thematic review

Thematic review series: Systems Biology Approaches to Metabolic and Cardiovascular Disorders

Multi-organ whole-genome measurements and reverse engineering to uncover gene networks underlying complex traits

Jesper Tegnér,[*,†,§] Josefin Skogsberg,[*,†] and Johan Björkegren[1,*,†]

The Computational Medicine Group,* the Atherosclerosis Research Unit, Center for Molecular Medicine, King Gustaf V Research Institute, Department of Medicine, Karolinska Institutet, Karolinska University Hospital, Solna, SE-171 76 Stockholm, Sweden; Unit of Computational Medicine,† Clinical Gene Networks, Fogdevreten 2b, SE-171 77, Stockholm, Sweden; and Division of Computational Biology,§ Department of Physics, Linköpings Institute of Technology, Linköping University, SE-581 83 Linköping, Sweden

Abstract Together with computational analysis and modeling, the development of whole-genome measurement technologies holds the potential to fundamentally change research on complex disorders such as coronary artery disease. With these tools, the stage has been set to reveal the full repertoire of biological components (genes, proteins, and metabolites) in complex diseases and their interplay in modules and networks. Here we review how network identification based on reverse engineering, as applied to whole-genome datasets from simpler organisms, is now being adapted to more complex settings such as datasets from human cell lines and organs in relation to physiological and pathological states. Our focus is on the use of a systems biological approach to identify gene networks in coronary atherosclerosis. We also address how gene networks will probably play a key role in the development of early diagnostics and treatments for complex disorders in the coming era of individualized medicine.—Tegnér, J., J. Skogsberg, and J. Björkegren. **Multi-organ whole-genome measurements and reverse engineering to uncover gene networks underlying complex traits.** ***J. Lipid Res.*** **2007.** 48: **267–277.**

Supplementary key words global gene expression • coronary atherosclerosis • multicellular disease • computational modeling • individualized medicine

Candidate gene approaches, such as positional cloning (1), inherited from studies of single-gene disorders, have thus far generated fragmented knowledge of complex traits. Attention is now being redirected toward systems biological approaches. The general belief is that such approaches, unlike those based on candidate genes, can better take into account the inherent complexity of these disorders. Although systems theory has been around for quite some time (2), its applications in biology are flourishing because of the availability of whole-genome measurement technologies such as genomics (3) in combination with computational analysis and modeling (4). With an increasing number of research communities embracing systems biology, it is important to be clear about what this term means.

It is tempting to define systems biology as physiology or pathology—that is, the biological functions of an entire system rather than those of its molecular components. A stricter, and in our view more correct, definition of systems biology is research that focuses not on the molecular parts themselves (i.e., genes, proteins, metabolites) but on their interactions within networks. For such studies, the four Ms—manipulation, measurement, mining, and modeling—are key ingredients (4). Reverse engineering (the process of identifying gene networks from whole-genome data using an underlying computational model) of biological networks requires perturbations (i.e., manipulations) of the biological system followed by measurements of the system response using whole-genome measurement tools (5, 6). Then, mining (data interpretation, network identification) and modeling (model systems based on network architecture) are crucial for guiding the next round of experimental measurements

Manuscript received 14 November 2006 and in revised form 28 November 2006.

Published, JLR Papers in Press, December 1, 2006.
DOI 10.1194/jlr.R600030-JLR200

Abbreviations: CAD, coronary artery disease; LCM, laser-capture microdissection; ODE, ordinary differential equation; siRNA, small interfering RNA.

[1] To whom correspondence should be addressed.
e-mail: johan.bjorkegren@ki.se

in the most efficient way. The importance of iteration in systems biology, using decision making supported by computer models (**Fig. 1**), can easily be underestimated.

Thus, if systems biology alludes to the combined use of computer-supported network models and well-defined datasets of genome measurements, there is little doubt that this approach will prove extremely useful for unraveling cardiovascular and metabolic diseases and other complex disorders.

Initially, systems biological research was performed primarily in prokaryotic organisms (e.g., *Escherichia coli*) (7) and yeast (e.g., *Saccharomyces cerevisiae*) (8). However, in the last few years, it has been applied increasingly to mammalian and human cell lines (9–13). The consensus seems to be that this "bottom-up" approach of moving from simpler model systems to more complex settings is how systems biological approaches will eventually address complex disorders. We propose a parallel "top-down" approach to complex diseases. This approach does not refer to the DNA-RNA-protein phenotype hierarchy. Rather, it refers to moving from a disease in humans to animal models of that disease and eventually to relevant cellular models. We believe that disease-relevant hierarchal gene-gene interactions in complex diseases can be delineated by using reverse engineering, specifically by moving from the whole-body/organ level to the intercellular level and then to individual cell models. At the whole-body and intercellular level, this approach can be used to identify "principal networks" consisting of many (but not all) of the key interactions. The principal networks can then be delineated into complete biological networks using appropriate cell models of disease.

The principal networks of disease at the whole-body level will demonstrate how subphenotypes common to several complex disorders (e.g., inflammation, immunity, metabolism, cell proliferation, translation) integrate in a complex disease setting. In the next step, key aspects of the principal gene network can be investigated in relation to disease development at the intercellular level in animal models. In the end, the central aspects of the principal network isolated at the whole-body level and during disease development in animal models can be delineated in complete biological gene and protein networks at the cellular level by using both genetic perturbations, such as small interfering RNA (siRNA); gene deletion, overexpression, and variants; and environmental perturbations with compounds or metabolites.

In this review, we outline how this approach can be applied to identifying gene networks underlying complex traits, using coronary artery disease (CAD) as our primary example.

CORONARY ARTERY DISEASE

CAD is a degenerative disease that develops over decades from the stress of circulating blood cells and other plasma constituents that gradually alters the artery wall composition (cellular and extracellular), eventually leading to the formation of atherosclerosis plaques (**Fig. 2**) (14) The rate of atherosclerosis development depends both on environmental pressures and on the genetic makeup of the individual (15). Environmental pressures relevant to CAD are mainly mediated by airborne pollutants (including cigarette smoke), infections, and food intake (calories and cholesterol), and by behavioral factors, in particular the degree of stress and exercise. The net effect of environmental pressures filtered through the

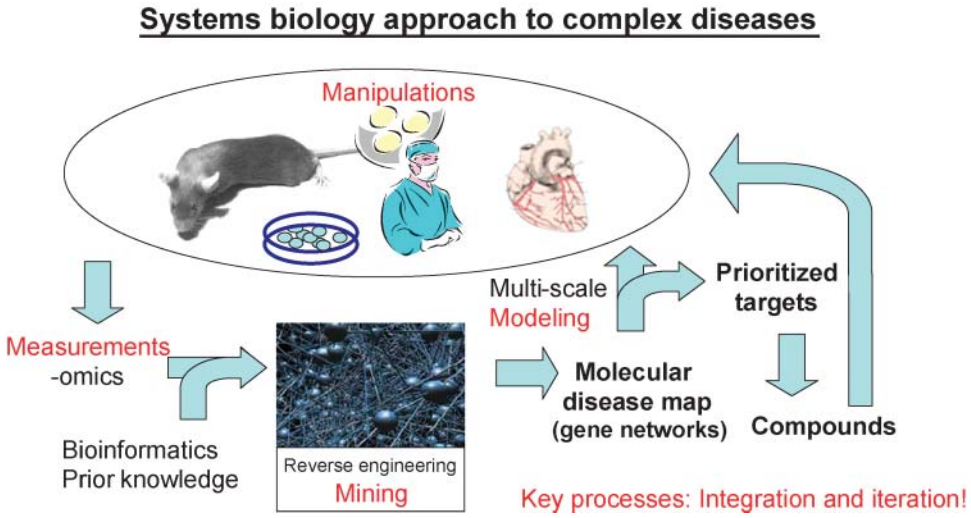

Fig. 1. A systems biological approach to complex diseases. By integrating whole-genome measurement (-omics) from clinical studies with experimental manipulations in disease-relevant model systems combined with prior knowledge (mining), reverse engineering can be used to infer regulatory gene networks and to generate molecular maps of disease development. These maps can be used to design computer multi-scale models of disease development that in turn can be used to design future experimental and clinical studies (iterative) and to prioritize disease targets (rather than validate single targets) against which compounds with better success rates can be identified. Identified compounds can also be taken into the iterative process.

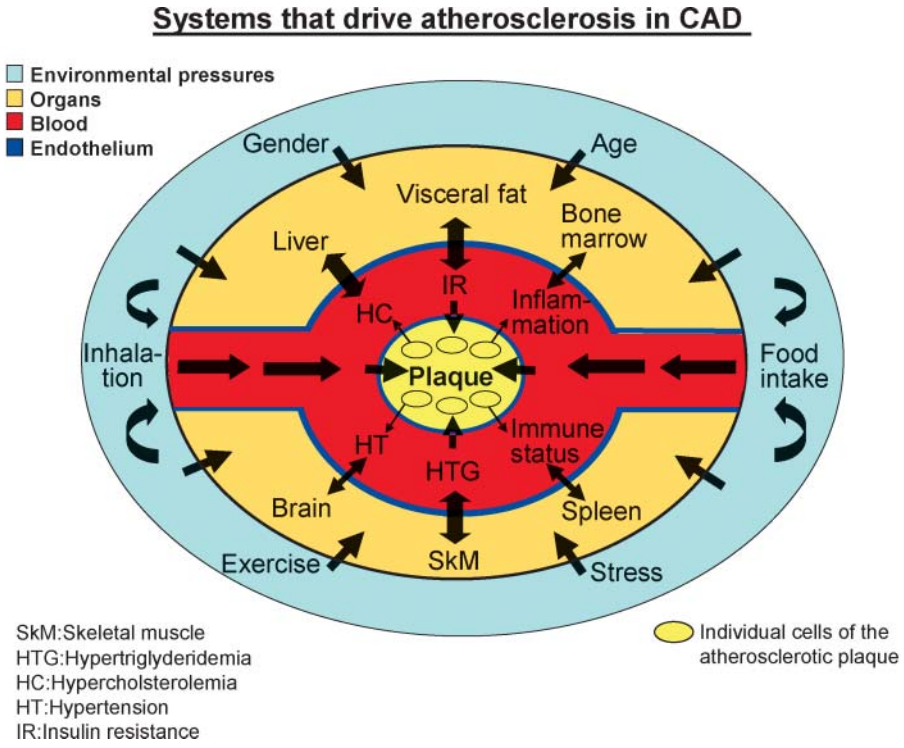

Fig. 2. Systems that drive atherosclerosis in coronary artery disease (CAD). The outside environment (light blue) affects the inside environment [bloodstream (red)] and organs (light brown) mainly through food intake and inhalation of airborne pollutants. The bloodstream carries the blueprint of these environmental pressures to alter organ function. Alterations in organ function are reflected back to the circulation, which in turn affects the development of atherosclerotic plaques (yellow).

individual genetic makeup is reflected by changes in blood flow and constituents.

Over years, environmental and lifestyle factors alter gene expression in organs. Changes in the expression of genes related to energy metabolism and inflammation in the liver, fat, or skeletal muscle are believed to be particularly relevant for CAD. In turn, alterations in gene expression are reflected in the circulation, where metabolic and inflammatory markers synthesized in these organs can be detected. Thus, measurements of plasma constituents (e.g., cholesterol and triglycerides), blood glucose and insulin levels, and inflammatory markers such as C-reactive protein are the standard way to detect hypertriglyceridemia, hypercholesterolemia, insulin resistance, diabetes, states of inflammation and immune activation, and other CAD phenotypes. These and most likely yet-unidentified constituents of blood and plasma determine the rate of atherosclerosis progression.

ATHEROSCLEROSIS

The first manifestation of atherosclerosis is the formation of foam cells in the intima of the arterial wall, leading to the histological appearance of fatty streaks. Briefly, circulating lipoproteins, mainly LDLs, adhere to the subendothelial matrix and undergo oxidative modifications that eventually alter gene and protein expression of endothelial cells. These changes lead to the recruitment of monocytes, which migrate to the intima of the arterial wall, differentiate into macrophages, and endocytose the modified LDL. These early steps are followed by additional inflammatory and immune responses (16), smooth muscle cell migration (17), and fibrosis, culminating in the formation of atherosclerotic plaques and apoptosis (18). The interplay of these biological processes, and probably others that have not been identified (14), underlies the development of atherosclerosis (**Fig. 3**).

A SYSTEMS BIOLOGICAL APPROACH TO CAD AND ATHEROSCLEROSIS VERSUS THE CANDIDATE GENE APPROACH

From a systems biological point of view, a key assumption is that if a CAD phenotype or an environmental pressure contributes to foam cell formation and plaque development, it must in some way be represented in cells, proteins, or metabolites in the circulation (not taking lymph or neuronal impact on the vessel wall into account) (Fig. 2). If so, it must also be true that constituents in the bloodstream reflect upstream organ activities and environmental pressures (Fig. 2). Some of the variables we monitor to group patients into CAD phenotypes—blood

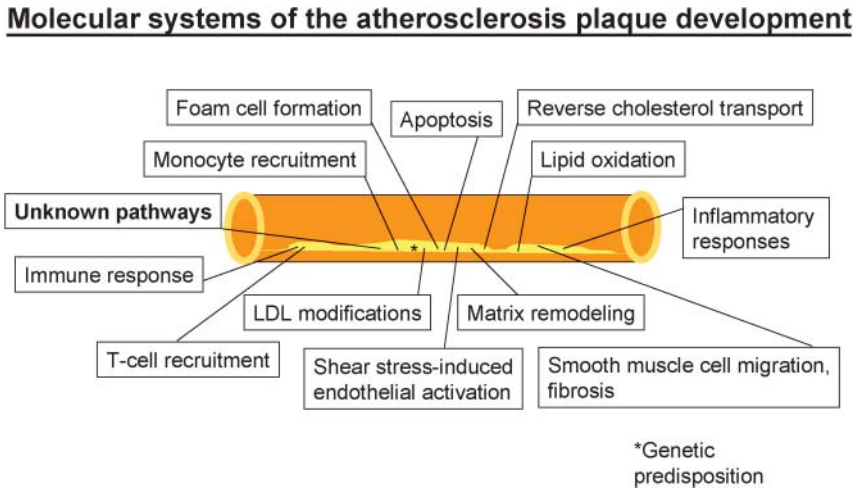

Fig. 3. Molecular systems of atherosclerosis development. Several known and most likely unknown pathways probably underlie the development of atherosclerosis. Depending on the environmental pressures from the bloodstream, the impact of individual pathways may vary.

constituents together with blood pressure, body mass index, sex, and hip-waist ratio—are established risk factors for CAD (14). However, it could be argued that the activity of organs central to CAD should reflect these and perhaps other unknown CAD phenotypes both at an earlier time and in greater molecular detail and in a fashion that reflects the status of environmental pressure over a longer period back in time. Thus, whole-genome measurements of CAD-relevant organs would give a more detailed picture of the risk for premature atherosclerosis.

The formation of fatty streaks and their transformation into plaques is a continuous, decades-long process. Although some of the pathological pathways involved in atherosclerosis are fairly well established (Fig. 3), their transcriptional regulation and interplay are not well understood. Moreover, the interplay between CAD phenotypes and the transcriptional regulation of atherogenesis over time is largely unexplored. From a systems biological standpoint, it should be possible to view these CAD phenotypes as perturbations that drive the development of atherosclerosis (see further section below, ''Gene network identification of atherosclerosis in humans'').

In some respects, our incomplete understanding of the pathology of atherosclerosis can be attributed to the candidate gene approach (15), which rests on the assumption that one or a few genes are the major contributors to disease development. The positional cloning approach was inherited from research on simple disorders, in which a single genetic mutation triggers the disease (19). In this setting, the candidate gene approach has been very successful in pinpointing the genetic causes of several diseases. In a similar fashion, this approach has been successful in delineating specific pathways of complex traits, which perhaps can be viewed as single-gene disorders within those traits. As an example, deCode has in this way elucidated several important candidate genes and thereby identified novel pathways and biological processes of complex traits (www.decode.com). However, there is growing concern that the candidate gene approach will not be sufficient to reveal the entire repertoire of genes and their interactions in networks. Such knowledge is probably essential for understanding complex traits.

Until recently, it must be said, there have been few alternatives to the candidate gene/pathway approach to complex diseases. Now that tools to measure the activity of the entire genome are available, systems biology approaches can be used to delineate the regulation and interplay between CAD phenotypes and pathways of atherosclerosis development, both known and unknown. In fact, using computer-supported algorithms, it should be possible to identify gene networks of atherosclerosis development.

MULTIPLE-CELL-TYPE VERSUS SINGLE-CELL-TYPE DISEASES

Atherosclerosis is a disease involving multiple cell types, including monocyte/macrophages, foam cells, endothelial cells, smooth muscle cells, and T cells. In contrast, cancer typically originates from a single type of cell, although if allowed to develop, almost all cancer cells become individual cell types (20). This is also partly true for atherosclerosis. For instance, gene expression profiles from foam cells isolated by laser-capture microdissection (LCM) (21) differ depending on where they were isolated (22). It is also likely that cell types involved in atherogenesis change their phenotype as the disease develops. The origin of smooth muscle cells is also debated: do they originate principally from the arterial media or from circulating progenitor cells of hematopoietic origin (23)?

It has been suggested that whole-genome measurements of complex traits like atherosclerosis should be performed on specific disease cell types separately (i.e., smooth muscle cells, foam cells, endothelial cells, and possibly T cells) in order to generate meaningful data (24). However, it is also quite possible—and in our view

necessary—to view the plaque or lesion as a single functional system. Whole-genome measurements of the entire system will provide information about how it is regulated in relation to changes in the environment outside the system. For instance, knowledge of the lesion expression phenotype is necessary for understanding gene expression changes induced by CAD phenotypes. The interpretation of lesion expression will also help to pinpoint specific cell types, pathways, and individual genes that merit further analysis in cellular models of disease. On the other hand, gene expression profiles of individual cell types are necessary to enable meaningful validation of atherosclerosis model systems in cell cultures. Clearly, it would be desirable to use a combination of whole-lesion and cell-type-specific whole-genome expression profiles.

WHOLE-GENOME ACTIVITY MEASUREMENTS

Various methods are available for measuring the entire repertoire of biological activities, and new methods are constantly being developed (25). It is evident that gene expression (mRNA levels) has many advantages in the first round of systems biological approaches to complex disorders. It is the most mature technique (26) and provides a fairly robust signal, and analytical tools for statistical analysis of the data are constantly being improved (27, 28). Moreover, mRNA levels take into account both degradation and synthesis. Whole-genome protein platforms are being refined but are still more suitable for studies of smaller, well-defined biological problems. No doubt, increasing the sensitivity of these technologies will eventually pave the way for researchers to identify biomarkers in plasma that will become useful for early diagnostics. Other technologies focus on metabolites (29). For instance, in this review series and elsewhere (30), lipidomics is being put forward in relation to cardiovascular and metabolic diseases.

Other kinds of whole-genome measurements are also emerging. Transcription factor binding and protein-protein interactions are particularly interesting in relation to network identification (31). Last but not least, whole-genome technology platforms to screen for single nucleotide polymorphisms are increasingly being used to study cardiovascular disease (32).

NETWORK IDENTIFICATION

A network is a graph defined by nodes and connecting edges. A protein interaction network is an example of an undirected network, because an edge only indicates whether two proteins bind to each other. A gene regulatory network, however, is a directed network, in which the directionality between two genes represents a mechanistic causality or a probabilistic dependency. For example, a transcription factor influences the target gene to which it binds. An edge in a gene network inferred from gene expression data (hereafter referred to as "gene networks") can thus directly reflect transcription factor activity or indirectly reflect protein-protein interaction between two genes affecting transcription or RNA degradation.

During the last 5 years, in silico studies (6, 33, 34) based on ordinary differential equations (ODEs) have demonstrated that it is possible to extract a directed graph from repeated measurements of the activities of nodes within the graph in response to perturbations such as knock-out or siRNA experiments. Thus, it appears that gene regulatory networks can be identified from systematic series of whole-genome expression profiles. The chief advantage of in silico studies is that computational methods for network identification can be systematically evaluated as knowledge is gained of the true network underlying the simulated gene expression data. Importantly, gene networks have been identified by using these algorithms to analyze gene expression data from *E. coli* (7) and *S. cerevisiae* (35). Furthermore, probabilistic Bayesian models have also been successfully applied to gene expression data (36). A Bayesian model captures the probabilistic dependency between genes, whereas combining an ODE model with experimental perturbations allows the causal relation between genes to be identified. Finally, calculating the correlations in gene expression values between genes across a number of samples provides a measure of the degree of coexpression but not causality, which, however, may be useful as a first step in a subsequent network analysis (37).

Several insights have surfaced from these computational and experimental studies. First, in designing microarray experiments, it is essential to incorporate a perturbation protocol. This allows differentially expressed genes to be identified that are directly or indirectly affected by the perturbation. Combining an underlying computational model of the gene regulatory system with several experimental perturbations is sufficient to identify the network. It has also been possible to recover gene networks from time series data in yeast (38), although the quality of the reconstructed networks has been more difficult to ascertain.

Another insight is that in several practical applications, the number of experiments is too small relative to the number of genes in the network of interest. For example, it is not feasible to identify the edges of a biological network with thousands of nodes from only a handful of experiments. Experience from computational analysis of gene regulatory networks demonstrates that it is essential to introduce other constraints on the types of solutions (networks) that we can expect from a set of measurements of the system. In the context of gene expression analysis, several types of prior knowledge can be used. For example, transcription factor and protein-protein binding data can be used to limit the number of possible edges within a network (6). An important simplification that facilitates the identification of a network is the notion that networks are sparse—that is, most genes have only a small number of edges (39). Text-mining algorithms operating on PubMed are an important data source for collecting putative edges.

Finally, gene expression data have been used to define the edges defined by a putative edge library that are ac-

tive under a given experimental condition. For example, Luscombe et al. (31) used gene expression data from yeast at different stages of the cell cycle to define the active edges from the library of putative edges defined by transcription factor and chromatin immunoprecipitation chip binding data. Similarly, de Lichtenberg et al. (40) defined a putative protein-protein interaction network from datasets, which they subsequently combined with gene expression data to define active protein subnetworks during the cell cycle.

Although useful when only a small set of experiments is available, this procedure is severely limited because novel edges cannot be detected. An important challenge for analyzing data in small samples will be to systematically integrate edges based on prior knowledge into an algorithmic network identification paradigm. This would not only constrain the number of possible solutions, but would also allow the detection of both previously characterized edges and novel edges for a given whole-genome expression dataset. Interestingly, new algorithms are being developed for reverse engineering of biological networks in human cell lines (11, 13) and in response to an inflammatory stimuli (12). A recent computational and experimental analysis of protein networks by Sachs et al. (41) employed a perturbation approach in which simultaneous measurements of multiple phosphorylated proteins and phospholipid components in human immune cells were analyzed by an underlying Bayesian model. Several of the inferred edges were experimentally verified, thus validating the applicability of a perturbation approach beyond gene networks and an ODE model. In conclusion, to translate these promising results and insights into an analysis of gene networks involved in multifactorial diseases, it is necessary to utilize prior knowledge in an algorithmic framework and to adapt the perturbation approach to a more complex disease-relevant setting.

Gene network identification in CAD-relevant organs

Can whole-genome expression profiling of CAD-relevant organs (i.e., liver, fat, and skeletal muscle) be the basis for network identification in CAD? Indeed, we believe reverse engineering can be used to identify principal (incomplete node representation) and biological networks (e.g., gene networks) of CAD and atherosclerosis by expression profiling of multiple organs.

First, a list of putative CAD genes must be established to distinguish gene activity related to CAD from that related to the normal function of a CAD-relevant organ. One possibility for extracting CAD-relevant gene expression is to use a case-control study design in which false discover rates are calculated by simple comparisons using multiple-testing-adjusted differential testing (28). Novel multivariate differential testing can capture additional relevant genes (42). Alternatively, an association study design can be used in which genes are related to surrogate measurements, such as the degree of coronary atherosclerosis obtained by quantitative coronary angiography (43) or magnetic resonance imaging (44), or the extent of carotid plaques determined by ultrasound examination of flow and intima media thickness (45). Regression calculations can then be performed to associate gene expression values to those of the surrogate measurement (and possibly to other subphenotypic data). Another alternative is to use clustering techniques (46). As an example, coupled two-way clustering (47) can be used to identify clusters based on gene activity and then examine whether any of the gene clusters also cluster the patients grouped according to the surrogate measurement. Cluster techniques are preferable because, unlike differential testing, they are unsupervised (i.e., the surrogate measure is not used in the analysis).

Once a list of putative CAD genes has been obtained (a cutoff is decided by the false discovery rate, typically 0.05), the network of these and related genes can be established by systematically integrating prior edges within an algorithmic network identification algorithm as described (see section "Network identification" above). The gene network established in this way will not be the biological gene network (i.e., genes that actually interact through their protein products), but it will reflect gene-gene interactions with none or several unidentified intermediate genes (principal or incomplete networks, **Figs. 4** and **5**). Nonetheless, the architectures of these networks will allow the identification of key aspects in the overall regulation of CAD gene activity in these organs, including activity in the circulating blood. These principal networks may also indicate biological functions, pathways, and possibly individual genes that are coactivated or coregulated in several CAD-relevant organs (Figs. 4 and 5).

As demonstrated in yeast, network wiring may be substantially altered under different environmental conditions (31). That is, the wiring of a general regulatory gene network of inflammation in macrophages, for example, may differ substantially depending on the particular site and context of the inflammation (e.g., organ-, cell-, or disease-specific in atherosclerosis, infection, obesity, or diabetes). Also, the extensive rewiring even within yeast suggests that high-hierarchal cross-species rewiring may even be greater. Rewiring is one reason why a clear disease focus may be desirable when using a systems biological approach to analyzing complex traits.

Gene network identification of atherosclerosis in humans

In published (24) and presumably in many unpublished or ongoing clinical trials, atherosclerotic plaques from patients undergoing carotid artery or coronary artery bypass grafting or specific atherosclerosis cell types obtained by LCM have been collected for whole-genome expression analysis to reveal genes central to atherosclerosis (21). Thus far, analyses of these datasets have been limited to traditional differential expression analysis and clustering.

However, these gene expression datasets should also be useful for network identification in a fashion similar to that described for CAD-related organs above. Another possibility is to use several separate whole-genome expression datasets related to a given disease to identify functional modules (48, 49). Modules define groups

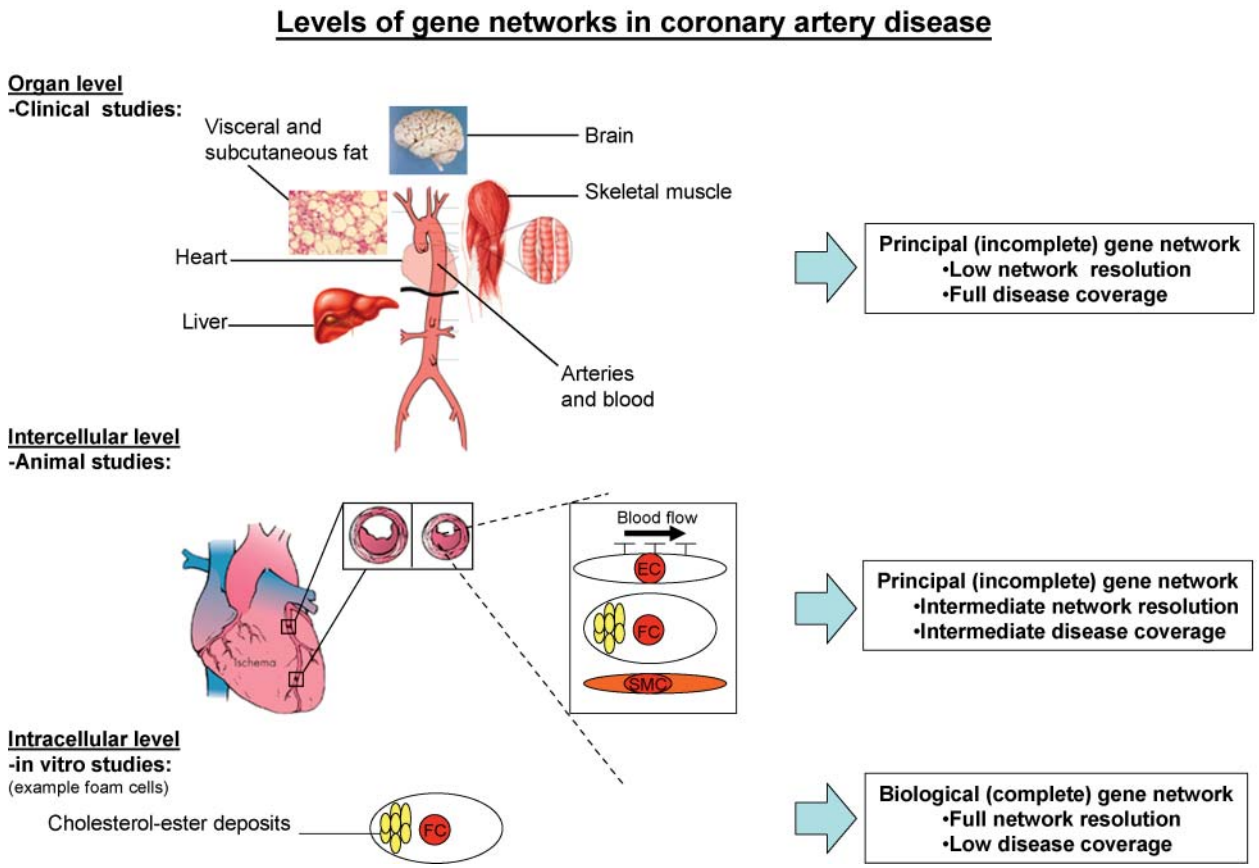

Fig. 4. Levels of gene networks in coronary artery disease. In clinical studies, whole-genome measurements can be extracted at the whole-body organ level. The principal (incomplete) gene network at this level will have high disease coverage and relevance but low network resolution (incomplete with many undetected intermediate nodes). Animal model systems can be used to study atherosclerosis at the intercellular level (e.g., in the lesion plaques) preferentially over time. The gene network inferred from these studies will reflect the combination of gene expression in the lesion cell types (endothelial cells, smooth muscle cells, and macrophages/foam cells; T cells are not shown) in combination or individually (obtained by laser-capture microdissection) and represent an intermediate resolution and coverage. At the cellular level, full biological gene networks can be defined by using combinations of perturbation techniques and whole-genome measurements (e.g., in a foam cell model).

of genes with common functions but, unlike principal networks, do not define individual gene-gene edges within the module.

As pointed out, perturbations of the system and monitoring of the response are key for identifying biological networks (6). In the setting of complex traits, the challenge is that clinical whole-genome datasets contain repeats of patients and not repeats of genetic manipulations, such as deletions of yeast genes (35). We propose an alternative way to achieve system perturbations by using the clinical phenotypes in humans (e.g., CAD phenotypes) as environmental perturbations (50) of the complex trait under investigation (e.g., atherosclerosis). For instance, by comparing whole-genome expression profiles of atherosclerosis in matched patients with and without diabetes, atherosclerosis genes central to diabetes could be identified. Similarly, it should be possible to identify a larger set of atherosclerosis genes by using several related CAD phenotypes (e.g., plasma glucose, insulin, and proinsulin) in the same manner. If all CAD phenotypes are used in this fashion, it should be possible to identify a large portion of the atherosclerosis network, including highly connected genes (so-called hubs) (51, 52), that underlies atherosclerosis development.

Moreover, this approach can be extended by using sets of whole-genome expression profiles of CAD-relevant organs. Hundreds of gene clusters related to CAD could be generated by using two-way clustering (47) on gene pair ratios of whole-genome expression profiles of CAD-relevant organs from patients with and without CAD. These gene expression clusters could then be used to group the patients into subgroups (i.e., two-way clustering), which in turn would serve as further perturbations to delineate additional aspects of the atherosclerosis network.

There are several caveats with this approach. First, subgrouping CAD patients based on clusters of CAD-relevant gene expression might not have a direct or even an indirect impact on atherosclerosis development. Second, CAD phenotypes may be more or less well-defined in individual patients. It is therefore of great importance to have enough patients in each CAD phenotype to ensure that the phenotype under examination in fact prevails.

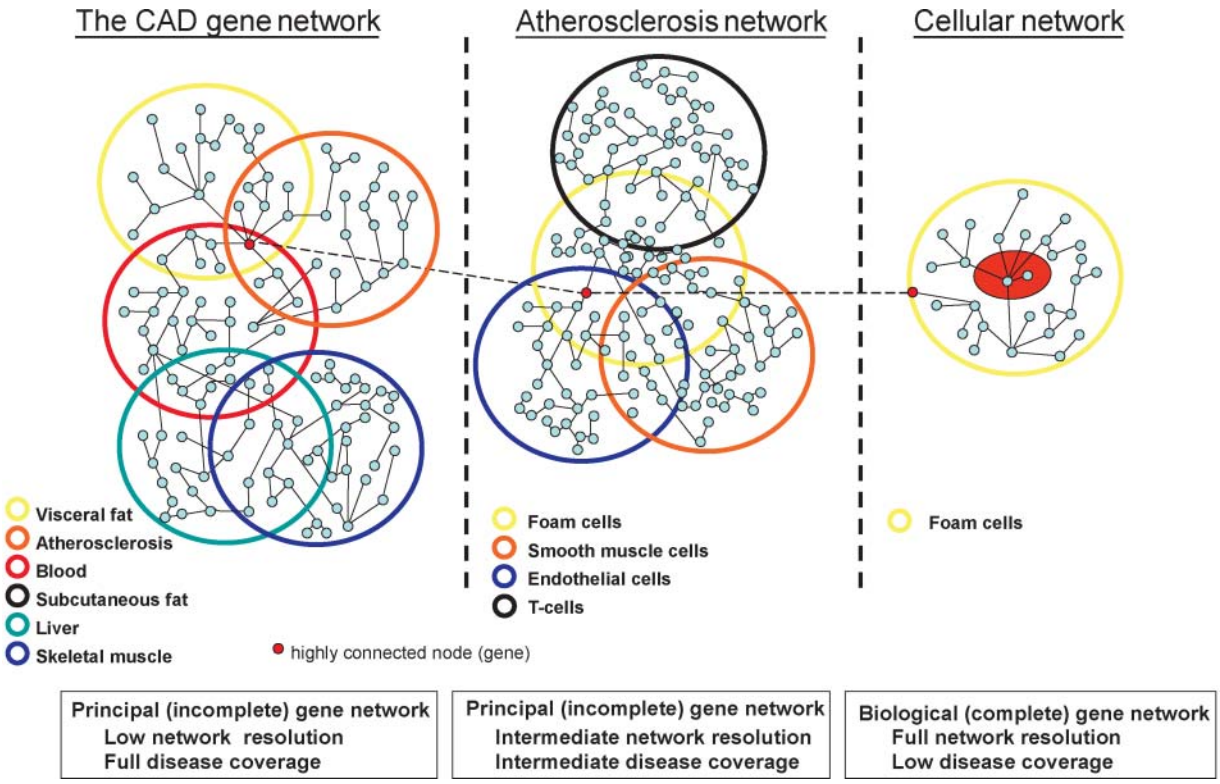

Fig. 5. The intersections between network levels in CAD. The principal gene network of CAD can help to highlight important structures (here, a highly regulated gene "hub" is given as an example) that can be delineated at the intermediate and full level in animal models and cellular systems, respectively.

This will require cohorts of hundreds and possibly thousands of patients. From our calculations of the average gene expression variation in CAD patients (unpublished observations), we estimate that a CAD cohort of up to 1,000 patients, including 100 non-CAD controls, will be required to ascertain that the phenotype of a subgroup of patients will prevail and thus function as an atherosclerosis perturbation.

Finally, specific gene, RNA, or protein perturbations (e.g., deletions or siRNA in cultured cells) differ from environmental perturbations. By default, changes in whole-genome activity induced by specific perturbations generate downstream causal effects. In contrast, the primary targets of environmental perturbations are unknown in most instances. Interestingly, however, ways to detect the primary target of environmental perturbations are beginning to emerge (e.g., by examining responses to environmental perturbations such as compounds and to metabolites) (53, 54). With modifications, these approaches should be applicable to delineating the primary targets of phenotypic perturbations.

The idea of using perturbations to uncover gene networks in complex disease has also been put forward in relation to the impact of genetic variants (of which the primary target is known) on so-called expression phenotypes (55). This topic is being reviewed in another section of this review series.

Gene network identification in model systems of atherosclerosis

Animal model systems can be used to study changes in the activity of gene networks over time in whole lesions or in individual atherosclerosis cell types obtained by LCM (21). For instance, there are algorithms that enable gene networks to be inferred from time series of whole-genome datasets (38). The gene networks in lesions can also be identified by screening the whole-genome lesion expression responses to environmental perturbations such as changes in plasma cholesterol or glucose levels or treatments with compounds. It should also be possible to use knock-out or transgenic animal models of atherosclerosis to disclose lesion networks.

When the key properties of the principal gene network have been identified from whole-genome expression profiles of human organ samples (Figs. 4 and 5) and validated in relation to disease development in animal models (Figs. 4 and 5), cellular model systems can be used to identify the complete biological networks (Figs. 4 and 5). Depending on the nature of the identified subnetwork, endothelial cells, smooth muscle cells, or macrophages can be used. In these systems, the perturbation approach for reverse engineering can be applied in full. By systematically silencing key genes with siRNA (56) in this subnetwork, the regulatory gene network can be identified (5) (Figs. 4 and 5).

Thus, in our view, perturbations are the central paradigm of systems biology (6). In **Fig. 6**, we summarize how perturbations may prove useful for network identification of atherosclerosis development in CAD.

THE USE OF GENE NETWORKS UNDERLYING COMPLEX TRAITS LIKE CAD: FUTURE PERSPECTIVES

Identifying gene networks of physiological or pathological biological processes may prove useful in several respects. First, gene networks can be the basis for computational models to predict biological responses (6). When thoroughly validated, such models can be used to make predictions that can be tested experimentally, as well as to explore questions that are not amenable to experimental inquiry (4). In the longer term, these models can be used to assess the impact of novel drugs, targets, and functional genetic variants. Indeed, models are being used for well-understood biological processes such as the cell cycle (57), cell growth (58), and metabolic analyses (59), and for comparative studies of the robustness of biological oscillation circuits (60).

Gene networks can also be useful for drug evaluation (61). First, targets can be selected based on their position in the gene network instead of on an isolated validation of a specific target. Network-based target evaluation sets the stage for validating all possible targets at once, so that the most promising targets can be prioritized. Moreover, a gene network can be used to calculate the mechanism of action of a given compound (7, 53, 54). In preclinical studies, network analysis will better predict possible side effects, for instance by assessing the impact of a compound on the gene networks of organs frequently involved in side effects, such as the liver, kidneys, skeletal muscle, and visceral fat deposits.

Eventually, gene networks could provide information on the genetic profiles and the current environmental pressures of individual patients. With this information, therapies could be tailored to the individual, and the individual network response to the treatment could be monitored. This development will require a vast improvement in the efficacy of DNA sequencing and expression profiling technologies, similar to what we have witnessed for the data storage capacities of computers during the last two decades. Eventually, when genome scans are requested by the general population, rapid, inexpensive, and secure technologies for DNA sequencing and gene expression profiling (and other "-omics") will follow.

CONCLUSIONS

Research in the field of gene network inference is rapidly growing. Most efforts reflect a "bottom-up" approach, in which reverse engineering paradigms from research on bacteria and yeast are being applied to eu-

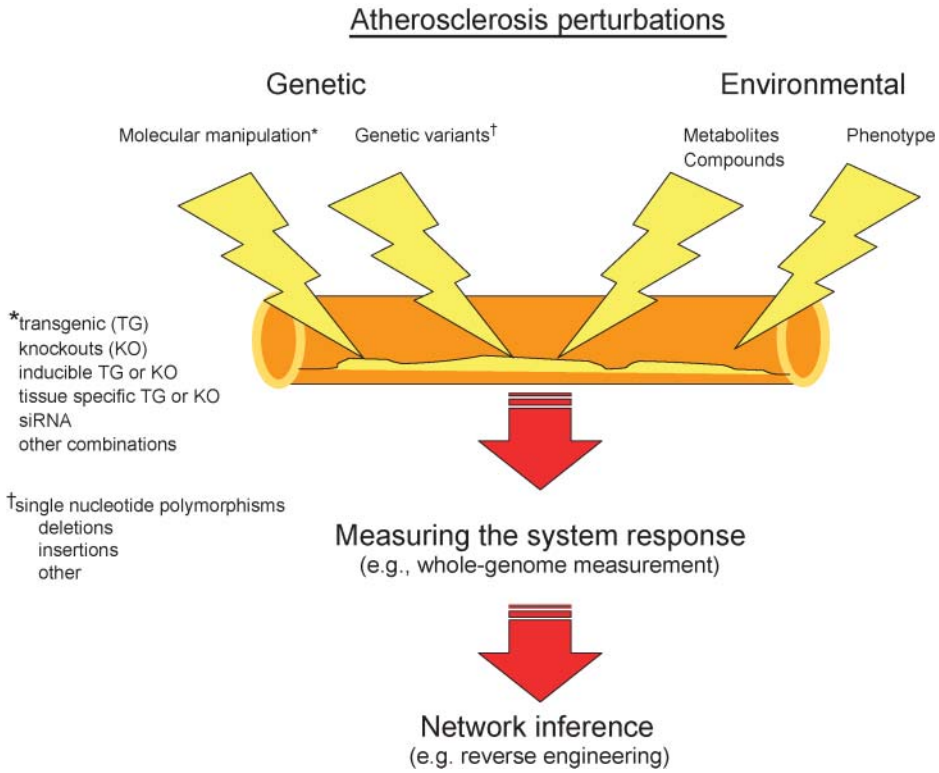

Fig. 6. Atherosclerosis perturbations to uncover gene networks. In clinical cohorts and animal and cell model systems of atherosclerosis, there are several ways of perturbing atherosclerosis that if monitored with whole-genome measurements can be used for network identification.

karyotic model systems under normal and pathological conditions. Several efforts around the globe are focusing on metabolic pathways and various aspects of inflammation. Moreover, Schadt and coworkers (55) are adopting an interesting strategy to use functional genetic variants in combination with gene expression profiles to infer directed networks. In this review, we have emphasized the need for a "top-down" systems biological approach to cardiovascular and metabolic disease, moving from a disease in humans, to animals models of that disease, and eventually to relevant cellular models. In the end, large datasets of whole-genome measurements from well-characterized human samples with a special disease focus will be crucial for identifying key and possibly disease-specific subsystems (indeed, *reducing* the number of systems to the most central). Computer models can be then be developed that are adjusted to the particular wiring diagram of the disease under investigation. In CAD, atherosclerosis can be triggered by many factors, each of which may affect the regulatory gene network of atherosclerosis differently. To develop effective individualized treatments, the genetic makeup of the individual has to be set in the context of the individual environmental pressures. To assess the individual risk of CAD, several computational models will therefore be required for evaluating the genetic and environmental factors that drive the development of disease in individual patients. We are now witnessing the emerging promise that the reactive approach to health care (treating diseases) will be replaced with a proactive approach (preventing diseases). Together with an increasingly computerized society, this uplifting possibility will open up completely new avenues for providing health care in the 21st century.

This work was supported by grants from the Swedish Research Council (J.B.), the Stockholm County Council (J.B.), the Swedish Heart-Lung Foundation (J.B.), the King Gustaf V and Queen Victoria Foundation (J.B.), the Swedish Society of Medicine (J.B., J.T.), the Hans and Loo Osterman Foundation for Geriatric Research (J.S., J.B.), the National Network for Cardiovascular Disease (J.S.), the Professor Nanna Swartz Fund (J.B.), the Foundation for Old Servants (J.B.), the Magnus Bergvalls Foundation (J.B.), Åke Wiberg Stiftelse (J.B., J.T), the Wennergren Foundation (J.T.), the Swedish Foundation for Strategic Research (J.B., J.T.), and the Karolinska Institutet PhD Program in Medical Bioinformatics (J.B., J.T.).

REFERENCES

1. Menzel, S. 2002. Genetic and molecular analyses of complex metabolic disorders: genetic linkage. *Ann. N. Y. Acad. Sci.* **967:** 249–257.
2. von Bertalanffy, L. 1971. General Theory of Systems: Application to Psychology. Walter de Gruyter, Berlin.
3. Ginsburg, G. S., M. P. Donahue, and L. K. Newby. 2005. Prospects for personalized cardiovascular medicine: the impact of genomics. *J. Am. Coll. Cardiol.* **46:** 1615–1627.
4. Kitano, H. 2002. Computational systems biology. *Nature.* **420:** 206–210.
5. Tegnér, J., M. K. Yeung, J. Hasty, and J. J. Collins. 2003. Reverse engineering gene networks: integrating genetic perturbations with dynamical modeling. *Proc. Natl. Acad. Sci. USA.* **100:** 5944–5949.
6. Tegnér, J., and J. Björkegren. 2007. Perturbations to uncover gene networks. *Trends Genet.* In press.
7. Gardner, T. S., D. di Bernardo, D. Lorenz, and J. J. Collins. 2003. Inferring genetic networks and identifying compound mode of action via expression profiling. *Science.* **301:** 102–105.
8. Mustacchi, R., S. Hohmann, and J. Nielsen. 2006. Yeast systems biology to unravel the network of life. *Yeast.* **23:** 227–238.
9. 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, et al. 2003. Genetics of gene expression surveyed in maize, mouse and man. *Nature.* **422:** 297–302.
10. Tian, Q., S. B. Stepaniants, M. Mao, L. Weng, M. C. Feetham, M. J. Doyle, E. C. Yi, H. Dai, V. Thorsson, J. Eng, et al. 2004. Integrated genomic and proteomic analyses of gene expression in mammalian cells. *Mol. Cell. Proteomics.* **3:** 960–969.
11. Basso, K., A. A. Margolin, G. Stolovitzky, U. Klein, R. Dalla-Favera, and A. Califano. 2005. Reverse engineering of regulatory networks in human B cells. *Nat. Genet.* **37:** 382–390.
12. Calvano, S. E., W. Xiao, D. R. Richards, R. M. Felciano, H. V. Baker, R. J. Cho, R. O. Chen, B. H. Brownstein, J. P. Cobb, S. K. Tschoeke, et al. 2005. A network-based analysis of systemic inflammation in humans. *Nature.* **437:** 1032–1037.
13. Nilsson, R., V. B. Bajic, H. Suzuki, D. di Bernardo, J. Bjorkegren, S. Katayama, J. F. Reid, M. J. Sweet, M. Gariboldi, P. Carninci, et al. 2006. Transcriptional network dynamics in macrophage activation. *Genomics.* **88:** 133–142.
14. Lusis, A. J., A. M. Fogelman, and G. C. Fonarow. 2004. Genetic basis of atherosclerosis: part I: new genes and pathways. *Circulation.* **110:** 1868–1873.
15. Watkins, H., and M. Farrall. 2006. Genetic susceptibility to coronary artery disease: from promise to progress. *Nat. Rev. Genet.* **7:** 163–173.
16. Hansson, G. K. 2005. Inflammation, atherosclerosis, and coronary artery disease. *N. Engl. J. Med.* **352:** 1685–1695.
17. Owens, G. K., M. S. Kumar, and B. R. Wamhoff. 2004. Molecular regulation of vascular smooth muscle cell differentiation in development and disease. *Physiol. Rev.* **84:** 767–801.
18. Tabas, I. 2005. Consequences and therapeutic implications of macrophage apoptosis in atherosclerosis: the importance of lesion stage and phagocytic efficiency. *Arterioscler. Thromb. Vasc. Biol.* **25:** 2255–2264.
19. Lusis, A. J. 2003. Genetic factors in cardiovascular disease. 10 questions. *Trends Cardiovasc. Med.* **13:** 309–316.
20. Spencer, S. L., R. A. Gerety, K. J. Pienta, and S. Forrest. 2006. Modeling somatic evolution in tumorigenesis. *PLoS Comp. Biol.* **2:** e108.
21. Emmert-Buck, M. R., R. F. Bonner, P. D. Smith, R. F. Chuaqui, Z. Zhuang, S. R. Goldstein, R. A. Weiss, and L. A. Liotta. 1996. Laser capture microdissection. *Science.* **274:** 998–1001.
22. Trogan, E., R. P. Choudhury, H. M. Dansky, J. X. Rong, J. L. Breslow, and E. A. Fisher. 2002. Laser capture microdissection analysis of gene expression in macrophages from atherosclerotic lesions of apolipoprotein E-deficient mice. *Proc. Natl. Acad. Sci. USA.* **99:** 2234–2239.
23. Bentzon, J. F., C. Weile, C. S. Sondergaard, J. Hindkjaer, M. Kassem, and E. Falk. 2006. Smooth muscle cells in atherosclerosis originate from the local vessel wall and not circulating progenitor cells in apoE knockout mice. *Arterioscler. Thromb. Vasc. Biol.* **26:** 2696–2702.
24. Bijnens, A. P. J. J., E. Lutgens, T. Ayoubi, J. Kuiper, A. J. Horrevoets, and M. J. A. P. Daemen. 2006. Genome-wide expression studies of atherosclerosis: critical issues in methodology, analysis, interpretation of transcriptomics data. *Arterioscler. Thromb. Vasc. Biol.* **26:** 1226–1235.
25. Joyce, A. R., and B. O. Palsson. 2006. The model organism as a system: integrating 'omics' data sets. *Nat. Rev. Mol. Cell Biol.* **7:** 198–210.
26. Canales, R. D., Y. Luo, J. C. Willey, B. Austermiller, C. C. Barbacioru, C. Boysen, K. Hunkapiller, R. V. Jensen, C. R. Knight, K. Y. Lee, et al. 2006. Evaluation of DNA microarray results with quantitative gene expression platforms. *Nat. Biotechnol.* **24:** 1115–1122.
27. Cleveland, W. 1979. Robust locally weighted regression and smoothing scatterplots. *J. Am. Stat. Assoc.* **74:** 829–836.
28. Efron, B., R. Tibshirani, J. Storey, and V. Tusher. 2001. Empirical Bayes analysis of a microarray experiment. *J. Am. Stat. Assoc.* **96:** 1151–1160.
29. Pelczer, I. 2005. High-resolution NMR for metabomics. *Curr. Opin. Drug Discov. Devel.* **8:** 127–133.

30. Wenk, M. R. 2005. The emerging field of lipidomics. *Nat. Rev. Drug Discov.* **4:** 594–610.
31. Luscombe, N. M., M. M. Babu, H. Yu, M. Snyder, S. A. Teichmann, and M. Gerstein. 2004. Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature.* **431:** 308–312.
32. Schadt, E. E. 2006. Novel integrative genomics strategies to identify genes for complex traits. *Anim. Genet.* **37(Suppl 1):** 18–23.
33. Friedman, N. 2004. Inferring cellular networks using probabilistic graphical models. *Science.* **303:** 799–805.
34. Gardner, T. S., and J. J. Faith. 2005. Reverse-engineering transcription control networks. *Physics of Life Reviews.* **2:** 65–88.
35. Hughes, T. R., M. J. Marton, A. R. Jones, C. J. Roberts, R. Stoughton, C. D. Armour, H. A. Bennett, E. Coffey, H. Dai, Y. D. He, et al. 2000. Functional discovery via a compendium of expression profiles. *Cell.* **102:** 109–126.
36. Friedman, N. 2004. Inferring cellular networks using probabilistic graphical models. *Science.* **303:** 799–805.
37. Gargalovic, P. S., M. Imura, B. Zhang, N. M. Gharavi, M. J. Clark, J. Pagnon, W. P. Yang, A. He, A. Truong, S. Patel, et al. 2006. Identification of inflammatory gene modules based on variations of human endothelial cell responses to oxidized lipids. *Proc. Natl. Acad. Sci. USA.* **103:** 12741–12746.
38. Gustafsson, M., M. Hörnqvist, and A. Lombardi. 2005. Constructing and analyzing a large-scale gene-to-gene regulatory network—lasso-constrained inference and biological validation. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2:** 254–261.
39. Barabasi, A. L., and Z. N. Oltvai. 2004. Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.* **5:** 101–113.
40. de Lichtenberg, U., L. J. Jensen, S. Brunak, and P. Bork. 2005. Dynamic complex formation during the yeast cell cycle. *Science.* **307:** 724–727.
41. Sachs, K., O. Perez, D. Pe'er, D. A. Lauffenburger, and G. P. Nolan. 2005. Causal protein-signaling networks derived from multiparameter single-cell data. *Science.* **308:** 523–529.
42. Xiao, Y., R. Frisina, A. Gordon, L. Klebanov, and A. Yakovlev. 2004. Multivariate search for differentially expressed gene combinations. *BMC Bioinformatics.* **5:** 164.
43. Austen, W. G., J. E. Edwards, R. L. Frye, G. G. Gensini, V. L. Gott, L. S. Griffith, D. C. McGoon, M. L. Murphy, and B. B. Roe. 1975. A reporting system on patients evaluated for coronary artery disease. Report of the Ad Hoc Committee for Grading of Coronary Artery Disease, Council on Cardiovascular Surgery, American Heart Association. *Circulation.* **51:** 5–40.
44. Larose, E., Y. Yeghiazarians, P. Libby, E. K. Yucel, M. Aikawa, D. F. Kacher, E. Aikawa, S. Kinlay, F. J. Schoen, A. P. Selwyn, et al. 2005. Characterization of human atherosclerotic plaques by intravascular magnetic resonance imaging. *Circulation.* **112:** 2324–2331.
45. Villines, T. C., and A. J. Taylor. 2005. Non-invasive atherosclerosis imaging: use to assess response to novel or combination lipid therapies. *Curr. Drug Targets Cardiovasc. Haematol. Disord.* **5:** 557–564.
46. D'Haeseleer, P. 2005. How does gene expression clustering work? *Nat. Biotechnol.* **23:** 1499–1501.
47. Getz, G., E. Levine, and E. Domany. 2000. Coupled two-way clustering analysis of gene microarray data. *Proc. Natl. Acad. Sci. USA.* **97:** 12079–12084.
48. Segal, E., M. Shapira, A. Regev, D. Pe'er, D. Botstein, D. Koller, and N. Friedman. 2003. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat. Genet.* **34:** 166–176.
49. Segal, E., N. Friedman, N. Kaminski, A. Regev, and D. Koller. 2005. From signatures to models: understanding cancer using microarrays. *Nat. Genet.* **37(Suppl):** 38–45.
50. 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.
51. Barabasi, A. L., and R. Albert. 1999. Emergence of scaling in random networks. *Science.* **286:** 509–512.
52. Jeong, H., S. P. Mason, A. L. Barabasi, and Z. N. Oltvai. 2001. Lethality and centrality in protein networks. *Nature.* **411:** 41–42.
53. Hallen, K., J. Bjorkegren, and J. Tegner. 2006. Detection of compound mode of action by computational integration of whole-genome measurements and genetic perturbations. *BMC Bioinformatics.* **7:** 51.
54. di Bernardo, D., M. J. Thompson, T. S. Gardner, S. E. Chobot, E. L. Eastwood, A. P. Wojtovich, S. J. Elliott, S. E. Schaus, and J. J. Collins. 2005. Chemogenomic profiling on a genome-wide scale using reverse-engineered gene networks. *Nat. Biotechnol.* **23:** 377–383.
55. Schadt, E. E., J. Lamb, X. Yang, J. Zhu, S. Edwards, D. Guhathakurta, S. K. Sieberts, S. Monks, M. Reitman, C. Zhang, et al. 2005. An integrative genomics approach to infer causal associations between gene expression and disease. *Nat. Genet.* **37:** 710–717.
56. Timmons, L., H. Tabara, C. C. Mello, and A. Z. Fire. 2003. Inducible systemic RNA silencing in Caenorhabditis elegans. *Mol. Biol. Cell.* **14:** 2972–2983.
57. Tyson, J. J. 1999. Models of cell cycle control in eukaryotes. *J. Biotechnol.* **71:** 239–244.
58. Ibarra, R. U., J. S. Edwards, and B. O. Palsson. 2002. *Escherichia coli* K-12 undergoes adaptive evolution to achieve *in silico* predicted optimal growth. *Nature.* **420:** 186–189.
59. Edwards, J. S., R. U. Ibarra, and B. O. Palsson. 2001. *In silico* predictions of *Escherichia coli* metabolic capabilities are consistent with experimental data. *Nat. Biotechnol.* **19:** 125–130.
60. Morohashi, M., A. E. Winn, M. T. Borisuk, H. Bolouri, J. Doyle, and H. Kitano. 2002. Robustness as a measure of plausibility in models of biochemical networks. *J. Theor. Biol.* **216:** 19–30.
61. Gerhold, D. L., R. V. Jensen, and S. R. Gullans. 2002. Better therapeutics through microarrays. *Nat. Genet.* **32(Suppl):** 547–551.