Análisis estadístico de polimorfismos genéticos en estudios epidemiológicos

Raquel Iniesta[a] **/ Elisabet Guinó**[a] **/ Víctor Moreno**[a,b]
[a]Servicio de Epidemiología y Registro del Cáncer, IDIBELL, Instituto Catalán de Oncología, L'Hospitalet de Llobregat, Barcelona. España; [b]Unidad de Bioestadística, Facultad de Medicina, Universidad Autónoma de Barcelona, Barcelona, España.

(Statistical analysis of genetic polymorphisms in epidemiological studies)

Resumen

El análisis de los polimorfismos genéticos permite identificar genes que confieren susceptibilidad a presentar enfermedades. En este trabajo se presenta la nomenclatura utilizada en la bibliografía de epidemiología genética y una estrategia básica de análisis estadístico de estudios epidemiológicos que incorporan estos marcadores. En primer lugar, se presenta el análisis descriptivo de un único polimorfismo y la evaluación del equilibrio Hardy-Weinberg. A continuación se presentan los métodos para evaluar la asociación con la enfermedad. Para ello se emplean modelos de regresión logística y se estudian los posibles modelos de herencia. Por último, se presentan métodos para el análisis simultáneo de múltiples polimorfismos: estimación de las frecuencias de haplotipos y análisis de asociación con la enfermedad.

Palabras clave: Epidemiología genética. Polimorfismo. Genotipo. Haplotipo. Análisis estadístico.

Abstract

Analysis of genetic polymorphisms allows the genes that confer susceptibility to diseases to be analyzed. This paper presents the nomenclature used in genetic epidemiology literature and a basic strategy for statistical analysis of epidemiological studies that use genetic markers. First, a descriptive analysis of a single nucleotide polymorphism is presented, with assessment of Hardy-Weinberg equilibrium. Next, methods to assess the association with disease are presented. To do this, logistic regression models are used and alternative models of inheritance are explored. Finally, methods for the simultaneous analysis of multiple polymorphisms are presented: haplotype frequency estimation and analysis of disease association.

Key words: Genetic epidemiology. Polymorphism. Genotype. Haplotype. Statistical analysis.

Introducción

Los polimorfismos genéticos son variantes del genoma que aparecen por mutaciones en algunos individuos, se transmiten a la descendencia y adquieren cierta frecuencia en la población tras múltiples generaciones. Se ha estimado que hay una variante en cada 1.000 pares de bases de los 3.000 millones que configuran el genoma humano. Los polimorfismos son la base de la evolución y los que se consolidan, bien pueden ser silentes o proporcionar ventajas a los individuos, aunque también pueden contribuir a causar enfermedades[1]. Se conocen muchas enfermedades determinadas genéticamente por mutaciones o variantes denominadas de «alta penetrancia», ya que los portadores de la variante suelen manifestar la enfermedad con una alta probabilidad. Estas variantes suelen ser de baja frecuencia en la población general, por ejemplo, las mutaciones heredadas en el gen supresor de tumores *APC* determinan la aparición de la poliposis familiar adenomatosa que a menudo degenera en carcinomas en el colon, pero esta entidad no explica más de un 1% del total de tumores de colon.

En la actualidad muchos investigadores centran sus trabajos en identificar genes con polimorfismos que se dan en la población con mayor frecuencia y que influyen en el riesgo de padecer una enfermedad, pero con baja probabilidad (son los llamados polimorfismos de «baja penetrancia»). También se les denomina variantes que confieren susceptibilidad genética a la enfermedad, y para que dicha variante genética se expre-

Correspondencia: Dr. Víctor Moreno.
Servicio de Epidemiología y Registro del Cáncer.
Instituto Catalán de Oncología.
Gran Vía, km 2,7. 08970 L'Hospitalet de Llobregat. Barcelona. España.
Correo electrónico: v.moreno@iconcologia.net

Recibido: 5 de octubre de 2004. *Aceptado:* 12 de enero de 2005.

se a menudo es necesaria la participación de una exposición[2]. Un ejemplo son las variantes nulas (así llamadas porque anulan la función) en genes que codifican enzimas glutatión-S-transferasas *(GSTT1, GSTM1)*. Los individuos fumadores y portadores de la variante nula podrían tener un riesgo aumentado de padecer cáncer de pulmón o de vejiga, posiblemente por ser incapaces de metabolizar los carcinógenos del tabaco[3-5], aunque estos hallazgos no siempre son consistentes[6].

Los polimorfismos más frecuentes son cambios de una única base. A éstos se les llama polimorfismos de un único nucleótido (*single nucleotide polymorphism* [SNP], pronunciado «esnip»). Por ejemplo, en el gen de la apolipoproteína E *(ApoE)* se han descrito varios polimorfismos frecuentes que consisten en cambios de una única base. Uno de ellos, denominado *ApoE* ε-4, resulta en un cambio en el aminoácido cisteína de la posición 112 por una arginina. Esta variante se asocia con la enfermedad de Alzheimer[7]. Otros polimorfismos son repeticiones, en un número variable de veces, de una secuencia corta (*variable number tandem repeat* [VNTR]). Por ejemplo, los individuos afectados por ataxia de Fredreich, una enfermedad autosómica recesiva, son portadores de variantes en el gen *frataxin* con un número elevado repeticiones del triplete GAA en el primer intrón. Los individuos normales suelen tener menos de 40 repeticiones, mientras que los afectados tienen entre 100 y 1.700 repeticiones[8]. En otras ocasiones, los polimorfismos se deben a deleciones o inserciones de secuencias cortas de nucleótidos. El cambio de un único nucleótido, si ocurre en una zona codificante como en el ejemplo de la *ApoE*, puede provocar un cambio de aminoácido en la proteína resultante, y ello puede resultar en una modificación de su actividad o función. Los cambios también pueden ocurrir en zonas del promotor de un gen y modificar su expresión. Estas zonas promotoras modulan el proceso de transcripción del ADN en ARN (la transcripción es el primer paso de la decodificación de un gen a una proteína). Lo mismo puede ocurrir si el cambio se produce en un intrón, como el ejemplo de la ataxia de Friedrich[8]. Aunque los intrones no se traducen a proteína, cambios en su estructura pueden modular la expresión del gen. Otras veces, probablemente la mayoría, los cambios son silentes y no tienen repercusiones funcionales. Mientras que sólo estudios moleculares específicos pueden poner de manifiesto si los polimorfismos son funcionales, los estudios epidemiológicos son fundamentales para valorar si hay efectos en la salud de la población[1,9,10].

Cuando el objetivo de un estudio es identificar un polimorfismo o variante en un gen que esté relacionado con una enfermedad se pueden emplear diferentes estrategias. En primer lugar, es importante obtener evidencia de que al menos una fracción de la enfermedad está determinada genéticamente. Para ello son útiles los estudios de agregación familiar, los de gemelos o los de emigrantes. En segundo lugar, hay que identificar dónde están los genes de interés para la enfermedad. En esta fase se realizan estudios denominados de ligamiento *(linkage)*, que emplean como marcadores genéticos una serie de polimorfismos repartidos por todo el genoma. En estos estudios se suelen emplear familias grandes con varios miembros afectados y sus análisis permiten identificar zonas del genoma de interés, pero tienen poca resolución. En esas zonas identificadas puede haber centenares de genes interesantes y miles de polimorfismos candidatos. Para identificar con mayor precisión los genes de interés y, dentro de esos genes, el o los polimorfismos responsables, se emplean estudios de asociación, en los que se compara la frecuencia relativa de las diferentes variantes de una serie de polimorfismos entre los individuos afectados y un grupo control adecuado. Estos estudios suelen seleccionar «genes candidatos» (aquellos cuya función puede estar relacionada con la enfermedad de interés), y dentro de esos genes se busca como marcadores genéticos a determinados polimorfismos, normalmente de tipo SNP, repartidos a lo largo del gen.

En cuanto a la metodología de estudio, se suelen emplear diseños epidemiológicos clásicos basados en individuos no relacionados, como estudios de casos y controles o de cohortes. También se pueden emplear diseños basados en familias, en los que los individuos de control son parientes de los casos, como los diseños de casos y hermanos sanos o tríos (caso y padres)[11-13].

En esta revisión presentaremos la estrategia habitual de análisis de estudios basados en diseños con individuos no relacionados que incluyen información de polimorfismos. Para simplificar, en los ejemplos trataremos el caso de polimorfismos tipo SNP con una única variante, pero los métodos sirven para marcadores más complejos. En primer lugar revisaremos el análisis de un único polimorfismo y, a continuación, el análisis simultáneo de múltiples polimorfismos. Para una revisión de la nomenclatura empleada en epidemiología genética es recomendable la introducción de Elston[14].

Análisis descriptivo de un polimorfismo

Un polimorfismo se caracteriza porque diferentes individuos presentan distintos nucleótidos o variantes en una posición concreta del genoma, que se denomina *locus*. A cada posible variante se le denomina alelo. Si se trata de un SNP, normalmente serán 2 los posibles alelos en un *locus*: por ejemplo, el cambio de T por C (T > C). Si el *locus* corresponde a un cromosoma autosómico (del 1 al 22), cada individuo es portador de 2 alelos, uno en cada copia del cromosoma, que se he-

redan del padre y madre de manera independiente. La pareja de alelos observada en un individuo se denomina genotipo y, para el *locus* T > C del ejemplo, las 3 posibilidades de parejas de alelos son: TT, TC y CC. Los individuos con los 2 alelos idénticos, sean TT o CC, se denominan homocigotos y los que tienen diferentes alelos (TC), heterocigotos. En general se considera variante al alelo menos frecuente, pero esto puede diferir de una población a otra.

La descripción estadística de un polimorfismo consiste, en primer lugar, en estimar la prevalencia en la población de cada alelo y de cada genotipo posible, lo que en nomenclatura genética se denomina estimar las frecuencias alélicas y genotípicas, respectivamente. En general, las técnicas de laboratorio permiten determinar el genotipo de cada individuo. Las frecuencias genotípicas, por tanto, se estiman directamente calculando la proporción de individuos con cada genotipo. Para estimar las frecuencias alélicas simplemente se duplica la muestra tomando como unidad de observación el cromosoma (cada individuo contribuye con 2 cromosomas) y se calcula la proporción de cada alelo. Por ejemplo, si en una muestra de 200 individuos observamos los siguientes genotipos:

110 TT, 75 TC y 15 CC,

las frecuencias genotípicas serán:

0,55 TT, 0,38 TC y 0,07 CC,

y las frecuencias alélicas serán:

0,74 T y 0,26 C [0,74 = (110 (2 + 75)/(200 (2)]

Equilibrio de Hardy-Weinberg

El principio de equilibrio de Hardy-Weinberg determina qué frecuencias deben observarse en la población para cada genotipo en función de las frecuencias de los alelos. En condiciones habituales, si la transmisión de los alelos de los progenitores a los descendientes es independiente y no ocurren fenómenos distorsionadores, como la aparición frecuente de nuevas mutaciones o la selección de alelos, la probabilidad de observar una combinación de alelos concreta (un genotipo) depende del producto de las probabilidades (frecuencias) de cada alelo. En nuestro ejemplo, si llamamos p a la frecuencia de T, el primer alelo, y q a la frecuencia de C, el segundo (suponiendo que sólo hay 2 alelos posibles), las frecuencias esperadas de cada genotipo son:

Np^2 TT, $2Npq$ TC y Nq^2 CC,

donde N es el tamaño de muestra. Si en nuestro ejemplo sustituimos p y q por los valores estimados:

0,74 y 0,26 respectivamente, las frecuencias esperadas serán:

109,52 TT, 76,96 TC y 13,52 CC.

Estas frecuencias esperadas se pueden comparar con las observadas utilizando el test de la χ^2 $HW = \Sigma (O-E)^2/E$, con 1 grado de libertad. Para el ejemplo HW vale 0,20, que corresponde a una p de 0,64, por lo que es compatible con el equilibrio de Hardy-Weinberg.

Antes de realizar un análisis de asociación se debe comprobar si se cumple el principio de equilibrio de Hardy-Weinberg en la muestra de controles (como representantes de la población general). En el caso de que se observara una desviación del equilibrio se debería revisar el método de genotipificación, pues en ocasiones se producen sesgos al interpretar los resultados por ser más fácil de detectar un genotipo que otros. Otras posibilidades son que los individuos no sean independientes (p. ej., por consanguinidad) o que se dé una selección de un alelo (p. ej., por estar asociado con la longevidad). Tampoco debe olvidarse que si empleamos un nivel de significación del 5%, por azar puede observarse falta de ajuste al esperado, aunque la condición de transmisión de alelos con independencia sea correcta en la población del estudio. En la muestra de casos es posible que no se cumpla el equilibrio de Hardy-Weinberg; ello puede ser indicativo de que el polimorfismo pueda estar asociado con la enfermedad.

Análisis de asociación de un polimorfismo con la enfermedad

Desde el punto de vista estadístico, un polimorfismo constituye una variable categórica con varios genotipos posibles y se suele considerar como categoría de referencia al grupo de individuos homocigotos para el alelo más frecuente. Para evaluar la asociación de un polimorfismo con la enfermedad se construye la tabla de contingencia correspondiente y se puede contrastar la hipótesis de asociación mediante un test de la χ^2. También se pueden calcular las *odds ratios* (OR) de cada genotipo respecto de la referencia para cuantificar la magnitud de la asociación.

Si es necesario ajustar los análisis por posibles variables de confusión, entonces es preferible emplear modelos de regresión logística por su versatilidad. Además, estos modelos permiten evaluar fácilmente si hay interacciones entre el polimorfismo y otros factores.

Llamaremos ahora p a la probabilidad de ser caso, G al polimorfismo (que codificará los diferentes geno-

tipos: TT, TC, CC) y Z a una o más variables por las que se desea ajustar el modelo. El modelo logístico se define por la ecuación:

$$\log[p/(1-p)] = \alpha + \beta G + \gamma Z$$

donde α, β y γ son parámetros estimados.

Supongamos que el polimorfismo G es un SNP en el que el alelo variante C modifica el riesgo de la enfermedad de interés. Ya que cada individuo posee una pareja de alelos, el riesgo asociado con cada genotipo puede depender del número de copias de C, lo que permite definir varios modelos de herencia posibles cuya verosimilitud se puede explorar mediante una adecuada codificación de los genotipos (tabla 1).

Los 4 modelos principales de herencia posibles, son:

Modelo dominante. Supone que una única copia de C es suficiente para modificar el riesgo y que ser portador de 2 copias lo modifica en igual magnitud; es decir, heterocigotos TC y homocigotos CC tienen el mismo riesgo. Se puede comparar la combinación de estos 2 genotipos respecto a los homocigotos TT:

$$\log[p/(1-p)] = \alpha + \beta Do + \gamma Z$$

Modelo recesivo. Supone que son necesarias 2 copias de C para modificar el riesgo; por tanto, heterocigotos TC y homocigotos del alelo más frecuente TT tienen el mismo riesgo. Se compara la combinación de ellos respecto a los homocigotos del alelo variante CC:

$$\log[p/(1-p)] = \alpha + \beta Re + \gamma Z$$

Modelo aditivo. Supone que cada copia de C modifica el riesgo en una cantidad aditiva (en escala logit); por tanto, los homocigotos CC tienen el doble de riesgo que los heterocigotos TC. Se compara la combinación ponderada, donde se da peso 1 a los heterocigotos TC y peso 2 a los homocigotos CC:

$$\log[p/(1-p)] = \alpha + \beta Ad + \gamma Z$$

Tabla 1. Codificación de variables indicadoras para evaluar diferentes modelos de herencia

	Codominante		Dominante	Recesivo	Aditivo
Genotipo[a]	He	Va	Do	Re	Ad
TT	0	0	0	0	0
TC	1	0	1	0	1
CC	0	1	1	1	2

[a]Genotipos posibles para un polimorfismo en un *locus* bialélico T > C

Modelo codominante. Es el más general. Cada genotipo proporciona un riesgo de enfermedad diferente y no aditivo. Se comparan heterocigotos (He) y homocigotos variantes (Va) por separado respecto a los homocigotos del alelo más frecuente. Este modelo emplea 2 coeficientes (grados de libertad).

$$\log[p/(1-p)] = \alpha + \beta_1 He + \beta_2 Va + \gamma Z$$

No es fácil hallar un criterio para establecer el modelo de herencia más adecuado para un polimorfismo concreto. Habitualmente se suele comparar el ajuste del modelo codominante, que es el más general (2 parámetros), con los demás modelos (1 parámetro). Estas comparaciones pueden realizarse mediante el test de la razón de verosimilitudes. Aun así, a menudo no es posible diferenciar entre varios modelos y se elige el que tenga menor valor del criterio de información de Akaike (AIC = –2log[L] + #parámetros), donde L es la verosimilitud del modelo. Este criterio pondera el ajuste del modelo (–2logL) con la complejidad (número de parámetros). Los resultados obtenidos al aplicar este criterio, puramente estadístico, deben considerarse provisionales. Lo adecuado sería que obtuviesen validación en el laboratorio y que fuesen replicados por otros estudios epidemiológicos.

Los modelos de interacción añaden un término producto entre el genotipo y una variable ambiental:

$$\log[p/(1-p)] = \alpha + \beta G + \gamma Z + \delta G \times Z$$

También pueden incluir el producto de 2 genotipos si se exploran interacciones gen-gen:

$$\log[p/(1-p)] = \alpha + \beta_1 G_1 + \beta_2 G_2 + \gamma Z + \delta G_1 \times G_2$$

A partir de los coeficientes β y δ de los modelos se puede calcular las OR de asociación entre cada genotipo y la enfermedad y los correspondientes intervalos de confianza del 95%.

Algunos investigadores analizan el riesgo comparando la distribución de alelos entre casos y controles. Para ello, cada individuo contribuye con 2 observaciones en la muestra, una por cada cromosoma, y se compara el riesgo del alelo variante respecto al más frecuente. Este análisis sólo es correcto en el caso de que la distribución de los alelos sea independiente en la población (equilibrio de Hardy-Weinberg en controles). El análisis equivalente que es correcto en cualquier ocasión es el modelo aditivo de genotipos[15,16].

Ejemplo de análisis

Supongamos que se ha realizado un estudio de casos y controles en el que se evalúa si un polimorfis-

Tabla 2. Análisis del riesgo de un polimorfismo en función del modelo de herencia

Modelo[a]	Genotipo[b]	Controles		Casos		OR	IC del 95%	*p*[c]	ΔCo[d]
		N	%	N	%				
	TT	210	65,0	225	62,2	1		0,07	
Co	TC	104	32,2	114	31,5	1,02	0,74-1,42		
	CC	9	2,8	23	6,4	2,38	1,09-5,22		
Do	TT	210	65,0	225	62,1	1		0,43	0,034
	TC-CC	113	34,9	137	37,8	1,13	0,83-1,55		
Re	TT-TC	314	97,2	339	93,6	1		0,024	0,89
	CC	9	2,79	23	6,3	2,37	1,09-5,14		
Ad						1,21	0,93-1,57	0,14	0,08

[a]Modelos de herencia: codominante (Co), dominante (Do), recesivo (Re), aditivo (Ad). [b]Genotipos y sus agrupaciones para un polimorfismo en un *locus* bialélico T > C. [c]p del test de asociación (comparación con el modelo nulo). [d]p del test que compara con el modelo codominante.

mo en un gen relacionado con la reparación del ADN modifica el riesgo de padecer cáncer colorrectal. En la tabla 2 se presentan los 4 modelos principales de riesgo posibles. El modelo codominante muestra que sólo los individuos CC tienen un riesgo aumentado. Los heterocigotos TC tienen un riesgo similar a los homocigotos TT, lo que sugiere que el modelo recesivo es el adecuado.

Análisis simultáneo de múltiples *loci*

Con frecuencia son varios los polimorfismos que se analizan simultáneamente en un gen o región candidata de un gen. El motivo es que el polimorfismo realmente responsable de influir o modificar el riesgo de la enfermedad puede ser desconocido; por ello se analizan varios polimorfismos para intentar identificarlo. Entre diferentes polimorfismos localizados en el mismo cromosoma y relativamente próximos entre sí suele observarse cierto grado de correlación o asociación estadística denominada desequilibrio de ligamiento *(linkage disequilibrium)*. Ello es debido a que en el proceso de meiosis que genera los gametos, los cromosomas que se transmitirán serán copias exactas de los del progenitor, a excepción de los entrecruzamientos que generan recombinación. Es decir, cada cromosoma transmitido a la descendencia estará formado por una composición de fragmentos largos que son una copia exacta de los del progenitor, pero combinando partes del cromosoma paterno y del materno. La frecuencia de entrecruzamientos por cromosoma es pequeña, de 1 a 4, y depende de su tamaño. La probabilidad de que entre 2 *loci* cercanos se dé una recombinación es baja y, por ello, se observa el desequilibrio de ligamiento, que tiende a disminuir en sucesivas generaciones hasta llegar al equilibrio (independencia estadística).

El desequilibrio de ligamiento es muy útil, pues permite localizar polimorfismos relacionados con la enfermedad. Si aparece una mutación que genera un polimorfismo responsable de la enfermedad, es posible que otros polimorfismos cercanos también estén asociados con ella. De hecho, como lo que se transmite de padres y madres a sus hijos son cromosomas, suele ser interesante identificar el conjunto de alelos que se transmiten conjuntamente en cada cromosoma, de manera que sea más fácil así identificar el polimorfismo causal. A este conjunto de alelos que se transmiten conjuntamente se le denomina haplotipo. Un individuo, para un conjunto de *loci* cercanos, posee 2 haplotipos, cada uno en un cromosoma. Identificar los haplotipos a partir de los genotipos de cada *locus* suele ser fácil, aunque hay algunos casos excepcionales. En la tabla 3 se presentan los haplotipos para el caso de 2 *loci* bialélicos, es decir, 2 polimorfismos cercanos, cada uno de ellos con 2 alelos posibles, el primero T > C y el segundo A > G. De las 9 combinaciones de genotipos po-

Tabla 3. Posibles haplotipos en función de los genotipos de 2 *loci* bialélicos

Genotipos T > C / A > G	TT	TC	CC
AA	T A + T A	T A + C A	C A + C A
AG	T A + T G	T A + C G T G + C A	C A + C G
GG	T G + T G	T G + C G	C G + C G

Un individuo es portador de una pareja de haplotipos. Para los individuos con ambos *loci* heterocigotos (TC/AG) no es posible identificar las parejas de haplotipos.

sibles, todas ellas permiten identificar los haplotipos a excepción de una, que corresponde al caso en que ambos *loci* sean heterocigotos. Por ejemplo, si los genotipos para un individuo son TC y AA, sus haplotipos serán T-A y C-A, es decir, el individuo será portador de una pareja de cromosomas, con cada una de estas combinaciones de alelos. En el caso de que los 2 genotipos sean heterocigotos: TC y AG, la combinación de alelos en cromosomas puede ser T-A y C-G o bien T-G y C-A. No se puede saber qué combinación es la correcta para un individuo si no se conocen los genotipos de los progenitores o se emplean técnicas de laboratorio muy sofisticadas que permitan identificar haplotipos. En la práctica, para realizar análisis estadísticos de asociación se recurre a métodos de estimación que tienen en cuenta variables con incertidumbre para resolver este problema.

Análisis descriptivo de haplotipos

La estimación de las frecuencias para cada haplotipo sería sencilla si no se dieran casos de incertidumbre, es decir, casos en que no es posible determinar la pareja de haplotipos que lleva el individuo debido a que éste tenga 2 o más *loci* heterocigotos. Si, además, hay valores perdidos en la determinación de alguno de los genotipos, la incertidumbre aumenta. Para estimar la frecuencia de los haplotipos en estos casos hay varias posibilidades:

1. Métodos basados en reglas. Consisten en observar la distribución de alelos en los haplotipos que no presentan incertidumbre e imputar los casos con incertidumbre a aquellas combinaciones observadas con más frecuencia[17]. El principal inconveniente de este método es que la solución puede no ser única en función de cómo se inicie el algoritmo.

2. Estimación estadística mediante el algoritmo esperanza-maximización[18] (EM). Éste es un método estadístico genérico[19] para tratar variables no observadas (en este caso, la proporción de individuos con cada posible haplotipo en caso de incertidumbre). El algoritmo itera repetidamente entre 2 pasos hasta conseguir convergencia. En primer lugar, se usan unas frecuencias iniciales, no necesariamente correctas, para cada posible haplotipo. En el supuesto de que esas frecuencias fuesen correctas, y asumiendo que hay equilibrio de Hardy-Weinberg, se podría calcular la frecuencia esperada de cada combinación de genotipos con incertidumbre (paso E). Con las frecuencias de cada combinación de genotipos se pueden obtener las frecuencias de cada haplotipo maximizando la función de verosimilitud de los haplotipos (paso M). Este paso consiste en contar los haplotipos compatibles con cada combinación de genotipos. El algoritmo converge a valores estables de frecuencia de combinaciones de genotipos con incertidumbre y de frecuencia de haplotipos, que es nuestro objetivo. Para los individuos con incertidumbre, a cada pareja de haplotipos posible se le puede calcular una probabilidad. Una limitación de este método es que no proporciona directamente varianzas de las estimaciones y, si se analizan simultáneamente muchos polimorfismos (más de 20) con un alto grado de incertidumbre, pueden obtenerse soluciones incorrectas, pues el algoritmo puede tender hacia un máximo local. En estas situaciones es importante repetir el algoritmo empleando diferentes valores iniciales de las frecuencias de haplotipos. En nuestra experiencia, para el análisis simultáneo de menos de 10 polimorfismos este método siempre ha proporcionado soluciones coherentes.

3. Estimación estadística mediante métodos Montecarlo basados en cadenas de Markov (MCMC). Se trata de una aproximación bayesiana al problema que ha permitido mejorar los resultados en caso de analizar muchos polimorfismos[20]. Este método proporciona como información adicional interesante una muestra aleatoria de la distribución a posteriori de la frecuencia de cada haplotipo, que puede ser empleada para calcular la varianza, además del valor esperado. De esta manera podemos conocer directamente la precisión de la estimación de la frecuencia de cada haplotipo.

Análisis de la asociación de los haplotipos con la enfermedad

Cualquiera de los métodos anteriores permite identificar o estimar, para cada individuo, la pareja de haplotipos que posee en función de los genotipos. Estos haplotipos pueden analizarse entonces en relación con la enfermedad mediante modelos de regresión logística. Normalmente se realiza un análisis de cromosomas y no de individuos, es decir, se duplica la muestra, de manera que cada individuo contribuye con 2 observaciones, una para cada haplotipo, y se compara el riesgo asociado con los diversos haplotipos entre sí, tomando como referencia el más frecuente. Este análisis es más sencillo de analizar e interpretar que el de parejas de haplotipos y no parece que requiera la realización ajustes, pues los haplotipos se heredan de manera independiente; no obstante, este aspecto no está completamente estudiado.

En caso de incertidumbre por múltiples heterocigotos o valores perdidos, los métodos basados en el algoritmo EM o MCMC proporcionan una lista de parejas de haplotipos compatibles, cada una con una probabilidad de aparición asociada. Entonces, estos individuos contribuyen en los datos con más de 2 observaciones

Tabla 4. Asociación de cada polimorfismo con la enfermedad

Locus[a]	Genotipo[b]	Controles		Casos		OR	IC del 95%	p[c]
		N	%	N	%			
A	GG	284	96,3	317	96,6	1		0,80
	GA/AA	11	3,7	11	3,6	0,90	0,38-2,10	
B	TT	278	96,5	307	95,6	1		0,57
	TG/GG	10	3,5	14	4,4	1,27	0,55-2,90	
C	AA	158	49,2	141	39,1	1		0,0099
	AC	137	42,7	172	47,7	1,41	1,02-1,94	
	CC	26	8,1	48	13,3	2,07	1,22-3,51	0,002[d]
D	TT	126	39,4	112	31,7	1		0,11
	TC	139	43,4	169	47,9	1,37	0,97-1,92	
	CC	55	17,2	72	20,4	1,47	0,95-2,27	0,038[e]

[a]Los *loci* están ordenados según su posición en el genoma. [b]Para los *loci* A y B se emplean modelos asumiendo un efecto dominante, pues sólo se ha observado 1 caso homocigoto variante. [c]p para el test de asociación. [d]p para el modelo aditivo. [e]p para el modelo dominante.

y se emplean las probabilidades de aparición de cada haplotipo como pesos en el modelo de regresión logística. El análisis combinado de genotipos y haplotipos suele ser más informativo que cualquiera de los 2 aislados, aunque hay cierta discusión sobre qué análisis es preferible en cada situación[21,22].

Ejemplo de análisis de genotipos y haplotipos

Supongamos un estudio de casos y controles sobre cáncer colorrectal en que se analizan 4 polimorfismos en un gen relacionado con el metabolismo de carcinógenos. En la tabla 4 se muestra el análisis de cada polimorfismo por separado para el modelo codominante. Los polimorfismos en los *loci* A y B son raros y sólo se ha observado un caso homocigoto variante para cada uno de ellos, por lo que se analizan con el modelo dominante. El polimorfismo C muestra una asociación fuerte con la enfermedad y las OR crecientes indican que el modelo aditivo es el que mejor ajusta. El polimorfismo D también está asociado con la enfermedad, pero en menor magnitud, y como los OR para heterocigotos y homocigotos variantes son similares, el modelo dominante es el más verosímil.

El análisis de haplotipos que se muestra en la tabla 5 revela que, aunque podría haber $2^4 = 16$ combinaciones de alelos diferentes, sólo se observan 7. De hecho, 2 haplotipos acumulan una frecuencia del 92%. En los análisis de asociación, los haplotipos raros normalmente se agrupan en una categoría. El punto de corte para definir un haplotipo raro suele depender del tamaño de muestra global. En nuestro caso, con unos 700 individuos limitaremos el análisis a los 5 haplotipos con frecuencia superior al 1%.

El haplotipo 1, el más frecuente, se toma como referencia. El segundo haplotipo en frecuencia contiene variantes para los polimorfismos C y D, simultáneamente. La correlación (desequilibrio de ligamiento) entre estos 2 polimorfismos es muy alta. Este haplotipo es el único que muestra un riesgo aumentado significativo y recoge la información observada en el análisis de genotipos por separado. Sin embargo, el haplotipo 3, con una frecuencia del 4%, sólo difiere del más frecuente por la variante en el polimorfismo D, y este haplotipo muestra una asociación inversa, no significativa, con la enfermedad. Ello permite deducir que el aumento de riesgo observado en el análisis del polimorfismo D es espúreo, reflejo de la correlación (desequilibrio de ligamiento) con el polimorfismo C. Se trata de un fenómeno similar al de la confusión en otros contextos de análisis epide-

Tabla 5. Análisis de frecuencias de haplotipos y su asociación con la enfermedad

Haplotipo	*Locus* A	B	C	D	fr[a]	OR	IC del 95%	p[b]
Hap 1	G	T	A	T	0,59	1		
Hap 2	G	T	C	C	0,33	1,50	1,15-1,96	0,003
Hap 3	G	T	A	C	0,04	0,59	0,31-1,10	0,095
Hap 4	G	G	A	C	0,019	1,63	0,65-4,10	0,30
Hap 5	A	T	A	C	0,018	0,88	0,37-2,10	0,78
Hap 6	G	T	C	T	0,002	–		
Hap 7	G	G	A	T	0,002	–		

[a]Frecuencia estimada de cada haplotipo.[b]p del test que compara cada haplotipo con frecuencia superior al 1% con el haplotipo más frecuente (Hap 1).

miológico. Otra manera de confirmar esta observación es realizar un análisis de los genotipos del polimorfismo D ajustado por el C. La asociación desaparece (OR modelo dominante = 0,99; IC del 95%, 0,55-1,77; p = 0,99), mientras que el análisis del polimorfismo C ajustado por el D mantiene los estimadores y la significación estadística. En este ejemplo, los polimorfismos A y B son poco frecuentes y apenas aportan información.

El hecho de que el polimorfismo C domine la asociación no permite concluir que es el realmente causante de modificar el riesgo de la enfermedad. Para ello sería necesario contar con información de estudios funcionales que demostraran un efecto biológico diferente en la variante en comparación con el alelo más frecuente. La asociación observada podría ser debida a otro polimorfismo no estudiado cercano al C que estuviese altamente correlacionado. Es evidente que si no contáramos con la información del polimorfismo C y tuviésemos sólo la del D podríamos caer en la tentación de imputar el efecto a D, algo que ahora sabemos que es falso.

Software para análisis de genotipos y haplotipos

Los análisis de este trabajo se han realizado con el paquete *haplo.Stats*[23] del software de uso libre *R*[24]. La función haplo.em permite estimar las frecuencias de haplotipos mediante el algoritmo EM. La función haplo.glm permite ajustar modelos de regresión logística empleando las probabilidades de cada pareja de haplotipos con incertidumbre como pesos en el modelo. En R hay otros paquetes útiles para análisis de epidemiología genética (genetics, hapassoc y gap) y casi todos emplean una adaptación del programa SNPHAP de Clayton[25]. Otros programas útiles de uso libre son Arlequin[26], Haploview[27] y EH[28] y se puede encontrar un listado exhaustivo de software para análisis genéticos en la Universidad Rockefeller[29].

Agradecimientos

Este trabajo ha recibido financiación del FIS (expedientes 96/0797, 03/0114) y del ISCIII (redes de centros C03/09 y C03/10).

Bibliografía

1. Guttmacher AE, Collins FS. Genomic medicine–a primer. N Engl J Med. 2002;347:1512-20.
2. Porta M. The genome sequence is a jazz score. Int J Epidemiol. 2003;32:29-31.
3. McWilliams JE, Sanderson BJ, Harris EL, Richert-Boe KE, Henner WD. Glutathione S-transferase M1 (GSTM1) deficiency and lung cancer risk. Cancer Epidemiol Biomarkers Prev. 1995; 4:589-94.
4. Sorensen M, Autrup H, Tjonneland A, Overvad K, Raaschou-Nielsen O. Glutathione S-transferase T1 null-genotype is associated with an increased risk of lung cancer. Int J Cancer. 2004;110:219-24.
5. Hung RJ, Boffetta P, Brennan P, et al. GST, NAT, SULT1A1, CYP1B1 genetic polymorphisms, interactions with environmental exposures and bladder cancer risk in a high-risk population. Int J Cancer. 2004;110:598-604.
6. Benhamou S, Lee WJ, Alexandrie AK, Boffeta P, Bonchardy C, Butkiewicz P, et al. Meta- and pooled analyses of the effects of glutathione S-transferase M1 polymorphisms and smoking on lung cancer risk. Carcinogenesis. 2002;23:1343-50.
7. Strittmatter WJ, Weisgraber KH, Huang DY, Dang LM, Salvesen GS, Pericak-Yance, et al. Binding of human apolipoprotein E to synthetic amyloid beta peptide: isoform-specific effects and implications for late-onset Alzheimer disease. Proc Natl Acad Sci USA. 1993;90:8098-102.
8. Pandolfo M. Friedreich's ataxia: clinical aspects and pathogenesis. Semin Neurol. 1999;19:311-21.
9. Caporaso NE. Why have we failed to find the low penetrance genetic constituents of common cancers? Cancer Epidemiol Biomarkers Prev. 2002;11:1544-9.
10. Tabor HK, Risch NJ, Myers RM. Opinion: candidate-gene approaches for studying complex genetic traits: practical considerations. Nat Rev Genet. 2002;3:391-7.
11. Cardon LR, Bell JI. Association study designs for complex diseases. Nat Rev Genet. 2001;2:91-9.
12. Zhao H. Family-based association studies. Stat Methods Med Res. 2000;9:563-87.
13. Gauderman WJ, Witte JS, Thomas DC. Family-based association studies. J Natl Cancer Inst Monogr. 1999;26:31-7.
14. Elston RC. Introduction and overview. Statistical methods in genetic epidemiology. Stat Methods Med Res. 2000;9:527-41.
15. Schaid DJ. Disease-marker association. En: Elston RC, Olson JM, Palmer L, editors. Biostatistical genetics and genetic epidemiology Chichester: Wiley; 2002. p. 206-17.
16. Clayton D. Population association. En: Balding DJ, Bishop M, Cannings C, editors. Handbook of statistical genetics. Chichester: Wiley; 2001. p. 519-40.
17. Clark AG. Inference of haplotypes from PCR-amplified samples of diploid populations. Mol Biol Evol. 1990;7:111-22.
18. Excoffier L, Slatkin M. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. Mol Biol Evol. 1995;12:921-7.
19. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc Ser B. 1997;39:1-38.
20. Stephens M, Smith NJ, Donnelly P. A new statistical method for haplotype reconstruction from population data. Am J Hum Genet. 2001;68:978-89.
21. Clayton D, Chapman J, Cooper J. Use of unphased multilocus genotype data in indirect association studies. Genet Epidemiol. 2004;27:415-28.
22. Cordell HJ, Clayton DG. A unified stepwise regression procedure for evaluating the relative effects of polymorphisms within a gene using case/control or family data: application to HLA in type 1 diabetes. Am J Hum Genet. 2002;70:124-41.
23. Lake SL, Lyon H, Tantisira K, et al. Estimation and tests of haplotype-environment interaction when linkage phase is ambiguous. Hum Hered. 2003;55:56-65.
24. The R Project [Consultado 5 Ene 2005]. Disponible en: http://www.r-project.org

25. SNPHAP. A program for estimating frequencies of large haplotypes of SNPs. Cambridge: Department of Medical Genetics, Cambridge Institute for Medical Research [Consultado 5 Ene 2005]. Disponible en: http://www-gene.cimr.cam.ac.uk/clayton/software/snphap.txt
26. Schneider S, Roessli D, Excoffier L. Arlequin: a software for population genetics data analysis. Ver 2.000. Geneva: Genetics and Biometry Lab, Department of Anthropology, University of Geneva [Consultado 5 Ene 2005]. Disponible en: http://lgb.unige.ch/arlequin/
27. Barrett JC, Fry B, Maller J, Daly MJ. Haploview: analysis and visualization of LD and haplotype maps. [Consultado 5 Ene 2005]. Disponible en: http://www.broad.mit.edu/personal/jcbarret/haploview/
28. EH linkage analysis software. New York: Rockefeller University [Consultado 5 Ene 2005]. Disponible en: http://linkage.rockefeller.edu/ott/eh.htm
29. An alphabetic list of genetic analysis software. New York: Rockefeller University [Consultado 30 Sep 2004]. Disponible en: http://linkage.rockefeller.edu/soft