Universidad Nacional Autónoma de México

Facultad de Ciencias

Determinación de factores genéticos asociados a obesidad y metabolismo de lípidos en una muestra de población nativa mexicana

T E S I S

QUE PARA OBTENER EL TÍTULO DE:
MATEMÁTICO

PRESENTA:
CARLA MÁRQUEZ LUNA

DIRECTOR DE TESIS:
DRA. SANDRA ROMERO HIDALGO

2012

1. Datos del alumno	1. Datos del alumno
Apellido paterno	Márquez
Apellido materno Nombre(s)	Luna
Teléfono	5544375193
Universidad Nacional Autónoma de México	Universidad Nacional Autónoma de México
Facultad de Ciencias	Facultad de Ciencia
Carrera	Matemáticas
Número de cuenta	407008463
2. Datos del tutor	2. Datos del tutor
Grado	Dra
Nombre(s)	Sandra
Apellido paterno	Romero
Apellido materno	Hidalgo
3. Datos del sinodal 1	3. Datos del sinodal 1
Grado	Dra
Nombre(s)	Eliane
Apellido paterno	Regina
Apellido materno	Rodrigues
4. Datos del sinodal 2	4. Datos del sinodal 2
Grado	Dr
Nombre(s)	Pedro Eduardo
Apellido paterno	Miramontes
Apellido materno	Vidal
5. Datos del sinodal 3	5. Datos del sinodal 3
Grado	Dra
Nombre(s)	María del Pilar
Apellido paterno	Alonso
Apellido materno	Reyes
6. Datos del sinodal 4	6. Datos del sinodal 4
Grado	Dr
Nombre(s)	Eduardo
Apellido paterno	Gutierrez
Apellido materno	Peña
7. Datos del trabajo escrito	7. Datos del trabajo escrito
Título	Determinación de factores genéticos asociados a obesidad y metabolismo de lípidos en una muestra de población nativa mexicana
Número de páginas	87 p
Año	2010

A mis abuelitas

Agradecimientos

A ti, oh Dios de mis padres, te doy gracias y te alabo, porque me has dado sabiduría y fuerza. Daniel 2:23

Primeramente, quiero agradecer a mi asesora la Dra. Sandra Romero Hidalgo. Le agradezco por su paciencia, su buena disposición a enseñar y por haberme brindado la oportunidad de trabajar con ella.

Con esta tesis concluyo mis estudios de licenciatura en matemáticas, pero el significado de este trabajo no sólo se acota al contenido del mismo, ya que con éste pretendo terminar una etapa de mi vida. Etapa llena de retos, emociones, aprendizaje y experiencias, que no hubiera sido lo mismo sin el apoyo de mi familia. Le quiero agradecer a mi mamá por su infinito cariño y por saber ser la perfecta madre y amiga. A mi papá por su constante apoyo y sus sabios consejos. Y a mi hermana por ser tan alegre y porque a su corta edad es un modelo a seguir.

Agradezco a Pablo por ser el mejor compañero que pude haber pedido. También le doy las gracias a aquellos dos amigos que con el tiempo se han convertido en hermanos, a Juancris y a Omar.

Por último a mi perrito Negro, que me acompañó en las noches de desvelo y que día con día me hace sonreir.

Distrito Federal,
Agosto 2012

Carla Márquez Luna

Índice general

Introducción

En los últimos años los estudios de asociación de genoma completo han permitido la identificación de numerosas variantes genéticas que contribuyen al desarrollo de enfermedades complejas. Sin embargo, la mayor parte de estos estudios se han hecho en población europea, estudiar otras poblaciones brinda la posibilidad de encontrar nuevas variantes de riesgo.

La población mexicana tiene una alta prevalencia en enfermedades metabólicas y se ha sugerido que la susceptibilidad genética que presenta esta población proviene del componente indígena.

El propósito general de este trabajo es la realización de un estudio piloto, donde por medio de un estudio de asociación de genoma completo, se pretende identificar si se replican SNPs[1] previamente asociados a distintos rasgos metabólicos en una muestra real de población indígena. Además, se evalúa la posibilidad de identificar asociaciones de SNPs nuevos en una muestra de tamaño moderado. El trabajo consta de tres capítulos y un glosario.

El capítulo 1 describe los conceptos básicos del campo de la genética necesarios para la comprensión conceptual de los análisis realizados. Consta de cinco secciones: la primera describe el componente biológico de la información genética y el concepto de recombinación; la segunda aborda el tema de variabilidad genética y la forma en que se mide; la tercera trata acerca de los patrones de herencia de las enfermedades; la cuarta describe el concepto de desequilibrio de ligamiento; y la última describe algunos proyectos que han marcado la pauta en el estudio de las enfermedades.

El capítulo 2 presenta las características principales del análisis estadístico realizado con el fin de introducir los métodos utilizados. Por ello, inicia con una presentación general de los métodos estadísticos convencionales para identificar genes que predisponen a enfermedades; para proseguir con la descripción detallada de las etapas que comprenden los estudios de asociación de genoma completo.

El capítulo 3 se integra como un reporte científico, por lo que presenta los objetivos del estudio, el método seguido para analizar los datos, los resultados obtenidos y la discusión de los resultados.

[1] Polifomorfismo de un solo nucleótido o SNP (Single Nucleotide Polymorphism, por sus siglas en inglés) es una variación en la secuencia de ADN que sucede cuando una sola base es sustituida por otra.

En este trabajo destacan tres etapas fundamentales: se eligió entre las pruebas estadísticas convencionales la que mejor se ajuste a las características de la población objeto de estudio; se estudió la estructura de la población para identificar la homogeneidad de la misma; y se determinó el nivel de significancia apropiado para reportar los resultados.

Capítulo 1

Conceptos de genética

Cuando se estudian los antecedentes genéticos de un fenotipo[1] o enfermedad, la primera pregunta que se debe entender es ¿A qué se hace referencia cuando se habla de antecedentes genéticos? Este capítulo trata de responder esta pregunta. En la primera sección se describe cuál es el componente biológico de la información genética, en qué parte del cuerpo humano se encuentra, qué significa y cómo se transmite. Con el concepto de recombinación se explica una de las causas de que la información genética sea distinta para cada persona.

La siguiente sección trata el tema de la variabilidad genética; ésta es clave al estudiar la arquitectura genética de una enfermedad y puede ser medida con marcadores genéticos.

Posteriormente, se abordan las leyes de segregación de Mendel, la transmisión de fenotipos entre familias y qué tipo de complicaciones afectan a los patrones de herencia. De ahí derivamos en lo que se conoce como enfermedades complejas. En la última parte de esta sección pasamos al principio fundamental de la genética de poblaciones, el equilibrio de Hardy-Weinberg.

En la siguiente sección se define el concepto de desequilibrio de ligamiento, concepto fundamental en los estudios de asociación de genoma completo.

Debido al desarrollo de alta tecnología en los últimos años, se ha podido estudiar la composición genética de enfermedades a un gran nivel de detalle. La última sección se dedica a algunos proyectos que han marcado la pauta en el estudio de enfermedades. Primero se habla acerca del proyecto del genoma humano, cuyo propósito fue tener la secuencia completa del ADN humano. El siguiente paso natural fue ver cómo se puede utilizar esta información para estudiar su participación en el desarrollo y progresión de enfermedades. Es así que surgieron nuevos proyectos complementarios como lo son el proyecto internacional de HapMap y 1000 genomas.

El material de este capítulo se obtuvo de los libros [19, 32, 54, 56, 62]. En los libros citados en [32, 54, 56] se abordan los conceptos desde un punto de

[1] Un fenotipo es cualquier característica o rasgo observable de un organismo el cual se puede conocer de la observación directa de la apariencia externa de un organismo.

vista más biológico. Mientras que en los libros citados en [19, 62] la definición de conceptos es mas concreta ya que los libros estan más enfocados al estudio de enfermedades desde un punto de vista estadístico. Las imagenes son cortesía de *National Human Genome Research Institute*.

1.1. Estructura, función y herencia del genoma humano

1.1.1. Estructura de ADN

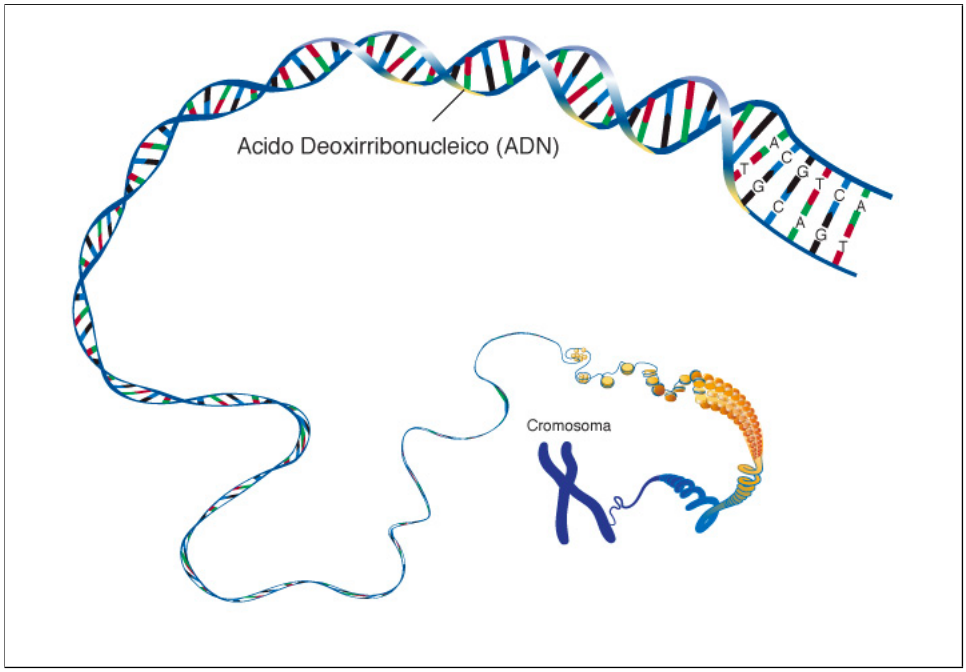

Figura 1.1: ADN es el nombre químico de la molécula que contiene la información genética en todos los seres vivos. El cuál está empaquetado en 23 pares de cromosomas.

Todos los organismos (con excepción de algunos tipos de virus) utilizan ácido desoxirribonucleico (ADN) como material genético. El ADN es una macromolécula que juega dos papeles biológicos centrales:

- Porta las instrucciones para crear los componentes de células (en su mayoría proteínas, las que a su vez manufacturan otros componentes);
- Provee significado a este conjunto de instrucciones que a su vez son pasadas a las células hijas cuando la célula se divide.

La estructura de una molécula de ADN consiste en dos cadenas que se envuelven una alrededor de la otra como si fuera una escalera de caracol, cuyos lados están compuestos de azúcar y moléculas de fosfato y están conectados por bases nitrogenadas. Cada cadena es un arreglo lineal de unidades repetitivas llamadas nucleótidos, compuestos por un azúcar, un fosfato y una base nitrogenada. Existen cuatro variedades de nucleótidos que forman el ADN y contienen las bases adenina (A), guanina (G), citosina (C) y timina (T), y es la secuencia de estas partes de moléculas nucleótidas que cargan con la información genética. Adenina y guanina, moléculas compuestas de dos anillos, son purinas; y citosina y timina, moléculas compuestas de un anillos, son pirimidinas. Cada base está unida a una molécula de azúcar, desoxirribosa, y cada desoxirribosa tiene pegado a un grupo de fosfato; el azúcar y el fosfato sólo juegan un papel estructural en el ADN, y no portan información. Las dos cadenas de ADN se mantienen unidas por enlaces de hidrógeno entre bases complementarias formando así pares de bases. Se llaman bases complementarias ya que cada base tiene una única pareja posible; adenina se empareja sólo con timina (par $A - T$) y citosina se empareja sólo con guanina (par $C - G$). La longitud de las moléculas de ADN son descritas en unidades de pares de bases (bp), y para moléculas más grandes se utilizan las unidades: kilobase (kb) equivalente a $1,000$ de pares de bases o megabase (Mb) equivalente a $1,000,000$ de pares de bases. Se calcula que el genoma humano contiene 3 billones de pares de bases.

1.1.2. Genes y cromosomas

Algunos segmentos de ADN contienen instrucciones para la síntesis de proteínas, estos segmentos se llaman genes. Cada molécula de ADN contiene muchos genes. Se define *gen* como una secuencia específica de bases de nucleótidos. Es decir, un segmento de una molécula de ADN localizado en un posición particular de un cromosoma en específico, cuya secuencia carga la información requerida para la síntesis de proteínas, proporcionando los componentes estructurales de células y tejidos, así como de encimas para reacciones químicas esenciales. Las distintas formas de un mismo gen que surgen por mutaciones se conoce como *alelos*.

La producción de proteínas de un gen no viene directamente de ADN. Primero, el ADN es transcrito para hacer una molécula de ácido ribonucléico o ARN, conocida como ARN mensajero (mARN), y luego esta actúa como referencia para la producción de proteínas. La producción de muchas copias de mRNA por parte de un solo gen amplifica el número de copias de las proteínas correspondientes que se pueden hacer y también provee muchas oportunidades para procesos regulatorios que influencíen la cantidad final y las propiedades de la proteína activa. El ARN no sólo difiere del ADN en el tipo de azúcar que contiene (ribosa, en vez de desoxirribosa), sino también en una de sus bases. El ARN contiene una base de pirimidina llamada uracil (U) en vez de la timina (T). Al igual que la timina, uracil se empareja con adenina (A).

Cabe señalar que cerca del 98.7 % del ADN no comprende secuencias codificantes de genes, y aproximadamente el 70 % del ADN no se transcribe. En sí,

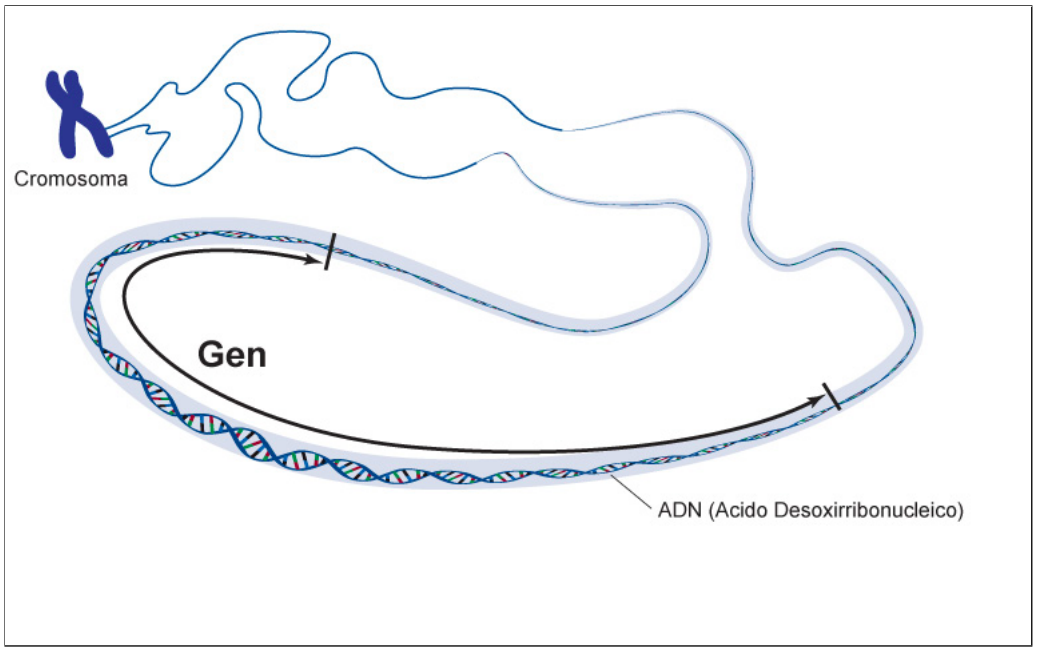

Figura 1.2: El gen es la unidad física básica de la herencia. Los genes se transmiten de los padres a la descendencia y contienen la información necesaria para precisar sus rasgos. Los genes están dispuestos, uno tras otro, en estructuras llamadas cromosomas.

la función de este material no génico es desconocida.

La naturaleza lineal del ADN implica que grandes genomas, tal como el del ser humano, corresponde a moléculas muy grandes. Cada una de las células somáticas contiene dos copias de este genoma en su núcleo, los cuales estan empacados de manera eficiente. El material genético de las células somáticas es dividido en 46 moléculas separadas que se llaman *cromosomas*.

Todos los genes se encuentran en orden lineal a lo largo de estos cuerpos microscópicos que se sitúan en el núcleo celular. Los cromosomas se presentan en pares homólogos, por lo que se tienen 23 pares en total, uno derivado de la madre y el otro del padre, y el número de pares es constante para cada especie. La totalidad de estos pares constituye el *genoma* de un organismo particular. Uno de los pares de cromosomas en el genoma son los cromosomas sexuales, normalmente denotados por X y Y, y determinan el sexo del individuo. A los restantes se les conoce como autosomas e influyen en casi todos los demás rasgos.

Se llama *locus* a un segmento específico de ADN en una posición particular del cromosoma, cuyo plural se conoce como *loci*. Como los cromosomas vienen en pares, los locus y los genes que los constituyen también vienen en pares. En la sección 1.2 se habla acerca de cómo se dan las diferencias en un locus.

1.1.3. Mitosis y meiosis

Cuando las células del cuerpos se dividen y multiplican, el núcleo de dichas células pasa por un proceso de división llamado *mitosis*, que da como resultado que dos células hermanas tengan un conjunto completo de pares de cromosomas idénticos a los de la célula madre. Sin embargo, en la producción de células reproductivas, *gametos*, el proceso es diferente y se llama *meiosis*. En este proceso se asegura que sólo un cromosoma de cada par homólogo pase a cada gameto; las células que pasan se conocen como el *haploides* ya que tiene sólo la mitad del número de cromosomas de la célula progenitora, en contraste con el número de huevo fertilizado, *cigoto*, el cual es *diploide*. La meiosis consiste en una ronda de duplicación cromosómica y dos rondas de división celular. A continuación se describen los pasos que componen el proceso meiótico:

1. El proceso comienza con una célula diploide regular ($2n$), lo cual significa que hay dos cromosomas: uno paterno y uno materno. Los cromosomas paternos y maternos que coinciden se conocen como *homólogos*.

2. Replicación de ADN: El ADN es duplicado, y como resultado, cada cromosoma ahora contiene dos dobles hélices de ADN idénticas. A este último se conoce como *cromátidas hermanas*. Hay un total de cuatro dobles hélices ($4n$) en la célula.

3. Formación de bivalentes: Cromosomas homólogos son conectados para formar *bivalentes*.

4. Posible entrecruzamiento: En esta etapa, es posible el intercambio de material genético entre las cadenas maternas y paternas.

5. División meiótica I: Las cromátidas no hermanas son separadas, mientras que las cromátidas hermanas se mantienen emparejadas. Esto da como resultado a dos células diploides que contienen a las cromátidas hermanas.

6. División meiótica II: Las cromátidas hermanas son separadas, dando como resultado cuatro células haploides (gametos).

7. Durante la fertilización, la fusión del material genético de los dos gametos conlleva a la restitución del estatus de haploide.

Un factor importante de la divisón meiótica es que cromosomas homólogos son distribuidos aleatoriamente e independientemente uno del otro en los gametos. Por lo que el gameto resultante normalmente tiene algunos cromosomas heredados de su padre y otros heredados de su madre, pero la combinación específica del cromosoma es aleatoria. Existe un gran número de posibles combinaciones de cromosomas en una sola célula. Se tiene un total de 23 pares de cromosomas, por lo que el número de combinaciones posibles en un gameto de un padre es 2^{23}.

Sin embargo, existe un segundo nivel de modificación en el paso de material genético al gameto, y se conoce como *entrecruzamiento*. Durante la meiosis, los

cromosomas homólogos paternos y maternos se alinean e intercambian segmentos a través de recombinaciones. Este proceso es recíproco y no existe pérdida de información genética. El entrecruzamiento asegura que cualquier gameto producido por un hombre o mujer es genéticamente diferente de cualquier otro. El entrecruzamiento no ocurre durante una división mitótica normal.

1.1.4. Recombinación

El proceso de entrecruzamiento se da durante la división meiótica I, donde las cromátidas no hermanas se entrecruzan y forman quiasmas visibles. En los puntos de entrecruzamiento, las cromátidas se pueden romper en puntos homólogos y reunirse con su cromátida no hermana. Esto puede ocurrir más de una vez. Se conoce como *recombinación* cuando se puede distinguir si hubo entrecruzamiento o no. Supóngase que se tienen que tenemos dos segmentos de cromosomas A y B en las cromátidas 1 y 2; una posiblilidad es que no exista entrecruzamiento entre los segmentos. Entonces, $A1$ se va a mantener en la misma cromátida con B1, y los mismo para $A2$ y $B2$. Otra posibilidad es que haya un entrecruzamiento; como resultado se intercambiarían los segmentos cromosómicos $A1$ y $B2$ en uno, y en el otro $A2$ y $B1$. De esta manera ha ocurrido la recombinación entre los segmentos A y B. Una tercera posibilidad es que ocurran dos entrecruzamientos, aquí el segmento intermedio es cambiado pero se tiene el mismo resultado en los segmentos superiores e inferiores, es decir, $A1$ y $B1$ están en la misma cromátida y $A2$ y $B2$ en la segunda. Generalizando, el fenomeno de recombinación entre dos segmentos cromosómicos se puede observar sólo cuando hay un número impar de entrecruzamientos. Entre mayor sea la distancia entre dos segmentos, mayor es la probabilidad de que exista entrecruzamiento. Con base a este fenomeno se define la *fracción de recombinación* θ como la probabilidad de que exista una recombinación entre dos loci. Se puede utilizar para medir la distancia genética entre dos segmentos cromosómicos. Si los segmentos están localizados muy cerca uno del otro, entonces casi nunca van a ser separados, por lo que θ se aproxima a 0. Por otra parte, si están situados en distintos cromosomas, en la mitad de los casos la primera división meiótica los va a distribuir en distintas células, dando así un valor de θ de 0.5. Similarmente, dos segmentos en el mismo cromosoma pero muy lejos uno del otro son propensos a ser objetos de recombinación debido al número alto de entrecruzamiento intermedios.

También se puede expresar la distancia genética en unidades de recombinación llamada *centiMorgans*: un centiMorgan entre un par de marcadores representa que se espera 0.01 % de entrecruzamientos entre ellos.

La observación directa entre alelos (secuencias homólogas de ADN) o la observación de diferencias manifestadas indirectamnete como fenotipos (como enfermedades) que surgen a partir de una secuenca particular, permite que la herencia de segmentos de ADN que va de un padre a un hijo puedan ser seguidos en pedigrís[2] humanos. Los eventos de recombinación pueden ser detectados cuando

[2]Un pedigrí es una descripción pictórica de un árbol familiar. El cuál proporciona infor-

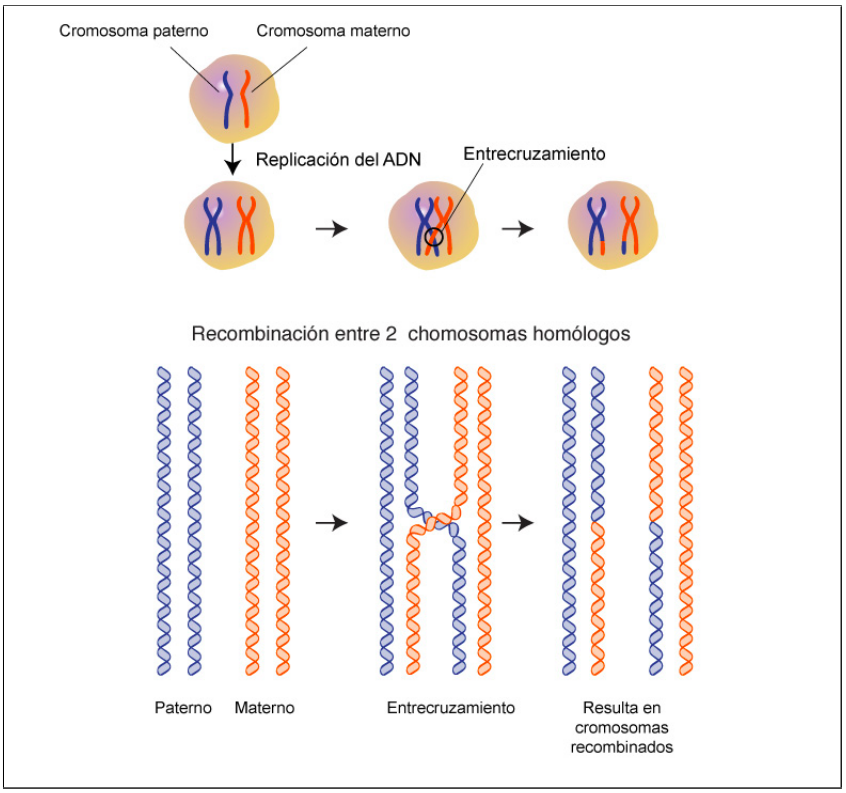

Figura 1.3: Ejemplo de recombinación genética que ocurre durante la meiosis. Los cromosomas apareados de los progenitores masculino y femenino se alinean de forma que secuencias similares del ADN se entrecruzan. Este entrecruzamiento produce un intercambio de material genético, el cual es causa importante de la variabilidad genética que se observa en la descendencia.

se interrumpe la cosegregación de marcadores moleculares, o la cosegregación de un marcador con un fenotipo. Reconocer estos eventos de recombinación permite la apreciación del orden de los marcadores genéticos a lo largo del ADN, y el conteo de los eventos de recombinación permite la estimación de la distancia genética entre marcadores, se le conoce como *mapeo genético*.

1.2. Variación de información genética

Existen dos fuentes principales de variación en una secuencia de ADN [62]. La primera se da durante la división celular, cuando un emparejamiento anormal de cromátidas hace que exista un reacomodo de los segmentos del cromosoma.

mación acerca de las relaciones biológicas de los individuos en la familia, su historia médica, patrones de hereditarios de algún trastorno genético en la familia, entre otros.

Esto da a lugar deleciones, inversiones, translocaciones, inserciones, o duplicaciones de segmentos completos de un cromosoma[3]. Si los nucleótidos son insertados o eliminados del ADN, se tiene que distinguir dos casos. Primero, la inserción o eliminación de un múltiplo de tres nucleótidos deriva en que más o menos aminoácidos sean codificados. Segundo, cualquier número de inserciones o eliminaciones conduce a un cambio en el marco de lectura; por tanto, puede sintetizar una proteína diferente.

La segunda fuente de variación se da durante la replicación de ADN, cuando ocurren mutaciones puntuales con la substitución de nucleótido por otro. Las *mutaciones* se definen como un estado que cambia la función de un gen permanentemente. Estas se dan cuando la replicación de ADN no es absolutamente perfecta, en general, la mayoría de los errores se corrigen inmediatemente. Pero si un error no es detectado o reparado, entonces la mutación ocurre. Como resultado, se pueden obtener transiciones o transversiones. La primera se refiere al intercambio de una purina por otra o el intercambio de una pirimidina por otra. Y el segundo se refiere al intercambio de una purina por una pirimidina o vice versa, lo que deriva en una secuencia de bases sinónimas o no sinónimas. Se dice que una mutación es sinónima cuando un nucléotido es substituido por otro, y la redundancia del código genético puede producir que la misma proteína sea sintetizada. En cambio, se dice que es no sinónima cuando se codifica a un aminoácido diferente por lo que la proteína sintetizada va a ser diferente. La tasa de mutación por locus va de 10^{-6} a 10^{-4} por generación [44]. Por lo tanto, la secuencia de ADN entre cualesquiera dos individuos difieren con una frecuencia de 1 base por cada 1000.

En ocasiones se intercambia el concepto de mutación con el de polimorfismo, se define *polimorfismo* como una variación en la secuencia de ADN que tiene frecuencia de al menos 1 % en al menos una población humana.

1.2.1. Tipos de medición de la variabilidad genética: Marcadores genéticos

Un marcador genético es un segmento de ADN con una ubicación física identificable, en donde al menos una base difiere entre al menos dos individuos.

1.2.2. Tipos de marcadores genéticos

Existen varios tipos de marcadores genéticos. Los que más se usan en la actualidad son: los microsatélites o repetidos en tándem y los polimorfismos de un solo nucleótido (SNPs por sus siglas en inglés).

[3]Una deleción suponen la pérdida de un segmento de un cromosoma, lo que origina un desequilibrio. Una inversión se produce cuando un cromosoma sufre dos roturas y vuelve a reconstruirse con el segmento entre las dos roturas invertidos. Las translocaciones consisten en un intercambio de segmentos entre dos cromosomas, generalmente no homólogos. Una inserción es un tipo de translocaci]'on que ocurre cuando un segmento desprendido de un cromosoma se inserta en otro cromosoma en su orientación usual o invertido. Una duplicación es la repetición de un fragmento de cromosoma a continuación del fragmente original.

Los microsatélites (STRs por sus siglas en inglés), consisten en secuencias cortas de ADN que se repiten consecutivamente, normalmente van de dos a cuatro nucleótidos. Basándose en el número de nucleótidos repetidos se denominan mononucleótidos, dinucleótidos, trinucleótidos o tetranucleótidos. El tamaño total de expansión de los STRs van de 10 a 100 kilobases. A lo largo de las regiones genómicas, las repeticiones de mononucleótidos son las más frecuentes, le siguen las repeticiones de dinucleótidos y trinucleótidos.

Por otra parte, el tipo de variación más abundante que ocurre en el genoma humano son los SNPs. Éstos representan aproximadamente el 90 % de la variación total existente. Básicamente estas variaciones ocurren en una sola base, es decir, que una base es substituida por otra. Para que un polimorfismo sea SNP debe de cumplir con que la frecuencia del alelo menor o la frecuencia de alelo más raro, tiene que ser de al menos 1 % en la población.

A pesar de que los SNPs son muy frecuentes, su frecuencia no está distribuida equitativamente a lo largo de todo el genoma. Los SNPs tienen una baja tasa de mutación y se dice que muchos de éstos surgieron antes de que emergieran las diferentes poblaciones humanas. Se dice que todas las poblaciones humanas comparten aproximadamente el 85 % de los SNPs, difieriendo en frecuencias alélicas. Los SNPs están distribuidos a lo largo de todo el genoma. Existen SNPs codificantes (cSNPs). Es decir, aquellos que se encuentran en regiones codificantes de genes, así como SNPs no codificantes que son aquellos que se encuentran en regiones no codificantes y que están en un vecindario inmediato alrededor de ciertos genes. También están los SNPs no codificantes aleatorios que se encuentran en una región intergénica. Los SNPS codificantes hacen que los cambios sinónimos que no alteran al aminoácido codificado sean menos consecuentes que aquellos cambios no sinónimos que derivan en un diferente aminoácido codificado.

Constratando STRs con SNPs, en cuanto a la informatividad, se dice que entre más alelos tenga un marcador más informativo será, por lo que los STRs son más informativos. Sin embargo, aunque un solo SNP tiene un grado de informatividad muy bajo, en conjunto cubren al genoma con la densidad necesaria para algunas aplicaciones. Los STRs tiene la ventaja que son altamente polimórficos, aunque por lo general hay sólo uno o dos alelos altamente frecuentes. Su alto nivel polimórfico se debe a su tasa alta de mutación, que es de 10^{-2} a 10^{-6} eventos por locus por generación. Los SNPs son mucho más estables.

A pesar de que la mayoría de los SNPs están en regiones no codificantes, algunos de estos están en genes o en el promotor de genes, por lo que pueden verse como variaciones candidatas para enfermedades. Es así que los SNPs son considerados como los catalizadores del estudio de la base genética de enfermedades complejas.

1.3. Variación genética y el fenotipo

Las diferencias que existen entre las personas en general, se deben en gran parte a la base genética: diferencias en el fenotipo causadas por las diferencias en

el genotipo. Se define *genotipo* a la constitución genética que posee un individuo. Y como *fenotipo* al conjunto de características observables de un individuo que son resultado de la interacción de su genotipo con el ambiente.

Algunas de estas diferencias se pueden observar a simple vista, como color de ojos, piel, cabello. Otras diferencias son más sutiles pero, inclusive, más importantes ya que nos afectan en un sentido médico. Éstas van desde tipo de sangre, factores que indiquen la respuesta a ciertos medicamentos, o ser propensos a contraer ciertas enfermedades como malaria, o trastornos como diabetes, asma, esquizofrenia, entre otras. Algunas diferencias genéticas son las causales directas a ciertas enfermedades como lo puede ser la enfermedad de Huntington o fibrosis quística. La base genética para algunas diferencias fenotípicas, tal como si un individuo sufre de fibrosis quística, se puede observar directamente y ya han sido estudiadas y entendidas. Por otra parte, la base genética para rasgos como la predisposición a esquizofrenia o a diabetes es bastante compleja. En éstas influyen más de un gen así como el ambiente.

Al estudiar las enfermedades de una población se puede hacer uso del cambio de frecuencias alélicas y genotípicas dentro de poblaciones. Para fines prácticos se puede definir a los genes como unidades discretas cuyas características biológicas son transmitidas de padres a hijos. Los genes normalmente son transimitidos sin cambios de generación en generación, y esto suele ocurrir en pares. Si dicho par consiste en genes similares, se dice que el individuo es *homocigoto* con respecto al gen en cuestión, y si los genes no son similares se dice que el individuo es *heterocigoto*. Para ilustrar las definiciones anteriores considérese dos alternativas posibles de genes, A_1 y A_2, en un determinado *locus*. Entonces tenemos dos opciones de homocigotos, A_2A_2 y A_1A_1, y un posible heterocigoto, A_2A_1. Dichas alternativas se llaman *alelos*, es decir, son distintas versiones de un mismo gen. Con un solo par de alelos existen tres posibles tipos de organismos representados por los tres genotipos A_1A_1, A_2A_2 y A_1A_2.
Con los tres genotipos anteriores, supóngase que en una población existen estas tres posibilidades. La *frecuencia relativa* de un genotipo se denota como x_{ij}, como se define en la siguiente tabla:

Genotipo:	A_1A_1	A_1A_2	A_2A_2
Frecuencia relativa:	x_{11}	x_{12}	x_{22}

Donde los valores x_{ij} cumplen que $0 \leq x_{ij} \leq 1$ y

$$x_{11} + x_{12} + x_{22} = 1.$$

Por otra parte, la *frecuencia alélica* de un alelo en particular, digamos A_1 es la probabilidad de que, al tomar un alelo de forma aleatoria en la población, éste sea A_1. El acto de tomar un alelo de forma aleatoria se divide en dos pasos: tomar un genotipo aleatoriamente de una población y tomar aleatoriamente un alelo del genotipo escogido. Siguiendo el ejemplo anterior, se tienen tres genotipos por lo que la frecuencia alélica p de A_1 es

$$p = (x_{11} \cdot 1) + (x_{12} \cdot \tfrac{1}{2}) + (x_{22} \cdot 0).$$

El primer término representa la probabilidad de que se tome el genotipo A_1A_1 multiplicado por la probabilidad de que se tome el alelo A_1, la cual claramente es 1. El segundo término representa la probabilidad de que se tome el genotipo A_1A_2 multiplicada por la probabilidad de que en ese genotipo salga el alelo A_1, la cual es $\frac{1}{2}$ ya que hay dos posibilidades, A_1 o A_2. Por último, se tiene la probabilidad de que se tome el genotipo A_2A_2 multiplicada por la probabilidad de que se tome el alelo A_1 en dicho genotipo, la cual es 0 porque este genotipo no contiene a dicho alelo.

1.3.1. Leyes de Mendel

Gregorio Mendel realizó experimentos de cruzamiento con guisantes o semillas, y observó la transmición de rasgos fáciles de distinguir tales como los colores de la flor o de la semilla. En base a esto dedujo conceptos que posteriormente fueron denominados como las leyes mendelianas. La primera de éstas se conoce como la ley de la uniformidad que establece que si se cruzan dos razas puras para un determinado carácter, los descendientes de la primera generación serán todos iguales entre sí fenotípica y genotípicamente. La segunda, conocida como la ley de la segregación, establece que durante la formación de los gametos, cada alelo de un par se separa del otro miembro para determinar la constitución genética del gameto filial.

Las dos leyes anteriores describen cómo la distribución de un rasgo en particular es controlada por un solo gen. En casos simples, después de haber cruzado dos homocigotos diferentes la primera generación contiene heterocigotos que se ven igual. Por ejemplo, en el caso del color de los guisantes, Mendel cruzó guisantes amarillos homocigotos con guisantes verdes homocigotos, y toda la descendencia resultaron ser heterocigotos de color amarillo. Sin embargo, al cruzar la generación resultante entre sí se generaron guisantes de color amarillo y de color verde.

Por último, la tercera ley de Mendel conocida como la ley de independencia, establece que dos factores genéticos son transmitidos de manera independiente uno del otro. Esto quiere decir que, por ejemplo, la madre tiene los alelos 1 y 2 en un locus, y tiene A y B en otro, sus hijos pueden heredar las combinaciones $1A$, $1B$, $2A$ y $2B$ con probabilididad igual para todas. De igual forma el caso del padre. En base a lo visto en la sección 1.1.3 sabemos que esta ley no se cumple para humanos a menos que los loci estén en diferentes cromosomas. Cuando están en distintos cromosomas es cuestión de azar si después de la meiosis dos loci terminan en el mismo gameto o no. Sin embargo, si están en el mismo cromosoma la tercera ley no se cumple ya que depende de la distancia a la que estén los loci uno del otro.

1.3.2. Patrones de herencia mendeliana

Aparte de los factores genéticos, la influencias ambientales son las que normalmente juegan un mayor papel. Basándose en estos factores podemos distinguir a las enfermedades mendelianas o monogénicas de las complejas. Las enfer-

medades mendelianas siguen directamente un patrón de herencia mendeliana. En contraste, las enfermedades complejas son causados por múltiples factores genéticos y ambientales. Sin embargo, estas diferencias son un tanto arbitrarias ya que no existe una transición clara entre una clasificación y la otra. En sí, cada caso presenta más o menos complicaciones que interrumpen el esquema básico de herencia.

Si sólo un factor causa la enfermedad, éste debe de tener un efecto grande y debería de ser fácil de revelar. Pero si muchos factores contribuyen a la enfermedad, el efecto de cada uno por separado debe de ser moderado o pequeño. Y se requerirán muestras grandes para detectar los efectos.

Los patrones de herencia observados por Mendel pueden aplicar también a enfermedades en seres humanos. Pueden ser:

Herencia autosómica dominante. Se da cuando el gen que determina la enfermedad es dominante y está en uno de los 22 cromosomas autosómicos. Y el individuo sólo necesita portar un alelo para tener la enfermedad.

Herencia autosómica recesiva. Se necesitan dos alelos causantes de enfermedad para que la enfermedad se manifieste.

Cuando el gen que influye en fenotipo de interés se encuentra en un cromosoma sexual. Es importante considerar las característica de transmisión de dichos cromosomas. Para el cromosoma X, las madres siempre transmite uno sin importar el sexo del hijo; el padre siempre transmite uno a las hijas y ninguno a los hijos. El cromosoma Y, dado que no lo portan las mujeres, sólo los padres lo transimiten y sólo lo heredan los hijos.

Herencia dominante del cromosoma X. En este caso, si el padre porta el alelo causante se lo transmite a todas las hijas y se verán afectadas, mientras que ningun hijo correrá el riesgo ya que no lo herederán. Si la madre lo porta ella se lo trasnmitirá a la mitad de sus hijos sin importar su género.

Herencia recesiva del cromosoma X. Es más común en hombres ya que los hombres son hemicigotos (solo tienen un alelo), por lo que no importa si es dominante o recesivo, un solo alelo basta para determinar la expresión de la enfermedad. Entonces un hombre puede verse afectado si la madre es portadora, mientras que una mujer se ve afectada sólo si cada padre le hereda el alelo de la enfermedad.

Herencia del cromosoma Y. No importa si es dominante o recesivo ya que solo los hombres lo heredarán y son hemicigotos.

1.3.3. Enfermedades complejas

El término "rasgo complejo" se refiere a cualquier fenotipo que no presenta herencia mendeliana clásica, ya sea de tipo recesiva o dominante atribuida a un locus particular. En general estas complejidades se presentan cuando se rompe la correspondencia entre genotipo y fenotipo. Este quiebre puede darse cuando un mismo genotipo da diferentes enfermedades (debido al azar, ambiente o interacciones con otros genes) o distintos genotipos dan el mismo fenotipo. Es casi imposible encontrar marcadores genéticos que muestren una perfecta cosegregación en un rasgo complejo. Esto se le atribuye a los siguientes problemas.

Penetrancia incompleta y fenocopia. Se dice que hay penetrancia incompleta cuando algunos individuos que heredan un alelo con predispocisión y enfermedad no manifiesten la enfermedad. Fenocopia se refiere a la situación en la que algunos individuos que no portan alelo con predisposición a la enfermedad desarrollan la enfermedad como resultado de causas ambientales o el azar. Lo que implica que el genotipo en un locus dado puede afectar la probabilidad de contraer una enfermedad pero no determina por completo el resultado. Es así que la función de pentrancia, que representa la probabilidad de contraer enfermedad dado el genotipo, puede depender de factores no genéticos como edad, sexo, ambiente y otros genes.

Heterogeneidad genética. Mutaciones en un conjunto de genes pueden causar fenotipos idénticos donde dicho conjunto comparten algún proceso bioquímico o estructura celular. Cuando sucede esto, no hay forma de saber si dos pacientes sufren la misma enfermedad por distintas causas genéticas, al menos hasta que los genes sean mapeados.

Herencia poligénica. Algunos rasgos requieren de mutaciones simultáneas en múltiples genes. La herencia poligénica complica el mapeo genético ya que no se requiere de ningún locus en particular para producir el rasgo.

1.3.4. Equilibrio de Hardy-Weinberg

En organismos diploides como los humanos, se tienen dos alelos A y a en un mismo locus, con frecuencias p y q respectivamente, y pueden ser combinadas para formar tres genotipos: AA, Aa y aa. Si se conoce la frecuencia de esos dos alelos en una población ideal podemos predecir las proporciones de los genotipos en la siguiente generación al combinar los gametos (que contienen sólo un alelo) de formar aleatoria. A este postulado se le conoce como *principio de Hardy-Weinberg*. Entonces la proporción de cada genotipo en la siguiente generación es:

$$AA = p^2 \quad Aa = 2pq \quad aa = q^2$$

Si las proporciones de genotipo de la siguiente generación son calculadas de esta forma, y se encuentra que son indistinguibles de aquellas de la generación parental, se dice que no está ocurriendo evolución (definida como el cambio de frecuencias alélicas), y que la población está en *equilibrio de Hardy-Weinberg*.

Para que se puedan estimar las proporciones de genotipos de una generación a la siguiente, la población debe de componerse de un número infinito de organismos diploides que se estén reproduciendo sexualmente de forma aleatoria. Sin embargo, para que el equilibrio de Hardy-Weinberg sea observado en una población, idealmente se tienen que tener las siguientes condiciones adicionales:

- no selección
- no mutación
- no migración

Esto es debido a que dichos factores pueden cambiar las frecuencias alélicas. De lo contrario, si las proporciones de genotipo calculadas no están en equilibrio de Hardy-Weinberg, podemos concluir que esa población está bajo el proceso de evolución (cambio de frecuencias alélicas) y que uno o varios de los factores anteriormente mencionadas está operando.

Por medio del estadístico denotado por F, se mide la desviación del grado de heterocigosidad de una población con respecto al valor esperado si estuviera en equilibrio de Hardy-Weinberg.

El valor de F se calcula como

$$F = 1 - \frac{\#\text{ observado de Aa}}{\mathrm{E}(f(Aa))}, \tag{1.1}$$

donde el número de heterocigotos esperado bajo la hipótesis de estar en equilibrio de Hardy-Weinberg es $\mathrm{E}(f(Aa)) = 2pq$. Si la población estudiada esta bajo equilibrio de Hardy-Weinberg entonces se espera un valor de $F \approx 1$.

1.4. Desequilibrio de ligamiento

La recombinación meiótica es consecuencia de la reproducción sexual y mejora la habilidad de las poblaciones a adaptarse a su ambiente a través de la combinación de alelos ventajosos en diferentes loci. La recombinación puede ser estudiada a un nivel poblacional por medio del desequilibrio de ligamiento. Antes de abordar el concepto de desequilibrio de ligamiento se define el concepto de haplotipo.

1.4.1. Haplotipo

Un haplotipo se refiere a la combinación de alelos de marcadores polimórficos a lo largo de una misma molécula de ADN. Los lugares pueden incluir cualquier clase de polimorfismo de ADN. En el cromosoma Y se deriva directamente un haplotipo ya que estas moléculas son haploides. En los marcadores del cromosoma X, en el caso de hombres, se deriva tambien un haplotipo directamente, debido a que los hombres sólo cargan con un cromosoma X. En el caso de mujeres sólo se obtiene un haplotipo si ambos cromosomas tienen el mismo haplotipo (homocigosidad). Debido a que el cromosoma Y no es recombinante, la variación que se encuentre en ésta se debe a mutaciones.

En el resto del genoma, aparte de la mutación, la diversidad haplotípica se debe a la recombinación: el haplotipo presentado en una generación puede romperse y dar un nuevo haplotipo en la siguiente. El efecto de la recombinación es incrementar la diversidad haplotípica. Como ya se mencionó anteriormente, la probabilidad de recombinación entre dos marcadores autosómicos depende de sus posiciones relativas.

1.4.2. Desequilibrio de ligamiento

La tendencia de que dos alelos particulares en dos loci sean cosegregados debido a una baja tasa de recombinación entre ellos conlleva a asociación entre alelos en una población. A esta propiedad se le conoce como *desequilibrio de ligamiento* (LD por sus siglas en inglés). Cuando se realizan escaneos de genoma completo para encontrar asociaciones entre marcadores y enfermedades, se ha mostrado un interés especial en estudiar patrones de LD. En particular, se dice que ocurre LD cuando dos alelos son encontrados juntos en un mismo cromosoma más veces que las que se esperarían ver si estos alelos segregaran al azar. Existen varias medidas para evaluar el LD, y éstas difieren en sus propiedades y utilidades.

Primeramente, considérese un marcador **A** con alelos A y a, donde p_A denota la frecuencia alélica de A. Si este SNP no está en desequilibrio de ligamiento con un segundo marcador **B** que tiene alelos B y b con frecuencia alélica p_B para el alelo B, entonces la frecuencia p_{AB} del haplotipo AB es igual al $p_A p_B$. Esto es debido a que segregan de manera independiente, es decir, un evento es independiente del otro por lo que la probabilidad de que aparezcan los dos es igual al producto de sus respectivas probabilidades.

La medida más simple de LD es D, la cual se define como la diferencia entre la frecuencia observada de un haplotipo de dos loci y la frecuencia esperada (basandose en frecuencias alélicas) si los alelos fueran segregados aleatoriamente. Entonces, para este caso se define $D = D_{AB}$ y su medida de LD:

$$D_{AB} = P_{AB} - P_A P_B. \tag{1.2}$$

Si D es significativamente diferente de cero entonces se dice que existe LD. La significancia estadística se verifica utilizando la prueba exacta de Fisher. Que sea positivo o negativo sólo depende del etiquetado arbitrario de los alelos. En la tabla 1.1, se resumen las probabilidades de posibles haplotipos entre el marcador **A** y el marcador **B**, donde si hay desequilibrio de ligamiento las frecuencias haplotípicas son modificados por D. Las frecuencias de las combinaciones específicas de alelos son representadas por p_{11}, p_{12}, p_{21} y p_{22}.

	Marcador **B**		
marcador **A**	B	b	Total
A	$p_{11} = p_A p_B + D$	$p_{12} = p_A(1 - p_B) - D$	p_A
a	$p_{21} = (1 - p_A)p_B - D$	$p_{22} = (1 - p_A)(1 - p_B) + D$	$1 - p_A$
Total	p_B	$1 - p_B$	1

Tabla 1.1: Tabla de probabilidades haplotípicas para dos marcadores bialélicos

En base a la tabla 1.1 se puede ver que D es la covarianza entre dos marca-

dores dialélicos.

$$\begin{aligned} D_{AB} &= p_{11} - p_A p_B \\ &= p_{11} - (p_{11} + p_{12})(p_{11} + p_{21}) \\ &= p_{11} - p_{11}(p_{11} + p_{21} + p_{12}) - p_{12}p_{21} \\ &= p_{11}p_{22} - p_{12}p_{21} \end{aligned} \tag{1.3}$$

Aunque esta definición es intuitiva, la dependencia en las frecuencias alélicas hace que las comparaciones entre diferentes valores de D tengan una utilidad limitada. Los valores de D para dos pares de loci genéticos sólo son comparables si las frecuencias alélicas son similares. Para superar esta dificultad se han sugerido distintas formas de estandarizar D. Una de éstas es la medida $|D'|$, la cual es el valor absoluto de D dividido entre el máximo valor posible dadas las frecuencias alélicas entre dos loci. Es decir,

$$D' = \frac{D}{D_{\max}} \tag{1.4}$$

donde

$$D_{\max} \leq \begin{cases} \max\{p_A q_B, q_A p_B\} & \text{si } D_{AB} > 0 \\ -\min\{p_A p_B, q_A q_B\} & \text{si } D_{AB} < 0. \end{cases}$$

Tiene como propiedad, que $|D'| = 1$ si y solo si dos alelos no han sido separados por recombinación durante la historia en la muestra analizada. Este caso se conoce como LD completo. Los valores de $|D'|$ pueden estar inflados cuando la muestra es pequeña, y cuando las frecuencias de alelo menor son bajas puede causar indicadores falsos de estado de LD.

Otra medida es r^2, el cuadrado del coeficiente de correlación entre dos loci, que se obtiene al dividir D^2 entre el producto de las cuatro frecuencias alélicas en esos dos loci, es decir,

$$r^2 = \frac{(p_{11}p_{22} - p_{12}p_{21})^2}{p_A(1-p_A)p_B(1-p_B)}. \tag{1.5}$$

Esta medida tiene varias propiedades de utilidad que la han hecho el parámetro más popular cuando se trata de comparar LD en estudios de asociación genética a enfermedades. El valor $r^2 = 1$ representa perfecto desequilibrio de ligamiento. Sucede si y solo si los alelos no han sido separados debido a recombinación y tienen la misma frecuencia alélica. La fórmula de r^2 presenta menos inflación que $|D'|$ cuando la muestra es pequeña.

1.5. Proyectos internacionales

1.5.1. Proyecto del genoma humano

El proyecto del genoma humano [10] inició formalmente en 1990, y fue un esfuerzo de 13 años. Fue coordinado por el Departamento de Energía de EE.UU.

y los Institutos Nacionales de Salud de EE.UU. El proyecto originalmente fue planeado para durar 15 años, pero los rápidos avances tecnológicos aceleraron la fecha de finalización para 2003. Tuvo como objetivos: identificar todos los aproximadamente 20,000 - 25,000 genes del ADN humano, determinar las secuencias de los 3 mil millones de pares de bases químicas que componen el ADN humano, almacenar esta información en bases de datos, mejorar las herramientas de análisis de datos, transferir tecnologías relacionadas con el sector privado, y abordar las cuestiones éticas, jurídicas y sociales (ELSI) que puedan surgir del proyecto.

Una vez concluido en 2003, el siguiente paso es utilizar este vasto depósito de datos para explorar cómo el ADN y las proteínas trabajan juntos e interactuan con el ambiente para crear sistemas complejos y dinámicos de vida. Algunas aplicaciones actuales y potenciales de la investigación del genoma incluyen:

- **La medicina molecular.** A partir del proyecto del genoma humano se desarrollaron nuevas tecnologías y fuentes de información, que han tenido un gran impacto en la investigación biológica y medicina clínica. Los mapas de genoma que se han desarrollado con gran detalle han facilitado la búsqueda de genes asociados a docenas de condiciones genéticas.
- **Las fuentes de energía y aplicaciones ambientales.** La información obtenida de la caracterización de genomas completos microbianos conduce a ideas sobre el desarrollo de nuevas biotecnologías relacionadas con la energía, los sistemas fotosintéticos, los sistemas microbianos que funcionan en ambientes extremos, y organismos que pueden metabolizar fácilmente los recursos renovables disponibles y desechar material con la misma facilidad.
- **Evaluación de riesgos.** La comprensión del genoma humano tiene un enorme impacto en la capacidad de evaluar los riesgos que se plantean a personas por la exposición a agentes tóxicos. Se sabe que las diferencias genéticas hacen a algunas personas más susceptibles y a otras más resistentes a dichos agentes.
- **Bioarqueología, la antropología, la evolución y la migración humana.** La comprensión de la genómica ayuda a entender la evolución humana y la biología común que se comparte con el resto de los organismos. La genómica comparativa entre los seres humanos y otros organismos tales como ratones ya ha llevado a genes similares asociados a enfermedades y rasgos. Otros estudios comparativos ayudan a determinar la función aún desconocida de miles de genes.
- **Forense de ADN (identificación).** Cualquier tipo de organismo puede ser identificado por medio del análisis de secuencias de ADN únicas para esa especie.
- **La agricultura, la ganadería, y bioprocesamiento.** Comprender los genomas de plantas y animales permite crear plantas y animales más fuertes, más resistentes a las enfermedades, permite la reducción de costos a

la agricultura así como proporcionar a los consumidores alimentos más nutritivos y libres de pesticidas.

1.5.2. Proyecto internacional de HapMap

El HapMap es un catálogo de las variantes genéticas comunes que ocurren en los seres humanos. Describe cómo se comparten entre las personas dentro de las poblaciones y entre poblaciones en diferentes partes del mundo. El Proyecto Internacional HapMap [12] no tiene como propósito establecer conexiones entre determinadas variantes genéticas y enfermedades. Por el contrario, el proyecto de HapMap tiene como objetivo principal la creación de una herramienta que facilite los estudios de asociación entre factores genéticos y enfermedades comunes. Este proyecto junto con el Proyecto del genoma humano proporcionó información de aproximadamente 10 millones de variantes comunes, en su mayoría SNPs. La información que proporcionan estos SNPs junto con sus patrones de LD han facilitado, por medio de los estudios de asociación de genoma completo, la identificación de cientos de regiones genómicas nuevas que contribuyen al desarrollo de enfermedades [13].

El proyecto de HapMap se inició oficialmente en octubre de 2002 con financiamiento de Canadá, China, Estados Unidos, Inglaterra y Japón. La meta fue realizar un mapa completo de haplotipos de 270 personas formados por 90 africanos de grupo Yoruba de Nigeria; 90 asiáticos que incluyen a 45 japoneses y 45 Chinos del grupo Han; y 90 estadounidenses de Salt Lake City con ancestros de norte y oeste de Europa. Los datos están disponibles de manera gratuita en internet [51].

1.5.3. Proyecto de 1000 genomas

El proyecto HapMap Internacional proporcionó un catálogo de variantes genéticas comunes y sus patrones de correlación, a lo largo de varias poblaciones en 3.5 millones de SNPs. Para el 2008, el catálogo público de los variantes (dbSNP 129) contenía aproximadamente 11 millones de SNPs y 3 millones de inserciones y deleciones cortas (indeles). Estas herramientas ayudaron a que la primera generación de descubrimientos de genes asociados a enfermedades se cumpliera. Para tener una comprensión profunda de la contribución genética a los fenotipos todavía falta mucho por hacer. Una vez que se ha identificado una región de riesgo, se tiene que se requiere estudiar detalladamente todas las variantes dentro de ese locus para poder determinar cuáles son las variantes causales, cuantificar su contribución a la susceptibilidad de la enfermedad, y visualizar su papel en las vías funcionales. Las variantes de frecuencia baja y de frecuencia rara (0.5 % a 5 % MAF[4], y menores a 0.5 % de MAF, respectivamente) sobrepasan en número a las variantes comunes, y también contribuyen significativamente en la arquitectura genética de las enfermedades, pero estas todavía no han sido estudiadas sistemáticamente [11]. Como prerrequisito para

[4]MAF significa, por sus siglas en inglés, frecuencia de alelo menor.

entender completamente el papel de las variantes comunes y de baja frecuencia, se necesita un catálogo de variación del DNA humano más completo. Ésta es la misión que el Proyecto de 1000 genomas tiene.

El plan para el proyecto completo es obtener la secuencia de cerca de $2,500$ muestras a una alta cobertura. El primer grupo de muestras para la secuenciación de muestras incluye $1,167$ que ya existían o que podían ser recolectados con rapidez, a partir de 13 poblaciones, para la secuenciación en 2010 y principios de 2011. El segundo conjunto incluye 633 muestras que se están recolectando, a partir de 7 poblaciones, para la secuenciación de principios de 2011. El tercer conjunto, que consta de 700 muestras, que están disponibles para la secuenciación desde finales de 2011.

Los propósitos principales del Proyecto de 1000 genomas son los siguientes:

- Descubrir, genotipificar[5] y proporcionar información precisa de los haplotipos y todas las formas de polimorfismo de ADN humano en múltiples poblaciones humanas.
- Caracterizar más del 95 % de las variantes que se encuentran en regiones genómicas accesibles a las actuales tecnologías de secuenciación de alto rendimiento y que tienen frecuencia alélica de 1 % o superior en cada uno de los cinco principales grupos de población (poblaciones de ascendencia de Europea, Asia Oriental, Asia Meridional, África Occidental y América).
- También incluir alelos de menor frecuencia (abajo a 0.1 %) en regiones codificantes, ya que muchos los alelos funcionales tienen frecuencias alélicas bajas y se encuentran a menudo en estas regiones.

Entre los beneficios que proporciona el Proyecto de 1000 genomas se encuentra la posibilidad de imputación. Por imputación nos referimos al proceso de substitución de valores de genotipos faltantes. Cuando se están estudiando enfermedades se pueden utilizar datos de 1000 genomas en dicho proceso. De tal manera, se combinan los datos de 1000 genomas con los datos de genotipo del estudio para obtiener millones de variantes adicionales de las que directamente se genotiparon. Los datos imputadas permiten localizar las regiones asociadas a la enfermedad con mayor precisión, y ahorran enormes cantidades de dinero a los investigadores ya que no tienen que genotipar directamente las variantes adicionales. Una vez que se identifica la region de interés, 1000 genomas proporcionará datos sobre casi todas las variantes con una frecuencia de al menos 1 % en las poblaciones estudiadas. Esto ahorrará tiempo y dinero al no tener que

[5]Genotipificar significa caracterizar y analizar el genotipo (la constitución genética) de un organismo, en uno o más locus y por medios diversos (genéticos, moleculares, inmunológicos, etc.), utilizando células, tejidos u organismos enteros. **Observación:** en castellano el verbo tipificar lo registra la vigésima segunda edición del diccionario académico de la lengua española con tres acepciones de uso distintas, una de las cuales, la que más se aproxima a la idea de caracterización, es la segunda: "dicho de una persona o de una cosa: Representar el tipo de la especie o clase a que pertenece" (en el sentido de que una persona o una cosa caracteriza o representa el tipo de la clase a la que pertenece). Si bien el verbo genotipificar no se utiliza en ese sentido, no resulta difícil imaginar cómo se ha popularizado en la práctica con el significado que aquí se indica (caracterización del genotipo) [31].

secuenciar sus propias muestras. Aunque no se determine exactamente qué variantes causan del aumento del riesgo de la enfermedad, sí reducen la lista de posibilidades.

Otro beneficio es que ayuda a comparar las frecuencias alélicas y los patrones de LD que se encuentran en los estudios propios con los de 1000 genomas. También se puede utilizar los datos de 1000 genomas para estudiar la recombinación, la selección natural y la estructura de la población.

Capítulo 2

Estudios de asociación de genoma completo

Este capítulo se enfoca en desarrollar los conceptos básicos para entender cómo se efectúa un estudio de asociación de genoma completo. El propósito de esto es dar un panorama general en lo que se refiere a la disección genética de las enfermedades complejas. Se comienza con una introducción a las estrategias que existen para indentificar marcadores genéticos asociados a enfermedades: análisis de ligamiento y análisis de asociación.

Una vez definidos los estudios de asociación, se comienza a describir a los estudios de asociación de genoma completo como tal. Se dedica una sección completa a cada uno de los pasos que involucran estos estudios. En la siguiente sección se describen los distintos tipos de controles de calidad. Posteriormente se describe los tipos de medidas de asociación que existen y los tipos de prueba estadísticas que se utilizan para realizar los análisis de asociación genética. Después se hace enfásis en dos fenómenos de gran influencia en los estudios de asociación: estratificación poblacional y relaciones crípticas.

Por último se describen los métodos existentes para determinar el nivel de significancia en los estudios de asociación de genoma completo.

2.1. Estrategias para identificar marcadores genéticos asociados a enfermedades

La susceptibilidad que se tiene a una enfermedad tiene cierta base genética, de aquí surge el interés para buscar genes relevantes. El mejor enfoque depende principalmente de la fuerza de la susceptibilidad genética, la cual es revelada si la enfermedad muestra evidencia de herencia mendeliana o no.

Dado que la mayoría de las enfermedades tienen un componente genético involucrado, es muy importante identificar genes relevantes porque nos ayudan a entender el mecanismo biológico detrás del desarrollo y progresión de las

mismas. Hay enfermedades monogénicas cuyo componente genético es determinante y hay enfermedades complejas en donde intervienen factores genéticos y ambientales que al interactuar confieren susceptibilidad.

Para los trastornos mendelianos, ya es rutina proceder con un análisis de ligamiento. Los análisis de ligamiento identifican la ubicación física aproximada del gen responsable en un cromosoma y para ello se utilizan individuos relacionados.

Si la enfermedad no muestra patrón de herencia mendeliana, va a ser difícil identificar genes relevantes [37]. Como opción más atractiva están los análisis de asociación en individuos no relacionados.

A continuación se describiran los conceptos básicos que subyacen al análisis de ligamiento y de asociación.

2.1.1. Análisis de ligamiento

Ligamiento se refiere a la existencia de una conexión entre dos loci en un mismo cromosoma que están suficientemente cerca para que sus alelos cosegreguen. El ligamiento mide desviaciones de las leyes de Mendel de segregación independiente.

El principio del análisis de ligamiento es simple. Todos los cromosomas vienen en pares, uno es heredado del padre y otro de la madre. Cada par de cromosomas contiene los mismos genes en el mismo orden pero las secuencias no son idénticas. Esto significa que debería de ser posible distinguir qué secuencia viene de la madre y qué secuencia viene del padre. A estas variantes se les llama alelos maternos y paternos.

En el caso del gen responsable de la enfermedad, se asume que el gen es bialélico (alelo normal y alelo de enfermedad) y que tiene un patrón de herencia mendeliano. La habilidad para determinar el origen paterno o materno de una secuencia de ADN en la descendencia nos permite determinar si hay eventos de recombinación.

Básicamente el análisis de ligamiento intenta explicar si hay cosegregación de los fenotipos y los genotipos observados en un pedigree. Es el método ideal cuando se estudian rasgos simples con patrón de herencia mendeliano y penetrancia arriba del 80 %. Sin embargo, aplicarlos a rasgos complejos puede ser más problemático ya que puede ser díficil de encontrar un modelo preciso que explique de manera adecuada los patrones de herencia.

Los genes de enfermedad son mapeados al medir la tasa de recombinación contra un panel de diferentes marcadores a lo largo de todo el genoma. Si la tasa de recombinación es igual a 0.5 entre el gen de la enfermedad y el marcador entonces se infiere que estos están lejos; mientras que si es menor a 0.5 entonces se dice que están ligados debido a su proximidad. De manera ideal, se identifican marcadores cercanos que flanquean el gen del la enfermedad y se define una región candidata del genoma de 1 a 5 megabases de longitud.

El análisis de ligamiento consiste en comparar el modelo M_1, que postula una ubicación específica de un gen causante de enfermedad, con un modelo M_0, el cual corresponde a la hipótesis nula de no ligamiento al gen causante de

enfermedad en la región. Dicha comparación se realiza por medio de la razón de verosimilitudes

$$LR = \frac{P(\text{datos}|M_1)}{P(\text{ datos }|M_0)}$$

o equivalentemente con el *lod score* $z = \log_{10}(LR)$ [37].

El modelo M_1 depende de parámetros como la frecuencia alélica del marcador y del gen de la enfermedad, la función de penetrancia y la probabilidad de transmisión de padres a hijos. Las frecuencias alélicas y la función de penetrancia son conocidas, por lo que el modelo es una función que depende sólo de la fracción de recombinación. El modelo M_0 bajo la hipótesis de no ligamiento supone que la fracción de recombinación es igual a 0.5.

El modelo de máxima verosimilitud M_1 es aceptado (en contraste con M_0) cuando el valor de la fracción de recombinación que maximiza la función es al menos 1000 veces más verosímil que el valor bajo la hipótesis nula. Es decir, se busca que $z \gg 3$.

2.1.2. Análisis de asociación

Se observa asociación genética en un alelo específico si es más frecuente en un grupo de afectados que en un grupo de no afectados.

Los estudios de asociación se diferencian de los estudios de ligamiento en varios aspectos. Por una parte, los primeros se enfocan en estudiar la relación de un alelo con una enfermedad, mientras que los segundos se enfocan en identificar una región o locus que cosegrega con la enfermedad. Por otro lado, los primeros se enfoca en estudiar poblaciones y los segundos estudian la transimición basándose en familias.

Aunque existen estudios de asociación genética en familias [55], el diseño común para este tipo de análisis utiliza individuos no relacionados y pertenecientes a la misma población. Por no relacionados, se refiere a que las relaciones son desconocidas o muy distantes, de tal manera que no se puede rastrear transimiciones de fenotipos a lo largo de las generaciones.

Las asociaciones dependen de la historia de la población. Supongamos que dos individuos no relacionados heredan de un ancestro común un alelo con susceptibilidad a una enfermedad. Durante muchas generaciones y muchas meiosis que los separan de ese ancestro común, el efecto de múltiples recombinaciones generan un reducción del segmento cromosómico. Por lo que sólo se comparte una región muy pequeña que está en alto desequilibrio de ligamiento con el locus susceptible a la enfermedad. Para un locus que tiene una fracción de recombinación θ con el locus susceptible, una proporción θ de cromosomas ancestrales se va a perder en cada generación y permanecerá una proporción $(1-\theta)$. Después de n generaciones, una fracción de $(1-\theta)^n$ de cromosomas conservará la asociación. Este cálculo es algo crudo pero sí muestra cómo las asociaciones alélicas dependen de la población estudiada. Debido a que el desequilibrio de ligamiento es un fenómeno de corto alcance, si se encuentra una asociación, ésta define una pequeña región candidata para buscar la susceptibilidad de un gen.

Posibles causas de asociación

- **Causa directa.** Tener el alelo A causa susceptibilidad a la enfermedad D. La posesión del alelo A no es necesaria ni suficiente para desarrollar la enfermedad D. Sin embargo, sí incrementa la probabilidad de desarrollarla. En este caso se espera que sea el mismo alelo asociado a la enfermedad en cualquier población.
- **Selección natural.** Personas con la enfermedad D tienen mayor probabilidad a sobrevivir y tener descendientes si tiene el alelo A.
- **Estructura poblacional.** La población contiene varios subconjuntos. Y el alelo A y la enfermedad D coinciden en ser frecuentes en un subconjunto.
- **Artefacto estadístico.** Debido a las múltiples pruebas estadísticas que se hacen, si no se aplican los niveles de significancias adecuados puede que se encuentren asocaciones espurias que no se repliquen en los análisis subsecuentes.
- **Desequilibrio de ligamiento.** La proximidad puede producir asociación alélica a nivel poblacional, consecuencia de que casi todos los cromosomas portadores de la enfermedad en la población provienen de uno o pocos cromosomas ancestrales. Si el desequilibrio de ligamiento es la causa de asociación, debería de haber algún gen cercano al locus $\widehat{A}$ que tiene mutaciones que contribuyen con el desarrollo de la enfermedad D. El alelo particular A en el locus $\widehat{A}$ que es asociado con la enfermedad D puede ser diferente en distintas poblaciones.

La causa directa y la selección son poco probables si el alelo asociado es una variante en el ADN no codificante y no está cercanamente asociado a algún gen. Los artefactos estadísticos se previenen utilizando métodos adecuados para la determinación del nivel de significancia [56].

Tipos de estudios de asociación

Existen distintos tipos de estudios de asociación, los cuales se pueden clasificar de la siguiente manera:

- **Polimorfismos candidatos:** Estos estudios se enfocan en estudiar un polimofismo particular que sea sospechoso de estar implicado en la enfermedad.
- **Gen candidato:** Estos estudios se enfocan en una cantidad de genes escogidos en base a análisis de ligamiento previos, o en base a caraterísticas funcionales. En dichas regiones se procura incluir la secuencia codificante y las regiones de flanqueo. En estos estudios se pueden incluir de $5 - 50$ SNPs que se encuentran dentro de un gen, hasta incluir cientos de genes, tantos que el estudio se puede asemejar a un análisis de genoma completo.

- **Fine mapping:** Son análisis basados en una región previamente identificada que va de $1 - 10$ Mb y pueden incluir varios cientos de SNPs. Busca determinar el alelo causal. La region candidata puede ser indentificada por medio de un estudio de ligamiento o de asociación y puede llegar a incluir de $5 - 50$ genes.

- **Genoma completo:** Tiene como objetivo identificar variantes asociadas a lo largo del genoma, y requiere una cantidad de al menos $300,000$ SNPs. La genotipación de dicha cantidad de marcadores es posible gracias a proyectos de como el Proyecto Internacional de HapMap y los avances tecnológicos para la caracterización de genotipos.

Debido a los propósitos de este texto, las siguientes secciones se concentran sólo en los estudios de asociación de genoma completo.

2.2. Estudios de asociación de genoma completo

Los estudios de asociación de genoma completo, o GWAS por sus siglas en inglés (Genomewide Association Studies), son un método poderoso para identificar genes con susceptibilidad a ciertas enfermedades. Por medio de GWAS se han podido identificar una gran cantidad asociaciones robustas entre loci específico y enfermedades humanas, y en los últimos años se ha convertido en un método popular para la identificación de variantes asociadas a enfermedades. Este análisis requiere de una exploración en miles de muestras, de miles de marcardores localizados a lo largo de todo el genoma humano. Normalmente como marcadores genéticos se utilizan SNPs porque son fáciles de etiquetar y abundan en el genoma humano [63].

Este enfoque se basa en la base de datos producidos por el Proyecto Internacional HapMap. La estructura haplotípica del genoma humano significa que es posible estudiar la variabilidad común del genoma asociada con el riesgo de enfermedades con tan solo genotipificar unos $500,000$ marcadores elegidos cuidadosamente el genoma en varios miles de individuos.

El poder de detectar la asociación entre una variante genética y la enfermedad es una función de varios factores, incluyendo la frecuencia del alelo de riesgo o genotipo, el riesgo relativo conferido al alelo o genotipo asociado a la enfermedad asociada, la correlación entre marcador genotipificado y el alelo de riesgo, el tamaño de la muestra, la prevalencia de la enfermedad y la heterogeneidad genética de la muestra de la población. Mientras que los tres primeros factores son desconocidos antes de realizar el GWAS, su impacto puede ser influenciado por el diseño del estudio.

El éxito que se vaya a obtener en un GWAS depende de si se genotipificó directa o indirectamente un polimorfismo causal. Que se obtenga directamente quiere decir que se genotipificó el polimorfismo causal. Cuando se dice que se obtiene indirectamente es porque marcadores genéticos cercano y altamente correlacionados con el polimorfismo causal fueron etiquetados.

Se dice que los alelos en dos o más loci están en desequilibrio de ligamiento (LD) si estos están correlacionados o están asociados de forma no aleatoria. Que tengan ancestría común se refiere a que los alelos cercanos a cierto loci tienden a ser heredados juntos en el mismo cromosoma, con una combinación específica de alelos que se conocen como haplotipos. En los GWAS, convencionalemente se etiquetan SNPs comunes a alta densidad a lo largo del genoma, y aunque es raro que se etiquete la variante causal, sí es probable que se haya etiquetado una que esté en alto desequilibrio de ligamiento con ésta. En algunos casos puede que la variante causal ni siquiera sea un SNP sino que sea una inversión, inserción, deleción o variante de número de copia [3].

Los GWAS tiene varias ventajas sobre otros métodos de descubrimiento de genes de enfermedades. En comparación con los estudios de genes candidatos, los GWAS permite un análisis completo del genoma de una manera imparcial y por lo tanto tienen la posibilidad de identificar factores de susceptibilidad totalmente nuevos.

En comparación con los análisis de ligamiento en familias los estudios de asociación tienen dos ventajas fundamentales. En primer lugar, son capaces de sacar provecho de todos los eventos de recombinación meiótica en una población, y no sólo los de las familias estudiadas. Debido a esto, las señales de asociación son localizadas en pequeñas regiones del cromosoma que contiene sólo un gen de un conjunto de unos cuantos genes, lo que permite la rápida detección del verdadero gen con susceptibilidad a la enfermedad. En segundo lugar, los GWAS permiten la identificación de los genes de la enfermedad con sólo un modesto incremento en el riesgo, la cual es una limitación grave en los estudios de ligamiento.

Gracias a estas ventajas, los GWAS puede identificar múltiples genes de la enfermedad que en muchos casos interactúan entre sí, dando así una comprensión integral de la etiología[1] de la enfermedad.

En los estudios de caso-control es fundamental considerar que los individuos que conforman los grupos de casos y de controles no proveen estimaciones de frecuencias alélicas sesgadas sobre la distribución subyascente a los grupos de individuos afectados y no afectados. De lo contrario, los hallazgos van a ser consecuencia de los sesgos causados por el diseño del estudio.

2.2.1. Etapas del GWAS

Para realizar un estudio de GWAS, primero se tiene que seleccionar la enfermedad o rasgo que sea apropiada para el estudio. Es más probable tener un análisis existoso si el fenotipo de intéres puede ser medido o diagnosticado de manera específica. Posteriormente, se requiere de recolectar cientos de muestras de individuos, que involucren tanto casos como controles. Luego se tienen que genotipificar miles de SNPs a lo largo de todo el genoma, para lo cual se utilizan

[1] Etiología se refiere a el estudio de las causas de las enfermedades.

microarreglos de ADN[2] producidos por Illumina o Affymetrix[3]. Aunque los marcadores SNP pueden llegar a tener hasta cuatro alelos nucleótidos, normalmete se utilizan dialélicos ya que su tasa de mutación es baja [3].

Una vez ya producida la base de datos de SNPs, estos son sometidos a controles de calidad. Después se realiza un análisis de asociación con la enfermedad o rasgo a los SNPs que sobrevivieron los controles de calidad. Los resultados usualmente se visualizan por medio de las gráficas *Manhattan*, las cuales grafican el logaritmo negativo de los valores de p contra la posición cromosómica.

Debido a la gran cantidad de pruebas estadísticas que se realizan, existe una gran cantidad de posibles falsos positivos. Dependiendo del número de pruebas se determina la significancia estadística; usualmente se reportan como estadísticamente significativos los valores de p menores o iguales a 5.0×10^{-8} [30].

Los pasos mencionados anteriormente se resumen en los siguientes puntos y se desarrollaran a más detalle en las siguientes secciones.

- **Diseño del análisis:** Escoger la enfermedad y la población a estudiar.
- **Variable:** Hacer selección de los marcadores genéticos (SNPs).
- **Tecnología:** Escoger el tipo de microarreglo que se va a utilizar.
- **Control de Calidad:** Exclusión de SNPs de baja calidad, exclusión de individuos con baja tasa de genotipificado, etc.
- **Análisis estadístico:** Hacer selección del modelo genético, si es dominante, recesivo o aditivo, seleccionar el tipo de prueba estadística a realizar y determinar el nivel de significancia.

2.3. Controles de calidad

Los datos utilizados en GWAS son obtenidos a través de múltiples procesos de escala industrial. Debido a la gran manipulación que sufren las muestras antes de obtener la base de datos final, realizar controles de calidad se vuelve una actividad necesaria. Existen dos tipos de controles de calidad: los que se realizan durante la producción de los datos y los que se realizan posproducción. Durante la producción, los controles de calidad consisten en una serie pasos que se llevan a cabo para monitorear y controlar la calidad del producto que se está fabricando y los posproducción consisten en una revisión de la calidad del producto, donde el *producto* es el conjunto de datos a utilizar para detectar asociaciones de genotipo-fenotipo.

[2]Un microarreglo de ADN es una colección de puntos microscópicos de ADN posicionados en una superficie sólida. Se usan para determinar el genotipo en múltiples regiones del genoma. También se conoce como chip de ADN.

[3]Illumina y Affymetrix son compañías manufacturadoras de microarreglos de ADN.

La necesidad de realizar controles de calidad, los problemas que bajan la calidad del producto, han sido estudiados por diversos autores [2, 7, 38]. Entre estos problemas están las diferencias entre las frecuencias alélicas que hay entre casos y controles. Debido a que los GWAS suele implicar grandes tamaños de muestras para detectar efectos pequeños y cientos de miles de polimorfismos son estudiados, existe la posibilidad en las que pequeñas diferencias de frecuencia alélica puedan generar resultados falsos-positivos. Diferencias en la estructura poblacional y el hecho de que en el mismo conjunto de individuos a estudiar se encuentren individuos de distintas poblaciones puede ocasionar que existan alelos con mayor frecuencia en una subpoblación que en la otra causando así falsos positivos. Diferencias de calidad de ADN entre las muestras pueden causar diferencias entre las frecuencias de genotipos. También cuando los estudios contienen individuos relacionados, los métodos que asumen implícitamente la independencia entre los sujetos puede inflar la cantidad de resultados falsos positivos. Los controles de calidad posproducción juegan un papel importante para la identificación de sesgos que pueden ser reducidos o eliminados durante la fase de análisis de GWAS [7].

Los problemas que enfrentan los controles de calidad se pueden dividir en dos secciones: específicos de SNP y específico de individuos. Antes de continuar con la descripción los distintos tipos de controles de calidad, es importante discutir si existe un posible orden de estos. Este problema se puede plantear en términos de qué tipo de datos se prefieren perder en mayor proporción. Es decir, cuando uno filtra individuos "malos" uno tiende a salvar SNPs que peligraban de ser eliminados, y vice versa, si uno filtra primero los SNP "malos". Sin embargo, si se piensa que se tienen cientos de miles de SNPs contra sólo unas cuantas centenas de individuos, es justo pensar el problema en términos de proporción de datos perdidos y ver qué es peor: perder 5 % de los individuos o 5 % de los SNPs, por decir un ejemplo. Al perder el 5 % de los individuos se va a perder potencia para detectar las señales de asociación. La relación entre tamaño de muestra y poder de la señal es sigmoidal, por lo que para algunas señales la pérdida de potencia va a ser despreciable, mientras que para otras puede ser considerable. Por otra parte, perder el 5 % de los SNPs puede ser catastrófico si una de las señales de asociación es mucho más notable en uno de los SNPs perdidos. Sin embargo, este efecto es mitigado por la existencia de desequilibrio de ligamiento que tiene este con SNPs vecinos que posiblemente sí hayan sobrevivido a los controles de calidad. Por lo tanto, la decisión acerca del orden que deben tomar los controles de calidad depende de cada proyecto en particular, depende si la base de datos tiene suficiente potencia estadística, si la población tiene niveles altos de desequilibrio de ligamiento, si el panel de SNP es de alta densidad, etc...

2.3.1. Gráficas de $Q - Q$

Al realizar la prueba estadística respectiva para el análisis de asociación se obtiene el valor de p para cada SNP en la base de datos. Los valores de p generados bajo la hipótesis nula deberían, por definición, ser extraídos de una

distribución uniforme entre 0 y 1. Se sigue que si un conjunto de m valores de p son ordenados del menor al mayor, entonces el valor del cuantil observado del j-ésimo valor ordenado debería ser, en general, igual al cuantil esperado correspondiente a m valores tomados al azar de una distribución Uniforme$(0, 1)$ (se puede probar que este valor esperado es $j/(m + 1)$). Por lo tanto, al hacer la gráfica de $Q - Q$ de los cuantiles observados contra los cuantiles esperados, se esperaría observar una línea recta tosca que empieza en el origen y tiene pendiente unitaria, incluyendo un poco de variación aleatoria a lo largo de ésta. Cabe señalar que se espera este patrón bajo la hipótesis nula, aún cuando existe cierta dependencia entre los valores de p[4] (por ejemplo, debido al desequilibrio de ligamiento local que hay entre SNPs). En este caso, la expectativa general es la misma aunque se espera que presente un poco de inflación sobre la línea unitaria. En un estudio de análisis de asociación, se espera que la gráfica $Q - Q$ se parezca lo más posible a la línea con pendiente unitaria pero que al final existan unos cuantos puntos que se despegan de la línea. Estos puntos se conocen como *hits*, son los SNPs que no cumplen con la hipótesis de no asociación.

Cuando existe un conjunto de m valores de p contiene elementos que no son extraídos de la hipótesis nula sino de la hipótesis alternativa, la distribución de los valores de p va a estar sesgada de una distribución Uniforme$(0, 1)$. Y estos valores son los que se separan de la línea recta al ser graficados. Estos valores de p son muy bajos, al transformar con logaritmo se enfatiza más los valores pequeños, por lo que en la práctica es común graficar los valores de p en una escala logarítmica negativa. Resulta que este escalamiento es equivalente a convertir la distribución esperada de una Uniforme$(0, 1)$ en una χ^2 con dos grados de libertad [60].

2.3.2. Tasa de información faltante

Las pérdidas de datos son uno de los grandes problemas en los controles de calidad de los GWAS. El hecho que exista cierta cantidad de información faltante en los genotipos puede ser indicador de mala calidad de la muestra. Si se omite este paso, se pueden obtener resultados falsos positivos. Estos surgen si la calidad del ADN difiere con el fenotipo, llevando a diferencias en las frecuencias de los genotipos. El aumento de falsos positivos es común en los estudios de casos y controles, donde se acostumbra recolectar y/o genotipificar de forma separada las muestras de casos y controles y son menos comunes en los estudios que tratan con fenotipos cuantitativos. Por otra parte, los falsos negativos surgen si las señales actúan en la dirección opuesta a la señal real o cuando se reduce la potencia al reducir el tamaño de la muestra para los valores no perdidos.

La tasa de información faltante se analiza en base a individuos y en base a SNPs. Es uno de los pasos obligados de los controles calidad debido a la fuerte

[4]Valor de p es el nivel más pequeño de significancia α para el cual la información observada indica que la hipótesis nula debe de ser rechazada.

correlación de las pérdidas de información con la calidad de los SNPs y el impactos de las pérdidas informativas de señales de asociación. Para ilustrar estos controles de calidad, en la figura 2.1 se grafica la tasa de información faltante a nivel de SNPs e individuos. Se graficó uno menos la proporción de SNPs faltantes por individuo y la proporción de muestras faltantes por SNP, respectivamente. El punto donde se quiebran las líneas de las gráficas denota el punto límite en el que no se gana más al aumentar la rigurosidad de los controles de calidad. Es decir, al ser más exigentes con los controles de calidad se va a perder un mayor número de SNPs o individuos y el cambio en cuanto a la eficacia de los controles de calidad va a ser mínimo. Por lo general, este punto ocurre entre un 97 % o 98 % de la tasa de ocurrencia total de SNPs en individuos, y en el 95 % del grado total covertura de SNPs. Estos valores se utilizan como guías para determinar los valores que se emplearán en el análisis.

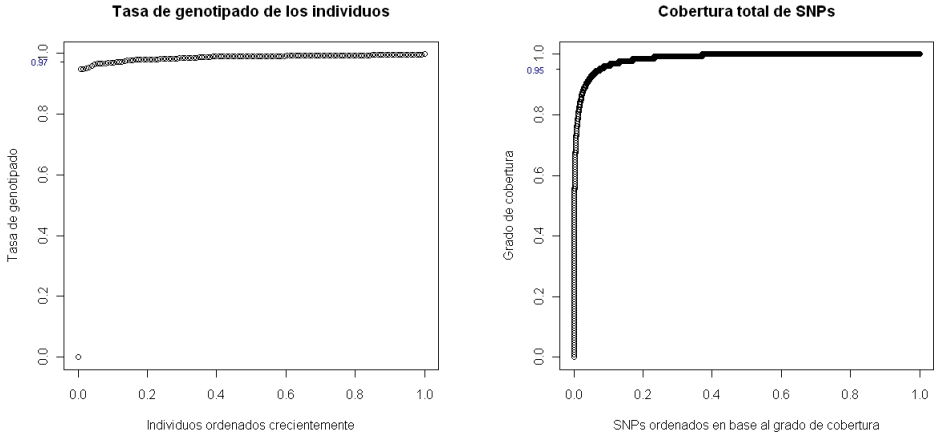

Figura 2.1: Gráficas del complemento de la tasa de información faltante (uno menos la tasa de información faltante) contra la frecuencia cumulativa (datos ordenados de menor a mayor)

2.3.3. Comprobación de género

Con los datos de GWAS, utilizando los datos de los cromosomas X o Y, es fácil detectar los individuos que son genéticamente hombres pero que fueron etiquetados como mujer, o vice versa. Normalmente se utilizan sólo los datos del cromosoma X ya que en muchos estudios no se acostumbra genotipificar el cromosoma Y.

Los hombres sólo portan un cromosoma X por lo que sus genotipos hemicigotos (sólo poseen un alelo) son codificados como si fueran homocigotos. Por

el contrario, las mujeres poseen dos cromosomas X por lo que se espera que sean muy heterocigotas. En base a esto, el procedimiento que se utiliza para filtrar esta clase de errores utilizando los datos del cromosoma X es detectando el grado de heterocigosidad. Esto se mide utilizando el estadístico F definido en la sección 1.3.4 utilizando la fórmula 1.1 donde se obtienen valores cercanos a cero para mujeres y valores cercanos a uno para hombres [52, 60].

Uno de los usos de esta herramienta es para asegurarse que las bases de datos de los fenotipos y de los genotipos estan alineadas correctamente. Si cerca del 50 % de las muestras no coinciden, hay evidencia de que existe una catastrófica aleatoriedad de las etiquetas de los individuos en alguno de los conjuntos de muestras. En otras ocasiones, aún cuando las bases de datos están bien alineadas, se puede indentificar y evaluar el número de errores que existen. En muchos casos, pero no siempre, es recomendable remover estos individuos del análisis.

En algunas ocasiones, las disparidades pueden ser reales, y son debidas a individuos con condiciones médicas no comunes. De ser así estos individuos deberían de ser excluidos ya que son relativamente atípicos para el resto de la muestra.

Los errores de género, normalmente, ocurren con una frecuencia alta en las bases de datos de GWAS, digamos $\sim 1\,\%$ de la frecuencia, una magnitud mayor a lo esperado si todo fuera atribuído a condiciones médicas. Por lo general, la mayoría de estos errores se deben a fallas en el etiquetado. Sin embargo, es imposible determinar si los errores de etiquetados sólo coresponden a la etiqueta del género (caso en el que no hay daño si el individuo se queda en la muestra después de ser corregida la etiqueta) o si es un error de etiquetado más grave, por ejemplo que se esté vinculando una muestra de ADN errónea con un registro clínico equivocado. Es así que surge la pregunta de qué hacer en estos casos. Desde un punto de vista muy conservador, la respuesta es asumir lo peor y excluir todos los individuos que presentan errores. Sin embargo, existen casos en los que se pueden ignorar, por ejemplo: si es un estudio de casos y controles, y estos han sido recolectados de forma separada y han sido genotipificados con protocolos que previenen los intercambios de ADN en el laboratorio, entonces es posible argumentar que el vínculo entre el ADN y los estados de caso-control es seguro y, en consecuencia, el individuo se puede conservar.

Otras veces se puede observar que se identifica un género intermedio basándose en los datos del cromosoma X; contiene muchos SNPs heterocigotos para ser masculino, y muchos SNPs homocigotos para ser femenino. En muchos de estos casos, hay evidencia de contaminación de la muestra. Dado el impacto biológico de este fenómeno es mejor excluir estos individuos con valores de F intermedios.

2.3.4. Frecuencias alélicas

La razón por la que se argumenta el filtramiento por frecuencias de alelo menor, o MAF por sus siglas en inglés (minor allele frequency), es que la potencia para detectar una señal de asociación decrece con un MAF decreciente. Solamente va a aumentar el número de pruebas realizadas y, por tanto, va decrecer la potencia para detectar señales de asociación en otros SNPs ya que se

incrementa la penalización por múltiples pruebas.

2.4. Medidas de asociación

Dado un marcador genético que consiste en un locus bialélico con alelos a y A, los genotipos posibles son $a|a$, $a|A$ y $A|A$. La penetrancia asociada a la enfermedad con un genotipo dado es el riesgo de que los individuos que tengan ese genotipo tengan la enfermedad. Entre los modelos estándar para la penetrancia de enfermedad que implican una relación específica entre fenotipo y genotipo se encuentran los modelos multiplicativos, aditivos, recesivos y dominantes.

Se denomina γ ($\gamma > 1$) al parámetro de penetrancia genética: un modelo multiplicativo indica que el riesgo de la enfermedad incrementa $\gamma-$veces para el genotipo $a|A$ y γ^2 para el genotipo $A|A$; un modelo aditivo indica que el riesgo de la enfermedad incrementa $\gamma-$veces para el genotipo $a|A$ y $2\gamma-$veces para el genotipo $A|A$; un modelo recesivo indica que dos copias del alelo A son requeridas para que incremente $\gamma-$veces el riesgo de la enfermedad, y el modelo dominante indica que basta con una copia del alelo A para que incremente $\gamma-$veces el riesgo de la enfermedad.

Una medida usual e intuitiva del grado de intensidad de asociación es el riesgo relativo (RR), el cual compara las penetrancias de la enfermedad entre los individuos expuestos a diferentes genotipos. En la tabla 2.1BASstats. se definen las relaciones existentes entre *RRs* para los modelos mencionados anteriormente. Se muestra las funciones de penetrancia para los genotipos $a|a$, $A|a$ y $A|A$, y el riesgo relativo de los genotipos $A|a$ y $A|A$ en comparación con el genotipo $a|a$.

	Penetrancia			Riesgo Relativo	
Modelo de enfermedad	$a\|a$	$A\|a$	$A\|A$	$A\|a$	$A\|A$
Multiplicativo	f_0	$f_0\gamma$	$f_0\gamma^2$	γ	γ^2
Aditivo	f_0	$f_0\gamma$	$2f_0\gamma$	γ	2γ
Recesivo	f_0	f_0	$f_0\gamma$	1	γ
Dominante	f_0	$f_0\gamma$	$f_0\gamma$	γ	γ

Tabla 2.1: Tabla de penetrancia y RR para distintos modelos de enfermedad. Donde $f_0 > 0$ denota la penetrancia del genotipo $a|a$.

Sin embargo, las estimaciones de riesgo relativo basadas en penetrancias sólo pueden ser calculadas directamente de un estudio de cohorte, cuando se les da un seguimiento a los grupos de pacientes expuestos y no expuestos para ver si han desarrollado la enfermedad. En los estudios de caso-control, donde el investigador controla la proporción de casos y controles, no es posible hacer cálculos

directos de penetrancia de la enfermedad, y por ende de RR. En este tipo de estudio, el grado de intensidad de asociación es calculado por medio de la razón de momios (OR, por sus siglas en inglés). El ORenotípico describe la asociación entre la enfermedad y el genotipo al comparar la probabilidade de que un individuo que porte cierto genotipo tenga la enfermedad contra la probabilidad de que un individuo que no porte dicho genotipo presente la enfermedad. Dichas probabilidades se calculan directamente de las frecuencias de exposición entre casos y controles [9]. Formalmente, si se define D^+ y D^- como el estado de tener la enfermedad y de no tenerla, y E^+ y E^- representa estar expuesto o no estar expuesto al genotipo, respectivamente, en este caso, tener uno u otro genotipo. Entonces el OR genotípico es

$$OR = \frac{P(D^+|E^+)/[1 - P(D^+|E^+)]}{P(D^+|E^-)/[1 - P(D^+|E^-)]} \tag{2.1}$$

Análogamente, el OR alélico describe la asociación entre la enfermedad y el alelo al comparar la probabilidad de que un individuo portador del alelo A tenga la enfermedad contra la probabilidad de que un individuo portador del alelo a la tenga. Cabe señalar que es posible estimar valores de OR utilizando técnicas de análisis multivariado como regresión logística y otros modelos log-lineales, los cuales permiten incorporar otros factores de confusión como variables clínicas, etc...

2.5. Tipos de pruebas estadísticas para el análisis de asociación genética

En esta sección se da un repaso de las pruebas estadísticas que permiten estudiar la asociación entre variables. El tipo de prueba estadística que se escoge para hacer la prueba de asociación se hace en base a la hipótesis y al tipo de población de estudio. También la prueba de asociación puede ser ajustadas o no ajustadas por covariables. Las pruebas de asociación genética son usualmente realizadas de manera independiente para cada SNP.

2.5.1. Estadísticos de prueba para rasgos dicotómicos

Cuando se estudian rasgos dicotómicos, usualmente se hace la prueba de asociación utilizando tablas de contingencia de 2×3, donde en las filas se tienen los genotipos y en las columnas los estados (ver la siguiente sección). Los estadísticos de prueba más comunes son la prueba χ^2 de Pearson con 2 grados de libertad (2 gl) o la prueba exacta de Fisher. Estas dos funcionan de manera muy similar y tienen una potencia razonable sin importar el riesgo subyacente.

Las tablas de contingencia tienen como ventaja que se pueden ajustar dependiendo del modelo de penetrancia de la enfermedad, lo que se hace al agrupar la cuentas genotípicas de distintas maneras [9]. Por ejemplo, en el modelo dominante con solo portar un alelo de riesgo (alelo A) existe predisposición a la

enfermedad. Por lo que la tabla de contingencia se puede resumir en una tabla de 2×2, donde en una fila van las frecuencias del genotipo $a|a$ y en la otra las frecuencias de los genotipos $A|a$ y $A|A$. Bajo la misma lógica se construye la tabla para un modelo recesivo.

Para los rasgos complejos, en general se cree que la contribución de los SNPs al riesgo de enfermedad tiende a ser aditiva [3]. En este caso las pruebas generales de χ^2 de Pearson (2 gl) y Fisher tienen una potencia razonable pero no son tan poderosas como las pruebas que se adaptan a este escenario. Una forma de combatir dicho problema y aumentar la potencia para detectar riesgos de tipo aditivo, es contar alelos en vez de genotipos. De esta manera cada individuo contribuye dos veces a las tabla de contingencia de 2×2 y se aplica una prueba χ^2 de Pearson con un grado de libertad. Sin embargo, es necesario mencionar que este método funciona bajo el supuesto de que los casos y controles combinados están en equilibrio de Hardy-Weinberg.

Adicionalmente, la prueba de Cochran-Armitage es similiar a la prueba basada en la cuenta de alelos, sin tener como condición que se sostenga el equilibrio de Hardy-Weinberg entre los individuos. La idea es poner a prueba la hipótesis de pendiente cero para una línea que se ajusta a las tres mejores estimaciones de riesgo genotípico.

No existe regla general que responda cuál es la prueba estadística más conveniente a usar. Se podría decidir con mayor facilidad si sabemos qué proporción de las variantes aún no descubiertas que predisponen a la enfermedad funcionan de forma aditiva y cuáles funcionan de manera dominante o recesiva. En caso de ignorar lo anterior, la decisión recae en el buen juicio del investigador. A continuación se dará una descripción más detallada de los estadísticos de prueba anteriormente mencionados. Pero primero se da una descripción de las tablas de contingencia que se utilizan.

Tablas de contingencia para cuentas genotípias y alélicas

Cuando se quiere investigar la dependencia entre dos variables categóricas, es conveniente clasificar los datos por medio de una tabla de contingencia. En el caso de hacer una evaluación de asociación de un genotipo en un SNP cuando se tienen estados de casos y controles, lo más natural es hacer una tabla de contingencia de frecuencias de genotipos/alelos entre casos y controles, ya que el factor de riesgo que se está evaluando es el genotipo o el alelo en un marcador específico. Dada la tabla de contingencia, el objetivo es probar la hipótesis nula de que el tipo de genotipo/alelo es independiente del estado de salud de la persona (si porta enfermedad o no). En términos generales, como hipótesis nula deseamos probar independiencia entre las filas y las columnas.

Los datos para cada SNP cuyo alelo menor es a y alelo mayor es A en grupos de n individuos de caso-control pueden ser descritos en una tabla de contingencia de estatus de enfermedad de $2 \times k$, donde $k = 2$ para alelos y $k = 3$ para genotipos.

Cuando se cuenta por alelos, la tabla de contingencia corresponde a la tabla 2.2.

Alelo	a	A	Total
Caso	m_{11}	m_{12}	$m_{1\bullet}$
Control	m_{21}	m_{22}	$m_{2\bullet}$
Total	$m_{\bullet 1}$	$m_{\bullet 2}$	$2n$

Tabla 2.2: Tabla de contingencia para cuentas alélicas en un estudio de asociación de caso-control.

La celda m_{ij} representa cuántos individuos que pertenecen al grupo i, con $i =$ caso o control, tienen el alelo de tipo j, $j = a$ o A. La celda $m_{i\bullet}$ representa al total de individuos que pertenecen grupo i, $m_{\bullet j}$ representa el total de individuos que tienen el alelo tipo j.

Con los datos obtenidos en la tabla de contingencia es posible calcular el valor de OR. El OROR alélico es estimado por

$$OR_A = \frac{m_{12}m_{21}}{m_{11}m_{22}} \tag{2.2}$$

Si la prevalencia de la enfermedad en un individuo control portador del alelo a puede ser estimada y denotada como P_0, entonces el riesgo relativo de la enfermedad en los individuos que tienen un alelo A comparado con un alelo a es estimado por

$$RR_A = \frac{OR_A}{1 - P_0 + P_0 OR_A} \tag{2.3}$$

Cuando la cuenta es genotípica la tabla de contingencia se puede ver en la tabla 2.3.

Genotipo	$a\|a$	$A\|a$	$A\|A$	Total
Caso	n_{11}	n_{12}	n_{13}	$n_{1\bullet}$
Control	n_{21}	n_{22}	n_{23}	$n_{2\bullet}$
Total	$n_{\bullet 1}$	$n_{\bullet 2}$	$n_{\bullet 3}$	n

Tabla 2.3: Tabla de contingencia para cuentas genotípicas en un estudio de asociación de caso-control.

El OR genotípico relativo para el genotipo $A|A$ sobre el genotipo $a|a$ estimado por

$$OR_{AA} = \frac{n_{13}n_{21}}{n_{11}n_{23}} \tag{2.4}$$

Y el OR para el genotipo $A|a$ sobre el genotipo $a|a$ estimado por

$$OR_{Aa} = \frac{n_{12}n_{21}}{n_{11}n_{22}}.$$

Si la prevalencia de la enfermedad en un individuo control portador del genotipo $a|a$ puede ser estimada y denotada como p_0, entonces el riesgo relativo de la enfermedad en los individuos que tienen el genotipo $A|A$ comparado con el genotipo $a|a$ es estimado por

$$RR_{AA} = \frac{OR_{AA}}{1 - p_0 OR_{AA}}. \tag{2.5}$$

Y el análogo para el genotipo $A|a$.

Prueba χ^2 de Pearson

Retomando las tablas de contingencia 2.2 y 2.3, se quiere probar que las dos variables son independientes entre sí. De ser el caso, cada probabilidad por celda es igual al producto de sus respectivas probabilidades de renglón o columna.

El estimador de máxima probabilidad (MLE) para cualquier probabilidad de renglón o columna se encuentra como sigue. Con n_{ij} se denota la frecuencia observada en el renglón i y la columna j de la tabla de contingencia dada y con p_{ij} se denota la probabilidad de que una observación esté en esta celda. Si las observaciones son independientes, entonces las frecuencias por celda tienen distribución multinomial y el MLE de p_{ij} es simplemente la frecuencia relativa observada para esa celda. Esto es, $\widehat{p}_{ij} = \frac{n_{ij}}{n}$ con $i = 1, 2, \ldots, r$ y $j = 1, 2, \ldots, c$. Del mismo modo, viendo el renglón i como una misma celda la probabilidad para el renglón i está dada por p_i, y como $n_{i\bullet}$ denota el número de observaciones en el renglón i, $\widehat{p}_i = \frac{n_{i\bullet}}{n}$ es el MLE de p_i. Análogamente para el caso de columna.

El valor esperado de la frecuencia por celda, n_{ij} para una tabla de contingencia es igual a $\mathrm{E}(\widehat{n}_{ij}) = \mathrm{E}(\frac{n_{i\bullet} n_{\bullet j}}{n})$. Por último la determinación de los grados de libertad de una tabla de contingencia que tenga r renglones y c columnas será igual a $(r-1)(c-1)$.

Bajo la hipótesis nula de no asociación con la enfermedad, se espera que las frecuencias relativas de alelos o genotipos sean las mismas en los grupos de casos y controles. Entonces, la prueba de asociación es efectuada con una simple prueba χ^2.

En la prueba convencional de χ^2 para una tabla de contingencia 2×3 de frecuencias de genotipos de los casos y controles, se asume que cada uno de los genotipos tiene una asociación independiente con la enfermedad y por lo tanto la prueba de asociación genotipos tiene dos grados de libertad.

Por medio de la prueba χ^2 de Pearson se puede hacer una prueba de asociación entre el genotipo y el fenotipo [20, 21]. Cuando la tabla de contingencia es de 2×2, una prueba de no asociación entre las filas y las columnas es equivalente a la prueba de la hipótesis nula, $H_0 : OR = 1$. La prueba de asociación genotípica basada en una simple χ^2 es dada por

$$X^2 = \sum_{i=1}^{2} \sum_{j=1}^{3} \frac{(n_{ij} - E[n_{ij}])^2}{E[n_{ij}]} \tag{2.6}$$

donde $E[n_{ij}] = \frac{n_{i\bullet}}{n_{j\bullet}}$. El estadístico X^2 tiene distribución χ^2 con $(r-1)(c-1)$ grados de libertad bajo la hipótesis nula de no asociación, para $r = 2$ y $c = 3$. Este estadísitico de prueba es una función de los cuadrados de las desviaciones de las cantidades observadas respecto de sus valores esperados, ponderados por los recíprocos de sus valores esperados.

Prueba exacta de Fisher

La prueba exacta de Fisher es una alternativa a la prueba χ^2 de Pearson y se utiliza cuando se tienen tamaños de muestra pequeños [20, 21]. También utiliza tablas de contingencia; sin embargo, se dice que es exacta ya que la significancia estadística de la desviación a la hipótesis nula se calcula exactamente, a diferencia de otras pruebas donde dicho cálculo se basa en aproximaciones asintóticas para tamaños de muestra grandes.

Dados los genotipos observados para un marcador específico, se tiene nuevamente una tabla de contingencia de 2×3. En esta prueba se toman la frecuencias marginales de los valores observados como fijas. Basándonos en esta restricción construimos la siguiente tabla de contingencia para un caso particular de la tabla 2.3.

Genotipo	$a\|a$	$A\|a$	$A\|A$	Total
Caso	$\widehat{n}_{11}$	$\widehat{n}_{12}$	$\widehat{n}_{13}$	$\overline{n}_{1\bullet}$
Control	$\widehat{n}_{21}$	$\widehat{n}_{22}$	$\widehat{n}_{23}$	$\overline{n}_{2\bullet}$
Total	$\overline{n}_{\bullet 1}$	$\overline{n}_{\bullet 2}$	$\overline{n}_{\bullet 3}$	n

Tabla 2.4: Tabla de frecuencia genotípica para un SNP específico.

Esta prueba obtiene la probabilidad exacta bajo la hipótesis nula de tener una tabla de contingencia al menos tan extrema como la observada, asumiendo una probabilidad igual para cada permutación de la variable dependiente. Las probabilidades exactas de las frecuencias $\widehat{n}_{11}$, $\widehat{n}_{12}$, $\widehat{n}_{13}$, $\widehat{n}_{21}$, $\widehat{n}_{22}$, $\widehat{n}_{23}$ pueden ser derivadas de la distribución de probabilidad hipergeométrica:

$$\begin{aligned} \mathrm{P}(\widehat{n}_{11}, \widehat{n}_{12}, \widehat{n}_{13}, \widehat{n}_{21}, \widehat{n}_{22}, \widehat{n}_{23}) &= \frac{\binom{\overline{n}_{\bullet 1}}{\widehat{n}_{11}}\binom{\overline{n}_{\bullet 2}}{\widehat{n}_{12}}\binom{\overline{n}_{\bullet 3}}{\widehat{n}_{13}}}{\binom{n}{\overline{n}_{1\bullet}}} \\ &= \frac{\binom{\overline{n}_{\bullet 1}}{\widehat{n}_{21}}\binom{\overline{n}_{\bullet 2}}{\widehat{n}_{22}}\binom{\overline{n}_{\bullet 3}}{\widehat{n}_{23}}}{\binom{n}{\overline{n}_{2\bullet}}} \\ &= \frac{(\overline{n}_{\bullet 1}!\overline{n}_{\bullet 2}!\overline{n}_{\bullet 3}!)(\overline{n}_{1\bullet}!\overline{n}_{2\bullet}!)}{n!\widehat{n}_{11}!\widehat{n}_{12}!\widehat{n}_{13}!\widehat{n}_{21}!\widehat{n}_{22}!\widehat{n}_{23}!} \end{aligned} \tag{2.7}$$

Para cada muestra de n individuos existen distintas posibles combinaciones de frecuencias marginales cuyas sumas de n. A la probabilidad de observar la frecuencias dadas en la tabla 2.4 la llamamos P_{obs} y se obtiene con la formula (2.7). Entonces lo que hace la prueba exacta de Fisher es evaluar la probabilidad

de observar P_{obs}, así como la probabilidad de observar todas las tablas cuya combinación de valores sume las frecuencias marginales $\{\overline{n}_{j\bullet}\}$ y $\{\overline{n}_{\bullet i}\}$. Cabe señalar que se restringe a la probabilidades que son menores o iguales a P_{obs}. Si la suma de estas probabildades es menor o igual a un nivel de significancia establecido entonces la hipótesis es rechazada. Es decir, el valor de p exacto es dado por la probabilidad tener un evento tan extremo, o más extremo, en dirección de la hipótesis alternativa, de lo que ya fue observado.

Prueba de Cochran-Armitage

La prueba de Cochran-Armitage se utiliza en el análisis de datos categóricos cuando el objetivo es evaluar la presencia de una asociación entre una variable con dos categorías y una variable con k categorías [21, 22]. Se distingue de la prueba de χ^2 de Pearson por incorporar un presunto orden de los efectos de las k categorías de la segunda variable.

Considérese la tabla 2.3, donde se tiene la tabla de contingencia de los genotipos contra estado caso y control, con A como alelo de riesgo. En la primera columna se tiene que no existe la presencia de alelos de riesgo, en la segunda hay un alelo de riesgo y en la última estan presentes dos.

Supongamos que (n_{11}, n_{12}, n_{13}) tiene una distribución trinomial, dichos eventos son mutuamente excluyentes y las probabilidades para los genotipos aa, aA y AA son p_0, p_1 y p_2, respectivamente. De igual forma (n_{21}, n_{22}, n_{23}) tiene una distribución trinomial con probabilidades q_0, q_1 y q_2.

Con base a la notación dada, como hipótesis nula se tiene que todas las entradas en la tabla son proporcionales, es decir, $H_0 : p_j = q_j$ para $j = 0, 1, 2$; no hay diferencias de probabilidades entre casos y controles. Y se quiere probar contra la hipótesis alternativa que dice que, dentro de cada columna, el valor absoluto de la diferencia entre la probabilidad de una observación clasificada como "caso" o "control" crece monótonamente a través de la tabla.

Para probar H_0 utilizando la prueba de Cochran-Armitage, un conjunto de valores $x = (x_0, x_1, x_2)$ deben de ser asignados a genotipos (aa, aA, AA). Considerando A como el alelo de riesgo, para el modelo aditivo se asigna $x = (0, 1, 2)$, para el modelo dominante se asigna $x = (0, 1, 1)$, y para el modelo recesivo se tiene $x = (0, 0, 1)$.

Dado el valor x, el estadístico para la prueba de Cochran Armitage se puede escribir como

$$Z_T^* = \frac{U}{[\mathrm{var}_{H_0}(U)]^{1/2}} \tag{2.8}$$

con

$$U = \frac{1}{n}\sum_{i=0}^{2} x_i(n_{2\bullet}n_{1i} - n_{1\bullet}n_{2i}) \tag{2.9}$$

y

$$\mathrm{var}_{H_0}(U) = \frac{n_{1\bullet}n_{2\bullet}}{n}[\sum_{i=0}^{2} x_i^2 q_i - (\sum_{i=0}^{2} x_i q_i)^2] \tag{2.10}$$

que se calcula bajo el supuesto de H_0 donde $p_j = q_j$, y $E(U) = 0$.

Desarrollando la ecuación (2.8), la prueba de Cochran-Armitage para evaluar la asociación entre la enfermedad y el marcador es dada por

$$Z_T^* = \frac{[\sum_{i=1}^{3} x_i(n_{1i}n_{2\bullet} - n_{2i}n_{1\bullet})]^2}{\frac{n_{1\bullet}n_{2\bullet}}{n}[\sum_{i=1}^{3} x_i^2 n_{\bullet i}(n - n_{\bullet i}) - 2\sum_{i=1}^{2}\sum_{j=i+1}^{3} x_i x_j n_{\bullet i} n_{\bullet j}]} \tag{2.11}$$

donde $x = (x_1, x_2, x_3)$ son los pesos escogidos para detectar un tipo de asociación en particular.

Cualquier modelo en donde se especifique la tendencia de riesgo con respecto al incremento de numero de alelos A, como los modelos aditivos, recesivos y dominantes, pueden ser examinados utilizando la prueba de Cochran-Armitage. En las pruebas de asociación genética en las que el modelo genético subyacente es desconocido, usualmente se usa la versión aditiva del modelo.

2.5.2. Uso de modelos lineales para análisis de asociación a enfermedades

Las pruebas estadísticas anteriores no toman a consideración covariables adicionales. En caso de que se sospeche que existen factores externos que puedan estar afectando el efecto del genotipo en el rasgo, es necesario incluir dichas covariables en el análisis. Los modelos lineales tienen la ventaja de que es posible agregar diversos factores que modifican el efecto. Adicionalmente, se pueden ajustar a rasgos cuantitativos y cualitativos.

Mediante el uso de modelos de regresión lineal simple o múltiple, se puede formalizar el uso de la relación entre dos variables para predecir el valor de una usando el valor de la otra, de tal manera que obtengamos las mejores predicciones posibles. Además, esta metodología resulta útil para explicar la variación de una variable como consecuencia de su relación con otra u otras variables, sobre todo si se cree que el efecto del genotipo en el rasgo varía en función del valor de otra variable.

Modelos de regresión lineal

Un modelo estadístico lineal [21, 43] que relaciona una respuesta Y con un conjunto de variables independientes fijas x_1, x_2, ..., x_k es de la forma

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \epsilon \tag{2.12}$$

donde β_0, β_1, ..., β_k son parámetros desconocidos, ϵ es una variable aleatoria y las variables x_1, x_2, ..., x_k toman valores conocidos. Supóngase que $\text{var}(\epsilon) = \sigma_\epsilon^2$ y $\text{E}(\epsilon) = 0$, por lo tanto se tiene que

$$\text{E}(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k \tag{2.13}$$

Decimos que el valor esperado de Y es $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k$ (una función de las variables independeintes x_1, x_2, ..., x_k), más un error aleatorio

ϵ. Desde un punto de vista práctico, ϵ reconoce la incapacidad para dar un modelo exacto por naturaleza. Y varía alrededor de E(Y) de un modo aleatorio, con media igual a 0 y una varianza constante entre los diferentes valores de las variables independientes, porque no se ha incluido en el modelo toda la gran cantidad de variables que pueden afectar a Y.

Un procedimiento para estimar los parámetros de cualquier modelo lineal, el método de mínimos cuadrados, se puede ilustrar con sólo ajustar una recta a un conjunto de puntos. Supóngase que se quiere ajustar el modelo de regresión lineal simple, cuya forma es

$$\mathrm{E(Y)} = \beta_0 + \beta_1 x \tag{2.14}$$

a un conjunto de puntos en el plano, es decir se quiere ajustar una recta que pasa por un conjunto de puntos. Se postula que $Y = \beta_0 + \beta_1 x + \epsilon$, donde ϵ tiene alguna distribución de probabilidad con $\mathrm{E}(\epsilon) = 0$. Si $\widehat{\beta}_0$ y $\widehat{\beta}_1$ son estimadores de los parámetros β_0 y β_1, entonces $\widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 x$ es claramente un estimador de E(Y).

El procedimiento de mínimos cuadrados para ajustar una recta que pase por un conjunto de n puntos es semejante al método que puede usar al ajustar una recta a simple vista; esto es, se requiere que las diferencias entre los valores observados y lo puntos correspondientes en la recta ajustada sean "pequeñas" en un sentido general. Una forma cómoda de lograr esto y que proporciona estimadores con buenas propiedades, es minimizar la suma de cuadrados de las desviaciones verticales a partir de la recta ajustada. Entonces, si

$$\widehat{y_i} = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$$

es el valor pronosticado del i-ésimo valor de y (cuando $x = x_i$), entonces la desviación, en ocasiones llamada *error*, del valor observado de y_i a partir de $\widehat{y_i} = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$ es la diferencia $y_i - \widehat{y_i}$ y la suma de los cuadrados de las desviaciones a minimizar es

$$\mathrm{SSE} = \sum_{i=1}^{n} (y_i - \widehat{y_i})^2 = \sum_{i=1}^{n} [y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 x_i)]^2.$$

La cantidad SSE también recibe el nombre de *suma de cuadrados del error*. Si la ecuación SSE tiene un mínimo, ocurrirá para valores de β_0 y β_1 que satisfagan las ecuaciones, $d\mathrm{SSE}/d\widehat{\beta}_0 = 0$ y $d\mathrm{SSE}/d\widehat{\beta}_1 = 0$. Estas ecuaciones se denominan ecuaciones de mínimos cuadrados para estimar los parámetros de una recta. Y forman un sistema de ecuaciones lineales, por lo que se pueden resolver simultáneamente. Las soluciones son

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2} \tag{2.15}$$

$$\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x} \tag{2.16}$$

La solución simultánea para las dos ecuaciones minimizan SSE. La expresión $\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$ se calcula al sumar los productos de los valores de x menos

su media y los valores de y menos su media y la expresión $\sum_{i=1}^{n}(x_i - \overline{x})^2$ se calcula al sumar los cuadrados de los valores de x menos su media.

Se puede escribir la observación para el $i-$ésimo individuo y_i como

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + \epsilon_i \tag{2.17}$$

donde x_{ij} es el ajuste de la $j-$ésima variable independiente para la $i-$ésima observación, $i = 1, 2, \ldots, n$. Ahora se definen las siguiente matrices con $x_0 = 1$:

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, X = \begin{bmatrix} x_0 & x_{11} & x_{12} & \ldots & x_{1k} \\ x_0 & x_{11} & x_{12} & \ldots & x_{1k} \\ \vdots & \vdots & \vdots & \vdots & \\ x_0 & x_{n1} & x_{n2} & \ldots & x_{nk} \end{bmatrix}, \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \epsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

Y representa a n observaciones independientes $y_1, y_2, ..., y_n$ y $\varepsilon_1, \ldots, \varepsilon_n$ son variables aleatorias independientes e idénticamente distribuidas. Por ejemplo, en el caso de una regresión lineal para un rasgo cuantitativo, el vector Y respresenta los valores de los fenotipos para los n individuos.

Entonces, las n ecuaciones que representan y_i como función de las x, las β y las ε se pueden escribir simúltaneamente como

$$Y = X\beta + \epsilon. \tag{2.18}$$

Para n observaciones desde un modelo lineal simple de la forma

$$Y = \beta_0 + \beta_1 x + \varepsilon, \tag{2.19}$$

se tiene

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}, \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}, \epsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

De acuerdo con las ecuacion de mínimos cuadrados (2.15) y (2.16) para β_1 y β_0 se tiene que

$$\begin{gathered} n\widehat{\beta}_0 + \widehat{\beta}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \\ \widehat{\beta}_0 \sum_{i=1}^{n} x_i + \widehat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i. \end{gathered}$$

Como

$$X^T X = \begin{bmatrix} 1 & 1 & \ldots & 1 \\ x_1 & x_2 & \ldots & x_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} = \begin{bmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{bmatrix}$$

y

$$X^T Y = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \end{bmatrix}, \text{ si } \widehat{\beta} = \begin{bmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \end{bmatrix},$$

se ve que las ecuaciones de mínimos cuadrados están dadas por

$$(X^T X)\widehat{\beta} = X^T Y.$$

Por tanto, las soluciones de mínimos cuadrados para un modelo lineal general estan dadas por

$$\widehat{\beta} = (X^T X)^{-1} X^T Y.$$

Aplicaciones de regresión lineal para la determinación de asociación genética a enfermedades

Si se considera el siguiente modelo de regresión lineal simple:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \tag{2.20}$$

donde $i = 1, ..., n$ indica el individuo. En este modelo, el parámetro β_1 es definido como la cantidad de cambio que ocurre en y por cada unidad de cambio que ocurre en x y es el que por medio de una prueba de hipótesis indicará si hay evidencia de asociación genética entre el fenotipo y y el marcador x. Por ejemplo, si x es un indicador de presencia de una variante alélica en un SNP locus dado y y es el nivel de triglicéridos, entonces β_1 es la diferencia de la media del nivel de triglicéridos entre los individuos con y sin esa variante alélica. Las estimaciones de mínimos cuadrados de β_0 y de β_1 son dados por

$$\widehat{\beta}_0 = (\sum_i y_i - \widehat{\beta}_1 \sum_i x_i)/n \tag{2.21}$$

y

$$\widehat{\beta}_1 = \frac{n \sum_i x_i y_i - \sum_i x_i \sum_i y_i}{n \sum_i x_i^2 - (\sum_i x_i)^2}. \tag{2.22}$$

Aquí, el coeficiente de cambio β_1 captura la información en la medida que la relación entre x y y es una linea recta.

El modelo lineal múltiple es una generalización del modelo dado por (2.20) en el cual otras variables independientes adicionales pueden ser incluidas a la ecuación. Por ejemplo, supongamos que tenemos m covariables, dadas por $z_{i1}, ..., z_{im}$ para el individuo i; z_{i1} puede ser el genero y z_{i2} puede ser IMC (índice de masa corporal), etc... En estos casos se utiliza el siguiente modelo

$$y_i = \beta_0 + \beta_1 x_i + \sum_{j=1}^{m} \alpha_j z_{ij} + \epsilon_i \tag{2.23}$$

Ahora la estimación de los parámetros β_0 y β_1 toma en cuenta las variables adicionales del modelo. Estas variables adicionales pueden ayudar a explicar la

variabilidad del rasgo, y es importante incluirlas para poder realizar conclusiones válidas acerca del efecto del genotipo en el rasgo.

Para probar la significancia de un coeficiente de regresión en particular, es decir, cuando se tiene como hipótesis nula $H_0 : \beta_j = 0$ y como hipótesis alternativa $H_1 : \beta_j \neq 0$, El estadístico utilizado es

$$t_0 = \frac{\widehat{\beta}_j}{\sqrt{\widehat{\sigma}^2 c_{jj}}} \tag{2.24}$$

donde c_{jj} representa a la varianza del j-ésimo coeficiente de regresión estimado $\widehat{\beta}_j$ y correspode al elemento de la diagonal de la matriz de varianza-covarianza

$$C = \widehat{\sigma}^2 (X^T X)^{-1}.$$

El valor de $\widehat{\sigma}^2$ se obtiene por medio del error cuadrático medio (MSE), que es igual a $\text{MSE} = \frac{\text{SSE}}{n-(k+1)}$, donde el divisor corresponde a los grados de libertad asociados a SSE, n es igual al número total de observaciones y k es igual a número total de predictores.

El denominador de (2.24) se llama error estándar del coeficiente de regresión $\widehat{\beta}_j$, definido como $\widehat{se}$. Donde $\widehat{se}(\beta_j) = \sqrt{\widehat{\sigma}^2 c_{jj}}$, por lo que una forma equivalente de escribir el estadístico de prueba

$$t_0 = \frac{\widehat{\beta}_j}{\widehat{se}(\widehat{\beta}_j)}.$$

Si $H_0 : \beta_j = 0$ no se rechaza, quiere decir que x_j puede ser eliminado del modelo. Nótese que ésta es una prueba parcial ya que el coeficiente de regresión β_j depende en todas las variables regresoras x_i $(i \neq j)$ que están en el modelo.

Para estos modelos lo que nos interesa es estimar los parámetros del modelo y evaluar las hipótesis relativas en la población. Por ejemplo, para el modelo aditivo dado por la ecuación (2.23), nos interesa evaluar la hipótesis nula de no asociación entre el genotipo y el rasgo, dado por $H_0 : \beta_1 = 0$.

Los procedimientos de inferencia para la prueba de hipótesis basados en regresión lineal generalmente suponen que el rasgo está distribuido normalmente [21]. En este caso es común utilizar la transformación logarítmica para el rasgo y así normalizar los datos. En este caso la regresión lineal sería

$$\ln(y_i) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \epsilon$$

y transformando los valores se tiene que

$$y_i = \exp(\beta_0)\exp(\beta_1 x_1)\exp(\beta_2 x_2) \ldots \exp(\epsilon_i)$$

Aquí se ve que el efecto de x_i es multiplicativos dado que el efecto de una unidad de cambio en, por ejemplo, x_1 es un incremento $\exp(\beta_1)$ en y, lo mismo para el resto de las x_i. Sin embargo el efecto de x_1 en y no depende del nivel de los otros x_i's, no importa si existe i tal que $x_i = 0$ ó $x_i = 1$, la unidad de cambio en x_1 resulta ser un incremento de $\exp(\beta_1)$ en y. Lo mismo para el resto de las x_i.

Regresión logística

Cuando se trata de estudiar un fenotipo binario la regresión lineal no se puede aplicar directamente, ya que los estados caso-control no se distribuyen normalmente. Estos problemas son resueltos con la regresión logística [20, 43].

Supóngase que se tiene un modelo de la forma

$$y_i = x_i^{'}\beta + \epsilon_i \tag{2.25}$$

donde $x_i^{'} = [1, x_i, x_{i2}, \ldots, x_{ik}]$, $\beta^{'} = [\beta_0, \beta_1, \ldots, \beta_k]$ y la variable de respuesta y_i toma el valor 0 ó 1. Asumimos que la variable y_i es una variable aleatoria de Bernoulli con distribución como sigue:

y_i	Probabilidad
1	$P(y_i = 1) = \pi_i$
0	$P(y_i = 0) = 1 - \pi_i$

Ahora como $E(\epsilon_i) = 0$, el valor esperado para la variable de respuesta es

$$E(y_i) = 1(\pi_i) + 0(1 - \pi_(i)) = \pi_i$$

Esto implica que $E(y_i) = x_i^{'}\beta = \pi_i$, lo que significa la respuesta esperada por esta función es simplemente la probabilidad de que la variable de respuesta sea igual a 1. Existen varios problemas con el modelo de regresión (2.25). El primero es que dado que las variables de respuesta son binarias entonces los términos de error ϵ_i sólo pueden tomar dos valores

$$\begin{aligned} \epsilon_i &= 1 - x_i^{'}\beta \quad \text{cuando} \quad y_i = 1 \\ \epsilon_i &= -x_i^{'}\beta \quad \text{cuando} \quad y_i = 0, \end{aligned}$$

por lo que los errores en este modelo no pueden tener distribución normal. El segundo punto es que la varianza del error no es constante, ya que

$$\sigma^2_{y_i} = E(y_i - E(y_i))^2 = (1 - \pi_i)^2\pi_i + (0 - \pi_i)^2(1 - \pi_i) = \pi_i(1 - \pi_i).$$

Por lo general, cuando la respuesta es binaria, la forma de la función de respuesta no es lineal. La función de respuesta logística puede ser linealizada fácilmente. Esto se resuelve defniniendo $\nu = x^{'}\beta$ como un predictor lineal donde ν es definida por la transformación

$$\nu = \ln\frac{\pi}{1 - \pi} \tag{2.26}$$

Esta transformación se conoce como "transformación logit" de la probabilidad π. La transformación logit es muy popular para modelar datos binomiales, y es una transformación del intervalo $[0, 1]$ a los reales. La razón $\pi/(1 - \pi)$ se llama momios. La estimación de los parámetros se puede hacer por medio de métodos de máxima verosimilitud.

En la interpretación que tiene la regresión logística en un estudio de asociación, la transformación $\text{logit}(\widehat{\pi}) = \log(\frac{\widehat{\pi}}{(1-\widehat{\pi})})$ es aplicada a π_i, donde $\pi_i =$

$P(y_i = 1|x_i)$ y representa el riesgo a enfermedad del $i-$ésimo individuo. La variable de respuesta y_i representa el indicador del estado de enfermedad de i-ésimo individuo. El valor de logit(π_i) es igualado a β_0, β_1 o β_2, de acuerdo al genotipo del individuo i (con β_1 heterocigoto). Y como hipótesis nula se tiene H_0: $\beta_0 = \beta_1 = \beta_2$.

Modelos mixtos

Hasta ahora los modelos estadísticos que se han visto sólo incluyen efectos fijos. Los modelos lineales mixtos [43] son el modelo visto en la ecuación (2.20) con términos de efectos aleatorios adicionales, es decir, tiene efectos fijos y aleatorios. Los efectos aleatorios se atribuyen a un conjunto infinito de niveles de un factor, los cuales son obtenidos de una distribución que modela estos valores. Los efectos fijos se atribuyen a un conjunto finito de niveles de un factor que ocure en los datos. Se dice que es un factor fijo si los valores son fijos, constantes desconocidas. Entonces, en estos datos se considera que tiene dos fuentes de variación: la varianza que mide efectos aleatorios y la varianza de error.

La motivación que se tiene al utilizar modelos mixtos en los análisis de asociación es poder incluir estratificación poblacional[5], estructura familiar y relaciones crípticas[6] dentro del modelo. Básicamente lo que hace es modelar los fenotipos utilizando una mezcla de efectos fijos y efectos aleatorios. Dentro de los efectos fijos se incluye al SNP candidato y las covariables opcionales, como puede ser sexo, edad, etc. Mientras tanto, los efectos aleatorios están basados en una matriz de covarianza fenotípica, que modela la suma de la variación aleatoria heredable y no heredable. Como bien lo demuestran los autores del método EMMAX [34], los métodos que modelan explícitamente la estructura poblacional, estructura familiar y relaciones crípticas tienen un mejor desempeño cuando estos factores están presentes en la población a estudiar que los métodos que no los tienen implementados. Por medio del uso de modelos mixtos los valores de control genómico[7] (sección 2.6.2) son menores a 1.01 [48].

Los modelos lineales mixtos representan al fenotipo Y como una función de los efectos fijos X más efectos aleatorios u:

$$Y = XB + u + \epsilon \quad \text{con} \quad \text{var}(u) = \sigma_g^2 K \quad \text{y} \quad \text{var}(\epsilon) = \sigma_e^2 I,$$

Y es un vector de $n \times 1$ de los fenotipos observados. X es la matriz de $n \times q$ de los efectos fijos, donde se encuentran los datos de genotipos y otras variables confusoras como edad y sexo, B es el vector $q \times 1$ que representa a los coeficientes

[5]Estratificación poblacional es la presencia de una diferencia sistemática en las frecuencias alélicas entre las subpoblaciones de una población, posiblemente debido a las diferente ascendencias de dichas subpoblaciones.

[6]Relaciones crípticas se presentan cuando dos o más individuos en un estudio tienen una relación parental oculta ya sea no reportada o desconocida, por ejemplo primos segundos.

[7]Control genómico (λ) es definido como la mediana de los estadísticos de prueba de los SNPs dividido entre la median de la distribución nula. Se utiliza como una medida para detectar la presencia de estratificación poblacional, valores de $\lambda \approx 1$ indica que no hay estratifiación poblacional, mientras que $\lambda > 1$ indca que existe estratifación poblacional o la presencia de otros factores de confusión.

de los efectos fijos, u representa al componente de la varianza de ruido total $u+\epsilon$ que se distribuye de acuerdo a la matriz de parentesco K, σ_g^2 es el parámetro de varianza genética aditiva. Por lo tanto, u representa el componente heredable de variación aleatoria y ϵ conocido componente de error ambiental, representa el componente no heredable de variación aleatoria.

La matriz de parentesco K es definida de acuerdo a las similaridades genotípicas que hay entre parejas de individuos, por lo que su estructura es influenciada por la estructura poblacional, la estructura familiar y las relaciones crípticas. El parámetro σ_g^2 relaciona la matriz K con el fenotipo Y. Donde σ_g^2 captura el grado en el que individuos genéticamente similares son fenotípicamente similares, eliminando así los factores confusores.

Debido a la alta cobertura que se tiene de datos de genotipo a lo largo de todo el genoma, es posible estimar el grado de la matriz de relación o parentesco entre los sujetos independientes y en ausencia de información genealógica.

Se asume que u y ϵ no están correlacionados, es decir, $\text{cov}(u, \epsilon) = 0$. En otras palabras, se omite la composición poligénica del rasgo, con el fin de que el modelo sea más trabajable [34]. Entonces, al calcular $\text{var}(Y) = \text{var}(XB) + \text{var}(u) + \text{var}(\epsilon)$, se obtiene

$$\text{var}(Y) = \sigma_g^2 K + \sigma_e^2 I \tag{2.27}$$

ya que $\text{var}(XB) = 0$ porque XB es la porción de efectos fijos.

Al hacer la prueba de hipótesis de asociación se tiene que probar la hipótesis $H_0 : \beta_k = 0$ para cada loci, evaluando un locus a la vez. Particularmente, para el locus k en el individuo i se tiene la ecuación

$$y_i = \beta_0 + \beta_k X_{ik} + \nu_{i\overline{k}} \tag{2.28}$$

Donde β_k es el tamaño de efecto del marcador k y X_{ik} representa a las frecuencias de alelo menor del marcador K. El término de error se define como $\nu_{i\overline{k}} = \sum_{s \neq k} \beta_s X_{ik} + \varepsilon$. Donde X_{is} corresponden a las frecuencias de alelo menor de los SNPs restantes, β_s son sus efectos correspondientes y ε es una variable aleatoria de error que representa los efectos ambientales en el fenotipo. Nuevamente, los valores de $\nu_{i\overline{k}}$ se asumen como independientes e idénticamente distribuidos.

Cuando se habla de rasgos cuantitativos muchos SNPs son los que contribuyen al rasgo, y la contribución particular de cada SNP a la varianza total es prácticamente despreciable. Es por eso que los componentes de varianza de $\nu_{i\overline{k}}$ se pueden aproximar a $\nu_i = \sum_{s=1}^{M} \beta_k X_{ik} + \varepsilon_i$ y no tienen que ser estimados por separado para cada SNP. Primero se estiman los componentes de varianza σ_g^2 y σ_e^2 en la ecuación (2.27). Después se dejan fijos y se estiman los parámetros β_k utilizando el método de mínimos cuadrados.

El estimador de mínimos cuadrados común $b = (X'X)^{-1}X'y$ de los parámetros de regresión no es óptimo cuando $V \neq \sigma_e^2 I$. Ahora la solución óptima es el estimador de mínimos cuadrados generalizado, definido como $b_{\text{GLS}} = (X'V^{-1}X)^{-1}X'V^{-1}y$, cuando V es conocida.

Publicaciones recientes han propuesto que los modelos lineales mixtos pueden corregir efectivamente la estructura de una población en los estudios de asociación con rasgos cuantitativos [61]. Los modelos lineales mixtos incorporan directamente al modelo estadístico el grado de relación genética que comparten todos las parejas de individuos. De esta manera, es más probable que dos individuos genéticamente similares estén correlacionados que los fenotipos de los individuos genéticamente diferentes.

El programa EMMAX [34] corrige los problemas producidos por la estructura de la muestra en GWASs utilizando un modelo lineal mixto. En éste se modela la correlación entre los fenotipos y los individuos con una matriz de relación genética que es estimada a partir de la matriz de IBS[8] que mide información compartida entre parejas de individuos. La ventaja que tiene este programa sobre los métodos basados en otras técnicas, como componentes principales (el cual se abordará en la siguiente sección), es que la matriz de relación empírica codifica un amplio rango de las estructuras de la muestra, incluyendo relaciones crípticas y estratificación poblacional.

2.6. Problemas que se enfrentan en GWAS: estratificación poblacional y relaciones crípticas

Los métodos convencionales de GWAS tienen como hipótesis que los individuos de la población a estudiar consisten en individuos no relacionados y que comparten el mismo antecedente poblacional. Sin embargo, en la práctica controlar estos factores en las poblaciones puede resultar casi imposible, sobre todo cuando cuando se estudian poblaciones donde la endogamia es común. Estos fenómenos que vienen complicando el análisis de asociación se conocen como estratificación y relaciones crípticas. Las pruebas estadísticas convencionales de independencia entre marcadores genéticos y rasgos fenotípicos son propensas a asociaciones falsas ya que el marcador y el fenotipo tienen tendencia a estar correlacionados debido a que la estructura poblacional viola la supocisión de independencia bajo la hipótesis nula [35].

Se han realizado diversos estudios que revisan los efectos que tiene el no tomar en cuenta los dos fénomenos mencionados anteriormente [28, 45, 59] .

2.6.1. Relaciones crípticas

Las relaciones crípticas ocurren cuando existen parejas de individuos que están más estrechamente relacionados que el resto de la población en promedio, indicando así que son miembros cercanos de familia. Individuos relacionados a un mayor nivel que los otros inducen correlación estructural que va a alterar los resultados de asociación. Cuando estos factores no se toman en cuenta se

[8]Dos o más alelos son idénticos por estado o IBS (por sus siglas en ingés *identical-by-state*) si tienen la misma composición de ADN pero no necesariamente provienen del mismo ancestro.

tiene una tendencia a obtener una cantidad exagerada de falsos positivos [45], sobre todo cuando se utilizan métodos que asumen que los individuos no estan relacionados.

El efecto de relaciones crípticas pueden darse en caso en que exista un mayor grado de parentesco entre los casos que entre los controles debido a que comparten una enfermedad genética.

Con el fin de eliminar los efectos de las relaciones crípticas, se puede estimar la proporción de alelos idénticos por descendencia (IBD) que comparte cada pareja de individuos del conjunto total de individuos y excluir aquellos individuos que aparecen con relaciones muy cercanas.

2.6.2. Estratificación poblacional

Se entiende por estratificación poblacional a la inclusión de individuos provenientes de distintas poblaciones en la misma muestra de estudio.

La estratificación poblacional se da bajo dos condiciones: primero, los casos y controles provienen de distintas poblaciones étnicas; segundo, los marcadores estudiados están distribuidos de manera diferente en estas poblaciones. Cabe señalar que, aparte de predisponer a un resultado falso positivo, también puede enmascarar a una asociación verdadera, como consecuencia de reducción de potencia para detectar efecto genético.

Para resolver los problemas ocasionados por la estratificación poblacional se han desarrollado diversos métodos para detectar y tratar con ella. Un enfoque común es el control genómico [17], donde se utiliza la distribución de las pruebas estadísticas del análisis de marcadores para estimar el factor de inflación, λ, posteriormente se hace un rescalamiento de las pruebas, y así se restringe el riesgo de falsos positivos.

La información genotípica que se tiene de la muestra ayuda mucho para poder conocer mejor la estructura de la población, y se han desarrollado diversos programas que la toman en cuenta para poder determinar la estructura de la población al incluirla en los análisis de asociación [47,50,62]. El más utilizado es el método de componentes principales. Este método de reducción de dimensiones logra capturar la estructura de la población; posteriormente se pueden incluir dichos componentes como covariables o se utilizan para ajustar a los genotipos y los fenotipos y luego realizar el análisis de asociación con los nuevos valores [47].

Control genómico

El control genómico se utiliza para detectar y compensar la existencia de estratificación poblacional durante la asociación. Se realiza bajo la premisa de que si la muestra estudiada presenta estratificación poblacional entonces la tasa de falsos positivos se incrementa. Lo que trata de hacer es corregir dicha inflación en el estadístico de prueba.

El cálculo de control genómico se hace considerando que los estadísticos de prueba son χ^2 que se distribuye aproximadamente χ^2_1. Bajo la hipótesis nula de no estratificación poblacional, el valor esperado de los estadísticos de prueba

respectivos es 1. La estratificación poblacional es tratada como un efecto aleatorio que hace que la distribución del estadístico de prueba χ^2 tenga una varianza inflada y una media más alta de la que se observaría de otra forma. De aquí que el factor de inflación λ refleja la desviación respecto al valor esperado. En las pruebas estadísticas se asume que se van a ver afectadas uniformemente por el factor de inflación λ, cuya magnitud es estimada al comparar la mediana de los valores de las pruebas estadísticas del conjunto de marcadores con la mediana de los valores de pruebas estadísticas bajo la hipótesis de no estratificación poblacional. La definición de λ para un modelo aditivo simple es dada por

$$\lambda = \frac{\text{mediana}(\chi_1^2, \chi_2^2, ..., \chi_L^2)}{0.456}, \tag{2.29}$$

donde χ_i^2 es el estadístico de prueba para el $i-$ésimo locus no asociado, y el valor 0.4549 es el 50 % cuantil de la distribución χ_1^2. Es así que, si el valor de λ es mayor a 1, entonces existe la presencia de estratificación poblacional dentro de la muestra, y la corrección se hace al dividir el valor del estadístico de asociación por λ. Esta fórmula supone que el factor de inflación es constante a lo largo de todo el genoma [17]. Posteriormente, el estadístico de prueba ajustado por el control genómico, denotado como χ_{GC}^2, se obtiene para cada locus c de la siguiente manera: λ con $\chi_{GC}^2 = \chi_c^2/\lambda$, el cual se distribuye asintóticamente χ_1^2.

En los estudios de genoma completo, la estimación no sesgada de λ puede ser determinada al utilizar todos los marcadores genotipificados; el efecto del factor de inflación sobre SNPs con asociación potencial a la enfermedad es considerado como despreciable. El control genómico sufre de pérdida de potencia cuando el efecto de la estructura poblacional es grande [35].

Componentes principales

Cuando se utiliza el método de componentes principales para estudiar la estratificación poblacional existen dos pasos fundamentales. Primero se realiza un análisis de componentes principales al conjunto de datos de genotipo para deducir ejes de variación genética contiguos. Intuitivamente, estos ejes de variación reducen los datos a un número pequeño de dimensiones, describiendo toda la variabilidad posible. Posteriormente, al realizar el análisis de asociación se incluyen los primeros componentes principales como covariables, normalmente se usan de 2 a 10 componentes [48]. El número utilizado queda a criterio del investigador. Las ventajas de este método son: la utilización de ejes continuos provee la descripción mas útil acerca de la variación genética, el hecho de que los ejes de variación son ortogonales, no influye el número de ejes inferido, y, por último, este procedimiento es computacionalmente viable para estudios de genoma completo.

El software EIGENSTRAT [47] hace uso de análisis de componentes principales (PCA) para detectar y describir la estructura de la muestra y ha sido utilizado en múltiples GWAS. Este programa captura los ejes de variación principales, atrapando así la mayor variabilidad posible. Sin embargo, se presentan

desventajas. Por ejemplo, entre algunos componentes principales pueden surgir grandes diferencias entre los individuos, y no es claro cómo implementar el resto de los componentes principales en la interpretación de la estructura de la muestra.

El análisis de componentes principales asume un tamaño pequeño de población ancestral, sólo hace una captura parcial de los múltiples niveles de la estructura poblacional y la relación genética que hay entre los individuos.

Normalmente, en los análisis de asociación se utiliza una combinación de los métodos anteriores. Primero se identifican los individuos relacionados y se quitan de la muestra, luego corrige la estructura poblacional ya sea utilizando los componentes principales o la información espacial, y, finalmente, se corrige la inflación residual con el control genómico.

En las siguientes secciones se utiliza el método de componentes principales, por eso a continuación se dará una descripción detallada de dicho método.

2.6.3. Análisis de Componentes Principales

El Análisis de Componentes Principales (PCA, por sus siglas en inglés) es un herramienta que permite reducir las dimensiones de un conjunto de datos complejos perdiendo la menor cantidad de información posible, y así poder encontrar una estructura simplificada de estos [19].

La meta principal es encontrar la base (en un sentido algebraico) que mejor represente a estos datos y sea tal que tenga el menor ruido posible y sea capaz de revelar su estructura oculta. Los nuevos componentes principales serán una combinación lineal de las variables originales, y además no estarán correlacionadas entre sí.

Con el PCA se quiere re-expresar los datos como una combinación lineal de sus vectores base. Es decir, tenemos la matriz de datos originales X, y Y es una matriz relacionada a ésta por medio de una transformación lineal. Se define Y como:

$$Y = P^{-1}XP, \tag{2.30}$$

donde Y es una nueva representación de los datos. Los renglones de Y se van a conocer como los componentes principales.

El problema se reduce a encontrar un cambio de base apropiado. Algebraicamente, los componentes principales son un conjunto específico de combinaciones lineales los elementos de X. En la interpretación geométrica estas combinaciones lineales representan un nuevo sistema de coordenadas cuyos ejes representan las direcciones en las que se maximiza la variabilidad.

Factores a considerar

Por otra parte, con el fin de poder garantizar la obtención de la matriz que mejor represente los datos, se tiene que considerar los factores que reduzcan el ruido y la redundancia.

Primeramente consideremos el ruido, no existe una escala absoluta de ruido, sino que el ruido se cuantifica en relación con la intensidad de la señal. Una

medida común es la relación señal-ruido o SNR (por sus siglas en inglés *signal-to-noise ratio*), y se define como

$$\text{SNR} = \frac{\sigma^2_{\text{señal}}}{\sigma^2_{\text{ruido}}}. \tag{2.31}$$

Donde la ecuación (2.31) reprensenta la proporción de la varianza de la señal sobre la varianza de ruido. Un alto SNR ($\gg 1$) indica una alta precisión mientras que un bajo SNR indica que se tienen datos con mucho ruido. Se asume que las direcciones con las varianzas más grandes en el espacio estudiado son las que contienen las dinámicas de mayor interés y por consiguiente son las que tienen mayor SNR. Esta suposición sugiere que busquemos una base que maximice la varianza.

El segundo punto que tenemos que considerar es la redundancia. Es decir, evitar la información repetida. Esto lo podemos medir con el valor absoluto de la covarianza ya que la covarianza mide el grado de relación lineal que tienen dos variables, y lo que nosotros queremos es que no exista relación alguna entre las variables.

Método

Una vez que ya se tienen claros los objetivos podemos comenzar a explicar el método.

Considérese la matriz X, con elementos x_{ij}, la matriz de observaciones "centrada", es decir, a cada observación x_j se le resta su media $\overline{x}$. Cada fila representa el conjunto total de valores que hay de una medición específica y cada columna representa el conjunto de mediciones que se le hizo a una muestra en particular. En este caso, cada fila representa a un SNP particular y cada columna a un individuo. Supóngase que se tiene m SNPs y n individuos, por lo que $\overline{X}$ es una matriz de $m \times n$.

Ahora se define la matriz de covarianza de X, denotada por C_X, como

$$C_X \equiv \frac{1}{n} X X^T. \tag{2.32}$$

Nótese que el ij-ésimo elemento de la matriz C_X es el producto punto entre el vector con valores correspondientes al SNP i con el vector de valores correspondientes al SNP j.

En la diagonal de C_X están contenidos los valores de las varianza de cada SNP ya que estos elementos tienen la forma $\frac{1}{n}\sum_j x^2_{ii}$ lo cual por definición es igual a la $\sigma^2_{x_i}$. Y fuera de la diagonal se encuentra la covarianza entre todos los posibles pares de SNPs, ya que los elementos tienen la forma $\frac{1}{n}\sum_j x_{ij}x_{i'j} = \sigma_{x_i x_{i'}}$.

Se ve que en la matriz de covarianza se encuentra la información buscada, ya que valores grandes en la diagonal indican que hay poco ruido y por consiguiente que se pueden encontrar estructuras interesantes. Fuera de la diagonal las magnitudes grandes corresponden a alta redundancia.

Supóngase que se puede manipular esta matriz de covarianza y encontrar una matriz C_Y. Para comenzar lo ideal sería minimizar la redundancia por lo que se quiere que los terminos fuera de la diagonal, las covarianzas, fueran iguales a 0. Es decir, se quiere que la matriz C_Y sea una matriz diagonal.

PCA utiliza el método más sencillo para diagonalizar ya que supone que todos los elementos $p_1, .., p_m$ son ortonormales, es decir, P es una matriz ortonormal. La idea es sencilla; se asume que los componentes principales con mayor varianza asociada representan a las estructuras más interesantes, mientras que aquellas con menor cantidad de varianza representan ruido. Por lo que el método se dirige a encontrar los componentes que maximicen la varianza.

Se escoge un vector normalizado en el espacio m-dimensional que cumpla que la varianza de X sea maximizada. A este vector lo llamamos p_1. Posteriormente se busca otra dirección a lo largo de la cual se maximiza la varianza, y así sucesivamente. Hay que recordar que se está condicionando a la ortogonalidad, por lo que la búsqueda se restringe a encontrar sólo en los vectores ortogonales al seleccionado previamente. A cada uno se le llama p_i, donde i corresponde al orden en que se tomó. Se repite este procedimiento hasta tener m vectores. Una ganancia que se tiene a través de este algoritmo es que por construcción las varianzas asociadas a cada dirección p_i cuantifican que tan "principal" es cada dirección.

Básicamente la elección de los factores se realiza de tal forma que el primero recoja la mayor proporción posible de la variabilidad original; el segundo factor debe recoger la máxima variabilidad posible no recogida por el primer componente, y así sucesivamente. Del total de factores se elegirán aquellos que recojan el porcentaje de variabilidad que se considere suficiente, los cuales serán los componentes principales.

El procedimiento se resume en lo siguiente:

Encontrar la matriz ortonormal P tal que $Y = P^{-1}XP$ y $C_Y \equiv \frac{1}{n}Y^TY$ es una matriz diagonal, y donde las filas de Y son los componentes principales de X.

Se empieza reescribiendo C_Y en términos de la variable desconocida.

$$\begin{aligned} C_Y &= \frac{1}{n}Y^TY \\ &= \frac{1}{n}(P^{-1}XP)^T(P^{-1}XP) \\ &= \frac{1}{n}(P^TXP)^T(P^TXP) \\ &= \frac{1}{n}(P^TX^TP)(P^TXP) \\ &= P(\frac{1}{n}XX^T)P^T \\ &= PC_XP^T, \end{aligned} \tag{2.33}$$

dado que P es ortonormal, $P^{-1} = P^T$. Nótese que se ha identificado la matriz de covarianza de X.

Se puede suponer sin pérdida de generalidad que dada cualquier matriz simétrica A, esta es diagonalizada por la matriz ortogonal de sus eigenvectores. Para una matriz simétrica A se tiene que $A = EDE^T$, donde D es una matriz diagonal y E es una matriz de eigenvectores de A arreglados como columnas.

Utilizando lo anterior, el siguiente paso es definir la matriz P tal que cada fila p_i es un eigenvector de $\frac{1}{n}XX^T$. Con esto se tiene que $P \equiv E^T$.

$$\begin{aligned} C_Y &= PC_XP^T \\ &= P(E^TDE)P^T \\ &= P(P^TDP)P^T \\ &= (PP^T)D(PP^T) \\ &= (PP^{-1})D(PP^{-1}) \\ C_Y &= D \end{aligned} \tag{2.34}$$

Es evidente que dada la selección de P diagonalizamos C_Y, y de esta manera los resultados de PCA se resumen en las matrices P y C_Y.

Se puede concluir que si C_X de $m \times m$ es la matriz de covarianza con parejas de eigenvalores-eigenvectores $(\lambda_1, e_1), \ldots, (\lambda_m, e_m)$. Entonces el i-ésimo componente principal es dado por

$$Y_i = e_1^T\mathbf{x} = e_{1i}x_1 + \cdots + e_{mi}x_m,$$

con $i = 1, \ldots, m$, donde $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_m \geqslant 0$. Y donde $\text{var}(Y_k) = \lambda_k$, con $k = 1, 2, \ldots, m$ y $\text{cov}(Y_k, Y_i) = 0$, con $k \neq i$.

En la práctica, cuando se quiere calcular PCA el procedimiento se resume en los siguientes pasos.

1. Organizar los datos como una matriz de m x n, donde m es el número total de los distintos tipos de mediciones que se hicieron y n es el número de muestras.
2. A cada tipo de medición substraerle su media.
3. Obtener P calculando los eigenvectores de C_X.
4. Calcular Y.

2.7. Nivel de significancia

2.7.1. Conceptos básicos sobre la prueba de hipótesis.

La prueba de hipótesis es un método esencial para la toma de decisiones.

Cualquier prueba de hipótesis estadística está compuesta por los siguientes elementos esenciales:

- Hipótesis nula, H_0
- Hipótesis alternativa, H_a
- Estadístico de prueba
- Región de rechazo

En el caso del análisis de asociación, como hipótesis nula se tiene que el marcador estudiado no está asociado al rasgo o enfermedad estudiada. La hipótesis alternativa indica que sí hay asociación.

El *estadístico de prueba* es una función de las mediciones muestrales en las que la decisión estadística estará basada.

Con base en la distribución del estadístico de prueba que se está utilizando, se especifica una región, llamada región crítica o región de rechazo. El tamaño de esta región se simboliza con α. Y ubica tradicionalmente en 0.05 o 0.01. La *región de rechazo* especifica los valores del estadísitico de prueba para el cual la hipótesis nula ha de ser rechazada a favor de la hipótesis alternativa.

Para cualquier región de rechazo fija, dos tipos de errores se pueden cometer al llegar a una decisión. Podemos decidir a favor de H_a cuando H_0 es verdadera (*error tipo I*), o podemos decidir a favor de H_0 cuando H_a es verdadera (*error tipo II*). La probabilidad de un *error tipo II* se denota por β. La probabilidad de un *error tipo I* está denotada por α. También se conoce como *nivel de significancia* o *nivel* de la prueba. La región está diseñada para cumplir con dos criterios. Primero, la probabilidad que se tome un valor dentro de la región crítica *si la hipótesis nula es verdadera* es pequeña. Esto puede asegurarse escogiendo un valor suficientemente pequeño para α. Si $\alpha = 0.05$, entonces, la probabilidad de que el estadístico de prueba tome un valor dentro de la región crítica es únicamente de 0.05. De tal forma, es improbable que el estadístico de prueba manifieste un valor en la región crítica *si la hipótesis nula es verdadera*. Segundo, la probabilidad de que el estadístico de prueba tome un valor dentro de la región crítica debe incrementarse cuando la hipótesis nula sea falsa. La lógica de la prueba de hipótesis establece la hipótesis nula como la condición por rechazar.

Decidir si la hipótesis alternativa será elegida depende de si el estadístico de prueba toma o no un valor dentro de la región crítica. Si esto sucede, la hipótesis nula se rechaza en favor de la alternativa, porque es improbable que el estadístico de prueba esté dentro de la región crítica si la hipótesis nula es verdadera. De hecho, la probabilidad es únicamente α. Por otra parte, si la hipótesis nula es falsa, la probabilidad se incrementa. Por lo tanto, si el estadístico de prueba está dentro de la región crítica, la explicación lógica para este suceso es que la hipótesis nula es falsa.

Para reportar el rechazo de la hipótesis nula se declara que la prueba fue **estadísticamente significativa**, mientras que la decisión de no rechazarla se reporta como **estadísticamente no significativa**.

2.7.2. Niveles de significancia alcanzados o valores de p

Una vez tomada una decisión sobre un estadístico de prueba, a veces es posible presentar el valor de p o el nivel de significancia alcanzado y que está relacionado con una prueba. Esta cantidad es un estadístico que representa el valor más pequeño de α para el cual se puede rechazar la hipótesis nula.

Si W es un estadístico de prueba, el *valor de p*, o *nivel de significancia alcanzado*, es el nivel más pequeño de significancia α para el cual la información observada indica que hipótesis nula debe ser rechazada.

Cuanto más pequeño sea el valor de p, es más fuerte la evidencia de que la hipótesis nula debe ser rechazada. Entonces, la hipótesis nula debería ser rechazada para cualquier valor de p por abajo de α incluyendo el valor de α. De otro modo, si α es menor que el valor de p, la hipótesis nula no puede ser rechazada.

Para calcular los valores de p, si se va a rechazar H_0 en favor de H_a para valores pequeños de un estadístico de prueba W, por ejemplo {Región de rechazo : $w \leq k$}, el valor de p relacionado con un valor observado w_0 de W está dado por valor de p = P($W \leq w_0$, cuando H_0 es verdadera). Análogamente, si fuéramos a rechazar H_0 a favor de H_a para valores grandes de W, por ejemplo {Región de rechazo : $w \geq k$}, el valor de p relacionado con el valor observado w_0 es valor de p = P($W \geq w_0$, cuando H_0 es verdadera).

2.7.3. Potencia de las pruebas

La *potencia* de una prueba se utiliza para evaluar el desempeño de una prueba. Básicamente, la potencia de una prueba es la probabiliad de que la prueba rechace la hipótesis nula. Supóngase que W es un estadístico de prueba, la potencia de la prueba denotada como potencia(θ), se define como potencia(θ) = P(W esté en la región de rechazo cuando el valor del parámetro es θ).

Supóngase que se quiere probar la hipótesis nula $H_0 : \theta = \theta_0$ y que θ_a es un valor particular de θ escogido para H_a, la potencia de la prueba para $\theta = \theta_0$ es igual a la probabilidad de rechazar H_0 cuando ésta es verdadera. Es decir, potencia(θ_0) = α, la probabilidad de cometer un error de tipo I. Y para $\theta = \theta_a$, la potencia de la prueba mide la capacidad para detectar que la hipótesis nula es falsa. Dado que β es la probabilidad de aceptar H_0 cuando $\theta = \theta_a$. Se puede relacionar la potencia de la prueba para θ_a y la probabilidad de un error tipo II como sigue: $\beta = 1 -$ potencia(θ_a).

La prueba más potente se define como una prueba que hace mínimo β, el tamaño de error tipo II.

El lema de Neyman-Pearson se refiere a pruebas de tamaño α de una hipótesis $H_0 : \theta = \theta_0$ contra una hipótesis alternativa $H_a : \theta = \theta_a$, estas preubas se basan en una muestra $x_1, ..., x_n$ de tamaño n fijo de una población con función de densidad de probabilidades $f(x; \theta)$. Y con $l(\theta) = f(x_1; \theta)f(x_2; \theta) \dots f(x_n; \theta)$ como la función de verosimilitud. Entonces, para un valor dado de α, la prueba que maximiza la potencia en θ_a tiene una región de rechazo determinada por $\frac{l(\theta_0)}{l(\theta_a)} < k$. Tal prueba será la más potente para H_0 frente a H_a.

2.7.4. Determinación de nivel de significancia en GWAS

Dada la enorme cantidad de comparaciones que se realizan en los estudios de GWA es necesario prestar particular atención a las medidas utilizadas para reportar significancia estadística, de lo contrario se pueden obtener muchos falsos positivos y, si el margen de corrección es muy conservador, pueden resultar muchos falsos negativos.

Los análisis de genoma completo involucran muchas variantes polimórficas y gran variedad de modelos de enfermedad posibles. Por ello, dada una variante cualquiera (o conjunto de variantes) es muy poco probable que ésta esté asociada con cualquier fenotipo bajo el modelo supuesto, por lo que tener una evidencia fuerte es requerida afirmar que está asociado. Como se mencionó anteriormente, la probabilidad de tener errores de tipo I es usualmente controlada al establecer un nivel de significanca α a la prueba estadística, normalmente $\alpha = 0.05$. Esta alternativa frecuentista para controlar los errores tipo I es: si n SNPs son evaluados y las pruebas son independietes, y α' es el nivel de significacia apropiado por SNP, entonces

$$P(\text{Estudio tenga errores tipo I}) = 1 - (1 - \alpha)^n,$$

la cual es una función de n, el número de comparaciones realizadas, así como de α. Lo anterior lleva a la corrección de Bonferroni $\alpha' \approx \alpha/n$. La corrección de Bonferroni es una corrección de comparaciones múltiples cuando varias pruebas estadísticas son realizadas simultáneamente (ya que mientras un valor α dado puede ser apropiado para cada comparación individual, no es para el conjunto de todas las comparaciones).

Con el fin de evitar una gran cantidad de falsos positivos, el valor α debe ser reducido para tener en cuenta el número de comparaciones se realizan. Por ejemplo, si probamos una hipótesis nula que es verdadera utilizando un nivel del significancia de 0.05, tenemos que hay una probabilidad de 0.95 de que no sea rechazada, es decir, de obtener la conclusión correcta. Por otra parte, si se estudian dos eventos cuyas hipótesis nulas son verdaderas, la probabilidad de que ninguna de las pruebas estadísticas resulte significativa es $0.95 \times 0.95 = 0.90$. Entonces si se prueban 20 eventos cuyas hipótesis nula son verdaderas, la probabilidad de que ninguno resulte significativo es $0.95^{20} \approx 0.36$, por lo que da una probabilidad de $1-0.36 = 0.64$ de tener un resultado significativo. Es así que es más probable tener un resultado significativo que no tenerlo. En general, si se tienen k pruebas independientes significativas, que a nivel se significancia α la hipótesis nula no se rechaza para todas. La probabilidad de no obtener resultados significativos es $(1 - \alpha)^k$. Si se hace α lo suficientemente pequeño se puede lograr que ninguna de las pruebas por separado resulte significativa con una probabilidad de 0.95. Como α va a ser muy pequeño se tiene que $(1-\alpha)^k \approx 1 - k\alpha$. Si $k\alpha = 0.05$, entonces $\alpha \approx \frac{0.05}{k}$ y tenemos una probabilidad de 0.05 de que uno de las k pruebas tenga valor de p menor a α si la hipótesis nula es verdadera.

La aplicación directa a GWAS es que con un nivel de significancia típico como $\alpha = 0.05$, $\alpha = 0.01$ o $\alpha = 0.001$, la probabildad de no cometer errores de

tipo I es muy pequeña. Al ajustar con Bonferroni un nivel de significancia de $\alpha = 0.05$ podemos estar un 95 % seguros que ninguno de los resultados obtenidos fue debido al azar. Basándose en la corrección de Bonferroni y suponiendo que hay millones de variantes independientes a lo largo de todo el genoma humano, el margen estándar de evidencia significativa en GWAS para asegurar la identificación de asociaciones de fenotipo-genotipo es considerada como $p < 5 \times 10^{-8}$ o $p < 1 \times 10^{-8}$, para $\alpha = 0.05$ y $\alpha = 0.01$, respectivamente. Sin embargo, el evitar errores de tipo I puede inflar la ocurrencia de errores tipo II.

En GWAS el poder para detectar asociaciones es determinado, en parte, por las frecuencias alélicas y el tamaño de sus efectos; dado que estas variables son constantes, sólo se puede ajustar el tamaño de la muestra. Entre más aumente el tamaño de la muestra, aumenta la potencia para detectar bajas frecuencias y/o efectos pequeños en las variantes genéticas. También, la naturaleza dependiente de los datos genéticos, donde SNPs en desequilibrio de ligamiento (LD) están correlacionados en cierto grado, puede llevar a una sobre-corrección con la utilización de los ajustes de Bonferroni, ya que una de las suposiciones de la corrección de Bonferroni es que todas las comparaciones son independientes. SNPs vecinos en un cromosoma tienden a heredarse en bloques y no son independientes, por lo que el ajuste de Bonferroni termina siendo muy conservador.

El problema de la determinación del umbral de significancia tiene dos raíces [8]. La primera, la determinación o aproximación de la distribución de la prueba estadística bajo la hipótesis nula. La segunda, es dada por las múltiples pruebas de hipótesis que están implícitas al hacer una búsqueda a lo largo de todo el genoma para encontrar genes candidatos. También se incluyen factores que pueden variar entre experimentos y puede influenciar la distribución de la prueba estadística. Entre estos factores se incluye: tamaño de la muestra, tamaño del genoma del organismo en estudio, la densidad del mapa genético, la proporción y patrón del los datos faltantes, entre otros.

Una alternativa para encontrar el nivel de significancia adecuado es el método de las permutaciones [8]. El procedimiento es estadísticamente correcto cuando es utilizado en un conjunto de pruebas estadísticas basadas en regresiones o probabilidades y para cualquier distribución de cualquier rasgo. Dado que el procedimiento es empírico, va a reflejar automáticamente las características del experimento al que es aplicado.

El enfoque que tiene este método de estimación del nivel de significancia se basa en la simple observación de la asociación fenotipo-marcador. Se plantea de la siguiente manera: si los datos indican asociación entre marcador-fenotipo entonces se eliminará dicha asociación al reasignar los valores del fenotipo a cada individuo, y donde cada individuo conserva su mapa genético, dado que se está rompiendo la relación genotipo-fenotipo. Por otra parte, si no existen asociaciones indicadas en específicas regiones del genoma, al alternar aleatoriamente los valores de los fenotipos sobre los individuos no alterará la distribución de la prueba estadística. Cualquier asociación debería de ser pequeña y atribuida al azar. Al calcular el valor de una prueba estadística a cada marcador en el conjunto de datos permutados, básicamente se muestrea la distribución nula correspondiente a la hipótesis de no asociación entre los rasgos y los mapas

genético. Dado que los valores de los mapas genéticos y los valores de los rasgos no son alterados al ser permutados, la distribución va a tomar en cuenta de manera automática las características particulares del experimento en cuestión.

Para la estimación del nivel de significancia, primero se etiqueta a todos los individuos del experimento de 1 a n. Los datos son alternados al hacer una permutación aleatoria de los índices $1, ..., n$ y asignando el i-ésimo valor de fenotipo a aquel individuo cuyo índice es dado por el i-ésimo elemento de la permutación. Posteriormente, los datos permutados son analizados para detectar asociaciones entre los marcadores y los fenotipos. Los resultados de las pruebas estadísticas para cada SNP son almacenados y el procedimento anterior (la permutación y análisis) es repetido nuevamente N veces. Al final de este procedimiento se tienen almacenados los resultados los análisis de asociación en N conjuntos de datos permutados.

El nivel de significancia buscado es un valor crítico de $100(1-\alpha)\,\%$ válido para todos los puntos del análisis de manera simultánea y se obtiene de la siguiente manera. Para cada permutación realizada se registra el valor mínimo de p de todos los SNPs a lo largo del genoma. El conjunto de valores mínimos de p es utilizado para estimar una distribución sin corregir de valores de p bajo la hipótesis nula de ninguna asociación real en el estudio. Este es ordenado y su percentil $100(1-\alpha)$ es el valor crítico buscado.

Sin embargo, el método de permutaciones presenta dos grandes desventajas. Primero, es computacionalmente exhaustivo, debido a que para cada permutación se tiene que realizar un análisis de asociación, y para realizar la estimación del valor crítico se recomienda no hacer menos de $10,000$ permutaciones. Segundo, cuando se tiene una muestra con subestructura poblacional, al momento de hacer hacer las permutaciones se rompe con esta subestructura, dando no fiabildad al cálculo.

Una forma de atacar el problema que se tiene con la corrección de Bonferroni es detectar cuántas pruebas estadísticas independientes se están realizando. En el contexto del cálculo de componentes principales en datos de genotipo, se define el número de comparaciones independientes en términos del número de componentes principales que abarcan una gran proporción de la varianza de los datos (se sugiere $95\,\%$). El conjunto de SNPs informativos representados por estos componentes puede ser usado para inferir la estructura del resto del conjunto de datos con un alto grado de fidelidad, y hacer un ajuste a la corrección de Bonferroni. De esta forma se puede obtener una estimación del nivel de significancia menos estricto

$$\alpha_{\text{GWAS}} = \frac{\alpha}{n_{\text{informativa}}}. \tag{2.35}$$

Donde α_{GWAS} se define como el nivel de significancia que se establecerá a para todos los estadísticos de prueba realizados en todo el genoma, y la $n_{\text{informativa}}$ se refiere al número de pruebas independientes obtenido por medio de PCA. Este último enfoque fue implementado en el programa SimpleM [23–25], y es una alternativa atractiva cuando no se tiene la suficiente potencia estadística para alcanzar el nivel de significancia de $p < 5 \times 10^{-8}$. También se ha establecido por

convención que a partir del nivel $p < 1 \times 10^{-5}$ existe la posibilidad de asociación genética, a este nivel se le conoce como *nivel de significancia sugestivo* [14].

Capítulo 3

Estudio piloto para la identificación de marcadores genéticos asociados a rasgos metabólicos en población indígena mexicana a través del escrutinio completo del genoma

3.1. Antecedentes

Los estudios asociación de genoma completo han sido muy útiles para identificar alelos de riesgo asociados a diferentes fenotipos. Los GWAS se hacen a través de análisis de microarreglos que contienen desde 100 mil y hasta 2.5 millones de SNPs comunes distribuidos a lo largo de todo el genoma; son identificados utilizando como referencia las poblaciones del proyecto de HapMap (subsección 1.5.2). Son libres de hipótesis, y dado que los SNPs comunes tienen un efecto muy pequeño se requiere de un gran número de individuos para contar con suficiente potencia estadística para alcanzar el nivel de significancia necesario (valor aproximado a $P < 10^{-5}$) [41].

En los últimos años por medio de GWAS se han identificado cientos de variantes genéticas asociadas a más de 80 enfermedades y rasgos [29]. Entre ellas han sido estudiadas la obesidad, diabetes tipo 2, colesterol y otras enfermedades metabólicas. Sin embargo, los GWAS se han concentrado mayormente

en poblaciones europeas. Sin embargo, la población europea contiene solo un subconjunto de la variabilidad genética humana. Es importante estudiar otras poblaciones, especialmente cuando se trata de enfermedades que tiene una mayor prevalencia en ciertas poblaciones que en las poblaciones europeas. Distintas poblaciones tienen distintas frecuencias alélicas y esto puede afectar la detección de variantes de riesgo, puede ser ciertos eventos de recombinación o mutaciones en alguna población en particular ayuden a detectar alguna variante asociada a una enfermedad.

Las enfermedades metabólicas son todas aquellas que están relacionadas con la perturbación del metabolismo, que es el proceso de convertir comida en energía a nivel celular. Miles de enzimas participan en numerosas vías metabólicas interdependientes para llevar a cabo este proceso. Las enfermedades metabólicas afectan la habilidad de la célula para realizar reacciones bioquímicas críticas que incluyen el procesamiento y transporte de proteínas (amino ácidos), carbohidratos (azúcares y almidones) o lípidos (ácidos grasos).

En México, el aumento en la prevalencia de la obesidad en los últimos siete años ha sido alarmante, ya que más del 70 % de la población adulta presenta sobrepeso u obesidad. De manera importante, se ha considerado a la obesidad como uno de los principales factores de riesgo para el desarrollo de hipertensión; displisemias, enfermedades cadiovasculares y diabetes tipo 2. Las últimas dos son las principales causas de muertes de México (ENSANUT, 2006). Debido a la alta prevalencia de enfermedades metabólicas observada en la población mexicana, se ha sugerido que la susceptibilidad genética que presenta esta población proviene del componente indígena [40], [15].

A la fecha sólo se han realizado algunos pocos GWAS de baja densidad en mexicanos [5], mexicano-americanos [27] y poblaciones nativas de Estados Unidos [26], [39], [58] identificando distintos loci asociados a diversos parámetros metabólicos, lo que sugiere que aun cuando los microarreglos utilizados no incluyen información del componente nativo americano, éstos pueden ser útiles para la identificación de genes asociados a rasgos metabólicos.

3.2. Objetivos

Dado que se trata de una prueba piloto, el objetivo general de este proyecto consiste en replicar SNPs previamente asociados a distintos rasgos metabólicos como son índice de masa corporal (IMC), colesterol total (c-Total), colesterol HDL (c-HDL), colesterol LDL (c-LDL), triglicéridos y glucosa en una muestra de población indígena. Como objetivo secundario se tiene evaluar si es posible identificar SNPs nuevos en una muestra de tamaño moderado.

3.3. Materiales y Métodos

3.3.1. Descripción de la población

El estudio incluyó muestras de sujetos no relacionados de cuatro poblaciones indígenas mexicanas (zapoteco de Oaxaca, nahua y totonaco de Puebla y seri de Sonora).

Como criterios de inclusión se consideró individuos adultos ($\geq$ 18 años), residentes de las comunidades estudiadas con ambos padres biológicos del mismo origen étnico. Se excluyeron individuos sujetos a tratamiento farmacológicos para la diabetes, displidemias o que aceleren la pérdida de peso, mujeres gestantes, sujetos con diagnóstico previo de problemas endocrinológicos, síndromes genéticos u obesidad monogénica, cáncer, nefropatía diabética, entre otros.

Los participantes se les aplicaron cuestionarios estandarizados para obtener información referente al nivel socioeconómico, historia clínica, antecedentes familiares, actividad física, uso de medicamentos, alcoholismo y tabaquismo. A los individuos que cumplieron con los criterios de inclusión se les realizaron mediciones antropométricas y bioquímicas. Se tomaron mediciones antropométricas como peso y talla. El IMC se obtuvo a través de la ecuación peso/talla^2. Se obtuvieron determinaciones bioquímicas como perfil de lípidos (c-Total, c-HDL, c-LDL, triglicéridos) y glucosa.

Con la finalidad de incrementar la potencia estadística de esta prueba piloto, se seleccionaron individuos con base en las niveles de c-HDL, es decir, se seleccionaron aquellos individuos con niveles de HDL ≤ 35 e individuos con niveles de HDL > 50.

3.3.2. Genotipificación y controles de calidad

El DNA genómico se obtuvo a partir de una muestra de sangre periférica, utilizando reactivos comerciales (QIAamp DNA Blood Maxi Kit 50, Quiagen). La genotipificación se realizó con el microarreglo SNP 6.0 de Affymetrix. Que incluye $906,703$ de SNPs comunes.

Se aplicaron controles de calidad a los datos con base a la frecuencia alélica, consistencia de sexo y tasa de información faltante por SNP y por individuo. Para lo anterior se utilizó el software PLINK [52].

3.3.3. Análisis estadístico

Los rasgos metabólicos analizados fueron c-Total, c-HDL, c-LDL, triglicéridos y glucosa. Cada una de estas variables fue analizada de manera independiente.

Las variables de c-HDL y glucosa se dicotomizaron para su análisis, es decir, se consideraron como casos individuos con valores de c-HDL ≤ 35 y glucosa > 126, y se consideraron controles individuos con valores de c-HDL ≥ 50 y glucosa ≤ 126.

Los valores de c-Total, c-LDL y triglicéridos fueron analizados como variables continuas. Para estas últimas, se revisó la normalidad de los datos con la prueba de Lilliefors, la cual es una adaptación de la prueba de Kolmogorov-Smirnov. En las pruebas de normalidad se observó que los valores de triglicéridos no tuvieron una distribución normal, por lo que los valores fueron transformados con logaritmo base 10. Los valores de triglicéridos ya transformados cumplieron con las pruebas de normalidad.

Para cada rasgo se revisó que no hubiera individuos con valores extremos. Sólo se encontró un individuo con valor extremo en IMC y este valor se ajustó al máximo valor dentro del rango de los valores extremos.

Se hicieron pruebas de medias no paramétricas para examinar la existencia de diferencias significativas entre grupos de la población analizada. Se utilizó la prueba de Wilcoxon para analizar diferencias entre hombres y mujeres, y la prueba de Kruskal-Wallis para evaluar si había diferencias entre las poblaciones.

Con el fin de detectar si había estratificación poblacional, se analizó la estructura de la población por medio del cálculo de componentes principales, método implementado en el programa EIGENSTRAT [47].

Para el análisis de asociación se utilizaron modelos lineales mixtos los cuales incluyen como efecto aleatorio fenómenos tales como relaciones crípticas, estratificación poblacional y otras variables confusoras. Dicho método está implementado en el programa EMMAX [34]. Las variables confusoras que se consideraron fueron edad, sexo y los 2 primeros componentes principales (para tomar en cuenta que los individuos provienen de distintos grupos étnicos). Estas covariables fueron consideradas para el análisis de asociación de c-HDL, y a los otros rasgos se les agregó el valor de c-HDL como covariable también, ya que lo sujetos fueron seleccionadas en base a los niveles de c-HDL.

El análisis de asociación tuvo dos etapas:

1. Replicar SNPs previamente reportados.

2. Realizar un análisis de asociación de genoma completo.

La primera etapa consistió en examinar la replicación de regiones previamente reportados (las listas de SNPs para IMC se obtuvieron de [16] , de glucosa se obtuvieron de [18], y los datos de c-HDL, c-LDL, triglicéridos, y c-Total se extrajeron de [57]). Se identificaron los SNPs reportados en nuestra base de datos, si alguno no estaba se buscó un SNP cercano que estuviera en alto desequilibrio de ligamiento ($r^2 > 0.8$) con él. Luego, se definieron intervalos de aproximadamente 5 Mb alrededor de cada SNP, se calcularon haplotipos de la población estudiada y si las fronteras de dichos intervalos estaban dentro de un haplotipo, se recorrió la frontera al extremo superior o inferior del haplotipo, segun fuera el caso. Lo anterior se hizo con la ayuda del software Haploview [4] y PLINK [52]. Ya definidos los intervalos se realizó el análisis de asociación utilizando EMMAX y ajustando por las covariables anteriormente mencionadas. Para reportar que una región fue replicada se utilizó la corrección de Bonferroni, considerando como pruebas independientes el número de regiones analizadas.

La segunda etapa consistió en un análisis de asociación de genoma completo, utilizando EMMAX y ajustando por las covariables anteriormente mencionadas. Para determinar el nivel de significancia se utilizó un método menos estricto que la corrección de Bonferroni ya que por medio de componentes principales determina el número de factores independientes que hay en los datos de estudio. Dicho método está implementado en el programa SimpleM [25]. Una vez teniendo el número de factores independientes se calculó la corrección de Bonferroni a nivel de significancia nominal de 0.05.

Finalmente, para tener una mejor visualización de los datos en las dos etapas se hicieron gráficas de Manhattan y de Q-Q, con ayuda del programa R [1].

3.4. Resultados

Descripción de fenotipos

El estudio piloto incluyó 126 individuos de cuatro poblaciones indígenas (53 Nahuas, 24 Seris, 25 Totonacos y 24 Zapotecos). En las tablas 3.1 y 3.2 se muestran los valores de cada uno de los rasgos estratificado por género y por población respectivamente. En la primera sólo se identifican diferencias estadísticamente significativas para IMC. Y en la segunda, se observan diferencias significativas para glucosa, triglicéridos y c-LDL. En los tres casos destaca la población Seri con los valores más altos. También cabe destacar que la población Nahua tiene una mayor representación dado que el muestreo se realizó en dos localidades diferentes.

	Hombre	Mujer
N	54	72
IMC (kg/m^2)	25.9 ± 4.2	$27.5 \pm 4.6*$
Glucosa (mg/dL)	117.26 ± 52.08	135.00 ± 69.78
c-total (mg/dL)	163.3 ± 50.4	185.2 ± 67.6
Triglicéridos (mg/dL)	193.9 ± 94.6	205.3 ± 95.5
c-LDL (mg/dL)	103.6 ± 28.4	107.7 ± 33.2
c-HDL (mg/dL)	40.67 ± 14.26	42.01 ± 14.21

Tabla 3.1: Fenotipos estudiados, en la población estratificada por sexo. Se indica con $*$ aquellos que tuvieron $P < 0.05$.

Controles de calidad

Se comenzó con un total de $906,703$ SNPs. Se excluyeron $272,686$ SNPs que tuvieron una frecuencia alélica por debajo de $0.02\,\%$; $67,333$ SNPs con tasa información faltante mayor al $95\,\%$. Al final, quedaron $593,823$ SNPs en total.
En cuanto a los exclusión de individuos, no se excluyeron individuos con base

	Nahua	Seri	Totonaco	Zapotecos	
N	53	24	25	24	
IMC (kg/m^2)	26.34 ± 3.97	27.11 ± 5.62	26.73 ± 4.92	27.79 ± 3.98	
Glucosa (mg/dL)	112.26 ± 47.10	163.42 ± 74.71	115.56 ± 42.99	$137.12 \pm 83.93*$	
c-Total (mg/dL)	165.30 ± 56.38	172.62 ± 76.52	173.68 ± 39.47	204.38 ± 69.46	
Triglicéridos (mg/dL)	176.92 ± 76.90	216.25 ± 46.31	175.72 ± 103.54	$262.29 \pm 126.82*$	
c-LDL (mg/dL)	105.52 ± 31.34	125.56 ± 24.03	95.44 ± 35.27	$101.25 \pm 21.08*$	
c-HDL (mg/dL)	38.79 ± 13.70	42.62 ± 8.72	43.12 ± 14.97	44.33 ± 18.17	

Tabla 3.2: Fenotipos estudiados estratificados por población. Se indica con $*$ aquellos que tuvieron $P < 0.05$.

en información faltante ya que todos los individuos tuvieron una tasa de genotipificación mayor al 94 %. Adicionalmente, se corrigieron las inconsistencias de sexo encontradas, 4 en total, poniendo el género determinado por el grado de heterocigosidad del cromosoma X.

En la tabla 3.3, se presenta el valor de control genómico λ que se obtuvo al aplicar controles de calidad. Se ve cómo se van corrigiendo los valores de λ al ir aplicando controles de calidad y al utilizar distintos métodos que toman en cuenta factores que pueden causar sesgo al realizar el análisis de asociación. En la primera columna se muestra el factor de inflación al hacer un análisis de asociación a los datos en crudo; se puede observar que se presentan valores de λ muy altos, en particular, para los rasgos glucosa, c-HDL y c-LDL. Al aplicar controles de calidad se siguen presentado valores de $\lambda > 1$, indicando que existen factores (como estratificación poblacional, relaciones crípticas, entre otros) que estan causando inflación en los resultados. Finalmente, al realizar el análisis de asociación con EMMAX, se puede notar que todos los valores de $\lambda \approx 1$ demostrando así que éste corrigió de manera adecuada todos los factores que estaban causando la inflación; en especial se puede apreciar que en la población sí se encontraban relaciones crípticas presentes y tambien había efectos de estratificación poblacional. De igual forma, se puede corrobar lo anterior en la figura B.2, donde notamos que los valores de p se ajustan a la distribución esperada.

	Datos en crudo		Datos con controles de calidad		EMMAX + covariables
Rasgos	SNPs	λ	SNPs	λ	λ
IMC	906703	0.973664	593823	0.976148	0.99
Glucosa	906703	1.3634	593823	1.30842	1
Triglicéridos (logaritmo)	906703	1.08754	593823	1.26809	0.97
c-Total	906703	1.07476	593823	1.0326	0.96
c-HDL	906703	1.4319	593823	1.13793	0.99
c-LDL	906703	1.36158	593823	1.34717	0.98

Tabla 3.3: Valor de control genómico λ para cada una de las etapas del análisis de asociación.

Estratificación poblacional

En la figura 3.1, a través de un análisis de componentes principales, se muestra la estructura de los distintos grupos poblacionales en la muestra. Se puede

obsevar que las poblaciones no son homogéneas, en particular la población Seri está muy apartada de las otras. Estos dos componentes se tomaron en cuenta al momento de realizar el análisis de asociación.

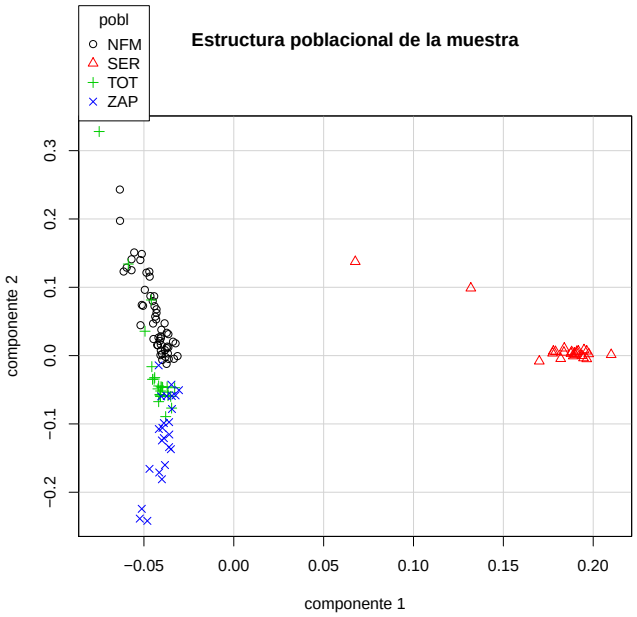

Figura 3.1: Gráfica donde se muestra la estrucutra poblacional de los 126 indígenas incluidos en la muestra. La abreviación NFM denota a la población nahua, SER denota a la población Seri, TOT denota a la población totonaco y ZAP a la población zapoteco.

Análisis de asociación

Etapa 1: Replicar SNPs previamente reportados

En el caso de c-HDL, se replicaron 11 regiones (gráfica a) de la figura A.1 de las 46 estudiadas. Se puede ver en la tabla A.1 en varias ocasiones el SNP detectado como más significativo se encontraba en más de dos vencidades creadas. Los intervalos que se abrieron para *hit*[1] rs11776767 y para el *hit* rs9987289 se intersectaron, es así, que el SNP que resultó ser más significativo para ambas regiones resultó ser el mismo. Los mismo para los *hits* rs4759375, rs4765127 y rs838880. En el gen CETP se encontró el SNP con el nivel de significancia más alto de $1.37\text{x}10^{-05}$.

[1]Llamamos *hit* al SNP con mayor significancia estadística en una región.

Para IMC, se replicaron 6 regiones de 32 regiones estudiadas 3, véase la gráfica b) de la figura A.1 y la tabla A.1. La región que alcanzó a tener el nivel de significancia más alto fue en el cromosoma 9 para el gen LRRN6C.

En cuanto a c-LDL, como se puede ver en la gráfica c) de la figura A.1 y la tabla A.1, 5 regiones de 37 se replicaron. La región con el nivel de significancia más alto fue la correspondiente al gen DNAH11.

En c-Total se replicaron 4 regiones de 51, como se puede ver en la gráfica d) de la figura A.1 y en la tabla A.1.

En triglicéridos se identificaron 4 regiones replicadas de 32 estudiadas; véase la gráfica e) de la figura A.1 y en la tabla A.1.

Por último, para glucosa se analizaron 17 regiones anteriormente reportadas, como se puede ver en la gráfica f) de la figura A.1 y en la tabla A.1; de éstas sólo se replicó una en el cromosoma 3.

Etapa 2: Análisis de asociación a nivel de todo genoma completo

A nivel de todo el genoma, en la caso de c-HDL se reportan 3 SNPs que sobrepasan el nivel de significancia correspondiente; véase la gráfica a) de la figura B.1 y en la tabla B.1.

Para BMI se encontraron 4 regiones que sobrepasaron el nivel de significancia sugestivo, mientras que ninguno sobrepasó el nivel de significancia correspondiente (gráfica b) de la figura B.1 y en la tabla B.1) .

En c-LDL sólo dos SNPs sobrepasaron el nivel de significancia sugestivo (gráfica c) de la figura B.1 y en la tabla B.1).

Para c-Total solo un SNP sobrepasó el nivel de significancia establecido correspondiente (gráfica d) de la figura B.1 y en la tabla B.1).

Como se puede ver en la gráfica e) de la figura B.1 y en la tabla B.1, en triglicéridos sólo dos regiones sobrepasaron el nivel de significancia sugestivo.

Mientras que en glucosa se encontraron tres regiones que sobrepasan el nivel de significancia sugestivo de $1x10^{-5}$, como se puede observar en la gráfica f) de la figura B.1 y en la tabla B.1.

3.5. Discusión

Por medio de los componentes principales se pueden obtener una especie de"mapa sintético", que refleja el "aislamiento por distancia" que hay entre las poblaciones, es decir, se pone en evidencia la similitud genética que existe entre las poblaciones, donde poblaciones vecinas presentan mayor similitud por lo que en el mapa van a estar cercanas unas de las otras, y para las poblaciones lejanas se va a presenciar el efecto contrario [53]. Con lo anterior se puede interpretar la figura 3.1, y explicar por qué la población Seri se encuentra tan apartada de las otras poblaciones. Geográficamente esta población está muy distante de las otras ya que se encuentra en el estado de Sonora, el norte del país, mientras que las restantes están concentradas en regiones centrales de México.

Por otra parte, se utilizaron componentes principales para identificar diferencias de ancestría entre las distintas poblaciones estudiadas. Dado que nuestra población no es homogénea, fue importante incluir estas diferencias como covariables al momento de realizar el análisis de asociación, así como, utilizar un método que tuviera contemplado la existencia estratificación poblacional, como lo es EMMAX. De lo contrario, podrían surgir errores de tipo I, causados por las diferencias de frecuencias alélicas que hay debido a las diferencias de ancestría que hay entre las poblaciones más que por ser asociaciones con la enfermedad. Especialmente para los rasgos glucosa, triglicéridos y c-LDL, ya que éstos presentaron diferencias en las distribuciones entre las poblaciones estadísticamente significativas (tabla 3.2). Cabe señalar que bajo otras circunstancias lo más recomendable sería excluir a las poblaciones que se presentraran muy apartadas al resto, en este caso la población Seri (figura 3.1). Sin embargo, debido a la pequeña cantidad de individuos que se tienen, se escogió no excluirlos y así no perder más potencia estadística.

En cuanto al análisis de asociación enfocado a regiones reportadas anteriormente, es notable ver que con sólo 126 individuos se logra obtener señales de réplica para cada uno de los fenotipos. Sobretodo si observamos que la mediana usual del tamaño de muestra inicial más réplicas es de $7,858$ (rango, $146 - 91,749$) individuos [29].

Particularmente en c-HDL, más del 20 % de las regiones estudiadas replican en este estudio, lo cual es consistente con el diseño original de valores extremos de c-HDL. Le sigue IMC donde se replica el 18.7 %, luego sigue c-LDL con el 13.5 % de réplicas y triglicéridos cuenta con el 12.5 %. Los que menor proporción de réplicas presentaron fueron c-Total con 7.8 % y glucosa con el 5.8 %. El hecho de que se haya logrado replicar una buena proporción de las regiones anteriormente reportadas en otros estudios da validez al estudio, ya que se espera que la biología se repita.

La ocurrencia de las variantes de riesgo puede cambiar entre poblaciones. También puede suceder que la misma variante esté presente en diversas poblaciones pero que sus frecuencias alélicas cambien. Por estas razones fue importante abrir un intervalo alrededor del hit reportado y capturar la mayor cantidad de variantes que estuvieran en desequilibrio de ligamiento con ese hit. De tal manera, si el hit detectado en otros estudios, por tanto en otras poblaciones, tiene una frecuencia lo suficientemente baja en nuestra población para no ser detectado por medio de las pruebas estadísticas, al capturar un intervalo de SNPs que estén en alto desequilibrio de ligamiento con éste, existe la posibilidad que alguno de estos sí pueda ser detectado.

Para el análisis de asociación de genoma completo, se encontraron regiones de interés para IMC, c-Total y c-HDL.

En el caso de IMC, el SNP rs12987572 en el cromosoma 2 está dentro del gen LRP1B que pertenece a la familia de receptores de c-LDL. Dicho gen se asoció a IMC en población europea en un estudio que incluyó $250,000$ individuos [16].

Para c-Total el SNP rs2631959 en el cromosoma 2 sobrepasa el nivel de significancia de sugestivo y esta muy cerca del establecido por SimpleM, se puede observar que es toda una región la que sube, es decir, que toda una región

que está en desequilibrio de ligamiento sugiere estar asociada a la enfermedad. No se puede establecer la exisitencia de asociación a la enfermedad porque no sobrepasa el nivel de significancia establecido, sin embargo, el hecho de que toda una región suba sugiere la posibilidad de no ser un falso positivo. Este SNP está dentro del gen ALK codifica a un receptor tirosin-kinasa que pertenece a la superfamilia de receptores de insulina. Se han identificado rearreglos, mutaciones o amplificaciones en tumores neoplásicos pero no ha sido asociado a nada más.

En c-HDL, se encontró el SNP rs170860 en el cromosoma 13 que sobrepasa el nivel de significancia sugestivo, este SNP se encuentra en el gen IRS2 (es un receptor de insulina). Y se han identificado variantes raras asociadas a obesidad en niños hispanos [6]. Se ha estudiado que en la mayoría de los casos la resistencia a la insulina se asocia con bajas concentraciones de c-HDL. Bajas concentraciones de c-HDL (por debajo de 35 mg/dL) suponen un aumento del riesgo de enfermedades cardiovasculares, especialmente para las mujeres.

La resistencia a la insulina no es una enfermedad, es una anormalidad fisiológica que, con otras alteraciones, pueden llevar al desarrollo de varias enfermedades cardiovasculares, relacionadas a hipertensión arterial sistémica, obesidad y diabetes, entre otras. Se han encontrado como causa de la resistencia a la insulina, defectos en receptores de insulina. Por lo que, las regiones con receptores de insulina encontradas si se relacionan con los rasgos estudiados y es una señal positiva de la posibilidad de encontrar señales reales asociadas a los rasgos estudiados.

3.6. Conclusión

En base a los resultados obtenidos en las dos etapas se puede concluir que sí es factible identificar nuevas asociaciones si se aumenta el tamaño de la muestra a un tamaño moderado, ya que en la primera etapa se replican resultados ya reportados lo cual da validez al estudio, y en la segunda etapa se lograron identificar genes relacionados a los rasgos estudiados.

Apéndice A

Resultados etapa 1

Tabla A.1: Replicas de SNPS previamente reportados para cada rasgo.

Rasgo	Gen cercano	HIT	Cromosoma	Alelo de riesgo	MAF	Valor de P
c-HDL	ARL15	rs6450176	5	A (G)	0.45	6.79×10^{-4}
	LPA	rs1084651	6	A (C)	0.2	4.99×10^{-4}
	PINX1	rs11776767	8	T (G)	0.46	1.29×10^{-4}
	PPP1R3B	rs9987289	8	T (G)	0.46	1.30×10^{-4}
	SBNO1	rs4759375	12	C (G)	0.5	8.21×10^{-4}
	ZNF664	rs4765127	12	C (G)	0.5	8.21×10^{-4}
	SCARB1	rs838880	12	C (G)	0.5	8.21×10^{-4}
	CETP	rs3764261	16	T (C)	0.38	1.37×10^{-5}
	HNF4A	rs1800961	20	G (A)	0.021	4.60×10^{-4}
	PLTP	rs6065906	20	G (A)	0.021	4.60×10^{-4}
IMC	FANCL	rs887912	2	T (A)	0.328	1.7×10^{-4}
	ETV5	rs9816226	3	G (A)	0.244	4.3×10^{-4}
	FLJ35779, HMGCR	rs2112347	5	T (C)	0.09	1.1×10^{-4}
	LRRN6C	rs10968576	9	T (C)	0.025	3.3×10^{-5}
	GPRC5BC, IQCK	rs12444979	16	A (G)	0.267	1.4×19^{-3}
	FTO	rs1558902	16	A (C)	0.050	2.2×10^{-4}
c-LDL	APOB	rs1367117	2	C (T)	0.25	6.9×10^{-4}
	ABCG5	rs4299376	2	G (C)	0.33	3.75×10^{-4}
	DNAH11	rs12670798	7	T (G)	0.20	1.93×10^{-4}
	PPP1R3B	rs9987289	8	G (T)	0.05	4.53×10^{-4}
	CETP	rs3764261	16	G (T)	0.44	9.38×10^{-4}
c-Total	MOSC1	rs2642442	1	G (A)	0.1441	3.02×10^{-4}
	ABCG5	rs4299376	2	C (G)	0.395	2.51×10^{-4}
	PPP1R3B	rs9987289	8	A (G)	0.22	6.71×10^{-4}
	HNF4A	rs1800961	20	G (A)	0.2185	7.84×10^{-4}
Triglicéridos	KLHL8	rs442177	4	T (A)	0.203	9.07×10^{-4}
	TRIB1	rs2954029	8	G (T)	0.40	1.12×10^{-3}
	ZNF664	rs4765127	12	G (T)	0.28	5.64×10^{-4}
	CYP26A1	rs20688884	10	C (G)	0.026	7.43×10^{-4}
Glucosa	SLC2A2	rs11920090	3	A (G)	0.36	3.32×10^{-3}

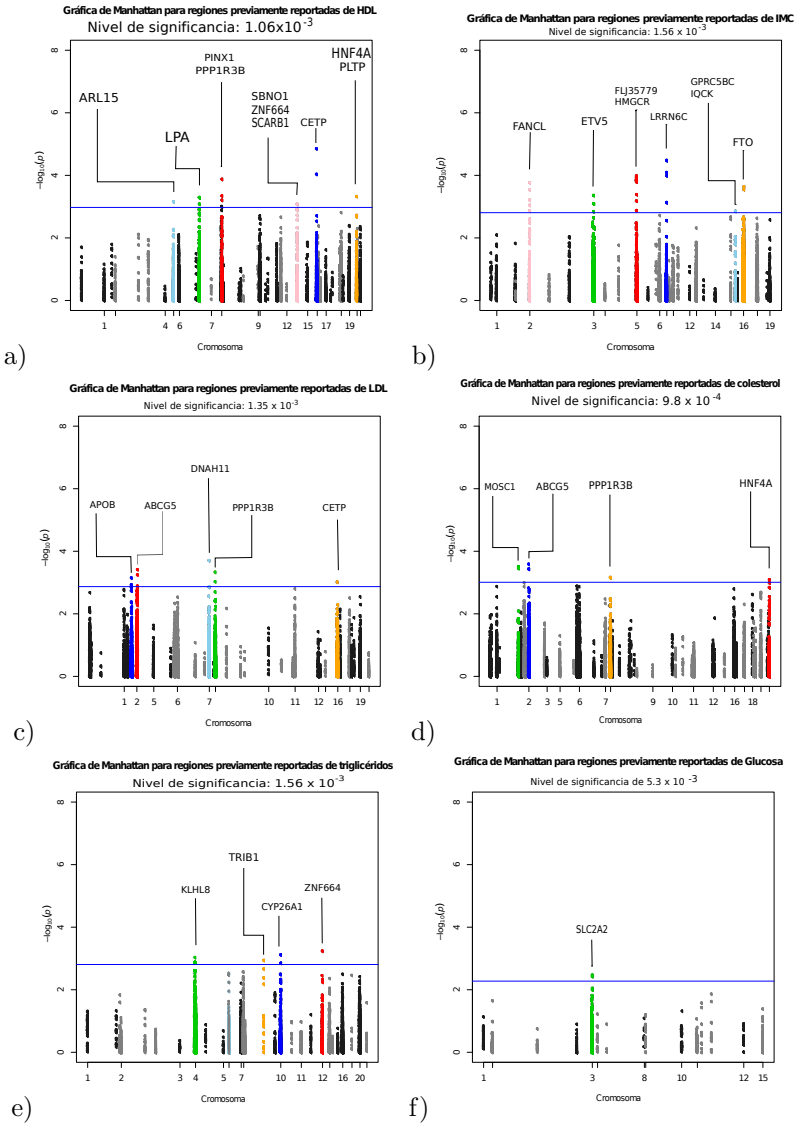

Figura A.1: Gráficas de Manhattan para el análisis de réplicas de SNPs. En la gráfica de Manhattan se grafica la posición del SNP contra el $-$logaritmo(valor de p).

Apéndice B

Resultado etapa 2

Tabla B.1: Resultados para el análsis de asociación de genoma completo. Se indica con $*$ los valores de p que sobrepasaron el nivel de significancia establecido por SimpleM.

Rasgo	Cromosoma	SNP	BP	Alelo de riesgo	MAF	P
c-HDL	12	rs10431603	40335111	G (A)	0.09	$1.36 \times 10^{-06}*$
	13	rs170860	109268933	A (G)	0.11	9.69×10^{-06}
	14	rs2236402	89804131	C (G)	0.23	$8.29 \times 10^{-06}*$
	23	rs5950746	424354	G (T)	0.3	$2.83 \times 10^{-06}*$
IMC	2	rs12987572	141341880	C (A)	0.2415	2.68×10^{-06}
	8	rs9918894	60406762	G (A)	0.3051	8.69×10^{-06}
	12	rs11057971	124052750	C (G)	0.1398	1.32×10^{-06}
	21	rs2826204	20650828	A (T)	0.4	6.56×10^{-06}
c-LDL	5	rs2961831	50490690	C (A)	0.3697	9.42×10^{-06}
	23	rs5979111	9472556	T (C)	0.1935	6.15×10^{-06}
c-Total	2	rs2631959	29822499	C (T)	0.3782	3.11×10^{-07}
	6	rs9480358	156954648	G (C)	0.04622	$1.70 \times 10^{-07}*$
	7	rs17776650	95585553	A (G)	0.1624	7.37×10^{-06}
	14	rs11851797	94239471	T (C)	0.03846	4.02×10^{-06}
	15	rs4246301	98502440	T (G)	0.2605	2.31×10^{-06}
	18	rs3922561	69528506	T (C)	0.1387	3.59×10^{-06}
	18	rs8087246	69540505	A (T)	0.1398	3.10×10^{-06}
Triglicéridos	1	rs16823935	37498561	T (A)	0.05462	2.98×10^{-06}
	5	rs36834	109266956	G (T)	0.08403	6.39×10^{-06}
Glucosa	2	rs6705717	46061470	A (G)	0.2437	5.24×10^{06}
	6	rs6920301	78412826	T (C)	0.1723	5.24×10^{06}
	7	rs972346	82298471	T (C)	0.1597	7.30×10^{06}

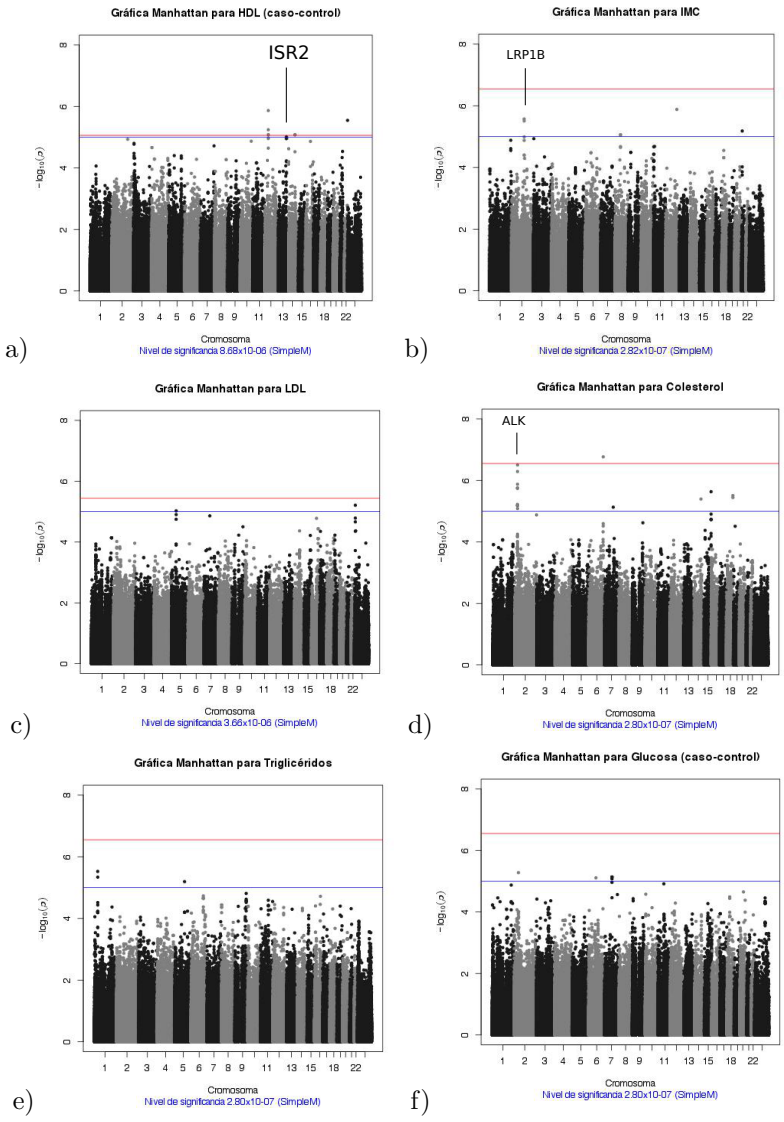

Figura B.1: Gráficas de Manhattan para el análisis de asociación de genoma completo. En la gráfica de Manhattan se grafica la posición del SNP contra el $-$logaritmo(valor de p).

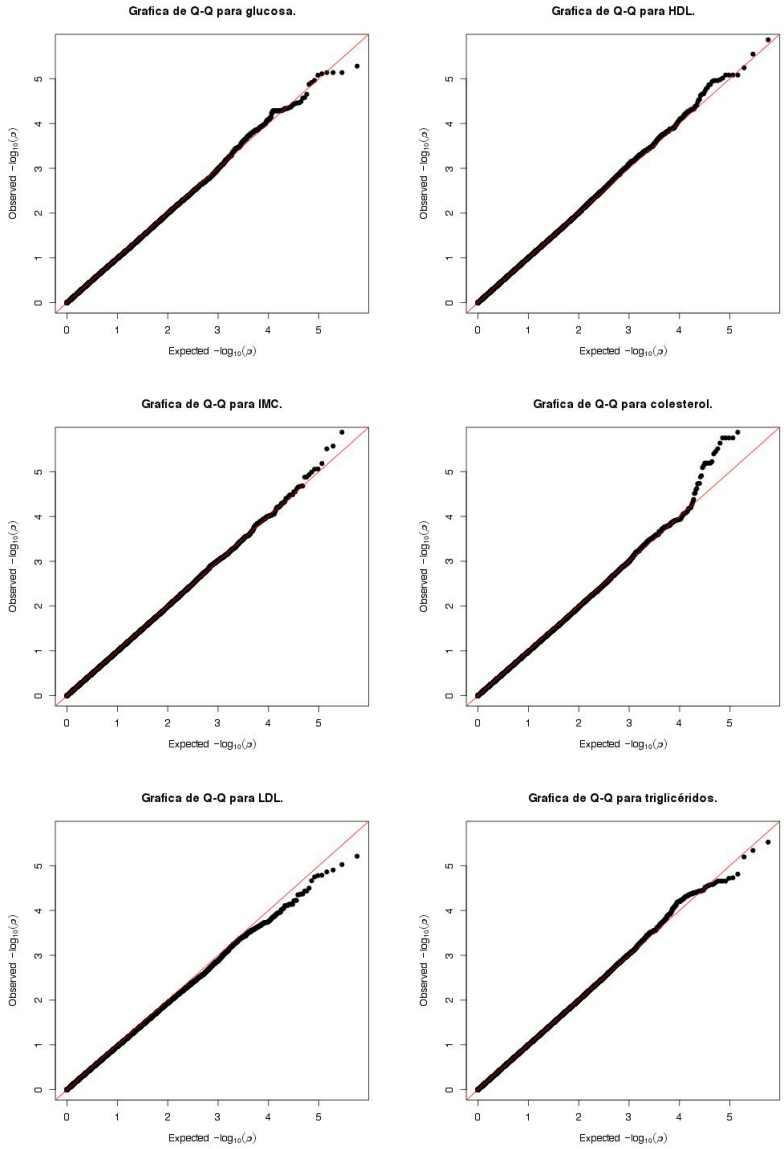

Figura B.2: Gráficas de Q-Q para cada uno de los rasgos estudiados. En la gráfica de Q-Q se grafica el $-$logaritmo(valor de p esperado bajo la hipótesis nula) contra el $-$logaritmo(valor de p observado).

Bibliografía

[1] R Development Core Team (2005). R: A language and environment for statistical computing, reference index version 2.14.0. http://www.R-project.org, 2005.

[2] CA. Anderson, FH. Pettersson, GM. Clarke, et al. Data quality control in genetic case-control association studies. *Nature Protocols*, 2010.

[3] DJ. Balding. A tutorial on statistical methods for population association studies. *Nature reviews. Genetics*, 2006.

[4] JC. Barrett, B. Fry, MJ. Daly, et al. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics*, 2005.

[5] JE. Below, ER. Gamazon, JV. Morrison, A. Konkashbaev, et al. Genome-wide association and meta-analysis in populations from starr county, texas, and mexico city identify type 2 diabetes susceptibility loci and enrichment for expression quantitative trait loci in top signals. *Diabetologia*, 2011.

[6] M. Cardellini, R. Menghini, M. Federici, et al. Decreased IRS2 and TIMP3 expression in monocytes from offspring of type 2 diabetic patients is correlated with insulin resistance and increased intima media thickness. *Diabetes*, 2011.

[7] C. Cathy, X. Zheng, BS. Weir, et al. Quality control and quality assurance in genotypic data for genome-wide association studies. *Genetic Epidemiology*, 2010.

[8] GA. Churchill and RW. Doerge. Empirical threshold values for quantitative trait mapping. *Genetics*, 1994.

[9] ME. Clarke, CA. Anderson, FH. Pettersson, et al. Basic statistical analysis in genetic case-control studies. *Nature Protocols*, 2011.

[10] International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 2001.

[11] The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature*, 2010.

[12] The International HapMap Consortium. A second generation human haplotype map of over 3.1 million snps. *Nature*, 2007.

[13] The International HapMap 3 Consortium. Integrating common and rare genetic variation in diverse human populations. *Nature*, 2010.

[14] The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 2007.

[15] N. Cossrow and B. Falkner. Race/ethnic issues in obesity and obesity-related comorbidities. *The Journal of Clinical Endocrinology and Metabolism*, 2004.

[16] C. Cotsapas, EK. Speliotes, I. Hatoum, MJ. Daly, et al. Common body mass index associated variants confer risk of extreme obesity. *Human molecular genetics*, 2009.

[17] B. Devlin and K. Roeder. Genomic control for association studies. *Biometrics*, Â1999.

[18] J. Dupuis, C. Langenberg, I. Prokopenko, et al. New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. *Nature Genetics*, 2010.

[19] RC. Elston, JM. Olson, and L. Palmer. *Biostatistical Genetics and Genetic Epidemiology*. Wiley, 2003.

[20] JL. Fleiss, B. Levin, and M. Cho Paik. *Statistical methods for rates and proportions*. Wiley, 2003.

[21] AS. Foulkes. *Applied statistical genetics with R*. Springer, 2009.

[22] B. Freidlin, G. Zheng, Z. Li, and JL. Gastwirth. Trend tests for case-control studies of genetic markers: Power, sample size and robustness. *Human heredity*, 2002.

[23] X. Gao. Multiple testing corrections for imputed SNPs. *Genetic Epidemiology*, 2011.

[24] X. Gao, LC. Becker, DM. Becker, et al. Avoiding the high Bonferroni penalty in genome-wide association studies. *Genetic Epidemiology*, 2010.

[25] X. Gao, J. Starmer, and ER. Martin. A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms. *Genetic epidemiology*, 2008.

[26] RL. Hanson, C. Bogardus, D. Duggan, et al. A search for variants associated with young- onset type 2 diabetes in American Indians in a 100K genotyping array. *Diabetes*, 2007.

[27] MG. Hayes, A. Pluzhnikov, K. Miyake, et al. Identification of type 2 diabetes genes in mexican americans through genome-wide association studies. *Diabetes*, 2007.

[28] A. Helgason, B. Yngvadttir, B. Hrafnkelsson, et al. An icelandic example of the impact of population structure on association studies. *Nature Genetics*, 2005.

[29] LA. Hindorff, P. Sethupathy, TA. Manolio, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *PNAS*, 2009.

[30] CJ. Hoggart, TG. Clark, DJ. Balding, et al. Genome-wide significance for dense snp and resequencing data. *Genetic Epidemiology*, 2008.

[31] Asociación internacional de tradructores y redactores de medicina y ciencias afines. Tremédica. http://www.medtrad.org/glosarios/bio_molecular/Glosario/G.html, 2005.

[32] M. Jobling, M. Hurles, and C. Tyler-Smith. *Human evolutionary genetics: origins, peoples & disease*. Garland Science, 2004.

[33] R. Johnson and W. Wichern. *Applied multivariate statistical analysis*. Pearson Prentice Hall, 2007.

[34] HM. Kang, JH. Sul, E. Eskin, et al. Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics*, 2010.

[35] HM. Kang, NA. Zaitlen, E. Eskin, et al. Efficient control of population structure in model organism association mapping. *Genetics*, 2008.

[36] S. Kobes, C. Bogardus, LJ. Baier, et al. PCLO variants are nominally associated with early-onset type 2 diabetes and insulin resistance in Pima indians. *Diabetes*, 2008.

[37] ES. Lander and NJ. Schork. Genetic dissection of complex traits. *Science*, 1994.

[38] CC. Laurie, KF. Doheny, DB. Mirel, et al. Quality control and quality assurance in genotypic data for genome-wide association studies. *Genetic epidemiology*, 2010.

[39] X. Li, KL. Monda, HH. Goring, et al. Genome-wide linkage scan for plasma high density lipoprotein cholesterol, apolipoprotein A-1 and triglyceride variation among American Indian populations: The Strong Heart Family Study. *Journal of Medical Genetics*, 2009.

[40] C. Lorenzo, M. Serrano-Rios, MT. Martinez-Larrad, et al. Was the historic contribution of spain to the mexican gene pool partially responsible for the higher prevalence of type 2 diabetes in mexican-origin populations? the spanish insulin resistance study group, the san antonio heart study, and the mexico city diabetes study. *Diabetes Care*, 2001.

[41] MI. McCarthy, GR. Abecasis, JN. Hirschhorn, et al. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Review Genetics*, 2008.

[42] W. Mendehall and D. Wackerly. *Estadística matemática con aplicaciones*. CENGAGE Learning, 2010.

[43] RH. Myers, DC. Montgomery, and G. Vining. *Generalized Linear Models: With Applications in Engineering and the Sciences*. Wiley, 2002.

[44] M Nachmana and S. Crowell. Estimate of the mutation rate per nucleotide in humans. *Genetics*, 2000.

[45] DL. Newman, M. Abney, MS. McPeck, et al. The importance of genealogy in determining genetic association studies. *American journal of human genetics*, 2001.

[46] AD. Paterson, D. Waggott, AP. Boright, et al. A genome-wide association study identifies a novel major locus for glycemic control in type 1 diabetes, as measured by both A1C and glucose. *Diabetes*, 2010.

[47] A. Price, R. Patterson, M. Plenge, N. Weinblatt, et al. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, Â2006.

[48] AL. Price, D. Reich, N. Patterson, et al. New approaches to population stratification in genome-wide association studies. *Nature Reviews. Genetics*, 2010.

[49] JK. Pritchard and NA. Rosenberg. Use of unlinked genetic markers to detect population stratification in association studies. *American journal of human genetics*, 1999.

[50] JK. Pritchard, M. Stephens, P. Donelly, et al. Association mapping in structured populations. *American journal of human genetics*, 2000.

[51] International HapMap Project. http://hapmap.ncbi.nlm.nih.gov/, 2005.

[52] S. Purcell, B. Neale, K. Todd-Brown, et al. PLINK: a tool set for whole-genome association and population-based linkage analysis. *American journal of human genetics*, 2007.

[53] D. Reich, AL. Price, and N. Patterson. Principal component analysis of genetic data. *Nature Genetics*, 2008.

[54] R. Robinson. *Genetics*. The Macmillan Science Library, 2003.

[55] JG. Smith, JK. Lowe, S. Kovvali, et al. Genome-wide association study of electrocardiographic conduction measures in an isolated founder population: Kosrae. *Heart Rhythm*, 2009.

[56] T. Strachan and P. Andrew. *Human Molecular Genetics*. Wiley-Liss, 1999.

[57] TM. Teslovich, K. Musunuru, AV. Smith, et al. Biological, clinical and population relevance of 95 loci for blood lipids. *Nature*, 2010.

[58] M. Traurig, J. Mack, RL. Hanson, M. Ghoussaini, et al. Common variation in SIM1 is reproducibly associated with BMI in Pima Indians. *Diabetes*, 2009.

[59] BF. Voight and JK. Pritchard. Confounding from cryptic relatedness in case-control association studies. *PLoS Genetics*, 2005.

[60] ME. Weale. *Quality Control for Genome-Wide Association Studies*. Genetic Variation: Methods in Molecular Biology. Springer, 2010.

[61] J. Yu, G. Pressoir, W. Briggs, et al. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics*, Â2005.

[62] A. Ziegler and Konig. *A statistical approach to genetic epidemiology*. Wiley-VCH, 2006.

[63] A. Ziegler, IR. Konig, and JR. Thompson. Biostatistical aspects of genome-wide association studies. *Biometrical journal. Biometrische Zeitschrift*, 2008.

Glosario

ADN
: Ácido desoxirribonucleico que ocupa un rol centra en la célula ya que porta la infromación genética necesaria para el desarrollo y funcionamiento de todos los organismos vivos conocidos.

alelo
: Representa una de las distintas alternativas de un gen, causadas por diferencias en la secuencia de ADN.

alelo menor
: Se define como el alelo menos común en una población.

control genómico
: Es definido como la mediana de los estadísticos de prueba de los SNPs dividido entre la median de la distribución nula.

desequilibrio de ligamiento
: Asociación entre los alelos presentes en dos sitios en un genoma.

equilibrio de Hardy-Weinberg
: Postula que las frecuencias alélicas en una población permanecen constantes a lo largo del tiempo en ausencia de fuerzas que las cambien.

estratificación poblacional
: Es la presencia de una diferencia sistemática en las frecuencias alélicas entre las subpoblaciones de una población, posiblemente debido a las diferente ascendencias de dichas subpoblaciones.

fenotipo
: Son el conjunto de características observables de un individuo resultantes de la interacción de su composición genética con el ambiente.

gen
: Son unidades funcionales de ADN que contienen las instrucciones necesarias para crear proteínas o ARN.

genoma
El total de material genético de una célula u organismo.

genotipificar
Significa caracterizar el genotipo (la constitución genética) de un organismo, en uno o más locus y por medios diversos (genéticos, moleculares, inmunológicos, etc.), utilizando células, tejidos u organismos enteros.

genotipo
Es el par de bases de ADN observado en un lugar del genoma. Normalmente representado una variable categórica que toma valores de un conjunto predefinido de carácteres.

GWAS
Estudio de asociacion de genoma completo (GWAS, por sus siglas en ingés) es la investigación de asociación entre fenotipo-genotipo que implica la caracterización de genotipos a lo largo de todo el genoma.

haplotipo
Combinación específica de alelo que estan alineadas en sólo un cromosoma de cada par homólogo.

hit
Es el SNP con mayor significancia estadística en una región.

IBD
Dos o más alelos son idénticos por descendencia o IBD (por sus siglas en ingés *identical-by-descent*) si son copias idénticas de mismo alelo ancestral.

IBS
Dos o más alelos son idénticos por estado o IBS (por sus siglas en ingés *identical-by-state*) si tinen la misma composición de ADN pero no necesariamente provienen del mismo ancestro.

LD
Abreviación de desequilibrio de ligamiento por sus siglas en inglés.

loci
Plural de locus.

locus
Segmento específico de ADN en una posición particular del cromosoma, cuyo plural es loci.

MAF
Significa, por sus siglas en inglés, frecuencia de alelo menor.

marcador genético
Es un segmento de ADN con una ubicación física identificable en un cromosoma y cuya herencia genética se puede rastrear.

microarreglo de ADN
Es una colección de puntos microscópicos de ADN posicionados en una superficie sólida. Se usan para medir simúltaneamente los niveles de expresión de un gran número de genes o para determinar el genotipo en múltiples regiones del genoma. También se conoce como chip de ADN.

mutación
Es un cambio hereditario en el genoma de un organismo.

OR
La razón de momios o OR, por sus siglas en ingés, es una medida de asociación común. Se define como la razón de la probabilidad de los individuos de desarrollar la enfermedad si están expuestos a cierto factor sobre la probabilidad de que desarrollen la enfermedad individuos que no están expuestos.

PCA
Análisis de componentes principales por sus siglas en inglés.

polimorfismo
Una variación en la secuencia de ADN que tiene frecuencia de al menos 1 % en al menos una población humana.

rasgo
Una variante fenotípica de un organismo que puede ser heredable, determinada por el ambiente o una combinación de ambas.

rasgo complejo
Son aquellos que son influenciados por mas de un factor. Los factores pueden ser genéticos o ambientales.

relación críptica
se presenta cuando dos o más individuos en un estudio tienen una relación parental oculta ya sea no reportada o desconocida, por ejemplo primos segundos.

SNP
Polifomorfismo de un solo nucleótido o SNP (Single Nucleotide Polymorphism, por sus siglas en inglés) es una variación en la secuencia de ADN que sucede cuando una sola base es sustituida por otra.

valor de p
Es el nivel más pequeño de significancia α para el cual la información observada indica que la hipótesis nula debe de ser rechazada.