Emerging Science Reviews are published on an occasional basis to highlight areas of research that are very recent and at the cutting edge of cardiovascular biology. The goal of these articles is to bring attention to promising new topics that are not yet well developed.

“Good Enough Solutions” and the Genetics of Complex Diseases

James N. Weiss, Alain Karma, W. Robb MacLellan, Mario Deng, Christoph D. Rau, Colin M. Rees, Jessica Wang, Nicholas Wisniewski, Eleazar Eskin, Steve Horvath, Zhilin Qu, Yibin Wang, Aldons J. Lusis

Abstract: In this Emerging Science Review, we discuss a systems genetics strategy, which we call gene module association study (GMAS), as a novel approach complementing genome-wide association studies (GWAS), to understand complex diseases by focusing on how genes work together in groups rather than singly. The first step is to characterize phenotypic differences among a genetically diverse population. The second step is to use gene expression microarray (or other high-throughput) data from the population to construct gene coexpression networks. Coexpression analysis typically groups 20 000 genes into 20 to 30 modules containing tens to hundreds of genes, whose aggregate behavior can be represented by the module’s “eigengene.” The third step is to correlate expression patterns with phenotype, as in GWAS, only applied to eigengenes instead of single nucleotide polymorphisms. The goal of the GMAS approach is to identify groups of coregulated genes that explain complex traits from a systems perspective. From an evolutionary standpoint, we hypothesize that variability in eigengene patterns reflects the “good enough solution” concept, that biological systems are sufficiently complex so that many possible combinations of the same elements (in this case eigengenes) can produce an equivalent output, that is, a “good enough solution” to accomplish normal biological functions. However, when faced with environmental stresses, some “good enough solutions” adapt better than others, explaining individual variability to disease and drug susceptibility. If validated, GMAS may imply that common polygenic diseases are related as much to group interactions between normal genes, as to multiple gene mutations. (Circ Res. 2012;111:493-504.)

Key Words: systems genetics | genetics of complex diseases | scale-free networks | hybrid mouse diversity panel | computational biology

The Human Genome Project completed in 2001 is one of the seminal achievements in modern science and ushered in an era of unbridled optimism about our prospects for deciphering the genetic basis of human diseases, and beyond.¹ A decade later, more than a thousand genetic loci conferring increased risk of a wide variety of diseases have been identified, successfully revealing new genes and novel pathways not previously suspected to be involved.² Although these findings are a major advance, associations of single nucleotide polymorphisms (SNPs) with common polygenic diseases, have, in some respects, been disappointing³: generally, SNPs typically confer only a modestly increased risk (average, 1.3-fold) and explain only a small fraction of the genetic component (<20%). Moreover, since the loci are identified out of context, elucidating the role such genes play in disease can take years (eg, the role of...
accounts for a small fraction of the disease risk, astronomically large samples are required to identify these variants. Straightforward combinatorial mathematics illustrates the problem. To identify one SNP out of a million associated with a disease is already looking for a needle in a haystack. However, the number of ways in which a million SNPs can be combined in pairs is approximately 500 billion; for triplets, this number mushrooms to $1.7 \times 10^{12}$. For 16 SNP combinations, the number of possible combinations reaches $10^{92}$, greater than the estimated number of atoms in the observable universe ($10^{80}$)!

As an example, GWAS performed to identify the genetic basis of coronary heart disease involved more than 100,000 individuals, and still these studies accounted for only about 4% variance of the trait. A discussion of the lack of power of GWAS for identifying genetic interactions is presented in Zuk et al.\(^4\)

If the bottom-up GWAS approach of looking at SNP (or gene) combinations to explain complex disease susceptibility has serious limitations, an alternative is to develop a top-down approach. In analyzing a business failure, this would entail a multistep process: first, group (cluster) employees into divisions (modules); second, relate the divisions to performance measures; third, identify underperforming employees within poorly performing divisions. In this article, we provide a perspective illustrating how we might take advantage of new “–omic” technologies to view how genes work in groups, rather than as single agents, and use this information to develop an analogous top-down strategy for elucidating the genetic basis of complex polygenic diseases. In the first section, we set the stage by borrowing the “good enough solution” concept from the field of evolutionary biology.\(^5,6\) This concept is based on the idea that in complex systems, many different combinations of the system’s parameters can produce a nearly identical output (ie, “good enough solutions” that meet the required output specifications). Thus, a wide range of individual gene expression patterns may all be perfectly adequate for normal function, but have different abilities to adapt to an environmental stress, accounting for differential susceptibility to common diseases and drug reactions. This raises the intriguing hypothesis that susceptibility to common diseases and drug reactions may be more related to the way in which normal genes interact with each other than to additive effects of multiple gene mutations. Following on this rationale, we then describe how network theory can be applied to gain insight into biological systems. We explain how a randomly-interacting, growing modular network facing selection pressures naturally evolves toward a scale-free network topology conferring robustness, adaptability, and efficiency, through small-world properties, to the system. We then turn to the gene networks specifically and discuss how systems genetics approaches can be used to discover how genes are naturally grouped into modules, each containing tens to hundreds of genes, linked in an approximately scale-free modular network. We explain how the functions of these modules can often be discerned from known genes (eg, apoptosis genes, metabolic genes, cell cycle genes) and cell markers in which they are enriched. We then illustrate how novel system genetics resources, such as the hybrid mouse diversity panel (HMDP), can be used to test the

### Non-standard Abbreviations and Acronyms

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>EC coupling</td>
<td>excitation-contraction coupling</td>
</tr>
<tr>
<td>(G_{Kr})</td>
<td>rapid component of the delayed rectifier K channel conductance</td>
</tr>
<tr>
<td>GMAS</td>
<td>gene module association study</td>
</tr>
<tr>
<td>GWAS</td>
<td>genome-wide association study</td>
</tr>
<tr>
<td>HMDP</td>
<td>hybrid mouse diversity panel</td>
</tr>
<tr>
<td>SNP</td>
<td>single nucleotide polymorphism</td>
</tr>
<tr>
<td>WGCNA</td>
<td>weighted gene co-expression network analysis</td>
</tr>
</tbody>
</table>

apolipoprotein E in Alzheimer disease). Even for single gene disorders (such as Huntington disease), a mechanistic understanding of the effects of the genetic variation remains challenging. For common complex diseases such as hypertension, diabetes, atherosclerosis, heart failure, cancer, and Alzheimer disease, SNPs associated with increased risk have small effect sizes and only account for a small fraction of the genetic risk.\(^3\) Thus, the majority of patients at risk for developing common diseases are unlikely to be identified by SNPs.

This is not really surprising, given that most common diseases are polygenic, caused by modest effects of multiple genes interacting with environmental factors. Hundreds or thousands of loci can influence common polygenic diseases and many of these loci have individual effect sizes too weak to be identified individually by genome-wide association studies (GWAS). Since GWAS focus on identifying SNPs that for the most part covary with single gene mutations, it is not surprising that the genetic basis of common diseases remains unsolved for the bulk of affected patients.

To break this impasse will require moving beyond the single gene perspective to view how groups of genes work together. A worthwhile analogy may be the similarity between a human organism and a business organization. Both have modular designs: a human is composed of physiological modules regulating metabolism, cell cycling, apoptosis, differentiation, signaling, structural integrity, and so forth, whereas a large business consists of divisions, such as production, distribution, marketing, public relations, hiring, legal, and so forth. If you were asked to analyze why a business is failing, the equivalent to GWAS would be a bottom-up approach, reviewing the personnel file of each employee (gene), one at a time, to decide whom to pin the blame on. Occasionally, a business failure may be attributed to a single rogue employee (analogous to a monogenic disease), but most business failures are multifactorial (analogous to complex polygenic diseases), in which key divisions are underperforming or not integrating their activities effectively with other divisions. In the human genome, there are millions of SNPs with a prevalence exceeding a few percent of the general population.\(^2\) It is the combined effect of many variants which contribute to an individual’s disease risk. In principle, a bottom-up GWAS approach could be applied to analyze these SNPs and identify the variation involved in disease. However, since each SNP involved in disease only
hypothesis that variation in gene module expression patterns between different strains of mice reflect different “good enough genetic solutions” that are all comparably adequate for normal function but exhibit differential adaptability to disease-inducing stresses. Conceptually, this allows for a gene module association study (GMAS) to explore associations between gene module expression patterns and disease susceptibility, using the same biostatistical techniques developed for GWAS. In the final section, we discuss some possible applications of this GMAS strategy, and speculate on how this approach might be used on a broader scale involving the other “–omics,” to further elucidate the causes of common human diseases and drug reactions.

“Good Enough Solutions”

In a series of fascinating studies, neuroscientist Eve Marder and colleagues have investigated the somato-gastric ganglion of lobsters and crabs. This ganglion, which consists of about 30 neurons with well-defined interactions, exhibits a characteristic neural bursting pattern controlling digestive activity (Figure 1A). When Marder and colleagues analyzed the densities of various ion channels regulating the bursting activity, they were surprised to find that despite very similar bursting patterns, the densities of the key ion channels differed substantially from lobster to lobster, which they also confirmed by measuring mRNA and protein levels. For example, some lobsters expressed Ca-activated K channels at a high level, whereas others relied on a low expression level, but both exhibited nearly identical bursting patterns under normal conditions (Figure 1B). When subjected to an environmental stress, such as histamine exposure, however, the responses could be quite different. When they constructed a mathematical model of the bursting behavior incorporating the relevant ion channels known to be present, and performed a random search of 17 parameters to identify combinations producing similar bursting patterns, of 594 510 models tested, they identified over more than a thousand “good enough solutions” that satisfied the basic requirements. On reflection, this is not really surprising, since it has been long appreciated that when models contain many adjustable parameters, many different combinations can be fitted to the same data (e.g., by requiring that the sum of squares of the differences between the predicted and actual data fall below some arbitrary minimum threshold). To paraphrase John von Neumann, “With four parameters, I can fit an elephant, and with five I can make him wiggle his trunk.”

Why do different lobsters choose different “good enough solutions” for the same bursting behavior? Marder and colleagues argued that it has the evolutionary advantage of enhancing the adaptability of the lobster population as a whole to environmental stresses. For example, imagine that a toxin blocking Ca-activated K channels is dumped into the lobsters’ habitat. Those lobsters whose “good enough solutions” are highly dependent on Ca-activated K channels will be poisoned, and may die. However, lobsters whose “good enough solutions” are minimally dependent on Ca-activated K channels will be spared, so that although the lobster population will be transiently diminished, it will not be extinguished. Thus, the “good enough solution” concept resolves the paradox that a living organism must be both robust, that is, impervious to environmental challenges, yet at the same time be able to adapt to environmental challenges. That is, the organism is resilient and stable in the face of most but not all environmental challenges. As long as all of the individuals in the population are not all susceptible to the same environmental challenges, this differential adaptability ensures that some individuals will survive to restore the population.

It is easy to imagine how the same principle might apply to patients in a clinical setting. For example, consider the human ventricular action potential model illustrated in Figure 2. In this computer modeling study, Sarkar and Sobie used a human cardiac action potential model to demonstrate that many different combinations of 16 ionic current parameters all produced “good enough solutions” to generate a “normal”
cardiac action potential. Figure 2A shows 2 different combinations of parameters that produced nearly identical action potentials (blue and red). For the blue trace, the conductance of rapidly activated delayed rectifier K channels (Gₖᵣ) is much larger than for the red trace, indicating that repolarization of the blue action potential depends more strongly on Gₖᵣ. Now imagine 2 patients, Tom and Jerry, whose genetically determined cardiac action potentials correspond to the blue and red “good enough solutions,” respectively. Tom and Jerry both have normal QT intervals on their electrocardiograms, since the summed ionic conductances are adequate to repolarize either action potential under normal conditions. However, suppose that Tom and Jerry develop infections and are treated with the antibiotic erythromycin, a known Gₖᵣ blocker. Tom’s QT interval may show dramatic QT interval prolongation, since his ventricular repolarization is strongly dependent on Gₖᵣ. Jerry, on the other hand, may show little change, since his ventricular repolarization is relatively insensitive to Gₖᵣ. Thus, Tom is at increased risk of developing life-threatening arrhythmias such as torsade de pointes when exposed to Gₖᵣ-blocking drugs, whereas Jerry is not.

If true in humans as well as lobsters and crabs, the “good enough solution” concept provides an appealing explanation for why some individuals but not others develop side effects from drugs, or, similarly, why only some individuals develop hypertension on a high salt diet or atherosclerosis on a high fat diet. This raises the possibility that “modifier genes,” the “good enough solutions” for normal cardiac function, might be determined more by how normal genes are grouped, rather than by additive effects of multiple gene mutations.

Networks
To understand how our genome produces different “good enough solutions,” we need to understand how gene expression is regulated at a global level, that is, how genes interact with each other within the context of an integrated biological network. In principle, our 23 000 genes could be individually regulated to produce a truly astronomical number of potential “good enough solutions.” However, mounting evidence indicates that biological systems are organized as modular networks, in which genes, proteins, metabolites, and other factors operate in groups rather than as single agents. For example, it is well-appreciated that many transcription factors, micro RNAs, DNA methylation, and chromatin remodeling regulate the expression of large numbers of genes in concert, consistent with genes being organized as a network of coexpression modules.

How can a gene coexpression network, if it exists, be revealed? In network theory, a complex system is represented in a highly abstract manner, dispensing with the fine details and considering the system simply as “nodes” connected by “links” or “edges.” This type of analysis has provided dramatic insights into how network structure (also called network topology) evolves in complex systems, ranging from technology and social networks to biological networks, including gene, metabolic, and protein-protein interaction networks.¹⁰

The process begins when a group of nodes form links with each other. If the process is random, then the majority of nodes will have close to the average number of links per node, and very few nodes will have many more, or many fewer, links than the average. But now suppose that the network is growing, with new nodes and new links being added over time. In this case, the original nodes, having been around longer, will have had more opportunities to capture links as they are added to the system. Thus, older nodes will have, on average, more links than newer nodes, eventually creating a small subset of the oldest nodes, called hub nodes, which are more highly connected than average node. Now imagine one more feature, that nodes that have already acquired a substantial number of links have an advantage in acquiring new links as they are added to the growing network, that is, “the strong get stronger.” This feature, called preferential attachment, further accentuates the connectivity of the hub nodes, generating a “scale-free” network, so-called be-
cause the number of links per node follows a power law distribution.

For a self-organizing system that needs to be robust, adaptable, and efficient to survive in a constantly changing environment full of daily stresses, such as any biological organism, the presence of highly-connected nodes in a network confers some extraordinarily useful properties. First, a network with many parallel connections between nodes has the advantage that should a link between two nodes \(a\) and \(b\) be destroyed, redundant pathways exist to allow the information to flow from \(a\) to \(b\) by an alternate route (eg, through a node \(c\) linked to both \(a\) and \(b\)). In a typical scale-free network, for example, up to 80% of nodes and links can be randomly destroyed before the network fails catastrophically.\(^{11}\) In contrast, a purely linear pathway (eg, \(a \rightarrow b \rightarrow c \rightarrow d\)) is disabled if even a single link or node is destroyed. Thus, the robustness of a network in the face of a constantly changing environment is a key property promoting survival.

The second key network property is the ease with which information flows quickly from one location to another. This “small-world property” is conferred by hub nodes that make long-range connections to other nodes, analogous to the “6 degrees of separation” effect in friendship networks. The ability to access any individual node from a multitude of alternative pathways makes a scale-free network inherently adaptable to changing environmental conditions, another key priority for living systems.

Since the properties described above involve the presence of multiple redundant pathways, one might think that a scale-free network, by favoring robustness and adaptability, sacrifices efficiency compared with a linear system in which the pathway between the input and output is very direct. This might be true if the system operated under nonvarying environmental conditions. However, to adapt to changing environmental conditions, a network is inherently more efficient than a strictly linear system. For example, consider the bacterium *E coli*, whose metabolic pathways include a total of 778 metabolites.\(^{12}\) To represent metabolism as a network, each metabolite is considered a node, and each enzyme which converts one metabolite to another is a link. If, hypothetically, *E coli*’s metabolites (nodes) were arranged in a linear chain, from No. 1 to No. 778, linked by enzymes converting No. 1 to No. 2, No. 2 to No. 3, and so forth (Figure 3A), then consider the following scenario. If metabolite No. 575 were a critical metabolite such as uridine, required for DNA synthesis and replication, its synthesis from nearby precursor metabolites such as No. 523 is highly efficient, requiring only a few steps. However, if only distant precursors such as No. 4 are available, hundreds of steps are now required, making uridine synthesis slow, inefficient, and energetically costly.\(^{11}\) In contrast, highly connected hub nodes (red) in a scale-free network allow efficient conversion of either No. 523 or No. 4 to uridine in only a few steps.

Figure 3. Small-world properties in metabolic networks. A, Hypothetical arrangement of *E coli*’s 778 metabolites in a linear chain. If No. 525 is uridine (required for DNA synthesis and replication), its synthesis from nearby precursor metabolites such as No. 523 is highly efficient, requiring only a few steps. However, if only distant precursors such as No. 4 are available, hundreds of steps are now required, making uridine synthesis slow, inefficient, and energetically costly. B, In contrast, highly connected hub nodes (red) in a scale-free network allow efficient conversion of either No. 523 or No. 4 to uridine in only a few steps.

Figure 4. A, The *E coli* metabolic network, displayed as a topological overlap map, in which metabolites are nodes (circles) and enzymes converting one metabolite to another are links (lines). Colors reflect the modules to which the nodes (metabolites) belong, corresponding to carbohydrate, protein, lipid, nucleic acid, aromatic, monocarbon, and coenzyme metabolism. Adapted from Ravasc et al,\(^{25}\) with permission. B, Log-log plot of the distribution of links per node \(P(k)\) versus \(k\), showing a power law distribution for both ingoing (blue) and outgoing (red) links. Adapted from Barabasi and Oltvai,\(^{50}\) with permission.

A Linear array

B Network with hubs
network displayed visually as a topological overlap map. Figure 4B shows the statistical distribution of the number of links per node on a log-log scale, demonstrating the power law relationship (ie, the defining characteristic of scale-free systems). For E. coli, the metabolic network of 768 nodes contains 5763 links. Due to the hub metabolites (such as ATP, pyruvate, glutamate, NADH, and others), the average number of links (distance) required to convert one metabolite to another is 3.2,12 that is, less than the “six degrees of separation” characterizing many social networks. So our hypothetical E. coli bacteria in Figure 3 would do just fine in a culture medium containing either metabolite No. 573 or No. 4 as its only available substrate for synthesizing uridine.

Barabasi and many others have gone on to show that, in addition to metabolism, both gene coexpression networks and protein-protein interaction networks display similar power law relationships suggesting at least an approximately scale-free topology.10 Although this could be merely coincidental, the convergence of modular systems, ranging from molecular to social networks, toward scale-free topology suggests a fundamental mechanism in biological evolution for optimizing robustness, adaptability, and efficiency in response to a changing environment, the essential qualities needed for a biological organism to thrive.

How to Construct a Gene Network

We now turn to analyzing genes from a modular network perspective. We start by assuming that if genes belong to the same module, they will be coregulated by the same factors, and therefore their expression levels should track each other. In addition, from the considerations enumerated above, we assume that this modular network is likely to exhibit, at least approximately, a scale-free topology. A number of analytical methods have been proposed to detect sets of interacting genes, including nonnegative matrix factorization,13 bayesian networks,14–17 ARACNe,18 Geronemo,19 and MINDy20 or weighted gene co-expression network analysis (WGCNA).21–23 The first step is to quantify gene expression levels in a genetically diverse population, which can be achieved using gene expression microarrays or RNA sequencing. The next step is to analyze this data using an appropriate network analysis algorithm. In the present report, we briefly summarize the approach using one of these approaches, WGCNA. WGCNA utilizes the natural quantitative variation in gene expression levels between individuals in the population to analyze how strongly each gene’s expression level correlates with every other gene’s expression level across the population. The result is a matrix containing all pairwise Pearson correlations between all genes. These Pearson correlations are then transformed to a measure of pairwise connection strength (adjacency), which is accomplished through a soft thresholding process, accomplished by raising the correlation to a fixed power $\beta$ known as the soft threshold (typically, the default $\beta=6$). This particular soft thresholding approach has several practical and theoretical advantages: (1) it results in a linear relationship between adjacency and correlation coefficients on a log scale; (2) $\beta$ is a single threshold parameter which is highly robust with respect to the choice of the threshold; and (3) the resulting correlation network facilitates a geometric interpretation of network statistics.24

A critical step in the network analysis is the definition of modules. Toward this end, the adjacency measure is often transformed into a more robust and biologically meaningful measure of network interconnectedness that takes into account the shared neighbors of each gene pair in the network, called topological overlap.21,25 From the topological overlap values calculated between each gene and every other gene, groups of genes with high topological overlap are hierarchically clustered to define modules.21 Modules correspond to branches of the resulting cluster tree, which are defined using dynamic tree cutting.26 A gene’s connectivity is rated by the sum of all of its adjacency values (or alternatively topological overlap values) with other genes in the same module (defining its intramodular connectivity), as well as with genes in different modules (defining its intermodular connectivity). Intramodular hub genes can be identified as the most highly connected genes in each module.

Since each gene module contains tens to hundreds of genes, the next key data reduction step is to identify a reliable surrogate which expresses the aggregate behavior of each gene module in the network. In WGCNA, the method used is principal component analysis (PCA), which summarizes the expression profile of the genes inside a coexpression module by the module eigengene (the first principal component of the adjacency covariance matrix). The module eigengene is the mathematically optimal summary of the module expression profiles, since it captures the maximum amount of variation.

It is important to emphasize that the construction of the network as outlined above is unbiased—that is, once the raw gene expression data has been processed, no assumptions are made beyond the soft threshold $\beta$ used for defining the adjacency matrix. Setting a low soft threshold is more inclusive, creating a small number of highly-interlinked modules, whereas setting a high threshold is more exclusive, creating a large number of sparsely-interlinked modules with more unlinked genes. How do we choose the most appropriate soft threshold $\beta$? There is no unequivocal answer, but several independent criteria are used in practice. The primary criterion assumes that Darwinian evolution should favor a scale-free network topology, to maximize robustness, adaptability, and efficiency through small-world properties, as discussed earlier. Thus, a practical criterion is to select the smallest value of $\beta$ that yields an approximately scale-free network topology.

Secondary criteria are based on whether the identified modules are biologically meaningful. Assuming that modules in a gene network, such as divisions of a business organization, are likely to have identifiable functions, each module can be examined for the genes with known functions, to determine if the module is enriched in a particular class of genes (eg, cell cycling, apoptosis, cytoskeleton, metabolism, signaling). Thus, from gene ontology (GO) enrichment, biological functions can often be assigned to different modules. Another strategy is to determine whether modules
correlate with known cell markers, since modules sometimes correspond to different cell types within a given tissue.27

In summary, a soft threshold choice yields a network with approximate scale-free topology containing modules with well-defined biological functions (based on gene ontology enrichment or cell marker enrichment) is biologically credible. The interested reader can access further details at: http://www.genetics.ucla.edu/labs/horvath/CoexpressionNetwork/ or in book form.28

Gene networks can be constructed using coexpression data obtained from any population exhibiting genetic diversity. The obvious challenge, however, is to demonstrate, in an independent manner, that a modular gene network is functionally meaningful and not just biostatistical fluke. The key functional criterion for biological validity is whether the expression of gene modules (i.e., the eigengene expression pattern) predicts phenotype. Although this aspect is still in its infancy, eigengene expression levels have been shown to correlate with traits in humans, animal models or cell lines for a variety of diseases including autism,29 Alzheimer disease,30 heart failure,31 obesity,32,33 hyperlipidemia,34 and atherosclerosis.35 Moreover, by associating eigengenes with specific traits, the gene module represented by the eigengene can then be broken down into its constituent submodules, which can be analyzed in detail to discover novel genes and pathways that regulate the phenotype.

**Models to Construct and Validate Gene Networks**

Validation of systems genetics approaches requires that both molecular phenotypes (such as global transcript levels) and clinically relevant phenotypes be concomitantly examined among the individuals of a population. What kinds of populations are available to test and validate the GMAS concept? Human studies are necessarily limited by the difficulty of accessing tissues other than blood and cell lines. Model organisms such as yeast, flies, and worms can serve as excellent resources, but their relevance to common human diseases is limited. Among mammals, rodents are currently the most well-developed models for systems genetics, since strains exhibiting wide genetic variation have been developed and there is a large body of knowledge of their physiology, genetics, and genomics. Sets of recombinant inbred rats, in particular, have proven particularly useful for systems genetic analyses of cardiovascular and metabolic traits.36 The use of mice for genetic studies was facilitated historically by the popularity of breeding mice as pets, called fancy mice, which originated in China and reached its height of popularity in Europe in the late 19th century. The genetic diversity of fancy mice spans more than a million years of murine evolution, when the ancestral Northern Asian, Southern Asian, and European lines first diverged.37

To link genotype to phenotype in rodents, a typical approach is to perform genetic crosses between 2 strains exhibiting different phenotypic responses to an environmental stressor, such as the predisposition to obesity when placed on a high fat diet. Due to homologous recombination during cross-breeding, large regions of chromosomes are randomly swapped among progeny of the two strains, which in turn results in differences in gene expression among the progeny. Drawbacks, however, are that each progeny is genetically unique, so that the correlation between gene expression and phenotype can only be assessed once per genetically-identical individual. Also, the incidence of homologous combination in mice is low, which limits resolution.

Recently, new resources has been developed to circumvent some of these limitations. A highly diverse set of mouse recombinant inbred strains, termed the Collaborative Cross, is now under development.39 Here, we focus on a similar resource called the Hybrid Mouse Diversity Panel (HMDP),38 which consists of 100 (or more) common and recombinant inbred strains which have been either entirely sequenced or densely genotyped with over 140 000 SNPs. The HMDP strains are commercially available and thus can be assayed for multiple phenotypes by different laboratories, providing cumulative biological insights. The ability to use classic inbred strains in this manner is dependent on correction for population structure using an algorithm such as the Efficient Mixed Model Algorithm.40 Since individuals within a given HMDP strain are genetically identical to each other, the reproducibility of gene expression and phenotype among individuals of the same strain can be assessed. Thus, intrastrain variability due to environmental factors can be separated from interstrain variability due to genetics, both with respect to gene expression and phenotypic responses to an environmental stressor. Studies to date have documented that intrastrain variability in gene expression and phenotype is much less than the interstrain variability.38 GWAS studies using the HMDP have successfully detected and finely mapped genes modulating adiposity, bone density, plasma lipids, and other complex traits.41 For example, in a network study examining the genetic basis of adiposity in mice, a set of curated pathways was shown to be enriched in genes differentially expressed between livers of fat and lean mice in a segregating mouse population.22 The identified pathways tended to center on the tricarboxylic acid cycle, which is central in energy production. Through the use of systems genetics, a set of 9 genes were predicted to be causally related to adiposity in mice.42 Their effects on adiposity were validated with transgenic approaches, and expression array analyses of the transgenic mice showed that the differentially regulated genes were enriched in many of the same tricarboxylic acid-centered pathways.42

With these encouraging GWAS results, the use of HMDP for a GMAS approach designed to link groups of genes, rather than single genes, to phenotype, is now beginning to be explored. Figure 5 shows the gene network generated from microarray analysis of heart tissue from 101 HMDP strains. Of 20 000 genes represented on the microarray chip, ~8000 genes were expressed at significant levels in the heart. In Figure 5A, the cardiac gene network is displayed in a topological overlap map. In this visualization, genes whose topological overlap (or adjacency) value exceeds the pre-defined threshold are defined as neighbors, which are color-coded and grouped into modules (clusters) connected by lines. Another useful display is a heat map representation shown in Figure 5B, in which modules containing the 8000 genes are stacked on the vertical axis, and the HMDP strains are arranged along the horizontal axis. The expression level
The patterns of gene expression within a given module in Figure 5B are very complex. To facilitate linking gene modules to phenotype, the aggregate behavior of the individual genes within each module can be represented by an “eigengene,” derived from principal component analysis. The eigengene is not a physical entity, as is a real gene. Rather, it is a phenomenological entity representing a read-out of the aggregate behavior of all the real genes in a module, in the same sense that the blood pressure is not a physical entity, but rather is a measurement or read-out of interactions between physical entities such as the heart, blood vessels, hormone levels, blood volume, and so forth, or like the refractory period is a read-out of the aggregate behavior of many real ion channel proteins determining tissue excitability. Thus, of the ~20,000 individual genes in the mouse genome, a given tissue such as the heart expresses ~8000 genes at significant levels, which can be condensed, using the WGCNA technique, into ~20 gene modules, each represented by an eigengene as the global read-out of the module’s expression level. Figure 5C shows a heat map of eigengene expression levels for the HMDP. The eigengene heat map emphasizes even more strikingly the variation gene module expression patterns among the different HMPD stains.

Given that most of the HMDP strains were bred in captivity and selected for traits such as hair color, a natural question to ask is, why are the patterns of gene module expression so markedly different? Natural selection for most fitness genes should have been similar and traits such as hair color require only a few genes, yet the expression levels of thousands of genes differ between HMDP strains, despite only a small variation among individuals of the same strain. These markedly different gene module (eigengene) expression patterns among HMDP strains appear to be equivalently well-suited for survival in captivity, since all strains are capable of being bred and maintained. One intriguing, although still speculative, possibility is that each HMDP strain represents a different “good enough solution” for survival in captivity under normal conditions. That is, within a given strain, each mouse has roughly the same “good enough solution,” reflecting strain-specific mutations in its regulatory DNA. However, other HMDP strains, with different strain-specific mutations in regulatory DNA, manifest different eigengene patterns encoding alternative “good enough solutions.”

To test the feasibility of this idea, we analyzed the variation in gene expression levels observed in the microarray analysis of the HMDP for five key cardiac proteins regulating normal cardiac excitation-contraction (EC) coupling in the mouse ventricle (the transient outward current Ito,f, the ryanodine receptor RyR; the Na-Ca exchanger NCX, the sarcoplasmic endoplasmic reticulum Ca ATPase SERCA; and the L-type Ca current Ica,L). The variation of mRNA expression for those 5 genes analyzed for one microarray (one heart) per strain for 100 strains is shown in Figure 6A. We then compared these findings to an exhaustive computational search of “good enough solutions” in a mouse ventricular action potential model (Figure 6B). The computational search was conducted by first generating 10,000 “trial solutions” constructed by assigning random values to the conductances corresponding to those five EC proteins, with each conductance varying from a value a few times smaller to a few times larger than its normal value. Not surprisingly, only a handful of trial solutions were good enough. To systematically find good enough solutions, we therefore used a minimization algorithm that iteratively varied all 5 conductances starting from the randomly chosen values until a positively defined “cost function,” which measures the departure from a normal electrophysiological phenotype (action potential and calcium transient), falls below a tolerance limit. This mathematical optimization yielded 1952 good enough solutions from the 10,000 trial solutions with the corresponding variation of conductances shown in Figure 6B.

Remarkably, the conductance variation of computer-generated good enough solutions matches reasonably well the variation of mRNA expression among the 100 HMDP strains. Only a rough agreement is expected at best since functional conductances may not always correlate well with mRNA expression levels due to posttranscriptional and/or posttranscriptional regulation, through alternative splicing, or posttranslational modifications.
Getting Down to Business: The GMAS Strategy

To unravel the genetic basis of common polygenic diseases, our goal is to devise a top-down strategy to analyze how genes work together in groups, rather than singly. Toward this end, the salient observations/assumptions discussed so far can be summarized as follows: (1) Applying WGCNA to the HMDP allows us to group the mouse genome of ≈20,000 genes into an approximately scale-free network comprised of ≈20 gene modules in a tissue such as the heart; (2) the aggregate behavior of the genes in a given module can be represented phenomenologically by the module’s eigengene, a read-out of the module’s expression level calculated from principal component analysis; (3) the expression pattern of eigengenes is very similar among individuals within a given HMDP strain but varies considerably between different HMDP strains, reflecting the concept that evolution has generated different “good enough solutions” adequate for normal cardiovascular function; (4) when the HMDP strains are exposed to an environmental stress, however, some strains may adapt better than others, consistent with some “good enough solutions” being better suited to adjust successfully to a specific stress than others; and (5) if the different eigengene expression patterns of HMDP strains truly correspond to different “good enough solutions,” then the eigengene pattern should predict the phenotypic response to the stressor.

If these observations/assumptions are valid, then the top-down strategy is straightforward: (1) Expose all of the ≈100 HMDP strains to an environmental stress of interest, for example, a high fat diet to induce atherosclerosis, chronic isoproterenol or angiotensin infusion to induce heart failure, or any stressor known to induce a disease phenotype; (2) functionally characterize the range of phenotypes in terms of quantitative traits exhibited by the ≈100 HMDP strains in response to the stressor; and (3) correlate the eigengene expression patterns of the HMDP strains with the panel of quantitative traits.

Thus, in place of the bottom-up approach of looking for associations between single genes (or SNPs) and phenotype, here we narrow the number of candidates from thousands of individual genes (or millions of SNPs) to a relatively small number of eigengenes (20–30), with each eigengene representing tens to hundreds of individual genes, depending on the size of the module. As in the case of a failing business organization, we are no longer looking one-by-one for a single rogue employee (gene) to blame but instead are viewing how the system’s modular divisions (represented by its eigengenes) interact and adapt to the stressor. Moreover, with only 20 or so eigengenes, combinations of eigengenes are much more feasible to evaluate. That is, searching for a correlation of a single eigengene with a quantitative trait involves evaluating 20 possibilities; correlating pairs of eigengenes involves evaluating \(20(19! \times 2!)=190\) possible pair combinations; for triplets the corresponding number is \(20(17! \times 3!)=1140\); and for groups of 16 eigengenes, the corresponding number is \(20(16! \times 4!)=4845\), still very manageable compared with the number of possible 16 SNP combinations, which exceeded \(10^{480}\) (the number of atoms in the observable universe), using a GWAS approach.

---

Figure 6. Correspondence between gene expression and “good enough solution” predictions. Left panels (blue) show histograms of gene expression levels for the same 5 key EC coupling proteins among 100 HMDP strains. The reasonable agreement suggests that hypothesis test if gene expression variation in HMDP strains reflects multiplicity of functional “good enough solutions” for normal murine EC coupling.
Caveats and Challenges

In the discussion above, we have sketched out a relatively straightforward scenario for a GMAS strategy, based on the assumptions that the basal eigengene pattern encodes a “good enough solution” for normal function, and some “good enough solutions” adapt more successfully than others to environmental stress. If these assumptions are correct, then it logically follows that the eigengene pattern should predict the phenotypic response to the stressor.

The challenge, of course, will be to validate these assumptions. Models such as the HMDP are valuable resources in this regard. HMDP has the advantage that the reproducibility of eigengene patterns and phenotypic response to stressors can be cumulatively tested in multiple genetically identical individuals to separate environmental influences from genetic determinants. However, the success of this approach ultimately depends on the accuracy of techniques such as WGCNA or alternative strategies at deciphering biologically meaningful gene networks. In other words, to what extent are the derived gene modules real or biostatistical flukes due to contamination by genetic noise? How does one decide on which genes to include in which module—are maximizing the scale-free topology and using gene ontology and cell marker enrichment to infer function the best criteria? Even if the gene network modular topology is accurate, are eigengenes calculated by principal component analysis the best way to represent the aggregate behavior of individual genes in a module? Both WGCNA and eigengene calculations are based on linear correlation measures. Whether more general nonlinear methods to group genes into modules can improve accuracy is yet to be determined.

It is also important to recognize that eigengene patterns reflect a mixture of both genetic and environmental (epigenetic) perturbations. Coexpression modules are imperfect representations of biology. They can be refined by experimental and interactive approaches, but they are limited by the fact that they represent only one of many scales contributing to overall function, and changes in gene expression patterns reflect both genetic and environmental perturbations. Thus, it will be important to expand the approach to datasets reflecting epigenetic control (DNA methylation, chromatin structure), alternative splicing (next generation sequencing), protein localization and post-translational modification (mass spectroscopy, imaging) and metabolism (mass spectroscopy). Nevertheless, the available evidence suggests that the gene network structure is relatively invariant—that is, individual genes do not jump to different modules in response to a stressor. Rather, the gene content of modules remains similar, but the expression level of the genes in the module (represented by the eigengene) changes. In this case, it may turn out that the change in the eigengene pattern, rather than the basal pattern, will be a better predictor of phenotype. These are all important issues that remain to be explored.

Clinical Implications

If the GMAS strategy outlined above for the HMDP proves successful in identifying gene module expression patterns that predict adaptability to environmental stressors, how might this help us better understand the genetic basis of common human diseases or drug reactions? With the exception of identical twins, all humans are genetically unique (ie, each is analogous to a different HMDP strain). However, the principle is the same: if each human’s genome encodes a different “good enough solution” for survival in a “normal” human environment, then some “good enough solutions” may adapt better than others when the environment changes (eg, from a subsistence diet to a high calorie, high fat diet, or from drug-free to a new drug state), accounting for the variable susceptibility to complex diseases. Using WGCNA, a human gene network can be constructed with microarray data from less than 100 patients. Thus, tissue-specific human and mouse gene networks, and their corresponding eigengene patterns, can be compared between human diseases and corresponding mouse disease models to search for commonalities. For example, we might expose the HMDP strains to an arrhythmia-inducing drug to search for specific eigengene patterns associated with electrocardiographic QT interval prolongation and cardiac arrhythmias; and if so, determine whether the corresponding human eigengene patterns can be similarly identified. Moreover, novel therapies for complex diseases may be suggested if we can discover interventions which alter eigengene expression patterns, and show that this also alters diseases susceptibility. At the present time, regulation of gene module expression levels is not well-understood, but interesting candidates to evaluate include hub and driver genes, microRNAs, DNA methylation or acetylation, and histone acetylation/chromatin remodeling. For example, microRNAs often coordinate the regulation of hundreds of individual genes. Do the genes regulated by a given microRNA belong to the same module? If so, then manipulation of specific microRNAs might provide a molecular approach to convert an eigengene pattern conferring susceptibility to a disease to one conferring resistance.

GMAS and GWAS approaches are also inherently complementary to each other. One of the major contributions of GWAS has been to identify novel pathways previously unsuspected to be involved in the pathogenesis of a disease. By linking pathways to discrete gene modules, GMAS provides an additional context for discovering how pathways are interactively linked in health and disease states. The synergy between GWAS and GMAS may thus enhance our ability to identify promising therapeutic targets among the molecular constituents of these pathways.

Realization of the potential clinical applications of GMAS will require much more work in this emerging research area which is still in its infancy. However, if a GMAS approach based on eigengene analysis or alternative strategies ultimately proves successful, then our view of how genetics influences susceptibility to common diseases and drug reactions may shift away from the traditional focus on searching for mutations, and toward studying the way in which normal genes are grouped. We already know that an identical set of genes can produce a markedly different phenotype depending

Downloaded from http://circres.ahajournals.org/ by guest on January 17, 2018
on the expression pattern (eg, the metamorphosis of a caterpillar into a butterfly). So perhaps it is not so far-fetched to imagine that whether or not an individual succumbs to a disease or drug reaction may depend more on the “good enough solution” reflected in the gene module expression pattern of completely normal genes, rather than on multiple gene mutations. Finally, gene networks are only one of many networks vital to a living organism. The ultimate and even more daunting challenge will be to map how information flows between gene networks, protein-protein interaction networks, metabolite networks, and so forth. The emerging science of network analysis, combined with “-omics” technologies, are powerful new tools whose further development holds great promise to meet these challenges.

Acknowledgments
We thank Jonathan Hoffman and Alan Garfinkel for participating in helpful discussions.

Sources of Funding
This work was supported by NIH grants P01 HL078931, P01 HL080111, P01 HL28481, P01 HL30568, R01 GM05656, R21 HL110667–01, R01 HL094322, 1DP3 D094311, R01 HL101228, Ruth L. Kirschstein National Research Service Award T32HL09766, NIH/NHLBI Contract No. HHSN268201000035C, by the Laubisch and Kawata Endowments, and a seed grant for interdisciplinary research from Northeastern University.

Disclosures
None.

References


"Good Enough Solutions" and the Genetics of Complex Diseases
James N. Weiss, Alain Karma, W. Robb MacLellan, Mario Deng, Christoph D. Rau, Colin M. Rees, Jessica Wang, Nicholas Wisniewski, Eleazar Eskin, Steve Horvath, Zhilin Qu, Yibin Wang and Aldons J. Lusis

Circ Res. 2012;111:493-504
doi: 10.1161/CIRCRESAHA.112.269084
Circulation Research is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231
Copyright © 2012 American Heart Association, Inc. All rights reserved.
Print ISSN: 0009-7330. Online ISSN: 1524-4571

The online version of this article, along with updated information and services, is located on the
World Wide Web at:
http://circres.ahajournals.org/content/111/4/493

Permissions: Requests for permissions to reproduce figures, tables, or portions of articles originally published in Circulation Research 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 Circulation Research is online at:
http://circres.ahajournals.org//subscriptions/