Abstract—The combination of systems biology and large data sets offers new approaches to the study of cardiovascular diseases. These new approaches are especially important for the common cardiovascular diseases that have long been described as multifactorial. This promise is undermined by biologists’ skepticism of the spider web–like network diagrams required to analyze these large data sets. Although these spider webs resemble composites of the familiar biochemical pathway diagrams, the complexity of the webs is overwhelming. As a result, biologists collaborate with data analysts whose mathematical methods seem much like those of experts using Ouija boards. To make matters worse, it is not evident how to design experiments when the network implies that many molecules must be part of the disease process. Our goal is to remove some of this mystery and suggest a simple experimental approach to the design of experiments appropriate for such analysis. We will attempt to explain how combinations of data sets that include all possible variables, graphical diagrams, complementation of different data sets, and Bayesian analyses now make it possible to determine the causes of multifactorial cardiovascular disease. We will describe this approach using the term causal analysis. Finally, we will describe how causal analysis is already being used to decipher the interactions among cytokines as causes of cardiovascular disease. (Arterioscler Thromb Vasc Biol. 2012;32:2821-2835.)

Key Words: analysis ■ atherosclerosis ■ Bayes ■ Bayesian ■ causality ■ genetics ■ graphical ■ heart failure ■ hypertension ■ multifactorial ■ systems biology

For many biologists, network analysis comes across as being a session with an expert user of the Ouija board. The expert, a bioinformatics analyst, uses seemingly magical approaches to derive hypotheses by passing massive quantities of data over what look like metabolic maps.

Nonetheless, the tools of systems biology and network analysis are essential because we now know that most common diseases have multiple causes. We excuse our own ignorance by calling these diseases multifactorial. In fact, multifactorial diseases are probably more common than diseases explained by even small numbers of causes. Genetic studies of hypertension, heart failure, breast cancer, and even schizophrenia show that inheritance of common diseases depends on combinations of multiple rare mutations. As shown in Figure 1, it is difficult or impossible to determine such causality using traditional hypothesis-based experiments.

Multifactorial Diseases
Cancer seems relatively simple when compared with cardiovascular disease. Most cancers are explained by a virus, by a mutation, or by an isolated environmental cause. By comparison, cancer biologists may see cardiovascular diseases as impossibly complex problems that rarely can be attributed to a single cause or even a well-defined sequence of causes. Lacking well-defined hypotheses, cardiovascular biologists resort to terms like multifactorial. Unfortunately, to resort to the term multifactorial probably means that we regard hypertension, heart failure, and atherosclerosis as too complex to understand.

Perhaps because of this frustrating lack of well-defined hypotheses, cardiovascular journals and our study sections are eager to avoid fishing expeditions. Reviewers are skeptical of conclusions not based on targeting a specific molecule,
and studies using the tools of systems biology are forced to produce evidence by a knockout or drug experiment targeting some specific molecule.

These reviewer concerns about systems biology are valid. The size of the data sets produced by these new tools is so large that we add the suffix -omic to their names. The implication of omic is that we have measured all the possible variables, so we want to believe that the answer is someplace in our data. Even if we believe the data sets are complete, we need help! To provide this help, bioinformatics software gives us diagrams that resemble the metabolic maps we have as posters on our laboratory walls. Bioinformaticians then reward our belief in these diagrams using skills that seem like those of ancient Chinese priests reading the cracks in oracle bones in the Shang dynasty (1700–1000 BCE).

Hopefully, biological maps are based on science that is well beyond the knowledge of auguries performed 3000 years ago. Since the explosion of biochemistry in the mid-20th century, biologists have rigorously assembled these maps, molecule by molecule. The result is a powerful set of network diagrams that account for a great deal of biology and allow us to design meaningful experiments. These network diagrams are the analytic tools of systems biology.

Genetic methods allow us to identify causes of disease by producing network diagrams that resemble our metabolic maps.\(^9,10\) Of course, not all causes are genetic, and the resemblance to metabolic maps could be deceptive. The central goal of this review is to explain the relationship between metabolic maps and the diagrams, or networks of the systems biologists.

Our review will focus on the genetics of gene expression. We will show how the unbiased nature of natural variation in genetic sequence can be used to make causal inferences. These inferences allow us to attribute the effects to multiple genes interacting with one another. We will also show how this approach avoids the type of error common to typical biology experiments where only 1 or 2 molecules can be manipulated at the same time.

### A Different Approach to Statistics

Much of the skepticism about systems biology arises because the tools used to construct network diagrams are very different from statistical methods familiar to most biologists. Familiar concepts, such as rejecting the null hypothesis, the 0.05 level, and complementation analysis, are replaced by uncomfortable concepts, such as partial correlation, data clustering, graphic diagrams, and Bayesian coefficients.

Collectively, these concepts may be called causal analysis. The statistical methods of causal analysis are well established in economics, ecology, and mathematical philosophy. Important names, such as Bayes, Le Place, Seward Wright, and Judea Pearl,\(^{11–13}\) are associated with concepts not usually taught in our statistics courses. A particularly good introduction to Bayesian concepts and causal analysis for most biologists may be found in Appendix 1 of the book *Causality* by Judea Pearl.\(^12\)

### Hypothesis-Based Science Can Be Misleading

Rigorous theorists, including Pearl and many of the new generation of systems biologists, are critical of hypothesis-based science. Their point is well taken. When we select a single hypothesis to test, we usually ignore an unknown number of other possible hypotheses. Put another way, hypothesis-based studies are biased.

This bias is obvious in biochemistry because biochemical experiments are done under well-controlled conditions that attempt to eliminate irrelevant variables. As we will discuss later, genetic analysis also provides an alternative to biochemistry. The underlying logic of causal analysis provides a third set of tools to define the interactions between 2 molecules even in a complex system as that shown in Figure 1.

These alternatives to biochemistry have become essential as we have acquired the ability to look at very large systems. Biological mechanisms that make sense at the level of 2 interacting molecules may not make sense when we attempt to understand many molecules interacting with each other. Worse, although Newton had 3 simple laws for all motion of the planets, metabolic maps relate pairs of molecules using diverse mathematical operations ranging from simple linear equations to the complex mathematics of stochastic probability. Physicists and chemists refer to our field as messy.

In summary, Pearl’s criticisms are not an indictment of experiments based on biochemistry. Rather, as shown in
Astronomy illustrates the value of knowing, or at least believing that we know, a data set is complete. By the 1840s, Newtonian mechanics, as applied in the Kepler equations, seemed to be able to account for orbits of all the planets. These claims were based on the belief that there were only 7 planets. In modern terms, the planet-omic data set was 7. Astronomical measurements based on 7 objects were close enough to being complete to allow a good test of the Kepler equations. However, once anomalies in the values for the 7 known planets were found, Urbain Le Verrier was able to predict the position of the then-undiscovered Neptune. Similar extensions of the model led to the discovery of Pluto by Lowell and Tombaugh. If Neptune, or some other similar cause, was never found, Newton’s laws would have been overturned.

How do we know today’s omic data sets include sufficient elements of a biological system to test our models? The DNA sequence seems to be a good example where we come close to an omic description. The completeness of other omic ideas is less convincing. For example, the completeness of expression data sets is suspect because of technical limitations of chip hybridization or sample size in an RNA sequencing analysis. Nonetheless, our assumptions about the number of genes may be false. We still measure the quantity of each gene using chips tagged with sequences from every known protein. This system, that is, the RNA expressome, now needs to be expanded to include an unknown number of noncoding RNAs as well as RNAs that complement noncoding but functional RNAs.15

Despite these concerns, we can be reassured by the astronomers’ success with their estimate of 7 planets as long as we understand that the biologist’s analysis is limited to answering questions about sequences that code for protein, that is if we confine our systems analysis to messenger RNA.

In contrast to messenger RNA, the omicity of proteomic data are more problematic. Although there are only ~20 000 sequences that code for polypeptides, we do not know how many proteins exist as a result of alternative start sites, splicing, or posttranslational processing. This does not mean that proteomics cannot be omic if we define a subset we believe we can measure. There are believable lists of all the proteins involved in particular functions; for example, the proteins of the ribosome, the proteins of the electron transport system, the components of the proteosome, or the transcription factors. Similarly, biochemical methods can identify all the proteins in complexes defined by immunoprecipitation or other methods of fractionation. The list of cytokines also fulfills these criteria. At the end of this review we will discuss causal approaches to analysis of cytokines using proteomic data.

- Dilemma 3. The third dilemma is called false error rate. The large numbers of comparisons being made using data sets with large numbers of variables means that the credibility of estimate of individual correlations between variables can become so low that significant results may not be seen. Statisticians call this the false error rate and attempt to correct for it by increasing the stringency of statistical tests made across large numbers of experiments.
False error corrections in large data sets are necessarily subjective. That means that although we know we have some error, the best we can do is agree on how much skepticism to tolerate. One startling example of this is the work based on serial analysis of gene expression. In this early form of deep sequencing, short 3′ serial message sequences were placed in tandem, leading to the identification of differentially expressed genes by frequency of expression. Sample sizes, that is, the numbers of tandem sequences, were in the low thousands, meaning that low abundance sequences were rarely represented. Nonetheless, some very low abundance important genes were found despite numbers well below what might have been acceptable by a rigorous statistical analysis.

Extending the null hypothesis to an analysis with large, or even omic numbers of variables, is not at all obvious. Even if we leave aside the dilemma posed by multiple comparisons of independent results, we need to ask whether 0.05 is appropriate when the correlations being studied involve multiple steps in a pathway. What does this criterion mean if the conclusions of the article require 10 separate experiments all tested at the 0.05 level? Should we believe the pathway is valid or use a statistical method that accounts for the 1/20 confidence level of each of the steps in the pathway?

**Empirical Modules as Phenotypes**

Variation of a single mRNA is easy to use as a phenotype. This is true even within the context of very large data sets where expression levels of each of the 20,000 genes may be measured for correlation with genetic variations in the genomes of hundreds of individuals. This concept will be discussed later under the topic Expression (as a) Quantitative Trait Locus (eQTL) in genetic studies. It is more difficult to understand how we can use a phenotype comprising a large number of mRNAs that may be coexpressed in different animals. A familiar example of mRNA data is the display of expression data sets as heat maps. In these maps, coexpressed genes show up as bands of color. Because the edges of these bands are fuzzy, the coexpressed mRNAs look like clouds. Some genes make up red clouds and others green clouds. These clouds of coexpressed mRNAs can be quantified using statistical tools. Such defined sets of correlated values are called modules. Defining parameters of a module, for example, its center of gravity, can be used as a phenotype just as we can use the value for an individual mRNA to define an eQTL. Because modules are made up of many measurements of different genes, we can focus our analysis on the center of gravity of the module. This should greatly reduce the error intrinsic in measuring individual genes because we can average the error of measurements of individual genes.

The low error within a module makes these phenotypes based on expression modules very powerful tools in detecting causal relationships. We will refer to modules defined only by the experimental data as empirical modules.

**Empirical Versus Ad Hoc Modules**

Use of the term module in bioinformatics is not limited to empirical data. In typical articles, modules are constructed not only from experimental data but from guesses derived from prior knowledge the investigator believes is true. Often, this ad hoc manipulation of the data was referred to as Bayesian. For example, Zhu et al used protein–protein association data to cluster mRNA expression data in yeast, presumably because they believed that proteins that interact with one another are more likely to be coexpressed. In other words, Zhu et al biased their data analysis by using a quantified, Bayesian estimate of the hypothesis “mRNA expression correlates with association in protein complexes.”

There is good reason to use such biases in constructing modules. For example, we can cluster genes involved in the cell death pathway and then ask whether expression of this death pathway module varies across different genetic strains of a model organism. As another example we could take all the proteins of the proteasome and ask whether their expression is coregulated. Of interest to cardiovascular biology, a recent review suggests that the proteasome is critical to angiogenesis and, of course, the proteasome is central to regulation of inflammation via its control of nuclear factor-κB.

A module defined by knowledge outside of the experiment itself comprises a hypothesis. We will refer to such modules as ad hoc modules.

It is very important to understand the difference between empirical and ad hoc construction of modules. Again, empirical modules have no bias. Finding an eQTL for an empirical module is sufficient evidence that some sequence in the locus is causing the variation in the module. In contrast, statistical tests are needed to determine whether the correlation of the ad hoc module with genetic loci is unusual. Our faith in the causal inference of any ad hoc module will depend on those tests.

Put another way, an ad hoc module is a form of hypothesis. The test of that hypothesis is whether the bias increases the correlation in an expected fashion. Later, we will suggest a method for refining ad hoc modules by focusing on biases that are well proven. We will also discuss how this approach requires that the biologist learn when it is appropriate to violate the traditional use of the null hypothesis.

In summary, biologists need to be wary of the difference between empirical and ad hoc modules. Identification of empirical modules has no bias and, therefore, identifies causal pathways. If causal conclusions, rather than vague guesses are to be made, all prior knowledge used to construct ad hoc modules needs to be defined in a quantitative manner.

**Traditional Approaches in Biology**

**Hypothesis Testing**

Biological cause and effect is defined when we believe we can show that variation of a measurement is dependent on changes between experimental groups where we control, or believe we control, all relevant variables. The usual argument is over whether the correlation is strong enough to imply causality. The convention is to accept the correlation as real if there is <1/20 or 0.05 probability that there is no difference between the 2 conditions.

Blood pressure experiments provide a very good example of the problems of control for the unknown variables.
A drop in mean blood pressure of 10 mm Hg after treating 12 animals with a drug would justify a claim that the drug caused the fall in pressure. To make this claim credible we need to have appropriate controls and show that a blood pressure change of 10 mm Hg is unlikely to be the result of chance. Usually this latter claim would center around a statistical test of the variation on blood pressure showing that there is <0.05 probability that the null hypothesis can be accepted. Even then, the blood pressure experiment may be challenged by the suggestion of an inadequate control. For example, a recent knockout experiment claimed that the gene RGS2 controls blood pressure.22 Later studies failed to reproduce the result, apparently because the effect depended on diurnal variations in the animals’ responses to environmental variables.23,24 For this and other reasons, biostatisticians insist that biological experiments can only show correlation, not causality.

Even if the experiments validating each step in a pathway have been done many times, there is another reason to be skeptical about the validity of any pathway based on hypothesis testing. Karl Popper, a philosopher of science, pointed out that the only questions that can be answered by hypothesis-based experiments are the ones we ask. Because other questions are not asked, hypothesis-based experiments may give misleading answers.

Returning to the example in Figure 1, even genetics experiments may have hidden sources of error. The common mouse used in knockout (KO) experiments, 129, has an unknown number of spontaneous, homozygous mutations. Knockout mice derived from these embryonic stem (ES) cells are then mated with a strain other than the ones the ES cells derive from, again without considering the underlying mutations in that strain. Even when the resulting mice are inbred, the result each time is a unique genotype. Because genes interact with other genes, a phenomenon called epistasis,26 the same gene in different knockout strains may have different phenotypes.

**Complementation Testing**

Long before we had any knowledge of DNA, geneticists believed in genes, that is, people knew that phenotypes, properties of different animals or plants with a species, were heritable. Complementation experiments, that is, experiments using crosses between genetically different strains of a species, have been used to identify causes since the time of Mendel.

The concept of complementation is derived from the use of genetics to define the steps involved in development in flies.26 To construct an example, the wild-type phenotype of red eye color is dominant and depends on the genes \(a\) and \(b\) in *Drosophila melanogaster*. Red eyes require that both genes are functional. If both copies of \(a\) and \(b\), or of both are defective, the eyes are white. In other words, \(a\) and \(b\) are complementary. Because the fly is diploid, red eyes will occur as long as at least 1 copy of \(a\) and 1 copy of \(b\) is functional. For convenience, the functional copies of the 2 genes will be referred to as \(A\) and \(B\); the nonfunctional forms are called \(a\) and \(b\). Loss of red eyes, that is, appearance of the white eye phenotype, depends on being homozygous for loss-of-function mutations in \(a\) or \(b\). For example, \(AA BB\), \(AA Bb\), \(Aa BB\), \(Aa Bb\) flies will all be red eyed. \(aa BB\), \(aa Bb\), \(aa bb\), \(AA bb\), will all be white-eyed.

Knowing this, the geneticist may perform a complementation test on 2 white-eyed fly strains. If the progeny all have red eyes then we know that the 2 strains complement, that they have mutations in 2 different genes responsible for wild-type eye color. In the above example, 1 strain must have the genotype \(aa BB\) and the other \(AA bb\), such that when the strains are crossed they yield progeny with the genotype \(AaBb\). In other words, the complementation test implies that some property of the gene causes the eye color. If the correlation is not perfect, we would likely ascribe the variation to experimental error or to some unknown cause.

Complementation tests can be used to identify very complex causal networks by connecting the gene pairs. The result is an epistatic network, a causal map of how one gene affects the functions of other genes. Because the genes found in the complementation assay can be identified using unbiased methods, for example, using saturation mutagenesis or large-scale RNAi knockdown screens, networks created in this way make no assumption about the molecular basis, that is, about the biochemistry of the molecules coded for by the genes. The only assumption made by the geneticist is his ability to define the set of locations, the genes in the genetic map. Using the fact of recombination of genes and the probability of recombination between genes, geneticists mapped these genes as entities separated along a line. That line, of course, is now DNA.

The epistatic network implies an underlying biochemical mechanism but tells us nothing about the functions of products encoded by the genes in the network. Put another way, this definition of gene refers only to a location, and the sequence at that location may not code for a protein. Indeed, some such genes identified in random mutagenesis screens do not code for proteins.

A very dramatic example of this process has been the work on the genetics of programmed cell death. A simplified, modern view of the cell death pathway is shown in Figure 2. Before 1990, efforts to explore the biochemistry of cell death were very confusing. Many of these experiments centered around hypothesis-derived epistemological arguments such as the point of no return.27 Biologists studying cell death argued that death must occur at some specific time and tried to determine the specific biochemical or metabolic changes occurring at the moment when a cell must inevitably die.

Because defining death was so difficult, investigators used morphological changes, called necrosis, that occur after a cell has died, as an end point.27 The investigators then performed experiments at earlier and earlier times to determine the identity of the common, critical event. For example, cells might be induced to die by anoxia; then various inhibitors would be tested at different times to see whether the proportion of cells reaching necrosis changed. There always seemed to be exceptions to the findings from each of these experiments. For example, although it may be obvious that enucleation is lethal, the fact that enucleate cells can survive for days made the nucleus as an end point problematic. Similarly, lysosomal...
release, mitochondrial disruption, and even loss of plasmalemmal integrity, all of which can obviously kill cells, were impossibly entangled in the effort to determine cause and effect.

This almost philosophical dilemma was elegantly resolved by the use of developmental genetics. In 1933, Morgan received the Nobel prize for the genetic map. Seven decades later, in 2002, the Nobel prize was awarded to Horvitz for his use of complementation to show that 4 genes, \(\text{egl-1}, \text{ced-3}, \text{ced-4},\) and \(\text{ced-9}\) determine cell death in the vulva of \textit{Caenorhabditis elegans} (\textit{C. elegans}). Generation of doubly mutant strains defined an epistatic network, that is, a causal map, shown in Figure 3.

The steps in Figure 3 show part of the logical process for deriving this causal diagram from a set of loss-of-function, recessive mutations. Loss of \(\text{ced-9}\) function results in an increase in cell death. This suggests that the wild-type function of \(\text{ced-9}\) is to inhibit death. In contrast, loss-of-function of \(\text{ced-4}, \text{ced-3},\) or \(\text{egl-1}\) all lead to a decrease in cell death. Again this suggests that these genes kill cells.

The complementation experiment leads to a different conclusion. Because loss of both \(\text{egl-1}\) and \(\text{ced-9}\) increase cell death, this indicates that the cell killing caused by loss of the protection conferred by \(\text{ced-9}\) does not require \(\text{egl-1}\) function. Thus, \(\text{ced-9}\) must act either in parallel to \(\text{egl-1}\) or downstream of the killing function of \(\text{egl-1}\) to protect against cell death. And if the latter, then \(\text{egl-1}\) must either directly or indirectly act as a negative regulator of \(\text{ced-9}\) function. In contrast, if loss of \(\text{egl-1}\) function killed cells independently of \(\text{ced-9}\) function then animals lacking both \(\text{egl-1}\) and \(\text{ced-9}\) function should show some effect of the \(\text{egl-1}\) mutation, that is, less programmed cell death, because the loss of \(\text{egl-1}\) function would
then have an effect regardless of the status of ced-9. Because this is not what happens, the genetics imply the hypothesis shown in C, that is, egl-1 is an inhibitor of ced-9. By the same process, epistasis analysis determines that normal ced-4 and ced-3 functions act downstream of or in parallel to ced-9 to cause cell killing.

Figure 4, adapted from Robert Horvitz’s published Nobel lecture, shows the complexity of the death pathways, as of that time, in the nematode C. elegans. The important lesson here is that this causal pathway was derived by Horvitz and his colleagues independently of traditional biochemical experiments.

**Beyond Hypothesis: Kuhn Versus Omic, Causal Networks**

The end result of using hypothesis-based experiments was predicted by another philosopher of science, the philosopher Thomas Kuhn. Kuhn pointed out that paradigms change because systematic, even boring experiments eventually fail. These failures drive scientific progress because they force us to develop new theories.

The ability to measure all the properties of a system, for example, omic data, is consistent with Kuhn’s view of how paradigms fail. If we believe we have an omic system and a model for how that system works, then experimental results should either be explainable by that model or require that we reconsider the rules that define the system.

**Relationship of Causal Pathways to the Pathway Posters on Laboratory Walls**

The models most familiar to biologists are the metabolic pathway posters on the walls of all biology laboratories. The posters need to be viewed like tapestries that decorated the walls of astronomers in the early 1800s. Those tapestries showed the 7 known planets. There was a good feeling that Newton had discovered how the universe worked. Our knowledge of the planets and the rules governing their motions worked just fine.

Unfortunately, as observations of the orbits improved, something was out of whack. Either Newton’s laws were wrong or there were other planets. The discovery of Pluto by Lowell and Tombaugh required a revision of the tapestry but preserved the laws of motion.

In comparison with finding new planets, finding new proteins is still not at all rare in biology. Usually these new proteins fit into our posters rather well. Problems with biological hypotheses rarely arise because our poster model of protein interactions is a lot more complex than Newton’s laws as seen on 19th century wall hangings.

The obvious problem with a biologist’s tapestry is the frequent discovery of unexpected interactions between the pathways. Returning to the death pathway, Horvitz’ diagram has remained valid. The biochemical, protease cascade model derived from genetics in C. elegans has, however, become interwoven with pathways that control cytokine activation, mitochondrial function, nuclear factor-xB, and mitogen-activated protein (MAP) kinases.29–34 The challenge now is to determine how all the embroidery added to Horvitz’ network can explain complicated phenotypes associated with the death pathway. Answering that challenge is likely to involve a more systematic approach than is possible by traditional studies of interactions among a few molecules.

A good example of the latter, unbiased approach is the synthetic genetic array analysis technique used to create networks in simple organisms, such as the fission yeast *Saccharomyces cerevisiae*.17 This method was developed to investigate genetic interactions between each possible pair of the ≈4700 nonessential genes of the yeast genome. In synthetic genetic array, 2 parental strains, each deleted for 1 known nonessential gene, are used to create recombinant double mutant progeny. Phenotypes of the double mutant colonies are compared with phenotypes of the individual single mutant colonies to identify those combinations of mutations that correct the phenotypic defect manifested by completely

![Figure 5. Synthentic gene array (SGA).](image)

**Table 1. Death Pathway in Worms and Animals**

<table>
<thead>
<tr>
<th>Worm</th>
<th>Mammal</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>EGL-1</td>
<td>BH3 domain</td>
<td>Contains the third domain of the 4 domains that define proteins in Bcl2 family, inhibits Bcl2 function, and directly promotes cell death.</td>
</tr>
<tr>
<td>CED-9</td>
<td>BCL2</td>
<td>Inhibits the ability of Apaf/CED-4 to activate caspase 3/CED-3</td>
</tr>
<tr>
<td>CED-4</td>
<td>Apaf1</td>
<td>Allosteric chaperone, activates caspases.</td>
</tr>
<tr>
<td>CED-3</td>
<td>Caspase 3</td>
<td>Executioner protease when active creates a feed forward loop of proteolytic events that kill cells.</td>
</tr>
</tbody>
</table>
defective or partially disabling mutations. Although this approach is based on a series of hypotheses, that is, testing the possible gene interactions between specific pairs, one can also claim that synthetic genetic array is omic because of the existence of genome-wide deletion libraries that specifically eliminate the function of all target genes.

As long as the phenotypes are fairly simple, automated methods available for handling yeast strains enable massively parallel experiments permitting a systematic analysis of all possible pairwise combinations. Dixon et al.\(^1\) have recently reviewed these and other systematic approaches to the detection of pairwise genetic interactions, for example, RNAi, in species as complex as the nematode worm \textit{C. elegans}. The result, as shown in Figure 5 is very much like a metabolic map except that causality, that is direction, is missing from the lines connecting gene pairs. Of course, we may have other information that allows us to annotate these lines. For example, we may know that one gene is an enzyme and another is a substrate.

The central goal of this review is to explain the logic behind the magic. We will use traditional genetics to define cause and use the discovery of the death pathway (Figure 2) to suggest that network analysis is central to solving the problems of multifactorial cardiovascular disease. Finally, we will use a familiar problem, interactions among cytokines, as an example of the opportunities for use of network analysis in our own field.

In summary, it is hard to imagine this elegant pathway emerging from the preceding, from the cumbersome hypothesis-based effort to control cell death. Once the causal network was defined by Horvitz and his colleagues, the traditional tools of hypothesis-based research identified the biochemical basis for each step. The resulting death pathway, shown in Table 1, is now a fundamental part of our understanding of biology.

Using this knowledge of the underlying biochemical mechanisms for a part of the network may allow us to derive the directions and posit likely biochemical mechanisms for the entire network. This is exactly the approach used by Zhu et al.\(^1\) to construct networks for yeast, except that they used the principle that any network that can be mapped to a genetic locus must have a causal origin at that locus because genetic variation is independent of all other functional properties of a gene. We will discuss this later under the concept of intrinsic variation of a variable.

Limitations of Networks Based on Hypothesis Data or on Complementation Test

Although these genetic experiments can define causal pathways, they do not necessarily provide quantitative models. Even in yeast, efforts to model something as simple as growth rate based on complementation maps have not been very successful.\(^1\) Extending this approach to the far more complex phenotypes represented by mammals seems unlikely.

Part of the problem with both metabolic maps and complementation pathways is turning causal relations into mathematical relationships between variables. Consider, for example, a model that assumes that a receptor shows the usual S-shaped response to an agonist. Although this may be correct for many receptors, we know that many receptors have coreceptors and, as a result the usual assumption of an S-shaped curve may need to be replaced by overlapping curves that have different signaling responses dependent on agonist and antagonist concentrations. Other unknown variables may also determine the abundance of components or activities represented in a signaling complex. A good example from the recent literature are the diverse outcomes of binding the type

---

**Figure 6.** A network. Is this a transcription network, or a spider web? This is a portion of a transcriptional network based on data from 3 mouse strains, B6, C3H, and CAST. The genetic variability of the strains, as input variables, has been used to determine the causal interactions between the same genes as transcription products (based on Yang et al\(^8\)).
1 interferon receptor in response to different type 1 ligands or Jak and Stat proteins.\textsuperscript{15}

In the following sections we will briefly introduce an approach that produces mathematical relationships even in complex pathways. This will begin with a discussion of an early use of causal analysis to predict the production of pigs in World War II.

**Sewall Wright: From Pig Farms to Genomes**

How can pigs be relevant to this discussion?

Graphical methods are not new. During World War II, Sewall Wright proposed to use this method to identify the best locations for pig farms so that the United States could optimize production of pigs.\textsuperscript{13} The problem is complex because the choice of where to grow pigs is affected by local food supplies and by the availability of railroads to deliver ham products to the war effort. Wright argued that a causal diagram based on existing data for farms in different locations could predict the optimal network to deliver ham to hungry US soldiers.

The common variable in Wright’s analysis was production of pigs. Perhaps unfortunately, for the efficient use of pigs and the quickness of our victory in Europe, objections from the traditionalist statisticians prevented his model from being used.

Wright used a form of graph called a directed graph. Each node in the directed graph has a value, for example, the number of pigs on a particular farm or the abundance of an mRNA transcribed by a particular gene. The nodes are connected by lines (also called edges) representing all the possible correlations between these values.

To simplify the math, most networks use directed acyclic graphs. Directed acyclic graphs simplify graph construction by side stepping feedback loops and assuming the graph is an instantaneous view of a system.

**Diagrams, Graphic Analyses, and Structural Equations**

In a metabolic map, $A \rightarrow B$ means that $A$ acts on $B$ but says little about the strength of the interaction or about the probability of the interaction’s occurring. No single equation can represent the interactions because $A \rightarrow B$, $C \rightarrow D$, $E \rightarrow F$, … $Y \rightarrow Z$, $Z \rightarrow A$ represent a large number of very different chemical processes and each letter may represent a very different type of molecule.

In a graphic analysis of expression data, $A \rightarrow B$ means that variations in abundance of $A$ correlate with variations in abundance of $B$. The length of the line $AB$ is the level of correlation between the 2 values. Further, because the arrow goes from $A$ to $B$, the line $AB$ asserts that the variation in $A$ somehow causes the variation in $B$ but $B$ does not cause $A$. Lines like $A \rightarrow B$, with or without the direction are called edges. These edges can be used to construct very large diagrams as in Figure 6. Unlike a metabolic map, a graphical network can be translated to a system of structural equations. It is important to understand that either $A-B$ or $A \rightarrow B$ implies an absolute correlation. Correlation can be partial. For example, the relationship $A \rightarrow B-C \rightarrow C$ means that $A$ and $C$ are both partially correlated with, and causal to $B$. The equations that describe these edges are called structural equations. Structural equations define the partial correlation between variables. In a map based on expression data, the equations describe the edges in terms of the effect that one quantity of one mRNA has on the amount of the other mRNA. Of course, depending on how the map is built, it may be causal or just correlative.

The equations taken together constitute a system of equations. Solving these systems of equations is an extension of the tricks we all learned in high school to solve equations with 2, 3, or >3 variables. Solutions of the equations constructed from omic data should reveal unexpected causal connections even between molecules where interactions have not yet been studied.

Such unexpected causal connections become obvious targets for hypothesis-based research. As we will see, experiments can be set up to explore systems we already understand. For example, we will show how current knowledge of the properties of promoters and transcription factors can be used to discover all of the pathways controlling transcription in yeast.

**Derivation of Structural Equations from Graphic Analysis: Need to Choose a Shared Attribute**

The mathematical derivation of equations to construct networks is beyond this review. However, Greenland et al’s\textsuperscript{36} review of a directed acyclic graph as it affects biological systems provides a good introduction. The diagram here, based on one used by Greenland et al,\textsuperscript{36} represents an analysis of the causes of asthmatic attacks in children.

In their diagram (left column in Table 2), $A$ represents air pollution, $B$ gender, $C$ bronchial reactivity, $E$ antihistamine, and $D$ is asthma. The model asserts that $A$, $B$, $C$, and $D$ are a complete explanation for whether a child does or does not get asthma.

Using the same diagram, we have constructed a hypothetical model for death in a mammalian cell system. The diagram is roughly based in the known apoptotic pathway controlled by the tumor necrosis factor $\alpha$ receptor, a death receptor.

### Table 2. Constructing A DAG to Define a Causal Network

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>B</td>
<td>A</td>
</tr>
<tr>
<td>C</td>
<td>D</td>
<td>E</td>
</tr>
<tr>
<td>E</td>
<td>D</td>
<td>C</td>
</tr>
</tbody>
</table>

- **A** Air pollution
- **B** Death receptor activity
- **C** Mitochondrial injury
- **D** Caspase 8 activity
- **E** Caspase 3 activity
- **Antihistamine administration** Transcription of A1, inhibitor of Caspase 3

DAG indicates directed acyclic graph.
pathway that can both stimulate and block cell death. Details of the death receptor are less important than the comparison with the model offered for asthma. In both cases the ability of the model to predict the results of experiments will depend on how accurately this model accounts for all the variables leading up to the phenotype.

Note here that like a metabolic map, this diagram connects different types of nodes, however each node has only 1 property and the edge represents the partial correlation between the 2 nodes. Metabolic maps can be shown as causal diagrams if these rules are followed and the quantitative attributes of each node are understood.

It is important that these models be unbiased. For example, looking at the asthma model, there is no edge connecting A and B … air pollution and gender. Greenlee et al do not show a line because they assume that there is no correlation between the two. Of course, this might not be true. For example, births in China today are disproportionately male because of the combination of traditional preferences for male children, the one child rule, and China’s massive effort at industrialization. Because air pollution (A) is dependent on industrial development, A → B may be important, at least in China.

The graphical analysis models depend on our ability to construct an unbiased diagram for any 2 variables, A and B, by asserting that there are only 3 possibilities for correlations between the variables. These are variability in B may be determined by internal variation in B, by the variability in A, or by variations that are external to our diagram. Depending on the math we wish to use to fit these correlations, we can describe all of these by equations. For example, it may be reasonable to propose that air pollution and bronchial reactivity are related by a simple linear equation. The same is true for relating activation of a death receptor to the level of proteolytic activation of caspase 8.

A graphical diagram can be converted to a large set of equations. Computer software now exists that can write sets of structural equations for very large graphic analyses. These models can then be validated by testing their ability to account for the values of all the variables in different data sets.

Systems, That Is, Causal Networks of Structural Equations, Violate the Traditional Use of the Null Hypothesis

Obviously, if we have large data sets we can create a model using graphical analysis and test the model by determining how well it fits the real data. A good fit means our model does define, at least for this data set, the causal relationships between variables. The process of finding that model is very different from the hypothesis-based experiment. Rather than accepting a model as real when there is <0.05 probability that a result could have occurred by chance, structural approaches accept a model as real when there is a minimal difference between the ability of one model to explain variation among data sets and the ability of any other model to explain that variation. Computers allow us to make recursive calculations of the system of equations and select the equations that converge on the real data.

A Simple Biological Example of Graphic Analysis

Figure 7 is an example of a very simple directed acyclic graph. This graph shows a system with variables we can think of as measures of some common property of 4 proteins. Typically, this might be abundance of the mRNA but it is worth considering omic measures, such as protein levels. The idea that these diagrams measure a common property is not as obvious as it may seem. For example, because the diagrams in Table 2 each depend on a small number of variables, the variable being measured may be different for each node. In contrast, an analysis with large numbers of variables will be easiest to construct if we restrict ourselves to some single property that is common to all nodes. In constructing the diagram it is essential that we agree that it is reasonable to imagine that the value for the property of A affects the same type of value measured in assessing the state of B.

Activity, that is a measure of the function of each protein, may be the most interesting property connecting genes. Unfortunately, omic measures of activity are complicated by the problem of normalizing activity measures for the very different functions of different proteins. As a result, networks are usually constructed from some common property, such as mRNA or protein abundance.

Of course, having levels of mRNA for 20,000 genes does not tell us which pairs of genes might be closely connected in a causal network. Typically, network analyses begin with guesses about expected correlation in mRNA levels for proteins with shared activities. These guess are informed hunches. Similar informed hunches may explain the successes of expert users of the Ouija board, the more knowledge the Ouija expert has of her subject, the more likely is that here answers will be convincing. For example, many networks based on mRNA levels assume connections derived from a metabolic map without questioning how reasonable it may be to expect that an enzyme and its substrate will be coregulated at the RNA level.

The importance of choosing a common property is shown in Figure 7. In this diagram, protein A is a kinase that directly phosphorylates and thereby activates a transcription factor, B. Activity of B in turn determines the transcription of C, an enzyme that cleaves a substrate protein (not shown). The cleavage releases a peptide that binds to and activates the enzyme D. Combining these 4 very different proteins to analyze a set of omic data are challenging. The most common property used in constructing networks is abundance of the mRNAs coding for the proteins. In this

**Figure 7.** A causal chain. In this diagram, A is a kinase that acts on a transcription factor B. B controls transcription of C. C is a protease that determines the activity of D, an enzyme that itself controls glycolysis. To turn this diagram into a causal pathway we need to specify the property changing at A, B, C, and D. In most network analyses, however, the only property we measure is one common to all the nodes, for example, the abundance of mRNAs coding for A, B, C, and D. Is it reasonable to go from the metabolic diagram to a causal diagram based on the message level for each protein?
diagram, we are assuming that the amount of A kinase will somehow determine the amount of B transcription factor. That assumption may not be correct even if the kinase is specific for B.

We will return to the use of diverse measures of activity when we discuss construction of diagrams and the use of modules, that is, clusters of genes, defined by function.

**Cis-Expression (as a) Quantitative Trait Locus**

The best-known example of causal analysis with large data sets is the use of transcript profiles to identify sequences that control the biochemical process of transcription. Particularly impressive is the identification of cis-acting quantitative trait loci, cis-eQTLs. In biochemical terms, finding a cis-eQTL means we have identified a cause without needing to do an experiment. At least in yeast, such cis-eQTLs identify causal mutations within 2kb 5′ or 3′ of the affected transcript >95% of the time. This is an independent verification of the huge amount of experimental evidence showing that mRNA abundance is controlled largely by promoter sequences and 3′ sequences contained in the messenger RNA. Similar findings are likely to be true in mammals because we now know so much about how cis sequences control expression.

The theory of identifying a cis-eQTL in inbred species is simple. One maps the genomic sequences of a genetically diverse collection of individuals or identifies frequent single nucleotide polymorphisms (SNPs) that differentiate individuals or strains of the species. These frequent markers identify patterns of recombination and imply that a nearby sequence is causal for any difference in expression. If the data come from recombinant inbred strains, then the only variations are in the SNPs. If we have sufficient crosses, we should be able to identify the sequence variation causing the variation in expression. Although finding funds for making sufficient crosses is expensive, as discussed in the last paragraph, it is very likely that we can identify likely causal sequences. We can then test this hypothesis by biochemical approaches.

Unlike inbred animals used in experimental genetics, humans are rarely inbred. This means that human alleles tend to be heterozygous, and sequences between human SNPs are very variable. As a result, finding a human cis-eQTL is less informative than identifying a cis-eQTL in an inbred animal. Nonetheless, we will finish this review with an example of the recent use of human cis-eQTL to identify pathways controlling complex human disease phenotypes.

**Adding Time to Causal Studies**

The discovery of the death pathway is a wonderful example of how genetic information can be used to determine biochemical causality. The cis-eQTL just discussed comprises another example of how we can identify causes even without doing temporal experiments.

Nonetheless, time is implicit in the common definition of cause. Temporal experiments can allow us to distinguish causal relations from simple correlations. This is especially important when analyzing complex systems. The notation A→B means that B happens after A. The logic of the experiment implies that if we mutate a murine cis-eQTL sequence, we will see a change in expression in the resulting cells. This is called temporal causality.

As an example of the value of temporal studies, imagine that we have identified 5 SNPs in inbred mice. These SNPs distinguish 2 strains that differ in expression of a collection of mRNAs, a module as defined in the discussion of Dilemma 1 above. None of the SNPs is a cis-eQTL. The challenge is to determine how these 5 loci interact, that is, to define a causal pathway.

Although we can imagine a complementation experiment like the one Horvitz used for the death pathway, the expression phenotype, that is the module, may fall apart if we breed each knockout to each of the other knockouts. As illustrated in Figure 1, the rules of network analysis tell us that every knockout can change the entire structure of a network.

As an alternative we could design a time series experiment using an agent that we know alters function of a specific node in the putative pathway. After applying this agent, we can look for changes in the properties of the other nodes over time. The sequence of changes over time can be used to add direction to the edges connecting nodes. In other words, the temporal data allow us to define a causal sequence.

These time series analyses may be thought of as a series of graphical diagrams that change over time. Although mathematics for analysis of time series is beyond this review, assuming temporal causality simplifies the development of graphical models because the network analysis can avoid considering feedback loops by assuming we are looking at causality in a very thin slice of time.

**Role of Noncoding Transcripts in Graphic Analysis of Biological Pathways**

The discussion so far has focused on parts of the genome that code for proteins. The problem of considering all possible causal pathways is apparent when we realize that a single expression array includes >20,000 expression values, 1 for each gene. The analysis will become much more difficult as we include other sequences, because we do not yet have enough knowledge to believe we can define the complete system of functional, noncoding RNA species.

**Rules for Constructing a Causal Diagram**

The next question is how do we begin to arrange the variables in a causal diagram, like the one in Figure 7 or the more complicated pathways in Figure 8? Although we might start with

![Figure 8](https://example.com/figure8.png)
the metabolic map, we want to rearrange the diagram to fit with our actual data.

This requires a set of logical rules. For example:

- Rule 1. ABC implies that A is indirectly causal for C, conditional on B. However, if B is fixed, that is if it is rendered unresponsive to the activity of A (the most obvious example being when B is rendered completely nonfunctional, as in a knockout experiment), then A is no longer causal for C. The challenge for applying this to biology is that fixing a variable, for example, by knocking it out or by overexpressing it, is the usual method of hypothesis-based research. Because biological reactions depend on mass action, however, the conclusions from a KO or overexpression may be misleading and lead to an incorrect causal diagram as in Figure 1.

- Rule 2. Path A→B←C is called a collider. A→B←C is a very different sort of causal path than A→B→C. The rules of graphic analysis tell us that fixing B, for example, by only looking at data where B has a specific value, makes the value of A and the value of C interdependent. That is, if we look at different values for A, we will be able to predict the value of C as if A were causing C even though we know this is not true. Although counter intuitive, the idea of a collider actually is what would be expected from biology. For example, imagine that A and C are both cofactors that activate B. If we know that B is activated and observe that A is OFF then we also know that C must be ON, thus there is a correlation between A and C although they are NOT causally connected.

- Rule 3. If A and B are correlated but all the variation in A is intrinsic to A, then AB, that is, A must be the cause of B. Note that this is actually using the same underlying reasoning as in genetic analyses already discussed. If the variability in A is due to sequence variation in the gene coding for A, and the values of A and B are related, variation in B must be caused by the molecules coded for by A. As we will see, Rule 3 is the basis for rooting of causal diagrams in the analyses of transcription.

Given recent studies attributing disease phenotypes to rare mutations occurring frequently in individual genes, Rule 3 suggests that we need statistical methods for comparing sequence variation with expression data.

When applied to expression data, at the end we have a diagram showing how mRNA levels of each gene’s product cause each other. How useful would this be in understanding a problem in systems biology? Remember that A→B→C→D is the causal diagram established by unbiased genetic reconstruction of the cell death pathway in the vulva of C. elegans: the state of the first gene, egl-1, informs the state of the second gene, ced-9, which then determines the state of the third gene, ced-4, which in turn controls the state of the fourth gene, ced-3, which will, if activated, kill the cell. Figure 2 is certainly causal, but this diagram would not likely be found from transcription levels.

There is one more rule to consider. The goal of structural equation modeling is to construct a set of equations best able to account for the covariation of all 20,000 different genes. This last sentence however, never states what is varying. It is important to note that the ONLY information in most network analyses is derived from transcription levels. This can be restated as a fourth rule:

- Rule 4. Causal analysis based in transcriptomes only identifies the abundance of the mRNA of 1 gene as a possible cause of the abundance of other mRNAs.

Rule 4 does not mean that transcription networks do not lead to hypotheses about other biochemical events. For example, using the relationships in Figure 7, biochemists may want to use hypothesis-based studies to determine how C determines the level of D mRNA. Perhaps, C prevents degradation of the D mRNA or phosphorylates some other gene that inhibits the D promoter? An entire new pathway may be discovered by such analysis.

**Experimental Design in Terms of Input Data and Output Data**

We are not suggesting that investigators learn how to construct causal diagrams. However, we do suggest that designing experiments that test a hypothesis with multiple, interacting causes requires understanding and using the Bayesian concept of priors to design the experiment. To simplify this we propose thinking of experimental design in terms of input data and output data.

We define input data as information that is known and intrinsic to each animal, cell or other experimental system before an experiment begins. The term intrinsic means data limited to differences in the properties of the animals, cells, or other experimental systems independently of any manipulation. In genetic experiments with animals, inbred strains allow us to limit the input data to differences between strains or to experimental manipulations.

Output data consist of properties of these same systems after any manipulation or as a result of differences in the data intrinsic to each of the experimental systems.

As long as the input data are omic, we can conclude that differences among the experimental groups, that is, differences in the output data, are caused by differences in the input data or by experimental manipulations.

Examples of input data might include the genetic sequence of different strains, epigenetic differences existing before the experimental manipulation, or mixtures of cytokines applied to otherwise identical members of an inbred mouse strain. Causality is implied because we claim to know all the relevant output data, and we assert that the only differences between experimental groups are either intrinsic to the groups or the result of defined experimental manipulations. Network methods allow us to determine the extent to which each of the input variables correlates with the output data.

**Performing a Causal Experiment Without a Network, the Value of Being Biased**

In the real world of biology, input data are always biased. For example, a QTL study of mouse strains is always biased by the breeding history of the mouse strains available to the
investigator. Mouse or yeast geneticists can estimate the bias because sequence differences between strains are a known part of an omic, input data set. Geneticists can determine multiple causes as long as the investigators choose to restrict the output data to phenotypes likely to be explained by specific sequences and define that bias. The Bayesian term for defining a bias is defining priors.

As long as the bias is defined precisely, then conclusions are valid. A good example comes from the systematic identification of sequences that determine expression phenotypes in yeast. Dr Lee, one of the authors of this article, began her work by stating that we know enough now about transcription to identify a very large part of the loci where variability is likely to cause changes in gene expression. Such loci include 5′ promoter regions, introns, 3′ tails that interact with interfering RNAs, and the known transcription factors. This set of genetic loci comprises a rule for expression comparable with the role of Newton’s laws in explaining the orbits of the planets.

This approach does not use diagrams because the priors are built into a hypothesis about the relative role of different components of the input data. A set of equations can be set up using guesses, that is, probability coefficients that estimate how strong an effect each of these input variables will have on the expression modules. Analyses based on different sets of probability coefficients will produce empirical modules biased only by our guesses. Computer software can do this repeatedly to identify the set of variables, called priors, that produces an optimum set of modules.

Lee et al used SNP maps of the yeast strains as input data. To simplify the analysis, they focused on loci selected as likely to control mRNA expression. A structural equation was set up with initial estimates of the possible causality of each locus. The power of the analysis was greatly enhanced by a computer program that identified sets of genes, that is, modules coregulated across the yeast genomes. The software performed recursive changes in the structural equations to determine a best estimate of the partial correlation of variation in the sequence of each gene locus with variation of the expression modules. Limitation of the input data to sites likely to control expression allowed the analysis to be performed rapidly with minimal computer resources. The result was the identification of specific SNPs that account for heritable changes in expression of each module in yeast.

The result is causal, that is, Lee et al empirically defined the set of loci that genetically determine expression patterns in yeast. As one might hope, not all the causal loci were expected. For example, Lee et al uncovered a novel, experimentally validated connection between Puf3—a sequence-specific RNA-binding protein—and P-bodies—cytoplasmic structures that regulate translation and RNA stability, as well as the particular causative polymorphism, an SNP in Mkt1, that induces the variation in the pathway.

We encourage readers skeptical of systems biology articles to realize that the approach identified the genetic controls of transcription without requiring a network diagram!

---

**Input Variables as Network Roots for Metabolic Maps**

Our desire to avoid reliance on the sort of biased and poorly informed models we have analogized to Ouija boards should not obscure the value of causal analyses to identify and detail networks beyond the reach of traditional approaches used to piece together metabolic maps.

As discussed, any causal analysis requires that variation in input data is intrinsic, that is, independent of causes external to the data set itself. Data like these are sometimes termed rooted.

Rooted data can be combined with metabolic maps or other known biological properties to construct networks that identify molecules or genes in pathways that extend beyond data available from genetic variation alone. These networks can produce novel hypotheses about pathways that account for multifactorial diseases, including coronary artery disease.

Although natural genetic sequence variation among individuals or strains is obviously rooted, there are other variables that meet this criterion. Another example of input data may be variables in a large-scale, carefully controlled experiment. Although determined by the investigator, these variations and the control conditions can be designed to minimize unknown sources of bias. Assuming we create adequate such variations, the variation in the input data is as intrinsic as naturally occurring mutations. As an example, one can imagine an in vitro experiment where the input variables are varied concentrations of multiple cytokines. The mathematical approach is the same for any set of data that meet the criteria discussed earlier.

Although Dr Lee’s systematic analysis of sequences causing transcription in yeast has not as yet been performed in mice or humans, the possible advantages of having a systematic causal map of sequences controlling gene expression are immense. As we will discuss later, a recent study in man has identified modules that are associated with multifactorial diseases and seem to be controlled by specific loci like those Lee identified in yeast.

**Application to Cardiovascular Phenotypes: Deriving a Causal Map From Cytokines**

If you have gotten this far, we hope we have offered a few simple ideas about how these causal approaches may be used to study diseases that are multifactorial. We will finish this exercise by proposing to define the role of cytokines in causing cardiovascular disease phenotypes. Examples of relevant disease phenotypes include atherosclerosis, the autoimmune vasculitides, hypertension, cardiac failure, fibrosis, diabetic vascular disease, and tumor angiogenesis.

Even in genome-wide association study (GWAS), loci may not identify any nearby locus with an obvious connection to known causes of disease. However, disease QTLs that localize on or near QTLs that we now know are causal for an expression modules become of great interest. If the locus causes the disease, then it is reasonable to expect that the mechanism of the disease may be found in the expression pathways controlled by the same locus.
The cytokine problem in vascular disease is ideal because we assume these diseases are multifactorial and often inflammation is implicated among these causes.

To design the study we first need to choose a source of data that fit the criteria for input variables. The most obvious input data come from DNA variations naturally occurring between mouse strains with known propensities to develop cardiovascular disease. Strains with defined genomes exist for atherosclerosis, myocardial failure, hypertension, autoimmune vasculopathies, and diabetes mellitus. By analogy to the yeast experiments discussed earlier, 7, 8 we can bias our input data by looking for genetic variation at sites likely to control expression of cytokines.

Although we could use complete expressomes as output data, we propose to simplify the study by focusing on 2 highly quantitative measures of cytokine expression as output data. mRNA level by quantitative real-time polymerase chain reaction and cytokine protein abundance by bead-based immunoassays are much more precise than the usual measures of gene expression by chip-based assays. Polymerase chain reaction has the advantage over hybridization analysis of being highly quantitative. Furthermore, relatively inexpensive assays already exist for the mRNAs encoding cytokines. Confining our proteomics to immunoassays of cytokines eliminates the usual problems of not knowing the full set of proteins. Finally, to further simplify the experiment, we propose to purify monocyte subsets and eliminates the cellular composition of tissues as a hidden variable in our analysis. This suggestion is not an idle one. A similar experiment using human data has recently been reported by Rotival et al. 9 The human mRNA was obtained from monocytes of 1490 Europeans and modules were examined for association with 675 350 SNPs.

Rotival et al. 10 identified 3 genomic regions associated with coregulated gene modules. One locus was correlated with type 1 diabetes locus, the second with type 1 diabetes mellitus, hypertension, and celiac disease. The third locus was related to cancer. The authors found several cis- and trans-eQTLs in each of these regions.

The Rotival study is not, strictly speaking, causal. Because the studies were done in noninbred humans the intervening sequences in haplotypes near the SNPs are not known. Obviously, we cannot be sure that the SNPs are not located close to other sequences that are the actual causes of coexpression. What Rotival et al. 10 do point out, however, is that a network analysis can point us to sites where interactions among multiple genetic sites are causal for an expression phenotype that is associated with disease.

If similar studies were done using inbred mice, the causal loci for expression could be identified by using recombinant inbred strains. What is unclear is how we would test the hypothesis that the loci that are causal for the expression modules are also causal for the disease. Traditional knockouts may not help because, as we have discussed, each knockout, or similar overexpression study, will affect the entire network. Using genetic engineering methods to manipulate all the loci that contribute to the expected phenotype at the same time is impractical. However, that experiment might be done by a breeding protocol where the putative disease loci are enriched by appropriate crosses. Such experiments are not unlike traditional efforts of breeding for a disease phenotype, for example the development of the spontaneously hypertensive rat. 11 Alternatively, if one has a database for expression responses to drugs, one might ask whether drugs with a similar effect on the transcriptome also affected the disease phenotype. 10

Summary

Our take home message for cardiovascular research has 2 parts. First, because most cardiovascular diseases are multifactorial, biologists who see network analysis only as fishing expeditions are not likely to discover the mechanisms for these diseases.

Second, we have tried to suggest simple criteria for the design of experiments that use network approaches to identify causes of disease. Although network approaches using genetic data are underway, other experiments, especially large-scale experiments done under controlled conditions, can meet the same criteria.

Sources of Funding

This work was supported by the following grants: NIH T32 HL07312, Experimental Pathology of Cardiovascular Disease; NIH PO1 HL03174, Mechanisms of Acute Vascular Reaction to Injury.

Disclosures

None.

References


A Systematic Approach to Multifactorial Cardiovascular Disease: Causal Analysis
Stephen M. Schwartz, Hillel T. Schwartz, Steven Horvath, Eric Schadt and Su-In Lee

Arterioscler Thromb Vasc Biol. 2012;32:2821-2835; originally published online October 18, 2012;
doi: 10.1161/ATVBAHA.112.300123
Arteriosclerosis, Thrombosis, and Vascular Biology is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231
Copyright © 2012 American Heart Association, Inc. All rights reserved.
Print ISSN: 1079-5642. Online ISSN: 1524-4636

The online version of this article, along with updated information and services, is located on the World Wide Web at:
http://atvb.ahajournals.org/content/32/12/2821

Permissions: Requests for permissions to reproduce figures, tables, or portions of articles originally published in Arteriosclerosis, Thrombosis, and Vascular Biology can be obtained via RightsLink, a service of the Copyright Clearance Center, not the Editorial Office. Once the online version of the published article for which permission is being requested is located, click Request Permissions in the middle column of the Web page under Services. Further information about this process is available in the Permissions and Rights Question and Answer document.

Reprints: Information about reprints can be found online at:
http://www.lww.com/reprints

Subscriptions: Information about subscribing to Arteriosclerosis, Thrombosis, and Vascular Biology is online at:
http://atvb.ahajournals.org//subscriptions/