Estudios de asociación en enfermedades complejas: problemas estadísticos relacionados con el análisis de polimorfismos genéticos

A. Salas[a] y Á. Carracedo[b]

[a] *Unidade de Xenética. Instituto de Medicina Legal. Facultad de Medicina. Universidad de Santiago de Compostela.* [b] *Grupo de Medicina Xenómica. Hospital Clínico Universitario. Santiago de Compostela. A Coruña.*

Los resultados positivos de asociación entre un polimorfismo genético y una determinada enfermedad compleja observados en los estudios de asociación caso-control raramente se replican. Frecuentemente, el diseño de estos estudios carece además de poder suficiente para detectar la asociación. En consecuencia, la prevalencia de errores tipo I y II en la literatura científica en el estudio de enfermedades complejas es extremadamente alta. Las causas que subyacen a esta desafortunada situación son muchas y variadas: estratificación poblacional, corrección por hipótesis múltiple, sesgo de publicación, etc. Esta revisión reflexiona sobre estos problemas y los procedimientos que ayudarían a minimizar sus efectos.

Salas A, Carracedo Á. Estudios de asociación en enfermedades complejas: problemas estadísticos relacionados con el análisis de polimorfismos genéticos. Rev Clin Esp. 2007;207(11):563-5.

Studies of association in complex diseases: Statistical problems related to the analysis of genetic polymorphisms

The positive results on the association between a genetic polymorphism and a specific complex disease -in population-based association studies (e.g. case-control) are rarely replicated in independent studies. Furthermore, the design of these studies often lacks sufficient power to detect the association. Therefore, the «prevalence» of both type I and II errors in the scientific literature dealing with the study of complex diseases is extremely high. There are many potential problems that underlie this current, unfortunate situation such as the effect of population stratification, deficient correction for multiple tests or publication bias, among others. The present review deals with all these problems and provides guidelines that can help to minimize their effects.

Introducción

La mayor parte de las enfermedades comunes son multifactoriales y están causadas por un número indeterminado de factores ambientales y/o genéticos. Este mecanismo contrasta con el modelo de enfermedad mendeliana, enfermedades de herencia de monogénica en donde generalmente la presencia o ausencia de un alelo causal predice completamente la presencia o ausencia de la enfermedad. Para las enfermedades complejas los alelos de riesgo son menos deterministas y el riesgo se mide de manera más probabilística. De acuerdo a la hipótesis «variante común-enfermedad común» los alelos de baja penetrancia responsables de enfermedades comunes se presentan con una frecuencia elevada en la población. El estudio de polimorfismos genéticos responsables de la susceptibilidad a enfermedades humanas complejas es un campo de interés creciente en la investigación genómica y, concomitantemente, la tecnología para la detección masiva de polimorfismos bialélicos sencillos, usualmente conocidos como *single nucleotide polymorphisms* (SNP), ha experimentado un gran desarrollo en estos últimos años.

Los estudios de asociación basados en poblaciones de individuos no relacionados (por ejemplo caso-control)[1,2] son uno de los mejores aliados para abordar las causas genéticas de la enfermedad compleja. La filosofía de un estudio de asociación caso-control es sencilla: los alelos de riesgo son más frecuentes en la muestra de casos que en la muestra de controles. Sin embargo, los estudios de asociación no están exentos de problemas. Prueba de ello es que en estos últimos años se ha observado una gran dificultad para reproducir los resultados de asociación positiva en un amplio espectro de enfermedades complejas[3].

Correspondencia: A. Salas.
Unidade de Xenética.
Instituto de Medicina Legal.
Facultad de Medicina.
Universidad de Santiago de Compostela.
San Francisco, s/n.
16502 Santiago de Compostela. A Coruña.
Correo electrónico: apimlase@usc.es

Aceptado para su publicación el 9 de abril de 2007.

Estratificación poblacional en estudios de asociación

La estratificación poblacional es un caso particular de *confounding by ethnicity* y representa seguramente la causa más importante de errores tipo I en los estudios de asociación. El problema de la estratificación

surge cuando el componente genético-poblacional de los casos y de los controles difiere significativamente; se trata por lo tanto de un problema de falta de apareamiento muestral-poblacional/étnico (*sample matching*) entre los casos y los controles.
La condición *sine qua non* para que la estratificación tenga efecto real sobre el estudio de asociación es que los grupos poblaciones (que subyacen a la población estratificada) presenten frecuencias alélicas diferentes para los marcadores genotipados. Si uno de los grupos poblaciones está más representado en los casos que en los controles (o viceversa), cualquier polimorfismo (neutral) que presente frecuencias alélicas diferentes (estadísticamente significativas) entre las dos poblaciones (muestras) será erróneamente observado como alelo de riesgo (o de protección) en el estudio de asociación (falso positivo). Otro factor coadyuvante es la existencia de diferencias en las prevalencias de la enfermedad en los dos grupos poblaciones. Por otro lado, cuando los genotipos de los SNP analizados no se encuentran en las proporciones esperadas de acuerdo al equilibrio Hardy-Weinberg (HW), se podría sospechar que la población está estratificada. Sin embargo, desequilibrios de HW indican a menudo problemas de genotipado que además podrían diferir entre los casos y los controles dependiendo de si los dos grupos de muestras se procesaron de formas distintas o simplemente como consecuencia de la diferente calidad de las mismas [4]. Existen métodos *ad hoc* para detectar y controlar el efecto de la estratificación en los estudios caso-control, tales como el *genomic control* [5,6] o el *structured association method* [7].

Correcciones para comparaciones (pruebas) múltiples

Cuando se efectúa más de un contraste estadístico en el análisis de datos, aumenta la probabilidad de que alguno sea estadísticamente significativo solamente por azar. El valor de significancia nominal (convencionalmente es 0,05) debe de ser ajustado en función del número de hipótesis ejecutadas. Una corrección inadecuada para prueba múltiple puede derivar en dos resultados igualmente indeseables: a) incremento de falsos positivos o error tipo I (debido a una corrección débil); o b) la no detección de los efectos reales de los marcadores sobre el fenotipo o error tipo II (debido a una corrección demasiado severa). El ajuste de Bonferroni se utiliza a menudo en los estudios de asociación; sin embargo, este método es extremadamente conservador y tiende a incrementar el error tipo II. Una de las mejores soluciones al problema de la corrección por prueba múltiple es el uso de correcciones basadas en procedimientos computacionales (tales como los métodos de permutación).

Errores de genotipado

Los estudios de asociación generalmente conllevan el genotipado de un gran número de SNP en una cantidad grande de muestras. En un estudio de asociación estándar, los errores de genotipado deberían afectar de la misma manera a las muestras de casos y de controles. Esto no tiene por qué ser sistemáticamente así; podemos imaginar una situación en donde los casos y los controles son genotipados en distintos laboratorios o en el mismo laboratorio pero en momentos diferentes o usando técnicas distintas. Esto podría dar lugar a *differential errors*, relacionados en general con la distribución no aleatoria de los errores de genotipado, que podrían tener especial relevancia cuando se procede por ejemplo a la imputación de los datos faltantes. Se ha comprobado que estos errores pueden tener consecuencias graves en el análisis y la interpretación de los resultados [8-10].

Poder de un estudio de asociación. Potencia de la prueba

Uno de los pasos fundamentales en el diseño y la planificación de un estudio de asociación es el cálculo de las muestras que serían necesarias para detectar un determinado efecto genotípico sobre una enfermedad concreta. Este cálculo es importante además si tenemos en cuenta el coste que generalmente conllevan estos estudios (en fenotipado y genotipado). Es posible que muchas asociaciones positivas sean reales pero no reproducibles debido a que el efecto de la variable de riesgo es muy débil. Si los estudios en donde se replica la asociación no presentan un poder adecuado para detectar un efecto débil, lo más probable es que el resultado de la asociación no sea estadísticamente significativo en la replicación. Esta dificultad es generalmente potenciada por el llamado efecto *jackpot*: el primer grupo de investigación que reporta una asociación débil es más probable que haya sobrestimado (y no lo contrario) el efecto real del polimorfismo. Este fenómeno ocurre realmente porque todos los estudios generalmente reportan estimas muy imprecisas del efecto de la variante debido fundamentalmente a variaciones de muestreo.
Otro factor que resta poder de detección de la asociación en el estudio de enfermedades complejas es la heterogeneidad de la enfermedad (*trait heterogeneity*), término que se refiere a la situación que surge cuando una enfermedad no ha sido bien definida y realmente se trata de distintas enfermedades subyacentes; esto conlleva generalmente una pérdida de poder.

Problemas en la determinación del fenotipo en la enfermedad compleja

Otro factor de ruido en los estudios de asociación es el fenómeno de la fenocopia (*phenocopy*): la presencia de una enfermedad (un fenotipo concreto) que puede surgir indistintamente por causas genéticas o ambientales. En ocasiones una patología no está definida con la especificidad suficiente y en la práctica engloba una serie de patologías relacionadas. La existencia de fenocopias pueden conducir a errores tipo II. El *phenocopy rate* se define como la proporción de fenotipos idénticos debidos a factores no genéticos y

puede variar significativamente entre grupos poblacionales. Es común ver que un porcentaje significativo de los pacientes usados como controles presenta alguna patología relacionada y en ocasiones son casos mal diagnosticados.

Interacción gene-gene y gen-ambiente

Otra fuente potencial de variabilidad y ausencia de replicación en los estudios de asociación es la epistasis. Cordell [11] identifica varias razones por las cuales la identificación de las interacciones puede llegar a ser un proceso complejo o incluso imposible. En la mayor parte de los estudios que pretenden detectar epistasis, los tamaños muestrales son sub-óptimos, de tal manera que no permiten estudiar el número masivo de hipótesis que normalmente se pueden llegar a formular, incluso considerando interacciones de dos dimensiones. Se cree que la epistasis puede jugar un papel importante en la enfermedad multifactorial pero, sin embargo, el esfuerzo por desarrollar algoritmos para la detección de la epistasis es todavía escaso [12].

Diferencias genéticas entre poblaciones

Dado que existen diferencias genéticas entre las poblaciones humanas, nos interesa saber si los alelos localizados en un mismo gen están asociados con la enfermedad de manera diferente en distintos contextos poblacionales; dicho con otras palabras, si la ancestralidad de una población puede tener influencia en el impacto de cada variante génica en el riesgo a la enfermedad. Si bien los marcadores genéticos varían significativamente en frecuencia entre las distintas poblaciones, el impacto biológico del riesgo para la enfermedad común parece ser generalmente consistente entre los grupos étnicos principales. No obstante, existen diferencias moderadas en los efectos biológicos que podrían explicar la falta de replicación en un porcentaje significativo de los estudios de asociación [13].

Metaanálisis y sesgo de publicación

Un metaanálisis es básicamente un resumen estadístico de los resultados obtenidos en otros estudios. Uno de los grandes problemas del metaanálisis es el denominado sesgo de publicación: la literatura tiende a reflejar solamente estudios de asociación positiva, mientras que un porcentaje elevado de los estudios negativos de asociación no se publican. Por lo tanto uno espera encontrarse con una tendencia sistemática hacia valores significativos de p, más a menudo que los que cabría encontrar únicamente al azar. La mayor parte de los resultados negativos no llegan a publicarse ya que en general no son de interés para los lectores.

Agradecimientos

Este trabajo ha sido parcialmente subvencionado por el programa Ramón y Cajal del Ministerio de Educación y Ciencia (RYC2005-3), el Ministerio de Sanidad y Consumo (PI030893; SCO/3425/2002), la Xunta de Galicia (PGIDIT06PXIB208079PR) y la Fundación de Investigación Médica Mutua Madrileña (2006/CL370).

BIBLIOGRAFÍA

1. Segado Soriano A, Santiago Dorrego C, Banares Canizares R, Fernández Álvarez E, Bandres Moya F, Gómez-Gallego F. Susceptibilidad genética al desarrollo de hepatitis alcohólica aguda: papel de las mutaciones genéticas en alcohol deshidrogenasa, aldehído deshidrogenas y citocromo p450 2e1. Rev Clin Esp. 2005;205:528-32.
2. Núñez C, Carbajal A, Belmonte S, Moreiras O, Varela G. Estudio caso-control de la relación dieta y cáncer de mama en una muestra procedente de tres poblaciones hospitalarias españolas. Repercusión del consumo de alimentos, energía y nutrientes. Rev Clin Esp. 1996;196:75-81.
3. Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K. A comprehensive review of genetic association studies. Genet Med. 2002;4:45-61.
4. Hosking L, Lumsden S, Lewis K, Yeo A, McCarthy L, Bansal A, et al. Detection of genotyping errors by Hardy-Weinberg equilibrium testing. Eur J Hum Genet. 2004;12:395-9.
5. Devlin B, Roeder K. Genomic control for association studies. Biometrics. 1999;55:997-1004.
6. Pritchard JK, Rosenberg NA. Use of unlinked genetic markers to detect population stratification in association studies. Am J Hum Genet. 1999;65: 220-8.
7. Pritchard JK, Donnelly P. Case-control studies of association in structured or admixed populations. Theor Popul Biol. 2001;60:227-37.
8. Pompanon F, Bonin A, Bellemain E, Taberlet P. Genotyping errors: causes, consequences and solutions. Nat Rev Genet. 2005;6:847-59.
9. Salas A, Carracedo Á, Macaulay V, Richards M, Bandelt HJ. A practical guide to mitochondrial DNA error prevention in clinical, forensic, and population genetics. Biochem Biophys Res Común. 2005;335:891-9.
10. Bandelt H-J, Salas A, Bravi CM. Problems in FBI mtDNA database. Science. 2004;305:1402-4.
11. Cordell HJ. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. Hum Mol Genet. 2002;11:2463-8.
12. Thornton-Wells TA, Moore JH, Haines JL. Genetics, statistics and human disease: analytical retooling for complexity. Trends Genet. 2004;20:640-7.
13. Ioannidis JP, Ntzani EE, Trikalinos TA. «Racial» differences in genetic effects for complex diseases. Nat Genet. 2004;36:1312-8.