Mapping expression in randomized rodent genomes

Karl W Broman

Gene expression microarray data on each individual in a pedigree or from an experimental cross enable the identification of the genetic determinants of variation in gene expression. Three new studies apply this approach to rodent recombinant inbred lines and provide new insights into the nature of variation in gene expression and its connection to disease.

There is much excitement about the mapping of genetic loci that underlie variation in gene expression, a concept called 'genetical genomics'[1] or 'expression genetics'. The basic idea is to apply gene expression microarray assays to each individual in a pedigree or from an experimental cross (such as an intercross of two inbred mouse strains) and then use quantitative trait locus (QTL) mapping techniques, treating the transcript abundance of each assayed gene as a quantitative trait, to identify QTLs that influence the expression of genes. In this issue are three reports on the results of this approach, all using recombinant inbred (RI) lines of rodents. Elissa Chesler and colleagues[2] (page 233) assayed gene expression in pooled brain tissue in RI lines formed from the C57BL/6 and DBA/2 mouse strains; Leonid Bystrykh and colleagues[3] (page 225) assayed gene expression in hematopoietic stem cells in this same panel of mouse RI lines. Norbert Hübner and colleagues[4] (page 243) assayed gene expression in both fat and kidney tissues in RI lines formed from the spontaneously hypertensive rat and Brown Norway rat strains.

This approach was previously applied in yeast[5,6], mouse[7] and human[8,9], identifying many features of the genetic basis of variation in gene expression: transcript abundance can be highly heritable; QTLs affecting transcript abundance can be mapped successfully; there are *cis*- and *trans*-acting effects, with some transcripts influenced by a locus at or near its own genomic location (a *cis*-acting QTL) and others influenced by loci mapping to completely different genomic locations (*trans*-acting QTLs); there are *trans*-acting QTLs that influence the expression of hundreds of transcripts; and the expression of a particular transcript may be influenced by multiple loci. The three reports in this issue[2–4] further document these features and explore the tissue specificity of *trans*-acting QTLs and the connection between gene expression and disease.

Genetic mosaics

The reports in this issue all used RI lines[10], which are formed by the crossing of two inbred strains followed by repeated sibling mating (or, in some organisms, self-mating) to produce a new inbred line whose genome is a mosaic of the two progenitor strains (**Fig. 1**). Data on multiple RI lines allow the mapping of genetic variants in the two progenitors that contribute to phenotypic variation. RI lines have a

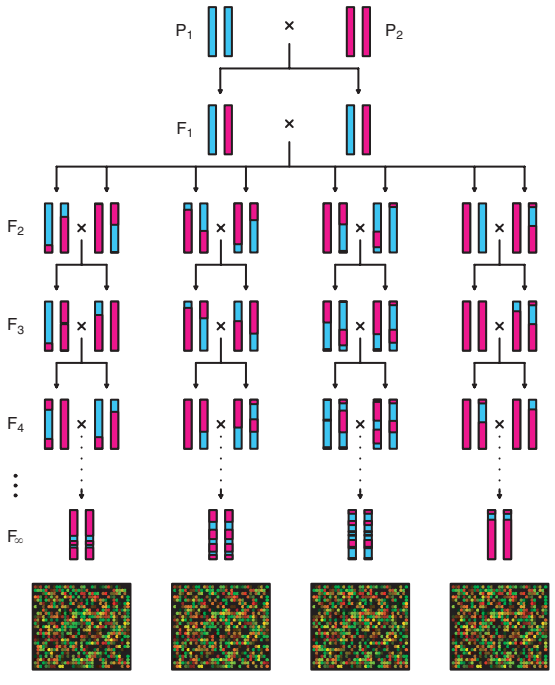

Figure 1 The production of RI lines by sibling mating: crossing of two inbred strains is followed by repeated sibling mating to produce a new inbred line whose genome is a mosaic of the two parental genomes. One chromosome pair is illustrated. Using gene expression microarray data on one or more individuals from multiple such lines, genomic regions containing polymorphisms that influence transcript abundance may be identified for each expressed transcript.

Karl W. Broman is in the Department of Biostatistics, Johns Hopkins University, 615 N. Wolfe Street, Baltimore, Maryland 21205, USA. e-mail: kbroman@jhsph.edu

number of advantages for genetic mapping. They can provide greater mapping resolution than intercrosses, as the breakpoints on RI chromosomes are more dense, having been accumulated in multiple generations. Multiple individuals from each line can be assayed to reduce individual, environmental and measurement variation. Essentially unlimited phenotypes may be obtained from each line; this is particularly advantageous for the study of gene-environment interactions, as the response of a single genome to multiple experimental interventions can be observed. Phenotype data from multiple investigators may be integrated: Chesler *et al.*[2] integrated their results on gene expression in brain tissue with more than 650 phenotypes previously obtained on the studied lines, including measures of behaviors, to identify new candidate genes underlying behavioral phenotypes. Bystrykh *et al.*[3] combined their data on gene expression in hematopoietic stem cells with the brain expression data from Chesler *et al.* to investigate the tissue specificity of *trans*-acting QTLs.

An important limitation of the available RI panels for the mouse and rat is their relatively small size, which limits the power to detect QTLs[11]. Chesler *et al.*[2] used 32 mouse RI lines, and Hübner *et al.*[4] used 30 rat RI lines. Members of the Complex Trait Consortium have recently proposed the development of a panel of 1,000 eight-way RI lines in the mouse[12], formed by intermating eight parental inbred strains. Such a panel would be a valuable community resource: with numerous researchers assaying the same set of mice, we could integrate data on gene expression in diverse tissues at diverse developmental stages under diverse experimental conditions with all possible disease phenotypes.

What do we hope to learn?

Because experiments in expression genetics are expensive, one should determine exactly what may be learned before embarking on such a project. Such studies provide new insight into basic biology, including the genetic basis of gene expression variation. These types of studies also have great potential for systems biology, allowing the identification of networks of coregulated genes and perhaps allowing us to dissect the pathways that connect genes. Many scientists are even more interested in the value of these data for improving our understanding of the etiology of disease phenotypes, such as metabolic syndrome, investigated by Hübner *et al.*[4]. Although efforts to identify genes contributing to complex diseases have been increasingly successful[13], the path from QTL to gene remains laborious and subject to chance. To narrow the number of candidate genes in the region of a QTL, one may focus on genes that have differential gene expression between two strains that also differ in the target phenotype[14]. Expression genetic data allow a refinement of this approach: a gene in a QTL region that has a *cis* effect (*i.e.*, its transcript abundance is associated with QTL genotype) is a more interesting candidate.

Limitations

With the expression genetics approach to understanding disease, it is crucial to identify the appropriate tissue and the appropriate developmental time point at which to assay gene expression. The changes leading to disease may have occurred in a different tissue at a different time point. Furthermore, gene expression data are inherently observational, and observed correlations are therefore insufficient for conclusions of causation. When the genotype at a locus is associated with a phenotype, we can be sure that the locus causes the phenotype; this gives gene mappers some degree of comfort. But with regard to the level of expression of a gene, we cannot distinguish whether the gene's response is part of the etiology of the disease or part of the pathology of the disease. In addition, transcript levels are not the whole story; a full understanding of disease may require more comprehensive models that include the states of proteins, metabolite concentrations and compartmentalization.

Mapping resolution and power are important limitations in any linkage study. Whether a particular *cis*-acting QTL truly reflects the influence on transcript abundance of a polymorphism in the corresponding gene or of another, tightly linked gene cannot easily be determined. In comparing mapping results from multiple tissues, great caution is required; lack of evidence for linkage is not the same as evidence for lack of linkage.

Challenges

The enormous size and scope of expression genetic data pose numerous conceptual, methodological and computational challenges. An obvious problem is that of test multiplicity: we must account for not only the multiple loci in a genome-wide search but also the tens of thousands of correlated phenotypes that have been mapped; findings from any such study must be viewed as preliminary and subjected to further investigation. A more important problem is the development of tools for making sense of these complex data, including their visualization. WebQTL, developed by Chesler and colleagues[2], is a step in this direction. Expression genetic data may not require a new breed of scientist, but they probably will require a new breed of collaboration. The focus of the computational biologist will need to change from the development of tools that answer specific questions to the development of general tools that enable biologists to carry out their own investigations—to explore, visualize and find biological signals in complex data.

1. Jansen, R.C. & Nap, J.P. *Trends Genet.* **17**, 388–391 (2001).
2. Chesler, E.J. et al. Nat. Genet. **37**, 233–242 (2005).
3. Bystrykh, L. *et al. Nat. Genet.* **37**, 225–232 (2005).
4. Hübner, N. *et al. Nat. Genet.* **37**, 243–253 (2005).
5. Brem, R.B. *et al. Science* **296**, 752–755 (2002).
6. Yvert, G. *et al. Nat. Genet.* **35**, 57–64 (2003).
7. Schadt, E.E. *et al. Nature* **422**, 297–302 (2003).
8. Cheung, V.G. *et al. Nat. Genet.* **33**, 422–425 (2003).
9. Morley, M. *et al. Nature* **430**, 743–747 (2004).
10. Silver, L.M. *Mouse Genetics: Concepts and Applications* (Oxford University Press, New York, 1995).
11. Belknap, J.K. *et al. Behav. Genet.* **26**, 149–160 (1996).
12. The Complex Trait Consortium. *Nat. Genet.* **36**, 1133–1137 (2004).
13. Korstanje, R. & Paigen, B. *Nat. Genet.* **31**, 235–236 (2002).
14. Wayne, M.L. & McIntyre, L.M. *Proc. Natl. Acad. Sci. USA* **99**, 14903–14906 (2002).