

Bioinformatics toolbox for narrowing rodent quantitative trait loci

Keith DiPetrillo, Xiaosong Wang, Ioannis M. Stylianou and Beverly Paigen

The Jackson Laboratory, 600 Main St, Bar Harbor, ME 04609, USA

Quantitative trait locus (QTL) analysis is a powerful method for localizing disease genes, but identifying the causal gene remains difficult. Rodent models of disease facilitate QTL gene identification, and causal genes underlying rodent QTL are often associated with the corresponding human diseases. Recently developed bioinformatics methods, including comparative genomics, combined cross analysis, interval-specific and genome-wide haplotype analysis, followed by sequence and expression analysis, each facilitated by public databases, provide new tools for narrowing rodent QTLs. Here we discuss each tool, illustrate its application and generate a bioinformatics strategy for narrowing QTLs. Combining these bioinformatics tools with classical experimental methods should accelerate QTL gene identification.

Introduction

Quantitative trait locus (QTL) analysis is a method to localize chromosomal regions harboring genetic variants that affect a continuously distributed, polygenic phenotype (including many common diseases) [1]. It is particularly important for biomedical research because QTLs detected in rodent models of disease often predict the location of human disease QTLs. For example, after *Rf-1* (renal failure QTL 1) was identified on rat chromosome (Chr) 1, two groups tested for and identified a kidney disease QTL in the syntenic region (Chr 10q23) of the human genome [2]. This finding of a conserved disease QTL in a syntenic region between rodents and humans, termed concordance, is apparent for many phenotypes, including plasma lipid concentrations [3,4], atherosclerosis [5], hypertension [6,7], bone mineral density [8] and kidney disease [2], and suggests that the causal genes underlying disease QTLs are conserved between rodents and humans. Thus, QTL analyses using rodent models can potentially identify genes that are important in human disease.

Despite the important role of QTL genes in many diseases and knowledge of thousands of rodent and human QTLs, investigators have identified only ~30 causal genes underlying QTLs since the advent of QTL analysis in the early 1990s [9,10]. Thus, the major obstacle to identifying QTL genes is not detection of a QTL, but rather the expensive and time-consuming process of narrowing a QTL to a few candidate genes that can be rigorously tested. Flint *et al.* [10] recently discussed the difficulties associated

with identifying causal polymorphisms underlying QTLs and reviewed several experimental strategies. In this article, we present a complementary bioinformatics strategy for narrowing QTLs made possible by the recent investment in public sequence, genotype and expression databases (Table 1). Although Patrinos and Brookes present a dim view of the utility of public databases for linking DNA polymorphisms with disease [11], we have employed several newly developed bioinformatics tools, based on these databases, to narrow QTLs effectively. Although each of these tools has limitations, combined use of these new bioinformatics techniques with the experimental methods reviewed previously [10,12] is a powerful way to narrow a QTL interval.

Comparative genomics

Genomic comparisons between rodents and humans (Figure 1) are effective for narrowing QTL intervals because of structural conservation among all mammalian genomes; those of rats, mice and humans are best defined. The gene content and linear organization of genes along chromosomal segments in mice correspond to that found in humans, with ~340 syntenic segments conserved between the two species [13]. A similar relationship exists between the rat and human genomes [14]. Although concordance between human and rodent QTLs has been observed for many traits, it has been quantified in only a few. When human QTLs were compared with those in mice, 93% of high density lipoprotein cholesterol QTLs, 100% of low density lipoprotein cholesterol QTLs, 80% of triglyceride QTLs and 63% of atherosclerosis QTLs were found to be concordant [4,5]. By aligning the concordant QTLs from these different species based on their genomic sequence, one can often identify a region of overlap common to the QTL. Assuming that the same causal gene underlies the QTL in all three species, the region of overlap is likely to contain the causal gene. Vitt *et al.* employed a comparative genomics strategy to narrow *Rf-1*, a kidney disease QTL [15]. *Rf-1* is localized to distal rat Chr 1, which is syntenic to parts of human Chrs 9, 10 and X. Human Chr 10 is linked to kidney disease, whereas human Chrs 9 and X are not. By focusing on only those regions of the *Rf-1* interval that are syntenic to human Chr 10, the authors narrowed *Rf-1* from 20 to 11.5 Mb [15].

Limitations

A rodent QTL can be homologous to multiple human regions and sometimes there are QTLs in more than one of

Corresponding author: Paigen, B. (bjp@jax.org).

Available online 13 October 2005

Table 1. Summary of bioinformatics tools for dissecting rodent QTLs

Bioinformatics tool	Summary	Resolution
Comparative genomics	Identifies regions of chromosomal synteny in QTLs that are concordant across species	10–20 Mb
Combined cross analysis	Recodes genotype information from multiple crosses detecting a shared QTL into one susceptibility and one resistance genotype to combine the crosses in a single QTL analysis	10–20 Mb
Interval-specific haplotype analysis	Detects regions of IBD within QTLs shared in multiple crosses	<5 Mb
Genome-wide haplotype analysis	Associates conserved haplotype patterns across the genome with a phenotype in inbred strains	<5 Mb
Sequence comparison	Searches strain-specific sequence databases for regulatory or coding polymorphisms within the QTL interval	10–20 genes
Expression comparison	Searches EST or microarray databases to identify genes expressed in an organ of interest or genes exhibiting differential expression between the strains of interest	10–20 genes

the homologous regions in humans, which complicates this approach. Comparative genomics assumes that the same causal gene underlies all of the concordant QTLs in the species examined. This assumption can be incorrect if a cluster of related genes map to the same region, but the rodent has a mutation in one gene and the human has a mutation in another. However, experimental evidence suggests that causal genes underlying rodent QTLs are often conserved as disease genes in humans. For example, the genes *Add1* (blood pressure; Refs [16,17]), *Ctla4* (type 1 diabetes; Refs [18]), *Angptl3* (atherosclerosis; Ref. [19]) and *Ox40l* (atherosclerosis; [20]) all underlie a rodent QTL and are also associated with the corresponding human disease (given in parenthesis), supporting the idea that the causal genes underlying concordant QTLs are conserved as disease genes in both rodents and humans.

Because chromosomal segments conserved between rodents and humans are often large, comparative genomic analysis only modestly narrows a QTL interval in most

cases. Because the average rodent QTL is 20 cM and corresponds to three or four homologous human regions, the resolution of comparative genomics is ~5–10 cM. Genomic comparisons that include data from additional vertebrate species, such as dogs and rabbits, will add more power for QTL narrowing because the chromosomal breakpoints are likely to be at different locations. However, QTL data from other vertebrates are limited, and the degree of concordance between QTLs identified in vertebrates other than rodents is not well established.

Combined cross analysis

Both combined cross analysis and haplotype analysis (discussed later) are based on the unusual nature of the inbred mouse genome, which is a mixture of DNA from different subspecies that hybridized in pet shops and laboratories. The laboratory mouse genome is a mosaic of segments derived primarily from *Mus musculus musculus* and *Mus musculus domesticus* ancestral sources, with a minor contribution from *Mus musculus castaneus* [21]. Most genetic variation between inbred mouse strains is ancestral variation [22]. Although individual single nucleotide polymorphisms (SNPs) are primarily bi-allelic, they sort into several different haplotype patterns in inbred mouse strains owing to recombination of the ancestral chromosomes, such that several haplotypes, comprising different combinations of the two alleles at each SNP, might be present for any gene (Figure 2). The allele conferring susceptibility to disease at any given QTL can come from one ancestral source and the allele conferring resistance can be derived from the other ancestral source. Because QTLs are often found at the same chromosomal location in multiple crosses using different inbred strains [2–5], all strains carrying the susceptibility allele at a QTL are assumed to share the same ancestral allele of the causal gene, whereas strains carrying the resistance allele are assumed to share the alternative ancestral allele.

Combined cross analysis is a statistical method to combine data from existing QTL crosses to narrow the QTL interval and does not require any additional experiments. The major assumption of combined cross analysis is that the same causal gene underlies the QTL in each cross, so combining multiple crosses increases the number of recombination events, leading to better resolution of the QTL interval. If this assumption is incorrect, the analysis can separate the QTL peak into two

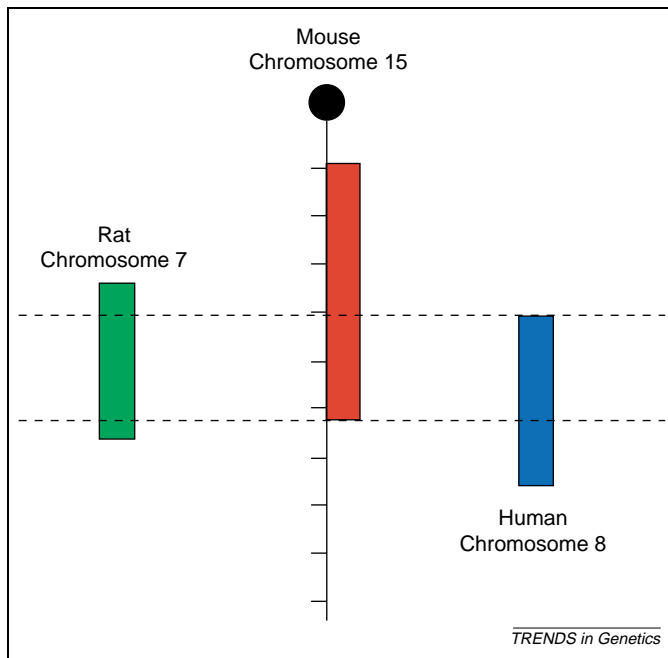


Figure 1. Comparative genomics. The red box depicts a blood pressure QTL interval on mouse chromosome 15. The green and blue boxes depict corresponding blood pressure QTL in the homologous regions of the rat and human genome, respectively. Based on the assumption that the same causal gene underlies the QTL in all three species, comparative genomic analysis of this mouse QTL localizes the gene to the region of overlap between the three species, represented by the area between the dashed lines.

SNP Position	Haplotype 1					Haplotype 2					Haplotype 3					Haplotype 4					Haplotype 5				
171 296 468	C	C	C	C	C	C	C	C	C	C	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
171 296 469	G	G	G	G	G	A	A	A	A	A	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G
171 296 480	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G
171 296 502	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	T
171 296 514	T	T	T	T	T	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
171 296 530	A	A	A	A	A	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G
171 296 567	C	C	C	C	C	T	T	T	T	T	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C
171 296 615	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	A	A	A	A	G	A	A	A	A
171 296 640	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	G
171 296 681	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	T

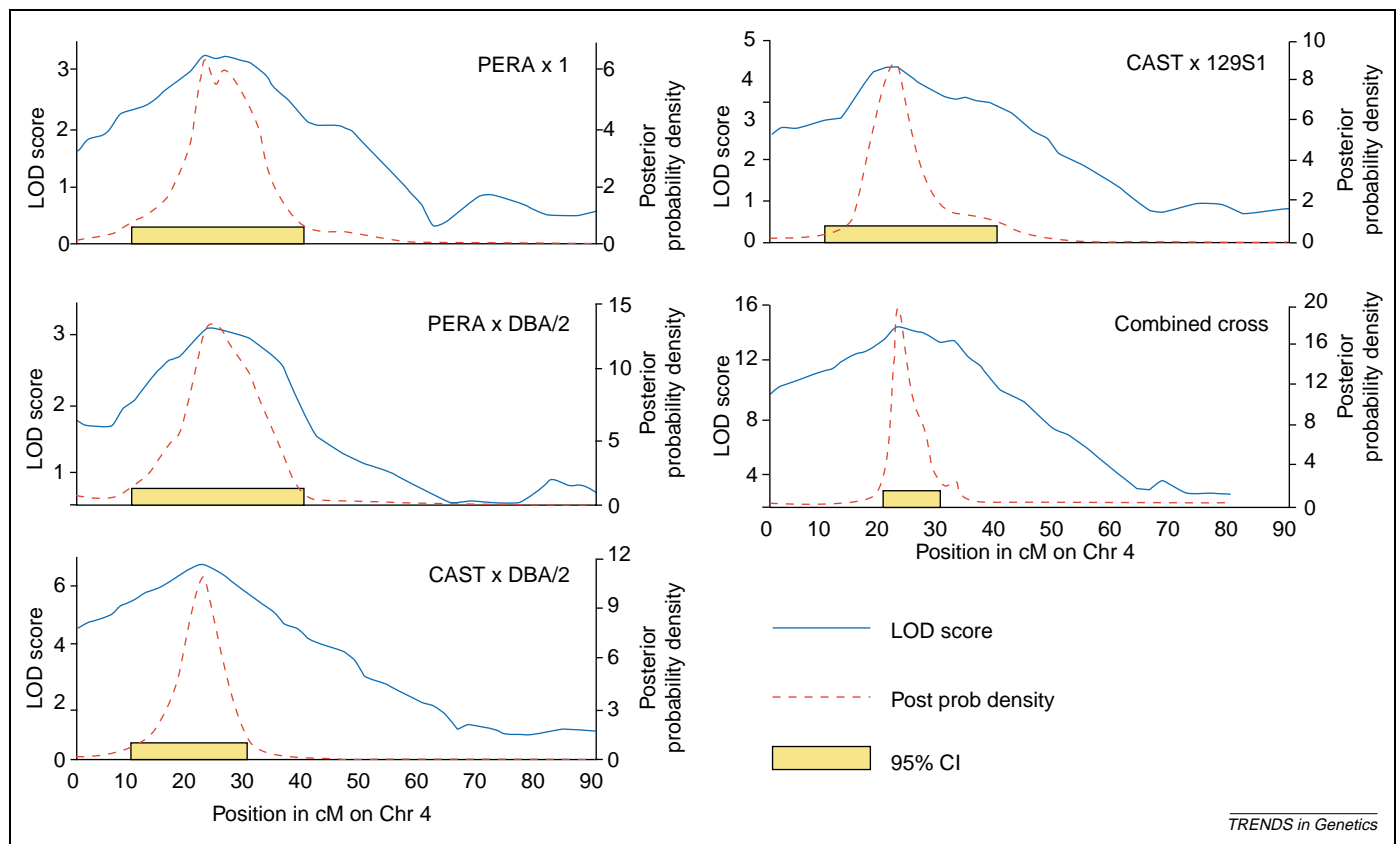
TRENDS in Genetics

Figure 2. SNPs sort into haplotype patterns among inbred mouse strains. Individual SNPs (the positions are listed in the left column) have two alleles, but these SNPs sort into different haplotype patterns. The figure shows five haplotype patterns across the *ApoA2* gene in inbred mouse strains.

closely linked but distinct peaks. Combined cross analysis provides increased power to detect QTLs with small effects, resolve closely linked QTLs and narrow QTL confidence intervals compared with individual QTL crosses (Figure 3) [23].

To combine the data from multiple crosses, the genotype information must be recoded from a strain-specific code, such as 'B' for C57BL/6, to a phenotype-specific code, such as 'H' for high blood pressure phenotype. For example, we identified a blood pressure

QTL on distal Chr 1 in separate crosses between C57BL/6J and A/J mice [7] as well as between C3H/HeJ and SWR/J mice [24]. C57BL/6J and SWR/J mice carry the high blood pressure allele at this QTL, whereas A/J and C3H/HeJ possess the low blood pressure allele. To combine these crosses statistically, we first recoded the C57BL/6J and SWR/J genotypes as a single 'high' allele and the A/J and C3H/HeJ genotypes as a single 'low' allele. We then combined the genotype information from both crosses into a single QTL analysis based on the combined population. Combination can include the entire cross or only the chromosome with the shared QTL. The resolution achieved by combined cross analysis depends on the number of crosses being combined, the number of animals and the density of genotype information to detect chromosomal recombinations in the crosses. We have used this method to narrow a blood pressure QTL on Chr 1 from 42 cM to 18 cM [24], to reduce a QTL affecting high-density lipoprotein (HDL) concentration identified in four crosses on Chr 4 from 30 cM to 10 cM [23] and to refine a bone density QTL on Chr 7 from 34 cM to 22 cM [25]. These examples show that combined cross analysis is an efficient way to use existing data to narrow a QTL interval. Raw data from numerous QTL crosses collected by Gary Churchill of The Jackson Laboratory is freely available on his web page (Table 2) to facilitate combined cross analysis.



TRENDS in Genetics

Figure 3. Combined cross analysis. The yellow rectangles in depict the confidence intervals (CI), from 20 to 40 cM, of QTLs detected at the same chromosomal location in four separate crosses. Combined cross analysis assumes that a common causal gene underlies these four QTLs. By recoding the genotype data for each strain carrying the susceptibility allele (PERA and CAST) as a single genotype and the genotype data for each strain carrying the resistance allele (I, DBA/2 and 129S1) into one alternative genotype, the data from all four crosses were combined into one QTL analysis. Combined cross analysis provides more chromosomal recombinations, often at different locations in each cross, resulting in a narrowed confidence interval (10 cM, yellow rectangle shown the lower right panel). This figure was redrawn using previously published data from Ref. [23].

Table 2. Web-based bioinformatics tools for dissecting rodent QTLs^a

Bioinformatics tool	Web page
Comparative genomics	http://www.informatics.jax.org/menus/homology_menu.shtml
	http://ecrbrowser.dcode.org
	http://www.rgd.mcw.edu/VCMAP/mapview.shtml
	http://pmrc.med.mssm.edu:9090/QTL/jsp/qtlhome.jsp
Combined cross analysis	http://www.jax.org/staff/churchill/labsite/datasets/qtl/qtlarchive
Haplotype analysis	http://mousesnp.roche.com
	http://snp.gnf.org/GNF10K
	http://rgd.mcw.edu/ACPHAPLOTYPED
	http://www.ncbi.nlm.nih.gov/SNP
	http://www.broad.mit.edu/snp/mouse
	http://aretha.jax.org/pub-cgi/phenome/mpdcgi?rtn=snp/door
	www.well.ox.ac.uk/mouse/INBREDS
	http://mouse.perlegen.com/mouse
	http://www.celeradiscovery.com
	http://www.informatics.jax.org/menus/expression_menu.shtml
Genome sequence	http://lena.jax.org/~dcb/ensRNA/exquest.html
Gene expression	http://mouse.biomed.cas.cz/sage
	http://www.ebi.ac.uk/arrayexpress
	http://genome-www.stanford.edu/microarray
	http://www.ncbi.nlm.nih.gov/geo
	http://symatlas.gnf.org
	http://pga.tigr.org
	http://www.genenetwork.org

^aA comprehensive list of molecular biology databases is located at <http://www3.oup.co.uk/nar/database/c/> [46].

Limitations

The major limitation of combined cross analysis is that most of the data from QTL analyses are not publicly available. The Churchill website serving as a public repository for QTL data sets will help to alleviate this limitation. In addition, the phenotype data must be comparable between the data sets for the crosses to be combined in a meaningful way. For example, body weight, body mass index, percent of body fat and total fat pad weight are measures of obesity, but these are not equivalent measurements and not easily combined [26].

Haplotype analysis

An alternative way to capitalize on the structure of the mouse genome to narrow QTL intervals is haplotype analysis. Cuppen recently reviewed the theory and practice of haplotype-based genetics [27]. There are two primary methods for applying haplotype analysis to QTLs: interval-specific haplotype analysis and genome-wide haplotype association. Both methods can be used to narrow experimentally derived QTLs.

Interval-specific haplotype analysis

Approximately 97% of the genetic variation between inbred mouse strains is ancestral [22], so regions of identity by descent (IBD) between two strains used to detect a QTL are highly unlikely to contain the causal genetic polymorphism underlying the QTL [28]. For example, a cross between C57BL/6J and A/J mice detected

a blood pressure QTL on Chr 1 [7]. Any region within the QTL interval where C57BL/6J and A/J mice share alleles derived from the same ancestral source is unlikely to contain the causal polymorphism, but any region where C57BL/6J and A/J mice contain alleles from different ancestral sources is likely to contain the causal polymorphism.

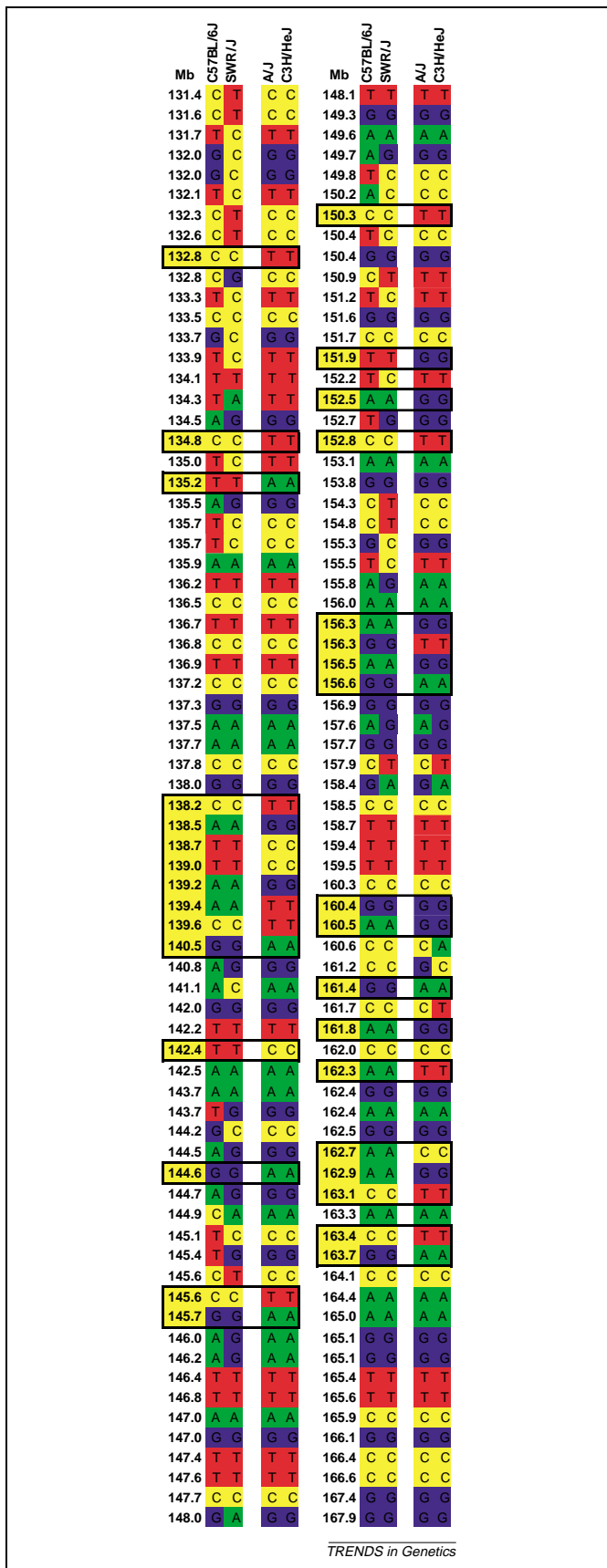
To perform interval-specific haplotype analysis, the genotype data covering the QTL interval can be grouped into the strains that carry the high allele and those that carry the low allele (Figure 4). Any region with genotypes that are shared between the high allele strains and different from the low allele strains is considered a haplotype region likely to contain the QTL gene.

Interval-specific haplotype analysis has been applied to rodent QTLs; for example, narrowing a metastasis QTL on mouse Chr 19 from >400 genes to ~25 candidate genes [29] and refining a blood pressure QTL on mouse Chr 1 from 42 cM to 2.3 cM [24]. These early examples required considerable laboratory work to generate haplotypes based on simple sequence length polymorphism (SSLP) markers, but the existence of public databases with SSLP or SNP marker information for numerous inbred mouse or rat strains will enable future haplotype analyses based solely on bioinformatics. Public databases with relatively dense marker coverage are currently available for haplotype analysis in both the mouse and rat (Table 2), and additional genotype data from Perlegen Sciences (<http://www.perlegen.com>) are forthcoming.

Haplotype analysis can be applied to any number of crosses that detect a QTL at the same approximate position, but it is most effective when applied to multiple crosses using different parental strains (Figure 4). Only those regions where the strains carrying the susceptibility allele share common ancestral DNA and differ from strains carrying the resistance allele are likely to contain the causal gene. Thus, increasing the number of crosses and strains available for haplotype analysis of a particular QTL increases the ability to narrow the QTL interval. This concept is elegantly shown by Wang *et al.*, who applied haplotype analysis to nine crosses that detected a QTL regulating plasma HDL concentrations (*Hdlq5*) to identify *ApoA2*, which encodes apolipoprotein A-II, as the causal gene [30].

Genome-wide haplotype association

Whereas interval-specific haplotype analysis requires *a priori* knowledge of a QTL interval, genome-wide haplotype association predicts the location of QTLs affecting a trait without *a priori* knowledge of the loci. The concept of genome-wide haplotype association is an extension of *in silico* QTL mapping, first proposed by Grupe *et al.*, who developed a computer algorithm to link phenotype to genotype in inbred mouse strains [31]. Smith *et al.* subsequently applied *in silico* QTL analysis to atherosclerosis susceptibility in six strains of *ApoE*-deficient mice [32], and found that four of five *in silico* QTLs overlapped with experimental QTLs. Although the initial foray into *in silico* QTL mapping was promising, there is substantial controversy about the number of strains and SNPs included and statistical approach used



TRENDS in Genetics

Figure 4. Haplotype analysis. SNP genotype data from four strains (C57BL/6, SWR, A and C3H) were used to identify overlapping blood pressure QTLs on chromosome 1 [24]. QTLs were detected by crossing C57BL/6×A mice and SWR×C3H mice. C57BL/6 and SWR mice contributed the high blood pressure allele at this locus. Haplotype analysis assumes that the QTL is caused by an ancestral allele, so the

causal gene should reside in a region where strains carrying the high blood pressure allele (C57BL/6 and SWR) share a common haplotype that differs from the strains carrying the low blood pressure allele (A and C3H). The boxed intervals represent regions of the QTL that fit the appropriate haplotype pattern that is likely to contain the causal gene.

in the original *in silico* QTL analysis method [33,34]. Nevertheless, the innovative idea of *in silico* QTL analysis stimulated the development of genome-wide haplotype association. Liao *et al.* provided a proof-of-concept demonstration for the ability of genome-wide haplotype association to identify the causes of monogenic trait differences in inbred strains [35]. Pletcher *et al.* extended its application to polygenic traits and improved the analytical method by including more strains with phenotypic data, increasing the number of SNP genotypes, and applying a generalized family-wise error-rate method to determine the significance of the results [36]. The authors grouped inbred strains based on the inferred haplotype, determined by three consecutive SNPs and tested for phenotype differences between the groups (Figure 5). They then moved the three SNP window by one SNP, regrouped the strains and tested for differences in the phenotype. They sequentially applied a rolling three SNP window across the genome to identify loci where the different inferred haplotype groups had significant differences in phenotype. Using this method, nine of the ten autosomal regions predicted by genome-wide haplotype association to influence plasma high-density lipoprotein cholesterol (HDL-C) concentration matched experimentally derived QTL intervals, a strong correlation between the computational and experimental results. The authors also applied the method to a second trait, gallstone formation, and found that seven of the 11 computationally predicted loci matched experimental QTLs. It is currently unclear whether the four computational loci that are predicted to affect gallstone formation without an experimental QTL match are false positives or are undetected by the few QTL crosses investigating gallstone formation.

In addition to predicting the loci that affect a trait, genome-wide haplotype-association studies can also be used to narrow experimentally derived QTLs. Using genome-wide haplotype association, Pletcher *et al.* identified an interval (89–94 Mb) on mouse Chr 8 that falls within an experimental QTL regulating HDL-C concentration. Thus, the authors focused on candidate genes within the associated haplotype interval and identified a plausible candidate gene containing a non-synonymous sequence polymorphism [36]. The haplotype containing this polymorphism is associated with higher average HDL-C concentrations in mice from 48 inbred strains [36], providing strong evidence that this polymorphism underlies this HDL-C QTL.

Another example is *Hdlq7*, an HDL-C QTL on mouse Chr 5, found in advanced intercross lines [37], which was narrowed from 8.6 Mb to 2.6 Mb using congenic mice. Genome-wide haplotype-association analysis further narrowed this QTL to a 0.4-Mb region containing only one gene (Figure 6a). This gene differed in mRNA expression between the strains, and injection of siRNA targeting this

causal gene should reside in a region where strains carrying the high blood pressure allele (C57BL/6 and SWR) share a common haplotype that differs from the strains carrying the low blood pressure allele (A and C3H). The boxed intervals represent regions of the QTL that fit the appropriate haplotype pattern that is likely to contain the causal gene.

Locus 1								
Strain	SDP 1			SDP 2				
	NZB/BINJ	CE/J	RIIS/J	PL/J	DBA/2J	LP/J	BTBR	SJL/J
SNP 1	C	C	C	T	T	T	T	T
SNP 2	C	C	C	T	T	T	T	T
SNP 3	G	G	G	A	A	A	A	A
HDL	134	72	48	80	75	80	90	61
Avg.	85			77				

No difference between haplotypes

Locus 2								
Strain	SDP 1				SDP 2			
	NZB/BINJ	LP/J	BTBR	PL/J	CE/J	DBA/2J	SJL/J	RIIS/J
SNP 1	T	T	T	T	G	G	G	G
SNP 2	A	A	A	A	G	G	G	G
SNP 3	A	A	A	A	G	G	G	G
HDL	134	80	90	80	72	75	61	48
Avg.	96				64			

Difference between haplotypes

TRENDS in Genetics

Figure 5. Genome-wide haplotype association. The top panel represents the strain distribution pattern (SDP) across three SNPs in eight strains at locus 1. NZB, CE and RIIS mice share an SDP, whereas PL, DBA/2, LP, BTBR and SJL share a different SDP. Genome-wide haplotype association first uses the SDP to infer common haplotypes among inbred strains, and then associates the haplotype with a phenotype. At locus 1, the average (Avg.) plasma HDL concentration is 85 for SDP 1 and 77 for SDP 2, so the haplotype is not linked to the phenotype. By contrast, the average phenotype for SDP 1 at locus 2 is 96 (NZB, LP, BTBR, and PL mice) versus 64 for SDP 2 (CE, DBA/2, SJL and RIIS mice), thus this haplotype is linked to plasma HDL concentration. Data were derived from Pletcher *et al.* [36]. Although this example shows two SDPs at each locus for simplicity, genome-wide haplotype association can test any number of SDPs at each locus for association with a phenotype.

gene into mice with higher expression decreased both their gene expression and plasma HDL-C levels, supporting this gene as an excellent candidate for *Hdlq7* (Wang X *et al.* unpublished). In some cases, both interval-specific haplotype analysis and genome-wide haplotype association can narrow the same experimental QTL. *Hdlq33* is a 34-Mb QTL, affecting plasma HDL-C levels, on distal mouse Chr 1 [3] that contains 181 genes. Interval-specific haplotype analysis narrowed the QTL to three regions, with a total length of 2.7 Mb, containing 13 genes. Genome-wide haplotype association further narrowed the interval to a 0.7-Mb region containing only four genes (Figure 6b). One of these genes is functionally relevant to HDL metabolism, differentially expressed between the strains, consistent with its potential function to down-regulate HDL, and contains a mutation in a conserved amino acid, making it a strong candidate for *Hdlq33* (Wang X *et al.* unpublished).

Limitations

Interval-specific haplotype analysis is based on the idea of IBD from common ancestors of the inbred strains. One limitation is that the mutation causing the QTL might have occurred after the strains were separated from common ancestors ~100 years ago and might be in a region of IBD; in such a case, focusing on non-IBD regions would miss the causal gene. This problem would occur if a QTL is detected between only a single pair of strains or between a single strain crossed to several other strains. However, detecting a QTL in multiple crosses using different parental strains strengthens the assumption that the causal polymorphism is ancestrally derived. Another limitation occurs for strains that have been

separated for thousands of years, such as strains recently derived from the wild (e.g. PERA, PWD). These strains might share ancestral mutations, but the regions are less likely to be IBD. Subspecies that made only minor contributions to the genomes of common inbred mice, such as *Molossinus* or *Castaneus*, are also less likely to be useful for haplotype analysis. Therefore, it is important to

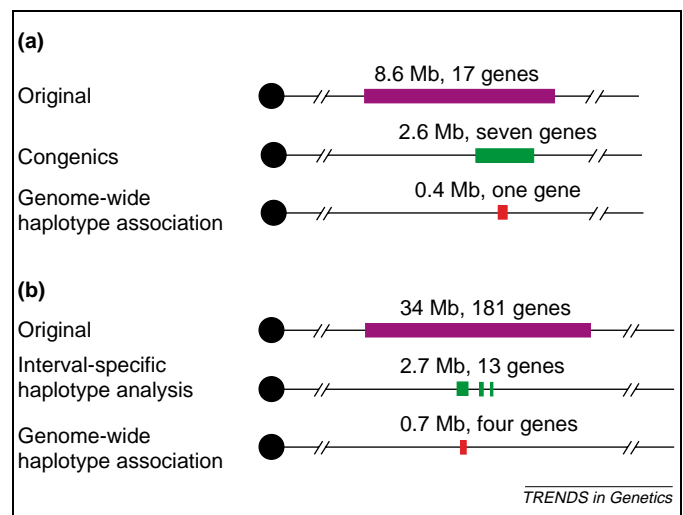


Figure 6. Using interval-specific haplotype analysis and genome-wide haplotype-association analysis to narrow experimental QTL. (a) The original QTL for plasma HDL-C levels (*Hdlq7*) on mouse chromosome 5 was found in (NZB×B6) F1 advanced intercross lines [37]. B6.NZB-*Hdlq7* congenic mice confirmed this QTL, and interval-specific and genome-wide haplotype association analyses were subsequently used to narrow this QTL further. (b) The original QTL for plasma HDL-C levels (*Hdlq33*), found in three mouse crosses (PERA×DBA/2, SM×A, B6×A), has a confidence interval of 34 Mb. Interval-specific haplotype analysis and genome-wide haplotype association analysis were used to narrow this QTL. The horizontal black lines represent chromosomes and the centromere is represented by a black circle to the left. The colored horizontal bars represent QTL intervals.

exclude mouse strains that were not recently derived from these ancestors from the analysis because they do not share IBD regions.

High-density genotype data are important for accurately narrowing a QTL region using interval-specific haplotype analysis because the haplotype pattern in any one strain can shift to a different pattern and back again over a relatively short distance [38]. This means that there are blocks of *M. m. musculus* and *M. m. domesticus* ancestral alleles across any given locus in a single inbred mouse strain, but the size of the blocks and the locations where the blocks change ancestry varies across the inbred strains at the same locus. High-density genotype data are important for accurately determining the haplotype pattern in each strain for interval-specific haplotype analysis (e.g. to determine if two strains that have the same QTL allele affecting a trait share common ancestral DNA). The increasing availability of public databases containing high-density SNP-genotyping data tested on many inbred strains should help address this limitation.

High-density genotype data are also crucial for genome-wide haplotype association. Because strains are grouped by inferred haplotype and the haplotypes are inferred from genotypes, high-density genotype data increases the likelihood that the inferred haplotype for any given strain accurately reflects the true haplotype. Genome-wide haplotype association is still a developing method and several questions remain unaddressed. For example, how many strains should be included in the analysis? Should wild-derived inbred strains be included? What is the minimum SNP density required to detect small haplotypes? Several groups are currently addressing these questions, so we can expect an improved method for genome-wide haplotype association by understanding the effect of changes in each of these parameters.

Sequence comparisons

DNA sequence polymorphisms affecting either expression or function of a gene product are the molecular bases for QTLs. Therefore, identifying sequence polymorphisms between strains used to detect a QTL is important for determining the causal gene [1]. The genome sequence of four inbred mouse strains (C57BL/6J, DBA/2J, A/J and 129S1/SvImJ) is nearly complete, and whole genome sequencing is ongoing for additional strains (AKR/J, BALB/cByJ, BTBR T⁺ tf/J, C3H/HeJ, CAST/EiJ, FVB/NJ, MOLF/EiJ, KK/HIJ, NOD/LtJ, NZW/LacJ, PWD/PhJ and WSB/EiJ) through the 'resequencing project' at the National Institute of Environmental Health Sciences (<http://www.niehs.nih.gov>). Genome sequence databases enable *in silico* sequence comparisons of whole QTL intervals to detect sequence polymorphisms. Marshall *et al.* developed a bioinformatics method to search strain-specific sequence databases for non-synonymous sequence polymorphisms within a QTL interval [39]. By comparing the sequences of 121 genes within an 8-Mb region on mouse Chr 1, the authors detected amino acid changes in only six genes [39], substantially reducing the number of candidate genes.

This method can also be used to detect sequence polymorphisms in regulatory regions. However, the utility of sequence comparisons to detect regulatory

polymorphisms is hindered by limited knowledge of functional regulatory elements. This approach can also be combined with classical QTL-narrowing methods, such as generating interval-specific congenic strains or selective phenotyping of recombinant progeny, to identify additional polymorphic markers for genotyping within the QTL interval [39].

Limitations

Although sequence databases provide a rapid way to screen many genes within a QTL interval for functional polymorphisms, errors in the database can hamper the utility of this approach. Variation in sequence coverage for different strains in a database leads to variable quality of finished sequence data for the different strains. Therefore, functional polymorphisms detected using sequence database comparisons should be confirmed by sequencing genomic DNA from the appropriate strains. In fact, identifying sequence variants that can be confirmed experimentally is the most efficient way to use sequence databases. The absence of a polymorphism in the database should not be taken as proof that the gene contains no functional polymorphisms, and plausible candidate genes should not be excluded from further consideration based solely on sequence databases.

Expression databases

As noted previously, a sequence polymorphism affecting either gene function or expression can underlie a QTL; therefore, searching gene expression databases is a complementary strategy to searching strain-specific sequence databases. Numerous expressed sequence tag (EST) and microarray databases have been generated to evaluate gene expression in many tissues from multiple species (Table 2).

One approach to searching expression databases is to identify those genes within a QTL interval that are expressed in a particular tissue. If a QTL is thought or known to affect the function of a particular tissue, it is logical to focus on candidate genes that are expressed in that tissue. For example, because kidney transplant studies demonstrated that the causal gene underlying *Rf-1* was likely expressed in the kidney, Vitt *et al.* searched EST databases to identify those genes within the QTL expressed in kidney (90 of 1029 genes), which substantially reduced the number of candidate genes [15]. ExQuest is a new bioinformatics tool for searching EST databases that reports tissue or developmental expression patterns for genes in the context of chromosomal intervals [40]. SymAtlas [41], based on direct measurement of gene expression by microarrays, provides an alternative resource to examine a chromosomal interval to screen positional candidate genes for expression in a tissue of interest.

An alternative method of searching expression databases is to identify genes that are differentially expressed between strains. Digital differential display compares the gene expression levels, based on Unigene abundance, of two or more selected cDNA libraries to identify genes that are differentially expressed. The Gene Expression Omnibus [42], ArrayExpress [43] and The Institute for Genome Research (<http://www.tigr.org>) each provide a searchable

database containing numerous microarray experiments examining strain-to-strain comparisons, the effects of stimuli on gene expression and developmental expression changes. These databases enable direct comparison of gene expression among strains for many tissues and will be increasingly useful as data from more tissues, strains and species are made available.

Expression QTL (eQTL) analysis is a new experimental strategy to link gene expression with the underlying genes that regulate expression levels by treating expression data in a segregating population as the phenotypic trait for QTL analysis [44]. Several exciting articles have connected eQTL with physiological QTL to identify differentially expressed candidate genes within the QTL interval that are regulated by *cis*-acting elements. Because expression data from recombinant inbred (RI) strains are accumulating in public databases, it is now possible to perform eQTL analysis based solely on bioinformatics. This feature is available on WebQTL (<http://www.genenetwork.org>), which contains genotype and expression data from several RI strain sets. The use of RI strain sets for this bioinformatics tool means that it will become more useful as additional expression studies are performed on the RI strains over time – a significant advantage of using RI strains, rather than intercross progeny, as the segregating population.

Limitations

As with sequence databases, the quality of data deposited in an expression database will determine the quality of results from bioinformatics tools based on the data. Thus, it is important to evaluate the quality control procedures governing the deposition of data in expression databases. More importantly, any differences identified using these bioinformatics methods should be confirmed experimentally. It is also important to determine which splice variants are represented in public databases because differential expression of splice variants could underlie a QTL. However, there is little public information regarding splice variants, which represents an area for future improvement in the databases.

The ability to determine differential expression between strains or to perform eQTL analysis requires that data from the appropriate strains and tissues be available in the database. Because inbred strains provide genetically reproducible offspring, data from inbred strains can accumulate over time to build up informative databases containing gene expression information from various tissues of inbred strains at different ages and under different environmental conditions.

An integrated bioinformatics strategy for narrowing QTLs

Comparative genomics provides an initial coarse analysis of a QTL interval to identify regions of chromosomal synteny in QTL across species. Generally, comparative genomic analysis can reduce confidence intervals by approximately half [15], although this will depend on the number of species used in the comparison and the size of syntenic blocks between the species. Combined cross analysis is also effective for reducing QTL intervals by

the same amount [23], depending on the number of crosses and strains combined and the genotype density. Despite both methods being equally effective for reducing a QTL interval, comparative genomic analysis requires only one QTL that is concordant in multiple species, whereas combined cross analysis requires a shared QTL from multiple crosses in only one species. Thus, these methods provide complementary bioinformatics approaches for initially dissecting a QTL interval.

Haplotype analysis is useful to refine further a QTL interval to several Mb. Whereas combined cross analysis compares intercross and backcross populations to refine a QTL region, interval-specific haplotype analysis focuses on regions that are not IBD within QTLs shared by multiple crosses. It is important to apply combined cross analysis to overlapping QTLs before interval-specific haplotype analysis to verify that the interval is narrowed and does not separate into distinct peaks to bolster the assumption that the same causal gene underlies both QTLs. Interval-specific haplotype analysis is often used to identify small regions (<5 Mb) within a QTL that are likely to contain the causal gene [24,28,29], although haplotype analysis can lead to the causal gene when the QTL is shared by many crosses [30]. Genome-wide haplotype association also assumes IBD, but analyzes haplotype patterns across the whole genomes of inbred strains surveyed for a phenotype to associate that phenotype with conserved haplotype patterns. The interval is determined by the size of the associated haplotype pattern and the SNP density available for testing its size, but genome-wide haplotype association can detect small (<5 Mb) regions associated with a phenotype [36]. Unlike interval-specific haplotype analysis, which is most useful for narrowing a QTL shared by multiple crosses, genome-wide haplotype analysis requires only phenotype information from many inbred strains and can effectively narrow a QTL identified in only one experimental cross [36].

After narrowing the QTL to an interval that is <5 Mb using these bioinformatics techniques or classical experimental methods, strain-specific sequence and gene expression comparisons are effective for focusing on a few strong candidate genes (Figure 7). Although these tools can be applied to a QTL interval of any size, they evaluate each positional candidate gene within the region, so a shorter input list of genes will result in a more focused output of candidate genes. It is virtually impossible to know *a priori* whether the polymorphism in the causal gene underlying a QTL affects function or expression, so both methods should be used together. Ideally, strain-specific databases containing mRNA and protein expression in numerous tissues could be interrogated, but such broad databases are currently unavailable. The ongoing development of expansive expression databases along with genomic sequence of additional mouse and rat strains will enhance the power of future sequence and expression comparisons for narrowing QTLs.

Conclusion: return to the bench

The bioinformatics methods described here and the public databases, which are improving, provide new tools for the

- 11 Patrinos, G.P. and Brookes, A.J. (2005) DNA, diseases and databases: disastrously deficient. *Trends Genet.* 21, 333–338
- 12 Darvasi, A. (1998) Experimental strategies for the genetic dissection of complex traits in animal models. *Nat. Genet.* 18, 19–24
- 13 Pennacchio, L.A. (2003) Insights from human–mouse genome comparisons. *Mamm. Genome* 14, 429–436
- 14 O'Brien, S.J. *et al.* (1999) The promise of comparative genomics in mammals. *Science* 286, 458–462, 479–481
- 15 Vitt, U. *et al.* (2004) Identification of candidate disease genes by EST alignments, synteny, and expression and verification of Ensembl genes on rat chromosome 1q43–54. *Genome Res.* 14, 640–650
- 16 Cusi, D. *et al.* (1997) Polymorphisms of α -adducin and salt sensitivity in patients with essential hypertension. *Lancet* 349, 1353–1357
- 17 Bianchi, G. *et al.* (1994) Two point mutations within the adducin genes are involved in blood pressure variation. *Proc. Natl. Acad. Sci. U. S. A.* 91, 3999–4003
- 18 Ueda, H. *et al.* (2003) Association of the T-cell regulatory gene *CTLA4* with susceptibility to autoimmune disease. *Nature* 423, 506–511
- 19 Korstanje, R. *et al.* (2004) Locating *Ath8*, a locus for murine atherosclerosis susceptibility and testing several of its candidate genes in mice and humans. *Atherosclerosis* 177, 443–450
- 20 Wang, X. *et al.* (2005) Positional identification of *TNFSF4*, encoding OX40 ligand, as a gene that influences atherosclerosis susceptibility. *Nat. Genet.* 37, 365–372
- 21 Wade, C.M. *et al.* (2002) The mosaic structure of variation in the laboratory mouse genome. *Nature* 420, 574–578
- 22 Frazer, K.A. *et al.* (2004) Segmental phylogenetic relationships of inbred mouse strains revealed by fine-scale analysis of sequence variation across 4.6 Mb of mouse genome. *Genome Res.* 14, 1493–1500
- 23 Li, R. *et al.* (2005) Combining data from multiple inbred line crosses improves the power and resolution of QTL mapping. *Genetics* 169, 1699–1709
- 24 DiPetrillo, K. *et al.* (2004) Genetic analysis of blood pressure in C3H/HeJ and SWR/J mice. *Physiol. Genomics* 17, 215–220
- 25 Ishimori, N. *et al.* Quantitative trait loci that determine bone mineral density in C57BL/6J and 129SvImJ inbred mice. *J. Bone Miner. Res.* (in press)
- 26 Perusse, L. *et al.* (2005) The human obesity gene map: the 2004 update. *Obes. Res.* 13, 381–490
- 27 Cuppen, E. (2005) Haplotype-based genetics in mice and rats. *Trends Genet.* 21, 318–322
- 28 Wiltshire, T. *et al.* (2003) Genome-wide single-nucleotide polymorphism analysis defines haplotype patterns in mouse. *Proc. Natl. Acad. Sci. U. S. A.* 100, 3380–3385
- 29 Park, Y.G. *et al.* (2003) Multiple cross and inbred strain haplotype mapping of complex-trait candidate genes. *Genome Res.* 13, 118–121
- 30 Wang, X. *et al.* (2004) Haplotype analysis in multiple crosses to identify a QTL gene. *Genome Res.* 14, 1767–1772
- 31 Grupe, A. *et al.* (2001) *In silico* mapping of complex disease-related traits in mice. *Science* 292, 1915–1918
- 32 Smith, J.D. *et al.* (2003) *In silico* quantitative trait locus map for atherosclerosis susceptibility in apolipoprotein E-deficient mice. *Arterioscler. Thromb. Vasc. Biol.* 23, 117–122
- 33 Darvasi, A. (2001) *In silico* mapping of mouse quantitative trait loci. *Science* 294, 2423
- 34 Chesler, E.J. *et al.* (2001) *In silico* mapping of mouse quantitative trait loci. *Science* 294, 2423
- 35 Liao, G. *et al.* (2004) *In silico* genetics: identification of a functional element regulating H2-Ealpha gene expression. *Science* 306, 690–695
- 36 Pletcher, M.T. *et al.* (2004) Use of a dense single nucleotide polymorphism map for *in silico* mapping in the mouse. *PLoS Biol.* 2, e393
- 37 Wang, X. *et al.* (2003) Using advanced intercross lines for high-resolution mapping of HDL cholesterol quantitative trait loci. *Genome Res.* 13, 1654–1664
- 38 Yalcin, B. *et al.* (2004) Unexpected complexity in the haplotypes of commonly used inbred strains of laboratory mice. *Proc. Natl. Acad. Sci. U. S. A.* 101, 9734–9739
- 39 Marshall, K.E. *et al.* (2002) *In silico* discovery of gene-coding variants in murine quantitative trait loci using strain-specific genome sequence databases. *Genome Biol.* 3, DOI:10.1186/gb-2002-3-12-research0078 (<http://genomebiology.com/2002/3/12/research/0078>)
- 40 Brown, A.C. *et al.* (2004) ExQuest, a novel method for displaying quantitative gene expression from ESTs. *Genomics* 83, 528–539
- 41 Su, A.I. *et al.* (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl. Acad. Sci. U. S. A.* 101, 6062–6067
- 42 Edgar, R. *et al.* (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* 30, 207–210
- 43 Brazma, A. *et al.* (2003) ArrayExpress – a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res.* 31, 68–71
- 44 de Koning, D.J. and Haley, C.S. (2005) Genetical genomics in humans and model organisms. *Trends Genet.* 21, 377–381
- 45 Hillebrandt, S. *et al.* (2005) Complement factor 5 is a quantitative trait gene that modifies liver fibrogenesis in mice and humans. *Nat. Genet.* 37, 835–843
- 46 Galperin, M.Y. (2005) The molecular biology database collection: 2005 update. *Nucleic Acids Res.* 33(Database issue), 5–24

Elsevier.com – Dynamic New Site Links Scientists to New Research & Thinking

Elsevier.com has had a makeover, inside and out.

As a world-leading publisher of scientific, technical and health information, Elsevier is dedicated to linking researchers and professionals to the best thinking in their fields. We offer the widest and deepest coverage in a range of media types to enhance cross-pollination of information, breakthroughs in research and discovery, and the sharing and preservation of knowledge. Visit us at Elsevier.com.

Elsevier. Building Insights. Breaking Boundaries