Reviews

Incorporating Mouse Genome
Mammalian Genome
Genes and Phenotypes

Integrating genetic and gene expression data: application to cardiovascular and metabolic traits in mice

Thomas A. Drake,[1] Eric E. Schadt,[2] Aldons J. Lusis[3]

[1]Department of Pathology and Laboratory Medicine, University of California, Los Angeles, Los Angeles, California 90095-1732, USA
[2]Rosetta Inpharmatics, 401 Terry Avenue North, Seattle, Washington 98109, USA
[3]Department of Human Genetics, Department of Medicine and Department of Microbiology, Immunology, and Molecular Genetics, and Molecular Biology Institute, University of California, Los Angeles, Los Angeles, California 90095-1679, USA

Received: 8 December 2005 / Accepted: 21 February 2006

Abstract

The millions of common DNA variations that occur in the human population, or among inbred strains of mice and rats, perturb the expression (transcript levels) of a large fraction of the genes expressed in a particular tissue. The hundreds or thousands of common *cis*-acting variations that occur in the population may in turn affect the expression of thousands of other genes by affecting transcription factors, signaling molecules, RNA processing, and other processes that act in *trans*. The levels of transcripts are conveniently quantitated using expression arrays, and the *cis*- and *trans*-acting loci can be mapped using quantitative trait locus (QTL) analysis, in the same manner as loci for physiologic or clinical traits. Thousands of such expression QTL (eQTL) have been mapped in various crosses in mice, as well as other experimental organisms, and less detailed maps have been produced in studies of cells from human pedigrees. Such an integrative genetics approach (sometimes referred to as ''genetical genomics'') is proving useful for identifying genes and pathways that contribute to complex clinical traits. The coincidence of clinical trait QTL and eQTL can help in the prioritization of positional candidate genes. More importantly, mathematical modeling of correlations between levels of transcripts and clinical traits in genetic crosses can allow prediction of causal interactions and the identification of ''key driver'' genes. An important objective of such studies will be to model biological networks in physiologic processes. When combined with high-density single nucleotide polymorphism (SNP) mapping, it should be feasible to identify genes that contribute to transcript levels using association analysis in outbred populations. In this review we discuss the basic concepts and applications of this integrative genomic approach to cardiovascular and metabolic diseases.

Correspondence to: Aldons J. Lusis, Department of Medicine/Division of Cardiology, 47-123 Center for the Health Sciences, 650 Charles E. Young Drive South, University of California, Los Angeles, Los Angeles, CA 90095-1679, USA; E-mail: jlusis@mednet.ucla.edu

Introduction

Cardiovascular and metabolic diseases develop as a consequence of a complex cascade of events, as depicted conceptually in Fig. 1. An individual inherits a set of alleles (genotype) from her or his parents and, in combination with environmental factors (lifestyle), these determine physiologic states such as lipoprotein levels, adiposity, and immune functions. These, in turn, can result in abnormalities such as vessel wall dysfunction, hypertension, and insulin resistance. Over many years these factors influence the development of chronic diseases such as diabetes, kidney disease, atherosclerosis, and myocardial infarction. Most of these interactions are likely to be influenced by genetic factors. A major goal of medical research is to define the various causal interactions and genetic factors involved. This knowledge would be useful in the rational design of maximally effective, minimally toxic drugs to treat these disorders. It would also provide a framework for understanding how genetic variations interact with an individual's sex and environment to influence the onset and progression of the diseases.

At present, our overall understanding of the pathways for metabolic and cardiovascular diseases

 DOI: 10.1007/s00335-005-0175-z • Volume 17, 466–479 (2006)

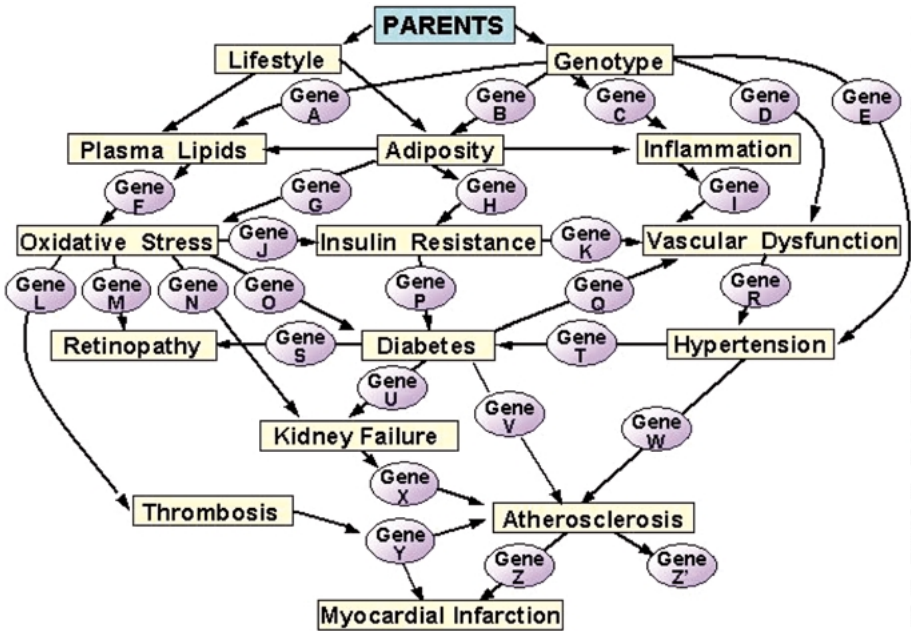

Fig. 1. Cascade of interactions in metabolic and cardiovascular disorders. This cartoon depicts some of the interactions thought to contribute to the development of diabetes, atherosclerosis, and their complications. Most of the interactions are likely to involve genetic factors (genes A–Z). Some genes may be markers of clinical disease (gene Z′).

is fragmentary, despite impressive advances in certain areas. Some of the elements involved have been defined in studies of Mendelian disorders and model organisms such as mice and rats. For example, over 60 Mendelian mutations are associated with increased atherosclerosis and over 200 are associated with diabetes (Rizvi et al. 2002). Transgenic studies in mice have also been very informative (Biddinger and Kahn 2005). For example, at present, well over 100 knockout or transgenic mice have been shown to exhibit differences in atherosclerotic lesion development. However, whether such extreme variations are physiologically relevant for common forms of the disorders is unclear, and many of these findings with transgenic animals may in fact represent artifacts resulting from mixed genetic backgrounds (discussed in Ghazalpour et al. 2004).

Efforts to dissect complex forms of metabolic and cardiovascular diseases have been only modestly successful. Positional cloning of genes for these diseases in humans has proven very difficult, although there have been some notable successes (Lusis et al. 2004). Candidate gene studies have revealed a number of genes that appear to be important, but most of the thousands of positive associations reported over the past 20 years have come from studies that were underpowered and a large fraction are likely to be attributable to problems such as ethnic admixture and publication bias. Studies in experimental organisms, primarily mice and rats, have resulted in the identification of hundreds of QTL but few genes or mechanistic insights (Ghazalpour et al. 2004). These studies have made it clear that metabolic and cardiovascular disorders are very complex, resulting from a large number of factors, each exhibiting a small effect.

Most studies of metabolic and cardiovascular disorders have used a classic ''one gene at a time'' approach. This approach has been very successful in clarifying Mendelian traits such as familial hypercholesterolemia and in elucidating individual pathways such as the regulation of sterol metabolism. However, complex traits such as the common forms of atherosclerosis present special problems. In particular, it is likely that most complex diseases result from the interactions of multiple genes and, therefore, cannot be realistically modeled by single-gene perturbations. One promising solution is a global strategy. Global approaches have been made possible by several technical and conceptual advances during the past decade. These include, of course, the Human Genome Project, which has provided a ''genetics parts list'' of the human genome and the genomes of many model organisms, including mice and rats. They also include rapid genotyping methods, DNA arrays for global quantitation of transcript abundances (both protein coding and noncoding), and the Internet, which make dissemination of global data sets possible. Global proteomic and metabolomic approaches are being developed (Weston and Hood 2004).

One of the most exciting applications of global data sets involves the combination of natural genetic variation and expression array analysis, sometimes referred to as ''genetical genomics'' (Jansen and Nap 2001). Just as environmental or single-gene

perturbations can be used in conjunction with genome-wide expression array analyses to identify regulatory links, the measurement of transcript levels in populations allows the identification of coregulated genes and of relationships between transcript levels and clinical traits. In addition, this integrative approach relates DNA variation to transcript abundance, allowing identification of primary (*cis*-acting) and secondary (*trans*-acting) effects controlling transcript abundance. The integration of genetic and transcript abundance data has recently been used to predict causal interactions in the complex trait of adiposity; this kind of analysis should be applicable to any complex trait (Schadt et al. 2005).

Although the insights that can be gained through the genetical genomics approach are extremely promising, the concepts and implications are often difficult to grasp, especially for those not working in the area of quantitative genetics. There are several reasons for this, but a major one is that the extensive data generated can be approached and analyzed in many ways, with different but often overlapping questions being asked and often times sophisticated and unfamiliar statistical approaches being used. For this reason, we review the basic underlying concepts and logic from a generalist perspective and then discuss a number of the applications that have been used to study metabolic and vascular disease in the mouse model. We conclude with a perspective on areas of difficulty and challenges for the future.

Basic concepts of interpreting expression QTL

The concept of mapping transcript levels is not new. Practitioners of QTL analysis appreciate that any quantitative trait can be mapped, and measuring relative transcript levels with reasonable accuracy has been feasible for quite some time. Initial applications were restricted to relatively few transcripts in any one study because of the technical difficulty in measuring multiple transcripts in tissue samples from hundreds of animals (Lan et al. 2003; Machleder et al. 1997). These studies revealed some of the basic observations made by larger-scale studies, namely, that transcript levels were often controlled by more than one locus and that these did not necessarily coincide with the location of the respective gene. However, the availability of microarray technology has made it possible to measure tens of thousands of transcripts simultaneously (in theory, all transcribed genes and noncoding RNA, were that to be known), though cost is a major impediment for most investigators. Analysis of so many traits raises unique technical issues related to processing and handling of large-scale data sets, as well as complex statistical concerns, which we will address as relevant in our discussion of applications below. More importantly, the simultaneous analysis of a large fraction of expressed genes in the setting of a genetic cross enables analyses and insights never before possible.

Just as with ''traditional'' traits studied in QTL analyses, the identification of a QTL for a gene transcript (an eQTL) implies that there is a genetic sequence variation within the genomic region encompassed by the QTL that directly or indirectly influences transcript levels (Flint et al. 2005). When an eQTL encompasses the physical location of the gene for that transcript, it is likely that the causative genetic variation resides within the gene itself (i.e., the transcript is being regulated in *cis*, and so the respective eQTL is referred to as a *cis*-eQTL). Conversely, when an eQTL does not encompass the physical location of the gene for that transcript, the causative genetic variation does not reside within the gene itself. The expression of the transcript in that case is regulated in *trans* (i.e., under the control of a different gene or genes physically located at that locus, and the eQTL in this case is referred to as a *trans*-eQTL).

The finding of an eQTL for a gene therefore indicates that the gene transcript levels are influenced by genetic factors. In the mouse F_2 populations that we have studied, at least 30% of all expressed genes in a given tissue have one or more eQTL. If one is interested in those genes that contain polymorphisms or mutations that are responsible for a specific phenotypic difference between individuals in the population being studied (e.g., susceptibility to atherosclerosis), then it is the set of genes with *cis*-eQTL that is of interest. In other words, these genes are the genetic drivers for the phenotypic variation among individuals. On the other hand, those genes with only *trans*-eQTL play a role in the expression of phenotypic variation, but they are responding to other genetic factors rather than being primary drivers.

For an extreme example of this concept, consider a single-gene mutation such as the *Ob* mutation (leptin deficiency due to a mutation in the *Leptin* gene) that leads to obesity and other consequences in affected mice. When compared with the background strain, the *Ob* mouse is genetically identical except for the *Leptin* gene mutation. An expression microarray analysis of liver, however, would show differences in the transcript levels of at least hundreds of genes between *Ob* mice and the background B6 strain. Some of these would be genes directly controlled by leptin, others would be genes far removed, with altered levels due to the massive obesity that

results. In this setting, the *Leptin* gene would be considered the *cis* variation, and all the reacting genes the *trans* variations. The same concept holds in an F_2 population, except that there are simultaneously many *cis* variations (*cis*-eQTL are one example), each with consequent effects on a variable number of secondarily affected genes (the *trans*-eQTL). In our experience with F_2 intercrosses between common inbred mouse strains, the fraction of total genes with *cis*-eQTL in a single tissue is approximately 10%, and the fraction with *trans*-eQTL is several-fold larger. When one considers multiple tissues, the fraction of total genes with *cis*-eQTL will be greater.

For practical purposes, we have assigned a gene as *cis*-acting if the eQTL mapped to within 20 cM of the physical location of the gene. As QTL for any trait are relatively broad and encompass from tens to hundreds of genes, the fact that the physical location of a gene coincides with an eQTL for its transcript does not exclude the possibility that it is in fact a *trans*-eQTL controlled by closely linked genes. We have experimentally validated that the majority of presumptive *cis*-eQTL are true *cis*-eQTL by the application of a classic *cis-trans* test in which (B6 × DBA)F_1 mice were analyzed for the relative levels of transcript from each allele (Doss et al. 2005), as depicted in Fig. 2. In a *cis*-regulated gene, the allelic expression in F_1 mice should be similar to the ratios observed in the two parental strains. If the gene is regulated in *trans*, we expect the allelic ratio of transcripts from each allele to be approximately 1:1, because a truly *trans*-acting regulator should act in a similar manner on both alleles. A gene could exhibit a combination of *trans*- and *cis*-regulation, although the *cis*-regulatory component should cause the ratio to be different than 1:1, albeit to a lesser extent than might otherwise be the case. Our finding confirmed that for analyses of F_2 data, it is reasonable to consider that an eQTL falling within 20 cM of the gene location is indicative of *cis*-acting regulation. Where critical, a *cis-trans* test can be used to definitively determine this. The ability to identify *cis*-acting genes is a key aspect of deriving causative associations from the genomic expression data as discussed below.

Levels of a given transcript may be significantly influenced by many genes, as one would expect. In this case, multiple eQTL may be identified, which may all be *trans*-eQTL or a combination of both *cis*- and one or more *trans*-eQTL (Fig. 3). In our studies there is a strong inverse relationship between LOD score and the likelihood of a given eQTL being *cis* or *trans* in nature. Thus, the number of eQTL detected

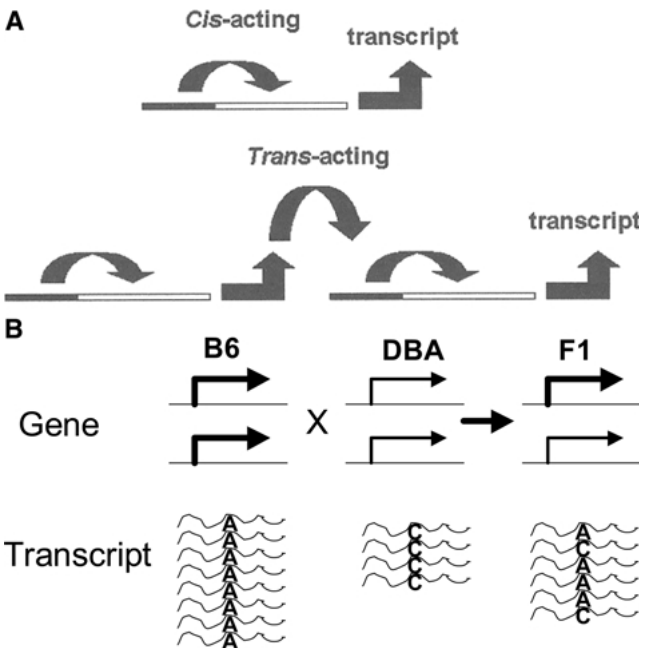

Fig. 2. *Cis-* and *trans*-regulation of transcript abundance. (**A**) In the case of *cis*-regulation, the eQTL for a transcript would map over the gene encoding the transcript. In the case of *trans*-regulation, DNA change influencing transcript abundance occurs in a gene different from that encoding the transcript, so that the eQTL would ordinarily not be expected to map over the gene encoding the transcript. (**B**) The classis *cis-trans* test examines the amount of product derived from each allele in an F_1 heterozygote. This can be done by distinguishing the transcripts using a SNP present in the transcript. In the example shown here, the B6 allele is twice as active and the DBA, and for a *cis*-regulated gene, the F_1 would have a 2:1 ratio of the B6 transcript to the DBA transcript.

per gene and the relative frequency of *cis*- versus *trans*-eQTL across all genes will be highly dependent on the LOD threshold one sets for defining an eQTL. As discussed below, in some settings it is desirable to have stringent thresholds and in other settings have relatively low thresholds. The determination of expected false discovery rates (FDRs) at different thresholds can assist in selecting the LOD threshold to use in a given situation.

Transcript levels may also be significantly affected by factors that do not result in eQTL but may influence the occurrence and magnitude of detected eQTL. Genetic regulation that is dispersed among many loci with small effects is one situation, and of course environmental influence is another. We have also observed pervasive effects of sex on gene expression and eQTL detection in multiple tissues and now routinely use a genetic model approach for QTL detection that incorporates sex as both an additive and an interactive factor in data sets that include both sexes (unpublished data).

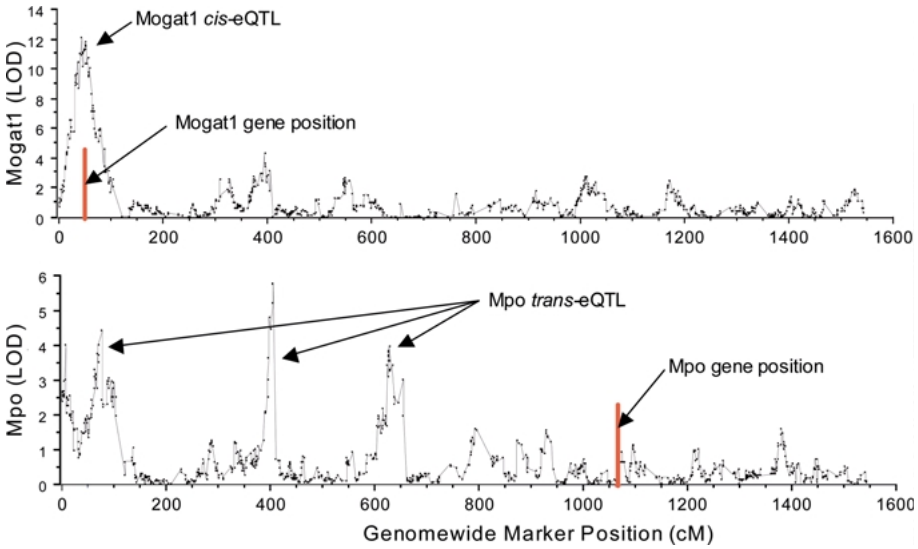

Fig. 3. Examples of *cis*- and *trans*-eQTL. For two representative genes, LOD curve plots are given showing the relationship of eQTL with anatomic gene position. The linkage map is shown for all autosomes, beginning at the top of Chromosome 1 and ending at the bottom of Chromosome 19. The *Mogat1* gene (red bar) is located at about 50 cM on Chromosome 1 and the only major LOD score peak for *Mogat1* transcript abundance is coincident with the gene, suggesting *cis*-regulation. The *Mpo* gene (red bar) is not coincident with the LOD score peaks for *Mpo* transcript abundance (located at about 80, 400, and 620 cM), indicating *trans*-regulation.

Applications of the genetical genomics approach relevant to disease models

The broad goal of genetical genomics experiments is to identify genes and pathways that contribute to complex clinical traits, and to understand how these function as a whole in normal and disease states. The wealth of data typically generated allows for many different, often overlapping, questions to be asked. In this section we discuss five specific applications of the genetical genomics approach that are relevant to metabolic and vascular disease (Fig. 4).

Prioritizing candidate genes responsible for clinical trait QTL. The traditional goal of QTL analyses has been the identification of the causative gene or genes responsible for a disease-related trait QTL (cQTL). The cQTL defines a genomic region that encompasses the causal gene, but the region defined is typically broad, with tens to hundreds of genes residing there. Applying traditional positional cloning methods to identify the causal gene is slow. Because variations in genes affecting function act through regulation of transcript levels in at least one half to two thirds of cases (and even in cases where variations affect protein function, as in the case of a nonsense mutation, transcript levels may still be affected due to nonsense-mediated decay, changes in nuclear transport, etc.), analysis of transcript levels for genes residing in the QTL region is informative in many instances. Such genes for which transcript levels are genetically regulated will have *cis*-eQTL colocalizing with the cQTL and therefore constitute a restricted set of candidate genes. Although in many instances colocalization of *cis*-eQTL and cQTL is obvious, it is best to use an objective statistical method of defining colocalization. We have developed and applied a ''close linkage versus pleiotropy'' test for this purpose and shown that tests for causality can also elucidate the relationship between eQTL and cQTL, which is applicable to other situations where colocalization of QTL is important to define (Drake et al. 2001; Schadt et al. 2005).

At this stage, the number of genes with *cis*-eQTL colocalizing with the cQTL is still greater than one can reasonably approach experimentally (e.g., through transgenic construction), and additional analyses are useful to further narrow the list. One would expect transcript levels of the causative gene to be significantly correlated with the clinical trait values, so determining Pearson or Spearman correlation coefficients between these is helpful. Another test we have applied is multitrait QTL analysis, where the clinical trait and the expression traits for genes with *cis*-eQTL are analyzed jointly. Finding a significant increase in LOD score with joint analysis is supportive of a causal relationship. Publicly available genomic and other data can also be brought to bear, as recently reviewed by Paigen and colleagues (Dipetrillo et al. 2005). The cQTL region under consideration can be mapped for regions that are genetically distinct between the parental strains versus regions that appear to be identical by descent (IBD) on the basis of SNP frequency (high frequency in the former, low in the latter) (Davis et al. 2005; Wade et al. 2002; Wiltshire et al. 2003). Candidate genes residing in a region with a high SNP frequency are more likely to be ''real'' than those that are IBD. Primary sequence data where available could be similarly used for determining IBD regions. Primary sequence can also be examined for functional mutations between the parental strains that may

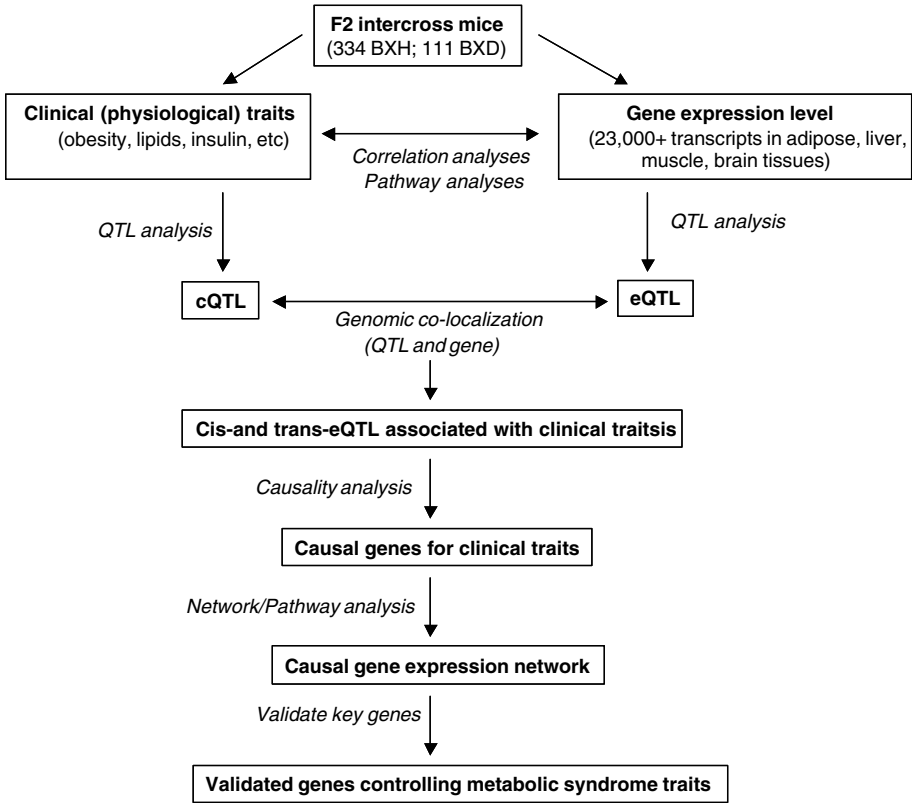

Fig. 4. Applications of integrative genomics. See text for discussion.

help prioritize among the remaining candidate genes. Finally, consistency of results for individual candidate genes among different crosses or strains for which phenotype, genotype, and expression data are available can be informative, as we have recently shown (Cervino et al. 2005).

Genetical genomic data can also be used in a converse manner to exclude potential candidate genes, as Attie and colleagues have shown (Lan et al. 2004). They demonstrated an unequivocal strong *cis*-eQTL (LOD 30) for a candidate gene for diabetes (*Pdi*), which did not colocalize with any QTL for diabetes-related traits, thus distinguishing covariation from potential causation.

Despite all of these available tools, the list of potential candidate genes may be relatively large, requiring finer mapping or other methods. This is one of the challenges for the future, as we discuss below. Also, depending on the microarray platform used, the coverage of the genes (i.e., the fraction of genes physically located in the cQTL region that are represented on the array) varies and is never 100%, necessitating additional work to be complete.

There are two important caveats in considering the relationship of eQTL and cQTL. The first is that gene expression patterns are cell and tissue specific. Complex clinical traits such as those related to metabolic and vascular diseases involve multiple organs and tissues. Therefore, the genetical genomic data obtained for any given tissue represent only a part of the whole, and one may be misled if a critical tissue is not examined. The second caveat is that regulation of biological processes occurs via control of gene expression in many, but certainly not all, situations; for example, missense or splicing variation could have functional consequences without altering transcript levels. When large-scale gene expression sets are obtained, it is not at all obvious from the data whether the trait QTL is a consequence of control at the level of gene expression.

Defining molecularly distinct subtypes within a population for a clinical phenotype. It has long been appreciated that complex disease phenotypes that appear similar among individuals may actually have distinct molecular subtypes. A classic example is diabetes mellitus, for which it took years to recognize the major distinctions between type I and type II forms. More recently, use of gene expression profiles in cancer has allowed identification of biologically distinct subtypes of disease, which could otherwise not be distinguished yet which carry significant therapeutic and prognostic implications (Bucca et al. 2004).

In a comparable manner, gene expression profiling can be applied to look for the presence of

biological subtypes for specific clinical traits relevant to metabolic and vascular diseases. We have done this for the trait of body fat in a (B6 × DBA) F_2 population in which expression microarray data were available from liver samples. Two hundred eighty transcripts were identified that showed the strongest correlation with the fat mass trait. For mice in the upper and lower 25th percentiles of the trait, these transcripts were then subjected to bidirectional hierarchical clustering. Among animals with high fat pad mass, two distinct subgroups were identified based on expression patterns (Fig. 5). When QTL analysis was performed for the two subgroups separately, distinct QTL were identified, one corresponding to the originally identified Chromosome 2 locus, the other to proximal Chromosome 19, a locus not identified on the initial QTL analysis of the full F_2 set. Thus, application of the eQTL approach identified what are presumably distinct physiologic subtypes of the fat pad mass trait. As discussed below, this subtyping was further supported by analysis of metabolic pathways in this data set. Thus, subgroup analysis can improve sensitivity of QTL analyses and enhance understanding of disease mechanisms.

Identifying known cellular or metabolic "pathways" that are involved in disease states. Differential gene expression between animals differing in extent of clinical trait expression can be used to identify pathways involved in disease pathogenesis. Changes in the transcriptome in association with complex metabolic traits typically involve sets of functionally related transcripts. The standard approach for identifying these is to determine which transcripts are statistically significantly different between two groups (e.g., between obese and lean mice) or are significantly correlated with a particular trait across the population, and then to determine whether particular categories of genes are overrepresented in this subset relative to all genes examined. The categories typically used include metabolic and signaling pathways and the GO ontology categories, among others (Diehn et al. 2003; Kanehisa et al. 2004). Most gene expression analysis software suites incorporate these analyses. However, with this approach, small but important consistent shifts in each individual gene composing a pathway set may not be recognized because of statistical considerations. Methods that examine a pathway as a set rather than as individual genes can detect significant coordinated changes. An analytic tool for this, termed GSEA (Gene Set Enrichment Analysis) has been developed by Mootha and colleagues (Mootha et al. 2003; Subramanian et al. 2005).

We have applied both the Fisher exact test for overrepresentation and the GSEA approach to data from the BXD cross to identify pathways associated with subcutaneous fat (Ghazalpour et al. 2005). As an initial step for both methods, 387 gene sets representing pathways or comparable functionally related genes were assembled from a variety of sources or compiled from primary sources. These included the KEGG database of metabolic pathways (Kanehisa et al. 2004), GenMapp sets (Dahlquist et al. 2002), and Biocarta signaling pathways (http://www.biocarta.com/genes/index.asp). The subsets of mice for study were selected as those in the top and bottom 15th percentiles of the subcutaneous fat mass trait, so that lean and obese sets containing equal numbers of animals were analyzed. For analysis of pathway overrepresentation by the Fisher exact test, a discrimination score was obtained for each transcript, determined by comparing the mean transcript levels between lean and obese mice using Student's t test. Transcripts with associated p values for the t test of less than 0.01 were identified as being differentially regulated in relation to the abdominal fat mass trait, and the Fisher exact test was applied as implemented in the EASE program to identify those gene sets that were overrepresented among these (Hosack et al. 2003). For analysis by the GSEA procedure, a subset of the genes on the array was identified that showed differential regulation (individual p value < 0.05) in 10% or more of all mice. Among this set of approximately 5000 differentially regulated genes, the microarray expression data were then ranked based on the magnitude of differences in transcript levels between the groups and an "Enrichment Score" calculated as described for each of the predetermined pathway/gene sets (Mootha et al. 2003). The calculated Gene Enrichment Score was then used to rank each pathway in our gene list. To determine if the ranking of a high-scoring pathway occurred by chance or if there was biological significance assigned to it, we permuted the class assignment of the high and low F_2 mice and recalculated the gene ranking, Gene Enrichment Scores, and ranking of each pathway. Similar pathways were identified for the most part by each approach, but the GSEA procedure had somewhat greater sensitivity for detecting significance. The majority of pathways identified are interrelated metabolically in that they feed into the tricarboxylic acid (TCA) cycle, and the second grouping of pathways was related to cholesterol metabolism (Ghazalpour et al. 2005).

The above analyses do not depend on the genetic data from the experiment, but genetic data can be incorporated by analyzing the eQTL of transcripts belonging to identified pathways or gene sets. If an

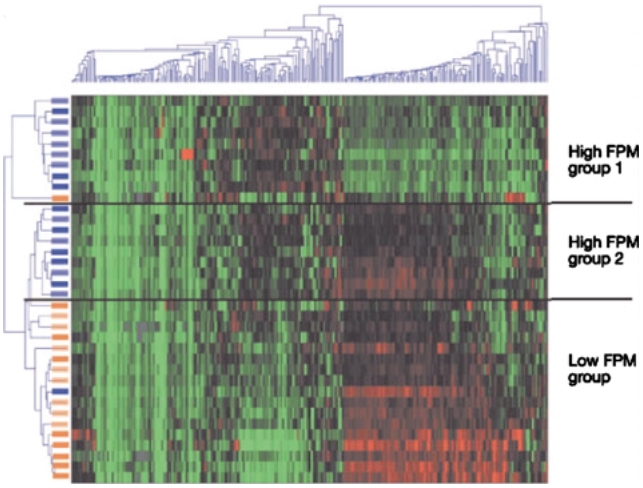

Fig. 5. Molecularly distinct subtypes within a population of F_2 mice for the trait of body fat. These data were produced by microarray analyses for about 23,000 transcripts using an Agilent platform. The F_2 cross was between strains DBA/2J and C57BL/6J and liver RNA was profiled. The color matrix display for hierarchically clustered genes (*x* axis) and extreme fat pad mass (FPM) (*y* axis). Dark/light blue bars indicate mice in the upper/lower of the high FPM group, and dark/light orange indicate mice in the lower/upper half of the low FPM group. Subdivision of mice in this manner defined groups in which FPM was influenced by distinct genetic loci (from Schadt et al. 2003, with permission)

entire pathway is under coordinated genetic control, then there ought to be colocalization of eQTL for pathway genes at the specific loci exerting that control. This is in fact what we observed. We also found that several of the identified loci corresponded to QTL for the obesity trait, as one would expect. These analyses also allow for assigning tentative function to uncharacterized genes that show strong correlation with pathway genes and have similarly colocalizing eQTL. A limitation of this approach is that it will assess only predefined gene sets, so novel pathways, or those in which too few member genes are known, will not be identified. It should also be noted that the most appropriate statistical way to assess whether a given pathway is enriched in a gene set is still an open question. This is because the Fisher exact test and other standard statistical tests like the Kolmogorov test (one of the standard tests that can be used as part of the GSEA procedure) are based on assumptions that are usually not true for gene expression data, related to the fact that traits are often correlated and the categories being searched can be hierarchical.

Identifying genes that are causally related to clinical trait expression. In complex clinical traits such as those associated with the metabolic syndrome, there are multiple QTL associated with each trait, and associated pathways involving expression of secondary genes, ultimately leading to the expression of the clinical trait. There will also be genes whose expression is influenced as a consequence of the trait itself (i.e., reactive) rather than being within the pathway leading to trait expression (i.e., causal). For example, referring to Fig. 1, genes H and J would be causal for the insulin-resistance trait, while genes K and P would be reactive. Typical gene expression experiments identify a set of genes whose expression is significantly correlated with a trait or disease of interest. However, these gene sets are composed of both causal and reactive genes, and because there is no way to differentiate between the two from the data itself, further experimentation is required.

The integration of gene expression with genetic data has led to the development of analytical approaches to distinguish causal from reactive genes (Schadt et al. 2005) (Fig. 6). The basis for this is grounded in two observations: (1) gene expression levels correlate with clinical trait measures across a population, and (2) eQTL colocalize with clinical trait QTL, where the QTL provide the causal anchors needed to infer the relationships among expression traits and between expression and clinical traits. Genes that meet both conditions are closely linked with the clinical trait, but they still may be either causal or reactive. However, assessing the conditional dependence of the clinical trait QTL on the state of candidate gene expression can indicate the likelihood of causality. If association of the clinical trait with the QTL genotypes is abolished when conditioned on the relative transcript abundance of the candidate gene, then the gene is supported as causative for the trait with respect to the QTL. If the association between the clinical trait and QTL genotypes is not affected by conditioning on the candidate gene, but the association between the QTL genotypes and expression trait vanishes after conditioning on the clinical trait, then the gene is supported as reactive to the trait. In cases where conditioning on the expression trait (or clinical trait) does not abolish the association between the QTL genotypes and the clinical trait (or expression trait), then the gene and trait are supported as independent of one another with respect to the QTL. Although there are a number of different ways this type of approach can be implemented mathematically, one of the first analytical applications of this concept was developed by Schadt et al. (2005), who used a maximum likelihood assessment of the three possible relationships at each locus between transcript and trait: causal, reactive, and independent.

This approach was applied to identify causal candidate genes for the trait of visceral (omental) fat in the BXD F_2 intercross. Of approximately 23,000 transcripts measured in livers of 111 F_2 mice, 4423 showed significant differential expression in 10% or more of the mice. Of these, 438 were significantly correlated (Pearson correlation coefficient) with omental fat pad mass ($p < 0.001$), while only 5 would have been expected by chance to have that degree of correlation. The omental fat trait had four QTL with LOD scores of 2 or greater. There were 114 of the set of 438 correlated transcripts with eQTL that colocalized with two or more of the omental fat trait QTL. By performing joint analysis for both expression levels and clinical traits, 267 pairs of eQTL and trait QTL were identified (corresponding to an FDR of 0.4%). Application of the causality test identified 134 genes where the causality test was accepted, and in the converse application of the reactive model, 23 genes were identified as reactive. These data allowed a ranking of the 114 genes, and we are in the process of completing validation studies of the ten top-ranked genes.

Among the ten top-ranked genes is *HSD1*, a gene that has previously been shown in a variety of studies to be associated with obesity and metabolic syndrome (Masuzaki and Flier 2003). In addition, we have evaluated several other top-ranked genes *in vivo*, using either transgenic or knockout approaches, and found an effect on fat mass in each of these, as recently published (Schadt et al. 2005). *C3ar1*$^{-/-}$ and *Tgfbr2*$^{+/-}$ mice were obtained from Deltagen, Inc., and a transgenic for *Zfp90* was constructed using a human BAC clone. All three models exhibited significant differences in adiposity compared with background strains. Evaluation of the remaining top-ranked genes is in progress through the construction of BAC transgenics for each of the genes. These data are strongly supportive of our approach, although *in vivo* evaluation of a larger number of predicted genes is necessary to determine the sensitivity of the method.

In the validation studies discussed above, we have focused on the genes identified as causal for the clinical trait being studied (obesity). However, the genes identified as reactive to the trait are also of interest because they may play a role in secondary effects of the primary trait such as insulin sensitivity or atherosclerosis. While detected as reactive in one cross, such genes may actually be causal in other contexts given feedback control mechanisms known to function in complex systems. It is also important to appreciate that this type of analysis greatly expands the information that can be derived from a given experimental intercross. The traditional approach of identifying those genes responsible for clinical QTL restricts the findings to a limited number of QTL (and hence a limited number of genes) and to only genes that have functionally significant sequence differences between the parental strains used in the cross (i.e., those that would generate *cis*-eQTL). In the approach described above, the genes identified need not have sequence differences between the parental strains and are typically *trans*-eQTL. The consequence is a much more efficient identification of causal genes for any given trait.

Constructing models of biological networks relevant to disease. As discussed above, there are methods and tools available to identify individual genes, sets of genes, and predefined ''pathways'' that play a role in disease. However, it is obvious that many genes and pathways are involved in trait expression, and the data by themselves do not indicate how they interact as a whole, or even which among them are likely to have a greater role or impact if therapeutically targeted. Developing realistic models for such complex diseases as diabetes and atherosclerosis that predict the role and function of each component remains far off. However, tools currently available that allow empirical gene expression ''network'' modeling are extremely promising for providing a large-scale picture of how sets of genes interact and which genes are more likely to play key roles.

There is substantial literature concerning network modeling of biological systems and a number of recent reviews (Barabasi and Oltvai 2004; Kitano 2004; Xia et al. 2004). Networks can be constructed from any large-scale data set where repeated measures are obtained under varying conditions. In a genetical genomic study, gene transcripts are measured across a set of animals, and the varying conditions are the consequence of the endogenous genetic variation among individuals in the population. Networks are frequently depicted graphically as a series of interconnected nodes, where the nodes are the individual gene transcripts and the lines connecting them (''edges'' in formal terminology) represent significant correlations in expression levels. A given gene transcript may be connected to few or many other transcripts, and the measure of this is referred to as connectivity. Analyses of various biological data sets have shown that all networks have a characteristic pattern referred to as ''scale free,'' where there are a limited number of highly connected nodes (termed ''hubs'') around which are many more nodes with fewer connections (Barabasi and Oltvai 2004). The full network derived from all the data is typically composed of subnetworks or

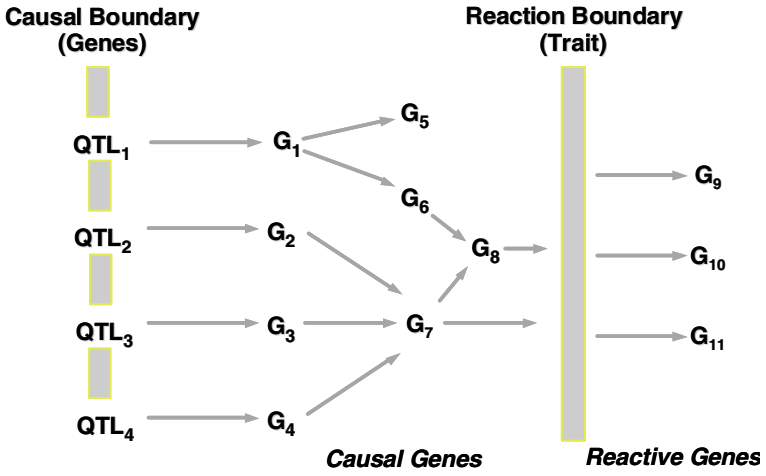

Fig. 6. Causal and reactive gene interactions in a complex trait. In this example, four loci (QTL_{1-4}) contribute to a complex trait. The DNA variations of the QTL directly influence the functions of genes G_{1-4}, which in turn perturb downstream genes that are causal for obesity ($G_{6,7,8}$) or reactive for the trait (G_9, G_{10}, G_{11}). Gene G_5 is independent of the trait but will nevertheless be correlated because its levels are controlled by a causal gene (G_1).

modules, where a set of nodes is highly interconnected, with relatively fewer connections to other modules in the network overall.

The gene expression network that one constructs from a data set is therefore an overall picture of how transcripts are related to each other. Transcripts that are in the same subnetwork and share connections are likely to function similarly. Transcripts that have many connections are likely to have a greater impact on the overall functioning of the network than those that have few connections. Transcripts that serve as points of connection between different modules or subnetworks are likely to be important in overall network stability and structure. Because the network is constructed empirically, without needing to know beforehand the established function of any specific transcript, transcripts for genes whose functions are unknown can be assigned tentative roles based on their close relationship with genes of known function. There are various analytical approaches one can take for these analyses, each of which has somewhat different assumptions and rules for constructing the network. These also vary in how one integrates phenotype information to relate the network to the disease process under study.

One of the approaches of particular interest for genetical genomic studies is Bayesian network analysis. Bayesian approaches allow for different types of data to be used in constructing networks and for incorporating directional relationships between nodes in a network. As discussed above, the integration of genetics allows for causal relationships to be established between transcripts and traits and among transcripts. Incorporating this information into network construction improves its power significantly (Zhu et al. 2004). Another approach for constructing networks from genetical genomic data is gene coexpression network analysis (Zhang and Horvath 2005). In this method, modules composed of sets of highly correlated genes are identified. These sets can be related to clinical trait expression and to common loci of genetic control.

Networks derived from genetical genomic data sets are in the early stages of being investigated and evaluated for their use and significance and there is limited validation as yet. Although subnetworks or modules are often enriched for genes of particular pathways or known function, it is far from a one–to-one correspondence, and genes of various functional categories according to established ontologies (functional categories) are often found to be closely related in a network. Understanding the relationship of gene expression networks to classical metabolic pathways or to networks derived from protein interactions or other elements will take much more work. However, at our current level of understanding, two very significant benefits that can be derived from gene expression network analyses are the ability to identify genes that have a high likelihood of controlling a clinical trait (the highly connected hub genes), and the ability to assign presumptive roles for otherwise uncharacterized genes.

Challenges and future directions

There are many challenges one could discuss and tremendous opportunities for future work. In this section we address several challenges that are of immediate importance and where we believe feasible approaches exist. These are (1) a need for rapid methods for validating candidate genes; (2) moving beyond genetic crosses to improve gene identification; (3) extending from global gene expression to include proteome and metabolome analyses; and (4) integrating genetical genomic data into publicly available databases.

Candidate gene prioritization and validation

A major challenge concerns the ''problem'' of having an abundance of high-likelihood candidate genes. The standard approach for validating whether a gene plays a role in a particular phenotypic process is to create mutants in which the gene has been deleted or overexpressed and examine whether they alter trait expression. This is time consuming and costly and, for various reasons, may not be as straightforward to interpret as one might think if negative results are obtained. In the long run, there are plans in the genomics community to systematically develop mutants for all expressed genes. However, at present only a relatively small fraction are available as extant strains or as banked gene-trapped embryonic stem (ES) cell lines. Even for such apparently ''readily available'' resources, the effort needed is considerable. Frequently, the administrative difficulties involved in obtaining mutant strains from other investigators are nontrivial and time consuming. Additional steps are needed to generate mutant mice from gene-trapped ES cell lines. In either case, expanding breeder pairs to generate a sufficient number of homozygous mice for study (at least 10–20 mice per group for traits as complex as atherosclerosis) adds more time. For validation of clinical traits such as obesity, diabetes, and atherosclerosis, mice usually need to undergo a protocol of 12–16 weeks of specified diet after weaning. Altogether, for any given gene this adds up to a minimum of one to two years of effort and considerable expense for validation.

Therefore, there is a need to develop efficient methods for ''screening'' candidate genes in a manner that allows prioritizing those most likely to be most physiologically relevant. As discussed above, the causality analyses and the network constructions are promising approaches, but these still yield relatively large numbers of candidate genes. We are currently assessing the use of a sequential approach that involves the use of surrogate end points based on gene expression profiles. Using the data set from which a candidate gene originated, a gene expression ''signature'' can be derived using the causal gene analysis technique described above. This can, in turn, be used as a validation test in various model systems by comparing the set of affected genes altered by a given intervention with the set defined by the causal expression signature. We have used this approach at a relatively advanced validation stage, comparing the expression signature obtained from intercross studies with those derived from knockout and/or inhibitor studies (Mehrabian et al. 2005). Results for two identified candidate genes, *HSD1* and *ALOX5*, indicate that there is a statistically significant conservation of the gene expression signatures between the original intercross mice and the targeted mutant mice. We are first screening candidate genes in cell culture systems—either cell lines or primary cell cultures appropriate for the gene and trait of interest—followed by comparable short-term *in vivo* studies if needed, where the end point is detecting conservation of the gene expression signature rather than the trait itself at this stage. The most promising genes identified would then be pursued in the traditional manner.

Use of outbred stocks

One of the limitations of traditional QTL mapping that remains an issue for transcript mapping is that the region defined by a QTL is large, a situation improved only modestly by dense genotype marker spacing. Even with the above-described approach of using *cis*-eQTL to enhance the ability to prioritize candidate genes, there frequently remain more candidate genes requiring followup than is feasible to handle. The basis resides in the limited number of crossovers that occur in any given animal of an F_2 intercross. Recent studies suggest that a promising approach to achieve highly accurate mapping is the use of outbred stocks. In particular, Flint and colleagues have demonstrated their method of finely mapping a quantitative behavioral trait (Yalcin et al. 2004, 2005). These mouse stocks are somewhat analogous to human populations, in which there are multiple alleles at any given locus, and it is likely that the spectrum of functional mutations observed in the available inbred strains will be represented in the outbred population. However, mice also demonstrate haplotype structures, and a limited number of haplotypes occur frequently. Haplotype blocks are much smaller and allow much finer mapping than could achieved in any intercross setting because they have accumulated over many tens to hundreds of generations. Extremely high-density genotyping would be necessary, but it is now technically feasible. Cheung et al. (2005) have recently reported using this approach in humans.

Incorporation of proteomic and metabolomic data

The concepts and approaches discussed above for transcript levels are directly applicable to protein (and metabolite) measures as well, since QTL analysis can be applied to any quantitative or semi-quantitative measurement and a protein has a corresponding gene encoding it. Thus, the finding of a QTL for variation in protein quantity or activity

that directly coincides with the physical location of the gene encoding that protein would suggest that the underlying genetic variation responsible resides in the coding or regulatory regions of that gene (i.e., is *cis*-acting). Finding both a *cis*-protein QTL and a *cis*-eQTL would be especially compelling. If the QTL were not at the location of the corresponding gene, then the genetic control of the protein measure must be through a different gene. In fact, one of the first applications of this approach was performed using measurement of protein instead of mRNA. Klose et al. (2002) used large two-dimensional gel electrophoresis to resolve and quantitate differences of mouse brain proteins from backcross mice, allowing a genetic analysis. Over 1000 proteins were shown to differ and over 600 were genetically mapped (i.e., QTL for levels of protein expression were determined, analogous to our transcript expression QTL). As with transcript QTL, some proteins mapped to the location of the gene for given protein (in "*cis*") and others mapped to one or more locations away from the physical location of the respective gene (in "*trans*"). The technologies for quantitative global protein measurement are far from the state available for mRNA, but there will undoubtedly be advances made in the coming years. The concurrent analysis of noncoding RNA expression with gene and protein expression studies is also expected to be highly informative and an area of strong interest.

Integrating genetical genomic data into publicly available databases

The "raw" data generated by genetical genomic studies contain information of broad interest and importance that extends beyond the immediate questions posed by the originating study itself. Much of these data are/will be made publicly available. Therefore, it is important that researchers who use mouse genetics approaches understand the data's applicability in conjunction with other data such as SNPs, haplotypes, strain and gene-related phenotypes, QTLs, and cross-species aspects of these. Data concerning *cis*-eQTL are of particular interest because these indicate that functionally significant sequence differences exist in the respective gene between the strains from which the data were derived. For example, we recently reported on conclusions derived from ultrafine mapping of SNPs for the C57BLKS/J strain in conjunction with eQTL data derived from a C57BL/6J and DBA/2J F_2 intercross (Davis et al. 2005). The C57BLKS/J strain (BKS) is genetically composed of primarily the C57BL/6J and DBA/2J genomes (approximately 70% and 20%, respectively), and it is particularly susceptible to diabetes and atherosclerosis compared with the respective originating strains. Identification of the genetic elements that predispose to these important diseases is of obvious interest. Given the genetic makeup of the BKS strain, those *cis*-regulated genes located in the DBA-like blocks of the BKS strain constitute primary candidates for genes that contribute to disease susceptibility. As discussed above, those *cis*-regulated genes whose expression levels are correlated with the trait of interest are primary candidates. As an example, we showed that the *Lipin* gene resides in a DBA-like region in the BKS strain, exhibited a *cis*-eQTL in the BXD cross with a five-fold difference in expression between parental genotypes, and was strongly correlated with adipose tissue-related phenotypes in that cross. Independently performed transgenic and knockout studies have shown that the *Lipin* gene shows significant influences on diabetes-related traits. By analogous logic, knowledge of *cis*-eQTL would be useful for prioritizing candidate genes residing in a QTL that had been identified in past experiments, so long as it was derived from a cross of the same inbred strains.

The Mouse Genome Informatics database (http://www.informatics.jax.org/) and the Rat Genome Database (http://rgd.mcw.edu/) include data related to QTL currently curated and incorporates mouse QTL studies and gene-related phenotype information. Also, the Rat Genome Database and some specific disease-related sites such as the Diabetes Genome Anatomy Project (http://www.diabetesgenome.org/) include microarray data sets as well, and as use of genetical genomic studies expands, these data will likely be incorporated.

Conclusions

The integrative genomics ("genetical genomics") approach is proving extremely useful for identifying genes and pathways that contribute to complex clinical traits. Clearly, the coincidence of clinical trait QTL and eQTL can help in the prioritization of positional candidate genes. More importantly, mathematical modeling of correlations between levels of transcripts and clinical traits in genetic crosses can allow prediction of causal interactions and the identification of "key driver" genes and can provide the data needed to develop models of biological networks that better explain disease pathogenesis. We anticipate that in the near future, the common variations influencing transcript levels both in *cis* and in *trans* will be defined for human populations and that the knowledge and experience gained from mouse models will complement human studies. It is likely that "genetical genomics" will have a revolutionary

impact on our understanding of complex traits such as cardiovascular disease and diabetes.

Acknowledgments

The authors thank the many students and colleagues who contributed to the work and ideas presented in this article. This work was supported by research grants from the National Institutes of Health (HL28481, HL70526, and HL30568), the UCLA Laubisch Fund, and the Iris Cantor-UCLA Women's Health Center.

References

1. Barabasi AL, Oltvai ZN (2004) Network biology: understanding the cell's functional organization. Nat Rev Genet 5, 101–113
2. Biddinger SB, Kahn CR (2006) From mice to men: insights into the insulin resistance syndromes. Annu Rev Physiol 68, 123–158
3. Bucca G, Carruba G, Saetta A, Muti P, Castagnetta L, et al. (2004) Gene expression profiling of human cancers. Ann N Y Acad Sci 1028, 28–37
4. Cervino AC, Li G, Edwards S, Zhu J, Laurie C, et al. (2005) Integrating QTL and high-density SNP analyses in mice to identify Insig2 as a susceptibility gene for plasma cholesterol levels. Genomics 86, 505–517
5. Cheung VG, Spielman RS, Ewens KG, Weber TM, Morley M, et al. (2005) Mapping determinants of human gene expression by regional and genome-wide association. Nature 437, 1365–1369
6. Dahlquist KD, Salomonis N, Vranizan K, Lawlor SC, Conklin BR (2002) GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. Nat Genet 31, 19–20
7. Davis RC, Schadt EE, Cervino AC, Peterfy M, Lusis AJ (2005) Ultrafine mapping of SNPs from mouse strains C57BL/6J, DBA/2J, and C57BLKS/J for loci contributing to diabetes and atherosclerosis susceptibility. Diabetes 54, 1191–1199
8. Diehn M, Sherlock G, Binkley G, Jin H, Matese JC, et al. (2003) SOURCE: a unified genomic resource of functional annotations, ontologies, and gene expression data. Nucleic Acids Res 31, 219–223
9. Dipetrillo K, Wang X, Stylianou IM, Paigen B (2005) Bioinformatics toolbox for narrowing rodent quantitative trait loci. Trends Genet 21, 683–692
10. Doss S, Schadt EE, Drake TA, Lusis AJ (2005) Cis-acting expression quantitative trait loci in mice. Genome Res 15, 681–691
11. Drake TA, Schadt E, Hannani K, Kabo JM, Krass K, et al. (2001) Genetic loci determining bone density in mice with diet-induced atherosclerosis. Physiol Genomics 5, 205–215
12. Flint J, Valdar W, Shifman S, Mott R (2005) Strategies for mapping and cloning quantitative trait genes in rodents. Nat Rev Genet 6, 271–286
13. Ghazalpour A, Doss S, Yang X, Aten J, Toomey EMV, et al. (2004) Thematic review series: The Pathogenesis of Atherosclerosis. Toward a biological network for atherosclerosis. J Lipid Res 45, 1793–1805
14. Ghazalpour A, Doss S, Sheth SS, Ingram-Drake LA, Schadt EE, et al. (2005) Genomic analysis of metabolic pathway gene expression in mice. Genome Biol 6, R59
15. Hosack DA, Dennis G Jr, Sherman BT, Lane HC, Lempicki RA (2003) Identifying biological themes within lists of genes with EASE. Genome Biol 4, R70
16. Jansen RC, Nap JP (2001) Genetical genomics: the added value from segregation. Trends Genet 17, 388–391
17. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M (2004) The KEGG resource for deciphering the genome. Nucleic Acids Res 32, D277–280
18. Kitano H (2004) Biological robustness. Nat Rev Genet 5, 826–837
19. Klose J, Nock C, Herrmann M, Stuhler K, Marcus K, et al. (2002) Genetic analysis of the mouse brain proteome. Nat Genet 30, 385–393
20. Lan H, Stoehr JP, Nadler ST, Schueler KL, Yandell BS, et al. (2003) Dimension reduction for mapping mRNA abundance as quantitative traits. Genetics 164, 1607–1614
21. Lan H, Rabaglia ME, Schueler KL, Mata C, Yandell BS, et al. (2004) Distinguishing covariation from causation in diabetes: a lesson from the protein disulfide isomerase mRNA abundance trait. Diabetes 53, 240–244
22. Lusis AJ, Mar R, Pajukanta P (2004) Genetics of atherosclerosis. Annu Rev Genomics Hum Genet 5, 189–218
23. Machleder D, Ivandic B, Welch C, Castellani L, Reue K, et al. (1997) Complex genetic control of HDL levels in mice in response to an atherogenic diet. Coordinate regulation of HDL levels and bile acid metabolism. J Clin Invest 99, 1406–1419
24. Masuzaki H, Flier JS (2003) Tissue-specific glucocorticoid reactivating enzyme, 11beta-hydroxysteroid dehydrogenase type 1 (11beta-HSD1)—a promising drug target for the treatment of metabolic syndrome. Curr Drug Targets Immune Endocr Metabol Disord 3, 255–262
25. Mehrabian M, Allayee H, Stockton J, Lum PY, Drake TA, et al. (2005) Integrating genotypic and expression data in a segregating mouse population to identify 5-lipoxygenase as a susceptibility gene for obesity and bone traits. Nat Genet 37, 1224–1233
26. Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, et al. (2003) PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat Genet 34, 267–273
27. Rizvi AA, Thompson PD, Pyritz R (2002) Genetic determinants of atherosclerotic heart disease and other occlusive disorders. In: *Principles and Practice of Medical Genetics*, Rimoin DL, Conner JM, Pyritz RE, Korp BE (eds.) (London: Churchill–Livingstone) pp 1519–1545

28. Schadt EE, Lamb J, Yang X, Zhu J, Edwards S, et al. (2005) An integrative genomics approach to infer causal associations between gene expression and disease. Nat Genet 37, 710–717
29. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, et al. (2005) Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA 30, 30
30. Wade CM, Kulbokas EJ, 3rd, Kirby AW, Zody MC, Mullikin JC, et al. (2002) The mosaic structure of variation in the laboratory mouse genome. Nature 420, 574–578
31. Weston AD, Hood L (2004) Systems biology, proteomics, and the future of health care: toward predictive, preventative, and personalized medicine. J Proteome Res 3, 179–196
32. Wiltshire T, Pletcher MT, Batalov S, Barnes SW, Tarantino LM, et al. (2003) Genome-wide single-nucleotide polymorphism analysis defines haplotype patterns in mouse. Proc Natl Acad Sci USA 100, 3380–3385
33. Xia Y, Yu H, Jansen R, Seringhaus M, Baxter S, et al. (2004) Analyzing cellular biochemistry in terms of molecular networks. Annu Rev Biochem 73, 1051–1087
34. Yalcin B, Willis-Owen SA, Fullerton J, Meesaq A, Deacon RM, et al. (2004) Genetic dissection of a behavioral quantitative trait locus shows that Rgs2 modulates anxiety in mice. Nat Genet 36, 1197–1202
35. Yalcin B, Flint J, Mott R (2005) Using progenitor strain information to identify quantitative trait nucleotides in outbred mice. Genetics 5, 5
36. Zhang B, Horvath S (2005) A general framework for weighted gene co-expression network analysis. Stat Appl Genet Mol Biol 4, Article 17
37. Zhu J, Lum PY, Lamb J, GuhaThakurta D, Edwards SW, et al. (2004) An integrative genomics approach to the reconstruction of gene networks in segregating populations. Cytogenet Genome Res 105, 363–374