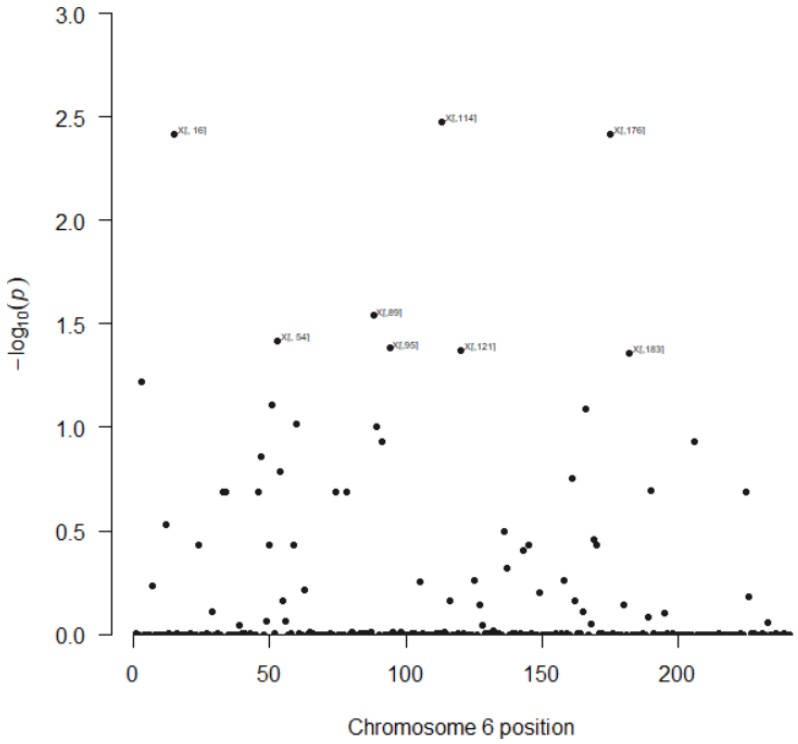


Figura 14. Gráfico de Manhattan.

CAPITULO IV ANÁLISIS DE LOS RESULTADOS

Los 510 genomas mitocondriales fueron alineados mediante el algoritmo ClustalW implementado en la herramienta MAFFT [35, 36]. Luego se inspeccionó visualmente la alineación con el software MEGA X. [60]. Se identificó una región de 241 pares de bases de largo que claramente presentaba mayor variabilidad que el resto del alineamiento. Esta región osciló entre la posición de pares de bases 16.170 y 16.410. Llamamos a esta región como R_v. Posteriormente, para estimar el grado de información proporcionada por Rv, se utiliza la entropía de Shannon [40] se calculó para toda la región; primero calculando la entropía para cada par de bases individual y luego sumando las entropías individuales en toda la región R_v. La entropía R_v resultante fue 16.83817. Posteriormente, se interrogó el resto del genoma tomando regiones contiguas del mismo tamaño que R_v (sin incluir R_v), y se calcularon estimaciones de entropía para cada región. En total, se encontraron 68 regiones (sin incluir R_v) con una entropía promedio de 1.64698. Se utilizó una prueba t estadística de igualdad de medias para verificar la significación estadística de la entropía de R_v, en comparación con el resto del genoma. La prueba dio como resultado un valor de $p = 2.2e^{-16}$, lo que indica que la entropía de R_v es significativamente mayor que la del resto del genoma. Luego seleccionamos R_v para realizar el resto del análisis.

El primer análisis en la región R_v fue visualizar la estructura de los datos buscando grupos de información que pudieran diferenciar entre individuos enfermos y sanos. Para ello se aplicó un Análisis de Componentes Principales [38]. Para realizar este análisis se utilizó el lenguaje estadístico R [61]. La Figura S2 (ver Figura complementaria S2) muestra una gráfica del Componente principal 1 (PC1) frente al Componente principal 2 (PC2). Del gráfico podemos observar que los datos parecen muy contradictorios, lo que significa que no hay una diferenciación

clara entre individuos enfermos y sanos. Por lo tanto, para realizar el análisis de asociación, sería necesario seleccionar una técnica capaz de analizar datos altamente mixtos.

Para encontrar la asociación de las posiciones de los pares de bases con la diabetes tipo 2, se utilizaron modelos de regresión logística. [42, 43]. El análisis se realizó en dos pasos. Primero, todos los individuos con diabetes tipo 2 fueron etiquetados con 0 y todos los individuos sanos fueron etiquetados con 1, para la variable dependiente del modelo. Luego se aplicó una regresión logística simple a cada par de bases individual dentro de Rv. Esto arrojó 8 posiciones estadísticamente significativas (valor de p <0,05). A continuación, se seleccionaron las 8 posiciones de pares de bases significativas y se aplicó una Regresión Logística Múltiple. Las posiciones de los pares de bases resultantes que fueron significativas a partir de la regresión múltiple se declararon asociadas con la diabetes tipo 2. La tabla 1 muestra los resultados. La primera columna muestra la posición del par de bases. La segunda columna contiene los valores p de la regresión simple. La tercera columna muestra los valores p resultantes de la regresión múltiple. La cuarta columna contiene el estado de asociación con DT2.

Tabla 11. Resultados de regresión logística simple y múltiple. La regresión simple y múltiple identificó los loci de posición del par de bases 16184, 16282 y 16344 como estadísticamente asociados con la diabetes tipo 2.

Posición Genómica (BP)	**Regresión Simple (valor-p)**	**Regresión Multiple (valor-p)**	**Asociado con DT2**
16,184	0.0038	0.0021	Si
16,222	0.0384	0.6592	No
16,257	0.0289	0.1037	No
16,263	0.0415	0.6937	No
16,282	0.0033	0.0064	Si
16,289	0.0426	0.4447	No
16,344	0.0038	0.0159	Si
16,351	0.0438	0.1983	No

Para definir si las posiciones de pares de bases asociadas son un factor de riesgo, se utilizan los criterios propuestos por Szumila, 2010 [46], que utiliza el Odds Ratio. Definimos una posición de par de bases como factor de riesgo si cumplía las siguientes tres condiciones: 1) el valor de p de la prueba estadística para el OR era ≤ 0.05; 2) el OR fue diferente de 1; y 3) el Intervalo de Confianza (IC) no incluyó el valor 1. Si un par de bases encontrado asociado cumple con estas tres condiciones y tiene OR > 1, entonces se declaró como un factor de riesgo asociado a una probabilidad alta. de diabetes tipo 2. Ahora bien, si la variable cumple las tres condiciones y tiene OR<1, entonces se declara como un factor de riesgo asociado a una pequeña probabilidad de padecer diabetes tipo 2. La tabla 2 muestra los resultados. La primera columna contiene la posición del par de bases. La segunda columna contiene el odds ratio y su IC del 95%. La tercera columna contiene el valor p. La cuarta columna contiene la declaración del factor de riesgo.

Tabla 12. Factores de riesgo de diabetes tipo 2. Posición genómica, valor P y odds ratio con intervalo de confianza del 95% para variantes asociadas con posibilidades altas y pequeñas de diabetes tipo 2.

Posición Base Par del Genoma Mitocondrial	**Odds Ratio (95% CI)**	**Valor-P**	**Factor de Riesgo**
16,184	4.301 (1.759 – 12.355)	0.00215	Sí, asociado con altas probabilidades de padecer DT2.
16,282	7.385 (2.014 – 41.343)	0.00641	Sí, asociado con altas probabilidades de padecer DT2.
16,344	0.580 (0.371 – 0.906)	0.01598	Sí, asociado con bajas probabilidades de padecer DT2.

Para localizar los genes relacionados con las posiciones de los pares de bases asociadas, se inspeccionó la anotación del genoma mitocondrial humano de la base de datos NCBI (//www.ncbi.nlm.nih.gov/). Se localizaron tres genes dentro de 3000 pares de bases de las posiciones asociadas. Estos genes son CYTB (también conocido como MT-CYB), que produce la

proteína Citocromo B y contribuye a la conversión de la energía de los alimentos en energía celular (Adenosín Trifosfato, ATP), el gen TRNP (también conocido como TRNP1), que es el ARNt de prolina y el gen TRNT (también conocido como MTTT), que es el ARNt de treonina. Luego, interrogamos a través de la herramienta en línea Genemania (https://genemania.org/) si estos genes interactúan [62]. muestra el gráfico resultante de interacciones entre genes. Podemos notar que Genemania no encontró el gen TRNT y que los genes CYBT y TRNT no muestran ninguna interacción conocida. Las Figuras complementarias S3, S4 y S5 presentan los informes de Genemania para los tres genes, el gen CYTB y el gen TRNP, respectivamente, las cuales se pueden encontrar en el Anexo C.

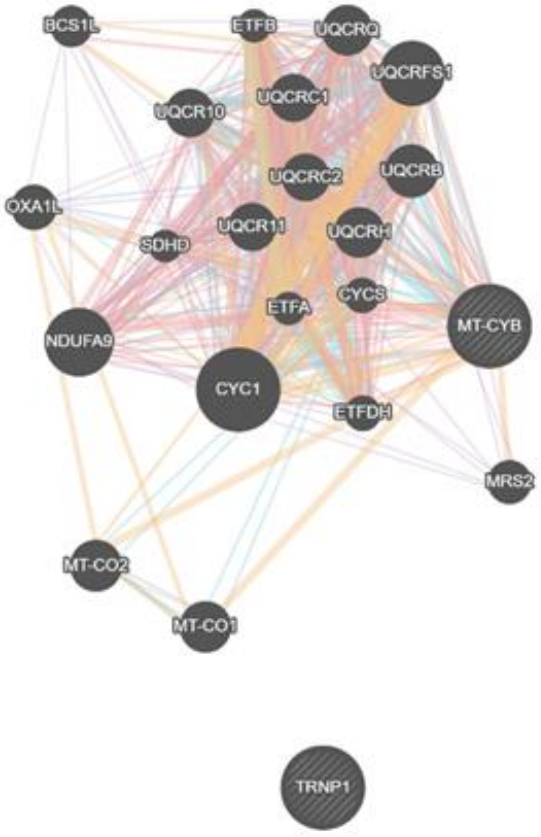


Figura 15. Gráfico de interacciones de genemania para los genes MT-CYB, TRNP1 y TRNT. No se encontró el gen TRNT. Y los genes MT-CYB y TRNP1 no muestran interacciones entre ellos.

El alineamiento de las 510 secuencias iniciales del genoma permitió visualizar y seleccionar una región de 241 pares de bases (R_v) con claramente más variabilidad que el resto del genoma. Sin embargo, la prueba t de una muestra implementada para comparar la entropía de R_v con la media del vector de entropías obtenido de todas las regiones del mismo tamaño de R_v permitió demostrar de manera sistémica que el Rv seleccionado es estadísticamente el más variante, y Como consecuencia, la región más informativa del genoma mitocondrial de la diabetes tipo 2. Entonces, el análisis de entropía ofrece un criterio adecuado para seleccionar regiones variantes para análisis específicos.

Después de una regresión logística múltiple, tres posiciones se asociaron con la diabetes tipo 2. Las posiciones asociadas fueron 16184, 16282 y 16344 con un valor p de 0.0021, 0.0064 y 0.0159, respectivamente. Luego, a partir de los tres genes encontrados dentro de los 3000 pares de bases de las posiciones asociadas, Ke Li, et al., 2020 informó previamente que TRNT estaba asociado con la heredabilidad materna de la diabetes tipo 2 en familias chinas [63]. En otro estudio realizado por Momiyama, et al., 2003 el mismo gen también se asoció, al igual que en nuestro estudio, con la posición genómica 16184; y declarado como una de las causas de hipertrofia ventricular izquierda en pacientes con diabetes tipo 2 en familias japonesas [64]. Sin embargo, los genes CYTB y TRNP no se habían asociado previamente con la diabetes tipo 2.

CAPÍTULO V CONCLUSIONES

En este estudio de asociación, se analizaron 510 genomas mitocondriales humanos, de los cuales 437 fueron de individuos con diabetes tipo 2 y 73 de individuos sanos. La selección de una región de 241 pares de bases que presentó una clara variabilidad en el alineamiento de los genomas completos, nos permitió realizar un estudio centrado en la región variante del genoma. Tres posiciones base-par resultaron asociadas después de un análisis de regresión logística. Y después de aplicar el criterio de Razón Odd tres posiciones base-par asociadas fueron declaradas factores de riesgo para diabetes tipo 2. Mediante una inspección de la anotación del genoma mitocondrial humano se localizaron, dentro de un rango de 3kb de las posiciones asociadas, tres genes. Estos genes fueron: el gen CYTB, que produce la proteína del citocromo B y contribuye a la producción de ATP; el gen TRNP, que es el ARNt de la prolina; y el gen TRNT, que es el ARNt de la treonina. Estudios previos por otros grupos de investigación reportaron que TRNT está asociado con DT2, mientras que con este trabajo de tesis son declarado por primera vez como asociados a DT2, los genes CYTB y TRNP. El análisis de predicción de interacción realizado con Genemania mostró que existe una interacción entre los genes CYTB y TRNP. Finalmente, este estudio de asociación proporciona nueva evidencia de asociación al proponer dos variantes novedosas estadísticamente asociadas con la diabetes tipo 2 que cumplen con los criterios de ser un factor de riesgo para la diabetes tipo 2.

REFERENCIAS

1. Chen, L., D.J. Magliano, and P.Z. Zimmet, *The worldwide epidemiology of type 2 diabetes mellitus--present and future perspectives.* Nat Rev Endocrinol, 2011. **8**(4): p. 228-36.
2. O´Farrill, L.C.L., Cuervo, A. M. d. S., Ferrer, R. L., & Valdés, M. T. L, *Interacción genoma-ambiente en la diabetes mellitus tipo 2.* Acta Médica del Centro, 2018. **12**(4).
3. Ayo Toye, D.G., *Source: genetics and functional genomics of type 2 diabetes mellitus.* Genome Biol, 2003. **4**: p. 241.
4. health, n.i.o., *FACT SHEET - Type 2 Diabetes.* 2010.
5. Tsai, F.J., et al., *A genome-wide association study identifies susceptibility variants for type 2 diabetes in Han Chinese.* PLoS Genet, 2010. **6**(2): p. e1000847.
6. Scott, L.J., et al., *A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants.* Science, 2007. **316**(5829): p. 1341-5.
7. Saxena, R., et al., *Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels.* Science, 2007. **316**(5829): p. 1331-6.
8. Sladek, R., et al., *A genome-wide association study identifies novel risk loci for type 2 diabetes.* Nature, 2007. **445**(7130): p. 881-885.
9. Zeggini, E., et al., *Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes.* Science, 2007. **316**(5829): p. 1336-41.
10. Burton, P.R., et al., *Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls.* Nature, 2007. **447**(7145): p. 661-678.
11. Zeggini, E., et al., *Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes.* Nature Genetics, 2008. **40**(5): p. 638-645.
12. Gudmundsson, J., et al., *Two variants on chromosome 17 confer prostate cancer risk, and the one in TCF2 protects against type 2 diabetes.* Nature Genetics, 2007. **39**(8): p. 977-983.
13. Saxena, R., et al., *Large-scale gene-centric meta-analysis across 39 studies identifies type 2 diabetes loci.* Am J Hum Genet, 2012. **90**(3): p. 410-25.
14. Baum, A.E., et al., *A genome-wide association study implicates diacylglycerol kinase eta (DGKH) and several other genes in the etiology of bipolar disorder.* Molecular Psychiatry, 2008. **13**(2): p. 197-207.
15. Galvan, A., et al., *Genome-wide association study in discordant sibships identifies multiple inherited susceptibility alleles linked to lung cancer.* Carcinogenesis, 2009. **31**(3): p. 462-465.
16. Forstbauer, L.M., et al., *Genome-wide pooling approach identifies SPATA5 as a new susceptibility locus for alopecia areata.* European Journal of Human Genetics, 2012. **20**(3): p. 326-332.
17. Wong, L.P., et al., *Deep whole-genome sequencing of 100 southeast Asian Malays.* Am J Hum Genet, 2013. **92**(1): p. 52-66.
18. Kong, A., et al., *Parental origin of sequence variants associated with complex diseases.* Nature, 2009. **462**(7275): p. 868-874.
19. Voight, B.F., et al., *Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis.* Nature Genetics, 2010. **42**(7): p. 579-589.
20. Dupuis, J., et al., *New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk.* Nature Genetics, 2010. **42**(2): p. 105-116.

21. Qi, L., et al., *Genetic variants at 2q24 are associated with susceptibility to type 2 diabetes.* Human Molecular Genetics, 2010. **19**(13): p. 2706-2715.
22. Yamauchi, T., et al., *A genome-wide association study in the Japanese population identifies susceptibility loci for type 2 diabetes at UBE2E2 and C2CD4A-C2CD4B.* Nature Genetics, 2010. **42**(10): p. 864-868.
23. Shu, X.O., et al., *Identification of new genetic risk variants for type 2 diabetes.* PLoS Genet, 2010. **6**(9): p. e1001127.
24. Kooner, J.S., et al., *Genome-wide association study in individuals of South Asian ancestry identifies six new type 2 diabetes susceptibility loci.* Nature Genetics, 2011. **43**(10): p. 984-989.
25. Cho, Y.S., et al., *Meta-analysis of genome-wide association studies identifies eight new loci for type 2 diabetes in east Asians.* Nature Genetics, 2012. **44**(1): p. 67-72.
26. Rossana De Lorenzi, C.G., *Towards the first recombiant drug.* ELLS – European Learning Laboratory for the Life Sciences.
27. J.C. Wiebe, A.M.W., F.J. Novoa Mogollón, *Genética de la diabetes mellitus.* Revista Nefrología, 2011. **2**(1): p. 1-119.
28. M.J. Picón, F.J.T., *Factores genéticos frente a factores ambientales en el desarrollo de la diabetes tipo 2.* Avances en Diabetología, 2010. **26**: p. 268-269
29. Esparza-Castro D, A.-A.F., Merelo-Arias CA,Cruz M, Valladares-Salgado A, *scaneo genómico completo en diabetes tipo 2 y su aplicación clínica.* Revista Medica Instituto Mexicano Seguro Social, 2015. **53**(5): p. 592-599.
30. Mauricio Hernández-Ávila, J.P.G., Nancy Reynoso-Noverón., *Diabetes mellitus en México. El estado de la epidemia.* Salud pública Méxicana, 2013. **55**.
31. García-Chapa, E.G., et al., *Genetic Epidemiology of Type 2 Diabetes in Mexican Mestizos.* BioMed research international, 2017. **2017**: p. 3937893-3937893.
32. Cruz, M., et al., *Candidate gene association study conditioning on individual ancestry in patients with type 2 diabetes and metabolic syndrome from Mexico City.* Diabetes Metab Res Rev, 2010. **26**(4): p. 261-70.
33. Gamboa-Meléndez, M.A., et al., *Contribution of common genetic variation to the risk of type 2 diabetes in the Mexican Mestizo population.* Diabetes, 2012. **61**(12): p. 3314-21.
34. Katoh, R., Yamada *Servicio en línea MAFFT: alineación de secuencias múltiples, elección y visualización interactivas de secuencias*. 2019 [cited 2020 junio]; Available from: https://mafft.cbrc.jp/alignment/software/algorithms/algorithms.html.
35. Kuraku, S., et al., *aLeaves facilitates on-demand exploration of metazoan gene family trees on MAFFT sequence alignment server with enhanced interactivity.* Nucleic Acids Research, 2013. **41**(W1): p. W22-W28.
36. Katoh, K., J. Rozewicki, and K.D. Yamada, *MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization.* Briefings in Bioinformatics, 2017. **20**(4): p. 1160-1166.
37. Konishi, T., et al., *Principal Component Analysis applied directly to Sequence Matrix.* Scientific Reports, 2019. **9**(1): p. 19297.
38. Mateos-Valenzuela, A.G., et al., *Risk factors and association of body composition components for lumbar disc herniation in Northwest, Mexico.* Scientific Reports, 2020. **10**(1): p. 18479.
39. Rodrigo, J.A. *Análisis de Componentes Principales (Principal Component Analysis, PCA) y t-SNE*. 2017 [cited 2020 Junio]; Available from: https://rpubs.com/Joaquin_AR/287787.

40. Delgado, A., A. Huamani, and B. Brillitt. *Applying Shannon Entropy to Analise Health System Level by departments in Peru*. in *2018 IEEE XXV International Conference on Electronics, Electrical Engineering and Computing (INTERCON)*. 2018.
41. *Shannon Diversity Index - The Diversity Function in R - ProgrammingR*. 2024.
42. Limeres, C.C. *REGRESIÓN LINEAL SIMPLE*. 2011 [cited 2022 19/Enero]; Available from: http://eio.usc.es/eipc1/BASE/BASEMASTER/FORMULARIOS-PHP-DPTO/MATERIALES/Mat_50140116_Regr_%20simple_2011_12.pdf.
43. Kavitha, S., S. Varuna, and R. Ramya. *A comparative analysis on linear regression and support vector regression*. in *2016 Online International Conference on Green Engineering and Technologies (IC-GET)*. 2016.
44. Zhang, Z., et al. *Multiple Linear Regression for High Efficiency Video Intra Coding*. in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2019.
45. *RPubs - Comparaciones múltiples: corrección de p-value y FDR*. 2024.
46. Szumilas, M., *Explaining odds ratios.* Journal of the Canadian Academy of Child and Adolescent Psychiatry = Journal de l'Academie canadienne de psychiatrie de l'enfant et de l'adolescent, 2010. **19**(3): p. 227-229.
47. Sladek, R., et al., *A genome-wide association study identifies novel risk loci for type 2 diabetes.* Nature, 2007. **445**(7130): p. 881-5.
48. Steinthorsdottir, V., et al., *A variant in CDKAL1 influences insulin response and risk of type 2 diabetes.* Nat Genet, 2007. **39**(6): p. 770-5.
49. Bouatia-Naji, N., et al., *A polymorphism within the G6PC2 gene is associated with fasting plasma glucose levels.* Science, 2008. **320**(5879): p. 1085-8.
50. Voight, B.F., et al., *Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis.* Nat Genet, 2010. **42**(7): p. 579-89.
51. Dupuis, J., et al., *New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk.* Nat Genet, 2010. **42**(2): p. 105-16.
52. So, H.C., et al., *Evaluating the heritability explained by known susceptibility variants: a survey of ten complex diseases.* Genet Epidemiol, 2011. **35**(5): p. 310-7.
53. Yasuda, K., et al., *Variants in KCNQ1 are associated with susceptibility to type 2 diabetes mellitus.* Nat Genet, 2008. **40**(9): p. 1092-7.
54. Unoki, H., et al., *SNPs in KCNQ1 are associated with susceptibility to type 2 diabetes in East Asian and European populations.* Nat Genet, 2008. **40**(9): p. 1098-102.
55. Yamauchi, T., et al., *A genome-wide association study in the Japanese population identifies susceptibility loci for type 2 diabetes at UBE2E2 and C2CD4A-C2CD4B.* Nat Genet, 2010. **42**(10): p. 864-8.
56. Kooner, J.S., et al., *Genome-wide association study in individuals of South Asian ancestry identifies six new type 2 diabetes susceptibility loci.* Nat Genet, 2011. **43**(10): p. 984-9.
57. Lettre, G., et al., *Genome-wide association study of coronary heart disease and its risk factors in 8,090 African Americans: the NHLBI CARe Project.* PLoS Genet, 2011. **7**(2): p. e1001300.
58. projects, C.t.W., *Sequence logo - Wikipedia*. 2023.
59. *How to determine the height/bits in a sequence logo?* 2024.
60. Kumar, S., et al., *MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms.* Mol Biol Evol, 2018. **35**(6): p. 1547-1549.
61. Team, R.C., *R: A Language and Environment for Statistical Computing*. 2020.

62. Warde-Farley, D., et al., *The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function.* Nucleic Acids Research, 2010. **38**(suppl_2): p. W214-W220.
63. Li, K., et al., *Maternally Inherited Diabetes Mellitus Associated with a Novel m.15897G>A Mutation in Mitochondrial tRNA(Thr) Gene.* J Diabetes Res, 2020. **2020**: p. 2057187.
64. Momiyama, Y., et al., *A mitochondrial DNA variant associated with left ventricular hypertrophy in diabetes.* Biochem Biophys Res Commun, 2003. **312**(3): p. 858-64.

ANEXOS

ANEXO A

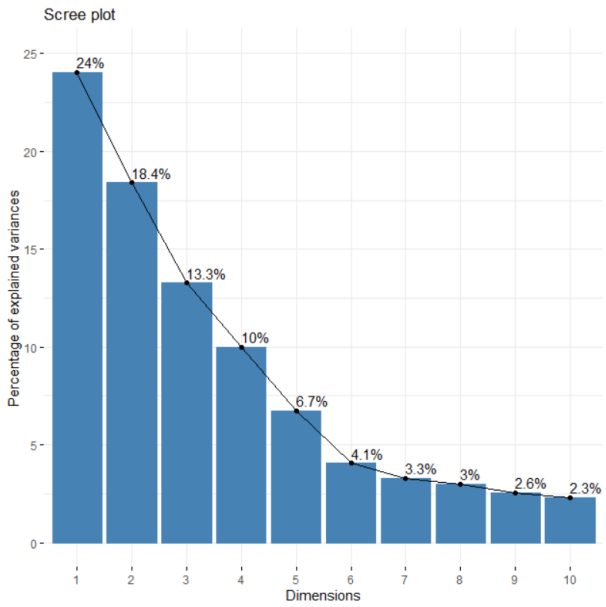
La lista de los identificadores (ID) de los genomas mitocondriales descargados de la base de datos de información sobre Diabetes Mellitus tipo 2: *http://www.type2diabetesgenetics.org/*, será anexado como un archivo adjunto de Excel.

ANEXO B

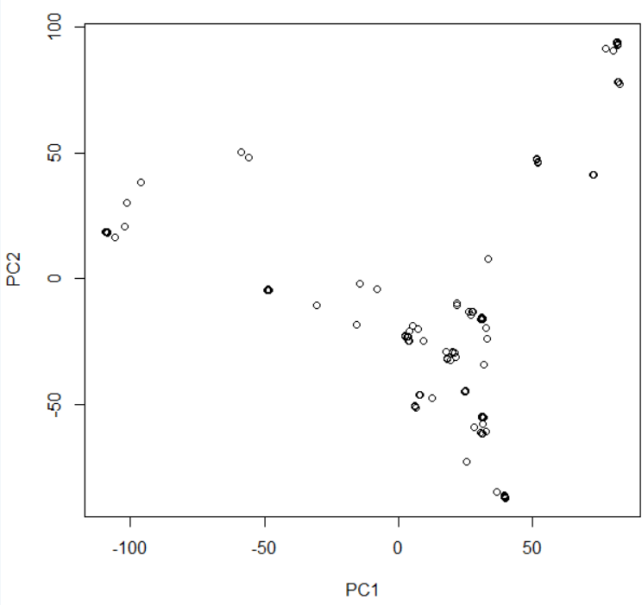
PCA con todas las Secuencias antes de Alinearse

El proceso de aplicar PCA fue realizado a través de lenguaje R. Esta actividad fue con todas las secuencias al mismo tiempo antes de ser alineadas.

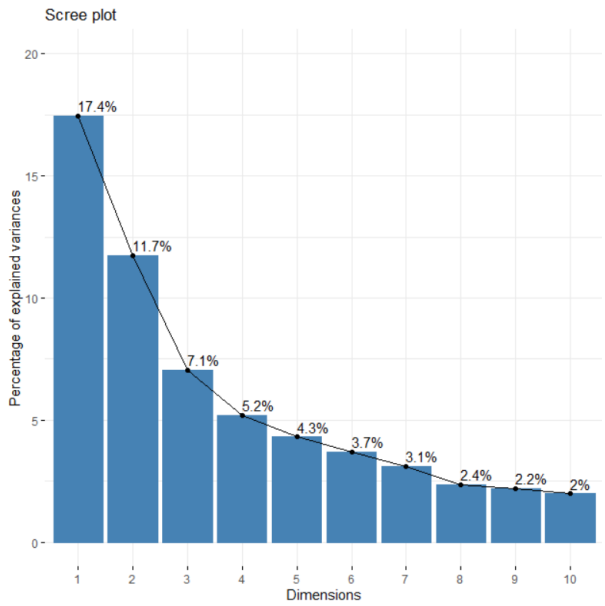
Para una mejor visualización de nuestros componentes principales los graficaremos para ver la aportación de cada uno de ellos, eso lo realizamos con el siguiente comando.



Como podemos ver de la figura anterior solo nos muestra los primeros 10 componentes principales, donde la primera y segunda componente son los que más varianza tienen. Ahora estos dos componentes los graficamos en una gráfica de dispersión para visualizar si existen cúmulos o clusters donde podamos apreciar que existen grupos marcados donde la mayor cantidad de datos proviene de personas sanas o con diabetes. Esta nueva grafica la realizamos con la siguiente función.

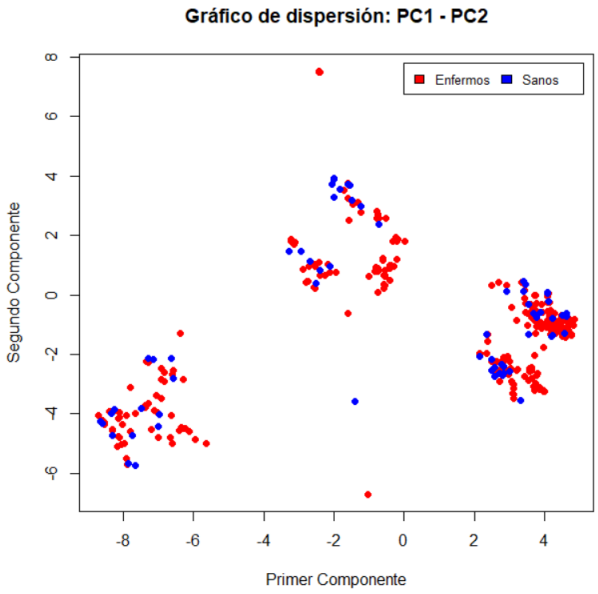


PCA a Secuencias Alineadas Completas

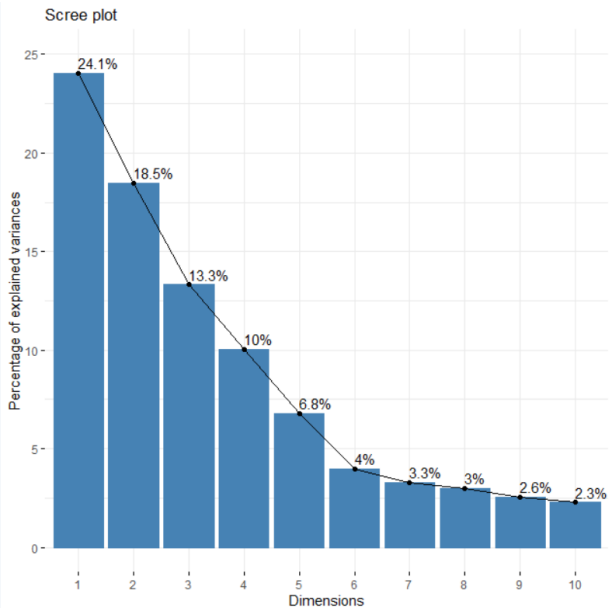


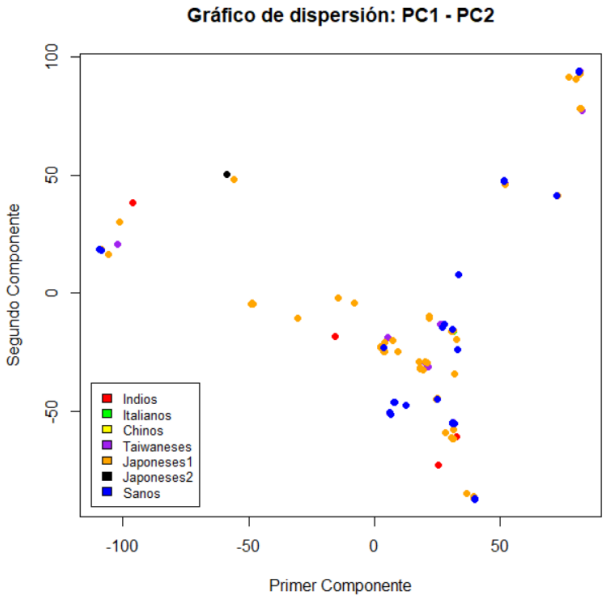
El etiquetado de los valores con las etiquetas de enfermos y sanos no se utilizó ya que causa mucha confusión por lo tanto solo se usará colores para distinguir entre ambas clases.

En esta grafica elimine los límites de los ejes x y y, porque si no cambiaba el tamaño de la gráfica y se deformaba la misma. Para corregir eso solo se tendría que revisar previamente los márgenes de la gráfica y colocar bien los limites.



Clasificación de Etnias con PCA usando todo el Genoma (secuencias sin alinear)





Clasificación de Etnias con PCA usando todo el Genoma (secuencias alineadas)

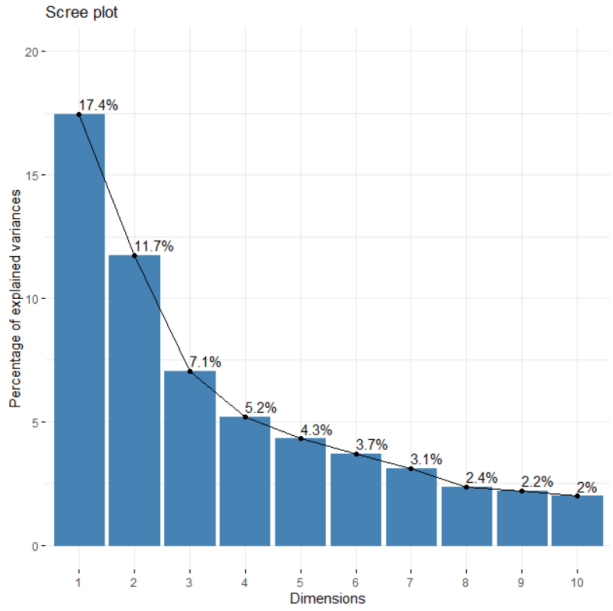
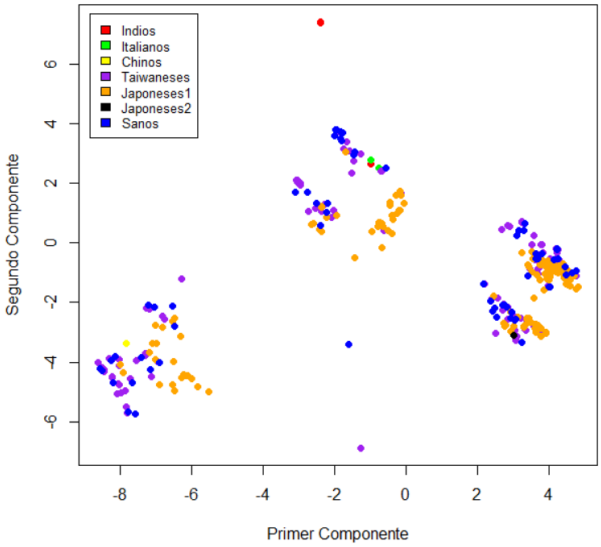
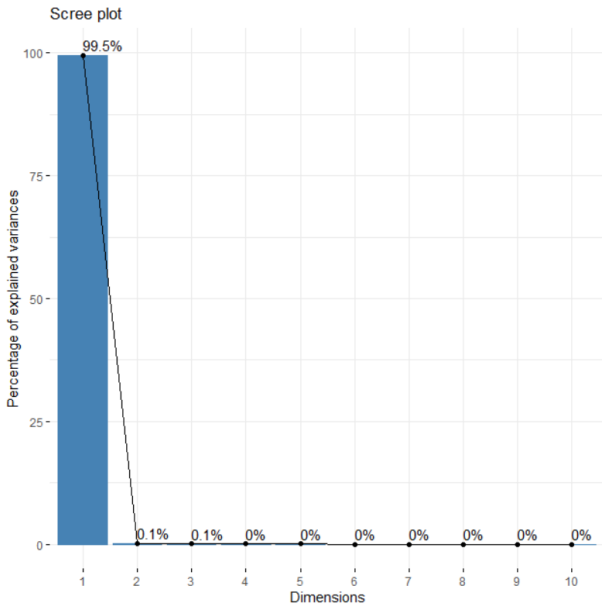
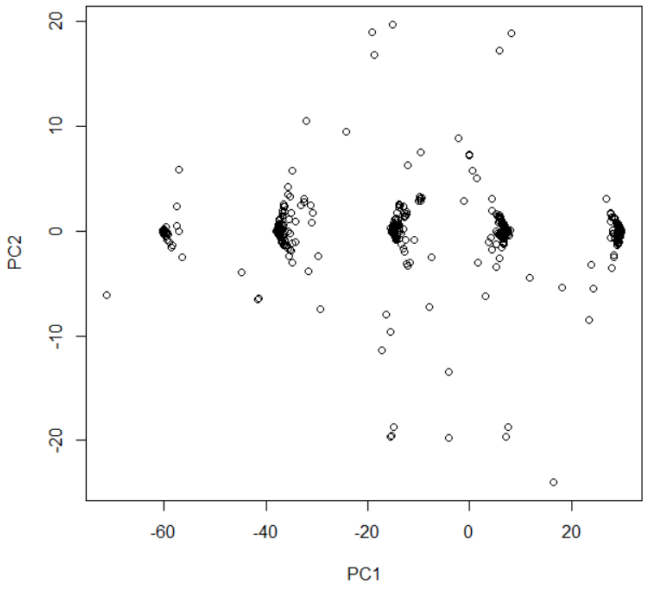


Gráfico de dispersión: PC1 - PC2

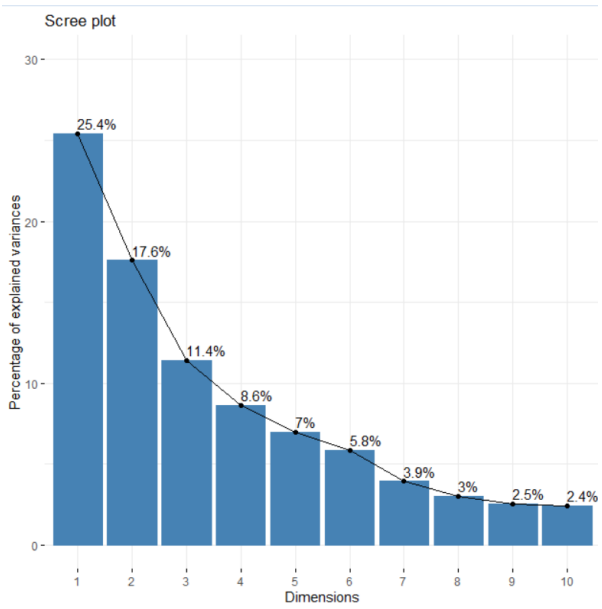


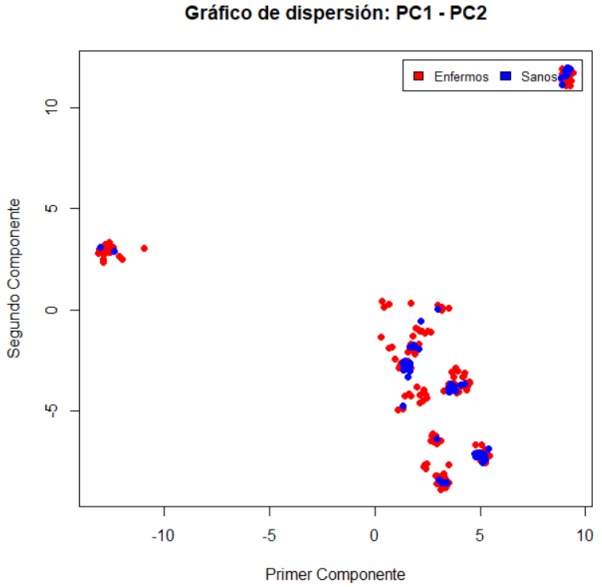
Graficas de PCA con Datos Transpuestos





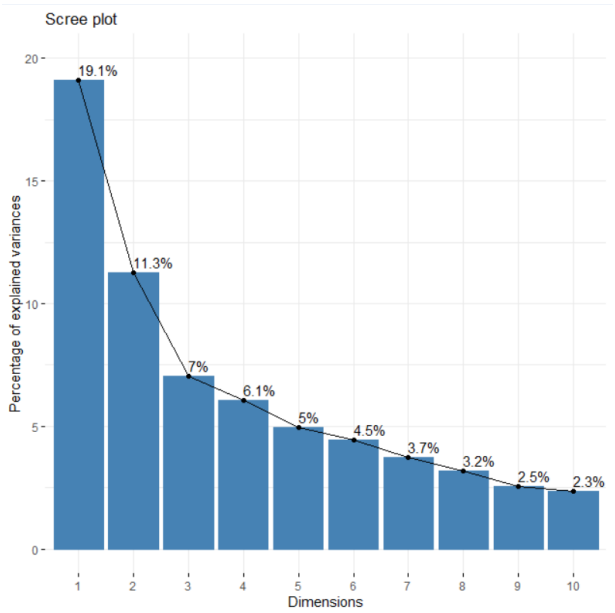
PCA en Secuencias Combinadas sin Alinear

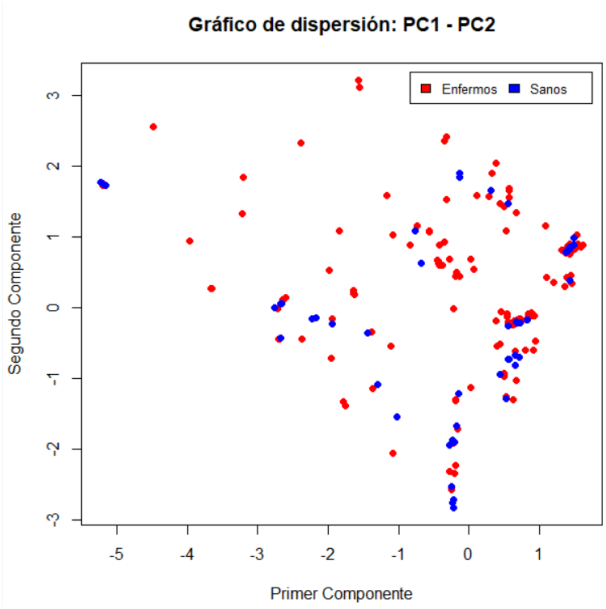




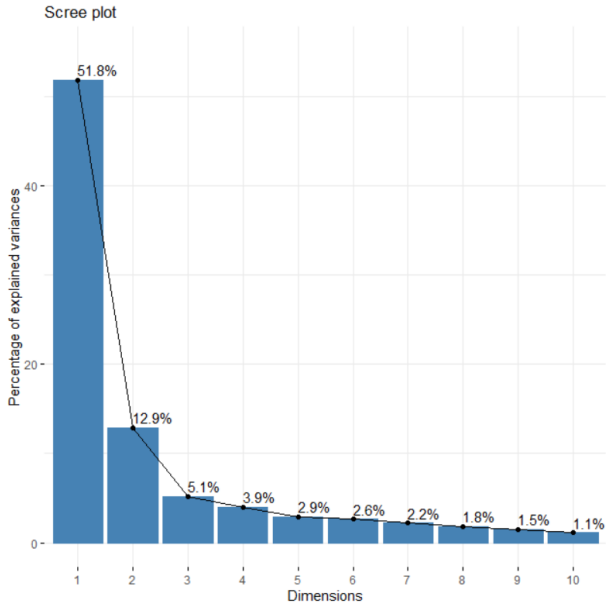
PCA en Secuencias Seccionadas Alineadas

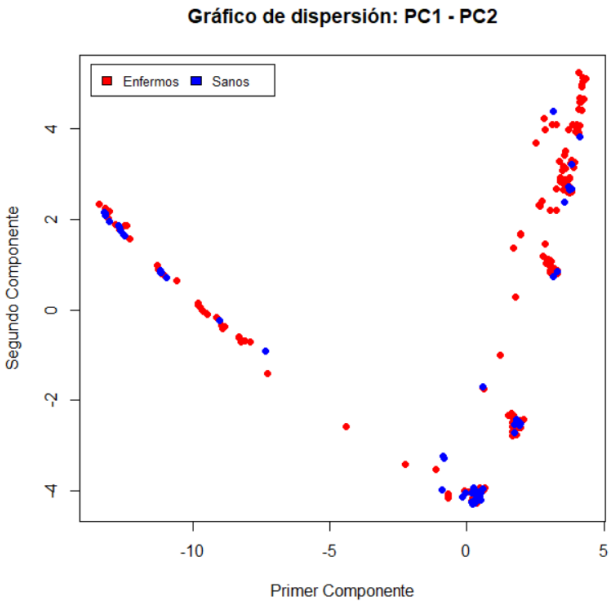
Estas secuencias primero fueron alineadas y después seccionadas. En los siguientes casos primero serán seccionadas y después alineadas con distintos métodos.



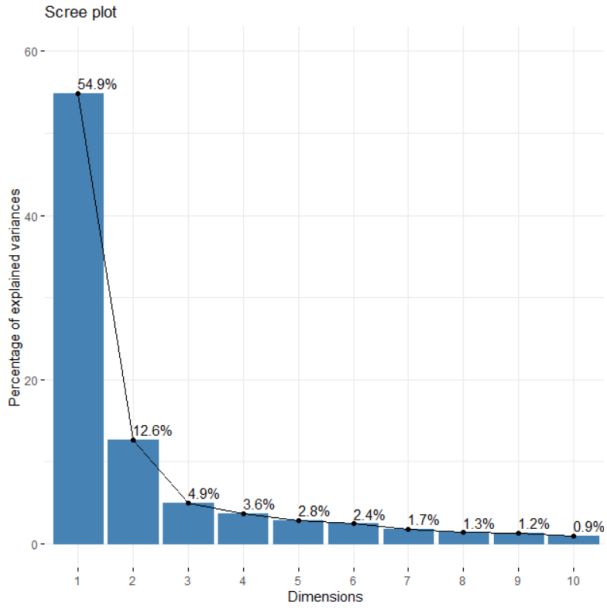


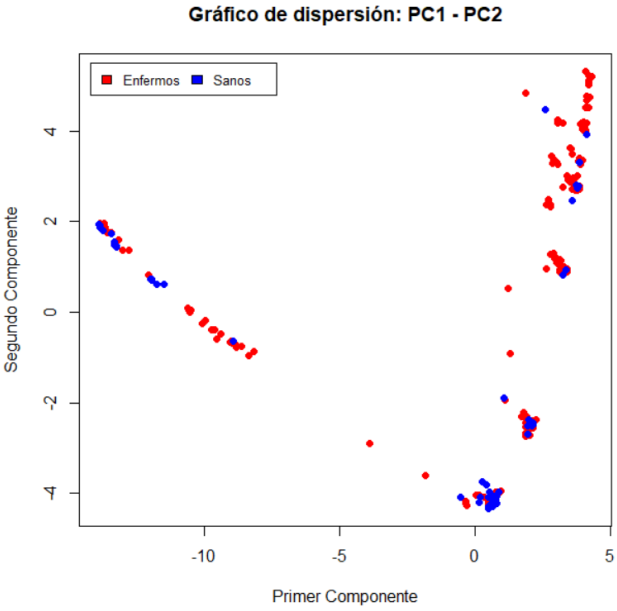
PCA en Secuencias Seccionadas Alineadas con MAFFT



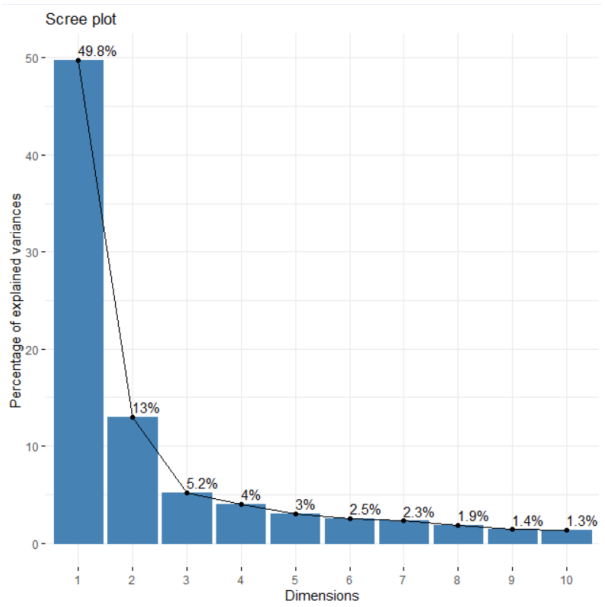


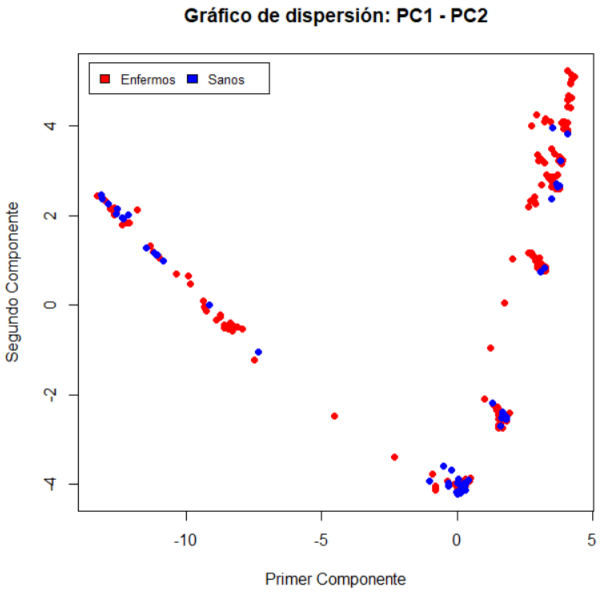
PCA en Secuencias Seccionadas Alineadas con ClustalW



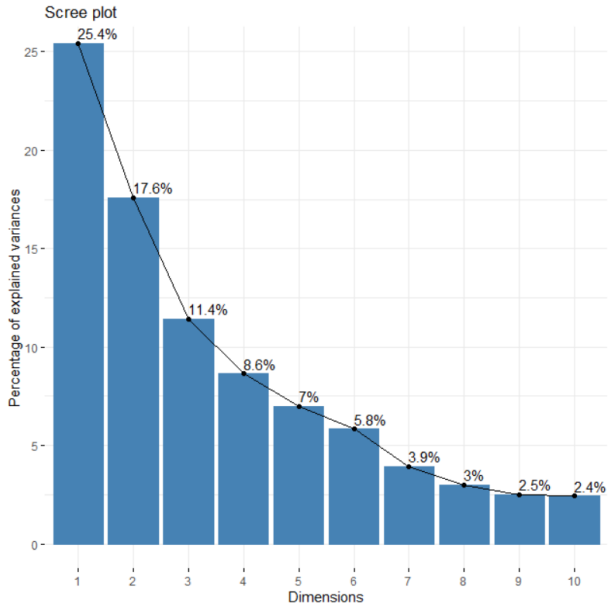


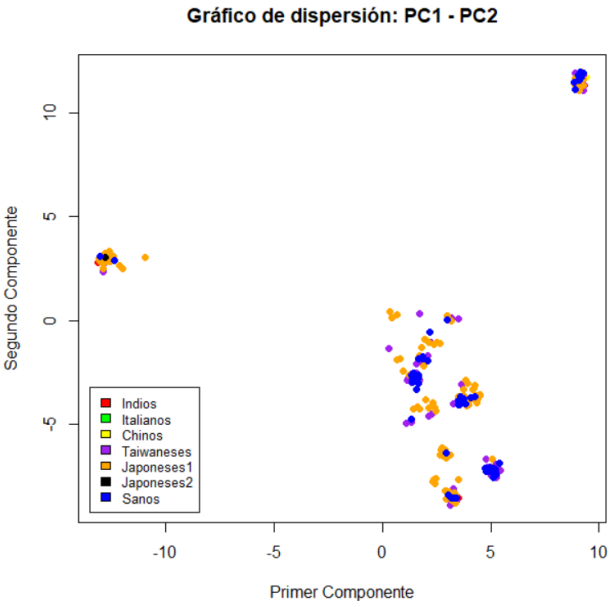
PCA en Secuencias Seccionadas Alineadas con Muscle



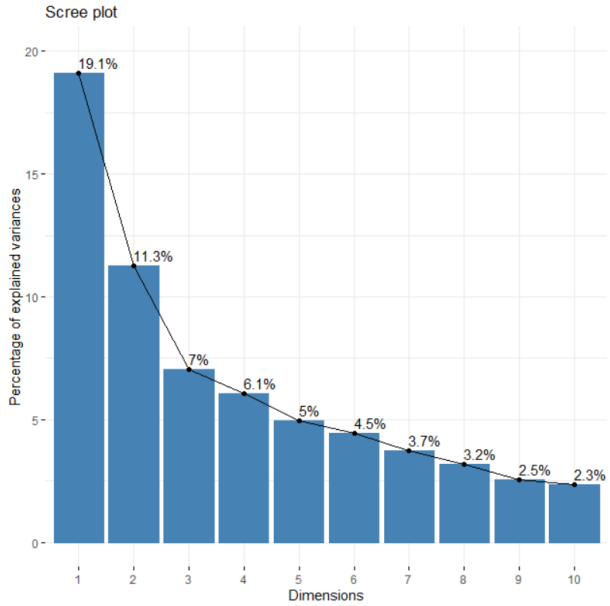


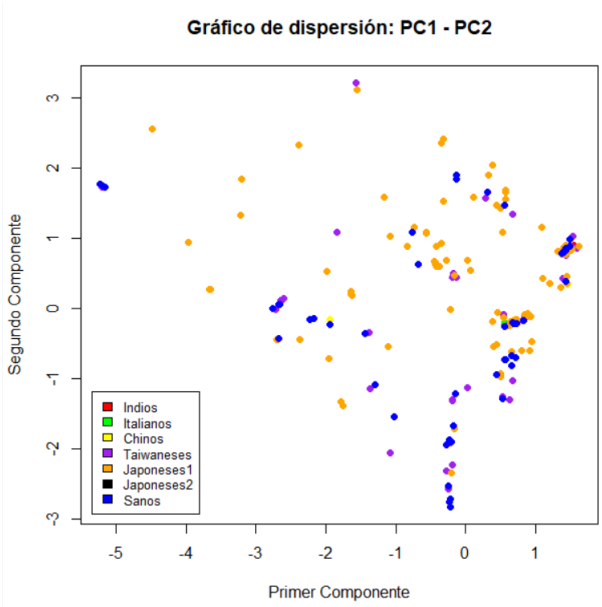
Clasificación de Etnias con PCA usando la región seccionada (secuencias sin alinear)





Clasificación de Etnias con PCA usando la región seccionada (secuencias alineadas)





Graficas de los primeros 5 PCs

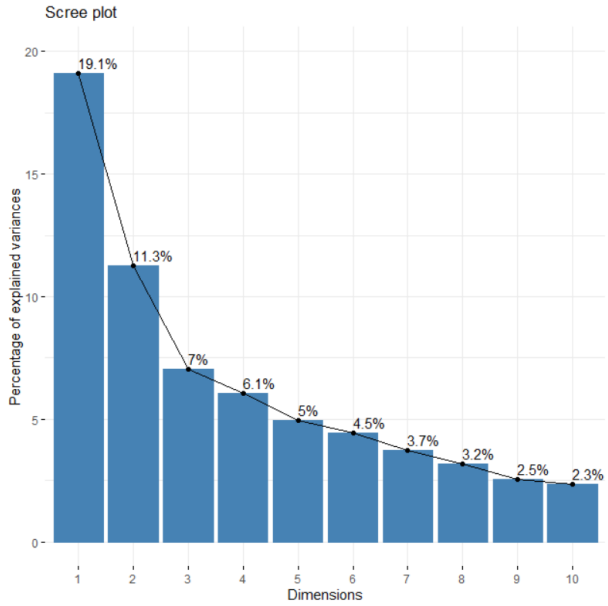


Gráfico de dispersión: PC1 - PC2

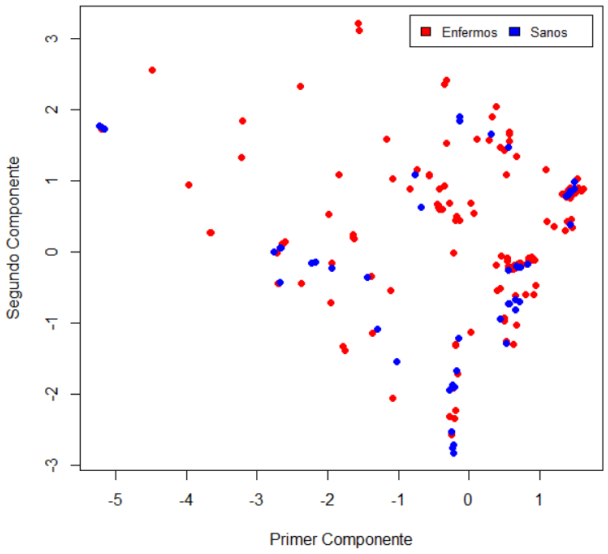


Gráfico de dispersión: PC1 – PC3

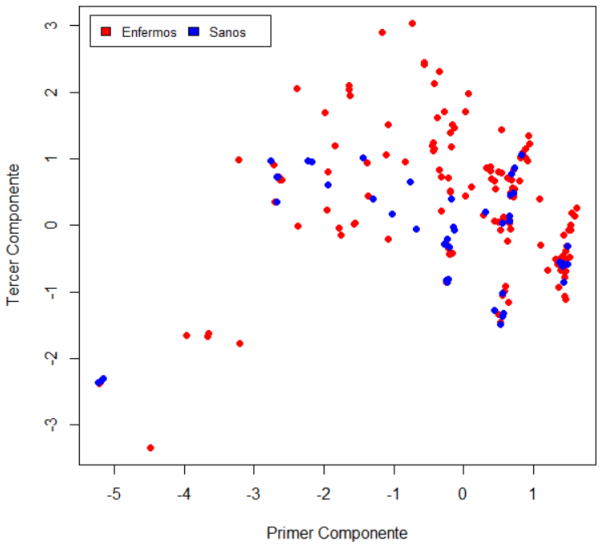


Gráfico de dispersión: PC1 – PC4

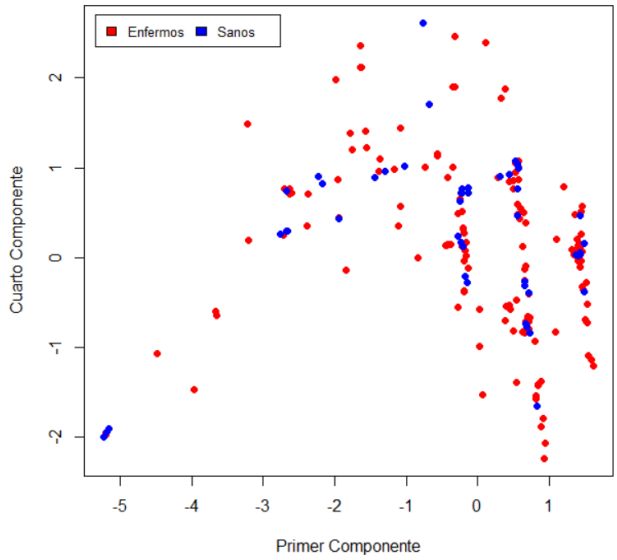


Gráfico de dispersión: PC1 – PC5

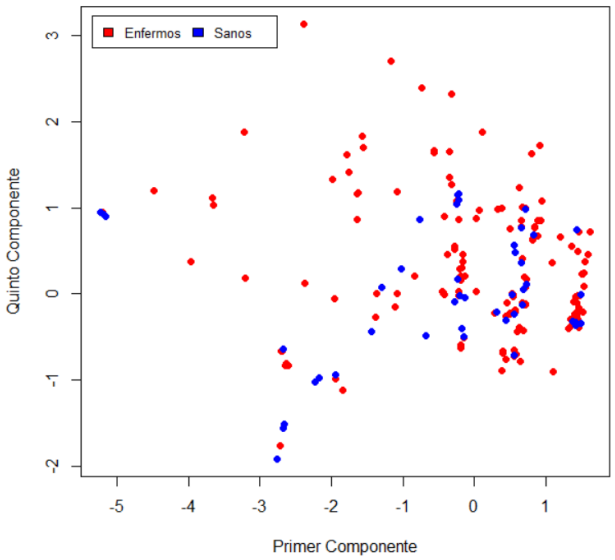


Gráfico de dispersión: PC2 – PC3

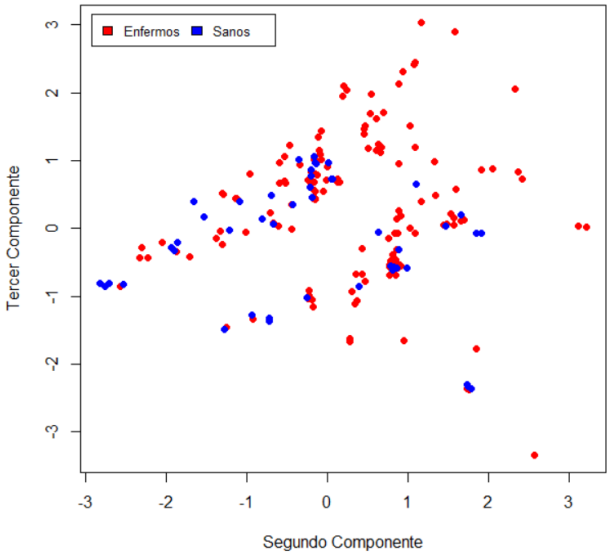


Gráfico de dispersión: PC2 – PC4

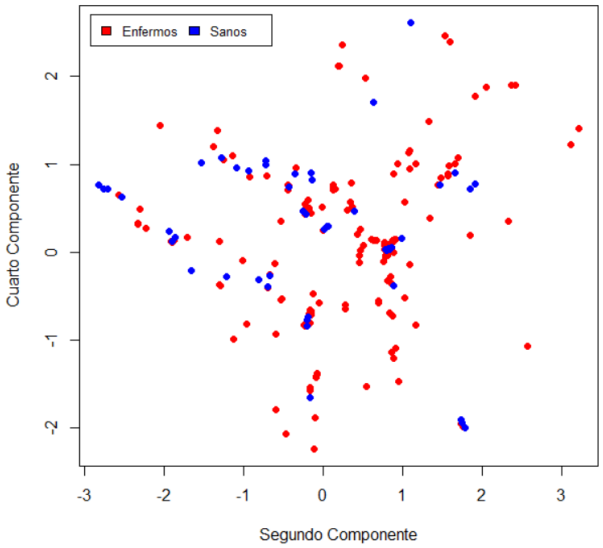


Gráfico de dispersión: PC2 – PC5

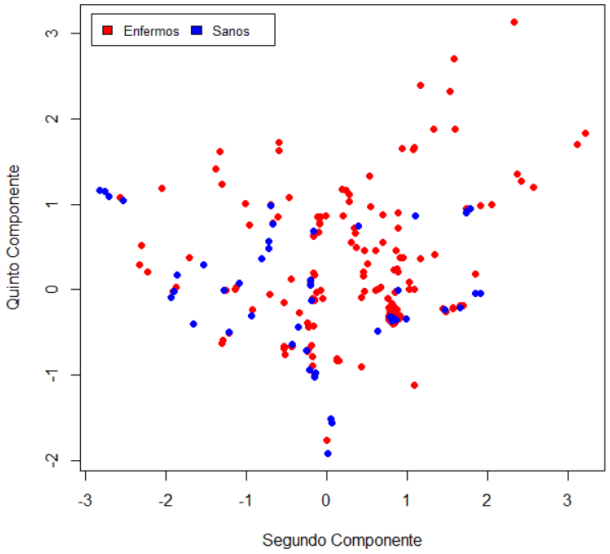


Gráfico de dispersión: PC3 – PC4

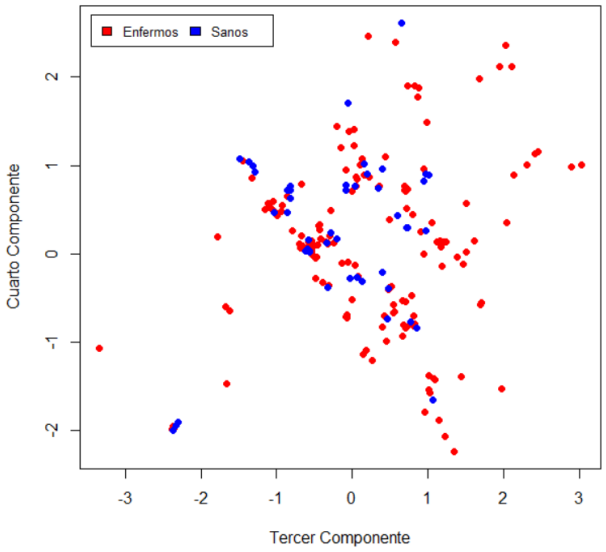
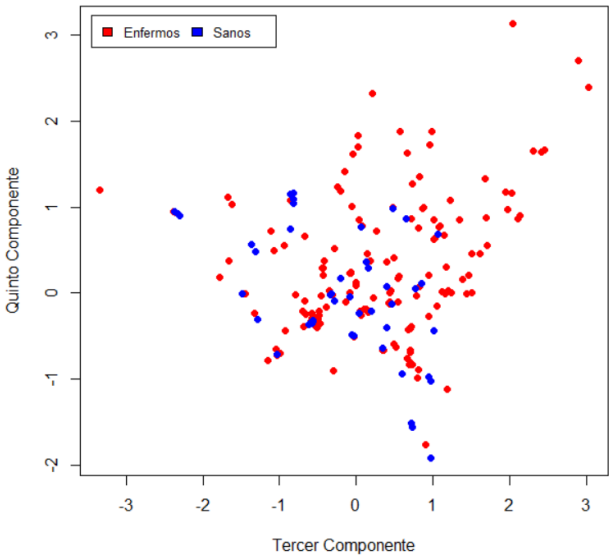
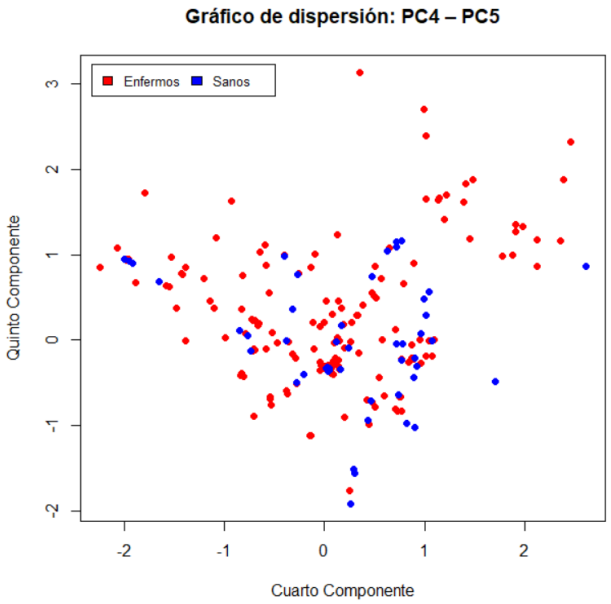
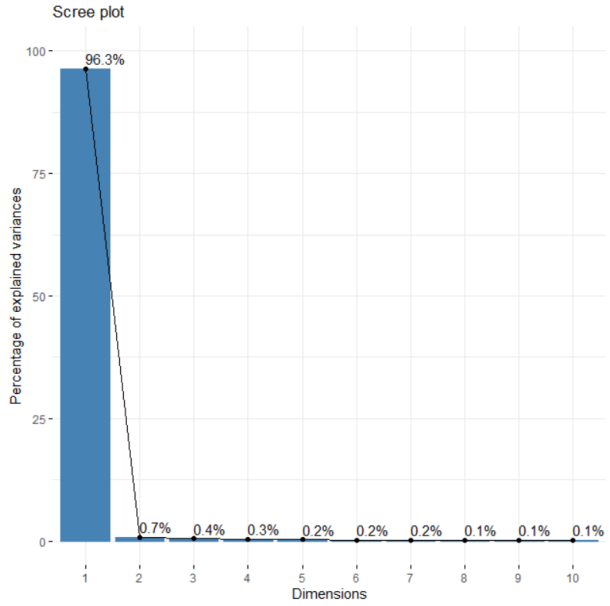


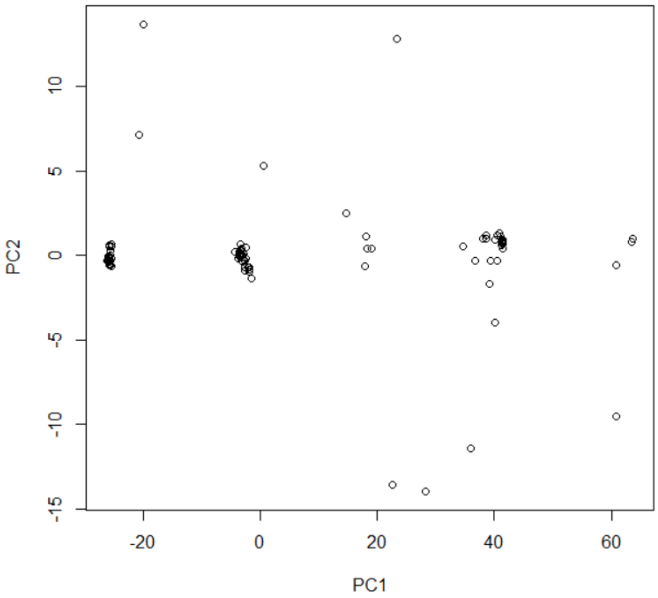
Gráfico de dispersión: PC3 – PC5





Graficas de PCA con Datos Transpuestos





ANEXO C

Los reportes de interacciones serán anexados como 3 documentos pdf aparta, debido a la extensión de cada uno estos reportes.

Productos Entregados

A continuación, se anexan los productos entregados en este doctorado.

Articulo

Genome-Wide Association of Diabetes Mellitus Type 2 in Mitochondrion Genome

Asociación de Genoma Completo de Diabetes Mellitus Tipo 2 en el Genoma Mitocondrial

J. A. Valdez[1,2], P. Mayorga[2], C. V. Angulo[1], R. V. Angulo[1]

Abstract — **Type 2 Diabetes Mellitus (DM2) is a complex and multifactorial disorder, it is currently classified as one of the epidemics of the 21st century due to the magnitude with which it is growing, as well as the cardiovascular effects it causes. This work presents an association study of base-pair positions of the human mitochondrial with Type 2 diabetes. The data consisted of 510 complete mitochondrial genomes, of which 437 were from patients with Type 2 Diabetes Mellitus and 73 were from healthy patients. An alignment algorithm allowed to visualize and select one with optional positions, a Principal Components analysis allowed to visualize the structure of the data and the regression analysis allowed to discover 3 associated base-pair regions. After inspecting the genome annotation, 3 genes were found as candidates to be associated, of which one was found in previous studies to be associated with Type 2 Diabetes Mellitus. This study provides new evidence of genomic positional association with Type 2 Diabetes Mellitus.**

Keywords — **Genome-Wide Association Study (GWAS), Type 2 Diabetes Mellitus (DM2), Logistic Regression.**

Resumen — **La Diabetes Mellitus Tipo 2 (DM2) es un trastorno complejo y multifactorial, actualmente es catalogada como una de las epidemias del siglo XXI por la magnitud con la que está creciendo, así como los efectos cardiovasculares que provoca. Este trabajo presenta un estudio de asociación de posiciones bases-par del genoma mitocondrial humano con diabetes Mellitus Tipo 2. Los datos consistieron de 510 genomas mitocondriales completos, de los cuales 437 fueron de pacientes con Diabetes Mellitus Tipo 2 y 73 fueron de pacientes sanos. Un algoritmo de alineación permitió visualizar y seleccionar una región con variabilidad alélica, un análisis de Componentes Principales permitió visualizar la estructura de los datos y análisis de regresión permitió descubrir 3 posiciones base-par asociadas. Después de inspeccionar la anotación del genoma, 3 genes fueron enconrtrados como candidatos a estar asociados, de los cuales uno fue encontrado en estudios previos como asociado a Diabetes Mellitus Tipo 2. Este estudio aporta nueva evidencia de asociación de posiciones genómica con Diabetes Mellitus Tipo 2.**

Palabras Clave — ***Estudio de asociación de genoma completo (GWAS), Diabetes Mellitus Tipo 2 (DM2), Regresion logística.***

I. Introduccion

La diabetes mellitus tipo 2 (DM2) es un trastorno complejo y multifactorial caracterizado por hiperglucemia crónica debido a la interacción de múltiples variantes genéticas y varios factores ambientales. Como resultado del envejecimiento de la población y la creciente prevalencia de obesidad e inactividad física, el número de pacientes con diabetes tipo 2 ha aumentado

notablemente en todo el mundo [1]. Es catalogada como una de las epidemias del siglo XXI, tanto por su creciente magnitud como por su impacto negativo en padecimientos cardiovasculares [2].

La DM2 es una enfermedad heterogénea, de etiología multifactorial, en la que se combinan la resistencia a la insulina y la inadecuada secreción de insulina compensatoria por células beta del páncreas; se manifiesta como una hiperglucemia crónica, acompañada por trastornos del metabolismo de carbohidratos, grasas y proteínas. La susceptibilidad de esta enfermedad está determinada por el efecto combinado de factores genéticos y ambientales [2].

El ambiente se refiere a todos los factores no genéticos que modulan el fenotipo y puede incluir factores del ambiente aleatorio (climáticos, geográficos, demográficos y socioeconómicos) así como el denominado estilo de vida (dieta, tabaquismo, alcoholismo y actividad física), que el individuo puede modificar [2].

La enfermedad se considera un trastorno poligénico, en el que cada variante genética confiere un efecto parcial y aditivo. Sólo del 5-10% de los casos de DM2 se deben a defectos de un solo gen; estos incluyen la diabetes de inicio en la madurez de los jóvenes, los síndromes de resistencia a la insulina, la diabetes mitocondrial y la diabetes neonatal [5]. El examen de los genes de susceptibilidad a la DM2 puede ser útil para la predicción, la prevención y el tratamiento temprano de la enfermedad.

Mediante la implementación de estudios de asociación de genoma completo (GWAS, por sus siglas en ingles), el número de variantes genéticas comunes asociadas con la DM2 ha aumentado rápidamente [6-12]. Además, se han identificado más de 40 loci genéticos asociados a la DM2; sin embargo, estos loci se han identificado principalmente en poblaciones europeas [13]. Las regiones genéticas identificadas solo explican una pequeña proporción de la heredabilidad estimada de la DM2, lo que sugiere que quedan por identificar factores genéticos adicionales. Una limitación de GWAS es la gran cantidad de hipótesis y el alto costo económico de estas investigaciones [14]. Varios estudios han abordado la viabilidad y eficacia de los GWAS basados en agrupaciones, con ahorros considerables en tiempo y costo [14-16]. Además, la secuenciación del genoma completo en múltiples muestras en una población brinda una oportunidad sin precedentes para caracterizar de manera integral las variantes polimórficas en la población [17].

El propósito de este trabajo fue realizar un estudio de asociación en el genoma mitocondrial para identificar posiciones genómicas de base-par (bp) estadísticamente asociadas a DM2. Un análisis de alineación permitió visualizar y seleccionar una región genómica con variabilidad alélica. Posteriormente, un Análisis de Componentes Principales (PCA, de sus siglas en inglés) fue utilizado para visualizar la complejidad de los datos; seguido de un análisis de regresión logística simple y múltiple que permitió descubrir posiciones base-par asociadas con DM2. Finalmente, una inspección de la anotación del genoma mitocondrial reveló 3 genes candidatos a estar asociados a DM2.

II. METODOLOGIA

A. Base de Datos

Un total de 510 genomas mitocondriales completos de humanos fueron utilizados en este estudio. Del total de los genomas, 437 fueron de personas con DM2 y 73 de personas sanas. Los datos fueron almacenados en un archivo con formato FASTA y los genomas fueron alineados utilizando el algoritmo CLUSTALW implementado en la herramienta MAFFT [35, 36]. La longitud total del alineamiento resultó de 16,610 nucleótidos.

En los genomas alineados, se realizó una inspección para localizar la región que presentaba mayor variabilidad, resultando la región comprendida de la posición 16,170 a la 16,410, con una longitud de 241 nucleótidos. Se extrajo esta región del alineamiento y el resto del análisis se realizó con estos datos.

B. Análisis de Componentes Principales (PCA)

La idea central del análisis de componentes principales (PCA, de sus siglas en inglés) es reducir la dimensionalidad de un conjunto de datos, que consta de un gran número de variables interrelacionadas, conservando al mismo tiempo la mayor cantidad posible de la variación presente en un conjunto de datos. Esto se logra transformando un nuevo conjunto de variables, las componentes principales (PCs), que no están correlacionadas y se ordenan de tal manera que las primeras conservan la mayor variación presente en todas las variables originales [38].

PCA es teóricamente la transformación óptima para un conjunto de datos dado, en términos de mínimos cuadrados. El procedimiento para obtener los componentes principales se puede resumir de la siguiente manera: Dado un vector X^T de n dimensiones, $X = [x_1, x_2, \ldots, x_n]^T$, cuyos vectores de medias (M), y covarianza (C), están descritos por: $M=E(X)= [m_1, m_2, \ldots, m_n]^T$ y $C = E [(X - M) (X - M)^T]$. Calcular los valores propios $\lambda_1, \lambda_2, \ldots, \lambda_n$ y los vectores propios $P_1, P_2, \ldots, P_n$; y ordenarlos según su magnitud $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$. Seleccionar d vectores propios para representar las n variables, $d < n$. Entonces $P_1, P_2, \ldots, P_d$ se denominan componentes principales [38].

Con el propósito de aplicar PCA a las secuencias, se realizó una trasformación de los nucleótidos, del formato ACGT a formato numérico. A cada nucleótido le fue asignado un valor entre 1 y 4 como sigue: A = 1; C = 2; G = 3 y T = 4. De igual forma, los espacios en blanco (GAP) = 5. A la matriz numérica resultante se le aplicó el análisis de PCA. El propósito de aplicar el análisis PCA fue analizar la estructura de los datos y buscar posibles cúmulos que diferenciaran los datos de personas enfermas y sanas.

C. Regresión

El objetivo de un modelo de regresión lineal es tratar de explicar la relación que existe entre una variable dependiente (variable respuesta) y un conjunto de variables independientes (variables explicativas) X_1, ..., X_n. En un modelo de regresión lineal simple tratamos de explicar la relación que existe entre la variable respuesta (Y) y una única variable explicativa (X). Mediante las técnicas de regresión de una variable Y sobre una variable X, buscamos una función que sea una buena aproximación de una nube de puntos (x_i, y_i), mediante una curva [42].

El modelo de regresión lineal simple tiene la siguiente expresión:

$$\boldsymbol{Y = \alpha + \beta X + \varepsilon} \qquad (1)$$

En donde α es la ordenada en el origen (el valor que toma Y cuando X vale 0), β es la pendiente de la recta (e indica cómo cambia Y al incrementar X en una unidad) y ε una variable que incluye un conjunto grande de factores, cada uno de los cuales influye en la respuesta sólo en pequeña magnitud, a la que llamaremos error. X e Y son variables aleatorias, por lo que no se puede establecer una relación lineal exacta entre ellas [42].

Una extensión natural del modelo de regresión lineal simple consiste en considerar más de una variable explicativa. Los modelos de regresión múltiple estudian la relación entre:

una variable de interés Y (variable respuesta o dependiente) y un conjunto de variables explicativas o regresoras X_1, X_2, . . . , X_p [65].

En el modelo de regresión lineal múltiple se supone que la función de regresión que relaciona la variable dependiente con las variables independientes es lineal, es decir:

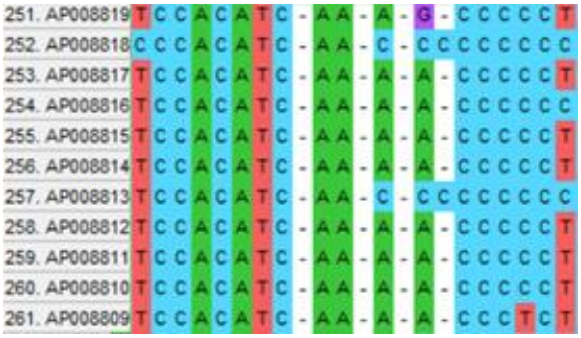
$$\boldsymbol{Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon} \qquad (2)$$

El propósito de utilizar regresión a los datos fue buscar SNPs asociado estadísticamente con la DM2.

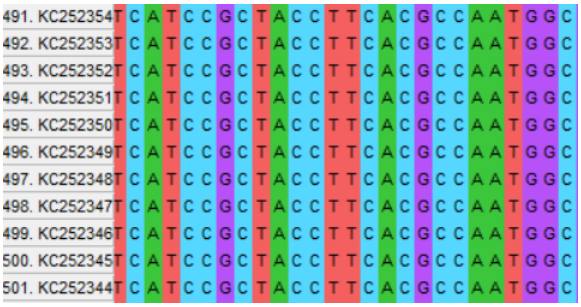
III. Resultados

Los genomas mitocondriales completos de los 510 pacientes fueron alineados con la herramienta MAFFT. El resultado de la alineación fue visualizado con el software MEGA X [60]. Mediante una inspección visual se observó una región con variabilidad, mientras que el resto mostró alineación perfecta. En la Fig. 1 (a) se muestra un fragmento de la región con variabilidad, y en la Fig. 1 (b) se muestra un fragmento de la región sin variabilidad.

Se eligió la región entre las posiciones 16,170 y 16,410, con una longitud de 241 nucleótidos, para realizar el resto de análisis.



(a)



(b)

Fig. 1. Fragmento de la alineación de los genomas mitocondriales de pacientes con DM2 y sanos. El (a) representa una región con variabilidad y el (b) representa una región con nula variabilidad.

Con el propósito de analizar la estructura de la información y buscar posibles cúmulos que diferenciaran los datos de personas enfermas y sanas, se aplicó PCA a la región alineada de alta variabilidad. Para realizar este análisis se utilizó el lenguaje estadístico R. La Fig. 2 muestra la gráfica del Componente Principal 1 (PC1) contra el Componte Principal 2 (PC2). Como podemos observar en la gráfica, la información aparece mezclada y no existe una diferenciación clara entre los grupos. Este análisis nos muestra la complejidad de los datos.

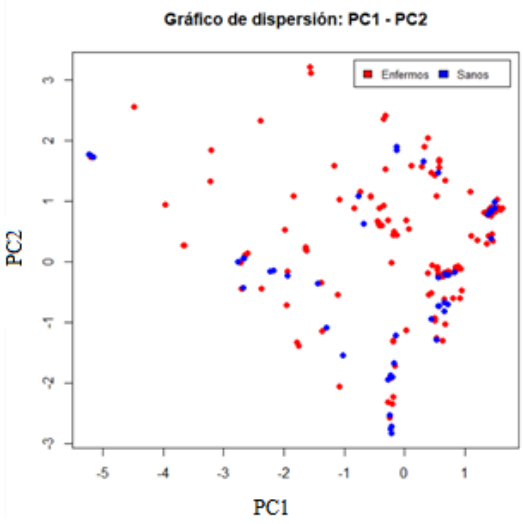


Fig. 2. Gráfico de comparación entre casos vs controles.

El análisis de asociación se realizó en dos pasos, por un lado, se aplicó regresión logística simple a cada posición base-par (bp) de la región variante (241 posiciones base-par), asignando un 1 a la variable dependiente, para todos los pacientes sanos, y un 0 a todos los pacientes con DM2. Se seleccionaron aquellas posiciones que resultaron estadísticamente significantes (valor $p<0.05$). Posteriormente se realizó una regresión logística múltiple agrupando las posiciones que resultaron significantes en la regresión simple. Aquellas que resultaron significantes en la regresión múltiple fueron declarados como posiciones asociadas con DM2. La TABLA XIII muestra las posiciones que resultaron significantes tanto en la regresión simple como en la regresión múltiple.

TABLA XIII

RESULTADOS DE LA REGRESIÓN SIMPLE Y MÚLTIPLE.

Posición genómica (BP)	Regresión Simple (Valor P)	Regresión Múltiple (Valor P)	Asociado con DM2
16,184	**0.0038**	**0.0021**	**Si**
16,222	0.0384	0.6592	No
16,257	0.0289	0.1037	No
16,263	0.0415	0.6937	No
16,282	**0.0033**	**0.0064**	**Si**
16,289	0.0426	0.4447	No
16,344	**0.0038**	**0.0159**	**Si**
16,351	0.0438	0.1983	No

IV. Discusion

Como se observa en la TABLA XIII, después de la regresión múltiple tres posiciones resultaron asociados con DM2. Las posiciones y sus valores *p* resultantes fueron: 16,184; 16,282 y 16,344 y 0.0021; 0.0064 y 0.0159, respectivamente. Con el fin de ubicar el gen asociado, se inspeccionó la anotación del genoma mitocondrial humano en la base de datos del Centro Nacional de Información Biotecnológica, (NCBI, de sus siglas en inglés) (https://www.ncbi.nlm.nih.gov/). Tres genes fueron ubicados en una distancia menor a 3,000 bases par (bp) de las posiciones asociadas. Estos genes son: el CYTB que produce la proteína Citocromo B y contribuye en la conversión de la energía de los alimentos a energía celular (Adenosin Trifosfato, ATP), el gen TRNP el cual es el tRNA de la Prolina, y el gen TRNT que es el tARN de la Treonina. En especial el gen TRNT, en un estudio realizado por Ke Li, et al., 2020 se encontró asociado con la heredabilidad materna de DM2 en familias chinas [63], y en otro estudio realizado por Momiyama, et al., 2003 este gen fue asociado, al igual que en nuestro estudio, con la posición genómica 16,184; y declarado como una de los causantes de la hipertrofia ventricular izquierda en pacientes con DM2 en familias japonesas [64].

V. Conclusion

En este estudio de asociación se analizaron 510 genomas mitocondriales completos. Del total de los genomas, 437 fueron de pacientes con DM2, y 73 de pacientes sanos. Una alineación de genoma completo permitió ubicar una región variable en su contenido alélico; un análisis PCA permitió visualizar la complejidad de los datos, y un análisis de regresión logística permitió encontrar 3 posiciones base-par asociados con DM2. Las posiciones asociadas fueron ubicadas en una cercanía menor a 3k bp de tres genes, de los cuales uno (gen TRNT) fue reportado por estudios previos como asociado a DM2. Finalmente, este estudio adhiere nueva evidencia de asociación de posiciones genómicas con DM2.

AGRADECIMIENTOS

Este trabajo se desarrolló dentro del programa de Maestría y Doctorado en Ciencias e Ingeniería (MYDCI) ofertado por la Universidad Autónoma de Baja California. Además, contó con el apoyo de una beca CONACYT.

Referencias

[1] L. Chen, D. J. Magliano, and P. Z. Zimmet, "The worldwide epidemiology of type 2 diabetes mellitus--present and future perspectives," (in eng), *Nat Rev Endocrinol*, vol. 8, no. 4, pp. 228-36, Nov 8 2011, doi: 10.1038/nrendo.2011.183.

[2] L. C. L. O´Farrill, Cuervo, A. M. d. S., Ferrer, R. L., & Valdés, M. T. L, "Interacción genoma-ambiente en la diabetes mellitus tipo 2," *Acta Médica del Centro*, vol. 12, no. 4, 2018. [Online]. Available: http://www.revactamedicacentro.sld.cu/index.php/amc/article/view/948.

[3] F. J. Tsai *et al.*, "A genome-wide association study identifies susceptibility variants for type 2 diabetes in Han Chinese," (in eng), *PLoS Genet,* vol. 6, no. 2, p. e1000847, Feb 19 2010, doi: 10.1371/journal.pgen.1000847.

[4] L. J. Scott *et al.*, "A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants," (in eng), *Science,* vol. 316, no. 5829, pp. 1341-5, Jun 1 2007, doi: 10.1126/science.1142382.

[5] R. Saxena *et al.*, "Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels," (in eng), *Science,* vol. 316, no. 5829, pp. 1331-6, Jun 1 2007, doi: 10.1126/science.1142358.

[6] R. Sladek *et al.*, "A genome-wide association study identifies novel risk loci for type 2 diabetes," *Nature,* vol. 445, no. 7130, pp. 881-885, 2007/02/01 2007, doi: 10.1038/nature05616.

[7] E. Zeggini *et al.*, "Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes," (in eng), *Science,* vol. 316, no. 5829, pp. 1336-41, Jun 1 2007, doi: 10.1126/science.1142364.

[8] P. R. Burton *et al.*, "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls," *Nature,* vol. 447, no. 7145, pp. 661-678, 2007/06/01 2007, doi: 10.1038/nature05911.

[9] E. Zeggini *et al.*, "Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes," *Nature Genetics,* vol. 40, no. 5, pp. 638-645, 2008/05/01 2008, doi: 10.1038/ng.120.

[10] J. Gudmundsson *et al.*, "Two variants on chromosome 17 confer prostate cancer risk, and the one in TCF2 protects against type 2 diabetes," *Nature Genetics,* vol. 39, no. 8, pp. 977-983, 2007/08/01 2007, doi: 10.1038/ng2062.

[11] R. Saxena *et al.*, "Large-scale gene-centric meta-analysis across 39 studies identifies type 2 diabetes loci," (in eng), *Am J Hum Genet,* vol. 90, no. 3, pp. 410-25, Mar 9 2012, doi: 10.1016/j.ajhg.2011.12.022.

[12] A. E. Baum *et al.*, "A genome-wide association study implicates diacylglycerol kinase eta (DGKH) and several other genes in the etiology of bipolar disorder," *Molecular Psychiatry,* vol. 13, no. 2, pp. 197-207, 2008/02/01 2008, doi: 10.1038/sj.mp.4002012.

[13] A. Galvan *et al.*, "Genome-wide association study in discordant sibships identifies multiple inherited susceptibility alleles linked to lung cancer," *Carcinogenesis,* vol. 31, no. 3, pp. 462-465, 2009, doi: 10.1093/carcin/bgp315.

[14] L. M. Forstbauer *et al.*, "Genome-wide pooling approach identifies SPATA5 as a new susceptibility locus for alopecia areata," *European Journal of Human Genetics,* vol. 20, no. 3, pp. 326-332, 2012/03/01 2012, doi: 10.1038/ejhg.2011.185.

[15] L. P. Wong *et al.*, "Deep whole-genome sequencing of 100 southeast Asian Malays," (in eng), *Am J Hum Genet,* vol. 92, no. 1, pp. 52-66, Jan 10 2013, doi: 10.1016/j.ajhg.2012.12.005.

[16] S. Kuraku, C. M. Zmasek, O. Nishimura, and K. Katoh, "aLeaves facilitates on-demand exploration of metazoan gene family trees on MAFFT sequence alignment server with enhanced interactivity," *Nucleic Acids Research,* vol. 41, no. W1, pp. W22-W28, 2013, doi: 10.1093/nar/gkt389.

[17] K. Katoh, J. Rozewicki, and K. D. Yamada, "MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization," *Briefings in Bioinformatics,* vol. 20, no. 4, pp. 1160-1166, 2017, doi: 10.1093/bib/bbx108.

[18] A. G. Mateos-Valenzuela, M. E. González-Macías, S. Ahumada-Valdez, C. Villa-Angulo, and R. Villa-Angulo, "Risk factors and association of body composition components for lumbar disc herniation in Northwest, Mexico," *Scientific Reports,* vol. 10, no. 1, p. 18479, 2020/10/28 2020, doi: 10.1038/s41598-020-75540-5.

[19] C. C. Limeres. "REGRESIÓN LINEAL SIMPLE." http://eio.usc.es/eipc1/BASE/BASEMASTER/FORMULARIOS-PHP-DPTO/MATERIALES/Mat_50140116_Regr_%20simple_2011_12.pdf (accessed 19/Enero, 2022).

[20] "Regresion Lineal Multiple." http://eio.usc.es/eipc1/BASE/BASEMASTER/FORMULARIOS-PHP-DPTO/MATERIALES/Mat_50140128_RegresionMultiple.pdf (accessed 19/Enero, 2022).

[21] S. Kumar, G. Stecher, M. Li, C. Knyaz, and K. Tamura, "MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms," (in eng), *Mol Biol Evol,* vol. 35, no. 6, pp. 1547-1549, Jun 1 2018, doi: 10.1093/molbev/msy096.

[22] K. Li, L. Wu, J. Liu, W. Lin, Q. Qi, and T. Zhao, "Maternally Inherited Diabetes Mellitus Associated with a Novel m.15897G>A Mutation in Mitochondrial tRNA(Thr) Gene," (in eng), *J Diabetes Res,* vol. 2020, p. 2057187, 2020, doi: 10.1155/2020/2057187.

[23] Y. Momiyama *et al.*, "A mitochondrial DNA variant associated with left ventricular hypertrophy in diabetes," (in eng), *Biochem Biophys Res Commun,* vol. 312, no. 3, pp. 858-64, Dec 19 2003, doi: 10.1016/j.bbrc.2003.10.195.

Capitulo de Libro

Book Chapter Template

Genome-wide association in the mitochondrial genome identifies two novel genes involved in diabetes mellitus type 2

J. A. Valdez[1,2], P. Mayorga[2], R. V. Angulo[1], C. V. Angulo[1]

[1]Engineering Institute, Universidad Autónoma de Baja California, Mexicali, B.C., México.

[2]Department of Electrical - Electronics, TecNM/ITMexicali, Mexicali, B.C., México.

Email: a1174500@uabc.edu.mx

Abstract

Diabetes Mellitus Type 2 (DM2) is a complex and multifaceted disorder currently listed as one of the epidemics of the twenty-first century due to its prevalence and the adverse cardiovascular effects it causes. This study examines the relationships between base pair positions in human mitochondrial genome and type 2 diabetes. The data included 510 complete mitochondrial genomes, of which 437 belonged to individuals with type 2 diabetes and 73 to healthy individuals. An alignment algorithm allowed inspecting and choosing a region with optional positions for analysis, a principal component analysis permitted viewing the data structure, and after a regression analysis, we declared three base pair positions associated to DM2. Upon examination of the genome annotation, three genes were identified as potential candidates for association, one of which was previously linked to type 2 diabetes according to previous studies. This study offers further proof of a possible genetic link between type 2 diabetes and metabolic syndrome.

Keywords: Genome-Wide Association Study (GWAS), Type 2 Diabetes Mellitus (DM2), Logistic Regression.

1. Introduction

A group of metabolic disorders known as diabetes mellitus are characterized by chronic hyperglycemia and can be caused by problems with insulin secretion, insulin action, or both. Alterations to the lipid and protein metabolism coexist with hyperglycemia. Long-term sustained hyperglycemia is linked to damage, dysfunction, and failure of many different organs and systems, particularly the heart, blood vessels, nerves, and the retina [66].

There are several types of diabetes and other categories of glucose intolerance. **Type 1 diabetes mellitus (DM1):** Its hallmark is autoimmune destruction of the β cell, which causes absolute insulin deficiency, and a tendency to ketoacidosis. Such destruction in a high percentage is mediated by the immune system, which can be evidenced by the determination of antibodies: Anti-GAD (anti glutamate decarboxylase), anti-insulin, and against the islet cell, with a strong association with the specific DQ-A alleles. and DQ-B of the major histocompatibility complex (HLA). DM1 can also be of idiopathic origin, where the measurement of the aforementioned antibodies gives negative results [66]. It usually manifests itself in the infant-juvenile age (before the age of 30) and the vast majority are of autoimmune origin. It is characterized by a defect in insulin secretion and constitutes 5-10% of all cases of diabetes. It is always a subsidiary of insulin treatment [67].

Diabetes type 2 (DM2): This is the most prevalent variety and is frequently linked to obesity or an increase in visceral fat. Ketoacidosis rarely develops spontaneously. The issue ranges from a predominant resistance to insulin, accompanied by a relative hormone deficiency, to a progressive malfunction in its secretion [66]. It is the most frequent form of DM2 since it represents between 90 and 95% of cases. It usually appears after the age of 40 and is associated with obesity, which is present in up to 80% of patients with type 2 DM. Its treatment requires diet and exercise alone or is associated with oral antidiabetics and/or insulin [67].

Gestational diabetes mellitus (GDM): Specifically, groups glucose intolerance detected for the first time during pregnancy. Hyperglycemia before twenty-four weeks of pregnancy is considered undiagnosed pre-existing diabetes [66]. It occurs in 1-14% of pregnant women and is associated with an increased risk of obstetric and perinatal complications [67].

Due to the interaction of numerous genetic variants and other environmental factors, diabetes mellitus type 2 (DM2) is a complex and multifaceted disorder characterized by chronic hyperglycemia. The prevalence of obesity and physical inactivity, together with the aging of the population, have all contributed to a significant rise in the number of people worldwide who have type 2 diabetes [1]. It's classified as one of the epidemics of the 21st century, both for its growing magnitude and for its negative impact on cardiovascular diseases [2].

DM2 is a heterogeneous disease of multifactorial etiology, in which insulin resistance and inadequate compensatory insulin secretion by pancreatic beta cells are combined; It manifests as chronic hyperglycemia, accompanied by carbohydrate, fat, and protein metabolism disorders. The susceptibility of this disease is determined by the combined effect of genetic and environmental factors [2].

Environment refers to all non-genetic factors that modulate the phenotype, which may include random environmental factors such as climate, geography, demographics, and socioeconomics; as well as the lifestyle that is made up of diet, smoking, alcoholism, and physical activity, which the individual can modify [2].

The disease is regarded as a polygenetic disturbance in which each genetic variety confers a partial and additive effect. Just 5-10% of cases of DM2 may be attributed to genetic defects; these cases include juvenile-onset diabetes, insulin-resistance syndromes, mitochondrial diabetes, and neonatal diabetes [5]. Examination of DM2 susceptibility genes may be useful for the prediction, prevention, and early treatment of the disease.

Through the implementation of genome-wide association studies (GWAS), the number of common genetic variants associated with DM2 has increased rapidly [6-12]. In addition, more than 40 genetic loci associated with DM2 have been identified; however, these loci have been identified primarily in European populations [13]. Still there are additional genetic factors to be discovered since the identified genetic regions only account for a small portion of the estimated heritability of DM2. The high economic cost and a large number of hypotheses in these studies are a limitation of GWAS [14]. Several research studies have examined cluster-based GWAS's viability and efficiency, with significant time and financial savings. [14-16]. In addition, whole genome sequencing across multiple samples in a population provides an unprecedented opportunity to comprehensively characterize polymorphic variants in the population [17].

Type 2 diabetes, as mentioned, is a complex illness brought on by numerous genetic and environmental factors; family-based and peers studies estimate that heredity ranges from 22% to 73%. Recent estimates placed the prevalence of DM2 in adults, adjusted for age, at 7,6% in European Americans, 14,9% in Afro-Americans, 4,3% to 8,2% in Asian Americans, and 10,9% to 15,6% in Hispanic Americans [68-71]. More than 40 genetic loci associated with DM2 have been identified, but so far, these locations have primarily been revealed through studies of people with European ancestry. The candidate gene association studies discovered a link between DM2 and nonsensical variants in PPARG (MIM 601487) and KCNJ11 (MIM 600937), which are targets for drugs to treat diabetes, and they implicated common genetic variants responsible for Mendelian forms of diabetes in DM2 [72-77].

The first genome-wide association studies (GWAS) for DM2 [6, 7, 9, 47, 48] and fasting glucose [49] successfully identified multiple associated loci. And, through recent GWAS meta-analyses for DM2 [50] and quantitative glycemic characteristics [51], the number of loci associated with DM2 have significantly increased in European populations; the majority of these variants act via defects in the function of beta-cells rather than insulin action. In total, known variants associated with DM2 account for 10% of genetic variation [50, 52], therefore it is likely that more locations and independent factors increase the risk of the disease.

Few people outside of Europe are aware of the genetic factors that contribute to type 2 diabetes. A new locus (KCNQ1 [MIM 607542]) was discovered based on a GWAS in a Japanese population [53, 54] and was later discovered to have separate alleles in people of European ancestry [50]. Most recently, GWAS in Chinese populations [5, 23], Japanese [55], and south Asian [56] discovered additional DM2 loci that exceed genome-wide significance. To date, GWAS in African Americans has been underpowered to detect new loci [57].

In a recent multiethnic meta-analysis, three DM2 risk loci in Europe (GATAD2A/CILP2/PBX4, TH/INS, and SREBF1), one DM2 risk locus in Africa (HMGA2), and one DM2 risk locus in multiple ethnic groups (BCL2) were associated confirming that an allele-based gene score exists. Hence, the multiethnic GWAS of DM2 should result in the discovery of additional genes associated with diabetes that are relevant to numerous ethnic groups [13].

There are still additional genetic factors to be discovered since the identified genetic regions only account for a small portion of the estimated heritability of DM2. The high economic cost and a large number of hypotheses in these studies are a limitation of GWAS [14]. Several studies have looked at the viability and effectiveness of GWAS based on clusters, with considerable time and cost savings [14-16]. In addition, whole genome sequencing across multiple samples in a population provides an unprecedented opportunity to comprehensively characterize polymorphic variants in the population [17].

The purpose of this work was to perform an association study in the mitochondrial genome to identify base-pair (bp) genomic positions statistically associated with DM2. An alignment analysis allowed visualization and select a genomic region with allelic variability. Subsequently, a Principal Component Analysis (PCA) was used to visualize the complexity of the data; followed by a simple and multiple logistic regression analysis that allowed the discovery of base-pair positions associated with DM2. Finally, an inspection of the mitochondrial genome annotation revealed 3 candidate genes to be associated with DM2.

2. Methodology

Next, the database used in this study will be explained, as well as the techniques used for the analysis of DNA sequences.

2.1 Database

We explored genetic variants of these type 2 diabetes-associated genes in different populations using genome-wide association analysis available in the Type 2 Diabetes Knowledge Portal database (http://www.type2diabetesgenetics.org/). The search criteria were: patients with DM2, considering a p-value <0.05 in the X2 test and an Odds Ratio> 1.0. Based on the results obtained, the variants were evaluated and identified in NCBI dbSNP (https://www.ncbi.nlm.nih.gov/snp/), and their registration was documented in ClinVar (https://www.ncbi .nlm.nih.gov/clinvar/). Related polymorphisms were explored in the UCSC Genome Browser (https://genome.ucsc.edu/) using the GRCh37/hg19 version and the change of polymorphisms as reference allele/effect and minimum allele frequency were compared with the information available in genomic databases. The allele frequency and the genotype frequency of the effect on heterozygotes and homozygotes were queried in the 1000 Genomes database using Ensembl (http://grch37.ensembl.org/index.html). Finally, samples of different tissues from patients with type 2 diabetes were analyzed with the Orange package (https://orange.biolab.si). To identify the differences in expression of this gene in different tissues, from GEO data sets (https://www.ncbi.nlm.nih.gov/gds) expression values of muscle, liver, and pancreas were obtained and the Differences were analyzed by Mann Whitney U Test considering p <0.05 significant.

To explore the prevalence and distribution of mitochondrial polymorphisms associated with DM2, the search for complete sequences of the mitochondrial chromosome (16,569 base pairs) was designed and minor fragments were considered; because most of the works on the subject are amplified for the control region (D-loop) with a size smaller than 1000 base pairs in the nucleotide database of the National Center for Biotechnology Information (NCBI) (https:/ /www.ncbi.nlm.nih.gov/nucleotide) where the GenBank, the most extensive collection of genetics and genomics available, is located. Changes in the sequence were identified, as well as the insertion, deletion, and heteroplasmy sites. We also estimated the number of mitochondrial single nucleotide polymorphisms (mtSNPs), the average number of distinct nucleotides among populations, as well as the number of fixed differences, shared polimorfisms, and mono- and polymorphic mutations between populations. This made it easier to identify the polimorfisms that are most prevalent in the control region and the mtDNA codifying region.

The database is separated into two fasta files, the first file has 437 whole mitochondrial genome sequences from type 2 diabetic human patients and the second file has 73 from healthy individuals (each sequence is the most common or dominant whole mitochondrial chromosome in each individual, with 16,569 bases each, although it may vary a few nucleotides between different individuals). The shortest sequence is 16,554 bases.

Both files were combined in a single file, in this new file the sick people were placed first and then the healthy ones. To have a total of 510 sequences. Once aligned, the sequences with the MEGA software have a length of 16609 data. After this, a visual analysis of all the already aligned sequences was carried out, looking for the region that presented the greatest disadvantage, the region resulting from position 16170-16410, a total of 241 positions or data.

In addition, the nucleotides (Adenine (A), Cytosine (C), Guanine (G), Thymine (T)) were changed by numbers as follows, this to perform a cluster analysis: A = 1, C = 2, G = 3 and T = 4, GAP(-) = 5. There were also other letters other than nucleotides such as R, Y, W, N, these letters were changed to the number 9. These last letters according to the nomenclature of the International Union of Pure and Applied Chemistry (IUPAC) correspond to:

R = GA (purine)

Y = TC (pyrimidine)

W = AT (weak bonds)

N = AGCT (any)

In addition, as extra information, we searched to which ethnic groups the people in the DNA sequences of the database belonged. Finding the following results: 239 sequences belong to Taiwanese people, 62 people are Indian, there are 6 Italians, 11 Chinese people, and 192 Japanese.

A total of 510 complete human mitochondrial genomes were used in this study. Of the total genomes, 437 were from people with DM2 and 73 from healthy people. The data was stored in a FASTA format file and the genomes were aligned using the CLUSTALW algorithm implemented in the MAFFT tool [35, 36]. The total length of the alignment was 16,610 nucleotides.

In the aligned genomes, an inspection was carried out to locate the region with the highest frequency, resulting in the detected region from position 16,170 to 16,410, with a length of 241 nucleotides. This region of the alignment was removed and the rest of the analysis was performed with these data.

2.2 Principal Component Analysis (PCA)

The main goal of principle component analysis (PCA) is to reduce the dimensionality of a set of data, which often consists of a large number of interrelated variables, while retaining all possible variation. This is accomplished by transforming a new group of variables known as the principal components (PCs), which are disassociated from one another and arranged so that the first few retain the most variation found in the total set of original variables [38].

Theoretically, PCA is the best least squares transformation of the given set of data. In order to obtaining the key components we provide a vector X^T of n dimensions, $X = [x^1, x^2, \ldots, x^n]^T$, whose mean vectors (M), and covariance (C) are described by $M=E(X)= [m_1, m_2, \ldots, m_n]^T$ and $C = E [(X - M) (X - M)^T]$. Then we calculate the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ and the eigenvectors $P_1, P_2, \ldots, P_n$; and order them according to their magnitude $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$. The d eigenvectors must be chosen to represent the n variables, $d < n$. Then $P_1, P_2, \ldots, P_d$ are known as principal components [38].

In order to apply PCA to the sequences, a transformation of the nucleotides was performed, from the ACGT format to the numerical format. Each nucleotide was assigned a value between 1 and 4 as follows: A = 1; C =2; G = 3 and T = 4. In the same way, the blank spaces (GAP) = 5. The PCA analysis was applied to the resulting numerical matrix. The purpose of applying the PCA analysis was to analyze the structure of the data and look for possible clusters that differentiated the data from sick and healthy people.

2.3 Entropy Analysis

Shannon's entropy theory, initially developed by Claude E. Shannon, is applied to measure the contrast between criteria and this information is used to make decisions. In this analysis it is indicated that for all p_i within a probability distribution P, there is a measure H, which satisfies the following properties [40]:

H is a continuous positive function,

If all p_i is equal and $p_i = 1/n$, then H should be an increasing monotonic function of n; and,

For all, $n \geq 2$,

$$H(p_1, p_2, \ldots, p_n) = h(p_1 + p_2, p_3 \ldots, p_n) + (p_1 + p_2) H\left(\frac{p_1}{p_1 + p_2}, \frac{p_2}{p_1 + p_2}\right)$$

Shannon proved that the only function that satisfies these conditions is:

$$\boldsymbol{H_{Shannon} = -\sum_i^n p_i \ \log(p_i)} \quad (8)$$

Where: $0 \leq p_i \leq 1$; $\sum_{i=1}^{n} p_i = 1$

2.4 Regression Models

The objective of a linear regression model is to try to explain the relationship between a dependent variable (response variable) and a set of independent variables (explanatory variables) X_1, ..., X_n.

In a simple linear regression model, we try to explain the relationship between the response variable (Y) and a single explanatory variable (X). Using the regression techniques of a variable Y on a variable X, we look for a function that is a good approximation of a cloud of points (x_i, y_i), by means of a curve [42].

The variable dependency can be a univariate or multivariate regression. Univariate regression identifies the dependency between a single variable as represented in Eq. (2)[43].

$$\boldsymbol{Y = \alpha + \beta X + \varepsilon} \quad (9)$$

Where y is a dependent variable, x is an independent variable with coefficient β (it is the slope of the line and indicates how Y changes when X increases by one unit) and α is a constant (it is the ordinate at the origin, the value which Y takes when X is 0), and ε a variable that includes a large set of factors, each of which influences the response only to a small magnitude, which we will call error. X and Y are random variables, so an exact linear relationship between them cannot be established. [42]. While multivariate regression is to identify the dependence between several variables simultaneously, it is represented in Eq. (3) [43].

$$\boldsymbol{Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon} \quad (10)$$

Where ε is the error term, β_0 is the intercept, β_1-β_k are partial regression coefficients, for example, β_i when $1 \leq i \leq k$ represents the change in the mean response corresponding to a unit change in x_i when the other variables remain constant.

Regression models predict the outcome of the dependent variables from the independent variables. Importance is considered in regression analysis to handle more complicated problems [43]. The objective of multiple linear regression is to solve the set of coefficients $\Theta = \{\beta_0, \beta_1, \ldots, \beta_k\}$ given the observations X and the objectives Y [44].

2.4.1 Linear Regression

Linear regression is the most common predictive model to identify the relationship between variables. It can be simple linear or multiple linear regression. Linear regression is described in Eq. (4)[43].

$$\boldsymbol{y = x\beta + \varepsilon} \quad \textbf{(}11\textbf{)}$$

In Eq. (4) y is the independent variable and can be a continuous or categorical value; x is a dependent variable that is always a continuous value. It analyzes a probability distribution and focuses mainly on conditional probability distribution with multivariate analysis. [43].

2.4.2 Simple Linear Regression

The simple linear regression process that is depicted in Figura 4. is a regression analysis that uses a single independent variable and is described in the equation (2). Similar to how correlation expands the relationship between two variables, regression lineal simple distinguishes between dependent and independent variables; however, correlation does not do so [43].

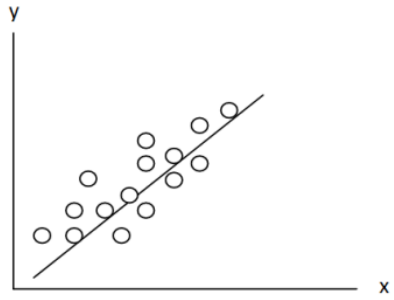


Figure 1. Simple Linear Regression *[43]*.

2.4.3 Multiple Linear Regression

Multiple or multivariate linear regression (MLR) depicted in Figura 5. is the prediction process with more than one independent or predictor variable that is similar to multivariate analysis as described in Eq. (3) [43].

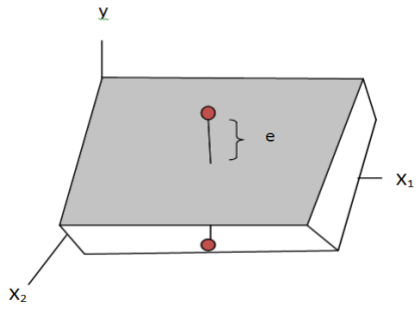


Figure 2. Multiple Linear Regression *[43]*.

A statistical technique known as multiple linear regression uses many explanatory variables to predict the outcome of a response variable. The multiple linear regression's goal is to model the relationship between the explanatory and response variables. The next model is a multiple linear regression model with k predictor variables, $x_1, \ldots, x_k$ [44].

The MLR problem is frequently resolved using least squares. If each predictor variable $x_1, x_2, \ldots, x_k$ has n observations. Then x_{ij} represents the i-th observation of the j-th predictor variable x_j.

Such as, x_{31} represents the first value of the third observation. Specifically, Eq. (3) above can be expressed as [44]:

$$\boldsymbol{y_j = \beta_0 + \beta_1 X_{j1} + \beta_2 X j_2 + \cdots + \beta_k X_{jk} + \varepsilon_j} \quad (12)$$

Where $1 \leq j \leq n$, yj is the j-th target value. The system of n equations can be represented as a design matrix as shown in Eq. (2), and describes the levels of the predictor variables acquired at each observations. All of the regression coefficients are contained in the vector β. The least squares estimates, which are stated below, are used to estimate to create the regression model β [44].

$$\boldsymbol{\widehat{\beta} = (X^T X)^{-1} X^T y} \quad (13)$$

Then the estimated value of y can be calculated as follows after obtaining $\hat{\beta}$ [44].

$$\begin{aligned} \hat{y} &= X\hat{\beta} \\ \epsilon &= y - \hat{y} \end{aligned} \quad (14)$$

The purpose of using regression data was to search for SNPs statistically associated with DM2.

2.5 Risk Factor's

A measure of the relationship between an exposure and a result is called an odds ratio (OR). The odds ratio (OR) shows the likelihood of an occurrence given a specific exposure in comparison to the likelihood of the outcome in the absence of that exposure. Case-control studies are the most frequent applications of odds ratios [46].

The odds ratio is used to compare the likelihood of an outcome (such a disease or disorder), because of exposure to a particular variable (eg, health characteristic, item of medical history). The odds ratio can also be used to assess if a specific exposure represents a risk for a specific outcome and to assess the relative importance of several risk variables for that outcome [46].

OR=1 Outcome probabilities are unaffected by exposure.

OR>1 Exposure is linked to bigger odds of success.

OR<1 Exposure is linked to a reduced likelihood of success.

It is calculated using the 95% confidence interval (CI) to determine the accuracy of the OR. A high OR precision is indicated by a small CI, while a low OR precision is shown by a large CI. It is important to note that the 95% CI does not provide information about a measure's statistical

significance, unlike the p-value. In reality, if the 95% CI does not overlap the null value (for instance, OR=1), it is frequently regarded as a marker of statistical significance. Therefore, it would be incorrect to interpret a 95% CI OR that encompasses the null as showing that exposure and outcome are not related [46].

To define risk factors, each base-pair positions found to be significant in the association analysis (regression analysis) was inspected. The Odds Ratio (OR) calculation criteria and definition of Risk Factor, as described in [46] were applied. The statistical significance, OR value, and 95% confidence range for each variable were examined based on the findings. Then, each base-pair position that satisfied the subsequent requirements was declared as a risk factor:

If the base-pair position statistical significance (p-value) was less than 0.05;

The odds ratio (OR) wasn't equal to 1; and

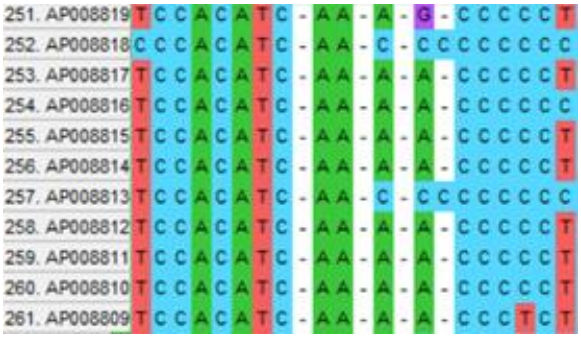
The 95% confidence range for the odds ratio did not contain 1.

Hence, if a base-pair position satisfied these three criteria and its OR > 1, it is declared as a risk factor associated with a higher probability of diabetes. In the same way, if the variable met the three conditions, and its OR < 1, it is declared as a risk factor associated with a lower probability of diabetes.

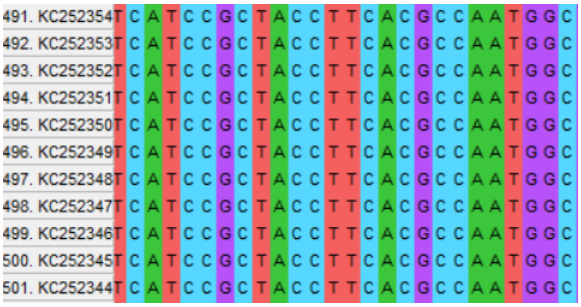
3. Results

The complete mitochondrial genomes of the 510 patients were aligned with the MAFFT tool. The result of the alignment was visualized with the MEGA X software [60]. By visual inspection, one region with variability was observed, while the rest showed perfect alignment. Figure 3 (a) shows a fragment of the region with variability, and Figure 3 (b) shows a fragment of the region without variability.

The region between positions 16,170 and 16,410, with a length of 241 nucleotides, was chosen to perform the rest of the analysis.



(a)



(b)

Figure 3. Fragment of the alignment of the mitochondrial genomes of patients with DM2 and healthy. (a) represents a region with variability and (b) represents a region with zero variability.

To analyze the structure of the information and look for possible clusters that would differentiate the data from sick and healthy people, PCA was applied to the aligned region of high variability. To carry out this analysis, the statistical language R was used. Figure 4 shows the graph of Principal Component 1 (PC1) against Principal Component 2 (PC2). As we can see in the graph, the information appears mixed and there is no clear differentiation between the groups. This analysis shows us the complexity of the data.

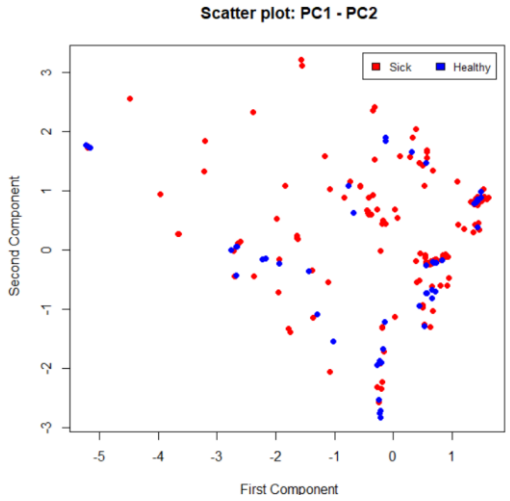


Figure 4. Comparison graph between cases vs controls.

The association analysis was performed in two steps, on the one hand, simple logistic regression was applied to each base-pair position (bp) of the variant region (241 base-pair positions), assigning a 1 to the dependent variable, for all healthy patients, and 0 to all patients with DM2. Those positions that were statistically significant (p-value <0.05) were selected. Subsequently, a multiple logistic regression was carried out grouping the positions that were significant in the simple regression. Those that were significant in the multiple regression were declared as positions associated with DM2. Table 1 shows the positions that were significant both in the simple regression and in the multiple regression.

Genomic Position (BP)	Simple Regression (P-value)	Multiple Regression (P-value)	Associated with DM2
16,184	0.0038	0.0021	Yes
16,222	0.0384	0.6592	No
16,257	0.0289	0.1037	No

16,263	0.0415	0.6937	No
16,282	0.0033	0.0064	Yes
16,289	0.0426	0.4447	No
16,344	0.0038	0.0159	Yes
16,351	0.0438	0.1983	No

Table 1. Simple and Multiple Regression Results.

4. Discussion

As observed in Table 1, after multiple regression three positions were associated with DM2. The positions and their resulting p-values were: 16,184; 16,282 and 16,344 and 0.0021; 0.0064 and 0.0159, respectively. To locate the associated gene, the human mitochondrial genome annotation in the National Center for Biotechnology Information (NCBI) database (https://www.ncbi.nlm.nih.gov/) was inspected. Three genes were located within 3,000 base pairs (bp) of the associated positions. These genes are: CYTB, which produces the Cytochrome B protein and contributes to the conversion of energy from food to cellular energy (Adenosine Triphosphate, ATP), the TRNP gene, which is the Proline tRNA, and the TRNT gene, which is the tRNA of Threonine. Especially the TRNT gene, in a study conducted by Ke Li, et al., 2020, was found to be associated with the maternal heritability of DM2 in Chinese families [63], and in another study carried out by Momiyama, et al., 2003, this gene was associated, as in our study, with the genomic position 16,184; and declared as one of the causes of left ventricular hypertrophy in patients with DM2 in Japanese families [64].

Conclusion(s)

In this association study, 510 complete mitochondrial genomes were analyzed. Of the total genomes, 437 were from patients with DM2, and 73 from healthy patients. A genome-wide alignment allowed locating a variable region in its allelic content; a PCA analysis allowed us to visualize the complexity of the data, and a logistic regression analysis allowed us to find 3 base-pair positions associated with DM2. The associated positions were located within 3k bp of three genes, one of which (TRNT gene) was reported by previous studies to be associated with DM2. Finally, this study adds new evidence of the association of genomic positions with DM2.

Acknowledgments

This work was developed within the Master's and Doctorate in Science and Engineering (MYDCI) program offered by the Autonomous University of Baja California. In addition, it was supported by a CONACYT scholarship.

References

1. Chen, L., D.J. Magliano, and P.Z. Zimmet, *The worldwide epidemiology of type 2 diabetes mellitus--present and future perspectives.* Nat Rev Endocrinol, 2011. **8**(4): p. 228-36.
2. O´Farrill, L.C.L., Cuervo, A. M. d. S., Ferrer, R. L., & Valdés, M. T. L, *Interacción genoma-ambiente en la diabetes mellitus tipo 2.* Acta Médica del Centro, 2018. **12**(4).

3. Ayo Toye, D.G., *Source: genetics and functional genomics of type 2 diabetes mellitus.* Genome Biol, 2003. **4**: p. 241.
4. health, n.i.o., *FACT SHEET - Type 2 Diabetes.* 2010.
5. Tsai, F.J., et al., *A genome-wide association study identifies susceptibility variants for type 2 diabetes in Han Chinese.* PLoS Genet, 2010. **6**(2): p. e1000847.
6. Scott, L.J., et al., *A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants.* Science, 2007. **316**(5829): p. 1341-5.
7. Saxena, R., et al., *Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels.* Science, 2007. **316**(5829): p. 1331-6.
8. Sladek, R., et al., *A genome-wide association study identifies novel risk loci for type 2 diabetes.* Nature, 2007. **445**(7130): p. 881-885.
9. Zeggini, E., et al., *Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes.* Science, 2007. **316**(5829): p. 1336-41.
10. Burton, P.R., et al., *Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls.* Nature, 2007. **447**(7145): p. 661-678.
11. Zeggini, E., et al., *Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes.* Nature Genetics, 2008. **40**(5): p. 638-645.
12. Gudmundsson, J., et al., *Two variants on chromosome 17 confer prostate cancer risk, and the one in TCF2 protects against type 2 diabetes.* Nature Genetics, 2007. **39**(8): p. 977-983.
13. Saxena, R., et al., *Large-scale gene-centric meta-analysis across 39 studies identifies type 2 diabetes loci.* Am J Hum Genet, 2012. **90**(3): p. 410-25.
14. Baum, A.E., et al., *A genome-wide association study implicates diacylglycerol kinase eta (DGKH) and several other genes in the etiology of bipolar disorder.* Molecular Psychiatry, 2008. **13**(2): p. 197-207.
15. Galvan, A., et al., *Genome-wide association study in discordant sibships identifies multiple inherited susceptibility alleles linked to lung cancer.* Carcinogenesis, 2009. **31**(3): p. 462-465.
16. Forstbauer, L.M., et al., *Genome-wide pooling approach identifies SPATA5 as a new susceptibility locus for alopecia areata.* European Journal of Human Genetics, 2012. **20**(3): p. 326-332.
17. Wong, L.P., et al., *Deep whole-genome sequencing of 100 southeast Asian Malays.* Am J Hum Genet, 2013. **92**(1): p. 52-66.
18. Kong, A., et al., *Parental origin of sequence variants associated with complex diseases.* Nature, 2009. **462**(7275): p. 868-874.
19. Voight, B.F., et al., *Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis.* Nature Genetics, 2010. **42**(7): p. 579-589.
20. Dupuis, J., et al., *New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk.* Nature Genetics, 2010. **42**(2): p. 105-116.
21. Qi, L., et al., *Genetic variants at 2q24 are associated with susceptibility to type 2 diabetes.* Human Molecular Genetics, 2010. **19**(13): p. 2706-2715.
22. Yamauchi, T., et al., *A genome-wide association study in the Japanese population identifies susceptibility loci for type 2 diabetes at UBE2E2 and C2CD4A-C2CD4B.* Nature Genetics, 2010. **42**(10): p. 864-868.

23. Shu, X.O., et al., *Identification of new genetic risk variants for type 2 diabetes.* PLoS Genet, 2010. **6**(9): p. e1001127.
24. Kooner, J.S., et al., *Genome-wide association study in individuals of South Asian ancestry identifies six new type 2 diabetes susceptibility loci.* Nature Genetics, 2011. **43**(10): p. 984-989.
25. Cho, Y.S., et al., *Meta-analysis of genome-wide association studies identifies eight new loci for type 2 diabetes in east Asians.* Nature Genetics, 2012. **44**(1): p. 67-72.
26. Rossana De Lorenzi, C.G., *Towards the first recombiant drug.* ELLS – European Learning Laboratory for the Life Sciences.
27. J.C. Wiebe, A.M.W., F.J. Novoa Mogollón, *Genética de la diabetes mellitus.* Revista Nefrología, 2011. **2**(1): p. 1-119.
28. M.J. Picón, F.J.T., *Factores genéticos frente a factores ambientales en el desarrollo de la diabetes tipo 2.* Avances en Diabetología, 2010. **26**: p. 268-269
29. Esparza-Castro D, A.-A.F., Merelo-Arias CA,Cruz M, Valladares-Salgado A, *scaneo genómico completo en diabetes tipo 2 y su aplicación clínica.* Revista Medica Instituto Mexicano Seguro Social, 2015. **53**(5): p. 592-599.
30. Mauricio Hernández-Ávila, J.P.G., Nancy Reynoso-Noverón., *Diabetes mellitus en México. El estado de la epidemia.* Salud pública Méxicana, 2013. **55**.
31. García-Chapa, E.G., et al., *Genetic Epidemiology of Type 2 Diabetes in Mexican Mestizos.* BioMed research international, 2017. **2017**: p. 3937893-3937893.
32. Cruz, M., et al., *Candidate gene association study conditioning on individual ancestry in patients with type 2 diabetes and metabolic syndrome from Mexico City.* Diabetes Metab Res Rev, 2010. **26**(4): p. 261-70.
33. Gamboa-Meléndez, M.A., et al., *Contribution of common genetic variation to the risk of type 2 diabetes in the Mexican Mestizo population.* Diabetes, 2012. **61**(12): p. 3314-21.
34. Katoh, R., Yamada *Servicio en línea MAFFT: alineación de secuencias múltiples, elección y visualización interactivas de secuencias*. 2019 [cited 2020 junio]; Available from: https://mafft.cbrc.jp/alignment/software/algorithms/algorithms.html.
35. Kuraku, S., et al., *aLeaves facilitates on-demand exploration of metazoan gene family trees on MAFFT sequence alignment server with enhanced interactivity.* Nucleic Acids Research, 2013. **41**(W1): p. W22-W28.
36. Katoh, K., J. Rozewicki, and K.D. Yamada, *MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization.* Briefings in Bioinformatics, 2017. **20**(4): p. 1160-1166.
37. Konishi, T., et al., *Principal Component Analysis applied directly to Sequence Matrix.* Scientific Reports, 2019. **9**(1): p. 19297.
38. Mateos-Valenzuela, A.G., et al., *Risk factors and association of body composition components for lumbar disc herniation in Northwest, Mexico.* Scientific Reports, 2020. **10**(1): p. 18479.
39. Rodrigo, J.A. *Análisis de Componentes Principales (Principal Component Analysis, PCA) y t-SNE*. 2017 [cited 2020 Junio]; Available from: https://rpubs.com/Joaquin_AR/287787.
40. Delgado, A., A. Huamani, and B. Brillitt. *Applying Shannon Entropy to Analise Health System Level by departments in Peru*. in *2018 IEEE XXV International Conference on Electronics, Electrical Engineering and Computing (INTERCON)*. 2018.
41. *Shannon Diversity Index - The Diversity Function in R - ProgrammingR*. 2024.

42. Limeres, C.C. *REGRESIÓN LINEAL SIMPLE*. 2011 [cited 2022 19/Enero]; Available from: http://eio.usc.es/eipc1/BASE/BASEMASTER/FORMULARIOS-PHP-DPTO/MATERIALES/Mat_50140116_Regr_%20simple_2011_12.pdf.
43. Kavitha, S., S. Varuna, and R. Ramya. *A comparative analysis on linear regression and support vector regression*. in *2016 Online International Conference on Green Engineering and Technologies (IC-GET)*. 2016.
44. Zhang, Z., et al. *Multiple Linear Regression for High Efficiency Video Intra Coding*. in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2019.
45. *RPubs - Comparaciones múltiples: corrección de p-value y FDR*. 2024.
46. Szumilas, M., *Explaining odds ratios.* Journal of the Canadian Academy of Child and Adolescent Psychiatry = Journal de l'Academie canadienne de psychiatrie de l'enfant et de l'adolescent, 2010. **19**(3): p. 227-229.
47. Sladek, R., et al., *A genome-wide association study identifies novel risk loci for type 2 diabetes.* Nature, 2007. **445**(7130): p. 881-5.
48. Steinthorsdottir, V., et al., *A variant in CDKAL1 influences insulin response and risk of type 2 diabetes.* Nat Genet, 2007. **39**(6): p. 770-5.
49. Bouatia-Naji, N., et al., *A polymorphism within the G6PC2 gene is associated with fasting plasma glucose levels.* Science, 2008. **320**(5879): p. 1085-8.
50. Voight, B.F., et al., *Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis.* Nat Genet, 2010. **42**(7): p. 579-89.
51. Dupuis, J., et al., *New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk.* Nat Genet, 2010. **42**(2): p. 105-16.
52. So, H.C., et al., *Evaluating the heritability explained by known susceptibility variants: a survey of ten complex diseases.* Genet Epidemiol, 2011. **35**(5): p. 310-7.
53. Yasuda, K., et al., *Variants in KCNQ1 are associated with susceptibility to type 2 diabetes mellitus.* Nat Genet, 2008. **40**(9): p. 1092-7.
54. Unoki, H., et al., *SNPs in KCNQ1 are associated with susceptibility to type 2 diabetes in East Asian and European populations.* Nat Genet, 2008. **40**(9): p. 1098-102.
55. Yamauchi, T., et al., *A genome-wide association study in the Japanese population identifies susceptibility loci for type 2 diabetes at UBE2E2 and C2CD4A-C2CD4B.* Nat Genet, 2010. **42**(10): p. 864-8.
56. Kooner, J.S., et al., *Genome-wide association study in individuals of South Asian ancestry identifies six new type 2 diabetes susceptibility loci.* Nat Genet, 2011. **43**(10): p. 984-9.
57. Lettre, G., et al., *Genome-wide association study of coronary heart disease and its risk factors in 8,090 African Americans: the NHLBI CARe Project.* PLoS Genet, 2011. **7**(2): p. e1001300.
58. projects, C.t.W., *Sequence logo - Wikipedia*. 2023.
59. *How to determine the height/bits in a sequence logo?* 2024.
60. Kumar, S., et al., *MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms.* Mol Biol Evol, 2018. **35**(6): p. 1547-1549.
61. Team, R.C., *R: A Language and Environment for Statistical Computing*. 2020.
62. Warde-Farley, D., et al., *The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function.* Nucleic Acids Research, 2010. **38**(suppl_2): p. W214-W220.

63. Li, K., et al., *Maternally Inherited Diabetes Mellitus Associated with a Novel m.15897G>A Mutation in Mitochondrial tRNA(Thr) Gene.* J Diabetes Res, 2020. **2020**: p. 2057187.
64. Momiyama, Y., et al., *A mitochondrial DNA variant associated with left ventricular hypertrophy in diabetes.* Biochem Biophys Res Commun, 2003. **312**(3): p. 858-64.
65. *Regresion Lineal Multiple*. 2011 [cited 2022 19/Enero]; Available from: http://eio.usc.es/eipc1/BASE/BASEMASTER/FORMULARIOS-PHP-DPTO/MATERIALES/Mat_50140128_RegresionMultiple.pdf.
66. Rojas de P, E., R. Molina, and C. Rodríguez, *DEFINICIÓN, CLASIFICACIÓN Y DIAGNÓSTICO DE LA DIABETES MELLITUS.* Revista Venezolana de Endocrinología y Metabolismo, 2012. **10**(1): p. 7-12.
67. Mediavilla Bravo, J.J., *la diabetes mellitus tipo 2.* Medicina Integral.
68. Cowie, C.C., et al., *Prevalence of diabetes and high risk for diabetes using A1C criteria in the U.S. population in 1988-2006.* Diabetes Care, 2010. **33**(3): p. 562-8.
69. Díaz-Apodaca, B.A., et al., *Prevalence of type 2 diabetes and impaired fasting glucose: cross-sectional study of multiethnic adult population at the United States-Mexico border.* Rev Panam Salud Publica, 2010. **28**(3): p. 174-81.
70. Lee, J.W., F.L. Brancati, and H.C. Yeh, *Trends in the prevalence of type 2 diabetes in Asians versus whites: results from the United States National Health Interview Survey, 1997-2008.* Diabetes Care, 2011. **34**(2): p. 353-7.
71. Bowden, D.W., et al., *Review of the Diabetes Heart Study (DHS) family of studies: a comprehensively examined sample for genetic and epidemiological studies of type 2 diabetes and its complications.* Rev Diabet Stud, 2010. **7**(3): p. 188-201.
72. Altshuler, D., et al., *The common PPARgamma Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes.* Nat Genet, 2000. **26**(1): p. 76-80.
73. Gloyn, A.L., et al., *Association studies of variants in promoter and coding regions of beta-cell ATP-sensitive K-channel genes SUR1 and Kir6.2 with Type 2 diabetes mellitus (UKPDS 53).* Diabet Med, 2001. **18**(3): p. 206-12.
74. Sandhu, M.S., et al., *Common variants in WFS1 confer risk of type 2 diabetes.* Nat Genet, 2007. **39**(8): p. 951-3.
75. Winckler, W., et al., *Evaluation of common variants in the six known maturity-onset diabetes of the young (MODY) genes for association with type 2 diabetes.* Diabetes, 2007. **56**(3): p. 685-93.
76. Winckler, W., et al., *Association of common variation in the HNF1alpha gene region with risk of type 2 diabetes.* Diabetes, 2005. **54**(8): p. 2336-42.
77. Winckler, W., et al., *Association testing of variants in the hepatocyte nuclear factor 4alpha gene with risk of type 2 diabetes in 7,883 people.* Diabetes, 2005. **54**(3): p. 886-92.
78. Larkin, M.A., et al., *Clustal W and Clustal X version 2.0.* Bioinformatics, 2007. **23**(21): p. 2947-8.
79. Rosner, B., *Fundamentals of biostatistics*. 2011: Seventh edition. Boston : Brooks/Cole, Cengage Learning, [2011] ©2011.
80. Warde-Farley, D., et al., *The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function.* Nucleic Acids Res, 2010. **38**(Web Server issue): p. W214-20.

Articulo
Article

Logistic regression in human mitochondrial genome sequences identifies two novel variants associated with Type 2 Diabetes

J. A. Valdez [1,2,*], P. Mahadevan[3], P. Mayorga[2] , C. Villa-Angulo[1], Raúl C. Baptista Rosas[4,5,6], Arieh Roldan Mercado-Sesma[7], Felipe de Jesus Orozco Luna[8], R. Villa-Angulo[1*]

Citation: To be added by editorial staff during production.

Academic Editor: Firstname Lastname

Received: date

Revised: date

Accepted: date

Published: date

[1] Instituto de Ingeniería, Universidad Autónoma de Baja California, Mexicali, B.C., México; a1174500@uabc.edu.mx

[2] Departamento de Eléctrica - Electrónica, Tecnológico Nacional de México/IT de Mexicali, Mexicali, B.C., México; jvaldezgonzalez@itmexicali.edu.mx

[3] Department of Biology, University of Tampa, Tampa, FL, United States of America (pmahadevan@ut.edu).

[4] Department of Health Sciences-Disease as an individual process, Centro Universitario de Tonalá,

Universidad de Guadalajara, México. raul.baptista@academicos.udg.mx

[5] Centro de Investigación Multidisciplinaria en Salud, Universidad de Guadalajara.

[6] Hospital General de Occidente, Secretaría de Salud Jalisco, México.

[7] Departamento de Ciencias de la Salud-Enfermedad como proceso individual, Centro Universitario de Tonalá, Universidad de Guadalajara, México.

[8] Centro de Análisis de Datos y Supercómputo / Universidad de Guadalajara, México.

* Correspondence: a1174500@uabc.edu.mx; rafael.villa@uabc.edu.mx; (J. A. Valdez & R. Villa-Angulo)

Abstract: Type 2 Diabetes (T2D) is a complex and multifactorial disorder. It is currently classified as one of the most significant epidemics of the 21st century due to the magnitude with which it is growing, as well as the cardiovascular effects it causes. This work presents a complete genome analysis for association of single base-pair positions of the human mitochondrial genome with T2D. The data consisted of 510 complete mitochondrial genomes, of which 437 were from patients with T2D and 73 were from control patients. First, multiple alignment was performed which allowed visualization and selection of a region with allelic variation. Then, a principal components analysis was used to visualize the structure of the data. Next, logistic regression analysis was performed to discover 3 base-pair positions associated with T2D. An Odds Ratio analysis revealed three associated positions as risk factors for T2D. An inspection of the mitochondrial genome

annotation identified three genes CYTB, TRNP and TRNT as associated with T2D. The gene TRNT was previously reported as associated, while the genes CYTB and TRNP have been identified in this work as associated with T2D for the first time. This study provides new evidence of association with two novel genes, which satisfy the criteria to be risk factors of T2D.

Keywords: Type 2 Diabetes (T2D), Mitochondrial Genome, mtDNA, Logistic Regression, Machine learning.

1. Introduction

Type 2 diabetes (T2D) is a complex and multifactorial disorder characterized by chronic hyperglycemia due to the interaction of multiple genetic variants and various environmental factors. As a result of the aging of the population and the increasing prevalence of obesity and physical inactivity, the number of patients with T2D has increased markedly throughout the world [1]. It is classified along with obesity, as one of the epidemics of the 21st century, both for its growing magnitude and for its negative impact on cardiovascular diseases [2]. The environment refers to all the non-genetic factors that modulate the phenotype and can include factors of the random environment (climatic, geographic, demographic, and socioeconomic) as well as lifestyle factors (diet, smoking, alcoholism, and physical activity), which the individual can modify [2]. Its heritability is estimated at 22% to 73% from twin and family studies. The age-adjusted prevalence of T2D in adults was recently estimated to be 7.6% in European Americans, 14.9% in non-Hispanic African Americans, 4.3% to 8.2% in Asian Americans, and 10.9% to 15.6% in Hispanics **[68-71]**.

T2D is a heterogeneous disease of multifactorial etiology, in which insulin resistance and inadequate compensatory insulin secretion by pancreatic beta cells are combined. It manifests as chronic hyperglycemia, accompanied by disorders of carbohydrate, fat, and protein metabolism [2]. The disease is considered a polygenic disorder, in which each genetic variant confers a partial and additive effect. Only 5-10% of T2D cases are due to defects in a single gene; these include maturity-onset diabetes of the young, insulin resistance syndromes, mitochondrial diabetes, and neonatal diabetes **[5]**. Examination of T2D susceptibility genes may be useful for the prediction, prevention, and early treatment of the disease.

Through the implementation of genome-wide association studies (GWAS), the number of common genetic variants associated with T2D has increased rapidly. The first GWAS studies for T2D **[6, 7, 9, 47, 48]** and fasting glucose **[49]** successfully identified more than 40 genetic loci associated with T2D. All these studies were carried out mainly in European populations **[13]**. Recently, through a GWAS meta-analysis study of T2D **[50]** and glycemic quantitative traits **[51]**, the number of genome-wide significant T2D-associated loci in European populations dramatically increased. Most of these variants act through defects in beta cell function rather than insulin action. Together, the variants known to be associated with T2D explain ~10% of the genetic variation **[50, 52]**, which indicates that additional loci and independent signals in known loci might be contributing to the disease. This large-scale gene-focused meta-analysis was performed on 39 multi-ethnic T2D association studies. It identified three European T2D risk loci (GATAD2A/CILP2/PBX4, previously known to have protective effects on lipids; TH/INS, previously known to have protective effects on T1D and SREBF1), an African-American T2D risk locus (HMGA2) and a multi-ethnic risk locus (BCL2). It also confirmed that a genetic score of T2D risk alleles influences T2D risk in multiethnic populations, including African American,

Hispanic, and Asian. Therefore, well-powered, multi-ethnic GWAS of T2D should lead to the discovery of additional diabetes-associated genes relevant to multiple ethnic groups **[13]**.

The genetic factors that contribute to T2D are less well understood in non-European populations. A novel locus (KCNQ1 [MIM 607542]) was identified based on a GWAS in a Japanese population **[53, 54]** and later it was shown that it harbors independent alleles in individuals of European descent **[50]**. Moreover, GWAS recently performed in Chinese **[5, 23]**, Japanese **[55]**, and South Asian **[56]** populations described additional T2D loci that exceed genome-wide significance. To date, T2D GWAS in African Americans has been underpowered to detect new loci. **[57]**.

The purpose of this work was identify variants in the mitochondrial genome looking for novel base pair (bp) positions statistically associated with T2D. An alignment analysis allowed visualization and selection of a genomic region with high allelic variability. An entropy analysis was used to verify that the selected genomic region was more highly informative than the rest of the genome. A Principal Component Analysis (PCA) was used to visualize the complexity of the data. Simple and multiple logistic regression analysis enabled the discovery of base-pair positions associated with T2D. Risk factors were also searched for using the Odds Ratio criteria. Finally, an inspection of the mitochondrial genome annotation revealed 3 candidate genes to be associated with T2D.

2. Materials and Methods

2.1 Database

A total of 510 complete human mitochondrial genomes were analyzed, from which 437 were from individuals with T2D, and 73 from healthy individuals. The sequences were located and downloaded from the nucleotide database of the National Center for Biotechnology Information (NCBI) (https://www.ncbi.nlm.nih.gov/nucleotide). The supplementary material (SP) contains the list of genome IDs. The individuals were from different nationalities. Genome samples were distributed as follows: 239 individuals were from Thailand (individuals with ID beginning with the letters KC); 62 individuals from India (individuals with ID beginning with the letters QH); 11 individuals from China (individuals with ID beginning with the letters KF); 192 individuals from Japan (one with ID beginning with the letters AB and 191 with AP); and 6 individuals from Italy (individuals with ID beginning with the letters JF). The data was stored in a FASTA format file and the genomes were aligned using the ClustalW algorithm implemented in the MAFFT tool (https://mafft.cbrc.jp/alignment/server/) **[35, 36]**. The total length of the alignment was 16,610 nucleotides. To carry out the data analysis, the nucleotides were replaced with numerical values, as follows: A = 1, C = 2, G = 3 and T = 4, GAP(-) = 5. In addition, we replaced the following residues: R (GA, purine), Y (TC, pyrimidine), W (AT, week bonds), and N (ACGT, any), with a value of 9. Supplementary Table S1 presents the list of individual sequences. The first column contains the list of individual ID sequences. The second column contains the disease condition with 0 indicating the individual is affected by T2D and 1 indicating no affection (control). The third column contains the individual nationality.

2.2 Entropy Analysis

By inspecting the 510 genome multiple sequence alignment produced by the *ClustalW* algorithm **[78]**, we identified a region that clearly presented greater variability than the rest. This region

comprises base-pair positions 16,170 to 16,410, with a length of 241 base pairs. This region was named as *Variant Region* (R_v). Then, in order to prove if R_v offered more information than the rest of the alignment, an entropy analysis was performed **[40]**. The complete alignment was traversed by taking contiguous regions of 241 nucleotides (the same length of R_v), and their entropy was obtained with the Shannon formula, as follows:

For each base-pair position in the selected region, the entropy of each residue was obtained, with the formula of Equation 1.

$$\boldsymbol{H_{Shannon} = -\sum_i^n p_i \log(p_i)} \quad (15)$$

Where: $\boldsymbol{0 \leq p_i \leq 1}$; $\boldsymbol{\sum_{i=1}^{n} p_i = 1}$

The sum of the entropy of each residue is the entropy of the base-pair position; and the sum of the entropies of all base-pair positions within a region is the total entropy of the region. It resulted in a vector of 68 region entropies, called V_r, excluding R_v . Then we implemented a statistical *t-test* for equality of means, as follows:

Null hypothesis (H_0) : $\boldsymbol{\bar{x}} = \mu_0$

Alternative hypothesis (H_1) : $\boldsymbol{\bar{x}} \neq \mu_0$

The t-statistic was calculated as follows: $\boldsymbol{t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}}$

Where: μ_0 = Entropy of R_v

$\boldsymbol{\bar{x}}$ = Mean of V_r.

s = Standard Deviation of V_r.

n = Size of V_r.

Once the *t* statistic was computed, a *p-value* of statistical significance was estimated as follows:

$$\boldsymbol{p\ value = \begin{cases} 2 * Pr(t_{n-1} \leq t), si\ t \leq 0 \\ 2 * [1 - Pr(t_{n-1} \leq t)], si\ t > 0 \end{cases}}$$

Where: t_{n-1} = Student's distribution with n-1 degrees of freedom.

A *p-value* = $2.2e^{-16}$ was obtained, so the region R_v was statistically *very highly significantly* more informative than the rest of the genome. The rest of the analysis was performed using R_v.

2.3 Principal Component Analysis (PCA)

The central idea of principal component analysis (PCA) is to reduce the dimensionality of a data set, which consists of a large number of interrelated variables, while retaining as much of the variation as possible that is present in the data set. This is achieved by transforming a new set of variables, the principal components (PCs), which are not correlated, and are arranged, in such a way that the former retain the greatest variation present in all the original variables **[38]**.

PCA is theoretically the optimal transformation for a given data set, in terms of least squares. The procedure to obtain the principal components can be summarized as follows: Given a vector X^T of n dimensions, $X = [x_1, x_2, \ldots, x_n]^T$, whose mean vectors (M), and covariance (C), are described by: $M=E(X)= [m_1, m_2, \ldots, m_n]T$ and $C = E [(X - M) (X - M)^T]$. Calculate the eigenvalues $\lambda 1, \lambda 2, \ldots, \lambda n$ and the eigenvectors $P_1, P_2, \ldots, Pn$; and order them according to their magnitude $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$. Select d eigenvectors to represent the n variables, $d < n$. Then $P_1, P_2, \ldots, P_d$ are called principal components **[38]**.

The purpose of applying the PCA analysis was to analyze the structure of the data looking for possible clusters that differentiated the sick and healthy individuals. Supplementary Figure S2 shows the scatter plot of Principal Component 1 (PC1) against Principal Component 2 (PC2). We can see that T2D and control appear mixed indicating a high complicity of data.

2.4 Regression Models

In general, Regression Models are mathematical methods to model the quantitative stochastic relationship between a variable of interest and one (or a set) of explanatory variables **[42, 43]**. Specifically, these models can be expressed as follows:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_p X_{pi} + \varepsilon_i \quad (16)$$

where:

Y_i: variable of interest, dependent or returning,

$X_{1i}, X_{2i}, \ldots, X_{pi}$: explanatory, independent, or regressor variables,

β_0: intersection or constant term,

$\beta_1, \beta_2, \ldots, \beta_p$: parameters, measure the influence that the explanatory variables have on the regressing,

p: number of independent parameters to take into account,

ε: observation error due to uncontrolled variables,

i: *1, 2, ..., n* number of observations of the variables.

With these models, it is possible to study linear relationships between multiple variables and the effect they have on the dependent variable. The β_i are estimated following the least squares criterion:

$$\min_{\substack{\beta \in n \\ i=1,\ldots,n}} \sum_{i=1}^{n} \left(Y_i - \beta_0 - \beta_1 X_{1i} - \beta_2 X_{2i} - \ldots - \beta_p X_{pi}\right)^2 \quad (17)$$

and the least squares estimators are obtained from the equation:

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X^T X})^{-1} \boldsymbol{X^T y}$$

(18)

In particular, logistic regression is a type of regression analysis used to predict the outcome of a categorical variable (a variable that can take on a limited number of categories) based on the independent or predictor variables. It is useful for modeling the probability of an event occurring as a function of other factors. The logistic regression analysis is part of the set of Generalized Linear Models (GLM) that uses the logit function as a link function. The probabilities describing the possible outcome of a single trial are modeled, as a function of explanatory variables, using a logistic function. The regression model can be expressed as follows **[79]**:

$$\log it(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_p X_{pi} + \varepsilon_i \quad (19)$$

where:

$$p_i = E\left(\frac{Y_i}{n_i} | X_i\right)$$

(20)

The purpose of using regression on the data was to search for base-pair positions statistically associated with T2D.

2.5 Risk factor's definition

An Odds Ratio (OR) is a measure of the association between an exposure and an outcome. The OR represents the odds that an outcome will occur given a particular exposure, compared to the odds that the outcome will occur in the absence of that exposure. Odds ratios are most commonly used in case-control studies **[46]**.

The odds ratio is used to compare the relative probabilities that the outcome of interest (e.g., disease or disorder) will occur, given exposure to the variable of interest (e.g., health characteristic, item of medical history). The odds ratio can also be used to determine whether a particular exposure is a risk factor for a particular outcome and to compare the magnitude of various risk factors for that outcome **[46]**.

OR=1 The exposure does not affect the probabilities of the result.

OR>1 Exposure associated with higher probabilities of outcome.

OR<1 Exposure associated with lower probabilities of outcome.

The 95% confidence interval (CI) is used to estimate the precision of the OR. A large CI indicates a low level of OR precision, while a small CI indicates a higher OR precision. However, it is important to note that, unlike the p-value, the 95% CI does not report the statistical significance of a measure. In practice, the 95% CI is often used as an indicator of the presence of statistical significance if it does not overlap the null value (for example, OR=1). However, it would be inappropriate to interpret a 95% CI OR encompassing the null as indicating a lack of association between exposure and outcome **[46]**.

To define risk factors, all base-pair positions found to be significant in the association analysis (regression analysis) were inspected. The Odds Ratio (OR) calculation criteria and definition of Risk Factor, as described in **[46]** was applied. From the results, the statistical significance, the OR value, and the 95% OR confidence interval for each variable were inspected. All the positions that met the following criteria were declared as risk factors:

If the variable was statistically significant (p-value ≤ 0.05);

Odds Ratio (OR) was different from 1; and

The odds ratio confidence interval (95% CI) did not include the 1.

Then, if an associated base-pair position meets these three conditions and its OR > 1, it is declared as a risk factor associated with a higher probability of T2D. In the same way, if the variable met the three conditions, and its OR < 1, it is declared as a risk factor associated with a lower probability of T2D.

2.6 Gene interaction prediction

GeneMANIA (http://www.genemania.org) is a flexible, user-friendly web interface for generating hypotheses about gene function, analyzing gene lists and prioritizing genes for functional assays. Given a query list, GeneMANIA extends the list with functionally similar genes that it identifies using available genomics and proteomics data. GeneMANIA also reports weights that indicate the predictive value of each selected data set for the query. Hundreds of data sets have been collected from GEO, BioGRID, Pathway Commons and I2D, as well as organism-specific functional genomics data sets. Users can select arbitrary subsets of the data sets associated with an organism to perform their analyses and can upload their own data sets to analyze. The GeneMANIA algorithm performs as well or better than other gene function prediction methods on yeast and mouse benchmarks. The high accuracy of the GeneMANIA prediction algorithm, an intuitive user interface and large database make GeneMANIA a useful tool for any biologist **[80]**.

The input to GeneMANIA is simple—the user enters a list of genes and, optionally, selects from a list of data sets that they wish to query. GeneMANIA then extends the user's list with genes that are functionally similar, or have shared properties with the initial query genes, and displays an interactive functional association network, illustrating the relationships among the genes and data sets. If the query genes are involved in disease, such as a mouse leukemia model, OMIM and phenotype data sets may receive high weight and GeneMANIA will output genes that likely are involved in the same process **[80]**.

Data sets are collected from publicly available databases, including co-expression data from Gene Expression Omnibus (GEO); physical and genetic interaction data from BioGRID; predicted protein interaction data based on orthology from I2D; and pathway and molecular interaction data from Pathway Commons, which contains data from BioGRID, Memorial Sloan-Kettering Cancer Center, Human Protein Reference Database, HumanCyc, Systems Biology Center New York, IntAct, MINT, NCI-Nature Pathway Interaction Database and Reactome. Individual data sets relevant to specific organisms are also collected, such as protein sub-cellular localization in yeast. Networks are produced from the data either directly (as in the case of protein or genetic

interactions) or using an in-house analysis pipeline to convert profiles to functional association networks **[80]**.

3. Results

The 510 mitochondrial genomes were aligned using the *ClustalW* algorithm implemented in the *MAFFT* tool [35, 36]. The alignment was then visually inspected with the *MEGA X* software [60]. A region of 241 base pair long that clearly presented greater variability than the rest of the alignment was identified. This region ranged from base-pair position 16,170 to 16,410. We named this region as R_v. Thereafter, in order to estimate the degree of information provided by R_v, the Shannon entropy [40] was calculated for the entire region; first calculating the entropy for each individual base-pair and then adding the individual entropies over the entire R_v region. The resulting R_v entropy was 16.83817. Subsequently, the rest of the genome was interrogated by taking contiguous regions of the same size as R_v, (not including R_v), and estimates of entropy were computed for each region. In total, 68 regions (not including R_v) with an average entropy of 1.64698 were found. A statistical *t-test* for equality of means was used to verify the statistical significance of the entropy of R_v, compared to the rest of the genome. The test resulted in a *p-value* = $2.2e^{-16}$, indicating that the entropy of R_v is very highly significantly larger than the rest of the genome. We then selected R_v to perform the rest of the analysis.

The first analysis in R_v region was to visualize the structure of the data looking for clusters of information that could differentiate between sick and healthy individuals. For this, a Principal Component Analysis was applied [38]. To carry out this analysis, the statistical language R was used [61]. Figure S2 (see Supplementary Figure S2) shows a plot of the Principal Component 1 (PC1) against the Principal Component 2 (PC2). From the plot we can notice that data appears highly mixed, which means that there is no clear differentiation between sick and healthy individuals. Therefore, in order to perform the association analysis, it would be necessary to select a technique capable of analyzing highly mixed data.

In order to find the association of base-pair positions with T2D, logistic regression models were used [42, 43]. The analysis was carried out in two steps. First, all individuals with T2D were labeled with 0, and all healthy individuals were labeled with 1, for the dependent variable of the model. A simple logistic regression was then applied to each individual base-pair within R_v. This yielded 8 statistically significant positions (p-value <0.05). Next, the 8 significant base-pair positions were selected and a Multiple Logistic Regression was applied. The resulting base pair positions that were significant from the multiple regression were declared as associated with T2D. Table 1 shows the results. The first column shows the base pair position. The Second column contains the p-values from the simple regression. The third column shows the p-values resulting from the multiple regression. The fourth column contains the status of association with T2D.

Table 2. Simple and multiple logistic regression results. The simple and multiple regression identified the base-pair position loci 16,184, 16,282 and 16,344 as statistically associated to T2D.

Genomic Position (BP)	Simple Regression (p-value)	Multiple Regression (p-value)	Associated with T2D
16,184	0.0038	0.0021	Yes
16,222	0.0384	0.6592	No
16,257	0.0289	0.1037	No
16,263	0.0415	0.6937	No
16,282	0.0033	0.0064	Yes
16,289	0.0426	0.4447	No
16,344	0.0038	0.0159	Yes
16,351	0.0438	0.1983	No

To define whether the associated base-pair positions are a risk factor, the criteria proposed by Szumila, 2010 [46], which uses the Odds Ratio, was applied. We defined a base-pair position as a risk factor if it met the following three conditions: 1) the p-value of the statistical test for the OR was ≤ 0.05; 2) the OR was different from 1; and 3) the Confidence Interval (CI) did not include the value 1. If a base-pair found to be associated meets these three conditions, and it has OR > 1, then it was declared as a risk factor associated to a high chance of T2D. Now, if the variable meets the three conditions, and it has OR<1, then it is declared as a risk factor associated to a small chance of T2D. Table 2 shows the results. The first column contains the base-pair position. The second column contains the odds ratio and its 95% CI. The third column contains the p-value. The fourth column contains the Risk Factor statement.

Table 3. Type 2 diabetes risk factors. Genomic position, P-value and Odds ratio with Confidence Interval at 95% for variants associated with high and small chances of T2D.

Mitochondrial Genomic base pair position	Odds Ratio (95% CI)	P-Value	Risk Factor
16,184	4.301 (1.759 – 12.355)	0.00215	Yes, associated with high chances of T2D.
16,282	7.385 (2.014 – 41.343)	0.00641	Yes, associated with high chances of T2D.
16,344	0.580 (0.371 – 0.906)	0.01598	Yes, associated with small chances of T2D.

In order to locate the genes related to the associated base pair positions, the human mitochondrial genome annotation from the NCBI *Nuclotide* database (https://www.ncbi.nlm.nih.gov/) was inspected. Three genes were located within 3,000 base pairs from the associated positions. These genes are CYTB (also known as MT-CYB), which produces the Cytochrome B protein and contributes to the conversion of energy from food to cellular energy (Adenosine Triphosphate, ATP), the TRNP (also known as TRNP1) gene, which is the Proline tRNA, and the TRNT (also known as MTTT) gene, which is the tRNA of Threonine. Then, we interrogated through the *Genemania* online tool (https://genemania.org/), whether these genes interact. **¡Error! No se encuentra el origen de la referencia.** shows the resulting graph of interactions between genes. We can notice that gene TRNT was not found by *Genemania*, and genes CYBT and TRNT show no known interaction. Supplementary Figures S3, S4 and S5 present the *Genemania* reports for the three genes, the CYTB gene, and TRNP gene, respectively.

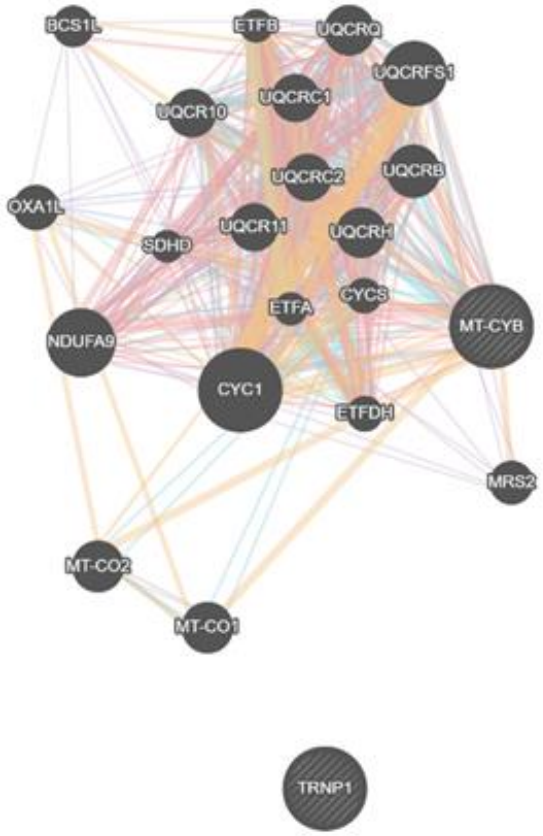


Figure 5. *Genemania* interactions graph for genes MT-CYB, TRNP1, and TRNT. Gene TRNT was not found. And genes MT-CYB and TRNP1 show no interactions between them.

4. Discussion

The alignment of the 510 initial genome sequences permitted to visualize and select a 241 base pair region (R_v) with clearly more variability than the rest of the genome. However, the one sample *t test* implemented to compare the entropy of R_v against the mean of the vector of entropies obtained from all regions of the same size of R_v permitted to demonstrate in a systemic way that

the selected R_v is statistically the most variant, and as a consequences, the most informative region in T2D mitochondrial genome. Then, entropy analysis offers a well-suited criterion for selecting variant regions for specific analysis.

After multiple logistic regression, three positions were associated with T2D. The associated positions were 16,184, 16,282 and 16,344 with a *p*-value of 0.0021, 0.0064 and 0.0159, respectively. Then, from the three genes found within 3,000 base pairs from the associated positions, TRNT was previously reported by Ke Li, et al., 2020 as being associated with maternal heritability of T2D in Chinese families [63]. In another study carried out by Momiyama, et al., 2003 the same gene was also associated, as in our study, with the genomic position 16,184; and declared as one of the causes of left ventricular hypertrophy in patients with T2D in Japanese families [64]. However, the CYTB and TRNP genes have not previously been associated with T2D.

5. Conclusions

In this association study, 510 complete human mitochondrial genomes were analyzed, from which 437 were from individuals with T2D, and 73 from healthy individuals. Selecting a 241 base-pair region that presented clear variability in the alignment of complete genomes allowed us to carry out a study focused on the variant region of the genome. Three base air positions resulted associated after a logistic regression analysis. Then, by applying the Odds Ration criterion the three associated positons were declares as risk factors for T2D. Next, by inspecting the human mitochondrial genome annotation the associated positions were located less than 3kb from the genes CYTB, which produces the Cytochrome B protein and contributes to ATP production, the gene TRNP, which is the tRNA of Proline, and the gene TRNT, which is the tRNA of Threonine. Of these three genes, TRNT was reported by previous studies as associated with T2D, while the CYTB and TRNP genes are associated with T2D for the first time in this study. The interaction prediction analysis performed with Genemania showed that there is an interaction between the CYTB and TRNP genes. Finally, this association study provides new evidence of association proposing two novel variants statistically associated with T2D that meet the criteria of being a risk factor for T2D.

Supplementary Materials: The following supporting information can be downloaded at: www.mdpi.com/xxx/s1, Figure S2: Plot of PC0 vs PC1 for the data of the Region R_v. Individuals appears highly mixed, indicating the high complexity of data; Table S1: List of the 510 Genome IDs downloaded from NCBI: (**https://www.ncbi.nlm.nih.gov/nucleotide**). Supplementary S3: *Genemania* interaction report including the three associated genes. Supplementary S4: *Genemania* interaction report for the gene CYTB. Supplementary S5: *Genemania* interaction report for the gene TRNP.

Author Contributions: Conceptualization, methodology, data curation, formal analysis and writing—original draft preparation, J.A.V. and R.V.A; supervision, writing—review and editing, C.V.A., P.M., P.M., R.C.B.R., A.R.M.S. and F.J.O.L. All authors have read and agreed to the published version of the manuscript.

Acknowledgments: We are grateful to the National Council of Science and Technology of México for supporting the Ph.D studies scholarship for J.A. Valdez.

References

1. Chen, L., D.J. Magliano, and P.Z. Zimmet, *The worldwide epidemiology of type 2 diabetes mellitus--present and future perspectives.* Nat Rev Endocrinol, 2011. **8**(4): p. 228-36.

2. O´Farrill, L.C.L., Cuervo, A. M. d. S., Ferrer, R. L., & Valdés, M. T. L, *Interacción genoma-ambiente en la diabetes mellitus tipo 2.* Acta Médica del Centro, 2018. **12**(4).

3. Cowie, C.C., et al., *Prevalence of diabetes and high risk for diabetes using A1C criteria in the U.S. population in 1988-2006.* Diabetes Care, 2010. **33**(3): p. 562-8.

4. Díaz-Apodaca, B.A., et al., *Prevalence of type 2 diabetes and impaired fasting glucose: cross-sectional study of multiethnic adult population at the United States-Mexico border.* Rev Panam Salud Publica, 2010. **28**(3): p. 174-81.

5. Lee, J.W., F.L. Brancati, and H.C. Yeh, *Trends in the prevalence of type 2 diabetes in Asians versus whites: results from the United States National Health Interview Survey, 1997-2008.* Diabetes Care, 2011. **34**(2): p. 353-7.

6. Bowden, D.W., et al., *Review of the Diabetes Heart Study (DHS) family of studies: a comprehensively examined sample for genetic and epidemiological studies of type 2 diabetes and its complications.* Rev Diabet Stud, 2010. **7**(3): p. 188-201.

7. Tsai, F.J., et al., *A genome-wide association study identifies susceptibility variants for type 2 diabetes in Han Chinese.* PLoS Genet, 2010. **6**(2): p. e1000847.

8. Saxena, R., et al., *Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels.* Science, 2007. **316**(5829): p. 1331-6.

9. Scott, L.J., et al., *A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants.* Science, 2007. **316**(5829): p. 1341-5.

10. Sladek, R., et al., *A genome-wide association study identifies novel risk loci for type 2 diabetes.* Nature, 2007. **445**(7130): p. 881-5.

11. Steinthorsdottir, V., et al., *A variant in CDKAL1 influences insulin response and risk of type 2 diabetes.* Nat Genet, 2007. **39**(6): p. 770-5.

12. Zeggini, E., et al., *Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes.* Science, 2007. **316**(5829): p. 1336-41.

13. Bouatia-Naji, N., et al., *A polymorphism within the G6PC2 gene is associated with fasting plasma glucose levels.* Science, 2008. **320**(5879): p. 1085-8.

14. Saxena, R., et al., *Large-scale gene-centric meta-analysis across 39 studies identifies type 2 diabetes loci.* Am J Hum Genet, 2012. **90**(3): p. 410-25.

15. Voight, B.F., et al., *Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis.* Nat Genet, 2010. **42**(7): p. 579-89.

16. Dupuis, J., et al., *New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk.* Nat Genet, 2010. **42**(2): p. 105-16.

17. So, H.C., et al., *Evaluating the heritability explained by known susceptibility variants: a survey of ten complex diseases.* Genet Epidemiol, 2011. **35**(5): p. 310-7.

18. Yasuda, K., et al., *Variants in KCNQ1 are associated with susceptibility to type 2 diabetes mellitus.* Nat Genet, 2008. **40**(9): p. 1092-7.

19. Unoki, H., et al., *SNPs in KCNQ1 are associated with susceptibility to type 2 diabetes in East Asian and European populations.* Nat Genet, 2008. **40**(9): p. 1098-102.

20. Shu, X.O., et al., *Identification of new genetic risk variants for type 2 diabetes.* PLoS Genet, 2010. **6**(9): p. e1001127.

21. Yamauchi, T., et al., *A genome-wide association study in the Japanese population identifies susceptibility loci for type 2 diabetes at UBE2E2 and C2CD4A-C2CD4B.* Nat Genet, 2010. **42**(10): p. 864-8.

22. Kooner, J.S., et al., *Genome-wide association study in individuals of South Asian ancestry identifies six new type 2 diabetes susceptibility loci.* Nat Genet, 2011. **43**(10): p. 984-9.

23. Lettre, G., et al., *Genome-wide association study of coronary heart disease and its risk factors in 8,090 African Americans: the NHLBI CARe Project.* PLoS Genet, 2011. **7**(2): p. e1001300.

24. Kuraku, S., et al., *aLeaves facilitates on-demand exploration of metazoan gene family trees on MAFFT sequence alignment server with enhanced interactivity.* Nucleic Acids Research, 2013. **41**(W1): p. W22-W28.

25. Katoh, K., J. Rozewicki, and K.D. Yamada, *MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization.* Briefings in Bioinformatics, 2017. **20**(4): p. 1160-1166.

26. Larkin, M.A., et al., *Clustal W and Clustal X version 2.0.* Bioinformatics, 2007. **23**(21): p. 2947-8.

27. Delgado, A., A. Huamani, and B. Brillitt. *Applying Shannon Entropy to Analise Health System Level by departments in Peru*. in *2018 IEEE XXV International Conference on Electronics, Electrical Engineering and Computing (INTERCON)*. 2018.

28. Mateos-Valenzuela, A.G., et al., *Risk factors and association of body composition components for lumbar disc herniation in Northwest, Mexico.* Scientific Reports, 2020. **10**(1): p. 18479.

29. Limeres, C.C. *REGRESIÓN LINEAL SIMPLE*. 2011 [cited 2022 19/Enero]; Available from: http://eio.usc.es/eipc1/BASE/BASEMASTER/FORMULARIOS-PHP-DPTO/MATERIALES/Mat_50140116_Regr_%20simple_2011_12.pdf.

30. Kavitha, S., S. Varuna, and R. Ramya. *A comparative analysis on linear regression and support vector regression*. in *2016 Online International Conference on Green Engineering and Technologies (IC-GET)*. 2016.

31. Rosner, B., *Fundamentals of biostatistics*. 2011: Seventh edition. Boston : Brooks/Cole, Cengage Learning, [2011] ©2011.

32. Szumilas, M., *Explaining odds ratios.* Journal of the Canadian Academy of Child and Adolescent Psychiatry = Journal de l'Academie canadienne de psychiatrie de l'enfant et de l'adolescent, 2010. **19**(3): p. 227-229.

33. Warde-Farley, D., et al., *The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function.* Nucleic Acids Res, 2010. **38**(Web Server issue): p. W214-20.

34. Kumar, S., et al., *MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms.* Mol Biol Evol, 2018. **35**(6): p. 1547-1549.

35. Team, R.C., *R: A Language and Environment for Statistical Computing*. 2020.

36. Li, K., et al., *Maternally Inherited Diabetes Mellitus Associated with a Novel m.15897G>A Mutation in Mitochondrial tRNA(Thr) Gene.* J Diabetes Res, 2020. **2020**: p. 2057187.

37. Momiyama, Y., et al., *A mitochondrial DNA variant associated with left ventricular hypertrophy in diabetes.* Biochem Biophys Res Commun, 2003. **312**(3): p. 858-64.