Del QTL al gen

Miguel Pérez-Enciso

Institut Català de Recerca i Estudis Avançats (ICREA), Pg. Lluis Companys 23, 08010 Barcelona
http://www.icrea.es/pag.asp?id=Miguel.Perez

Universitat Autònoma de Barcelona, Departament de Ciència Animal i Tecnologia dels Aliments, 08193 Bellaterra[1]
miguel.perez@uab.es

Resumen

Este trabajo resume algunos de los principales avances metodológicos de la era QTL y de los desafíos que esperan. La metodología del modelo mixto es, y seguirá siendo, una herramienta clave en el análisis genético de los caracteres cuantitativos. 'Sólo' debemos irle añadiendo piezas para adecuarlo a las necesidades de cada situación, lo que no quiere decir que sea una tarea fácil. En estos momentos, ya disponemos de una teoría bastante general para el análisis de QTL, aunque echamos de menos un análisis más detallado de las posibles epistasias en el genoma o métodos específicos para caracteres longitudinales o de supervivencia. Tampoco parece haber un consenso generalizado sobre qué criterio emplear para fijar los umbrales de significación. Uno de los temas más candentes en estos momentos es el de localizar una mutación lo más precisamente posible, debemos encontrar formas robustas de incorporar el desequilibrio de ligamiento. Desde un punto de vista experimental, son las razas sintéticas las más adecuadas para llevar a cabo un experimento de cartografía fina. Por ultimo, repasamos algunos de los desafíos que encontraremos próximamente: la genética genómica y los diversos proyectos hapmap. Urge un diálogo con los bioinformáticos y con los genetistas de poblaciones y evolutivos.

Antes del microsatélite (Henderson)

El paradigma clásico en Genética Cuantitativa Animal consta de tres herramientas principales: el modelo mixto, el modelo genético infinitesimal y el mendelismo. Las clásicas ecuaciones del modelo mixto son válidas sólo si aceptamos todos estos ingredientes. Bajo el modelo inifinitesimal y aceptando las reglas mendelianas de transmisión hereditaria, la única información relevante para calcular las probabilidades de que dos alelos sean idénticos por descendencia (IBD) es el pedigrí. Nótese, en todo caso, que estas probabilidades son siempre *condicionales* al pedigrí, y nunca absolutas. Si utilizáramos todo el pedigrí real desde el origen de una especie o de una raza, posiblemente la consanguinidad de todos los individuos sería 1. Cuando el objetivo es identificar las causas últimas de la variabilidad genética para caracteres cuantitativos, es evidente que el modelo infinitesimal no nos sirve, ya que implica que la probabilidad de IBD entre dos individuos es constante a lo largo del genoma. Son precisamente las fluctuaciones en IBD con respecto a lo esperado bajo el modelo infinitesimal las que nos permiten ir más allá e identificar los llamados QTL. En contraste con el modelo genético infinitesimal, el modelo mixto es, y seguirá siendo con toda probabilidad, una herramienta fundamental en el análisis genético de los caracteres 'complejos'. Por tanto, debemos adaptar el modelo mixto a las peculiaridades del análisis de caracteres cuantitativos mediante marcadores (QTL).

[1] Correspondencia a esta dirección

Durante el microsatélite (QTL)

Insistamos una vez más en que un QTL es sólo una asociación *estadística* entre una región del genoma y un carácter. Ir desde un QTL al gen en sí mismo es una tarea extremedamente ardua, la mayor parte de las veces ingrata. A los genéticos con una formación cuantitativa nos es difícil comprender que una región de 20 o 30 cM es absolutamente gigantesca desde el punto de vista molecular: puede contener centenas de genes y decenas de miles de polimorfismos. En la búsqueda de genes candidatos, es normal la angustia de los estudiantes al ver que el intervalo de confianza para un QTL varía de sólo unos cM 'insignificantes' al rehacer el análisis incluyendo unos pocos individuos más o cambiando ligeramente el modelo de análisis. La diversidad de resultados según el método estadístico empleado, además, no hace sino empañar la ya deslucida reputación de los metodólogos, una palabra quiza demasiado próxima de la de meteorólogo[2].

Dicho (o más bien escrito) esto, no cabe duda tampoco del éxito e interés científico que ha despertado la disponibilidad de un gran número de polimorfismos distribuidos por todo el genoma y relativamente baratos de genotipar, los microsatélites. Una de las ventajas de metodología de QTL es que los mismos principios se pueden aplicar a cualquier especie y carácter, no hace falta tener un conocimiento previo de la base genética del carácter. En los últimos diez años se ha avanzado mucho en el conocimiento de la arquitectura genética de los caracteres cuantitativos. Que sepamos muchísimo más no quiere decir, sin embargo, que comprendamos exactamente qué es lo que observamos. Un resultado general es que la mayoría de los QTL presentan una acción génica aditiva y que la epistasia no es muy frecuente. A la vista de la gran cantidad de QTL publicados, cunde un cierto pesimismo entre los cutelólogos en el sentido de que, al fin y al cabo, el modelo infinitesimal no era tan malo como lo pintaban. Creo que éste es un pesimismo pasajero hasta que vayamos conociendo las bases genéticas precisas de la variación cuantitativa. Recordemos que QTL es de los pocos vocablos genéticos que se usan tanto en genética humana, animal, o vegetal, sólo los microbios parecen carecer de QTL.

Desde un punto de vista estadístico, llama la atención la simplicidad de las estrategias utilizadas corrientemente para el análisis de QTL, sobre todo cuando se compara con la sofisticación que habían alcanzado los modelos mixtos utilizados en la evaluación de los reproductores, por ejemplo de las evaluaciones internacionales de los toros lecheros. En este sentido, los experimentos de QTL han supuesto un retroceso enorme en la artillería estadística de los mejoradores. Normalmente, los cruces entre poblaciones se han analizado suponiendo que los alelos de los QTL están fijados en cada raza y en los análisis dentro de razas 'puras' se han analizado las familias de (medios) hermanos como si estuvieran emparentadas. En ambos casos los métodos sencillos tipo mínimos cuadrados han sido más populares que estrategias más generales como la basadas en verosimilitud. De vez en cuando, además, ocurren situaciones curiosas con los fenotipos utilizados. Por ejemplo, en vacuno lechero se ha usado como fenotipo el valor BLUP que se obtiene asumiendo un modelo infinitesimal: sin duda, sería mucho más apropiado utilizar los datos brutos.

Es deseable profundizar en una teoría general y unificada que nos permita el anális genético de caracteres complejos. Dicha teoría puede estar basada en el modelo mixto. La formulación clásica se puede representar como

[2] Tal como puede comprobarse fácilmente con CLUSTAL, una sola deleción y un SNP separan ambas secuencias-palabra, lo que hace sospechar que se trata del mismo gen que hace poco tiempo sufrió una duplicación. Previsiblemente, las funciones se deben haber conservado o, cuando menos, ser muy similares.

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \mathbf{e}$$

En el caso que nos ocupa, el objetivo final es encontrar un modelo alternativo, que sería

$$\mathbf{y} = \mathbf{X}\,\mathbf{b} + \mathbf{Z}\,\mathbf{u} + \sum_{j=1}^{n_{loci}} \mathbf{W}_j\,\mathbf{g}_j + \mathbf{e}\,,$$

donde **W**j es una matriz de incidencia que relaciona los individuos con los alelos del locus j, mientras que **g** contiene los efectos alélicos para cada locus. Es evidente que la principal diferencia entre esta ecuación y la primera es que en la segunda las matrices de incidencia (**W**) no son conocidas, como no lo son su dimensión (el número de alelos) ni siquiera el número de loci que afectan a cada carácter. Si se conocieran, empero, el bagaje teórico del modelo mixto podría ser empleado sin ninguna modificación. En última instancia, podríamos aspirar al modelo

$$\mathbf{y} = \mathbf{X}\,\mathbf{b} + \sum_{j=1}^{n_{loci}} \mathbf{W}_j\,\mathbf{g}_j + \mathbf{e}\,,$$

esto es, aquél que nos permitiera individualizar todas causas genéticas sin recurrir a un residuo genético infinitesimal. Lógicamente, la particularidad de los análisis de QTL es que la matriz **W** es desconocida. La información de marcadores nos permite, sin embargo, estimarla. La forma en cómo utilicemos los marcadores para estimar **W** y en cómo definamos **g** resultará en distintos modelos que cubren la mayor parte, o la totalidad, de los diversos diseños experimentales de QTL. Dentro del paradigma del modelo mixto es inmediato, además, considerar modelos multicarácter donde no es necesario asumir que el mismo modelo para todos los caracteres.

Durante el SNP

En estos momentos, uno de los problemas más importantes a los que nos enfrentamos es el mapeo fino, es decir, la localización de los QTL lo más precisamente posible. Idealmente, deberíamos reducir el intervalo de confianza a menos de un cM si queremos encontrar candidatos posicionales con un mínimo de garantía. El principal factor limitante para la cartografía fina es la ausencia de recombinantes, más que el número de individuos genotipados o la densidad de marcadores. Desde el punto de vista experimental, las razas sintéticas son un material idóneo, especialmente si también se dispone de material genético de las líneas o razas fundadoras. En la mayor parte de las especies domésticas, además, todavía no disponemos de la secuencia del genoma, con la excepción de un borrador en el pollo. Esta situación es transitoria, pero afecta profundamente al tipo de estudios que se pueden abordar.

Uno de los temas más interesantes en estos momentos, desde el punto de vista metodológico, es combinar la información de ligamiento y la de asociación (desequilibrio, LD). Por desgracia, la forma óptima de utilizar el LD depende de la historia de la población, que en general no es conocida. Metodológicamente, los dos problemas principales del mapeo fino son identificar qué alelos del QTL porta cada individuo y calcular la probabilidad de IBD entre dos individuos en cada región del genoma. Por curioso que parezca obtener las probabilidades IBD es un problema extremadamente complejo que las técnicas de Monte Carlo alivian, pero no resuelven completamente. Hoy no sabemos siquiera si se puede resolver en un tiempo realista de computación (hablamos entonces de un problema 'NP', *non-*

polynomial time). La mayoría de los enfoques utilizan, explícita o implícitamente, aproximaciones y simplificaciones. Además, el éxito o fracaso de la cartografía fina depende también del grado de heterogeneidad genética, del número de loci realmente implicados en el carácter y de la fiabilidad en el registro de los fenotipos.

Una estrategia adicional al análisis clásico de QTL es identificar la huella de la selección a través del patrón de polimorfismos en determinadas regiones candidatas o a lo largo de todo el genoma. Desde hace décadas, pero con más ímpetu en los últimos años, se han inventado diversos tests que permiten establecer si la variabilidad de una secuencia de ADN es explicable sólo por deriva o si hay evidencia de selección. En general, la selección causa una disminución en la variabilidad genética. Una huella típica de la selección es un gran número de loci con alelos a baja frecuencia y un nivel de desequilibrio de ligamiento mayor que el esperado por azar. La mayoría de tests están basados en la teoría de la coalescencia, inventada por Kingman en 1982. Este enfoque se ha aplicado con éxito en la especie humana, donde la base pública de SNPs permite realizar estudios de la variabilidad a lo largo de todo el genoma. Pensamos que este enfoque es aún más prometedor en las especies domésticas que en la humana debido a la amplia diversidad entre razas, la intensidad de selección, el corto intervalo generacional, que permite almacenar ADN de muchas generaciones en un tiempo razonable, y la existencia de ancestros no domesticados.

Después (La posgenómica)
Es ésta, sin duda, la etapa más interesante. Quisiera incidir en dos desafíos que se nos presentarán de forma más o menos inmediata, por lo menos en Genética Humana. El primero se refiere a la Genética Genómica, el segundo, a la disponibilidad de grandes cantidades de polimorfismos (proyectos hapmap).

El término genética genómica, acuñado por Ritsert Jansen, se refiere al estudio conjunto de la variabilidad del transcriptoma y del polimorfismo en la secuencia. Podemos distinguir dos enfoques. El primero, trata de determinar la arquitectura genética del transcriptoma, en forma de miles de análisis QTL donde los fenotipos son los niveles de cDNA asociados a cada gen. Como ejemplo de este enfoque, véase el trabajo de Brem et al (2002). El segundo enfoque consiste en utilizar los datos de expresión como ayuda para la localización de genes candidato. Para que esta estrategia tenga sentido se deben cumplir dos condiciones: i) alguno de los niveles de expresión debe estar bajo control genético, al menos parcial, y ii) alguno de los niveles de expresión heredables debe estar correlacionado con el carácter de interés. En caso contrario, no sólo aumentamos enormemente el coste del experimento, sino que también disminuimos su potencia al añadir parámetros innecesariamente. El lector interesado puede consultar los trabajos de Mootha et al. (2003) y el nuestro (Pérez-Enciso et al., 2003).

¿Y si tuviéramos la secuencia completa de 10,000 toros lecheros y quisiéramos conocer los genes que afectan a la producción lechera? Un primer enfoque podría ser identificar las regeiones para las que los 'supertoros' tengan la misma secuencia. Sin embargo, un problema sería determinar cuál /cuáles de las regiones son importantes, por ejemplo, en una región 5' supuestamente reguladora ¿todas las diferencias en SNP son igualmente relevantes? Por otro lado, si existen bloques de haplotipos será imposible determinar la mutación causal última, sólo el bloque. Intentar responder a esta pregunta pone de manifiesto que no es obvio cuál es la mejor estrategia para la identificación de las mutaciones causales. Nótese que en esta situación la información del pedigrí es irrelevante, a no ser para determinar los haplotipos. A corto plazo, el tipo de información de la que dispondremos es de una gran cantidad de SNPs en poblaciones experimentales (p.e. estudios caso / control). Si el genotipado es

suficientemente denso, podemos aplicar técnicas basadas en la coalescencia para determinar si la región genotipada sufre una presión selectiva.

Conclusión: Algunas necesidades metodológicas

Como conclusión, me gustaría incidir en algunos aspectos que necesitan de nuevos avances metodológicos.

- Estudios de asociación masivos: No es descabellado pensar que tendremos estudios de genética genómica o similares a gran escala en un futuro próximo. Dispondremos, no sólo de gran cantidad de genotipos, sino también de fenotipos. Existen dos problemas principales en este caso: uno, encontrar el equilibrio entre potencia y porcentaje de falso positivos; y dos, encontrar nuevos fenotipos, combinaciones de los originales, cuya arquitectura genética sea más fácil de interpretar que la de sus componentes. Una herramienta atractiva podría ser el cálculo de factores de Bayes para la comparación masiva de modelos.

- Coalescencia para caracteres cuantitativos: Si bien hay técnicas para identificar estructuración dentro de una población (lo que nos sirve para asociar haplotipos y caracteres binarios, como enfermedades), la teoría sobre cómo proceder cuando se trata de un carácter continuo no está apenas desarrollada. Está por ver, además, si la coalescencia puede ser aplicada sin problemas al estudio de las razas domésticas animales, que sufren una fuerte intensidad de selección y admixturas frecuentes.

- Reconstrucción de haplotipos: Una gran parte de estudios se recogen muestras aisladas, sin pedigrí. Sin embargo, el conocimiento de las fases es una ayuda fundamental para sacar el máximo partido de la información genotípica. Necesitamos métodos que nos permitan reconstruir los haplotipos con la máxima fiabilidad y el mínimo de asunciones con respecto a la historia de la población.

- Bioinformática: A pesar de su espectacular desarrollo y de que es una disciplina que requiere de habilidades próximas a las utilizadas por un mejorador clásico, apenas contamos con bioinformáticos entre nuestras filas. Urge recrutarlos y formarnos.

Agradecimientos

Estoy muy agradecido a los organizadores de este congreso por la invitación, en particular a Juan Manuel Afonso López, así como a Miguel Toro por sus comentarios. Quisiera dedicar este trabajo a los sufridos estudiantes que se dedican a la excavación del genoma buscando QTL.

Bibliografía

Esta bibliografía no tiene por objetivo principal fundamentar el texto, sino dar una lista de referencias que permita profundizar en los temas tratados en este trabajo.

QTL

Abiola, O., J. M. Angel, P. Avner, A. A. Bachmanov, J. K. Belknap *et al.* 2003 The nature and identification of quantitative trait loci: a community's view. Nat.Rev.Genet 4: 911-916.

Pérez-Enciso, M., & Misztal, I. 2004. Qxpak: A versatile mixed model application for genetical genomics and QTL analyses. *Bioinformatics* en prensa. (disponible en http://www.icrea.es/pag.asp?id=Miguel.Perez).

SNP

Ardlie, K. G., L. Kruglyak, and M. Seielstad, 2002 Patterns of linkage disequilibrium in the human genome. Nat Rev Genet 3: 299-309.

Bamshad, M., and S. P. Wooding, 2003 Signatures of natural selection in the human genome. Nat Rev Genet 4: 99-111.

Hamblin, M. T., E. E. Thompson, and A. Di Rienzo, 2002 Complex signatures of natural selection at the Duffy blood group locus. Am J Hum Genet 70: 369-83.

Liu, J. S., C. Sabatti, J. Teng, B. J. Keats, and N. Risch, 2001 Bayesian analysis of haplotypes for linkage disequilibrium mapping. Genome Res. 11: 1716-24.

Meuwissen, T. H., & Goddard, M. E. 2001. Prediction of identity by descent probabilities from marker-haplotypes. Genet Sel Evol 33, 605-634.

Nordborg, M., and S. Tavaré, 2002 Linkage disequilibrium: what history has to tell us. Trends Genet 18: 83-90.

Sabeti PC, Nick Patterson, Trisha Vanderploeg, Steve F. Schaffner, Jared A. Drake, Matthew Rhodes, David E. Reich, and Joel N. Hirschhorn 2004. Genetic Signatures of Strong Recent Positive Selection at the Lactase Gene. Am J Hum Genet 74:1111

Posgenómica

Brem, R. B., Yvert, G., Clinton, R., & Kruglyak, L. 2002. Genetic Dissection of Transcriptional Regulation in Budding Yeast. *Science* 296, 752-755.

Hastie, T., Tibshirani, R., & Friedman, J. H. 2001. The Elements of Statistical Learning, Springer Verlag, New York.

Jansen, R. C., 2003 Studying complex biological systems using multifactorial perturbation. Nat.Rev.Genet. 4: 145-151.

Mootha VK, Lepage P, Miller K, Bunkenborg J, Reich M, Hjerrild M, Delmonte T, Villeneuve A, Sladek R, Xu F, Mitchell GA, Morin C, Mann M, Hudson TJ, Robinson B, Rioux JD, Lander ES 2003. Identification of a gene causing human cytochrome c oxidase deficiency by integrative genomics. Proc Natl Acad Sci U S A 100: 605-610

Nelson, M. R., S. L. Kardia, R. E. Ferrell, and C. F. Sing, 2001 A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation. Genome Res 11: 458-70.

Peña, D. 2002. Analisis de Datos Multivariantes. McGraw-Hill.

Pérez-Enciso, M., M. A. Toro, M. Tenenhaus, and D. Gianola, 2003 Combining gene expression and molecular marker information for mapping complex trait genes: a simulation study. Genetics 164: 1597-1606.

Pociot, F, Karlsen AE, Pedersen CB, Aalund M, Nerup J, 2004. Novel analytical methods applied to type 1 diabetes genome-scan data. Am. J. Hum. Genet. 74:647-660.

Stoll, M., A. W. Cowley, Jr., P. J. Tonellato, A. S. Greene, M. L. Kaldunski *et al.* 2001 A Genomic-Systems Biology Map for Cardiovascular Function. Science 294: 1723-1726.

Ueda, H., J. M. Howson, L. Esposito, J. Heward, H. Snook *et al.* 2003 Association of the T-cell regulatory gene CTLA4 with susceptibility to autoimmune disease. Nature 423: 506-511.