thematic review

*Thematic Review Series: The Pathogenesis of Atherosclerosis*

# Toward a biological network for atherosclerosis

**Anatole Ghazalpour,* Sudheer Doss,* Xia Yang,* Jason Aten,† Edward M. Toomey,* Atila Van Nas,* Susanna Wang,* Thomas A. Drake,§ and Aldons J. Lusis[1,*]**

Department of Human Genetics,* Department of Medicine, and Department of Microbiology, Immunology, and Molecular Genetics, and Molecular Biology Institute, University of California, Los Angeles, CA 90095-1679; Department of Biomathematics,† University of California, Los Angeles, CA 90095-1766; and Department of Pathology and Laboratory Medicine,§ University of California, Los Angeles, CA 90095-1732

**Abstract** **The goal of systems biology is to define all of the elements present in a given system and to create an interaction network between these components so that the behavior of the system, as a whole and in parts, can be explained under specified conditions. The elements constituting the network that influences the development of atherosclerosis could be genes, pathways, transcript levels, proteins, or physiologic traits. In this review, we discuss how the integration of genetics and technologies such as transcriptomics and proteomics, combined with mathematical modeling, may lead to an understanding of such networks.**—Ghazalpour, A., S. Doss, X. Yang, J. Aten, E. M. Toomey, A. Van Nas, S. Wang, T. A. Drake, and A. J. Lusis. **Toward a biological network for atherosclerosis.** ***J. Lipid Res.*** **2004.** 45: **1793–1805.**

**Supplementary key words** systems biology • transgenic mice • quantitative trait locus mapping • principal components • Bayesian networks • correlation coefficients • genetics • genomics • proteomics

Atherosclerosis involves a large genetic network, not a simple linear pathway. This network extends to interactions with the many known risk factors for the disease and involves many cell types and organ systems (**Fig. 1**). The connectedness of the various risk factors results in their clustering in populations, and these clusters have been given designations such as "the metabolic syndrome" and "the atherogenic lipoprotein phenotype." Experimentally, the network is commonly studied by perturbing a single element, as in knockout mice, in a single genetic background. Although this approach provides valuable information that is simple to interpret, it may not identify the key regulators. Knockout experiments, for example, are dependent on prior biological information about the candidate gene, and they are not an efficient screen for the many epistatic and pleiotropic interactions that are likely to be involved. Approaches involving multiple perturbations, as in crosses between two genetically distinct strains of mice, may provide greater power to elucidate relevant pathways.

In this review, we discuss progress toward unraveling the complex network that influences atherosclerosis. First, we discuss various approaches that have provided much of our present knowledge of the pathways in atherosclerosis. These include genetic studies in humans and in animal models, including transgenic studies. Second, we discuss the use of genomic and proteomic technologies, as well as nonclassical statistics, to identify genes and pathways contributing to atherosclerosis. The combination of genetics and gene expression promises to be a particularly powerful approach in the identification of the interactions underlying complex traits. Third, we discuss the general properties of biological networks and some early results related to networks for atherosclerosis.

## HUMAN STUDIES

Studies of a number of Mendelian traits have formed the basis for much of our understanding of the pathways contributing to atherosclerosis. Familial hypercholesterolemia taught us about the importance of cholesterol, homocystinuria led us to the importance of small variations in plasma homocysteine levels, and various Mendelian blood pressure disorders showed us the importance of salt balance in blood pressure. Interestingly, few of the genes identified in these studies appear to contribute importantly to the common forms of atherosclerosis (1).

Genetic contributions to common, complex forms of atherosclerosis (and of traits such as lipid metabolism that are relevant to the disease) were first studied by popula-

*Manuscript received 16 July 2004 and in revised form 27 July 2004.*

*Published, JLR Papers in Press, August 1, 2004.*
*DOI 10.1194/jlr.R400006-JLR200*

[1] To whom correspondence should be addressed.
e-mail: jlusis@mednet.ucla.edu

The online version of this article (available at http://www.jlr.org) contains an additional table.

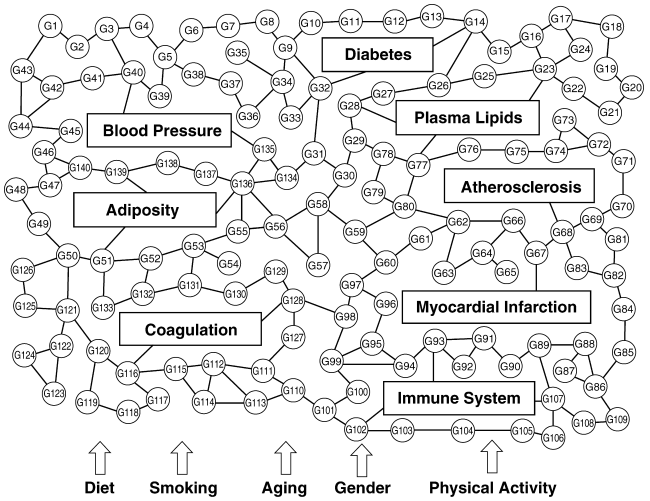



**Fig. 1.** A cartoon illustrating the concept of a genetic network for atherosclerosis and some of its risk factors. As discussed in the text, the actual network is undoubtedly much larger than this.

tion association with candidate genes based on our biochemical knowledge. One of the first important examples of this was apolipoprotein E (apoE), which was found to exhibit three common alleles in all populations studied (E2, E3, and E4). The E2 form was found to be strongly associated with a relatively uncommon dyslipidemia (type III) and with low cholesterol levels in the population, and the E4 allele was associated with increased cholesterol levels. Since the early 1980s, thousands of association studies with candidate genes have been performed for traits relevant to atherosclerosis. Of these, approximately a dozen have shown rather consistent findings [e.g., hepatic lipase with HDL levels and peroxisome proliferator-activated receptor γ (PPARγ) with type 2 diabetes], but most remain questionable, including many that have been studied in multiple, relatively large populations (2).

Since the mid 1990s, a common paradigm in human studies of complex traits has been to carry out linkage analysis in families to identify the regions of the genome harboring the most significant common genetic factors, followed by either linkage disequilibrium analysis of the region to pinpoint the underlying gene or the testing of "positional candidates" at that locus. The first successful example of this was the identification of calpain 10 in type 2 diabetes in a large set of Mexican-American families (3). Other similar studies have now identified several other loci and genes relevant to atherosclerosis (2). Linkage analysis has very limited power for complex traits and thus will reveal only the strongest and most common variations in the populations being studied (4). With the advent of cheaper methods for the detection of polymorphisms, genome-wide association studies are becoming feasible. For example, Ozaki et al. (5) carried out a study of single nucleotide polymorphisms in thousands of individuals in Japan who had been studied for coronary heart disease and identified several genes exhibiting strong evidence of association. These were then studied in a second set of families, and one gene, lymphotoxin-α, was found to be highly significant in the second set of individuals as well. A detailed map of common polymorphism haplotypes (HapMap) of the genome with a single nucleotide polymorphism (SNP) every kilobase or so should be completed by 2005, and this should greatly aid in the implementation of whole-genome association studies (www.hapmap.org).

Why have efforts to identify genes for the common forms of atherosclerosis been largely unsuccessful? One reason, of course, is that genes for the common forms have mostly modest effects that are difficult to detect in the background of many genetic and environmental perturbations. Another important reason is likely related to epistatic interactions. Thus, the effects of certain variations may influence phenotypes only in particular genetic backgrounds. This may explain why human studies frequently fail to replicate other human studies (different populations) or animal findings (different genetic context) (6). It seems unlikely that the goal of understanding in detail the genetic network involved in atherogenesis can be achieved in the foreseeable future by direct studies of human populations. Given the extensive conservation of gene structure and function among mammals (mice and humans differ by ~300 genes), the overall features of this network are likely to be similar between humans and other mammals. Therefore, the most useful approach will be to work out details of the network in animal models and then examine the corresponding features in human populations. It will be particularly important to define gene-gene and gene-environment interactions in animal models, because these will be the most challenging aspects of the problem.

Because atherosclerosis involves many cell types and important systemic influences, tissue culture studies will reveal only a subset of the important interactions. Nevertheless, such studies will importantly complement in vivo studies (7). In particular, expression array analyses of cells in response to genetic, nutritional, or pharmacologic perturbations should help in the formulation or validation of network models. For example, Johnson et al. (8) studied gene expression profiles of vascular smooth muscle cells in response to a polycyclic aromatic hydrocarbon present in tobacco smoke. Studies of cells obtained from individuals with various Mendelian or complex disorders may also be informative when subjected to genomic, proteomic, or metabolomic analyses.

## KNOCKOUTS AND OTHER SINGLE GENE MUTATIONS

Most of the atherosclerosis research community is focused on the use of single perturbations, primarily involving knockout or transgenic mice, in a single genetic background, usually strain C57BL/6J carrying a sensitizing mutation, such as a null mutation of apoE or the LDL receptor. Although this approach has taught us a great deal, we argue that other, genome-wide approaches will in some cases be more efficient for the identification of key interactions. As discussed above, a knockout perturbs just one branch of the very large genetic network contributing to

atherosclerosis, and the effect is dependent upon the particular genetic background employed. Because the network has so many connections, the fact that a knockout influences the development of lesions may be attributable to indirect effects. Also, a knockout is "unphysiological," and a positive result does not necessarily indicate a regulatory or modifier function of that gene in atherosclerosis. The knockout may also have a different effect in a different genetic background. For example, bone marrow transplantation of C57BL/6J mice with apoE null bone marrow cells enhances atherogenesis, whereas transplantation of C3H/HeJ mice with apoE null bone marrow reduces atherosclerosis, indicating that apoE is protective in one genetic context and proatherogenic in another (9). The failure to observe an effect of a knockout also does not necessarily exclude that gene from playing an important role in the network. For example, a gene may have both positive and negative effects (as with apoE above) and, in some genetic backgrounds, the effects may cancel each other. Or the knocked out gene could be part of an interconnected robust network that will adopt a new architecture and make the proper adjustments to ameliorate the effect of the perturbation and preserve the normal phenotype; in this instance, it may be necessary to simultaneously perturb two or more genes to observe an effect. A third explanation for the lack of phenotype could be that the conditions under which the gene of interest plays a role have not been tested yet.

At present, investigators primarily use transgenic approaches to study candidate genes, but with the completion of the genome sequences of human and various model organisms, including rat and mouse, an important future goal will be to define the functions of all 35,000 or so mammalian genes. For this, classic gene-specific approaches will be too laborious and time-consuming, and gene-trap mutagenesis or RNA interference (RNAi) approaches will be used instead. Already, several large gene-trap libraries of embryonic stem cells have been produced (10). Another approach for identifying genes relevant to specific processes involves the use of spontaneous or chemically induced mutations. Spontaneous mutations in mice, for example, have proven very useful for examining aspects of lipid metabolism (11). Several large-scale chemical mutagenesis screens of mice are being performed at present in the public and private sectors using ethyl nitrosourea, an alkylating agent that introduces point mutations at a high frequency (12).

Transgenic animals are usually characterized only with respect to a few phenotypes, such as the amount of atherosclerosis, the complexity of the lesions, the levels of plasma lipids, or the expression of selected candidate genes. Such results provide only a small fraction of the potential information that can be extracted with respect to networks. For example, genome-wide microarray analyses could be performed on a variety of tissues to provide a picture of the components of the transcriptional network that are perturbed. Such data could help in the formulation and validation of network models. For such studies, it may be preferable to examine animals in which the expression of a gene is altered but not totally ablated, because the latter condition may result in many nonphysiological alterations. Parallel studies in tissue culture cells can be used to complement or guide the animal studies (13).

## DISSECTION OF COMPLEX TRAITS IN ANIMAL MODELS

Whereas environmental and genetic backgrounds differ from individual to individual in human populations, these can be strictly controlled in studies with experimental animal models. Moreover, with animal models, it is possible to set up large, informative genetic crosses for purposes of mapping genes that contribute to complex traits. Mice and rats are by far the most useful mammals for genetic studies, with a great body of physiological, developmental, and genetic knowledge on which to build (14, 15).

Among inbred strains of mice and rats are variations relevant to most aspects of atherogenesis: plasma lipoprotein levels, blood pressure, diabetes, obesity, inflammation, atherosclerotic lesion development, lesion composition, lesion calcification, lesion-related medial destruction, and dietary responsiveness. Some of these variations are observed only in sensitized genetic backgrounds, such as hypercholesterolemia induced by null mutations for apoE or the LDL receptor. Recent studies have shown that hypercholesterolemic mice also show evidence of lesion rupture, although the occlusive thrombosis that is an important feature of the clinical disease has not been observed. These genetic variations tend to be very complex in rodents as well as in humans (16).

The genetic loci responsible for these variations can be mapped by linkage analysis [quantitative trait locus (QTL) mapping] in crosses between different strains (see, for example, **Fig. 2**). A recent review from the Complex Trait Consortium provides a clear and concise overview of QTL mapping (17). Generally, a hundred or more backcross or intercross progeny are generated and typed for the traits of interest and for genetic markers spaced at intervals along the genome. A variety of programs are available to perform linkage analysis, with features that permit interval mapping (testing for linkage between markers), calculation of statistical evidence of linkage, and analysis of epistasis and other interactions between loci. Such studies have shown that most of the traits relevant to atherosclerosis are highly complex and frequently exhibit epistasis. In crosses between a handful of strains, dozens of loci for plasma lipoprotein levels, body fat, and lesion size have been mapped [reviewed in ref. (16)].

QTL analysis is only capable of mapping a gene to a very large region (usually 50 Mb or more). If the resulting region does not contain any strong positional candidate genes, or if it contains many such genes, fine mapping must be performed. The most commonly used fine mapping strategy involves the isolation of individual QTLs in common genetic backgrounds (congenic strains) to remove other loci influencing the trait. This can be thought of as "Mendelizing" a complex trait. The congenic region can then be subdivided by further genetic crosses.

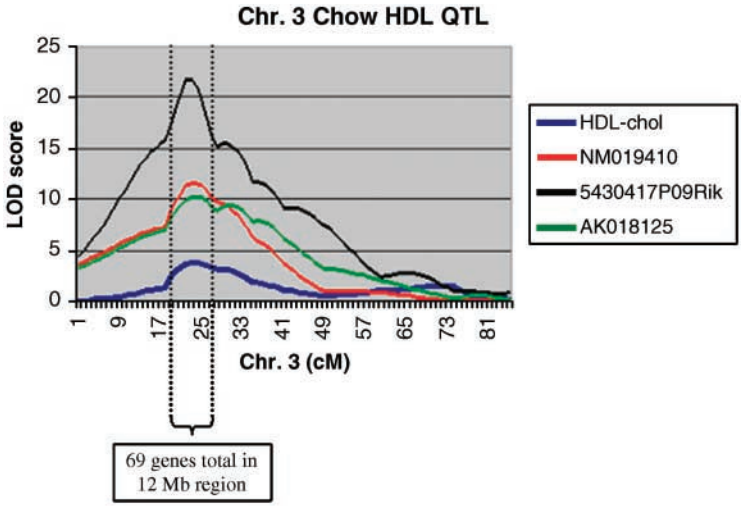



**Fig. 2.** Use of genetics of gene expression data to prioritize candidate genes underlying a locus of interest. A quantitative trait locus (QTL) for HDL cholesterol levels (HDL-chol) was previously described in a cross between strains DBA/2J and C57BL/6J on chromosome (Chr.) 3 at 22 centimorgan (cM) at the peak marker D3mit241 (44). Expression array analyses of livers from the mice in this cross were performed and used to predict *cis*-acting expression QTL (eQTL) (43). The *cis*-acting eQTLs that mapped to within 12 Mb of that marker were identified. This resulted in a prioritized candidate gene list of three genes (NM_019410, AK018125, and a previously uncharacterized gene, 5430417P09Rik). D3mit241, NM_019410, AK018125, and 5430417P09Rik are located at 68.06, 59.23, 69.04, and 69.05 mB, respectively. A LOD score is a statistical measure of the significance of linkage. In an F2 intercross in mice, a LOD score of 4.3 is considered significant.

In cases in which the effect of a QTL is very modest or the coefficient of variation of the trait is very large (as in the size of atherosclerotic lesions), progeny testing or the construction of subcongenic lines is required for fine mapping (18). The goal of fine mapping is to reduce the size of the critical region to ~1 or 2 Mb so that a relatively small number of candidate genes remain.

The construction of congenic strains is expensive and time-consuming, even when using a "rapid congenic" approach. Several whole-genome congenic libraries have now been constructed, allowing this step to be bypassed if the QTL alleles differ between the appropriate strains (19–21). Recently, Singer et al. (21) surveyed one set of "chromosome substitution strains" (congenic strains in which entire chromosomes are substituted) between strains A and C57BL/6 for several traits relevant to atherosclerosis, including plasma levels of cholesterol, campesterol, and sitosterol, weight gain in response to two different diets, and plasma levels of various metabolites. Altogether, in a survey of 53 traits, they identified ~150 different loci. These included loci for cholesterol on 8 different chromosomes, loci for sitosterol on 14 chromosomes, and loci for weight gain on 17 chromosomes. The authors suggest that direct surveys of such congenic strains provides a more sensitive way of locating QTLs compared with genetic crosses, because the latter exhibit "phenotypic noise" resulting from the simultaneous segregation of multiple QTLs (21).

It has been suggested that in silico SNP haplotype analysis (analysis of haplotypes that are available in databases) across inbred strains of mice might be a useful strategy for mapping complex traits (22). Although the approach is probably of limited utility for highly complex traits (23, 24), it can be very useful in conjunction with analysis of QTLs in multiple crosses to identify which strains are likely to share a common allele (25). Extensive SNP databases for a number of strains are now available and are rapidly expanding.

The identity of the gene underlying a QTL is normally confirmed by examining the effects of a knockout or a transgene on the phenotype. For this, one would normally first search the literature for previously engineered mice, including gene-trap libraries. If none can be identified, it may be possible to examine aspects of the phenotype in cultured cells. We have also used bacterial artificial chromosomes harboring candidate genes for the construction of transgenic mice, reasoning that for most quantitative traits, a 1- or 2-fold perturbation in the level of expression of a gene will influence the final phenotype (although this will not always be the case). The strongest evidence, of course, would be to replace one allele for another using a "knock-in" strategy, although this should not be required as "proof" of the identity of the underlying gene.

Although QTL mapping has great power to detect linkage, the identification of genes underlying the QTL has proven to be very difficult. For example, more than 20 different loci for atherosclerotic lesions have been identified in mice, but of these, only 2 genes, both positional candidates, have been confirmed using transgenic approaches (16). The recent completion of the sequencing of the mouse and rat genomes will considerably aid in the harvest of genes, but the identification of novel genes will still be limited by recombination intervals.

Williams, Haines, and Moore (6) recently proposed the construction of a very large (~1,000) set of recombinant inbred (RI) strains to provide a tool for rapid fine mapping of QTLs. RI strains are produced by crossing two or more inbred strains and then inbreeding the progeny to genetically fix particular combinations of alleles from the parental strains. The RI strains would be derived from eight highly diverse inbred strains to incorporate a great deal of naturally occurring variation and would be genotyped at a very high density, allowing resolution of ~100,000 unique recombination breakpoints with an average spacing of ~25 kb (26). Envisioned as a "collaborative cross" that would be used and maintained by multiple scientists and institutions, the RI set would be used for QTL analysis in three stages. First, a subset of RI strains would be studied to roughly map the QTL. Second, 100–200 strains with breakpoints in the interval of interest would be examined for the phenotype. Third, all mice with relevant breakpoints (including other QTLs for the trait) would be studied.

## GENOMICS, TRANSCRIPTOMICS, PROTEOMICS, AND METABOLOMICS

During the past decade, whole-genome microarrays have been widely used to survey differences in gene expression between tissues, between normal and disease states, and between different environmental conditions (27). Hierarchical clustering of such data allows genes to be grouped into classes of potential functional significance (28), and groups of genes can be tested for predictive value for disease (as in the assessment of the stage of a cancer). Also, high-throughput gene expression microarray analysis can be used to facilitate gene identification when combined with mapping information. This approach has now been applied to help identify genes for several QTLs. For example, Aitman and colleagues (29) used cDNA microarrays containing 10,000 randomly collected cDNA clones from a rat cDNA library to compare gene expression levels between the epididymal fat of the spontaneously hypertensive rat (SHR) and a congenic strain containing a locus for insulin resistance derived from the Brown Norway rat in the SHR background. Out of 13 clones exhibiting reduced hybridization from SHR fat, three belonged to the CD36 gene, which resided within the congenic interval. Transgenic overexpression of the CD36 gene in the SHR strain ameliorated insulin resistance and decreased plasma free fatty acid levels, confirming that CD36 was causal in the trait. Similarly, expression array analyses helped to identify genes underlying QTLs for asthma (30) and bone mass (31).

In the case of complex disorders, differences in gene expression may be subtle and thus difficult to detect, and human studies are likely to be complicated by genetic heterogeneity. For example, attempts to identify significant differences in the expression profiles of muscle from type II diabetics compared with normal volunteers have failed to reveal differences in individual genes. Mootha et al. (32) used an ingenious approach to the problem: rather than test for differences in the expression of individual genes, they tested for overall differences in expression patterns of various sets of genes in annotated pathways. In this study, they used 149 metabolic pathways and groups of functionally (or spatially) related genes and computed a score for each pathway/group based on the combined differential expression measure of the genes in each group. The score each pathway received was proportional to the number of genes enriched in the microarray profiling data. Pathways were then ranked based on the score they received, and the statistical significance of the score of top-ranking pathways was determined using a permutation test. The analysis revealed that groups of genes involved in oxidative phosphorylation and mitochondrial functions ranked highest, although the overall changes in gene expression in diabetics compared with controls were relatively modest (32). One of the genes downregulated in diabetic patients was PGC-1α, a primary regulator of metabolism. Overexpression of PGC-1α in a mouse skeletal muscle cell line resulted in increased expression of many of the oxidative phosphorylation genes in the identified pathways. Although this study did not result in the identification of the causal genes in type II diabetes, it did suggest that they act by perturbing oxidative phosphorylation. As discussed below, expression array analysis in combination with genetic or environmental perturbations provides a powerful approach not only for the identification of candidate genes underlying complex traits but also for the elucidation of causal interactions between genes and traits.

Gene expression, of course, will not capture many important interactions within a cell. Thus, the correlation between transcript levels and protein levels is poor for many proteins, and the activities of many proteins are further regulated by modifications such as phosphorylation or proteolysis. Moreover, structural variations such as missense mutations or alternative splicing are unlikely to be detected by standard expression arrays. Large-scale analysis of proteins has the potential to provide a more comprehensive understanding of complex biological processes, but methods for comprehensive screening for differences in protein levels or structures have not yet been developed. Two-dimensional polyacrylamide gel electrophoresis (2D PAGE) has very limited sensitivity (usually ~1,000 proteins). Nevertheless, several studies have used 2D PAGE to identify numerous differences in protein levels that occur during atherogenesis. An extension of 2D PAGE is differential in-gel electrophoresis, in which two pools of proteins are labeled with different fluorescent dyes, allowing detection of quantitative differences between the pools (33, 34). Mass spectrophotometric methods have great sensitivity but are difficult to apply on a genome-wide level. Protein microarrays have been designed to capture various features of functional proteomics, including protein levels, protein-protein interactions, and activity. These arrays are essentially high-throughput versions of enzyme-linked immunosorbent assays, in which characterized peptides or antibodies are immobilized on the surface of a chip and subsequently probed with the sample of interest.

A number of different applications have been developed to characterize protein-protein interactions, including the yeast two-hybrid system, a genetic assay in which binding is detected upon induction of reporter genes (35, 36). To facilitate the characterization of post-translational modifications, such as phosphorylation, several mass spectrophotometry-based techniques, including multi-dimensional protein identification technology, isotope-coded affinity tagging, and Fourier transform ion cyclotron resonance, have the capability to detect protein alterations. In the section on Biological Networks below, we discuss the results of genome-wide yeast two-hybrid analysis that has provided a comprehensive network of protein-protein interactions in several organisms.

Like the other "omic" technologies, metabolomics seeks to identify all gene products (transcripts, proteins, or metabolites) present in biological samples and to elucidate the quantitative dynamics of these products. The principal tools for metabolomics are gas-liquid chromatography coupled with mass spectrometry. Most progress in the metabolomics field has involved plant biology, but there are

now a number of reports relevant to atherosclerosis and diabetes. For example, Watkins et al. (37) carried out a comprehensive metabolic assessment of lipid metabolites to identify the specific effects of the PPARγ agonist rosiglitazone in a mouse model of type 2 diabetes. The authors demonstrated a large number of tissue-specific metabolic effects and proposed that metabolomics has excellent potential for the clinical assessment of responses to drug therapy (37). Metabolomics will be most powerful when coupled with other functional genomics approaches.

## COMBINING GENETICS AND GENE EXPRESSION

Transcript levels can be used as genetic traits in the same way as phenotypes, such as cholesterol levels (38, 39). The genetic loci controlling the levels of a transcript can thus be mapped by QTL analysis, and the loci thus identified have been termed expression QTL (eQTL). If the transcript level is controlled by structural variation of a gene that influences its rate of transcription or the maturation or stability of the transcript, the eQTL would be expected to map directly over the gene in question. Such an eQTL is termed a *cis*-acting eQTL. If, on the other hand, the levels of a transcript are determined by a genetic variation in a second gene (e.g., a variation in a transcriptional regulator of the first gene), the eQTL would map to the position of the regulator gene rather than to the gene whose transcript levels segregated in the cross. Such an eQTL would be termed a *trans*-acting eQTL. The approach of examining the segregation of transcript levels in a genetic cross should be distinguished from a study in which transcript levels are simply compared between two different strains. In the latter study, differences in transcript levels can be identified, but it is not possible to determine from such data whether any of the differences observed are the result of *cis*-acting genetic differences or *trans*-acting differences.

An example of this approach was the analysis of HDL levels in a cross between two strains differing in the response of HDL to an atherogenic diet. C3H mice maintain high levels of HDL on a high-fat diet, whereas strain C57BL/6 mice show a reduction in response to the diet. To test for the potential involvement of bile acid metabolism in this trait, Machleder et al. (40) quantified mRNA levels of cholesterol-7α-hydroxylase (CYP7A) as well as HDL levels in the cross. They observed three loci that segregated for HDL levels, and at each locus they also observed QTLs for the mRNA levels of CYP7A. Because the structural gene for CYP7A was located outside of any of these regions, it was clear that it was regulated in *trans* by several unlinked genes. The observation that the CYP7A transcript levels segregated with HDL suggested that it was involved in the HDL trait.

More recently, microarrays have been used to assess genome-wide transcriptional activity in segregating populations, offering a powerful tool to dissect causal relationships between genes and traits. As discussed by Jansen and Nap (38, 39), the analysis of gene expression in segregating populations with multiple genetic perturbations can potentially reveal much information about gene-gene and gene-clinical trait interactions. This approach was first applied to yeast, in which genome-wide analyses of transcript levels in a cross between two divergent strains revealed a large number of loci of both the *cis*-acting and *trans*-acting variety (41). Subsequently, two studies were performed in mice involving crosses of strains differing in diabetes-related traits (42, 43).

Lan et al. (42) studied the levels of expression of a number of candidate genes for insulin resistance and lipid metabolism segregating in a mouse cross. Using principal components analysis, they were able to identify groups of transcripts whose levels were explained by principal components (**Fig. 3**). Such principal components likely correspond to *trans*-acting factors influencing a set of genes.

Schadt et al. (43) used whole-genome expression arrays to examine transcript levels in livers in a cross between strains DBA/2 and C57BL/6. One surprising finding was the large number of genes whose transcript levels segregated in the cross. More than one-third of the transcripts exhibited evidence of segregation, as judged by hierarchical clustering and the fact that the transcript levels exhibited significant QTLs at one or more locations in the genome. The question of false positives with so many comparisons is obviously very important in such studies. The fact that *cis*-acting transcript QTLs happen to map over the corresponding gene suggests strongly that most of these are unlikely to be false positives, but the false-positive rate for *trans*-acting loci is unclear. A high fraction of false positives would likely complicate the use of the data in attempting to construct gene networks. In addition to mice, Schadt and colleagues also carried out combined expression array analyses on crosses between strains of maize and on cultured cells from several human pedigrees. The human studies did not yield data of sufficient quality to permit QTL analysis, but the maize data yielded results comparable to the yeast and mouse results.

One benefit of a "genetics of gene expression" approach is that it provides candidate genes for QTL studies. Figure 2 illustrates the use of *cis* eQTL underlying a phenotypic QTL to prioritize candidate genes. In this example, a QTL for plasma HDL levels was identified in a cross between DBA/2 and C57BL/6 (44). This region encompasses more than 69 genes in 12 Mb, but of these genes only 3 exhibited significant *cis*-acting eQTLs (Fig. 2). A genetics of gene expression approach can also be used to subclassify animals in a cross based on their expression profiles, similar to the use of microarrays for the classification of cancers. For example, Schadt et al. (43) identified genes that best distinguished thin from fat mice and showed that these fell into groups relating to different QTLs for body fat. Probably the most important application of the genetics of gene expression will be to construct gene networks for biological traits and identify causal interactions.

## STATISTICAL ANALYSIS OF DATA FOR COMPLEX TRAITS

Statistical analyses, such as analyses of variance, are pervasive in the biological sciences. However, the multifacto-

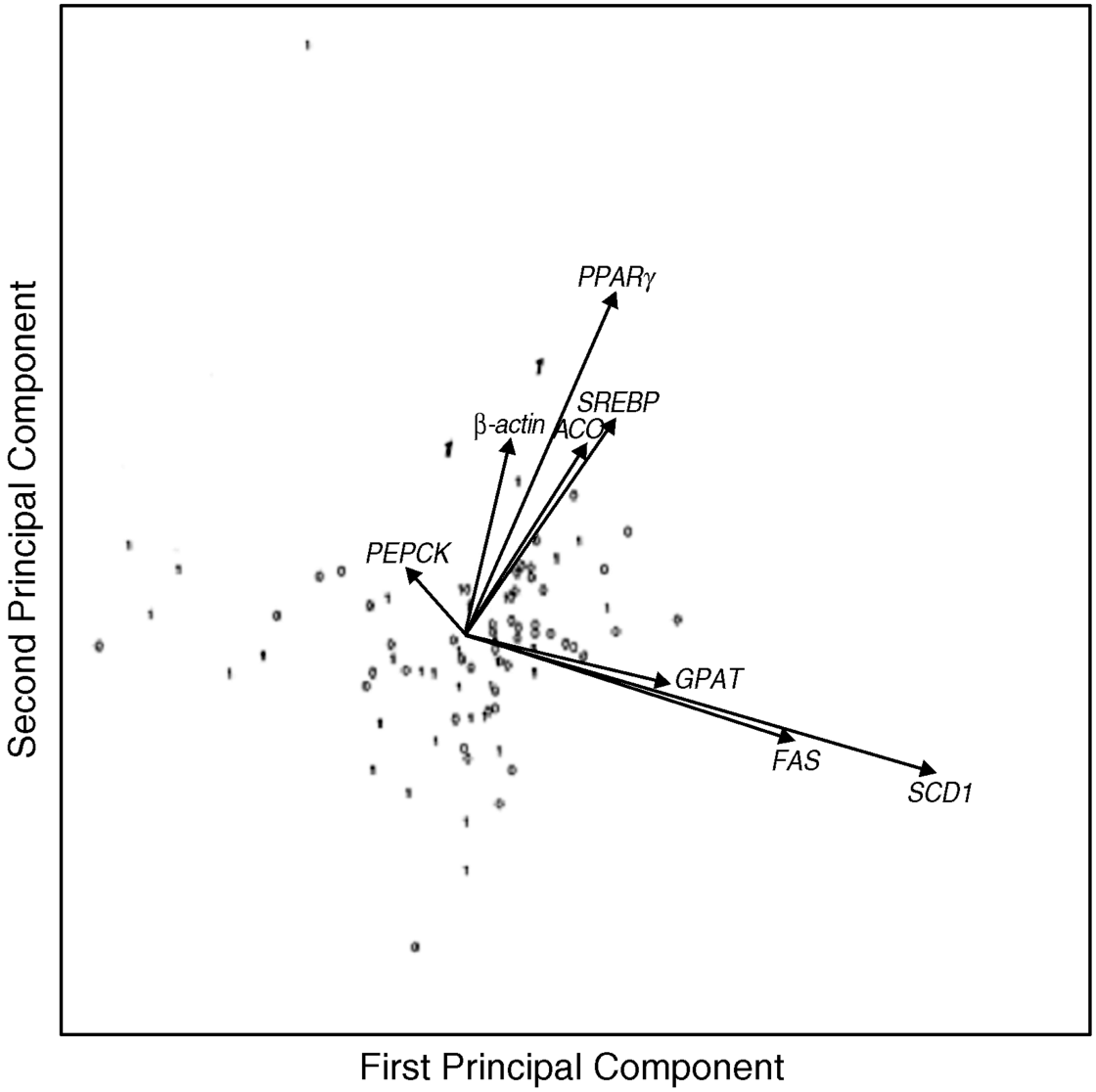



**Fig. 3.** An example of the use of principal component analysis to "reduce" data in a genetic study of diabetes (42). The mRNA traits of seven metabolic genes (SCD1, FAS, GPAT, PPARγ, SREBP, PEPCK, and ACO) as well as a normalization control for gene expression (β-actin) were determined in a genetic cross between two strains differing in susceptibility to diabetes (see text). The trait values were then subjected to principal component analysis, and the first two principal components are shown on a two-dimensional plot. [From Lan et al. (42), reprinted with permission of the Genetics Society of America, copyright © 2003.]

rial and multidimensional nature of complex traits frequently makes classic statistical methods insufficient for data analysis. In recent years, various other statistical approaches, such as principal component analysis, neural networks, and Bayesian networks, have been used increasingly to deal with large amounts of data and to analyze complex interactions.

The use of large data sets, involving thousands of genes and multiple traits, raises statistical issues such as false discovery rates and difficulties in integrating multidimensional information (42, 45). Dimension reduction techniques can simplify such data sets and avoid the issue of multiple comparisons (46, 47). One such technique is principal component analysis (42, 48). Principal component analysis captures orthogonal linear combinations of correlated variables such as gene expression values, and each combination is called a principal component (PC). PCs are ranked based on their significance in explaining the variance in a data set. Two- or three-dimensional plots can be constructed with the first two or three PCs that capture most of the information in the data. The resulting visual display may elucidate how the variables are grouped into clusters and how important each variable is in each PC. Figure 3 shows an example of principal component analysis in a study conducted by the Attie group (42). In this study, the expression levels of seven genes involved in metabolic pathways were analyzed against several phenotypes, including glucose level, insulin level, and body weight, in an F2-*ob/ob* cross between C57BL/6J and BTBR. Two PCs were identified, with the first PC encompassing mRNA levels of SCD1, FAS, GPAT, and PEPCK and the second encompassing mRNA levels of PPARγ, SREBP, and

ACO. The first principal component, mostly driven by the expression levels of SCD1 and FAS, was found to be strongly associated with the insulin trait. In this case, by performing QTL mapping for the two PCs instead of seven individual genes, the dimensions of the analysis were significantly reduced.

In studying complex biological traits, various data-mining tools have gained popularity. These methods can allow efficient and flexible integration of a large number of genetic and environmental factors as well as their interactions into the overall picture (49, 50). The essential technique of data mining used in such applications is pattern recognition, that is, extraction of hidden covariates of predictive value for a complex trait from a given data set. In addition, complex nonlinear interactions between the covariates may be detected. **Figure 4** describes two data-mining approaches, neural network analysis and tree-based recursive partitioning, that have proved useful in linkage and population association studies with various traits relevant to atherosclerosis (49–56). In addition to neural networks and tree-based methods, other data-mining tools, such as discriminant analysis, Bayesian variable selection, combinatorial partitioning, stepwise regression, and automated detection of informative combined effects, have shown promise in dissecting the genetics of complex traits such as myocardial infarction, hypertension, and cholesterol levels (57–59).

Bayesian inference has been referred to as bringing a "new revolution in genetics" because of its ability to incorporate both prior knowledge and sample data into the networks of complex biological processes (60). Bayesian inference is a probability model in which both data and model parameters are considered to be random variables with a joint probability distribution. Both prior knowledge about an event ($\phi$) and sample data (x) are used to calculate the posterior (or conditional) distribution of $\phi$ given data x using the equation $P(\phi|x) = [P(x|\phi)P(\phi)]/P(x)$, where $P(\phi|x)$ is the probability of $\phi$ given x, $P(x|\phi)$ is the probability of x given $\phi$, $P(\phi)$ is the probability of $\phi$, and P(x) is the probability of x. Graphical models that use Bayes's rule of inference, termed Bayesian networks, have been used increasingly to model complex biological processes such as metabolic and transcriptional regulatory pathways (61–64). An excellent introduction to Bayesian networks can be found at http://www.ai.mit.edu/~murphyk/Bayes/bnintro.html. Bayesian networks combine probability and directed graphs to visually depict conditional dependencies between large numbers of variables. **Figure 5** illustrates one possible Bayesian network for atherosclerosis. The network incorporates measurements of diet, genotype, obesity, diabetes, cholesterol, and atherosclerosis. Analysis of such a network (65–67) could be used to classify genes into functional categories. If the Bayesian network correctly captures the causal dependencies between variables, then given enough data, genes that mediate cholesterol's effect on atherosclerosis (gene A in Fig. 5) could be distinguished from those that act on atherosclerosis via obesity (gene B in Fig. 5), and both could be distinguished from those genes that act directly on atherosclerosis risk (gene C in Fig. 5). More complex models are possible and

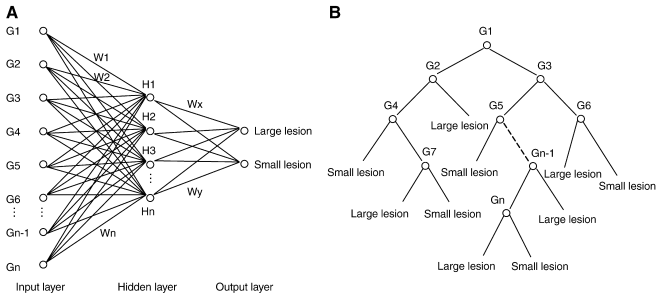



**Fig. 4.** Schematic representation of neural networks and tree-based partitioning. Atherosclerotic lesion size, a complex trait, is used as the target trait in this hypothetical example. A: A neural network is organized into input, hidden, and output layers. Each layer consists of multiple processing units called neurons (represented by open circles). In the input layer, neurons are labeled as G1, G2, . . . Gn, representing data from particular genes, loci, markers, or environmental factors that might be involved in lesion formation. The information received from the input layer is weighted, summed, and compared with a threshold using an activation function in the hidden neurons (labeled as H1, H2, . . . Hn) of one or more hidden layers. The results generated from the hidden layers (in this case, large or small lesion size of atherosclerosis) are transferred through another activation function to the neurons in the output layer. The strength of the connections between neurons in different layers are designated as weights (labeled as W1, W2, Wn, Wx, and Wy), which are adjusted during the training of the network using a training data set with known inputs and outputs. The weights are responsible for the flexibility and adaptability of a fitted network model. B: In tree-based modeling, a data set is recursively divided into more homogenous subgroups based on optimal split variables (tree nodes) such as genetic markers or genes. Tree nodes are represented by open circles and labeled as G1, G2, . . . Gn. Using certain splitting rules, each variable is measured for its potential to divide the data set into more parsimonious subsets based on lesion size. The explanatory variables with high potentials are added to the tree as nodes. The growth of the tree ends when each of the subsets becomes homogeneous to a predetermined degree. Based on the final tree, sets of rules can be deduced to determine which combinations of markers or genes have predictive values of large or small lesion size.

are the norm. Bayesian networks allow the integration of data from multiple studies, enable ready incorporation of medical and biochemical background knowledge, and can assess the consistency of observational and experimental data with different functional roles for genes.

## BIOLOGICAL NETWORKS

Genetic pathways were originally modeled after the enzymatic steps of intermediary metabolism. The inadequacy of such linear models was first revealed in attempts to use saturation mutagenesis to identify all of the steps in developmental pathways of the vulva in *Caenorhabditis elegans* and the compound eye in *Drosophila*. Such studies showed that many genes were involved, some specific for vulva or eye development, but many others, equally important, were involved in other processes as well. These and subsequent studies have indicated that the complex relationships among genes are best described as distrib-

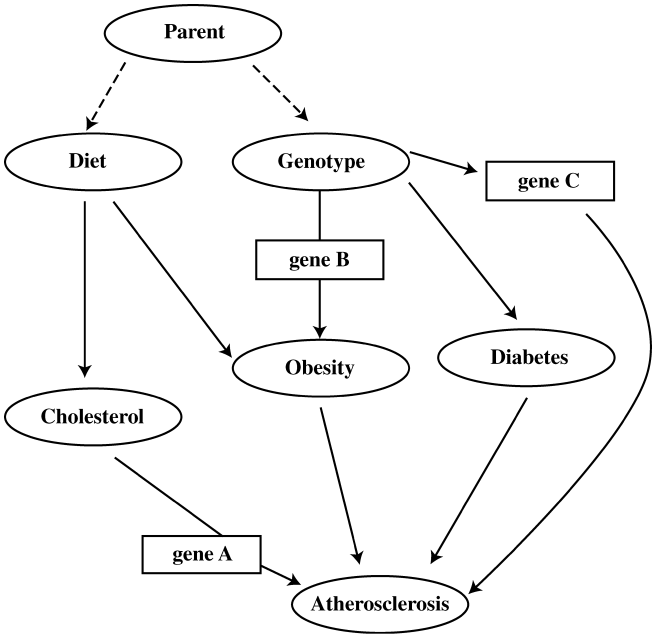


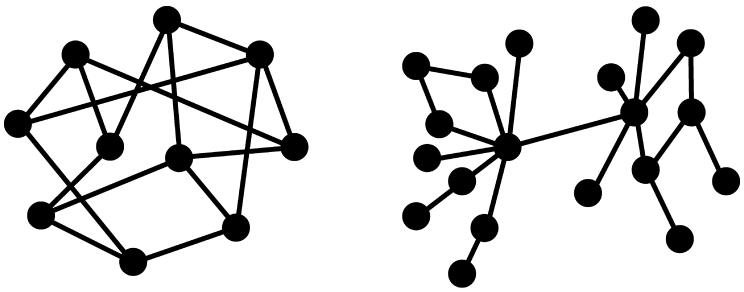

**Fig. 5.** A Bayesian network model of atherosclerosis can test the consistency of gene expression data with different functional gene roles. See text for discussion. Dashed lines represent unobservable flow of influence, solid lines represent observable dependencies, and lack of a direct arc between two nodes implies a conditional independence. Genes in roles A, B, and C can theoretically, given sufficient data and assuming no missing arrows, be differentiated in this Bayesian network.

uted networks. A network consists of "nodes" (elements such as genes) that exhibit specific interactions, termed "connections" (or "edges"). In a network of genes, functional variations of some genes will influence a phenotype more than those of other genes, but the effects will depend heavily on the context of other alleles that are present. This explains the common finding of very different phenotypes of a mouse knockout when studied in different genetic backgrounds. This does not imply that the interactions defined by studying atherosclerosis using knockouts in one genetic background are wrong; rather, they are simply part of a larger picture.

An important characteristic of networks compared with linear pathways is increased flexibility to respond to diverse conditions. For example, in a network the same output can be produced in multiple ways. This "buffering" capacity explains in part the common finding of knockouts with little or no apparent effects. Although "redundancy" is frequently invoked to explain the absence of phenotypes in knockouts, different genes cannot be completely redundant because natural selection would not maintain two genes for exactly the same function. The plasticity of a response is also greatly increased by multicellularity (as is the case with atherosclerosis). Thus, interactions between cells that are themselves nonidentical result in exponential increases in the possible combinations (68). Although such networks have increased buffering capacity and plasticity, their extensive interactions make them sensitive to many different perturbations. Thus, in the case of cardiovascular disease, large numbers of genetic and environmental factors are seen to influence susceptibility. This is strikingly observed in mouse models of atherosclerosis, in which more than 100 different knockouts have been observed to influence the development of lesions (2).

A recent review by Barabasi and Oltvai (69) highlights the emerging properties of biological networks. Networks can be constructed using various "nodes," including proteins, metabolites, or genes. Although networks have been studied in most detail in yeast and bacteria, the networks of all organisms appear to share similar global properties. Typically, most nodes in a network have few links, although some nodes have numerous links. Such networks are termed "scale-free." These contrast with "random networks," in which all nodes have similar connectivities (**Fig. 6**). In scale-free networks, nodes with numerous links, also referred to as hubs, play a central role in shaping the network's behavior. Scale-free networks are characterized by a high degree of robustness. That is, if a change occurs in nodes of the network with few connectivities, there would likely be strong resistance against perpetuation of the change throughout the network. Biologically, this means that mutations or environmental factors affecting a gene or a pathway will not result in drastic changes in the overall structure of the network. For example, knockout of a gene that happens to be a node with few connections to other genes will generally have a much smaller effect than knockout of a hub gene. Consistent with this notion, Jeong and colleagues (70) reported that in yeast knockouts of genes with many connections were much more likely to be lethal than knockouts of genes with few connections.

One example of the use of systems biology to construct networks is the yeast *GAL* gene interaction network (71).



**Random Network** **Scale-free Network**

**Fig. 6.** Network models. At left is a random network characterized by an almost equal number of connections between nodes. At right is a scale-free network, characterized by high connectivity for a few nodes and low connectivity for most other nodes.

In this study, the integration of genetics with gene expression profiling technology, along with data collected from protein-DNA and protein-protein interaction databases, resulted in a model of how galactose metabolism genes interact with other pathways. After experimental verification, it was determined that the assembled network was able to predict most interactions correctly, providing a concrete example of how systems biology could be used to unravel the interaction between biological pathways. Recently, several other genome-wide interaction networks have been assembled in various organisms ranging from prokaryotes to multicellular organisms (72–75).

One striking example is the *Drosophila* protein interaction network, assembled based on genome-wide yeast two-hybrid analysis and other data (75). In this study, the authors were able to examine local interactions and identify previously unrecognized motifs, assign pathway membership to uncharacterized proteins, assign subcellular locations to proteins, derive new links in signal transduction cascades, elucidate intercompartmental and intracompartmental interactions, and predict a mechanism of action for the ortholog of a human gene associated with B-cell lymphoma. An interesting observation was that after organizing the protein interactions according to cellular compartments (nuclear, cytoplasmic, membrane), the authors were able to demonstrate that interactions within compartments were much more frequent than those between compartments (75).

## CARDIOVASCULAR NETWORKS

A network of cardiovascular physiologic traits was constructed by Nadeau et al. (76) through the use of simple correlation (of traits) in a segregating set of mice. For example, if there are interactions between traits A, B, and C, such that B mediates entirely the interaction between A and C, the correlation in a segregating population between A and C should be equal to the product of the correlation between A and B and the correlation between B and C [corr(A,C) = corr(A,B) × corr(B,C)]. In general, causally linked traits will be correlated in a genetic cross, and direct interactions will tend to be more strongly correlated than indirect interactions. Nadeau et al. (76) examined the relationships of several cardiovascular traits

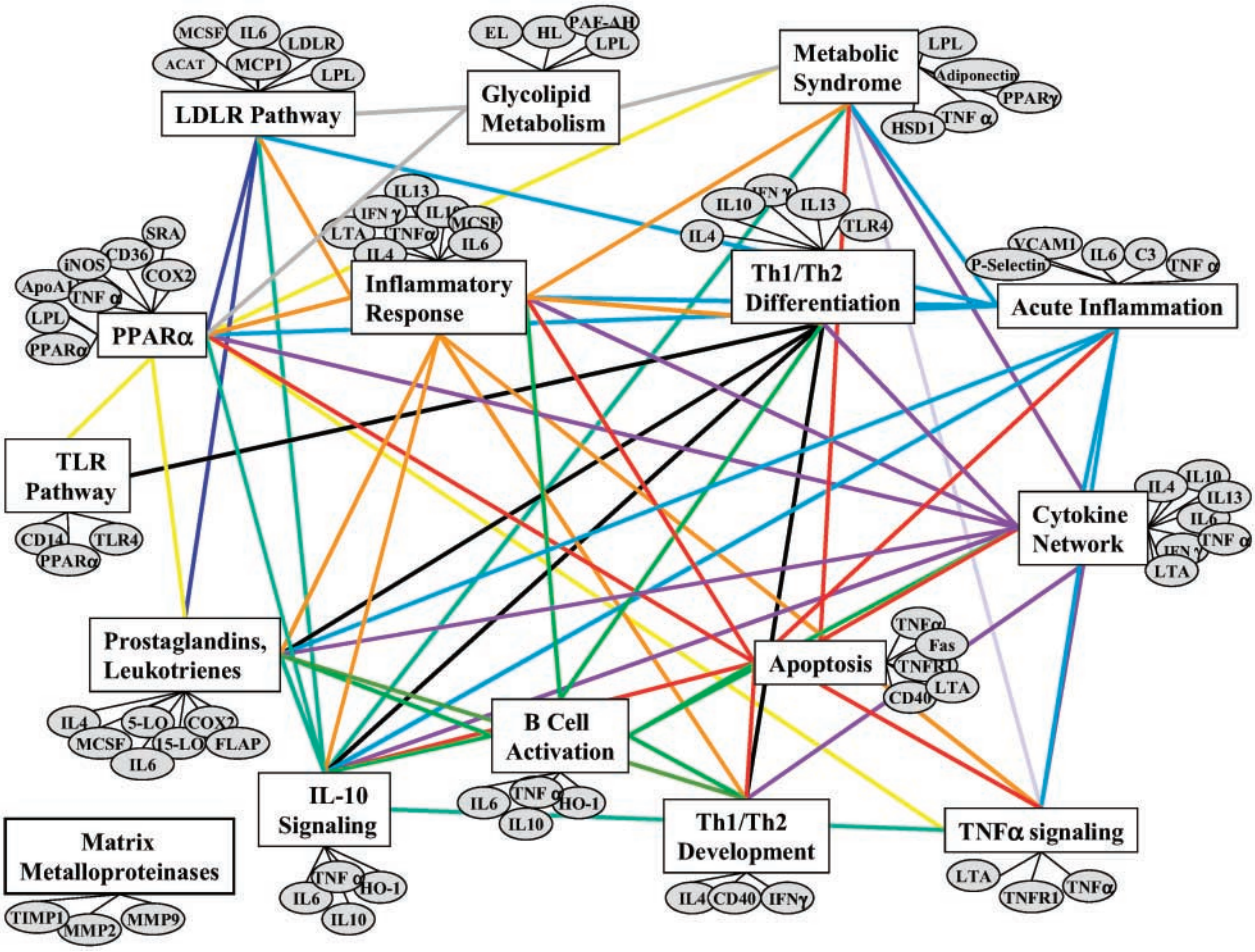



**Fig. 7.** Atherosclerosis pathway interaction network. Shown is a schematic representation of interactions between 16 biological and metabolic pathways (KEGG at http://www.genome.ad.jp/kegg and Biocarta at http://www.biocarta.com), each containing a minimum of three atherosclerosis genes. The connections between the pathways illustrate the presence of common atherosclerosis genes in two pathways. IL-10, interleukin-10; LDLR, LDL receptor; PPARα, peroxisome proliferator-activated receptor α; Th1/Th2, T helper cell 1/2; TLR, toll-like receptor; TNF-α, tumor necrosis factor-α.

among one set of RI strains of mice. Mean levels of many cardiovascular traits, from heart rate to vessel and chamber dimensions, were first measured in a number of RI strains derived from the parental strains A/J and C57BL/6J. Pairwise Pearson's correlations were then calculated for all traits over all different genetic backgrounds. As a control, the phenotypic data were simply permuted and the results of the permutation analysis used as a baseline value for the standard correlation coefficient. The conclusions of the study were entirely consistent with physiological studies. For example, ventricular size was related to vessel wall thickness and diastolic dimensions, and heart rate was found to have an inverse relationship with cardiac dimensions. This proof-of-principle paper reveals how correlation matrices with the proper threshold can unravel causal relationships.

In a similar study, Stoll et al. (15) constructed a map of correlated cardiovascular traits by combining physiological profiles (correlation matrices between "likely determinant phenotypes" of cardiovascular traits) and genetic linkage analysis to unravel potential functional interactions between these traits that were not apparent using linkage analysis alone. Similarly, naturally occurring variation affecting the expression of genes in segregating populations has the potential to establish causal relationships among genes and could be used to construct gene-gene and gene-phenotype interaction networks.

Another network, illustrated in **Fig. 7**, represents a "pathway interaction" network. This network was constructed by identifying annotated pathways that contain three or more genes previously implicated in atherosclerosis. In summary, a list of 92 genes (see supplementary table) associated with atherosclerosis was selected (2). The genes in this list either have been shown to affect atherosclerosis through studies in genetically altered animals (transgenic or gene-targeted mice) or have shown evidence of association with atherosclerosis-related traits in multiple population studies. Publicly available annotated biological and metabolic pathways at KEGG (http://www.genome.ad.jp/kegg) and Biocarta (http://www.biocarta.com) were then searched for the presence of these atherosclerosis genes. Each node in Fig. 7 represents a pathway that contains a minimum of three atherosclerosis genes. From the original 92 genes, 39 genes exist in pathways that contain a minimum of 3 atherosclerosis genes. The links between the nodes represent the co-occurrence of a gene (or genes) in two biological pathways. As shown in Fig. 7, there are 16 pathways containing 353 unique genes, several of which overlap in various pathways. This analysis reveals numerous genes for which no known function has previously been associated with atherosclerosis. Thus, these genes should be considered as potential candidates, particularly if they reside at loci identified using linkage analysis.

## PROSPECTS

An improved understanding of the networks involved in atherosclerosis will have several important benefits. First, it will clarify interactions between various traits related to atherosclerosis (such as the components of the metabolic syndrome) and also between atherosclerosis and related disorders such as diabetes and osteoporosis. Such information may have clinically relevant predictive value. Second, it will provide new candidate genes for genetic studies, independent of biochemical approaches, thus contributing to the goal of developing genetic tests to assess the risk of atherosclerosis and predict responses to therapies. Third, it will help pinpoint the "weakest links" in pathways contributing to disease; for example, perturbations of highly connected nodes in a network are most likely to affect the output of a pathway. Such information should be useful in identifying targets for the development of new therapies and in predicting potential side effects of therapeutic interventions. Fourth, an understanding of atherosclerosis networks should help unravel the interactions between genetic and environmental factors in the disease.

The elucidation of networks for atherosclerosis will certainly require genome-wide approaches such as microarray analyses. Because atherosclerosis involves many systemic influences and multiple cell types, cell-based studies will be able to reveal only a subset of the important interactions. Thus, animal models, most likely the mouse, will be central to such studies. The most promising approach at present appears to be the combination of genetics and expression array analyses. The multiple perturbations in genetic crosses should allow the modeling of networks, but the validation of such models will probably require defined perturbations such as knockouts or RNAi-based approaches.

The authors thank Eric Schadt, Desmond Smith, and Steve Horvath for valuable discussions. Work in the authors' laboratory was supported in part by National Institutes of Health Grants HL-28481, HL-30568, HL-70526, and HL-60030, the Laubisch Fund, the University of California, Los Angeles, and the Bristol-Myers Squibb Unrestricted Biomedical Research Award. The authors thank the American Society of Genetics for graciously allowing us to reproduce Fig. 3.

## REFERENCES

1. Lusis, A. J., R. Mar, and P. Pajukanta. 2004. Genetics of atherosclerosis. *Annu. Rev. Genet.* In press.
2. Lusis, A. J., A. M. Fogelman, and G. Fonarow. 2004. Genetic basis of atherosclerosis: new genes and pathways. *Circulation.* In press.
3. Horikawa, Y., N. Oda, N. J. Cox, X. Li, M. Orho-Melander, M. Hara, Y. Hinokio, T. H. Lindner, H. Mashima, P. E. Schwarz, L. del Bosque-Plata, Y. Oda, I. Yoshiuchi, S. Colilla, K. S. Polonsky, S. Wei, P. Concannon, N. Iwasaki, J. Schulze, L. J. Baier, C. Bogardus, L. Groop, E. Boerwinkle, C. L. Hanis, and G. I. Bell. 2000. Genetic variation in the gene encoding calpain-10 is associated with type 2 diabetes mellitus. *Nat. Genet.* **26:** 163–175.
4. Botstein, D., and N. Risch. 2003. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat. Genet.* **33 (Suppl):** 228–237.
5. Ozaki, K., Y. Ohnishi, A. Iida, A. Sekine, R. Yamada, T. Tsunoda, H. Sato, M. Hori, Y. Nakamura, and T. Tanaka. 2002. Functional SNPs in the lymphotoxin-alpha gene that are associated with susceptibility to myocardial infarction. *Nat. Genet.* **32:** 650–654.

6. Williams, S. M., J. L. Haines, and J. H. Moore. 2004. The use of animal models in the study of complex disease: all else is never equal or why do so many human studies fail to replicate animal findings? *Bioessays.* **26:** 170–179.
7. Qin, M., Z. Zeng, J. Zheng, P. K. Shah, S. M. Schwartz, L. D. Adams, and B. G. Sharifi. 2003. Suppression subtractive hybridization identifies distinctive expression markers for coronary and internal mammary arteries. *Arterioscler. Thromb. Vasc. Biol.* **23:** 425–433.
8. Johnson, C. D., Y. Balagurunathan, K. P. Lu, M. Tadesse, M. H. Falahatpisheh, R. J. Carroll, E. R. Dougherty, C. A. Afshari, and K. S. Ramos. 2003. Genomic profiles and predictive biological networks in oxidant-induced atherogenesis. *Physiol. Genomics.* **13:** 263–275.
9. Shi, W., X. Wang, K. Tangchitpiyanond, J. Wong, Y. Shi, and A. J. Lusis. 2002. Atherosclerosis in C3H/HeJ mice reconstituted with apolipoprotein E-null bone marrow. *Arterioscler. Thromb. Vasc. Biol.* **22:** 650–655.
10. Stryke, D., M. Kawamoto, C. C. Huang, S. J. Johns, L. A. King, C. A. Harper, E. C. Meng, R. E. Lee, A. Yee, L. L'Italien, P. T. Chuang, S. G. Young, W. C. Skarnes, P. C. Babbitt, and T. E. Ferrin. 2003. Baygenomics: a resource of insertional mutations in mouse embryonic stem cells. *Nucleic Acids Res.* **31:** 278–281.
11. Peterfy, M., J. Phan, P. Xu, and K. Reue. 2001. Lipodystrophy in the fld mouse results from mutation of a new gene encoding a nuclear protein, lipin. *Nat. Genet.* **27:** 121–124.
12. Kile, B. T., K. E. Hentges, A. T. Clark, H. Nakamura, A. P. Salinger, B. Liu, N. Box, D. W. Stockton, R. L. Johnson, R. R. Behringer, A. Bradley, and M. J. Justice. 2003. Functional genetic analysis of mouse chromosome 11. *Nature.* **425:** 81–86.
13. Geary, R. L., J. M. Wong, A. Rossini, S. M. Schwartz, and L. D. Adams. 2002. Expression profiling identifies 147 genes contributing to a unique primate neointimal smooth muscle cell phenotype. *Arterioscler. Thromb. Vasc. Biol.* **22:** 2010–2016.
14. Stoehr, J. P., J. E. Byers, S. M. Clee, H. Lan, I. V. Boronenkov, K. L. Schueler, B. S. Yandell, and A. D. Attie. 2004. Identification of major quantitative trait loci controlling body weight variation in ob/ob mice. *Diabetes.* **53:** 245–249.
15. Stoll, M., A. W. Cowley, Jr., P. J. Tonellato, A. S. Greene, M. L. Kaldunski, R. J. Roman, P. Dumas, N. J. Schork, Z. Wang, and H. J. Jacob. 2001. A genomic-systems biology map for cardiovascular function. *Science.* **294:** 1723–1726.
16. Allayee, H., A. Ghazalpour, and A. J. Lusis. 2003. Using mice to dissect genetic factors in atherosclerosis. *Arterioscler. Thromb. Vasc. Biol.* **23:** 1501–1509.
17. Abiola, O., J. M. Angel, P. Avner, A. A. Bachmanov, J. K. Belknap, B. Bennett, E. P. Blankenhorn, D. A. Blizard, V. Bolivar, G. A. Brockmann, K. J. Buck, J. F. Bureau, W. L. Casley, E. J. Chesler, J. M. Cheverud, G. A. Churchill, M. Cook, J. C. Crabbe, W. E. Crusio, A. Darvasi, G. de Haan, P. Dermant, R. W. Doerge, R. W. Elliot, C. R. Farber, L. Flaherty, J. Flint, H. Gershenfeld, J. P. Gibson, J. Gu, W. Gu, H. Himmelbauer, R. Hitzemann, H. C. Hsu, K. Hunter, F. F. Iraqi, R. C. Jansen, T. E. Johnson, B. C. Jones, G. Kempermann, F. Lammert, L. Lu, K. F. Manly, D. B. Matthews, J. F. Medrano, M. Mehrabian, G. Mittlemann, B. A. Mock, J. S. Mogil, X. Montagutelli, G. Morahan, J. D. Mountz, H. Nagase, R. S. Nowakowski, B. F. O'Hara, A. V. Osadchuk, B. Paigen, A. A. Palmer, J. L. Peirce, D. Pomp, M. Rosemann, G. D. Rosen, L. C. Schalkwyk, Z. Seltzer, S. Settle, K. Shimomura, S. Shou, J. M. Sikela, L. D. Siracusa, J. L. Spearow, C. Teuscher, D. W. Threadgill, L. A. Toth, A. A. Toye, C. Vadasz, G. Van Zant, E. Wakeland, R. W. Williams, H. G. Zhang, and F. Zou. 2003. The nature and identification of quantitative trait loci: a community's view. *Nat. Rev. Genet.* **4:** 911–916.
18. Bodnar, J. S., A. Chatterjee, L. W. Castellani, D. A. Ross, J. Ohmen, J. Cavalcoli, C. Wu, K. M. Dains, J. Catanese, M. Chu, S. S. Sheth, K. Charugundla, P. Demant, D. B. West, P. de Jong, and A. J. Lusis. 2002. Positional cloning of the combined hyperlipidemia gene Hyplip1. *Nat. Genet.* **30:** 110–116.
19. Iakoubova, O. A., C. L. Olsson, K. M. Dains, D. A. Ross, A. Andalibi, K. Lau, J. Choi, I. Kalcheva, M. Cunanan, J. Louie, V. Nimon, M. Machrus, L. G. Bentley, C. Beauheim, S. Silvey, J. Cavalcoli, A. J. Lusis, and D. B. West. 2001. Genome-tagged mice (GTM): two sets of genome-wide congenic strains. *Genomics.* **74:** 89–104.
20. Demant, P. 2003. Cancer susceptibility in the mouse: genetics, biology and implications for human cancer. *Nat. Rev. Genet.* **4:** 721–734.
21. Singer, J. B., A. E. Hill, L. C. Burrage, K. R. Olszens, J. Song, M. Justice, W. E. O'Brien, D. V. Conti, J. S. Witte, E. S. Lander, and J. H. Nadeau. 2004. Genetic dissection of complex traits with chromosome substitution strains of mice. *Science.* **304:** 445–448.
22. Grupe, A., S. Germer, J. Usuka, D. Aud, J. K. Belknap, R. F. Klein, M. K. Ahluwalia, R. Higuchi, and G. Peltz. 2001. In silico mapping of complex disease-related traits in mice. *Science.* **292:** 1915–1918.
23. Chesler, E. J., S. L. Rodriguez-Zas, and J. S. Mogil. 2001. In silico mapping of mouse quantitative trait loci. *Science.* **294:** 2423.
24. Darvasi, A. 2001. In silico mapping of mouse quantitative trait loci. *Science.* **294:** 2423.
25. Park, Y. G., R. Clifford, K. H. Buetow, and K. W. Hunter. 2003. Multiple cross and inbred strain haplotype mapping of complex-trait candidate genes. *Genome Res.* **13:** 118–121.
26. Threadgill, D. W., K. W. Hunter, and R. W. Williams. 2002. Genetic dissection of complex and quantitative traits: from fantasy to reality via a community effort. *Mamm. Genome.* **13:** 175–178.
27. Pravenec, M., C. Wallace, T. J. Aitman, and T. W. Kurtz. 2003. Gene expression profiling in hypertension research: a critical perspective. *Hypertension.* **41:** 3–8.
28. Rinn, J. L., J. S. Rozowsky, I. J. Laurenzi, P. H. Petersen, K. Zou, W. Zhong, M. Gerstein, and M. Snyder. 2004. Major molecular differences between mammalian sexes are involved in drug metabolism and renal function. *Dev. Cell.* **6:** 791–800.
29. Aitman, T. J., A. M. Glazier, C. A. Wallace, L. D. Cooper, P. J. Norsworthy, F. N. Wahid, K. M. Al-Majali, P. M. Trembling, C. J. Mann, C. C. Shoulders, D. Graf, E. St. Lezin, T. W. Kurtz, V. Kren, M. Pravenec, A. Ibrahimi, N. A. Abumrad, L. W. Stanton, and J. Scott. 1999. Identification of CD36 (fat) as an insulin-resistance gene causing defective fatty acid and glucose metabolism in hypertensive rats. *Nat. Genet.* **21:** 76–83.
30. Karp, C. L., A. Grupe, E. Schadt, S. L. Ewart, M. Keane-Moore, P. J. Cuomo, J. Kohl, L. Wahl, D. Kuperman, S. Germer, D. Aud, G. Peltz, and M. Wills-Karp. 2000. Identification of complement factor 5 as a susceptibility locus for experimental allergic asthma. *Nat. Immunol.* **1:** 221–226.
31. Klein, R. F., J. Allard, Z. Avnur, T. Nikolcheva, D. Rotstein, A. S. Carlos, M. Shea, R. V. Waters, J. K. Belknap, G. Peltz, and E. S. Orwoll. 2004. Regulation of bone mass in mice by the lipoxygenase gene Alox15. *Science.* **303:** 229–232.
32. Mootha, V. K., C. M. Lindgren, K. F. Eriksson, A. Subramanian, S. Sihag, J. Lehar, P. Puigserver, E. Carlsson, M. Ridderstrale, E. Laurila, N. Houstis, M. J. Daly, N. Patterson, J. P. Mesirov, T. R. Golub, P. Tamayo, B. Spiegelman, E. S. Lander, J. N. Hirschhorn, D. Altshuler, and L. C. Groop. 2003. Pgc-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.* **34:** 267–273.
33. Alban, A., S. O. David, L. Bjorkesten, C. Andersson, E. Sloge, S. Lewis, and I. Currie. 2003. A novel experimental design for comparative two-dimensional gel analysis: two-dimensional difference gel electrophoresis incorporating a pooled internal standard. *Proteomics.* **3:** 36–44.
34. Swatton, J. E., S. Prabakaran, N. A. Karp, K. S. Lilley, and S. Bahn. 2004. Protein profiling of human postmortem brain using two-dimensional fluorescence difference gel electrophoresis (2-D DIGE). *Mol. Psychiatry.* **9:** 128–143.
35. Gavin, A. C., M. Bosche, R. Krause, P. Grandi, M. Marzioch, A. Bauer, J. Schultz, J. M. Rick, A. M. Michon, C. M. Cruciat, M. Remor, C. Hofert, M. Schelder, M. Brajenovic, H. Ruffner, A. Merino, K. Klein, M. Hudak, D. Dickson, T. Rudi, V. Gnau, A. Bauch, S. Bastuck, B. Huhse, C. Leutwein, M. A. Heurtier, R. R. Copley, A. Edelmann, E. Querfurth, V. Rybin, G. Drewes, M. Raida, T. Bouwmeester, P. Bork, B. Seraphin, B. Kuster, G. Neubauer, and G. Superti-Furga. 2002. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature.* **415:** 141–147.
36. Phizicky, E., P. I. Bastiaens, H. Zhu, M. Snyder, and S. Fields. 2003. Protein analysis on a proteomic scale. *Nature.* **422:** 208–215.
37. Watkins, S. M., P. R. Reifsnyder, H. J. Pan, J. B. German, and E. H. Leiter. 2002. Lipid metabolome-wide effects of the PPARgamma agonist rosiglitazone. *J. Lipid Res.* **43:** 1809–1817.
38. Jansen, R. C., and J. P. Nap. 2001. Genetical genomics: the added value from segregation. *Trends Genet.* **17:** 388–391.
39. Jansen, R. C. 2003. Studying complex biological systems using multifactorial perturbation. *Nat. Rev. Genet.* **4:** 145–151.
40. Machleder, D., B. Ivandic, C. Welch, L. Castellani, K. Reue, and A. J. Lusis. 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.

41. Brem, R. B., G. Yvert, R. Clinton, and L. Kruglyak. 2002. Genetic dissection of transcriptional regulation in budding yeast. *Science.* **296:** 752–755.
42. Lan, H., J. P. Stoehr, S. T. Nadler, K. L. Schueler, B. S. Yandell, and A. D. Attie. 2003. Dimension reduction for mapping mRNA abundance as quantitative traits. *Genetics.* **164:** 1607–1614.
43. Schadt, E. E., S. A. Monks, T. A. Drake, A. J. Lusis, N. Che, V. Colinayo, T. G. Ruff, S. B. Milligan, J. R. Lamb, G. Cavet, P. S. Linsley, M. Mao, R. B. Stoughton, and S. H. Friend. 2003. Genetics of gene expression surveyed in maize, mouse and man. *Nature.* **422:** 297–302.
44. Colinayo, V. V., J. H. Qiao, X. Wang, K. L. Krass, E. Schadt, A. J. Lusis, and T. A. Drake. 2003. Genetic loci for diet-induced atherosclerotic lesions and plasma lipids in mice. *Mamm. Genome.* **14:** 464–471.
45. Sabatti, C., S. Service, and N. Freimer. 2003. False discovery rate in linkage and association genome screens for complex disorders. *Genetics.* **164:** 829–833.
46. Raychaudhuri, S., J. M. Stuart, and R. B. Altman. 2000. Principal components analysis to summarize microarray experiments: application to sporulation time series. *Pac. Symp. Biocomput.* 455–466.
47. Xia, X., and Z. Xie. 2001. AMADA: Analysis of microarray data. *Bioinformatics.* **17:** 569–570.
48. Chase, K., D. R. Carrier, F. R. Adler, T. Jarvik, E. A. Ostrander, T. D. Lorentzen, and K. G. Lark. 2002. Genetic basis for systems of skeletal quantitative traits: principal component analysis of the canid skeleton. *Proc. Natl. Acad. Sci. USA.* **99:** 9930–9935.
49. Costello, T. J., M. D. Swartz, M. Sabripour, X. Gu, R. Sharma, and C. J. Etzel. 2003. Use of tree-based models to identify subgroups and increase power to detect linkage to cardiovascular disease traits (Abstract). *BMC Genet.* **4 (Suppl. 1):** 66.
50. Pociot, F., A. E. Karlsen, C. B. Pedersen, M. Aalund, and J. Nerup. 2004. Novel analytical methods applied to type 1 diabetes genome-scan data. *Am. J. Hum. Genet.* **74:** 647–660.
51. Lucek, P., J. Hanke, J. Reich, S. A. Solla, and J. Ott. 1998. Multi-locus nonparametric linkage analysis of complex trait loci with neural networks. *Hum. Hered.* **48:** 275–284.
52. Yoon, Y., J. Song, S. H. Hong, and J. Q. Kim. 2003. Analysis of multiple single nucleotide polymorphisms of candidate genes related to coronary heart disease susceptibility by using support vector machines. *Clin. Chem. Lab. Med.* **41:** 529–534.
53. Zhang, H., C. P. Tsai, C. Y. Yu, and G. Bonney. 2001. Tree-based linkage and association analyses of asthma. *Genet. Epidemiol.* **21 (Suppl. 1):** 317–322.
54. Atkinson, E. J., and M. de Andrade. 2003. Screening the genome to detect an association with hypertension. *BMC Genet.* **4 (Suppl. 1):** 63.
55. Bureau, A., J. Dupuis, B. Hayward, K. Falls, and P. Van Eerdewegh. 2003. Mapping complex traits using random forests. *BMC Genet.* **4 (Suppl. 1):** 64.
56. Chen, C. H., C. J. Chang, W. S. Yang, C. L. Chen, and C. S. Fann. 2003. A genome-wide scan using tree-based association analysis for candidate loci related to fasting plasma glucose levels. *BMC Genet.* **4 (Suppl. 1):** 65.
57. Guo, Z., X. Li, S. Rao, K. L. Moser, T. Zhang, B. Gong, G. Shen, L. Li, R. Cannata, E. Zirzow, E. J. Topol, and Q. Wang. 2003. Multivariate sib-pair linkage analysis of longitudinal phenotypes by three step-wise analysis approaches. *BMC Genet.* **4 (Suppl. 1):** 68.
58. Oh, C., K. Q. Ye, Q. He, and N. R. Mendell. 2003. Locating disease genes using Bayesian variable selection with the Haseman-Elston method. *BMC Genet.* **4 (Suppl. 1):** 69.
59. Tahri-Daizadeh, N., D. A. Tregouet, V. Nicaud, N. Manuel, F. Cambien, and L. Tiret. 2003. Automated detection of informative combined effects in genetic association studies of complex traits. *Genome Res.* **13:** 1952–1960.
60. Beaumont, M. A., and B. Rannala. 2004. The Bayesian revolution in genetics. *Nat. Rev. Genet.* **5:** 251–261.
61. Sachs, K., D. Gifford, T. Jaakkola, P. Sorger, and D. A. Lauffenburger. 2002. Bayesian network approach to cell signaling pathway modeling. *Sci. STKE.* **2002:** PE38.
62. Bockhorst, J., M. Craven, D. Page, J. Shavlik, and J. Glasner. 2003. A Bayesian network approach to operon prediction. *Bioinformatics.* **19:** 1227–1235.
63. Savoie, C. J., S. Aburatani, S. Watanabe, Y. Eguchi, S. Muta, S. Imoto, S. Miyano, S. Kuhara, and K. Tashiro. 2003. Use of gene networks from full genome microarray libraries to identify functionally relevant drug-affected genes and gene regulation cascades. *DNA Res.* **10:** 19–25.
64. Green, M. L., and P. D. Karp. 2004. A Bayesian method for identifying missing enzymes in predicted metabolic pathway databases. *BMC Bioinformatics.* **5:** 76.
65. Jensen, F. V. 2001. Bayesian Networks and Decision Graphs. Springer-Verlag, New York.
66. Pearl, J. 1988. Probabilistic Reasoning in Intelligent Systems. Morgan Kaufmann, San Mateo, CA.
67. Pearl, J. 2000. Causality: Models, Reasoning, and Inference. Cambridge University Press, Cambridge, UK.
68. Greenspan, R. J. 2001. The flexible genome. *Nat. Rev. Genet.* **2:** 383–387.
69. Barabasi, A. L., and Z. N. Oltvai. 2004. Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.* **5:** 101–113.
70. Jeong, H., S. P. Mason, A. L. Barabasi, and Z. N. Oltvai. 2001. Lethality and centrality in protein networks. *Nature.* **411:** 41–42.
71. Ideker, T., V. Thorsson, J. A. Ranish, R. Christmas, J. Buhler, J. K. Eng, R. Bumgarner, D. R. Goodlett, R. Aebersold, and L. Hood. 2001. Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science.* **292:** 929–934.
72. Herrgard, M. J., M. W. Covert, and B. O. Palsson. 2004. Reconstruction of microbial transcriptional regulatory networks. *Curr. Opin. Biotechnol.* **15:** 70–77.
73. Tong, A. H., G. Lesage, G. D. Bader, H. Ding, H. Xu, X. Xin, J. Young, G. F. Berriz, R. L. Brost, M. Chang, Y. Chen, X. Cheng, G. Chua, H. Friesen, D. S. Goldberg, J. Haynes, C. Humphries, G. He, S. Hussein, L. Ke, N. Krogan, Z. Li, J. N. Levinson, H. Lu, P. Menard, C. Munyana, A. B. Parsons, O. Ryan, R. Tonikian, T. Roberts, A. M. Sdicu, J. Shapiro, B. Sheikh, B. Suter, S. L. Wong, L. V. Zhang, H. Zhu, C. G. Burd, S. Munro, C. Sander, J. Rine, J. Greenblatt, M. Peter, A. Bretscher, G. Bell, F. P. Roth, G. W. Brown, B. Andrews, H. Bussey, and C. Boone. 2004. Global mapping of the yeast genetic interaction network. *Science.* **303:** 808–813.
74. Li, S., C. M. Armstrong, N. Bertin, H. Ge, S. Milstein, M. Boxem, P. O. Vidalain, J. D. Han, A. Chesneau, T. Hao, D. S. Goldberg, N. Li, M. Martinez, J. F. Rual, P. Lamesch, L. Xu, M. Tewari, S. L. Wong, L. V. Zhang, G. F. Berriz, L. Jacotot, P. Vaglio, J. Reboul, T. Hirozane-Kishikawa, Q. Li, H. W. Gabel, A. Elewa, B. Baumgartner, D. J. Rose, H. Yu, S. Bosak, R. Sequerra, A. Fraser, S. E. Mango, W. M. Saxton, S. Strome, S. Van Den Heuvel, F. Piano, J. Vandenhaute, C. Sardet, M. Gerstein, L. Doucette-Stamm, K. C. Gunsalus, J. W. Harper, M. E. Cusick, F. P. Roth, D. E. Hill, and M. Vidal. 2004. A map of the interactome network of the metazoan *C. elegans*. *Science.* **303:** 540–543.
75. Giot, L., J. S. Bader, C. Brouwer, A. Chaudhuri, B. Kuang, Y. Li, Y. L. Hao, C. E. Ooi, B. Godwin, E. Vitols, G. Vijayadamodar, P. Pochart, H. Machineni, M. Welsh, Y. Kong, B. Zerhusen, R. Malcolm, Z. Varrone, A. Collis, M. Minto, S. Burgess, L. McDaniel, E. Stimpson, F. Spriggs, J. Williams, K. Neurath, N. Ioime, M. Agee, E. Voss, K. Furtak, R. Renzulli, N. Aanensen, S. Carrolla, E. Bickelhaupt, Y. Lazovatsky, A. DaSilva, J. Zhong, C. A. Stanyon, R. L. Finley, Jr., K. P. White, M. Braverman, T. Jarvie, S. Gold, M. Leach, J. Knight, R. A. Shimkets, M. P. McKenna, J. Chant, and J. M. Rothberg. 2003. A protein interaction map of *Drosophila melanogaster*. *Science.* **302:** 1727–1736.
76. Nadeau, J. H., L. C. Burrage, J. Restivo, Y. H. Pao, G. Churchill, and B. D. Hoit. 2003. Pleiotropy, homeostasis, and functional networks based on assays of cardiovascular traits in genetically randomized populations. *Genome Res.* **13:** 2082–2091.