Casa abierta al tiempo

UNIVERSIDAD AUTÓNOMA METROPOLITANA

Unidad Iztapalapa

DIVISIÓN DE CIENCIAS BÁSICAS E INGENIERÍA

MAESTRÍA EN CIENCIAS MATEMÁTICAS APLICADAS E INDUSTRIALES

Estudios familiares en epidemiología genética: diseño y análisis

Tesis que presenta:

Adriana Arely Regalado Rodríguez

Para obtener el grado de:

Maestra en Ciencias (Matemáticas Aplicadas e Industriales)

Dirigida por:

Dra. Hortensia Moreno Macias

Resumen

Cuando se realizan estudios de asociación genética basados en individuos independientes, considerando una población mestiza como la mexicana, se debe realizar un ajuste por estratificación poblacional. Los diseños familiares constituyen una alternativa para este problema. Es por ello que en el presente trabajo se analizan y comparan, en términos de ventajas y desventajas, los métodos estadísticos basados en familias para el análisis de asociación genética con enfermedades complejas.

Índice general

Introducción

La epidemiología genética se encarga de estudiar la interacción entre los factores genéticos y ambientales que originan las enfermedades del ser humano (Beaty et al., 1993). Poder identificar los factores involucrados en el desarrollo de enfermedades es de gran importancia para establecer políticas de prevención, tratamientos adecuados y mejorar la estrategia de diagnósticos.

Las personas compartimos el 99.5 % del genoma humano, eso nos hace biológicamente seres humanos. El 0.05 % restante es lo que nos hace totalmente individuales y genera variabilidad tanto en nuestro aspecto físico como en la susceptibilidad de desarrollar enfermedades. Esta variabilidad se origina por una asignación genética biológicamente aleatoria que depende de las características genéticas de los padres y de un complejo proceso biológico en la formación de un nuevo ser humano.

Usualmente para conocer los factores genéticos que se involucran con la enfermedad se realizan estudios de genes candidatos los cuales parten de una hipótesis biológica que justifica la potencial asociación entre la variante genética y la enfermedad. A pesar de que la transmisión genética necesariamente involucra familias usualmente la asociación se evalúa utilizando información de individuos independientes como los diseños tipo casos y controles.

En una población mestiza, como lo es la mexicana, el análisis de asociación genética considerando sujetos independientes puede generar fácilmente confusión por estratificación poblacional debido a la heterogeneidad étnica de la población. Los diseños familiares son robustos ante esta estratificación, (Evangelou et al., 2006).

Por lo anterior, el objetivo principal de este trabajo es presentar y analizar algunos métodos estadísticos aplicados a la epidemiología genética considerando diseños familiares. Nos daremos cuenta de la importancia que la estadística tiene en los métodos para evaluar asociación genética en estudios que consideras diseños familiares.

Estructura

Comenzamos en el capítulo 1 definiendo algunos conceptos genéticos básicos que nos ayudarán a describir y comprender la aleatoriedad y variabilidad involucrada en la biología de la herencia genética y así entender el papel que la estadística desempeña en el análisis de asociación genética.

Posteriormente, en el capítulo 2, presentamos tres diseños familiares (Tríos caso-padres, parejas de hermanos y familias extendidas) ya documentados y para los cuales en el capítulo 3

presentamos el desarrollo y análisis de las diferentes pruebas de asociación. El riesgo relativo del haplotipo, la prueba de desequilibrio de transmisión y sus diferentes variantes se presentan con detalle. Al final del capítulo 3 se discute el enfoque del modelo mixto y la forma en la que se incorporan los factores ambientales y se estudia un fenotipo cuantitativo.

En el capítulo 4 presentamos el análisis de asociación de 32 genes candidatos con el fenotipo de labio leporino y paladar hendido tomando información de 27 familias mexicanas, considerando algunos modelos descritos en el capítulo anterior. Finalmente, presentamos una conclusión sobre el papel relevante de la estadística en el análisis de asociación genética cuando se usan diseños familiares.

Capítulo 1

Conceptos básicos de genética

La genética es la ciencia que estudia la transmisión de la información hereditaria de una generación a la siguiente, su objetivo de estudio son los genes, los cuales pueden abordarse desde distintas perspectivas, molecular, bioquímica, celular, organísmica, familiar, poblacional o evolutiva.

El estudio de los organismos comienza con la observación de las características físicas que identifican a un ser vivo. Estas características se conocen como **fenotipos**, algunos ejemplos son: el color de las plumas de un ave, la forma de la cara de una persona, o el tamaño del caparazón de una tortuga. Formalmente el fenotipo se define como la expresión de la información de un genotipo determinado, en relación con el ambiente en el cual el organismo se desarrolla. Por su parte, el **genotipo** es la construcción genética de un ser vivo, representada por los genes que posee como miembro de una especie particular, por ejemplo, los seres humanos estamos constituidos por 23 pares de cromosomas en cada célula somática, 20,000 genes aproximadamente en el genoma, y 3 mil millones de pares de bases nitrogenadas. Esos pares de bases se agrupan en regiones dentro de los genes llamadas exones e intrones. Entonces, para entender la construcción y variabilidad genética del ser humano es necesario definir y explicar como se componen los cromosomas y los genes.

Los **cromosomas** son las unidades más condensadas del genoma de un organismo que se encuentran en el núcleo de todas las células del organismo, se presentan en pares, y cada miembro del par es heredado uno por el padre y otro por la madre. En la década de los 50's se realizó el cálculo correcto del número de cromosomas en la célula humana, anteriormente se pensaba que habían 48 cromosomas, pero en 1956 se determinó que en el núcleo de cada célula del cuerpo humano se encuentran 46 cromosomas (exceptuando las celulas reproductoras que solo tienen la mitad). Ahora se sabe que los humanos tenemos 23 pares de cromosomas (22 son autosomas y un par de cromosomas sexuales).

En el año 1905, Johannsen introdujo por primera vez el término **gen**, el cual definió como unidad de almacenamiento de información y unidad de herencia al transmitir la información. Todos los genes componen el genoma de un organismo, por lo tanto, estarán distribuidos entre los cromosomas de los cuales esté compuesto dicho organismo, correspondiéndole una posición precisa. Posteriormente en 1944, Oswald Avery mostró que un gen es un segmento corto de **DNA (ácido desoxirribonucléico)**, el cual poseé una secuencia específica de pares de bases

nitrogenadas, la cual codifica para una proteína específica. En el año de 1954, y gracias al trabajo de Rosalind Franklin, fue que J. Watson y F. Crick describen la estructura física del DNA, como una cadena de doble hélice. El DNA es el componente químico primario de los cromosomas y material en el que se encuentran la secuencia de todos los genes, es decir, el genoma, y por lo tanto es el portador de la información hereditaria y dicha información se encuentra codificada en la secuencia de bases nitrogenadas. En la Figura 1.1 se muestran los distintos niveles estructurales en los cuales se presenta el material genético en una célula eucariota[1], las cuales en conjunto originan organismos eucariontes, por ejemplo el del ser humano.

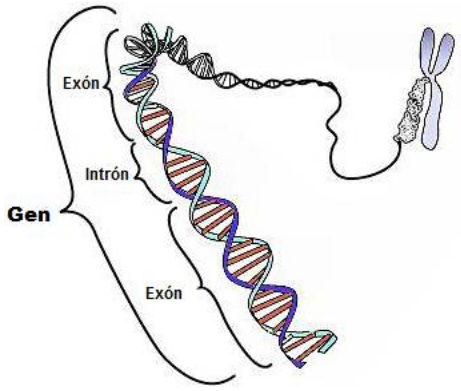


Figura 1.1: Composición del cromosoma
Fuente: https://es.wikipedia.org/wiki/Gen

En el diagrama esquemático de la figura 1.1 se muestran la estructura del gen y las regiones llamadas exones e intrones. Los exones tienen una actividad directa en la formación de proteínas (codifica a proteínas) no así los intrones.

Hasta ahora conocemos la estructura básica del organismo del ser humano, pero todavía no explicamos por que los seres humanos somos distintos entre nosotros. Para entender la variación genética, introduciremos algunos términos que nos serán útiles. A la posición fija del cromosoma que define la ubicación de un gen (o genes) se le conoce como **locus**(**loci**, para el plural). A cada versión diferente de un gen para la misma proteina, se le denominan **alelos**, y se clasifican en dominantes y recesivos[2]. El par de alelos que se encuentran en un determinado locus de un par de cromosomas homólogos definen el **genotipo** en ese locus. Si los alelos correspondientes a un genotipo son iguales se le denomina **homocigoto**, y si son diferentes **heterocigoto**, en Figura 1.2 se ilustran estos términos.

En el siguiente apartado discutiremos las fuentes que originan la variabilidad genética y sus consecuencias.

[1]Son aquellas células que tienen núcleo. Ejemplos de organismos eucariotas son: plantas, animales algas, hongos, entre otros.

[2]Si los alelos son diferentes, el alelo dominante se expresa, mientras que el efecto del otro alelo, denominado recesivo, queda enmascarado

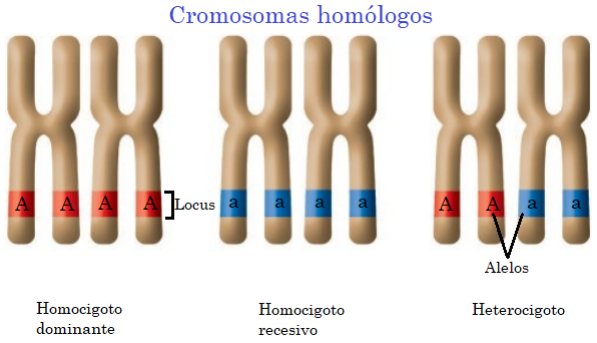


Figura 1.2: Genotipos homocigotos y heterocigotos
Fuente: www.biodiversidad472308521.wordpress.com/category/herencia/

1.1. Variación genética

La primera fuente de variabilidad genética ocurre durante la generación de las células germinales de los padres, a este proceso se le conoce como meiosis. La meiosis es un proceso de división celular a través del cual a partir de una célula diploide (un juego completo de pares de cromosomas homólogos; siendo en cada par heredado, un cromosoma paterno y el otro materno) se producen cuatro células haploides (gametogénesis), este proceso se ejemplifica en la figura 1.3. Las células haploides son aquellas que contienen un solo juego de cromosomas. Así, pues, el objetivo de la meiosis es generar células sexuales: ovocitos (gametos femeninos) en la ovogénesis y espermatozoides (gametos masculinos) en la espermatogénesis. También la información genética de los espermatozoides y óvulos creados es variable debido a la combinación aleatoria de los 23 cromosomas. Además, durante el proceso de la gametogénesis, se lleva a cabo un evento de suma importancia para la variabilidad genética: el entrecruzamiento del material genético entre las cromátidas hermanas de cada par de los cromosomas homólogos. Este proceso explica por que dos hermanos no son fisicamente iguales (a menos que sean gemelos homocigotos)

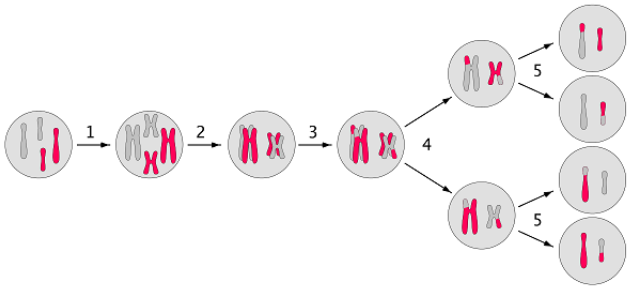


Figura 1.3: Del DNA a las proteínas
Fuente: Elaboración propia

Las leyes de Mendel son adecuadas para entender un poco mejor este tipo de variabilidad.

Mendel realizó cruces sobre distintos tipos de plantas de chícharo y observó la transmisión de rasgos físicos que se distinguían fácilmente, como el color de las semillas y las flores de chícharo. Basado en sus experimentos, Mendel observó ciertas regularidades que ahora se conocen como las *leyes de Mendel:*

1. *Ley de la uniformidad:* Ésta ley explica que hay alelos que dominan sobre otros, pero que la dominancia no incide en la transmisión, es decir, si un alelo es dominado por otro, este alelo también se puede heredar.

2. *Ley de la segregación:* Durante la formación de los genotipos, cada alelo de un par, es heredado uno por la madre y otro por el padre.

3. *Ley de la distribución independiente:* Diferentes rasgos son heredados independientemente unos de otros, no existe relación entre ellos, por tanto el patrón de herencia de un rasgo no afectará al patrón de herencia de otro.

En la ley de distribución de independencia Mendel supone que la herencia de cada variante genética o locus era independiente, pero el entrecruzamiento no cumple con esta característica, es te proceso los locus que físicamente están muy cerca no se separan, es decir, se heredan juntos por lo que no serían independientes.

Otra fuente de variabilidad surge durante la división celular, en el proceso de transcipción traducción. el cual inicia considerando la información genética del DNA que es utilizada, por medio de la **transcripción**, para generar el RNA mensajero (mRNA); posteriormente, mediante el proceso de **traslación**, el mRNA se decodifica y da origen a la proteína. El diagrama de este proceso se muestra en la Figura 1.4.

Con esto podemos entender que si ocurre un cambio en el DNA entonces las proteínas en el cuerpo podrían cambiar y generar un mal funcionamiento en el organismo.

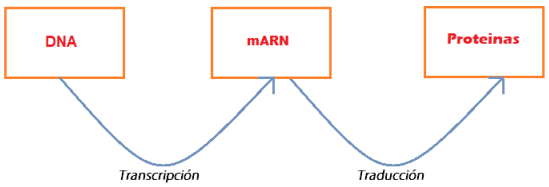


Figura 1.4: Del DNA a las proteínas
Fuente: Elaboración propia

Durante ese proceso se pueden dar cambios en la secuencia de neuclótidos del DNA, y por lo tanto en la secuencias protéicas. El cambio en estas secuencia puede o no causar un cambio en el fenotipo. Usualmente, cuando la variación ocurre en menos del 1 % de la población, se le denomina **mutación**, si la variación genética ocurre en un locus determinado, y se presenta en más del 1 % de la población, se le conoce como **polimorfismo**. Existen varios tipos de polimorfismos pero nosotros estaremos interesados en los de un solo neuclótido (**SNP**, por sus siglas en inglés), ya que constituyen hasta el 90 % de las variaciones genómicas humanas. Estos polimorfismos se caracterizan por afectar solo a una base nitrogenada del DNA. Estas definiciones pueden ser obsoletas y confusas (Karki et al., 2015), si entendemos como mutación a cualquier variación en la secuencia del ADN, tendremos que diferenciar dos, las mutaciones que se pueden heredar de los padres (**mutaciones de la linea germinal**), este tipo de mutaciones ocurren en los gametos. Dado que la descendencia se deriva inicialmente de la fusión de un óvulo y un espermatozoide, también se pueden encontrar mutaciones en la línea germinal de los padres en cada célula nucleada de su progenie. Otro tipo de mutaciones son las que se adquieren durante

la vida de la persona (**mutaciones somáticas**), siendo estas el principal conductor de enfermedades como el cáncer (Karki et al., 2015).

Exites enfermedades que son causadas únicamente por el cambio de un neuclótido. La anemia falciforme es un ejemplo de ello, pues se origina por un cambio de neuclótido en un gen que codifica la beta cadena de la proteína de hemoglobina. A este tipo de enfermedades se les conoce como **monogénicas** o mendelianas. Sin embargo hay enfermedades que involucran diferentes genes, y en las que puede influir los factores ambientales, a estas enfermedades poligénicas o multifactoriales se les conoce como **complejas**

El estudio de las enfermedades complejas desde el punto de vista genético es de gran importancia y ha aumentado en los últimos años, por lo que el objetivo de la tesis es dar a conocer las ventajas y desventajas de distintos métodos estadísticos aplicados a la epidemiología genética basada en estudios familiares que ayuden al análisis de la asociación genética con enfermedades complejas.

Capítulo 2

Diseños Familiares

Cuando hablamos de asociación genética a enfermedades complejas deseamos evaluar la frecuencia alélica entre la población que es afectada por determinada enfermedad, a la que se le denomina **casos**, y población no afectada, la cual se conoce como **controles**. Si después de ajustar por confusores, un SNP es más frecuente entre los casos que entre los controles, decimos que el SNP está asociado a la enfermedad.

Los estudios de asociación genética se dividen principalmente en dos, aquellos que incluyen personas sin parentesco es decir, independientes y los que toman en cuenta casos índices y controles consanguíneos, a los últimos se les denomina **estudios de asociación genética basados en familias** (EAGBF).

Ente los diseños de investigación epidemiológica más frecuentes que no consideran parentesco se encuentran: casos y controles, de cohorte y transversales. Estos diseños suelen tener mayor eficiencia estadística comparados con los que consideran familias, suelen ser menos costosos y permiten examinar asociaciones de múltiples genes con la enfermedad de interés, sin embargo, también presentan algunas desventajas, por ejemplo, son susceptibles a confusión por estratificación poblacional (Flores-Alfaro et al., 2012), esto ocurre cuando las tasas de ocurrencia de la enfermedad son diferentes por grupos étnicos,los controles difieren étnicamente de los casos, o las frecuencias alélicas varían entre etnias o razas. Esto sucede en poblaciones de mezcla reciente entre dos o más grupos étnicos que originalmente estaban separados, las poblaciones latino-americanas son de mezcla reciente.

Los EAGBF corrigen los problemas que se mencionan anteriormente de los diseños que no consideran parentesco, además aportan mayor información sobre enfermedades hereditarias y la prevalencia de la enfermedad en subgrupos de familias. Los diseños familiares más frecuentes son: tríos casos-padres, parejas de hermanos y finalmente familias extendidas, en las siguientes secciones analizaremos y ejemplificaremos cada uno de ellos.

El objetivo de este capítulo es únicamente describir la construcción de los diseños familiares más populares. En el capítulo posterior hablaremos de los métodos estadísticos empleados para el análisis de asociación considerando los diseños familiares.

2.1. Tríos Casos-Padres

El diseño tríos casos-padres (TCP) comienzan con un individuo afectado (caso) y el reclutamiento de sus padres, los cuales pueden estar afectados o no por la enfermedad. En éste diseño se comparan los alelos transmitidos de los padres a hijos contra los alelos no transmitidos. Por ejemplo, consideremos que en una familia el padre tiene genotipo (a,b) y la madre (c,d). Asumamos que el hijo recibe el genotipo (a,c), al par de alelos que conforman el genotipo se les conoce como alelos transmitidos, existen otros tres genotipos que el hijo pudo haber recibido, (a,d), (b,c) y (b,d), a éstos se les denomina genotipos no transmitidos. En la Figura 2.1 se representa una familia participante en un estudio TCP. A esta representación gráfica se le conoce como diagrama familiar o genograma, el cuadro representar a un miembro masculino, el circulo que se trata de un miembro femenino y un marco grueso índica el caso.

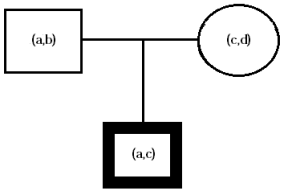


Figura 2.1: Ejemplo de diseño trío caso-padres.
Fuente: Elaboración propia

El diseño TCP examina numerosos tríos para evaluar si un alelo específico o una combinación de alelos es preferentemente transmitida a los casos, lo que sugiere una asociación entre el alelo correspondiente y la enfermedad.

Este diseño ataca el problema de estratificación poblacional y es ideal para analizar enfermedades de diagnóstico temprano. La principal desventaja de este diseño es que muchas veces no se cuenta con información de ambos padres, principalmente cuando se están analizando enfermedades de diagnóstico tardío (Hatemi et al., 2010).

2.2. Parejas de hermanos

En este diseño, cada caso es comparado con un hermano o más no afectados, los cuales tomaran el rol de controles, en la Figura 2.2 se ejemplifica una pareja participante en el estudio. En general los controles elegibles deben ser aquellos hermanos que alcancen la edad correcta para ser diagnosticados. Si se está estudiando los casos incidentes lo más probable es que los controles sean hermanos mayores. Es importante considerar que la diferencia de edad entre hermanos mayores y menores, podría llevar a exposiciones ambientales dependientes del tiempo o la tendencia (es decir, hermanos de diferentes edades pueden ser sujetos a diferentes variables de exposición que dependen del tiempo).

Este diseño, puede controlar múltiples variables de confusión, tanto ambientales como genéticas. Sin embargo esta ventaja puede resultar costosa, ya que al considerar más variables se puede perder poder estadístico. En algunos ejemplos, el uso de hermanos como controles requiere hasta el doble de muestra para mantener el poder de la prueba (Flores-Alfaro et al., 2012), sin embargo este estudio ayuda a determinar la asociación en enfermedades de diagnóstico tardío,

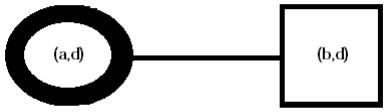


Figura 2.2: Ejemplo de diseño parejas de hermanos.
Fuente: Elaboración propia

como es el caso de algunos tipos de cáncer. Además, cuando se cuenta con la información de hermanos gemelos monocigótos y dicigótos , el diseño es muy útil para evaluar la *heredabilidad* de un rasgo o enfermedad, que es la proporción de variación observada que puede atribuirse a factores genéticos heredados (Merriam-Webster).

De acuerdo con Hatemi et al. (2010), una clara desventaja de este diseño es que no todos los casos cuentan con un hermano elegible para el estudio, en esta situación se podría asociar el caso con un primo hermano no afectado, este camino podría proporcionar una muestra más grande para el estudio sin embargo no proporciona la mejor solución para el problema de la estratificación poblacional, ya que el caso y el control estarían relacionados únicamente por un solo padre, más aún se puede perder el control de algunas variables de confusión.

En éste diseño se comparan las frecuencias, de los posibles alelos de susceptibilidad, entre hermanos afectados y no afectados.

2.3. Familias extendidas

El diseño de familias extendidas considera la mayor cantidad de familia disponible del caso. Este diseño es el más complicado de realizar ya que se requiere la colaboración de bastantes familiares y generaciones anteriores del caso índice, sin embargo ofrece mayor información genética y ambiental lo que propicia a una mejor medición del riesgo contraer la enfermedad en estudio. En la Figura 2.3 se muestra un ejemplo de una familia extendida.

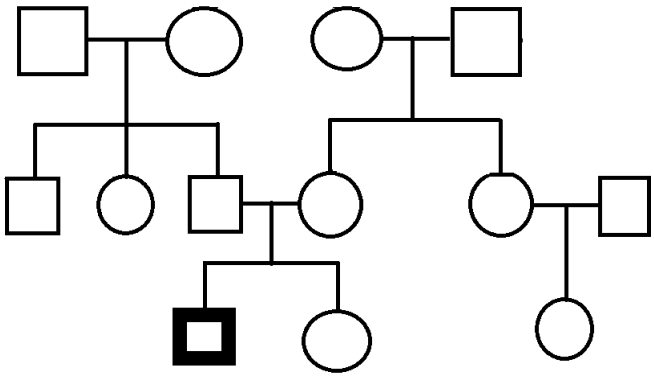


Figura 2.3: Ejemplo de familia extendida.
Fuente: Elaboración propia

Los análisis que se realizan utilizando diseños de familias extendidas suelen ser más ricos, ya que además de estudiar asociación genética, se puede estudiar le heredabilidad. Más aún, ocupando un diseño con información completa de la familia y definiendo de manera clara el probando y su relación con el resto de la familia, las pruebas realizadas suelen ser más poderosas (Park et al., 2015). Sin embargo, estos análisis suelen ser más complejos y costosos. Entre los métodos estadísticos que se pueden ocupar para realizar estos análisis, se encuentran las pruebas de hipótesis no paramétricas y los modelos de efectos mixtos.

La desventaja más clara de este diseño es que rara vez se cuenta con la información completa por generaciones familiares.

Capítulo 3

Pruebas de asociación basadas en familias

Para analizar la asociación de un alelo en un locus genético con un fenotipo se pueden realizar diferentes diseños de estudio. En el capítulo anterior mencionamos que la manera mas sencilla de comenzar con el análisis es considerando casos y controles y discutimos las ventajas y desventajas que este diseño proporciona. En esta sección examinaremos algunos métodos estadísticos para el análisis de asociación considerando familias.

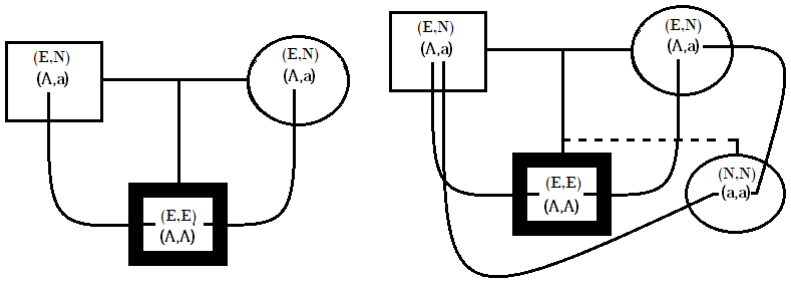
En la primera parte del capítulo consideraremos únicamente fenotipos dicotómicos (afectado o no afectado por la enfermedad). El primer diseño que analizamos es el de Tríos Casos-Padres (TCP) y comenzaremos describiendo la prueba de Riesgo Relativo del Haplotipo (HRR) abordando sus desventajas, las cuales se pueden solucionar con la construcción de la clásica prueba de desequilibrio de transmisión (TDT, por sus siglas en inglés). Después examinaremos algunas variaciones del TDT que consideran otros diseños familiares, como parejas de padre (o madre) e hijo, o conjuntos de hermanos. Posteriormente estudiaremos la construcción de una prueba de hipótesis para familias extendidas (PDT). Finalizaremos analizando una metodología basada en modelos lineales mixtos, la cual nos permitirá considerar tanto rasgos cualitativos como cuantitativos y ajuste por covariables.

3.1. Método del riesgo relativo del haplotipo (HRR)

Entenderemos por haplotipo, al conjunto de alelos simples estrechamente relacionados que tienden a heredarse juntos (NAL, 2013). En general, los loci que forman el haplotipo se localizan muy cerca el uno de otro. Se dice que están en desequilibrio de ligamiento o que "están ligados" si al heredarse, se heredan juntos rompiendo la ley de independencia de Mendel. El estudio considerando haplotipos es importante, ya que en el análisis de asociación podemos encontrar alguna variante significativa, sin embargo no sabremos si esa variante es la causal del fenotipo o las que están ligadas a ella ya que se heredaron juntas. Esto es de gran ayuda ya que nos permite acotar las regiones del gen que pueden causar la enfermedad.

El método HRR considera un haplotipo formado por dos loci: el causal de la enfermedad, con alelos (E, N) (donde N representa un alelo normal y E el alelo de la enfermedad), y el marcador, el cual nos interesa analizar si se encuentra asociado con la enfermedad, con alelos (A, a) además

de una familia TCP donde el hijo ha sido afectado y ambos padres son heterocigotos en ambos loci. Con ello podemos saber qué alelos fueron transmitidos al hijo y cuales no, como se muestra en la Figura 3.1a. La idea básica de este método es construir un "pseudo-control" con los alelos que no fueron transmitidos , como en la Figura 3.1b. Haciendo esta construcción, los pares de alelos no transmitidos representan los alelos de la población que no tienen la enfermedad, por lo que la construcción es válida sólo si la muestra aleatoria de los alelos no transmitidos es representativa. Esto requiere que no haya endogamia entre padres, no correlación entre los fenotipos de los padres y considerar un único diseño TCP por familia (Ziegler et al., 2010).



(a) Familia TCP con padres heterocigotos. Es posible reconocer que alelos se transmitieron al hijo.

(b) Construcción del pseudocontrol.

Figura 3.1: Familia TCP sin y con pseudo-control
fuente: Ziegler et al. (2010), pp.320

Utilizando la información de estos pseudo-controles se puede construir una tabla de contingencia 2x2 de las frecuencias alélicas observadas en el marcador genético di alélico para los casos y pseudo-controles (Cuadro 3.1).

	A	a	Total
Alelos transmitidos (Casos)	n_{TA}	n_{Ta}	n_T
Alelos no transmitidos (Pseudo-controles)	n_{NA}	n_{Na}	n_N
Total	n_A	n_a	$4n$

Cuadro 3.1: Conteo de alelos A y a observados en casos y pseudo-controles.

Gracias a la construcción de esta tabla de contingencia, se puede realizar una prueba de asociación utilizando el estadístico de prueba clásico χ^2_c. El cual, en esencia, probará si existe dependencia estadística entre los alelos y la clasificación de la población (casos y pseudo-controles).

$$\chi^2_c = \sum_{k=1}^{4} \frac{(o_k - e_k)^2}{e_k}$$
$$= \frac{(4n \cdot n_{TA} - n_T n_A)^2}{4n \cdot n_T n_A} + \frac{(4n \cdot n_{Ta} - n_T n_a)^2}{4n \cdot n_T n_a} + \frac{(4n \cdot n_{NA} - n_N n_A)^2}{4n \cdot n_N n_A} + \frac{(4n \cdot n_{Na} - n_N n_a)^2}{4n \cdot n_N n_a}$$

Donde o_k denota la frecuencia observada y e_k la frecuencia esperada.

El estadístico χ_c^2 no refleja la fuerza de asociación entre los alelos transmitidos a los casos y los transmitidos a los pseudo-controles, como lo pudiese hacer la razón de momios (*odds ratio*) $(\frac{n_{TA}n_{Na}}{n_{Ta}n_{NA}})$ o el determinante de la matriz de contingencia $(n_{TA}n_{Na} - n_{Ta}n_{NA})$. Estos indicadores muestran con mayor claridad el grado de asociación, pues de no haber es de esperarse que la razón de momios se aproxime a uno, o equivalentemente, el determinante se aproxime a cero. Es decir, la diferencia entre, la cantidad de alelos **A** transmitidos y alelos **a** no transmitidos, y la cantidad de alelos **a** transmitidos y alelos **A** no transmitidos a los casos, debe ser pequeña. Por lo que un estadístico más representativo es:

$$\chi^2_{HRR} = \frac{4n(n_{TA}n_{Na} - n_{Ta}n_{NA})^2}{n_T n_N n_A n_a}.$$

A continuación mostraremos que el estadístico χ^2_{HRR} es equivalente a χ_c^2.

Comenzaremos desarrollando $(o_1 - e_1)^2$ y ocupando que: $n_{TA} + n_{Na} + n_{Ta} + n_{NA} = 4n$.

$$\begin{aligned}
(o_1 - e_1)^2 &= \left(n_{TA} - \frac{n_T n_A}{4n}\right)^2 \\
&= \left(\frac{4n \cdot n_{TA} - n_T n_A}{4n}\right)^2 \\
&= \left(\frac{n_{TA}(n_{TA} + n_{Na} + n_{Ta} + n_{NA}) - (n_{TA} + n_{Ta})(n_{TA} + n_{NA})}{4n}\right)^2 \\
&= \frac{(n_{TA}n_{Na} - n_{Ta}n_{NA})^2}{(4n)^2}.
\end{aligned}$$

Un procedimiento análogo muestra que:

$$(o_2 - e_2) = (o_3 - e_3) = (o_4 - e_4) = \frac{(n_{TA}n_{Na} - n_{Ta}n_{NA})^2}{(4n)^2}.$$

De lo anterior se sigue:

$$\chi_c^2 = \left(\frac{1}{e_1} + \frac{1}{e_2} + \frac{1}{e_3} + \frac{1}{e_4}\right)\left(\frac{n_{TA}n_{Na} - n_{Ta}n_{NA}}{4n}\right)^2. \tag{3.1}$$

Desarrollando el primer factor de (3.1) obtenemos:

$$\begin{aligned}
\frac{1}{e_1} + \frac{1}{e_2} + \frac{1}{e_3} + \frac{1}{e_4} &= \frac{4n}{n_T n_A} + \frac{4n}{n_T n_a} + \frac{4n}{n_N n_A} + \frac{4n}{n_N n_a} \\
&= 4n\frac{n_N n_a + n_N n_A + n_T n_a + n_T n_A}{n_T n_N n_A n_a}. \\
&= 4n\frac{(n_N + n_T)(n_a + n_A)}{n_T n_N n_A n_a} \\
&= \frac{(4n)^3}{n_T n_N n_A n_a}
\end{aligned}$$

Sustituyendo la última igualdad en (3.1) obtenemos que $\chi^2_c = \chi^2_{HRR}$.

La construcción del estadístico χ^2_{HRR} es sencilla e intuitiva, sin embargo se debe satisfacer que los alelos parentales sean independientes, los alelos no transmitidos representen a la población que no presenta la enfermedad, y que la frecuencia alélica en el locus investigado sea similar en todos los tríos (Ziegler et al., 2010). Para evitar estas desventajas se construyó la prueba de desequilibrio de transmisión (TDT, por sus siglas en inglés). A continuación describiremos su construcción y analizaremos sus ventajas.

3.2. Prueba de desequilibrio de transmisión (TDT)

El estadístico estándar para trabajar con diseños TCP es el TDT, el cual fue propuesto por Spielman et al. (1993). Para entender la construcción de esta prueba también consideraremos un haplotipo con las mismas características que en la anterior, esta prueba además de analizar la asociación, probará ligamiento entre los loci que estamos considerando.

También utilizaremos familias tipo TCP, sin importar si son o no heterocigotos, es por ello que este estadístico no utiliza la construcción de pseudo-controles, en su lugar evalúa si la proporción de alelos transmitidos, de padres heterocigotos a hijos afectados, se desvía del 50 % esperado bajo la frecuencia mendeliana asumiendo que no hay ligamiento entre el alelo marcador y el alelo que confiere la enfermedad.

Para entender las características del TDT, analizaremos su construcción como proponen Ziegler et al. (2010). Comenzaremos considerando un haplotipo como en la prueba anterior. El locus causal de la enfermedad con alelos N y E con frecuencias di alélicas $P(E) = p$ y $P(N) = 1 - p = \bar{p}$ y el marcador dialélico con alelos A y a, cuyas frecuencias alélicas están dadas por $P(A) = q$ y $P(a) = 1 - q = \bar{q}$. También asumiremos que los alelos A y E tienen un **desequilibrio de ligamiento positivo**, es decir, la diferencia entre la frecuencia de un haplotipo (AE en nuestro ejemplo) y la frecuencia que debería tener si estuviera en **equilibrio** es positiva, es decir $P(EA) - pq = D > 0$, consecuentemente se satisface que:

$$\begin{aligned} P(EA) &= pq + D & P(NA) &= \bar{p}q - D \\ P(Ea) &= p\bar{q} - D & P(Na) &= \bar{p}\bar{q} + D \end{aligned}$$

Primeramente exploraremos los posibles haplotipos parentales, en nuestro caso, estamos considerando 2 locus di alélicos, (E, N) y (A, a), por lo que existen 16 posibles haplotipos parentales distintos, los cuales se muestran en la primer columna del Cuadro 3.2. Si consideramos un haplotipo parental específico, solo existen 8 combinaciones posibles de los alelos heredados y no heredados, en el Cuadro 3.2 se muestran las probabilidades de transmisión y no transmisión de haplotipos considerando un haplotipo parental, donde θ indica la frecuencia del genotipo, es decir si $\theta = \frac{1}{2}$ se espera que no haya asociación entre el locus y la enfermedad, pues la frecuencia mendeliana esperada del 50 % prevalece.

Usando esta información, podemos calcular la probabilidad de cualquiera de las 8 posibles combinaciones de haplotipos transmitidos y alelos no transmitidos. Por ejemplo, si queremos calcular la probabilidad de que el alelo E sea transmitido junto con el alelos A y al mismo tiempo otro alelo A no sea transmitido, debemos considerar la probabilidad de que el haplotipo parental sea $\begin{smallmatrix} E \\ A \end{smallmatrix} \Big| \begin{smallmatrix} E \\ A \end{smallmatrix}$ es $(pq + D)^2$, y que la probabilidad de que un haplotipo sea $\begin{smallmatrix} E \\ A \end{smallmatrix}$ y otro $\begin{smallmatrix} N \\ A \end{smallmatrix}$

es $(pq+D)(\bar{p}q-D)$, así la probabilidad de transmitir (T) el haplotipo $\begin{smallmatrix}E\\A\end{smallmatrix}$ y no transmitir (NT) el alelo A es:

$$P(T=\ \begin{smallmatrix}E\\A\end{smallmatrix}\ , NT=A) = (pq+D)^2 + 2\frac{1}{2}(pq+D)(\bar{p}q-D) = pq^2 + Dq.$$

Análogamente se pueden calcular las probabilidades de transmitir un haplotipo y no transmitir determinado alelo. En el Cuadro 3.3 se muestran estas probabilidades, observemos que en la última columna se calculan las probabilidades marginales de los posibles haplotipos. Notemos que θ se presenta únicamente en los casos donde se tiene información de un padre heterocigoto en ambos locus.

En toda la construcción que hemos hecho hasta ahora, hemos ignorado el diseño familiar, recordemos que estamos considerando familias TCP, donde los hijos presentan la enfermedad, por lo que nos enfocaremos unicamente en los haplotipos transmitidos que incluyan al alelo E. Por esta razón, únicamente consideraremos los primeros dos renglones del Cuadro 3.3, para calcular las probabilidades condicionales de transmitir un alelo, del locus que estamos investigando, y no transmitir el otro alelo, dado que el hijo presenta la enfermedad. Por ejemplo; $p_{A,a} = P(T=A, NT=a|E) = \frac{P(T=\ \begin{smallmatrix}E\\A\end{smallmatrix}\ , NT=a)}{P(E)} = \frac{pq\bar{q}+D(\bar{q}-\theta)}{p} = (q+\frac{D}{p})\bar{q} - \theta\frac{D}{p}$ se interpreta como la probabilidad de que un padre haya transmitido el alelo A, a un hijo afectado con la enfermedad, mientras que el alelo a no fue transmitido. En el Cuadro 3.4 se muestra el cálculo de las 4 posibles probabilidades condicionales.

Para probar la diferencia en la transmisión alélica, consideraremos solo padres heterocigotos, ya que son los que aportan información de ambos alelos. Por esto último, únicamente tomaremos en cuenta $p_{A,a}$ y $p_{a,A}$ para examinar si el alelo A es transmitido con una frecuencia diferente que el alelo a. Formalmente, realizaremos la prueba de hipótesis:

$$Ho: \frac{p_{A,a}}{p_{A,a}+p_{a,A}} = \frac{1}{2} \qquad vs \qquad Ha: \frac{p_{A,a}}{p_{A,a}+p_{a,A}} \neq \frac{1}{2}$$

Para este problema, una prueba de McNemar es adecuada. Esta se basa en observar las frecuencias de los alelos transmitidos y no transmitidos de un padre, como se observa en el Cuadro 3.5 . El estadístico estándar de la prueba de McNemar esta dado por:

$$T_{TDT} = \frac{(n_{A,a}-n_{a,A})^2}{n_{A,a}+n_{a,A}}.$$

Donde:

- $n_{A,a}$ es el total de padres heterocigotos que transmitieron el alelo A y no el transmitieron el alelo a.
- $n_{a,A}$ es el total de padres heterocigotos que transmitieron el alelo a y no el transmitieron el alelo A

Bajo la hipótesis nula, el estadístico T_{TDT} sigue una distribución χ^2_1.

A diferencia de la prueba chi-cuadrada estándar que se utilizó en el método de riesgo relativo del haplotipo, el T_{TDT} no prueba independencia, si no la consistencia de las respuestas. Estamos comparando si padres heterocigotos heredan con mayor frecuencia un alelo y no el otro. Además,

Haplotipos parentales		Haplotipos transmitidos | Haplotipos no transmitidos															
E A	$\cdot$ A	E A	$\cdot$ a	E a	$\cdot$ A	E a	$\cdot$ a	N A	$\cdot$ A	N A	$\cdot$ a	N a	$\cdot$ A	N a	$\cdot$ a
E A	E A	1													
E A	E a			$\frac{1}{2}$		$\frac{1}{2}$									
E a	E A			$\frac{1}{2}$		$\frac{1}{2}$									
E a	E a						1								
N A	E A	$\frac{1}{2}$						$\frac{1}{2}$							
E A	N A	$\frac{1}{2}$						$\frac{1}{2}$							
N A	E a			$\frac{\theta}{2}$		$\frac{1-\theta}{2}$				$\frac{1-\theta}{2}$		$\frac{\theta}{2}$			
E a	N A			$\frac{\theta}{2}$		$\frac{1-\theta}{2}$				$\frac{1-\theta}{2}$		$\frac{\theta}{2}$			
N a	E A			$\frac{1-\theta}{2}$		$\frac{\theta}{2}$				$\frac{\theta}{2}$		$\frac{1-\theta}{2}$			
E A	N a			$\frac{1-\theta}{2}$		$\frac{\theta}{2}$				$\frac{\theta}{2}$		$\frac{1-\theta}{2}$			
N a	E a						$\frac{1}{2}$							$\frac{1}{2}$	
E a	N a						$\frac{1}{2}$							$\frac{1}{2}$	
N A	N A							1							
N A	N a									$\frac{1}{2}$		$\frac{1}{2}$			
N a	N A									$\frac{1}{2}$		$\frac{1}{2}$			
N a	N a													1	

Cuadro 3.2: Posibles haplotipos parentales y probabilidades de haplotipo transmitido y no transmitido

fuente: (Ziegler et al., 2010), pp.323

Haplotipos transmitidos	Alelos no transmitidos A	a	Total
$\begin{matrix}E\\A\end{matrix}$	$pq^2 + Dq$	$pq\bar{q} + D(\bar{q} - \theta)$	$pq + (1-\theta)D$
$\begin{matrix}E\\a\end{matrix}$	$pq\bar{q} - D(q - \theta)$	$p\bar{q}^2 - D\bar{q}$	$p\bar{q} - (1-\theta)D$
$\begin{matrix}N\\A\end{matrix}$	$\bar{p}q^2 - Dq$	$\bar{p}\bar{q}q - D(\bar{q} - \theta)$	$\bar{p}q - (1-\theta)D$
$\begin{matrix}N\\a\end{matrix}$	$\bar{p}q\bar{q} + D(q - \theta)$	$\bar{p}\bar{q}^2 + D\bar{q}$	$\bar{p}\bar{q} + (1-\theta)D$
Total	q	$\bar{q}$	1

Cuadro 3.3: Probabilidades de transmitir un haplotipo y no transmitir determinado alelo
fuente: (Ziegler et al., 2010), pp.324

Alelo transmitido	Alelo no transmitidos A	a	Total
A	$p_{A,A} = (q + \frac{D}{p})q$	$p_{A,a} = (q + \frac{D}{p})\bar{q} - \theta\frac{D}{p}$	$q + (1-\theta)\frac{D}{p}$
a	$p_{a,A} = (\bar{q} + \frac{D}{p})q + \theta\frac{D}{p}$	$p_{a,a} = (\bar{q} - \frac{D}{p})\bar{q}$	$\bar{q} - (1-\theta)\frac{D}{p}$
Total	$q + \frac{\theta D}{p}$	$\bar{q} - \frac{\theta D}{p}$	1

Cuadro 3.4: Probabilidades de que los padres transmitan un determinado alelo y no transmitan el otro.
fuente: (Ziegler et al., 2010), pp.324

Alelo transmitido	Alelo no transmitidos A	a	Total
A	$n_{A,A}$	$n_{A,a}$	n_{TA}
a	$n_{a,A}$	$n_{a,a}$	n_{Ta}
Total	n_{NA}	n_{Na}	$2n$

Cuadro 3.5: Frecuencias observadas de alelos transmitidos y no transmitidos para el TDT, utilizando familias TCP
fuente: (Ziegler et al., 2010), pp.325

el TDT proporciona simultáneamente una prueba de asociación y ligamiento, esto se puede observar fácilmente contrastando las probabilidades teóricas que es están fuera de la diagonal del Cuadro 3.4, teniendo que $p_{A,a} - p_{a,A} = \frac{D}{p}(1-2\theta)$. Bajo Ho, la diferencia es cero, lo que implica que $D = 0$ o $\theta = \frac{1}{2}$. Si se satisface lo primero, entonces los alelos A y E están en equilibrio de ligamiento, es decir, se heredan de manera independiente. Si se cumple que $\theta = \frac{1}{2}$ entonces no hay asociación entre en el locus investigado y la enfermedad. Por lo tanto la hipótesis nula será rechazada si existe ligamiento y asociación, (Ziegler et al., 2010).

Para finalizar notemos que, a diferencia del HRR, no es necesario que ambos padres sean heterocigotos para realizar el TDT, sin embargo solo se ocupará la información de aquellos padres que si lo sean, pues son los que aportan información al estadístico. Esto representa una desventaja pues las familias con padres homocigotos no son informativas pero se han genotipado, lo que provoca una inversión de recursos que serán desperdiciados.

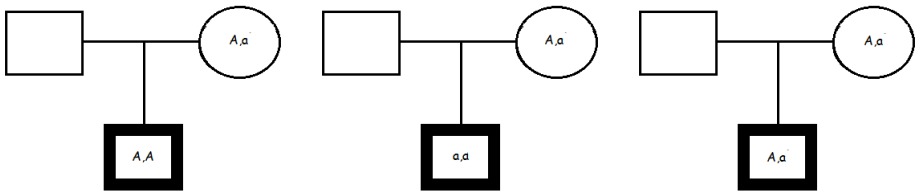
3.3. Análisis para diadas (Diseño caso-madre o padre)

En algunas ocasiones no se cuenta con información completa de familias TCP, ya sea por falta de información de la madre o el padre. (Chen, 2004) muestra que si la falta de alelos parentales es independiente del genotipo del hijo, entonces el TDT usual sigue siendo válido. Sin embargo, esto puede generar una perdida significativa de información.

Para recuperar algo de la información perdida, podemos tomar en cuenta los tres posibles casos que se tienen, si consideramos que se está analizando un locus di alélico con madre (o padre) heterocigoto e hijo afectado, como se muestra en la Figura 3.2. En los casos 3.2a y 3.2b es claro que los alelos los ha transmitido la madre. En el primero la madre transmitió el alelo A y no el alelo a. En el segundo caso, transmitió el alelo a y no A. El tercer caso no es informativo, pues no es posible distinguir cual alelo fue transmitido y cual no. (Curtis and Sham, 1995) muestran que descartar los casos que no son informativos puede llevarnos a realizar una prueba sesgada. Entender esto último no es complicado, consideremos que el alelo a es extremadamente raro en la población, esto provocará que, en la mayoría de los casos, los padres de los que no se tiene información tengan genotipo AA y por lo tanto hereden el alelo A. Entonces si la madre heterocigoto hereda el alelo a el hijo tendrá genotipo Aa y la diada no será informativa, pero si la madre hereda el alelo A entonces el hijo tendrá genotipo AA y la información de la diada se incluirá en el cálculo del TDT porque se infiere que el padre también heredó al hijo el alelo A. En esta situación, claramente concluiremos que existe evidencia en la diferencia de la transmisión alélica con preferencia en el alelo A sobre el alelo a, pero este resultado en realidad se debe a la diferencia en la frecuencia alélica de A y a en la población.

Con el objetivo de evitar sesgo en la información, (Sun et al., 1999) proponen el estadístico 1-TDT. Este estadístico se construye considerando los genotipos observados de las madres (o padres) de los que se tiene información y de los hijos afectados, esta información se resume en el Cuadro 3.6.

Comienzan a construir el estadístico considerando dos estimadores. El primero para el riesgo relativo entre individuos con genotipos Aa y aquellos con genotipos AA (λ_1), y el segundo para el riesgo relativo entre individuos con genotipos Aa y aquellos con genotipos aa (λ_{-1}), considerando que solo se tiene la información de un padre.



(a) Madre transmite alelo A. (b) Madre transmite alelo a. (c) Familia no informativa.

Figura 3.2: Diadas caso-madre, con madre heterocigoto

Genotipo del hijo	Genotipo parental		
	AA (0*)	Aa (1*)	aa (2*)
AA (0)	n_{00}	n_{01}	0
Aa (1)	n_{10}	n_{11}	n_{12}
aa (2)	0	n_{21}	n_{22}

Cuadro 3.6: Conteo de genotipos observados de los hijos considerando la información parental disponible. *Números de alelos a en el genotipo.

$$\hat{\lambda}_{-1} = \frac{n_{11} + n_{12} - n_{10}}{2 \cdot n_{21}} \qquad \hat{\lambda}_1 = \frac{n_{11} + n_{10} - n_{12}}{2 \cdot n_{01}}. \tag{3.2}$$

Estos estimadores son asintóticamente insesgados si se satisfacen dos supuestos:

- **Sup. 1:** Hombres y mujeres con el mismo genotipo en el locus investigado tienen la misma posibilidad de apareamiento.
- **Sup. 2:** Padre y madre en cada núcleo familiar está ausente con la misma probabilidad, pero en la muestra, en todas las familias, el ausente siempre es padre o siempre es madre.

Si la enfermedad no esta asociada con el locus investigado se cumple que $\lambda_{-1} = \lambda_1 = 1$. Así, si consideramos esto último como hipótesis nula (H_0,) entonces se debe satisfacer que $\hat{\lambda}_{-1} = \hat{\lambda}_1 = 1$, lo que implica que:

$$n_{11} + n_{12} - n_{10} - 2 \cdot n_{21} = 0 \tag{3.3}$$
$$n_{11} + n_{10} - n_{12} - 2 \cdot n_{01} = 0 \tag{3.4}$$

Restando (4) y (3) se tiene que: $n_{01} + n_{12} - (n_{10} + n_{21}) = 0$. Si definimos $b_1 := n_{01} + n_{12}$ y $c_1 := n_{10} + n_{21}$ concluimos que la enfermedad esta asociada con el locus investigado si $b_1 - c_1 = 0$.

Podemos interpretar a b_1 como el número de casos informativos en los que el padre ausente heredó el alelo A y c_1 como el número de casos informativos en los que el padre ausente heredo a. Por lo tanto, dado $b_1 + c_1$ y bajo H_0, se tiene que b_1 y c_1 siguen una distribución binomial, $b_1, c_1 \sim Bin(b_1 + c_1, \frac{1}{2})$. Consecuentemente, por el teorema de Moivre-Laplace, si $b_1 + c_1$ es suficientemente

grande, entonces b_1 y c_1 se aproxima a una distribución normal, $b_1, c_1 \sim N\left(\frac{1}{2}(b_1+c_1), b_1+c_1\right)$. De esto último se sigue que:

$$\frac{b_1 - c_1}{\sqrt{(b_1+c_1)}} \sim N(0,1) \tag{3.5}$$

El estadístico 1-TDT (T_1) se puede calcular directamente utilizando (5) o equivalentemente como:

$$T_1 = \frac{(b_1-c_1)^2}{b_1+c_1}. \tag{3.6}$$

Por lo que, bajo H_0, T_1 sigue una distribución χ^2 con un grado de libertad. Recordemos que este estadístico es válido sólo si se satisfacen los supuestos 1 y 2. En el caso donde una o ambas suposiciones no se cumplan, los autores proponen una alternativa para realizar la prueba 1-TDT. Toman en cuenta otros estimadores insesgados para los riesgos relativos que consideramos anteriormente, los cuales no requieren que ninguna de las 2 suposiciones se cumplan. Estos estimadores se obtienen reemplazando n_{ij} con $i, j = 0, 1, 2$ en (2) por;

$$n'_{i,j} = MP_{ij} + PM_{ij}.$$

Donde P y M y es el número de familias disponibles con padre y madre respectivamente, P_{ij} (M_{ij}) es el número de casos en donde el hijo heredó el genotipo i y el padre (madre) tiene genotipo j, estos datos se resumen en el Cuadro 3.7.

Genotipo del hijo	Genotipo del padre (madre) AA (0)	Aa (1)	aa (2)
AA (0)	P_{00} (M_{00})	P_{01} (M_{01})	0
Aa (1)	P_{10} (M_{10})	P_{11} (M_{11})	P_{12} (M_{12})
aa (2)	0	P_{21} (M_{21})	P_{22} (M_{22})

Cuadro 3.7: Conteo de genotipos observados de los hijos considerando familias con padres o madres ausentes.

Siguiendo nuevamente la idea de comparar ambos estimadores y suponiendo H_0, concluimos que se debe satisfacer que;

$$M(pb_1 - pc_1) + P(mb_1 - mc_1) = 0.$$

Donde;

$$pb_1 = P_{01} + P_{12} \qquad pc_1 = P_{10} + P_{21}$$
$$mb_1 = M_{01} + M_{12} \qquad mc_1 = M_{10} + M_{21}.$$

Si suponemos que las familias con información del padre disponible son independientes de aquellas con información de la madre disponible, y que dados pb_1+pc_1, mb_1+mc_1 suficientemente grandes, entonces bajo H_0 y por la construcción anterior, podemos concluir lo siguiente;

$$M(pb_1 - pc_1) \sim N\left(0, M^2(pb_1 + pc_1)\right)$$
$$P(mb_1 - mc_1) \sim N\left(0, P^2(mb_1 + mc_1)\right)$$

De esto último podemos concluir que el estadístico 1-TDT (T_2), para este caso, sigue una distribución χ^2 con un grado de libertad y se calcula como sigue;

$$T_2 = \frac{(M(pb_1 - pc_1) + P(mb_1 - mc_1))^2}{M^2(pb_1 + pc_1) + P^2(mb_1 + mc_1)}$$

(Sun et al., 1999) muestran que esta prueba resulta ser menos poderosa que T_1 si al menos uno de los dos supuestos se cumple.

3.4. Análisis para grupos de hermanos

En algunas ocasiones, cuando se estudian enfermedades complejas es complicado contar con información genética de los padres del hijo afectado, esto ocurre principalmente en enfermedades de diagnóstico tardío. Una forma de solucionar este problema es mediante la información de hermanos no afectados, los cuales podrían tomar el papel de controles.

Como se mencionó en la sección 2.2, este diseño requiere tomar la muestra de manera cuidadosa, pues factores como la edad del hermano pueden afectar la eficiencia de la prueba.

La primera prueba que analizaremos se le conoce como sib-TDT (S-TDT) y fue desarrollada por (Spielman and Ewens, 1998). Requiere familias en las cuales al menos un hijo padezca la enfermedad y uno que no la padezca, además que no tengan el mismo genotipo, como se observa en la Figura 3.3.

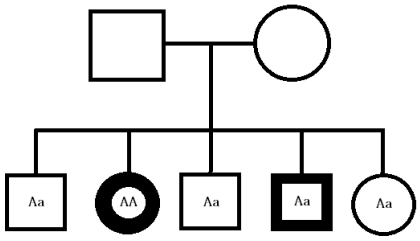


Figura 3.3: Hermandad ilustrativa para el análisis con grupos de hermanos

En resumen, el S-TDT determina si la frecuencia de un alelo (por ejemplo A), de un locus investigado en hijos afectados difiere significativamente de la frecuencia alélica en hijos no afectados. La asociación de la enfermedad con el alelo sin ligamiento no refleja la diferencia en la frecuencia alélica, por lo tanto el ligamiento debe estar presente, entonces la hipótesis nula es que la enfermedad y el locus no estén ligados.

Para realizar una prueba de hipótesis podríamos empezar por comparar las frecuencias alélicas del locus investigado en los hijos afectados y no afectados, estos datos se pueden resumir

como en en Cuadro 3.8. Para datos de este estilo lo usual es realizar una prueba χ^2 clásica, sin embargo, ésta no será válida por la dependencia de las observaciones entre hermanos afectados y no afectados.

	Número de alelos		
Estatus del hermano	A	a	Total
Afectado	$T_{Af}A$	$T_{Af}a$	$T_{Af}A + T_{Af}a$
No afectado	$T_{NAf}A$	$T_{NAf}a$	$T_{NAf}A + T_{NAf}a$

Cuadro 3.8: Frecuencia alélica observada en hermanos afectados y no afectados.

Para realizar una prueba válida podemos efectuar un procedimiento de permutación de Monte Carlo dentro de cada hermandad. En las pruebas de permutación se pretende estimar el valor-p. Es importante mencionar que esta no es una prueba estadística en el sentido usual, más bien pretende determinar el grado de significancia estadística de una hipótesis (Losilla Vidal, 2009).

La simulación Monte Carlo consiste en realizar la prueba considerando únicamente una muestra aleatoria de las permutaciones posibles, esta muestra aleatoria representa el modelo de independencia.

Para realizar la prueba con la información que ofrece este diseño familiar, comencemos considerando que se tienen n grupos de hermanos y que en cada grupo hay a_i afectados y u_i no afectados, con $i = 1, \ldots, n$. Las permutaciones se construyen dentro de cada familia, se ignora el estatus de los hermanos y se asigna aleatoriamente la categoría de afectado y no afectado, por ejemplo, si se sabe que en la familia i hay a_i hermanos afectados entonces, sin considerar el estatus de cada hermano, de manera aleatoria se asignan como 'afectados' a a_i de ellos y los demás serán no afectados. Posteriormente se calcula la frecuencia del alelo de interés en los hermanos afectados considerando esta permutación. Finalmente el valor-p se calcula considerando la proporción de replicas en las que la frecuencia alélica es igual o mayor a la observada en los datos reales.

El algoritmos para llevar a cabo la prueba es el siguiente:

1. Calcular la frecuencia del alelo de interés en los hermanos afectados para el conjunto de datos reales ($\hat{\theta}_0$).
2. Inicializar un contador ($cont = 0$).
3. Fijar el número de permutaciones (NP) que se realizarán en los datos.
4. Para $j = 1$ hasta NP.
   - 4.1 Para $i = 1$ hasta n. (Donde n es el total de grupos de hermanos.)
     - 4.1.1 $x_i^* = permutacion(x_i)$

4.2 Calcular la frecuencia alélica como en el paso uno ($\hat{\theta}_j^*$) pero utilizando los datos generados (x_i^*).

4.3 Si $\hat{\theta}_j^* > \hat{\theta}_0$ entonces $cont = cont + 1$

5. Calcular el grado de significancia (p-valor empírico) como sigue;

$$\frac{cont + 1}{NP + 1}$$

.

Con un número suficientemente grande de permutaciones el algorítmo anterior producirá un p-valor preciso (Spielman and Ewens, 1998). El método analítico, que describiremos a continuación, resulta ser esencialmente el mismo que la simulación por permutaciones si se cuenta con una muestra suficientemente grande.

Recordemos que nos interesa analizar la frecuencia del alelo A y estamos considerando n grupos de hermanos, cada uno de ellos con a_i hermanos afectados y u_i no afectados, donde $i = 1, 2, \ldots, n$, por lo que el total de hermanos en cada grupo es $t_i = a_i + u_i$. Supongamos que r_i y s_i es el número de hermanos afectados con genotipo AA y Aa en un grupo de hermanos respectivamente. Realizando el proceso de permutación en cada grupo de hermanos podemos definir a y_{1i} como la variable aleatoria que describe el número de hermanos afectados con genotipo AA en una muestra de tamaño r_i. Bajo H_0 y_{1i} sigue una distribución hipergeométrica ($y_{1i} \sim Hiper(t_i, a_i, r_i)$). Análogamente podemos construir y_{2i} como la variable aleatoria que describe el número de hermanos afectados con genotipo Aa.

Considerando la construcción de las variables aleatorias anteriores, podemos definir $y_i := 2y_{1i} + y_{2i}$ e interpretarla como el número total de alelos A en los hermanos afectados de una hermandad. Dado que conocemos la distribución de y_{1i} y y_{2i} bajo la hipótesis nula es fácil calcular la esperanza y varianza de y_i.

$$\begin{aligned} E[y_i|H_0] &= \frac{a_i(2r_i + s_i)}{t_i} \\ Var[y_i|H_0] &= 4Var[y_{1i}] + Var[y_{2i}] + 4Cov[y_{1i}, y_{2i}] \\ &= \frac{a_i u_i[4r_i(t_i - r_i - s_i) + s_i(t_i - s_i)]}{t_i^2(t_i - 1)}. \end{aligned}$$

Por lo tanto, la media y la varianza del número de alelos A, en hermanos afectados, considerando todas las hermandades y bajo H_0 son;

$$E[y] = \sum_{i=1}^{n} E[y_i|H_0] \qquad Var[y] = \sum_{i=1}^{n} Var[y_i|H_0]$$

Gracias a la construcción anterior se puede definir el estadístico S-TDT, el cual se distribuye asintóticamente normal estándar, como sigue;

$$T_{S-TDT} = \frac{y - E[y]}{\sqrt{Var[y]}} \qquad Donde; \quad y = \sum_{i=1}^{n} y_i.$$

Los autores del S-TDT mencionan que, en los estudios de enfermedades complejas, las pruebas basadas en ligamiento y asociación suelen ser más poderosas que las que se fundamentan únicamente en ligamiento. Esto representa una desventaja para esta prueba. Considerando este problema Horvath and Laird (1998) proponen otro estadístico, que contempla tanto ligamiento como asociación, y se conoce como *prueba de desequilibrio entre hermanos* (SDT por sus siglas en inglés) y requiere el mismo diseño familiar que el S-TDT.

En esencia el SDT compara el promedio de alelos A en hermanos afectados y no afectados. Para entender mejor por que se realiza esta comparación analizaremos la construcción de éste estadístico.

Comenzaremos considerando un locus dialélico (Aa). Posteriormente, para cada grupo de hermanos, denotaremos por m^A_{Aff} y m^A_{NAff} al promedio de alelos A en hermanos afectados y no afectados respectivamente, esto es;

$$m^A_{Aff} = \frac{Total\ de\ alelos\ A\ en\ hermanos\ afectados}{n_A}$$
$$m^A_{NAff} = \frac{Total\ de\ alelos\ A\ en\ hermanos\ no\ afectados}{n_U}$$

Donde n_A y n_U denotan el número de hermanos afectados y no afectados en una hermandad respectivamente.

Consideremos la diferencia $d^A = m^A_{Aff} - m^A_{NAff}$ en cada grupo de hermanos. Si $d^A = 0$ entonces esa hermandad se descartará para el análisis pues no aporta información. Fijando a b y c, como el número de hermandades tales que $d^A > 0$ y $d^A < 0$ respectivamente, se puede calcular el estadístico T_{SDT} como sigue;

$$T_{SDT} = \frac{(b-c)^2}{b+c}.$$

Horvath and Laird (1998) prueban que bajo la hipótesis nula de no asociación y no ligamiento b y c siguen una distribución binomial, ambas con parámetros $b+c$ y $1/2$ y por lo tanto T_{SDT} se distribuye asintóticamente χ^2.

Además, como se conoce la distribución de b bajo H_0 es posible calcular un p-valor exacto, por ejemplo en una prueba de dos colas estaría dado por;

$$p_{value} = 2\,\text{mín}\left[\sum_{i=0}^{b}\binom{b+c}{i}\left(\frac{1}{2}\right)^{b+c}, \sum_{i=b}^{b+c}\binom{b+c}{i}\left(\frac{1}{2}\right)^{b+c}\right].$$

Podemos pensar que esta prueba tiene mejores propiedades que la S-TDT pero, a pesar que en la mayoría de los casos resulta ser más poderosa, no siempre es así, por lo que siempre es bueno comparar ambas.

3.5. Análisis para familias extendidas (PDT)

La mayoría de las pruebas que hasta ahora hemos estudiado son válidas para asociación y ligamiento, pero requieren que las familias sean independientes, lo cual puede provocar perdida

de valiosa información.

Algunas veces es posible contar con información de familias extendidas, es decir, con familias con dos o más TCP o grupos de hermanos, como se muestra en la Figura 3.4. Con el propósito de construir una prueba que aproveche toda la información significativa de la familia y con ello obtener mayor poder, que en las pruebas que hemos analizado anteriormente, Martin et al. (2000) proponen el estadístico PDT. La prueba se basa en construir una variable aleatoria que mida el desequilibrio de ligamiento en la familia completa en lugar de tratar a cada grupo de hermanos o tríos del mismo pedigrí como independientes. La variable aleatoria se construye considerando una medida de desequilibrio de ligamiento para cada grupo de hermanos y núcleos familiares del pedigrí, posteriormente la media de estas cantidades representa el desequilibrio de ligamiento en la familia completa.

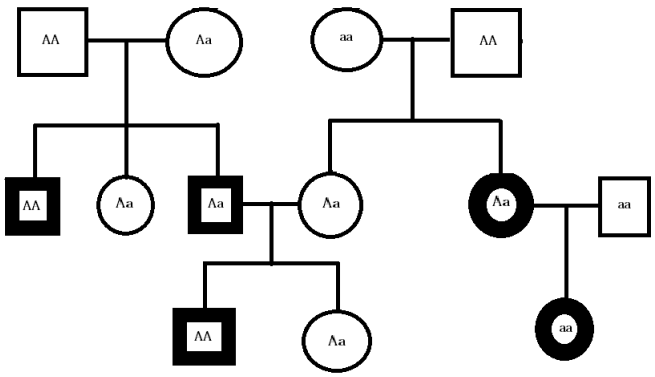


Figura 3.4: Ejemplo de familia extendida para la prueba PDT.

Existen dos tipos de familias informativas para esta prueba: los grupos de hermanos donde por lo menos hay uno afectado y uno no afectado y que no tengan el mismo genotipo, y los núcleos familiares donde se cuenta con la información de al menos un hijo afectado y ambos padres, donde por lo menos uno de ellos es heterocigoto. Se considera que la familia extendida es informativa si al menos un grupo de hermanos o núcleo familiar es informativo.

Al igual que en las pruebas anteriores, consideremos un locus con alelos A, a. Para cada familia con diseño TCP se puede saber cuántos alelos A se transmitieron y cuántos no se transmitieron. Se define una variable aleatoria X_T para cada triada dentro de un mismo núcleo familiar como la diferencia del número de alelos A transmitidos menos el número de alelos A no transmitidos. Similarmente se define la variable X_S, para cada pareja de hermanos discordante dentro de un grupo de hermanos, como la diferencia de alelos A en el hermano afectado menos número de alelos A en el hermano no afectado. Finalmente para una familia extendida con n_T triadas y n_S parejas de hermanos discordantes se define la variable aleatoria D que resume toda la información:

$$D = \frac{1}{n_T + n_S}\left(\sum_{j=1}^{n_T} X_{Tj} + \sum_{j=1}^{n_S} X_{Sj}\right)$$

Bajo la hipótesis nula de no desequilibrio de ligamiento, $E[X_T] = 0$ y $E[X_S] = 0$ para todas las triadas y todas las parejas de hermanos discordantes, por lo tanto $E[D] = 0$ para cualquier

pedigrí. Por lo que si consideramos N familias extendidas, informativas e independientes, y la variable D_i para cada una de ellas, entonces bajo la hipótesis nula se satisface que:

$$E\left[\sum_{i=1}^{N} D_i\right] = 0 \quad \text{y}$$

$$Var\left[\sum_{i=1}^{N} D_i\right] = \sum_{i=1}^{N} Var[D_i] = E\left[\sum_{i=1}^{N} D_i^2\right]$$

Por lo tanto bajo H_0, el estadístico:

$$T_{PDT} = \frac{\sum_{i=1}^{N} D_i}{\sqrt{\sum_{i=1}^{N} D_i}}$$

se distribuye asintóticamente normal estándar (Martin et al., 2000).

3.6. Asociación considerando covariables y rasgos cuantitativos

Hasta ahora las pruebas de hipótesis que hemos considerado prueban únicamente asociación entre un locus (o alelo) y la enfermedad (la que consideraremos como rasgo categórico o cualitativo, ya que puedes padecerla o no), sin embargo en múltiples ocasiones se desea probar asociación entre alelos y un rasgo cuantitativo, por ejemplo, se desea probar que genotipos se asocian a la cantidad de glucosa en la sangre, altura, entre otros.

En esta sección estudiaremos un modelo general que nos permite probar asociación sin importar si el rasgo es cuantitativo o cualitativo, considerando covariables por las cuales se puede ver afectado.

Comenzaremos explicando en que consiste un modelo lineal mixto, ya que es la base para probar asociación utilizando distintas técnicas considerando el tipo de rasgo en el que estemos interesados.

Modelo lineal mixto

Consideremos que un rasgo cuantitativo particular, como el Índice de Masa Corporal (IMC), se observa en n individuos agrupados por familias, esta información se puede almacenar en un vector y de tamaño $n \times 1$. Supongamos que tales observaciones se describen adecuadamente, de forma matricial, por el modelo lineal mixto;

$$y = X\beta + Zu + e$$

donde β es un vector de efectos fijos y u un vector de efectos aleatorios de tamaños $p \times 1$ y $q \times 1$ respectivamente. Usualmente β contiene información de la media poblacional del rasgo así como la de otros factores como el género, año de nacimiento, nivel de glucosa en la sangre, entre otros. u normalmente describe los efectos genéticos. Las matrices X y Z se conocen como matrices de incidencia y son de tamaño $n \times p$ y $n \times q$ respectivamente, finalmente e es un vector

que representa el error aleatorio en la relación y asumiremos que se distribuye independientemente de los factores genéticos aleatorios.

Para analizar la media y varianzas del modelo lineal mixto asumiremos que $E[u] = E[e] = 0$, por lo que $E[y] = X\beta$. Denotemos por R y G a las matrices de covarianzas de e y u respectivamente, dado que $y - E[y] = Zu + e$, $E[u] = E[e] = 0$ y u, e no están correlacionados entonces;

$$\begin{aligned} var(y) &= E[(Zu + e)(Zu + e)^t] \\ &= var(Zu + e) \\ &= var(Zu) + var(e) \\ &= ZGZ^t + R \end{aligned}$$

El término ZGZ^t es la contribución de los efectos genéticos aleatorios y R representa la contribución de los residuales. Generalmente se asume que los errores residuales tienen varianza constante y no están correlacionados, por lo que R es una matriz diagonal $R = \sigma_E^2 I$.

Para el modelo mixto, y, X y Z son variables observables, mientas que β, u, R y G son generalmente desconocidas. Consecuentemente en análisis del modelo mixto consiste en realizar dos estimaciones complementarias: Primero se estiman las matrices de covarianzas R y G y posteriormente los vectores de efectos fijos y aleatorios β y u.

Para comenzar con el análisis supondremos que X, Z, G y R son conocidas y con ello poder estimar β y u. Más adelante se realizará la estimación de G y R.

Estimación de los efectos fijos y predicción de los efectos aleatorios

Para comenzar con el análisis de la estimación de los efectos fijos y los predictores de los aleatorios supondremos que las matrices de covarianzas R y G son conocidas.

Existen distintos métodos para obtener los estimadores y predictores, sin embargo los más utilizados son los BLUE (mejor estimador lineal insesgado, BLUE por sus siglas en inglés) y BLUP (mejor predictor lineal insesgado). Se dicen que son los mejores en el sentido de minimizar la varianza muestral, lineales como función del fenotipo observado y, e insesgado ya que satisfacen $E[\text{BLUE}(\hat{\beta})] = \beta$ y $E[\text{BLUP}(\hat{u})] = u$.

Suponiendo que: X es de rango completo, los efectos fijos modelan la media de y, es decir, $E[y] = X\beta$, los efectos aleatorios múltiples se consideran independientes entre si, y que la variable dependiente y procede de una distribución multinormal. Lynch et al. (1998) muestra, usando máxima verosimilitud, que el estimador BLUE para β es:

$$\hat{\beta} = (X^t V^{-1} X)^{-1} X^t V^{-1} y \tag{3.7}$$

donde $V = ZGZ^t + R$. Además, si se asume que u y e siguen una distribución multinormal con media cero y varianza G y R respectivamente, (Henderson, 1963) muestra que el predictor BLUP de u es:

$$\hat{u} = GZ^t V^{-1}(y - X\hat{\beta}) \tag{3.8}$$

A $\hat{u}$ también se le conoce como predictor Bayesiano empírico. Debido a que estamos suponiendo que el vector de efectos aleatorios u sigue una distribución multinormal con media cero

y varianza G, entonces cada elemento del vector (u_i) sigue una distribución normal con media cero y varianza σ^2_G. Analizándolo desde el punto de vista Bayesiano, se puede interpretar que la distribución a priori tiene esperanza cero. La distribución a posteriori de los efectos aleatorios dados los datos satisface que $f(u_i|y) \propto f(y|u_i)f(u_i)$, por lo que la esperanza de la distribución a posteriori será:

$$E[u_i|y] = \int u_i f(u_i|y_i)du_i$$

.

(Fong et al., 2010) muestran que para el caso especial en el que se satisface $y|u \sim NMV(X\beta + Zu, \sigma^2 I)$, $u \sim NMV(0, G)$ la media a posteriori es:

$$\hat{u} = GZ^tV^{-1}(y - X\beta)$$

Utilizando $\hat{\beta}$ como estimador de β tendremos un predictor de u justificado como estimador Bayesiano.

La aplicación práctica del estimador $(\hat{\beta})$ y predictor $\hat{u}$, requiere que los elementos de la varianza G y R son conocidos, es por ello que las estimaremos mediante el método de máxima verosimilitud restringida (MVRE) y con ello poder hacer inferencias sobre los efectos fijos y aleatorios.

Estimación de los componentes de varianza

Existen distintos métodos para estimar los componentes de la varianza, la manera más común es realizar un análisis de varianzas (ANOVA) siguiendo el método de mínimos cuadrados para estimar β (el cual coincide con el estimador de máxima verosimilitud cuando G y R son conocidas) y con ello estimar los componentes de la varianza equiparando la suma de los cuadrados medios observados con las expresiones que describen sus valores esperados. Los estimadores de la varianza obtenidos con este método son insesgados independientemente de si los datos siguen una distribución normal, sin embargo existen dos limitaciones importantes. Primeramente, las observaciones que estamos considerando involucran registros de familias, tales como: grupos de hermanos, que no se pueden analizar conjuntamente con ANOVA. En segundo lugar, los estimadores ANOVA requieren que los datos obtenidos por familia estén bien balanceados, en un buen caso el número de integrantes por familias es el mismo, sin embargo, esto algunas veces no es suficiente. En situaciones prácticas es muy difícil lograr esto debido a que los núcleos familiares o grupos de hermanos no son del mismo tamaño, o simplemente las características de la familias difieren mucho entre si.

A diferencia de los estimadores ANOVA los obtenidos por máxima verosimilitud (MV) o máxima verosimilitud restringida (MVRE) no imponen ninguna exigencia especial respecto al balance y diseño de los datos. Estos estimadores son ideales para los diseños desbalanceados que surgen en los estudios genéticos basados en familias.

Entender la idea conceptual de los estimadores de MV es muy sencillo, sin embargo, un inconveniente de esta estimación es que asume que todos los efectos fijos son conocidos o estimados sin error, lo que puede producir un sesgo en las estimaciones que realicemos. Un ejemplo sencillo para ilustrar este echo, es cuando se calcula el estimador de MV de la varianza de una muestra aleatoria de variables normales, ya que cuando la media no es conocida y debe estimarse, dicha

estimación introduce un sesgo en el estimador MV de la varianza. A diferencia de los estimadores utilizando MV, los estimadores vía MVRE maximizan solo la parte de la verosimilitud que no depende de los efectos fijos, por lo que el sesgo producido en los estimadores MV puede ser eliminado, es por ello que el método de MVRE es preferido para analizar conjuntos de datos grandes con estructura compleja.

A pesar que el método de MVRE es preferido sobre MV, comenzaremos con los estimadores de MV, ya que la estimación vía MVRE puede expresarse como un problema de MV mediante una transformación lineal simple.

Consideremos el modelo lineal mixto general, $y = X\beta + Zu + e$, y asumamos que: $u \sim NMV(0, G)$ y $e \sim NMV(0, R)$. Bajo estos supuestos, se satisface que: $y \sim NMV(X\beta, V = ZGZ^t + R)$, por lo que la función log-verosimilitud para β y V, dados los datos observados X, y, es:

$$L(\beta, V|X, y) = -\frac{n}{2}\ln(2\pi) - \frac{1}{2}det(V) - \frac{1}{2}(y - X\beta)^t V^{-1}(y - X\beta). \tag{3.9}$$

Para continuar con el análisis, consideremos que $u = a$, donde a representa los valores genéticos aditivos. De acuerdo con (Falconer and Mackay, 1996), el valor genético aditivo es la suma del aporte individual que cada alelo hace al genotipo, y representa solo la parte que puede ser transmitida de los padres a su descendencia.

Los componentes de la varianza que intentamos analizar involucran a G y R, debido a que a representa los valores genéticos aditivos entonces podemos suponer que $G = \sigma_A^2 A$, donde A es la matriz de relación genética aditiva y se compone de las varianzas y covarianzas de los valores genéticos aditivos, además es un componente muy importante ya que representa la principal causa de semejanza entre familiares y es el único componente que puede ser fácilmente estimado dadas las observaciones, más adelante hablaremos de dicha estimación. Asumiremos que la varianza de los residuales de individuos diferentes es independiente y la misma para cada uno (propiedad de homocedasticidad), es decir, $R = \sigma_E^2 I$.

El enfoque que hemos dado hasta ahora facilita la estimación de los componentes de la varianza aditiva, sin embargo podemos definir un modelo más general en el cual se pueden incorporar varianzas de dominancia o no aditivas de la siguiente manera:

$$y = X\beta + \sum_{i=1}^{n} Z_i u_i + e \tag{3.10}$$

donde u_i representa los efectos aleatorios los cuales asumiremos que no están correlacionados y que siguen una distribución multinormal, $u_i \sim NMV(0, \sigma_i^2 B_i)$, con B_i matrices de constantes conocidas. La función los máxima verosimilitud para este modelo es la misma que la ecuación 3.9 pero con matriz de varianzas y covarianzas:

$$V = \sum_{i=1}^{n} \sigma_i^2 Z_i B_i Z_i^t + \sigma_E^2 I \tag{3.11}$$

donde solo se desconocen σ_E^2 y σ_i^2 con $i = 1 \ldots n$. Para fines prácticos, en la aplicación del método consideraremos únicamente el modelo aditivo el cual es un caso particular del modelo

general pues satisface que:

$$V = \sigma_A^2 A + \sigma_E^2 EI.$$

De acuerdo con el método de MV, para obtener los estimadores, debemos calcular $\frac{\partial L}{\partial \beta}$ y $\frac{\partial L}{\partial V}$. La derivada parcial respecto de beta, únicamente involucra el último término de la ecuación 3.9 el cual corresponde a una forma cuadrática, por lo que el resultado es:

$$\frac{\partial L(\beta V|X,y)}{\partial \beta} = X^t V^{-1}(y - X\beta) \tag{3.12}$$

Para obtener las derivadas parciales de σ_i^2 y σ_E^2 ocuparemos las siguientes propiedades.

Si M es una matriz cuadrada cuyas entradas son funciones de x, se satisface:

- $\frac{\partial \ln(\det(M))}{\partial x} = \mathrm{tr} M^{-1} \frac{\partial M}{\partial x}$
- $\frac{\partial M^{-1}}{\partial x} = -M^{-1} \frac{\partial M}{\partial x} M^{-1}$

donde tr es el operador traza. Ocupando la ecuación 3.11 es sencillo verificar que $\frac{\partial V}{\partial \sigma_i^2} = V_i$ con $V_i = I$ si $\sigma_i^2 = \sigma_E^2$ o $V_i = Z_i B_i Z_i^t$ en otro caso. Gracias a las propiedades anteriores resulta sencillo calcular la derivada parcial de la función log verosimilitud respecto a las varianzas desconocidas, obteniendo como resultado:

$$\frac{\partial L(\beta V|X,y)}{\partial \sigma_i^2} = -\frac{1}{2}\mathrm{tr}(V^{-1}V_i) + \frac{1}{2}(y - X\beta)^t V^{-1} V_i V^{-1}(y - X\beta) \tag{3.13}$$

Los estimadores se obtienen igualando a cero las ecuaciones 3.12 y 3.13 y resolviendo simultáneamente. Utilizando la ecuación 3.12 y ocupando la estimación de V es sencillo verificar que el estimador de MV para β es:

$$\hat{\beta} = (X^t \hat{V}^{-1} X)^{-1} X^t \hat{V}^{-1} y \tag{3.14}$$

Empleando la ecuación 3.13 podemos concluir que el estimador de MV para la varianza puede ser sesgado, ya que si $\hat{\beta} \neq \beta$ y sumando y restando $X\hat{\beta}$ de forma adecuada, tenemos que:

$$\begin{aligned}
\frac{\partial L(\beta V|X,y)}{\partial \sigma_i^2} &= -\frac{1}{2}\mathrm{tr}(V^{-1}V_i) + \frac{1}{2}(y - X\beta + X\hat{\beta} - X\hat{\beta})^t V^{-1} V_i V^{-1}(y - X\beta + X\hat{\beta} - X\hat{\beta}) \\
&= -\frac{1}{2}\mathrm{tr}(V^{-1}V_i) + \frac{1}{2}(y - X\hat{\beta})^t V^{-1} V_i V^{-1}(y - X\hat{\beta}) + \\
&+ \frac{1}{2}(\hat{\beta} - \beta)^t X^t V^{-1} V_i V^{-1} X(\hat{\beta} - \beta)
\end{aligned}$$

Los estimadores de MV de la varianza se obtienen asumiendo que $\hat{\beta} = \beta$ e igualando a cero la ecuación anterior, de esto se sigue que:

$$\mathrm{tr}(\hat{V}^{-1}V_i) = (y - X\hat{\beta})^t \hat{V}^{-1} V_i \hat{V}^{-1}(y - X\hat{\beta}) \tag{3.15}$$

Para simplificar la ecuación definamos la matriz P como:

$$P = V^{-1} - V^{-1}X(X^t V^{-1} X)^{-1} X^t V^{-1}$$

Denotaremos con $\hat{P}$ como estimador de P, destacando que depende de la varianza V que intentamos estimar. Ocupando el hecho de que $\hat{P}y = \hat{V}^{-1}(y - X\hat{\beta})$, la ecuación 3.15 se puede reescribir como:

$$\text{tr}(\hat{V}^{-1}V_i) = y^t\hat{P}V_i\hat{P}y. \tag{3.16}$$

En resumen los estimadores de MV satisfacen las ecuaciones 3.14 (para los efectos fijos) y 3.16 para las componentes de la varianza. Observemos que tomando en cuenta m efectos aleatorios y un residual debemos resolver simultáneamente un conjunto de $m+1$ ecuaciones para estimar las varianzas de los efectos aleatorios. Si utilizamos el modelo aditivo se deben resolver únicamente dos ecuaciones:

$$\begin{aligned} \text{tr}(\hat{V}^{-1}) &= y^t\hat{P}\hat{P}y && \text{para} \quad \sigma^2_E \\ \text{tr}(\hat{V}^{-1}ZAZ^T) &= y^t\hat{P}ZAZ^t\hat{P}y && \text{para} \quad \sigma^2_A \end{aligned}$$

Las soluciones de la ecuación 3.14 y las anteriores tienen dos propiedades poco favorecedoras. Primeramente, la solución para $\hat{\beta}$ no es cerrada, ya que depende de la matriz de varianzas y covarianzas. Además las ecuaciones involucran la matriz inversa de $\hat{V}$, por lo que no es una función lineal de los componentes de la varianza y su solución requiere procedimientos iterativos. Algunos métodos usados son el algoritmo de Newton Raphson modificado o el de maximización de la esperanza. Finalmente cabe destacar que estos estimadores ya no son BLUE y BLUP como en el caso anterior, donde suponíamos que G y R son conocidas.

Recordemos que el objetivo que estamos intentando alcanzar es realizar es realizar la estimación vía MVRE para evitar el sesgo provocado por suponer que $\hat{\beta} = \beta$. Este método consiste en aplicar una transformación lineal al vector de observaciones y que remueva los efectos fijos del modelo.

Para comenzar con el análisis imaginemos una matriz K asociada a la matriz X tal que $KX = 0$. Aplicando K al modelo mixto tenemos:

$$y^* = Ky = K(X\beta + Za + e = KZa + Ke)$$

No es difícil notar que los estimadores de MV se pueden aplicar al modelo modificado considerando las siguientes sustituciones:

$$Ky \quad \text{por} \quad y, \quad KX = 0 \quad \text{por} \quad X, \quad KZ \quad \text{por} \quad Z \quad \text{y} \quad KVK^t \quad \text{por} \quad V \tag{3.17}$$

En principio, realizar la estimación con este enfoque, requiere encontrar la matriz K, sin embargo, en Searle et al. (2009) se prueba que K satisface $P = K^t(KVK^t)^{-1}K$, ocupando este echo podemos notar que:

$$(y^*)^t(V^*)^{-1}y^* = (y^tK^t)(KVK^t)^{-1}(Ky) = y^tPy.$$

De esto último y reemplazando las sustituciones de 3.17 en 3.15 obtenemos las ecuaciones de MVRE:

$$\text{tr}(\hat{P}) = y^t\hat{P}\hat{P}y \qquad \text{para} \quad \sigma^2_E \tag{3.18}$$

$$\text{tr}(\hat{P}ZAZ^t) = y^t\hat{P}ZAZ^t\hat{P}y \qquad \text{para} \quad \sigma^2_A \tag{3.19}$$

Notemos que no podemos recuperar la estimación de β bajo este enfoque pues removimos los efectos fijos, por lo que su calculo consiste en sustituir las estimaciones obtenidas de las varianzas en $\hat{\beta}$ obtenido mediante MV.

Entender la construcción y estimación de los parámetros de un modelo lineal mixto con efectos aditivos es importante, ya que es la base del modelo que implementaremos, el cual, además de ser capaz de realizar diversos análisis genéticos cuantitativos, nos ayudará a construir una prueba de asociación. Dicho modelo es conocido como Poligénico, y en el siguiente capítulo explicaremos con un poco más de detalle sus características y cómo utilizarlo para probar asociación.

Modelo poligénico y asociación

De acuerdo con (Zhou et al., 2013), el objetivo de la modelación poligénica es entender mejor la relación entre la variación genética y la variación de las características observadas, incluyendo la variación en rasgos cuantitativos (tales como: el nivel de colesterol en los humanos o la producción de leche en el ganado) y la susceptibilidad a enfermedades.

El modelo poligénico que utilizaremos se basa en un MLM el cual permite utilizar diseños familiares considerando efectos aleatorios que modelan la correlación de miembros de la misma familia. Este modelo incluye efectos fijos β y tres efectos aleatorios: g, c y e.

$$y = X\beta + g + c + e \tag{3.20}$$

El primer término, g, explica el efecto genético aditivo, con matriz de covarianzas A. Como mencionamos en la sección anterior, esta matriz puede ser fácilmente estimada utilizando las observaciones, (Tier, 1990) sugiere que una buena estimación se puede hacer calculando la matriz de parentesco empírica K y multiplicarla por 2 es decir, $A \approx 2K$. La matriz de parentesco empírica proporciona una estimación probabilística de que un gen aleatorio de un sujeto dado, i, sea idéntico por descendencia a un gen en el mismo locus de un sujeto, j. Si consideramos una familia extendida de m personas estas probabilidades se describen en la matriz K de tamaño $m \times m$ la construcción de esta matriz se puede consultar en (Cavalli-Sforza and Edwards, 1967).

El segundo término, c, se denomina efecto doméstico pero en general está relacionado con el intercambio de cualquier factor no genético entre individuos que compartan un entorno familiar o algún efecto ambiental, por lo que hay pedigríes que pueden unirse a otros grupos familiares si existe una relación entre ellos, es decir contempla individuos que, aunque no estén en el mismo pedigrí, podrían estar en el mismo grupo ambiental o efecto domestico, suponiendo que $c \sim NMV(0, \sigma_c^2 H)$.

Finalmente el tercer término es el error residual con matriz de covarianzas $\sigma_e^2 I$.

Este modelo además de determinar la proporción de la varianza de los rasgos cuantitativos qué es explicada por las covariables (genéticas o ambientales) nos permite realizar pruebas de asociación entre uno o varios alelos de distintos locus y el rasgo cuantitativo, esto último mediante una extensión del modelo poligénico agregando una nueva covariable de efectos fijos (β_{snp}).

$$y = X\beta + \beta_{snp} * snp + g + c + e \tag{3.21}$$

El valor-p de la asociación, entre el rasgo cuantitativo y el alelo de riesgo, se basa en la prueba de razón verosimilitud (LRT por sus siglas en inglés), la cual compara el modelo poligénico

original (3.20) contra el modelo (3.21), es decir se prueba $H_o: \quad \beta_{snp} \neq 0$ vs $H_a: \quad \beta_{snp} = 0$.

Como mencionamos al inicio de este capítulo, este enfoque es muy útil para hacer análisis de rasgos cuantitativos, incluyendo el modelo de asociación. Sin embargo el objetivo principal de este trabajo es determinar si hay asociación entre uno o varios alelos y una enfermedad específica, en este caso el rasgo es cualitativo, dicotómico específicamente, pues indica si un individuo se muestra, o no, afectado por la enfermedad de interés. Para estos casos los modelos de umbral de riesgo son apropiados ((Falconer, 1967; Gottesman and Shields, 1972; Hayeck et al., 2017)).

Su objetivo principal es modelar los factores de riesgo que contribuyen a la enfermedad. En nuestro caso las variables son todos los genes y diferentes condiciones ambientales que pueden proteger o aumentar el riesgo de contraer la enfermedad. En la siguiente sección explicaremos con más detalle la construcción de estos modelos.

Modelos de umbral de riesgo

Comencemos suponiendo que la probabilidad de padecer una enfermedad de interés se ajusta como una función del riesgo latente y es continua.

Denotemos por l el riesgo y por D el estado de afectación, donde $D = 1$ si el individuo está afectado por la enfermedad y $D = 0$ si no lo está. Por lo que podemos expresar la probabilidad de que un individuo esté enfermo como:

$$p(D = 1) = \int_{-\infty}^{\infty} f(l)s(l)dl$$

donde $f(\cdot)$ es la distribución de probabilidad y $s(\cdot)$ una función de riesgo. Por otra parte, la probabilidad conjunta, de n individuos relacionados, de padecer o no la enfermedad es:

$$p(D_1, \ldots, D_n) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_n(l_1, \ldots, l_n) \times \prod_{i=1}^{n} s(l_i)^{D_i}[1 - s(l_i)]^{1-D_i} dl_i \cdots dl_n$$

$f_n(\cdot)$ denota la distribución conjunta de n variables continuas correlacionadas. Hasta ahora hemos descrito un modelo de riesgo general, pero nosotros estamos interesados en particular en el modelo de umbral de riesgo, el cual asume que el riesgo l sigue una distribución normal estándar y que la función de riesgo modela un umbral que determina si un individuo está afectado, si su riesgo latente l sobrepasa el umbral, es decir, la función $s(\cdot)$ se define como:

$$s(l) = \begin{cases} 1 & si \quad l > T \\ \\ 0 & si \quad l \leq T \end{cases}$$

Bajo este enfoque, el riesgo continuo l representa la suma de un gran número de factores genéticos y ambientales independientes. Además podemos considerar la susceptibilidad de cada individuo de contraer la enfermedad, sugiriendo T como una función de los rasgos fenotipicos, o bien considerar el mismo umbral para todos los individuos. Bajo los supuestos anteriores la distribución conjunta de en individuos de padecer la enfermedad o no es:

$$P(D) = \int_{I_1} \cdots \int_{I_n} \phi_n(l_1, \ldots, l_n; R) dl_1 \ldots dl_n$$

donde $D = \{D_1, \dots D_n\}$ es el conjunto de estados de afectación, $\phi(\cdot)$ indica la distribución de probabilidad normal con matriz de correlación R, y finalmente I_j denota el intervalo $(-\infty, T_j)$ si $D = 0$ o $[T_j, \infty)$ si $D = 1$, con T_j el umbral del individuo j, con $j = 1, \dots, n$.

Los umbrales y correlaciones del modelo son desconocidos y se pueden estimar vía MV (Thompson, 1972). Una vez hecho el ajuste de este modelo se prueba la asociación de los alelos con la enfermedad utilizando nuevamente una prueba LRT.

En este capítulo nos enfocamos en discutir las distintas pruebas estadísticas que nos ayudan a probas asociación genética. En el siguiente capítulo mostraremos una aplicación de algunos de estos métodos. Es importante resaltar que ya existen software que realiza estas pruebas en donde unicamente tenemos que ingresar los datos con la estructura correcta y definir los parámetros del modelo.

Capítulo 4

Aplicación de los métodos estadísticos a un diseño familiar real para labio y paladar hendido en una muestra de pacientes mexicanos

4.1. Estadística descriptiva

Los datos que analizaremos proceden de un estudio de asociación de genes candidatos en labio leporino y paladar hendido (LPH), realizado por la Unidad de Biología Molecula y Medicina Genómica del Instituto Nacional de Ciencias Médicas y Nutrición Salvador Zubirán. El estudio parte considerando un caso índice y se toman en cuenta la familia si también se tiene información de por lo menos un progenitor.

Para llevar a cabo las pruebas de asociación contamos con los datos de 27 familias mexicanas, que en total suman 98 individuos, 52 mujeres y 47 hombres. Del total de individuos, 41 corresponden a *no afectados* y 57 a *afectados*, de estos últimos el 49 % son hombres y el 51 % mujeres. Las familias se agrupan en los siguientes diseños familiares:

- 8 Familias Tríos Casos Padres
- 12 Núcleos familiares de 4 integrantes
- 3 Núcleos familiares de 5 integrantes
- 3 Diadas Madre-Hijo
- 1 Diada Madre-Hijo y 3 hermanos.

El labio leporino y paladar hendido ocurre cuando el labio superior o paladar no se desarrollan correctamente y esto sucede durante el primer trimestre de embarazo. Un bebe puede tener labio leporino, o paladar hendido o ambos, por lo que existen distintos aspectos de LPH. En nuestra muestra contamos con afectados de 12 distintos espectros, la información se resume en el Cuadro 4.1. Observemos que el espectro más frecuente es labio y paladar hendido bilateral y no hay diferencia entre hombres y mujeres. En general, se observa una proporción equilibrada entre el género de afectados en los distintos espectros de LPH, sin embargo, el espectro labio y

paladar hendido izquierdo es el único que, en la muestra, se presenta con mayor frecuencia en mujeres que en hombres.

Espectro LPH	Total	Hombres	Mujeres	Casos índice
Labio y paladar hendido bilateral	24	12	12	14
Labio y paladar hendido derecho	6	4	2	3
Labio y paladar hendido izquierdo	16	6	10	7
Labio hendido bilateral incompleto	2	2	0	1
Labio hendido derecho	3	2	1	0
Labio hendido incompleto	1	0	1	1
Labio hendido izquierdo	1	0	1	1
Labio hendido izquierdo incompleto	1	0	1	0
Labio hendido Minimicroforma	1	1	0	0
Microforma izquierda	1	0	1	0
Paladar hendido	1	0	1	0
Afectados	57	28	29	27

Cuadro 4.1: Frecuencia de espectros LPH entre hombres y mujeres

Examinando los datos de los casos índice, encontramos que 15 son mujeres y 12 hombres. Además, en 19 familias se registro un progenitor afectado, en 10 de ellas resultó ser la madre, y en 9 el padre. Esto es un indicio de que no hay relación entre el desarrollo del defecto congénito y el género del afectado, adicional a ello, los niños con al menos un padre afectado, parecen más susceptibles comparados con los que tienen padres no afectados.

4.2. Pruebas estadísticas

Para probar asociación entre algunos de los SNPs y LPH emplearemos dos pruebas que discutimos en la sección 3. La primera es la prueba TDT, en la cual consideramos 23 tríos caso padres tomando la información de los 15 núcleos familiares de 4 y 5 integrantes, considerando solo los casos además de los 8 tríos de la muestra.

Esta prueba se encuentra implementada en el software PLINK (como se indica en el apéndice B) y calcula el p_valor de tres pruebas, el estadístico TDT tradicional, la prueba de discordancia parental (PAR), la cual es muy similar al TDT tradicional pero realiza el conteo de alelos transmitidos y no transmitidos tomando en cuenta si el padre o madre padece o no la enfermedad, y finalmente una prueba combinada (COM) que considera los conteos del TDT tradicional y de la prueba de discondancia parental, las tres estadísticas siguen una distribución chi-cuadrada con un grado de libertad bajo la hipótesis nula.

La segunda prueba de asociación que emplearemos es la basada en el modelo mixto, en ella ocupamos la información completa de las 27 familias y utilizamos el software SOLAR para su implementación (ver apéndiceA).

En el Cuadro 4.2 mostramos los p_valores nominales obtenidos en las cuatro pruebas realizadas. Podemos observar que los resultados de todas las pruebas son los mismos para algunos pares de SNPs, como en el 8 y 9, o el par 7 y 16, que además cada par comparte el mismo gen. Esto indica que los alelos de ambos locus se encuentran en desequilibrio de ligamiento, es decir, se heredan juntos. Adicional a ello, notemos que en pocos casos los p_valores son significativos

SNP	TDT	TDT-PAR	TDT-COM	Modelo mixto
SNP 1	0.563	0.0832	0.149	0.134
SNP 2	0.563	0.020	0.220	0.374
SNP 3	0.827	0.593	0.612	0.957
SNP 4	0.131	0.654	0.133	0.400
SNP 5	0.654	0.033	0.449	0.806
SNP 6	0.438	0.157	0.601	0.598
SNP 7	0.808	0.006	0.026	0.019
SNP 8	0.049	0.637	0.078	0.952
SNP 9	0.049	0.637	0.078	0.952
SNP 10	0.563	0.083	0.149	0.134
SNP 11	0.563	0.020	0.220	0.374
SNP 12	0.827	0.593	0.612	0.957
SNP 13	0.131	0.654	0.133	0.400
SNP 14	0.654	0.033	0.449	0.806
SNP 15	0.438	0.157	0.601	0.598
SNP 16	0.808	0.006	0.026	0.019
SNP 17	0.032	0.256	0.275	0.229
SNP 18	0.617	0.654	0.512	0.211
SNP 19	0.414	0.414	1.000	0.646
SNP 20	0.393	0.593	0.738	0.725
SNP 21	0.763	0.414	0.466	0.835
SNP 22	0.563	0.108	0.432	0.953
SNP 23	0.563	0.763	0.531	0.326
SNP 24	0.637	1.000	0.723	0.855
SNP 25	0.405	0.479	0.827	0.883
SNP 26	0.781	0.179	0.345	0.467
SNP 27	1.000	0.317	0.563	0.819
SNP 28	0.365	0.781	0.414	0.626
SNP 29	0.205	1.000	0.285	0.580
SNP 30	0.049	0.637	0.078	0.952
SNP 31	0.365	0.157	0.818	0.426
SNP 32	0.827	0.466	0.516	0.843

Cuadro 4.2: p_valores nominales de las pruebas realizadas.

por lo menos a nivel de 0.1. Esto se puede apreciar con mejor detalle en el Cuadro 4.3,en él reportamos el nivel de significancia de asociación de los SNPs con el fenotipo nominalmente. Observemos que los SNPs 7 y 16 se asocian al fenotipo significativamente al 0.05 en tres de las cuatro pruebas. Sin embargo haciendo un ajuste de Bonferroni para hacer la comparación de todas las pruebas de manera simultánea resulta que ningún SNP está asociado con la enfermedad. Es importante mencionar que el tamaño de la muestra, con la que contamos, es pequeño y se está trabajando en aumentarla. Hasta el momento tenemos la información de 27 familias de las cuales solo en el modelo mixto ocupamos la información de todas ellas. Debido a que la prueba TDT únicamente considera familias TCP se utilizó la información de 23 familias con las que se podía formar la triada. Aquí es importante mencionar que las familias deben ser independientes, por lo que si en una familia se tiene la información de padre, madre y 3 hijos de los cuales 2 son afectados, en el TDT solo se considera la información de solo un hijo afectado con padre y madre, ignorando la información de la triada que el segundo hermano pudo haber formado.

Si bien con el ajuste de bonferroni no encontramos ningún SNP asociado a la enfermedad, los resultados nominales nos indican que vamos por buen camino, continuamos en espera de la información de más familias para realizar nuevamente las pruebas. Por otra parte, en el modelo mixto no incluimos covariables de efectos fijos además de los SNPs, sería importante contar con la información y hacer un ajuste para determinar si hay covariables que interfieran en la asociación genética y enriquecer el modelo.

SNP	TDT	TDT-PAR	TDT-COM	Modelo mixto
SNP 1		*		
SNP 2		**		
SNP 5		**		
SNP 7		***	**	**
SNP 8	**		*	
SNP 9	**		*	
SNP 10		*		
SNP 11		**		
SNP 14		**		
SNP 16		***	**	**
SNP 17	**			
SNP 30	**		*	

Cuadro 4.3: Asociacion nominal de cada SNP con el fenotipo.* Significancia al 0.1. ** Significancia al 0.05. *** Significancia al 0.01

Capítulo 5

Conclusiones

Los estudios de asociación genética más comunes se realizan considerando casos y controles que no estén relacionados genéticamente, hemos mencionado que con estos métodos se tiene que realizar un ajuste por estratificación poblacional cuando se estudian poblaciones de mezclas recientes. Por ello los estudios considerando diseños familiares representan una alternativa.

La estadística juega un papel importante en los estudios de asociación genética, sin embargo cuando se consideran familias tenemos que considerar que no contamos con independencia entre individuos y los métodos que se han desarrollado suelen ser más complejos que los utilizados en los EAG considerando casos y controles.

Los métodos estadísticos que se utilizan para los EAGBF se han modificado y enriquecido para no tratar solo la asociación del SNP con la enfermedad, los modelos mixtos han permitido estudiar la asociación considerando rasgos continuos y realizar ajustes por covariables, esto es de gran utilidad para entender las enfermedades complejas y analizar los factores genéticos y ambientales por las cuales se pueden ver afectadas.

A pesar de que los diseños familiares enriquecen los EAG, los métodos estadísticos tienen la desventaja de requerir un gran número de familias para la muestra, además de requerir la información de la historia familiar lo cual podría representar una desventaja si no se cuenta con esta información, esto puede ocurrir principalmente en las enfermedades de diagnostico tardío como la diabetes o en las enfermedades propias de edad avanzada como el Alzheimer. Aún queda trabajo por realizar, si bien el objetivo del trabajo era discutir los métodos estadísticos, y ejemplificarlos, continuamos trabajando en ampliar la muestra para mejorar nuestros resultados y poder realizar un mejor análisis.

Apéndice A

Instalación y prueba de asociación en SOLAR

Si bien SOLAR se puede instalar en los sistemas operativos más populares, es recomendable hacerlo en una computadora que tengan Linux como sistema operativo, ya que es más estable. Si su computadora no tiene este sistema operativo puede instalar una máquina virtual que le permitirá trabajar con Linux.

La distribución de SOLAR está disponible en la página web oficial *http://solar.txbiomedgenetics.org/*. Ahí encontrarás los archivos de instalación para Linux en el comprimido `solar_linux.tar.gz`. Una vez descargados y descomprimidos los archivos de instalación se va a encontrar con el archivo `README` y la secuencia de comandos `intall_solar` para instalar el programa. En particular, el script copia los archivos de biblioteca, binarios y documentación necesarios en un directorio de instalación definido por el usuario. Si la computadora esta configurada bajo un idioma diferente al inglés se puede presentar un problema para ejecutar el programa, esto se soluciona ejecutando el comando `export LC_ALL=C` para obligar a la aplicación a usar el idioma predeterminado. Una vez finalizado el de instalación se mostrará un mensaje similar al siguiente cuando SOLAR se ejecute.

```
SOLAR Eclipse version 8.4.2 (General), last updated on August 27, 2018
Developed at Maryland Psychiatric Research Center,
University of Maryland School of Medicine, Baltimore.
Visit our documentation and tutorial website www.solar-eclipse-genetics.org
Our download page https://www.nitrc.org/projects/se_linux
Our github page https://github.com/brian09/solar-eclipse
For questions email: pkochunov@gmail.com
Enter help for help, exit to exit, doc to browse documentation.
The software development is supported by NIH grant RO1EB015611
from The National Institute for Biomedical Imaging and Bioengineering.
Enter cite to see how to cite this software.

solar>
```

Solar se ejecuta sobre la terminal. Todas las funciones de solar las puede consultar en: *http://solar-eclipse-genetics.org/solar-commands.html*.

Para realizar la prueba de asociación recomiendo crear una carpeta en la cual guardemos los archivos que ocuparemos y los resultados obtenidos, para ello debemos crear la carpeta y cam-

biarnos de directorio una vez que hemos iniciado en solar con la instrucción `cd`. Como ejemplo yo cree la carpeta `Resulatados`, por lo que la instrucción se debe ejecutar `solar>cd Resultados`.

Para comenzar con la prueba tenemos que cargar el archivo de pedigrí, el de fenotipos y el de SNPs, los cuales deben ser `.csv`.

En el primero se deben incluir un *ID individual*, *ID del padre ID de la madre* y sexo, este último usualmente se codifica con 1 para hombre y 2 para mujer además de los SNPs a estudiar. La estructura de este documento se muestra en la Figura A.1. Para cargar el archivo en solar lo debemos hacer son la función `load pedigree <filename>`, la instrucción se ejecuta como `solar>load pedigree pedfile.csv`.

ID	FA	MO	SEX
100	101	102	2
101	0	0	1
102	0	0	2
200	201	202	2
201	0	0	1
202	0	0	2

Figura A.1: Ejemplo de archivo PED

En el archivo de fenotipos se deben incluir el *ID individual* y los fenotipos, nuestro caso estamos considerando un único fenotipo, padecer o no la enfermedad, el cual codificamos con 1 si la padecen y 0 si no la padecen (ver Figura A.2). Para cargar volvemos a ocupar la función `load`, la instrucción se ejecuta como `solar>load phenotypes phenofile.csv`.

ID	aff
100	0
101	0
102	1
200	0
201	0

Figura A.2: Ejemplo de archivo phenotypes

En el archivo donde se incluyen los SNPs se deben incluir los mismos campos que considera el archivo pedigrí aumentando la información de los SNPs (ver Figura A.3). La instrucción para leer el archivo es similar a la anteriores: `solar>load snp snpdata.csv`. Una vez que hemos cargado al archivo de datos de los SNPs, podemos ejecutar la instrucción `snp show` y calcula la frecuencia alélica de cada SNP en la lista.

ID	FA	MO	SEX	snp1	snp2	snp3
100	101	102	2	CT	AG	TC
101	0	0	1	CT	GG	TC
102	0	0	2	TT	AG	TT

Figura A.3: Ejemplo de archivo SNP

Para realizar el análisis de asociación debemos indicar la variable dependiente, es decir el rasgo del modelo, mediante la instrucción `trait rasgo`. Posteriormente debemos definir las covariables del modelo, en nuestro caso son los SNPs, para ello es necesario crear el archivo `snp.genocov`, el cual ingresaremos como parámetro cuando ejecutemos el análisis de asociación, este archivo se crea ejecutando la instrucción `snp covar` (para ello es necesario haber cargado

anteriormente el archivo con la información de los SNPs), esta función requiere el archivo de haplotipos cómo parámetros, pero en caso de contar con esa información se indica con la opción `-nohaplos`.
Una vez que hemos definido los SNPs como covariables y el rasgo podemos ejecutar la función `mga -files snp.genocov` la cuál realiza los cálculos del modelo de asociación. Es importante mencionar que nosotros no estamos considerando otras covariables, es caso de considerarlas se deben especificar con la función `covariate` antes de ejecutar la función `mga`. Los resultados del análisis se encuentran en una carpeta, que se crea al ejecutar la función `mga`, que lleva el nombre del rasgo. Los resultados del análisis los encontramos en el archivo `mga.out`.

La documentación completa de las funciones las encuentra en:
`http://solar-eclipse-genetics.org/solar-commands.html`

Apéndice B

Instalación y prueba de asociación en PLINK

Al igual que SOLAR recomendamos instalar PLINK en una distribución de LINUX (Ubuntu por ejemplo). Su instalación es muy sencilla. El programa se distribuye de manera gratuita y se encuentra en `http://zzz.bwh.harvard.edu/plink/download.shtml`. Una vez que haya descargado el archivo correspondiente a la plataforma de Linux, lo único que debe hacer es descomprimirlo y ejecutarlo sobre la línea de comandos.

Para comenzar con la prueba debemos ejecutar la instrucción `plink --file mydata`. Donde esperamos que en la carpeta donde se instalo plink se encuentren dos archivos: en este caso, mydata.ped y mydata.map

El archivo PED es un archivo delimitado por espacios en blanco (espacio o tabulación), donde las primeras seis columnas son obligatorias:

- ID Familiar
- ID Individual
- ID Parental
- ID Maternal
- Sexo (1 para hombre y 2 para mujer)
- Fenotipo

A diferencia del archivo PED de SOLAR, en PLINK se permite uno y solo un fenotipo en el archivo, y se deben incluir los genotipos de la columna 7 en adelante. Todos los marcadores deberán ser dialélicos y cada alelo se escribirá en una columna.

El archivo MAP también es un archivo delimitado por espacios en blanco y debe contener exactamente 4 columnas:

- Cromosoma

- rs# o el identificador del SNP
- Distancia Genética (morgans)
- Posición del par base (unidades de base)

La distancia genética se puede especificar en centimorgans con la bandera `--cm` . Alternativamente, puede usar un archivo MAP excluyendo la distancia genética agregando la bandera `--map3`, es decir, en la línea de comandos escribiremos: `plink --file mydata --map3`.

Una vez que se han creado ambos archivos solo falta ejecutar el análisis de TDT con la instrucción `plink --file mydata --tdt`, la cual generará un archivo con el nombre `plink.tdt` donde se indica el valor de la prueba estadística y el p_valor.

Bibliografía

Beaty, K. M. J., T. H. Cohen, and B. H
1993. *Fundamentos de epidemiología genética*, volume 22.

Cavalli-Sforza, L. L. and A. W. Edwards
1967. Phylogenetic analysis: models and estimation procedures. *Evolution*, 21(3):550–570.

Chen, Y.-H.
2004. New approach to association testing in case-parent designs under informative parental missingness. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*, 27(2):131–140.

Curtis, D. and P. Sham
1995. A note on the application of the transmission disequilibrium test when a parent is missing. *American Journal of Human Genetics*, 56(3):811.

Elston, R. C.
2000. Introduction and overview. *Statistical methods in medical research*, 9(6):527–541.

Evangelou, E., T. A. Trikalinos, G. Salanti, and J. P. Ioannidis
2006. Family-based versus unrelated case-control designs for genetic associations. *PLoS genetics*, 2(8).

Falconer, D.
1967. The inheritance of liability to diseases with variable age of onset, with particular reference to diabetes mellitus. *Annals of human genetics*, 31(1):1–20.

Falconer, D. and T. Mackay
1996. Introduction to quantitative genetics. essex. *UK: Longman Group*.

Flores-Alfaro, E., A. I. Burguete-García, and E. Salazar-Martínez
2012. Diseños de investigación en epidemiología genética. *Revista Panamericana de Salud Pública*, 31:88–94.

Fong, Y., H. Rue, and J. Wakefield
2010. Bayesian inference for generalized linear mixed models. *Biostatistics*, 11(3):397–412.

Gottesman, I. I. and J. Shields
1972. A polygenic theory of schizophrenia. *International Journal of Mental Health*, 1(1-2):107–115.

Hatemi, P. K., J. R. Hibbing, S. E. Medland, M. C. Keller, J. R. Alford, K. B. Smith, N. G. Martin, and L. J. Eaves
2010. Not by twins alone: Using the extended family design to investigate genetic influence on political beliefs. *American Journal of Political Science*, 54(3):798–814.

Hayeck, T. J., P.-R. Loh, S. Pollack, A. Gusev, N. Patterson, N. A. Zaitlen, and A. L. Price
2017. Mixed model association with family-biased case-control ascertainment. *The American Journal of Human Genetics*, 100(1):31–39.

Henderson, C. R.
1963. Selection index and expected genetic advance. *Statistical genetics and plant breeding*, 982:141–163.

Henry, A. and Sturtevant
1965. A history of genetics.

Horvath, S. and N. M. Laird
1998. A discordant-sibship test for disequilibrium and linkage: no need for parental data. *The American Journal of Human Genetics*, 63(6):1886–1897.

Jorde, L. B., J. C. Carey, and M. J. Bamshad
2015. *Medical genetics e-Book*. Elsevier Health Sciences.

Karki, R., D. Pandya, R. C. Elston, and C. Ferlini
2015. Defining "mutation" and "polymorphism" in the era of personal genomics. *BMC medical genomics*, 8(1):37.

Losilla Vidal, J. M.
2009. *MonteCarlo toolbox de Matlab: herramientas para un laboratorio de estadística fundamentado en técnicas Monte Carlo*. Universitat Autònoma de Barcelona,.

Lynch, M., B. Walsh, et al.
1998. *Genetics and analysis of quantitative traits*, volume 1. Sinauer Sunderland, MA.

Martin, E. R., S. A. Monks, L. L. Warren, and N. L. Kaplan
2000. A test for linkage and association in general pedigrees: the pedigree disequilibrium test. *The American Journal of Human Genetics*, 67(1):146–154.

Merriam-Webster
. https://www.merriam-webster.com/dictionary/heritability.

Montes, A. M. S., A. S. S. Rodríguez, and J. S. A. Borunda
2016. *Biología molecular*. McGraw-Hill Interamericana.

NAL
2013. National agricultural library, thesaurus.

Park, S., S. Lee, Y. Lee, C. Herold, B. Hooli, K. Mullin, T. Park, C. Park, L. Bertram, C. Lange, et al.
2015. Adjusting heterogeneous ascertainment bias for genetic association analysis with extended families. *BMC medical genetics*, 16(1):62.

Schnell, A. H. and J. S. Witte
2016. Family-based study designs. In *Molecular Epidemiology*, Pp. 33–42. CRC Press.

Searle, S. R., G. Casella, and C. E. McCulloch
2009. *Variance components*, volume 391. John Wiley & Sons.

Sevilla, S. D.
2007. Metodología de los estudios de asociación genética. *Insuficiencia cardíaca*, 2(3):111–114.

Spielman, R. S. and W. J. Ewens
1998. A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test. *The American Journal of Human Genetics*, 62(2):450–458.

Spielman, R. S., R. E. McGinnis, and W. J. Ewens
1993. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (iddm). *American journal of human genetics*, 52(3):506.

Sun, F., W. D. Flanders, Q. Yang, and M. J. Khoury
1999. Transmission disequilibrium test (tdt) when only one parent is available the 1-tdt. *American Journal of Epidemiology*, 150(1):97–104.

Thomas, D. C. et al.
2004. *Statistical methods in genetic epidemiology*. Oxford University Press.

Thompson, R.
1972. The maximum likelihood approach to the estimate of liability. *Annals of human genetics*, 36(2):221–232.

Tier, B.
1990. Computing inbreeding coefficients quickly. *Genetics Selection Evolution*, 22(4):419.

Zhou, X., P. Carbonetto, and M. Stephens
2013. Polygenic modeling with bayesian sparse linear mixed models. *PLoS genetics*, 9(2):e1003264.

Ziegler, A., I. R. König, and F. Pahlke
2010. *A Statistical Approach to Genetic Epidemiology: Concepts and Applications, with an E-learning platform*. John Wiley & Sons.