

Metodología de los estudios de asociación genética

Sergio D. Sevilla*

Los estudios de "asociación genética" buscan establecer la relación estadística entre variables genéticas poblacionales y un fenotipo determinado (ejemplo: rasgo, riesgo de enfermedad, etc.). Estos están siendo utilizados para descubrir el componente genético que subyace a las enfermedades comunes de alta prevalencia como la diabetes mellitus (DM), la enfermedad coronaria o la insuficiencia cardíaca. Se trata generalmente de estudios de cohortes prospectivas o de tipo casos-controles en los cuales se establece el peso relativo del componente genómico con respecto a otros factores como el ambiente, en el riesgo de desarrollar la enfermedad.

Habitualmente, se utilizan como marcadores genéticos a los polimorfismos simples puntuales (SNPs). Estas variaciones pueden ser en sí mismas funcionales y estar relacionadas a la fisiopatología de la enfermedad, pero en la mayoría de los casos, son utilizadas para el mapeo y ubicación de los verdaderos sitios relevantes. Los dos acercamientos posibles son el del "gen candidato" cuando existe evidencia previa de funcionalidad de la variante, o el de la "asociación indirecta". Actualmente, la técnica de asociación genómica amplia (*wide genome association*) está siendo utilizada para la realización de un *screening* del genoma completo con el fin de establecer posibles sitios de asociación.

Estos estudios se articulan en forma horizontal con estudios de modelos genómicos animales (ratones). Las posiciones posiblemente relacionadas pueden describirse inicialmente en estudios de mapeo genético en ratones u otras especies y luego ser explorados a través de estudios de asociación en humanos. O pueden haber sido descubiertos en estudios de *linkage disequilibrium* en familias de pacientes y luego comprobar la hipótesis fisiopatológica en una cepa de ratón transgénico para dicho gen.

(Rev Insuf Cardíaca 2007; vol. 2; 3:111-114)

Definiciones

Genotipificar. El genotipo es la descripción genética de la composición de un organismo. Al genotipificar se determina el alelo que corresponde a cada variante genética.

Mapa. Es un esquema que muestra la posición de un marcador genético o físico en el genoma.

Mapeo genético. Consiste en el uso de técnicas genéticas para construir un mapa genómico.

Marcador genético. Es un gen o posición en el genoma que existe en dos o más alelos distinguibles y cuya herencia puede ser, por lo tanto, seguida a través de un cruce genético, permitiendo mapear la posición de un gen a determinar. Los polimorfismos como los SNPs y los microsatélites son ejemplos de marcadores genéticos.

Rasgos complejos (complex trait). Rasgos o características fenotípicas de herencia compleja y poligénica, donde los factores ambientales y otros menos conocidos interactúan. Son ejemplos el índice de masa corporal (IMC), presión arterial, altura, etc.

Segregación. Es la separación de cromosomas homólogos o miembros de pares de alelos en diferentes gametas durante la meiosis.

Introducción

Como vimos en el artículo de introducción¹, las condiciones monogénicas han hecho un relevante aporte para comprender los mecanismos que subyacen tras la transmisión de las enfermedades comunes. Sin embargo, comprender el factor genético que influye en la manifestación de una enfermedad como la DM tipo II, la enfermedad coronaria o la insuficiencia cardíaca crónica (ICC) requiere de una metodología diferente. También vimos que la descripción completa del genoma humano, así como el agregado permanente de nuevos SNPs a las bases de datos públicas constituyen el sustrato que posibilita el avance de esta joven técnica de investigación.

A continuación discutiremos la metodología de los estudios de asociación genética.

La heredabilidad y los rasgos complejos

Muchas enfermedades comunes se agrupan en patrones que demuestran que el *background* genético de los individuos juega algún rol en la susceptibilidad². A nivel individual, el riesgo de un pariente de presentar una patología (λ) es una medida del componente de influencia genética que tiene dicha enfermedad. Por otra parte, se define como heredabilidad (h^2) a la

*Médico Internista y Farmacólogo (UBA)
Kendle Argentina, Team Cardiovascular y Respiratorio.
Ex Exchange Research Scientist-Innate Immunity genomics. National Institute of Environmental Health Science (NIH), USA

Correspondencia: Dr. Sergio D. Sevilla
E-mail: sergiodsevilla@yahoo.com.ar

Trabajo recibido: 27/06/2007
Trabajo aprobado: 12/07/2007

fracción de la variación poblacional que puede ser explicada por factores genéticos, trabajando conjuntamente y en forma aditiva. La heredabilidad de una enfermedad puede calcularse en estudios familiares o de gemelos y se encuentra habitualmente entre el 30 al 50% para enfermedades comunes como la DBT, o los rasgos complejos como el IMC o la presión arterial. En la mayoría de las enfermedades comunes de alta prevalencia (como la enfermedad coronaria, la hipertensión arterial, etc.), sabemos que múltiples factores genéticos y no genéticos interactúan para afectar el fenotipo. Estas enfermedades y rasgos reciben, desde el punto de vista genético, el nombre de "rasgos complejos" (*complex traits*), y para la mayoría de ellas la variación atribuible al componente genético subyacente sigue siendo desconocida.

Las dos principales metodologías usadas para mapear las variantes genéticas que influyen en el riesgo de enfermedad son los estudios de *linkage disequilibrium* y los estudios de asociación.

Estudios de *Linkage Disequilibrium*

En los estudios de *linkage disequilibrium* se genotifican unos cientos o miles de marcadores espaciados en millones de bases, en familias con varios parientes afectados^{3,4}. Los marcadores que se segregan en los familiares que presentan la enfermedad, más frecuentemente de lo esperado, son utilizados para localizar el gen causante. Esta técnica ha sido exitosa en encontrar alelos relacionados en todo el genoma, particularmente en los desórdenes monogénicos o de transmisión mendeliana. Sin embargo, estos *linkage disequilibrium analysis* han sido menos exitosos en encontrar genes asociados a enfermedades poligénicas y rasgos complejos. Esto es debido, quizás en parte, al poder limitado de la técnica para detectar el efecto de alelos comunes con modesta influencia en la enfermedad. Por lo tanto, estos estudios serán más poderosos para detectar alelos raros de alto riesgo con mecanismos de transmisión mendeliano. Los estudios de asociación, por otra parte parecen ser más potentes para detectar alelos de enfermedades comunes que confieren riesgo moderado. Por la relevancia e implicancia clínica que suponen las enfermedades comunes de herencia poligénica, dedicaré el resto de este artículo a revisar someramente esta metodología.

Estudios de asociación genética

Los estudios de asociación buscan relacionar un marcador genético particular con una enfermedad (o un rasgo complejo) a través de una población, más que dentro de familias⁵⁻¹³. Estos estudios tienen mucho más poder para detectar los efectos de las variantes comunes con respecto a los estudios de *linkage disequilibrium*.

Un ejemplo ha sido la descripción del alelo tipo III del receptor VNTR, el cual tiene una frecuencia aproximadamente del 70% y tiene un modesto efecto sobre el riesgo de DBT tipo 1 con un valor de $p = 10^{-22}$.

Veamos entonces como se diseñan los llamados estudios de "asociación genética".

Cuando se planea un estudio de asociación genética, se consi-

deran cuatro componentes mayores: a) la enfermedad o el rasgo a ser estudiado (punto final de estudio), b) el grupo de individuos en el cual el rasgo o enfermedad va a ser medido (diseño propiamente dicho), c) los marcadores genéticos que van a ser genotificados y, por último, d) el método analítico para testear la asociación entre el genotipo y el fenotipo (plan estadístico).

a- Selección del fenotipo (puntos finales)

Existen 2 tipos principales de rasgos: dicotómicos o cualitativos (ej.: dos fenotipos posibles: diabético vs no diabético) y continuos o cuantitativos (ej.: niveles sanguíneos de glucosa). Los rasgos de mayor impacto clínico directo suelen ser los rasgos de carácter dicotómico (salud vs enfermedad), y el análisis genético puede proveer una vía para comprender directamente la asociación entre la variable genética y la susceptibilidad.

La h^2 (ver arriba *heredabilidad*) puede ser baja si existen fuertes factores no-genéticos que contribuyen a la susceptibilidad, por lo tanto una baja h^2 reduce el poder de los estudios genéticos. Cuando se decide el fenotipo a ser estudiado, es bueno tener una idea estimada de la h^2 y de la λ . También, es importante reconocer que algunas enfermedades constituyen estados vagamente definidos o poco mensurables. Por ejemplo, el accidente cerebrovascular es una enfermedad con etiologías claramente diferentes. En estos casos, es conveniente utilizar, como puntos finales, rasgos continuos más que condiciones de enfermedad. Estos puntos finales pueden ser los conocidos como "fenotipos intermedios" o "endofenotipos".

Un punto final de estudio clínico puede ser pensado como la síntesis (o gran total) de muchos factores de riesgo, en los cuales, se expresan fenotipos intermedios como subtotaes. Un ejemplo de esto sería el perfil lipídico y el riesgo de infarto agudo de miocardio (IAM). En este caso, el número de variables que afectan los niveles de LDL son probablemente menores que el número de variables que afectan el riesgo de IAM. De este modo, al reducir las posibles variables confusoras, los factores genéticos que contribuyen con los fenotipos intermedios serán más fáciles de identificar, debido a la mejor relación señal/ruido en la fracción de la varianza que es explicada por un factor simple. Por lo tanto, es crucial reconocer de qué modo la definición del fenotipo puede afectar la perspectiva del análisis de asociación. Mientras que los estudios que utilizan sólo un punto final clínico intentan disparar a la luna con sólo una chance de éxito; es posible que los estudios que coleccionan varios fenotipos intermedios nos ayuden a entender la contribución de los factores genéticos a los componentes de la enfermedad. Aunque no sea posible establecer un efecto significativo en el gran punto final clínico (ej.: IAM).

Algunos fenotipos intermedios medibles pueden ser la expresión de ARN, determinados niveles proteicos, medidas de respuesta a drogas y metabolismo, entre otros. Es seguro que las nuevas técnicas proteómicas de alto rendimiento expandan dramáticamente la capacidad de coleccionar información de fenotipos intermedios. También, es importante tener en cuenta que, mientras que los genotipos son determinados con razonable precisión; por lo contrario, los factores de riesgo ambiental y los fenotipos son medidos generalmente en forma inexacta. La precisión de estos convariados será muy importante para evaluar las interacciones ambiente-gen. Por ejemplo, el IMC es un fe-

notipo intermedio más preciso (y útil) que una variable dicotómica como obeso vs no-obeso. En el mismo sentido, el metabolismo de la nicotina “cotinina” es una mejor medida del *status* de fumador que la declaración del mismo paciente.

En conclusión, cabe remarcar que un número modesto de muestras cuidadosamente fenotipificadas puede ser más válido que un gran número de muestras pobremente caracterizadas.

b- Enrolamiento (diseño del estudio)

Existen varios esquemas posibles de enrolamiento y diseño de un estudio de *asociación genética*. Los individuos pueden ser elegidos en base a su fenotipo (ej.: diabéticos vs individuos no afectados) o como miembros de una cohorte, dentro de la cual son seguidos longitudinalmente en el tiempo para el desarrollo de la enfermedad, mientras que otros fenotipos son también medidos. El tipo de fenotipo a estudiar va a influenciar significativamente el método por el cual se reclutan los sujetos.

En el caso de los rasgos cuantitativos, generalmente es suficiente una cohorte prospectiva ya que el rasgo puede ser medido en la mayoría o en todos los sujetos. Una ventaja adicional de las cohortes prospectivas es que puede medirse en forma no sesgada un factor de interacción ambiental (lo que permite el control estadístico por la influencia de dicho factor). Por otra parte, los estudios de caso-control retrospectivos pueden introducir sesgos debido al conocimiento del estado de la enfermedad.

Los factores de riesgo genético deben ser medidos igualmente en los estudios prospectivos y retrospectivos. Sin embargo, hay que considerar que las variantes genéticas que aumentan el riesgo de muerte puedan resultar en un sesgo hacia los sobrevivientes. Los estudios de asociación pueden ser conducidos en múltiples familias o en grupos de individuos no relacionados. El uso de familias requiere métodos analíticos que toman en cuenta la correlación esperada de genotipos entre individuos relacionados. El reclutamiento de pedigrees de sujetos relacionados puede ser difícil, caro y lento. Por otra parte, los individuos no relacionados son mucho más fáciles de reclutar, pero son susceptibles de estratificación poblacional.

c- Selección de marcadores genéticos

Una vez que los individuos y la información fenotípica han sido recolectadas, deben seleccionarse los marcadores genéticos para genotipificar. Pueden ser marcadores individuales, o que abarquen genes, regiones de cromosomas o en última instancia, el genoma entero.

Como vimos en un artículo previo¹, los SNPs son variaciones en la secuencia de DNA en la cual uno de los 4 nucleótidos es reemplazado por otro (ej.: C por A). Los SNPs son la forma más frecuente de polimorfismo en el genoma, y por lo tanto serán la vasta mayoría de los marcadores en un estudio de mapeo del genoma total. Los SNPs genotipificados en un estudio de asociación se denominan *tag SNP* o “*SNPs etiqueta*”.

El concepto de *linkage disequilibrium* (LD) describe la correlación no azarosa entre alelos de un par de SNPs. Los análisis de asociación pueden identificar variantes genéticas (o mutaciones) de riesgo de enfermedad, sólo cuando dichas variantes están fuertemente asociadas con un *tag SNP*.

Existen dos acercamientos para establecer la relación entre va-

riantes genéticas y riesgo de enfermedad: el estudio de “*SNP candidato*” y la “*asociación indirecta*”. El estudio de “*SNP candidato*” es un *test* directo de asociación entre una variante putativa funcional y el riesgo de enfermedad. En este caso, se establece un gen candidato de antemano en base a estudios previos o evidencia experimental biológica.

Por su parte, la “*asociación indirecta*” consiste en testear un mapa denso de SNPs para la asociación con la enfermedad, bajo la asunción de que si un polimorfismo de riesgo existe, éste será o bien tipificado directamente o se encontrará en fuerte LD con uno de los *tag SNPs*. La ventaja del análisis de asociación indirecta es que no requiere la determinación previa de cuál SNP podría ser funcionalmente importante. La desventaja es que se necesita genotipificar un número mucho mayor de SNPs. La disponibilidad actual de bases de datos de SNPs, así como de métodos de genotipificado de alto rendimiento, hacen posible utilizar esta metodología con razonable éxito.

Como se comentó en el artículo previo¹, el proyecto internacional *HapMap Project* tiene el objetivo principal de identificar adecuados grupos de *tag SNPs* que abarquen el genoma, facilitando en gran medida el acercamiento basado en LD.

Esencialmente, los marcadores son elegidos basados en 4 criterios: 1) la probabilidad previa de ser funcionales, 2) la correlación con potenciales variantes causales (LD), 3) ser variaciones de tipo *missense* (que producen la transcripción de un aminoácido alternativo) detectadas por secuenciamiento del ADN y, por último 4) debido a consideraciones tecnológicas.

Los SNPs elegidos según la probabilidad de ser funcionales pueden incluir *missense SNPs*, los cuales son más posiblemente deletéreos o beneficiosos y, por lo tanto, pueden más probablemente ser variantes que contribuyan a la enfermedad común. Como un suplemento a los *missense variations*, los SNPs pueden ser elegidos de regiones no-exónicas, las cuales están conservadas a través de las especies y pueden representar regiones regulatorias funcionalmente importantes. Sin embargo, el uso de variantes no codificantes para determinar variaciones funcionales es mucho más complejo ya que tenemos un conocimiento limitado de las secuencias regulatorias, con respecto al conocimiento que tenemos del código genético.

d- Métodos de análisis de asociación

Los *tests* de asociación pueden ser aplicados a individuos relacionados o no-relacionados. En el caso de los rasgos dicotómicos en individuos no-relacionados, pueden utilizarse pruebas de χ^2 para diferentes frecuencias de genotipo entre los afectados y los no-afectados. Los métodos de regresión lineal se utilizan comúnmente para testear la asociación con rasgos cuantitativos en individuos no-relacionados. Por el contrario, las pruebas para individuos relacionados comparan la distribución observada de genotipos en los parientes con respecto a la frecuencia esperada dada la relación familiar bajo la presuposición de no-asociación.

Estudios de *Wide Genome Association* (GWA)

Los estudios de *GWA* son una extensión del acercamiento de “*asociación indirecta*”¹⁴ (ver arriba). Utilizan cientos de mi-

les de SNPs marcadores y están revolucionando las posibilidades de identificar la influencia genética de los trazos complejos y las enfermedades comunes. A pesar de los costos millonarios, esta técnica se está imponiendo como una de las mejores formas de estudiar las bases genéticas de las enfermedades complejas, en diseños libres de hipótesis. Se realizan generalmente en tres fases: 1) se genotifican individualmente alrededor de 250.000 SNPs en cientos de miles de individuos, 2) se validan los SNPs que demuestran ser más significativos (decenas a miles de SNPs) por genotificado en nuevas cohortes y por último 3) se realiza el mapeo fino de los SNPs adyacentes a los SNPs validados (generalmente sólo unas pocas regiones).

Esta técnica permite un *screening* extenso y de alta densidad del genoma completo en busca de sitios de significativa asociación con el fenotipo estudiado.

Limitaciones de los estudios de asociación

Los estudios de asociación genética de rasgos complejos presentan hasta el momento algunos desafíos adicionales a los tecnológicos y logísticos, ya que habitualmente son mal interpretados y por lo tanto parecen ser pobremente reproducibles. Habitualmente se interpreta como significativa la asociación entre una variable genética y un fenotipo cuando el valor de $p < 0,05$. Sin embargo, la mayoría de tales asociaciones tienen dificultades para reproducirse consistentemente. En un estudio de revisión, sólo 6 de 166 asociaciones pudieron ser replicadas por al menos el 75% de los estudios subsiguientes. La posible causa de tales inconsistencias son los reportes falsos positivos o falso-negativos que fallan en replicar una asociación válida, o posiblemente una verdadera heterogeneidad entre los estudios comparados. Es posible también que la falla por replicar consistentemente las asociaciones se deban al modesto efecto que tienen las variantes causales en el riesgo de enfermedad. En la mayoría de los estudios de asociación, la variable causal se asocian con un 10-15% de aumento del riesgo de enfermedad, y por lo tanto se requieren tamaños de muestras de miles de pacientes para alcanzar un valor nominal de $p < 0,05$.

El mayor acceso a técnicas de genotificado de alto rendimiento, así como los avances en el proyecto *HapMap* y el continuo aporte de nuevos marcadores a las bases de datos, están ayudando a homogeneizar la metodología de los estudios de asociación y a hacerlos más consistentes.

La posibilidad de encontrar nuevas variantes genéticas nos permitirá en el futuro identificar sub-grupos de riesgo, así como, describir nuevos genes involucrados y, también, nos ayudaran en el diseño de nuevas moléculas terapéuticas.

Summary

Genomic association studies pursue to establish the statistical association between population genetic variants and a deter-

mined phenotype (i.e. a trait, the risk of disease, etc). These studies have been used to discover the genetic component of high prevalence diseases such diabetes, coronary hearth disease or cardiac failure. They are most commonly prospective cohort studies or case-control studies where the relative weight of genomic component, respect to other factors such the environment in the risk of developing a disease, is established. Commonly, single nucleotide polymorphisms (SNPs) are used as genetic markers. These variations can be functional and related to the physiopathology of the disease. However, in most of the cases, they are utilized as proxy markers for the mapping of the actual relevant genetic variant. The two possible approaches are "candidate gene", when a previous evidence of functionality exist for the variant, and "indirect association" Currently, "wide genome association" technique is being used to screen the whole genome for possible sites of association.

These type of studies articulate horizontally with animal genomic models studies (mice). Possibly related positions described in mice (or other species) genetic mapping studies could later be explored in human association trials. Or, on the other hand, discoveries done in linkage disequilibrium studies (in families of patients) could later be tested for physiopathological hypothesis in mice strains transgenic for that particular gene.

Referencias bibliográficas

1. Sevilla SD. Introducción a los estudios genómicos en insuficiencia cardíaca. *Rev Insuf Cardíaca* 2007;2:2: 70-72.
2. Hirschhorn JN. Genetic Approach to Studying Common Disease and Complex Traits. *Pediatric Research* 2005;57(5 Part 2):74R-77R.
3. Pulst S M. Genetic Linkage Analysis. *Arch Neurol* 1999; 56:667-672.
4. Feingold E. Methods for Linkage Analysis of Quantitative Trait Loci in Humans. *Theoretical Population Biology* 2001;60:167-180.
5. Carlson CS, Eberle MA, Kruglyak L and Nickerson DA. Mapping complex disease loci in whole-genome association studies. *Nature* 2004;429:446-452.
6. Cheh CN and Hirschhorn J. Genetic association studies of complex traits: design and analysis issues. *Mutation Research* 2005;573(1-2):54-69.
7. Hirschhorn et al. A comprehensive review of genetic association studies. *Genet Med* 2002;4(2):45-61.
8. Cardon LR, Bell JI. Association study designs for complex diseases. *Nat Rev Genet.* 2001;2(2):91-9.
9. Tabor HK, Risch NJ, Myers RM. Candidate-gene approaches for studying complex genetic traits: practical considerations. *Nat Rev Genet* 2002;3(5):391-397.
10. Kwok PY. Genomics. Genetic association by whole-genome analysis? *Science.* 2001;294(5547):1669-70.
11. Wang WY, Barratt BJ, Clayton DG, Todd JA. Genome-wide association studies: theoretical and practical concerns. *Nat Rev Genet* 2005;6(2):109-118.
12. Hirschhorn JN, Daly MJ. Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet.* 2005;6(2):95-108.
13. Lohmueller KE, Pearce CL, Pike M, Lander ES, Hirschhorn JN. Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nat Genet* 2003;33(2):177-182.
14. Pearson V et al. Identification of the Genetic Basis for Complex Disorders by Use of Pooling-Based Genomewide Single-Nucleotide-Polymorphism Association Studies. *Am J Hum Genet* 2007;80:126-139.