

**A systems-based framework for understanding complex metabolic and
cardiovascular disorders**

Sulin Wu¹, Aldons J Lusic^{1,2} and Thomas A Drake^{3,*}

**¹Department of Human Genetics, David Geffen School of Medicine at University of
California, Los Angeles, CA 90095-1679**

**²Department of Microbiology, Immunology, and Molecular Genetics, University of
California, Los Angeles, CA 90095-1679**

**³Department of Pathology and Laboratory Medicine, David Geffen School of
Medicine at University of California, Los Angeles, CA 90095-1679**

Characters (including spaces, References and figures): 33,377

References: 43

***Corresponding Author: Thomas A Drake, MD, ³Department of Pathology and
Laboratory Medicine, David Geffen School of Medicine at University of California,
Los Angeles, CA 90095-1679 USA. Phone: (310) 825-6975;
Email: TDrake@mednet.ucla.edu**

Abstract

Common forms of metabolic and cardiovascular diseases involve the interplay of numerous genes as well as important environmental factors. Traditional biochemical and genetic approaches generally attempt to dissect these diseases one gene at a time, for example, by analysis of Mendelian forms or genetically engineered experimental organisms. But, it is also important to understand how the genes interact with each other, and the environment, and how these interactions change in disease states. Technological advances, such as the development of expression arrays that allow quantification of all transcript levels in a cell or tissue, have made it feasible to globally monitor molecular phenotypes that underlie disease states. By applying statistical methods, relationship between DNA variation, gene expression patterns and diseases can be modeled.

Introduction

Common forms of metabolic and cardiovascular diseases, constituting by far the major cause of death in most of the world, are exceptionally complex. Although only a tiny fraction of the underlying genes have been identified (most in the last two years), it seems likely that common or rare variants of hundreds or even thousands of genes will be involved (1, 2). And, of course, environmental factors such as over-nutrition, sedentary lifestyle, and smoking play a crucial role. Most biochemical and genetic studies to date, such as those involving transgenic animals, have focused on identifying and characterizing the individual genes that contribute to disease (3). Likewise, genome-wide association (GWA) and other genetic approaches typically focus on finding the specific genes responsible for the association (4). While these approaches continue to be informative, it is also important to address the interactions between genes and the environment (5). In this brief review, we discuss how systems-based approaches involving the integration of genomic, molecular and physiological/clinical data, can complement traditional approaches to address the complexity of these disorders.

A systems perspective on disease envisions the integration of multiple elements, from genome through phenotype as depicted in Figure 1. The technological advances that allow large scale and high throughput quantification of elements at each level are critical. Thus far, only genome and transcriptome come near the level necessary, and so we focus on these as exemplary of the systems approach.

Identifying Disease-associated Genes by Integrating Genetic and Gene Expression

Information

Candidate gene studies and recent GWA studies have identified an impressive list of genes contributing to complex disorders such as atherosclerosis (6), hyperlipidemia (7, 8), obesity (9) and diabetes (10), but altogether these account for a minority of the genetic effect on disease development. This suggests that many genes carrying relatively small to modest effect contribute to complex disease, as has long been postulated. The established biology underlying these disorders has identified at least hundreds of genes for each. Genetic studies in humans and animal models have made significant contributions to elucidating the pathogenesis of complex diseases, though the number of

specific genes identified has been only a fraction of the total expected to be involved for any given disease. Accordingly, one of the major challenges in studying complex disease is to understand how genes that carry causal variants interact with each other, and with “downstream” genes, to regulate disease expression.

Traditional genetic studies in humans or animal models establish the relationship between genotype and phenotype, providing little to inference as to the intervening biology or, correspondingly, guidance for selecting likely causative genes. Incorporating global gene expression analyses into genetic studies has significantly enhanced both these aspects (1, 11, 12). Identification of genetic variants that control trait variation provides information to identify causal factors responsible for gene expression and phenotype changes. Gene expression can be measured as a quantitative trait and has been observed to be highly heritable (12-14). Linkage or association analysis can be employed to identify genetic loci or single nucleotide polymorphism (SNP) perturbing both abundance and activity of gene products. The linkage-based identification of loci linked with gene expression is termed eQTL mapping, and the same concept can be extended to discover SNPs associated with transcript abundance, termed eSNP (15, 16). Observation of eQTL for a gene indicates that genetic factors are partially responsible for its transcript abundance. According to proximity between genetic factors influencing gene expression and the location of the gene, eQTL can be categorized as *cis*- or *trans*-eQTL. When an eQTL localizes closely to the location of the gene coding the transcript, it is likely that the causative genetic variations resides within the gene or its regulator elements and directly influence transcription, splicing or protein stability. The presence of proximal genetic components regulating gene expression is referred to regulatory mechanisms acting in *cis* manner, termed *cis*-eQTL. Conversely, when eQTL does not compass the physical location of the gene and its flanking regions the gene is defined as being regulated in *trans* manner, termed *trans*-eQTL (Figure 2A). In studies in mice, for example, a 20Mb distance around the location of gene has been used to define the likely *cis*-acting manner of loci linked with transcript abundance. In several genetic crosses between inbred strains of mice, hundreds or thousands of expressed genes in a given tissue have at least one eQTL, including both *cis* and *trans* ones. Among all observed

eQTLs in crosses of several hundred or more mice, the total number of *trans*-eQTL is about ten times higher than *cis*-eQTL.

Cis-acting eQTL are obvious candidates for genes underlying a phenotypic trait (the quantitative trait gene or QTG) when the eQTL and the trait QTL coincide. Though causative mutations may not act by altering transcript levels, many do. Applying this concept can significantly reduce the time and effort involved in positional cloning of genes. For example, this was a critical factor in our discovery of *Abcc6* as the major causal gene in cardiac calcification in a genetically randomized mouse population (17). Coincidence of *trans*-eQTL and trait QTL may also be informative for downstream genes involved in trait expression as discussed below.

Integrative Genetics Allows Causal Inference Between Trait-Transcript Correlations

The application of global gene expression analyses to studies of cells and tissues has provided a wealth of data and new knowledge relevant to complex diseases. They have allowed investigators to search for novel relationships between genes and particular traits or processes, without the limitation of needing prior experimental data or hypotheses. Finding statistically significant correlations between a trait and particular genes suggests a biologic relationship between them (18, 19). Many databases and analytical tools have been developed to help identify and characterize which functional classes of genes may be involved in a given process. Gene Set Enrichment Analysis (GSEA) (20), Database for Annotation Visualization and Integrated Discovery (DAVID) (21, 22) and Ingenuity Pathway Analysis (IPA) are analytical tools developed to test the enrichment of particular biological processes and molecular functions of gene sets by examining information collected by databases such as Gene Ontology (GO) and KEGG pathway.

However, as has long been recognized in statistics, correlation does not prove causality. In complex diseases, genes may be correlated with a given trait because they are directly involved in the development of that process (i.e. “causal”) or because the process itself secondarily alters the expression of the genes (“reactive”). The causal genes are usually of primary interest, but with simple correlation analyses, one cannot distinguish these. A major contribution of integrative genetics has been the development

of analytical tools to allow causal inferences to be made between correlated genes and traits when genetic data are incorporated (23, 24). This elaborated in Figure 2B, which shows possible relationships between trait, transcript and genetic location when correlations among these are observed. Because genetic variation is for practical purposes always primary, this can be utilized to order the relationship between transcript and trait.

Several analytical approaches have been proposed to assess potential “causality” in this setting, and undoubtedly more will be developed given the importance of the problem. Schadt and colleagues developed the likelihood-based causality model selection (LCMS) procedure and applied it to predicting liver expressed genes causal for abdominal obesity in a mouse intercross setting (23). Subsequent studies using transgenic or knockout models have validated 8 of 9 predicted genes. Structural equation modeling has also been applied to assess directionality between correlated traits and transcripts, termed “edge orientation”. An algorithm based on this approach was developed by Aten et al., termed Network Edge Orientation (NEO) (25). By employing NEO algorithm, genetic markers are used for conditional correlation test to indicate the causal relationship between traits, including gene-gene and gene-phenotype pairs. As shown by Aten et al., examination of female liver gene expression data together with genotyping information from the segregating mouse population successfully retrieved a known causal relationship in the cholesterol biosynthetic pathway consisting of *Insig1* and some of its downstream genes, *Dhrc7* and *Fdft1*(25).

Constructing gene expression networks for complex diseases

Rather than uncovering single genes for complex diseases such as atherosclerosis, a systems-based perspective is interested in elucidating the interactions of genes and environment operating on a complex multicellular biological system (26). Such a “systems” approach involves modeling the relationship among elements of the system such as transcript levels in the form of a network. Two major modeling approaches have been employed to decipher network patterns underlying complex traits: forward and reverse engineering. Forward approaches apply a set of equations generated a priori from previously defined biologic relationships that are then tested and revised as needed. This approach generally is used with relatively small-scale network formation. The reverse

approach does not apply a pre-defined set of relationships. Rather it utilizes general mathematical tools for network construction, and lets the data itself define the relationships among the elements being studied, such as transcript level (27). This approach typically utilizes large data sets and is computationally intensive. In the setting of data obtaining from populations with genetic and or environmental variations, such analyses allow one to infer the relationship and interaction among all such elements. This is valuable for studying complex diseases, since we do not yet adequately understand the relationships between gene expression and trait variability.

Network analysis provides a useful framework to identify and visualize interactions among genes, by creating a graphic model. A network is composed of elements, such as specific gene transcripts (referred to technically as “nodes”), and connections (relationships) among these (“edges”). Edges can indicate a relationship between genes as at transcript level, protein interaction pattern, and any other measurement that describes a meaningful association between two elements of the system. A gene transcriptional network is composed of individual gene transcripts as nodes, while the edges represent a measure of pair wise correlation of transcript levels. A given gene can correlate with multiple genes, and a measure of the relative number of such connections is referred to as connectivity (28, 29). An important feature of network constructed with biological data is the “scale-free” nature. In a scale free network, there are a relatively small number of highly connected genes, and many more with far fewer connections. Such highly connected genes are often referred to as “hubs”. Targeting hub genes have been found to disrupt the structure of gene networks and are more likely to impact biological processes when disrupted in animal models. For example, in our analyses of co-expression networks for activated endothelial cells, *Atf4*, *Xbp1* and *Insig1* were identified as hub genes (among others) (30). Targeted knockout of these genes had been shown to be lethal in mouse models (31-34).

Similar to other biologic networks, genes in coexpression networks are found to organize into “modules”, which are clusters of genes which have higher degree of connectedness with other members of the same module than with genes in different modules (29). Genes composing a module therefore tend to behave more similarly to one another with regard to correlation with phenotypes (Figure 3), and are often enriched for

particular functional categories of genes. Using data reduction methods such as principle component analysis, the aggregate of genes in a module can be characterized by a single value to use for such analyses (35). One example from our work is the identification of the Unfold Protein Response (UPR) pathway as being important in the response of endothelial cells to oxidized lipids. Coexpression networks were constructed of transcripts induced by oxPAPC in primary endothelial cells (EC) isolated from human aorta (30, 36). Two of fifteen modules were strongly correlated with interleukin-8 induction. These modules were enriched for UPR pathway genes, and several of the most highly connected genes were members of the UPR pathway. “Knockdown” experiment with certain UPR genes revealed the role of the UPR pathway in the regulation of interleukin-8 and other cytokines. In other experiments, a hub gene in the UPR module, MGC4504, proved to contribute to an apoptosis response. These findings led to the discovery of a novel gene that is critical for UPR function in this process, based on its being a hub gene in the same module and closely associated with known UPR genes. By traditional analyses, it mostly would not have been recognized as being particularly important.

Gene coexpression network construction provides topological properties of biological networks by partitioning the transcriptome data into functional units (modules) made of co-regulated genes at transcript level (29). However, how genes interact with each other and how exactly the genetic information flows via regulatory or signaling mechanisms to influence traits is not explained by gene coexpression network alone. Accordingly, strategies leveraging biological and genetical information into gene network may be useful to provide possible explanation for trait variability (37, 38). In studies of common and complex diseases, the identification of regulatory components and explanation of interconnectedness among traits provide useful information to prioritize targets as to improve treatment, diagnosis and prevention, as in the format of personalized medicine (26, 37). Especially for late-onset diseases, ie. cardiovascular heart diseases, the identification and effective control of risk factors in early adulthood is valuable. Accordingly, the establishment of a functional network with directed edges among all genes together with their activity in the network is important in identifying

critical modulators causing the disease under the stress of environment stress and medication.

One promising approach to reconstruct a directed gene network is the Bayesian Modeling approach. Instead of examining strictly the gene connectivity and module formation, Bayesian modeling is useful in leveraging genetic information to infer causality among genes in the directed network (39). As a probabilistic model approach, Bayesian network reconstruction utilizes posterior probability to map traits with particular markers to exploit the increased information from joint mapping of correlated transcripts. Schadt and colleagues have incorporated the genetic data into Bayesian networks for hepatic gene expression from genetically randomized mouse populations, using the LCMS approach described above for causal gene detection (Yang X et al., manuscript under revision). This provides a significant improvement over the gene coexpression network constructed without the genetic data by incorporating predicted causal relationships among genes. Such "directed" networks can elucidate the mechanisms underlying phenotypes by which causal regulators give rise to changes in expression activity of various genes. More recently, using expression data from a genetically randomized yeast population, Zhu et al integrated noisy protein interaction data collected from various sources as well as genetic information into gene expression network by applying Bayesian modeling approach (40).

Recently, a genome-wide functional network for mouse population has also been established and validated, with which Bayesian integrative modeling brings the protein interaction pattern together with gene expression profiles to illustrate the network including probabilistic functional linkages among over 20,000 genes (38). Furthermore, a cross-species comparison of topological properties of functional networks revealed the characteristics of conserved sub-networks in different organisms, which indicates the value of establishing network model in various organisms to facilitate the medical studies in regard of human complex diseases. In addition to Bayesian probabilistic models, there are various ways to integrate biological information, especially transcriptional regulatory mechanisms into gene network underlying complex diseases. Success in identifying critical regulator, for example, acting as cell cycle regulator or component of a transcription factor complex, both have great impact on biological regulation and

dynamics of gene expression, have been demonstrated in various organisms by employing integrative genetics approaches (41). Furthermore, the involvement of tissue specificity and sex effect in complex disease formation can also provide further information in personalized medicine in the assist of systems biology approaches.

We are a long way from understanding complex diseases from a systems perspective. However, the use of high throughput global gene expression assays in the context of genetic analyses has shown how an integrative genetics approach can reveal higher order interactions for traits as complex as diabetes and heart disease (1,2,5,8). As analogous methods for the metabolomic and proteomic elements are developed, progressively richer models of complex disease will be developed (42, 43).

References

1. Lusis, A. J., A. D. Attie and K. Reue. 2008. Metabolic syndrome: from epidemiology to systems biology. *Nat Rev Genet.* **9**: 819-30.
2. Watkins, H. and M. Farrall. 2006. Genetic susceptibility to coronary artery disease: from promise to progress. *Nat Rev Genet.* **7**: 163-73.
3. Flint, J., W. Valdar, S. Shifman and R. Mott. 2005. Strategies for mapping and cloning quantitative trait genes in rodents. *Nat Rev Genet.* **6**: 271-86.
4. Kruglyak, L. 2008. The road to genome-wide association studies. *Nat Rev Genet.* **9**: 314-8.
5. Gibson, G. 2008. The environmental contribution to gene expression profiles. *Nat Rev Genet.* **9**: 575-81.
6. Chen, Y., J. Rollins, B. Paigen and X. Wang. 2007. Genetic and genomic insights into the molecular basis of atherosclerosis. *Cell Metab.* **6**: 164-79.
7. Suviolahti, E., H. E. Lilja and P. Pajukanta. 2006. Unraveling the complex genetics of familial combined hyperlipidemia. *Ann Med.* **38**: 337-51.
8. Wang, J., M. R. Ban, G. Y. Zou, H. Cao, T. Lin, B. A. Kennedy, S. Anand, S. Yusuf, M. W. Huff, R. L. Pollex, et al. 2008. Polygenic determinants of severe hypertriglyceridemia. *Hum Mol Genet.* **17**: 2894-9.
9. Bell, C. G., A. J. Walley and P. Froguel. 2005. The genetics of human obesity. *Nat Rev Genet.* **6**: 221-34.
10. Frayling, T. M. 2007. Genome-wide association studies provide new insights into type 2 diabetes aetiology. *Nat Rev Genet.* **8**: 657-62.
11. Johannes, F., V. Colot and R. C. Jansen. 2008. Epigenome dynamics: a quantitative genetics perspective. *Nat Rev Genet.* **9**: 883-90.
12. Phillips, P. C. 2008. Epistasis--the essential role of gene interactions in the structure and evolution of genetic systems. *Nat Rev Genet.* **9**: 855-67.
13. Emilsson, V., G. Thorleifsson, B. Zhang, A. S. Leonardson, F. Zink, J. Zhu, S. Carlson, A. Helgason, G. B. Walters, S. Gunnarsdottir, et al. 2008. Genetics of gene expression and its effect on disease. *Nature.* **452**: 423-8.
14. Dixon, A. L., L. Liang, M. F. Moffatt, W. Chen, S. Heath, K. C. Wong, J. Taylor, E. Burnett, I. Gut, M. Farrall, et al. 2007. A genome-wide association study of global gene expression. *Nat Genet.* **39**: 1202-7.
15. Schadt, E. E., C. Molony, E. Chudin, K. Hao, X. Yang, P. Y. Lum, A. Kasarskis, B. Zhang, S. Wang, C. Suver, et al. 2008. Mapping the genetic architecture of gene expression in human liver. *PLoS Biol.* **6**: e107.
16. Drake, T. A., E. E. Schadt and A. J. Lusis. 2006. Integrating genetic and gene expression data: application to cardiovascular and metabolic traits in mice. *Mamm Genome.* **17**: 466-79.
17. Meng, H., I. Vera, N. Che, X. Wang, S. S. Wang, L. Ingram-Drake, E. E. Schadt, T. A. Drake and A. J. Lusis. 2007. Identification of *Abcc6* as the major causal gene for dystrophic cardiac calcification in mice through integrative genomics. *Proc Natl Acad Sci U S A.* **104**: 4530-5.
18. Segal, E., N. Friedman, N. Kaminski, A. Regev and D. Koller. 2005. From signatures to models: understanding cancer using microarrays. *Nat Genet.* **37** **Suppl**: S38-45.

19. Park, J. and A. L. Barabasi. 2007. Distribution of node characteristics in complex networks. *Proc Natl Acad Sci U S A*. **104**: 17916-20.
20. Subramanian, A., P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, et al. 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*. **102**: 15545-50.
21. Huang da, W., B. T. Sherman, Q. Tan, J. R. Collins, W. G. Alvord, J. Roayaei, R. Stephens, M. W. Baseler, H. C. Lane and R. A. Lempicki. 2007. The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists. *Genome Biol*. **8**: R183.
22. Huang da, W., B. T. Sherman, Q. Tan, J. Kir, D. Liu, D. Bryant, Y. Guo, R. Stephens, M. W. Baseler, H. C. Lane, et al. 2007. DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. *Nucleic Acids Res*. **35**: W169-75.
23. Schadt, E. E., J. Lamb, X. Yang, J. Zhu, S. Edwards, D. Guhathakurta, S. K. Sieberts, S. Monks, M. Reitman, C. Zhang, et al. 2005. An integrative genomics approach to infer causal associations between gene expression and disease. *Nat Genet*. **37**: 710-7.
24. Degnan, J. H., J. Lasky-Su, B. A. Raby, M. Xu, C. Molony, E. E. Schadt and C. Lange. 2008. Genomics and genome-wide association studies: an integrative approach to expression QTL mapping. *Genomics*. **92**: 129-33.
25. Aten, J. E., T. F. Fuller, A. J. Lusis and S. Horvath. 2008. Using genetic markers to orient the edges in quantitative trait networks: the NEO software. *BMC Syst Biol*. **2**: 34.
26. Sieberts, S. K. and E. E. Schadt. 2007. Moving toward a system genetics view of disease. *Mamm Genome*. **18**: 389-401.
27. Schadt, E. E. and P. Y. Lum. 2006. Thematic review series: systems biology approaches to metabolic and cardiovascular disorders. Reverse engineering gene networks to identify key drivers of complex disease phenotypes. *J Lipid Res*. **47**: 2601-13.
28. Barabasi, A. L. and Z. N. Oltvai. 2004. Network biology: understanding the cell's functional organization. *Nat Rev Genet*. **5**: 101-13.
29. Zhang, B. and S. Horvath. 2005. A general framework for weighted gene co-expression network analysis. *Stat Appl Genet Mol Biol*. **4**: Article17.
30. Gargalovic, P. S., M. Imura, B. Zhang, N. M. Gharavi, M. J. Clark, J. Pagnon, W. P. Yang, A. He, A. Truong, S. Patel, et al. 2006. Identification of inflammatory gene modules based on variations of human endothelial cell responses to oxidized lipids. *Proc Natl Acad Sci U S A*. **103**: 12741-6.
31. Reimold, A. M., A. Etkin, I. Clauss, A. Perkins, D. S. Friend, J. Zhang, H. F. Horton, A. Scott, S. H. Orkin, M. C. Byrne, et al. 2000. An essential role in liver development for transcription factor XBP-1. *Genes Dev*. **14**: 152-7.
32. Carter, S. L., C. M. Brechbuhler, M. Griffin and A. T. Bond. 2004. Gene co-expression network topology provides a framework for molecular characterization of cellular state. *Bioinformatics*. **20**: 2242-50.
33. Tanaka, T., T. Tsujimura, K. Takeda, A. Sugihara, A. Maekawa, N. Terada, N. Yoshida and S. Akira. 1998. Targeted disruption of ATF4 discloses its essential role in the formation of eye lens fibres. *Genes Cells*. **3**: 801-10.

34. Engelking, L. J., G. Liang, R. E. Hammer, K. Takaishi, H. Kuriyama, B. M. Evers, W. P. Li, J. D. Horton, J. L. Goldstein and M. S. Brown. 2005. Schoenheimer effect explained--feedback regulation of cholesterol synthesis in mice mediated by Insig proteins. *J Clin Invest.* **115**: 2489-98.
35. Langfelder, P. and S. Horvath. 2007. Eigengene networks for studying the relationships between co-expression modules. *BMC Syst Biol.* **1**: 54.
36. Gargalovic, P. S., N. M. Gharavi, M. J. Clark, J. Pagnon, W. P. Yang, A. He, A. Truong, T. Baruch-Oren, J. A. Berliner, T. G. Kirchgessner, et al. 2006. The unfolded protein response is an important regulator of inflammatory genes in endothelial cells. *Arterioscler Thromb Vasc Biol.* **26**: 2490-6.
37. Lee, D. S., J. Park, K. A. Kay, N. A. Christakis, Z. N. Oltvai and A. L. Barabasi. 2008. The implications of human metabolic network topology for disease comorbidity. *Proc Natl Acad Sci U S A.* **105**: 9880-5.
38. Guan, Y., C. L. Myers, R. Lu, I. R. Lemischka, C. J. Bult and O. G. Troyanskaya. 2008. A genomewide functional network for the laboratory mouse. *PLoS Comput Biol.* **4**: e1000165.
39. Zhu, J., M. C. Wiener, C. Zhang, A. Fridman, E. Minch, P. Y. Lum, J. R. Sachs and E. E. Schadt. 2007. Increasing the power to detect causal associations by combining genotypic and expression data in segregating populations. *PLoS Comput Biol.* **3**: e69.
40. Zhu, J., B. Zhang, E. N. Smith, B. Drees, R. B. Brem, L. Kruglyak, R. E. Bumgarner and E. E. Schadt. 2008. Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. *Nat Genet.* **40**: 854-61.
41. Lee, S. I., D. Pe'er, A. M. Dudley, G. M. Church and D. Koller. 2006. Identifying regulatory mechanisms using individual variation reveals key role for chromatin modification. *Proc Natl Acad Sci U S A.* **103**: 14062-7.
42. Ferrara, C. T., P. Wang, E. C. Neto, R. D. Stevens, J. R. Bain, B. R. Wenner, O. R. Ilkayeva, M. P. Keller, D. A. Blasiolo, C. Kendzierski, et al. 2008. Genetic networks of liver metabolism revealed by integration of metabolic and transcriptional profiling. *PLoS Genet.* **4**: e1000034.
43. Chaibub Neto, E., C. T. Ferrara, A. D. Attie and B. S. Yandell. 2008. Inferring causal phenotype networks from segregating populations. *Genetics.* **179**: 1089-100.

Figure Legends

Figure 1. Genetic And Genomic Information Provide Useful Information in Complex Disease Studies

Information collected from various dimension can all contribute to complex diseases in various way. And the interaction among them can be used as a clue to discover factors driving the disease. Genetic mutation can lead to changes in transcript abundance and/or activity variations of gene products. Gene product carrying error information can be translated into a dysfunctional protein, while the protein interaction can be disrupted or modified. Consequentially, the homeostasis regulated at protein level can be perturbed and lead to changes in metabolite profiling that is one step closer to be responsible for development of pathological phenotypes associated with complex diseases.

Figure 2. Genetic loci controlling transcript levels and their application to causal modeling.

A. Example of cis- and trans-acting QTL of Genes Expressed in Mouse Liver.

Expression quantitative trait locus (eQTL) analysis of two genes, Ucp2 and Mc4r, was performed by examining genetic data with gene expression profiles of whole livers collected from a genetically randomized population. Ucp2 (uncoupling protein 2) locates at 107Mb on Chromosome 4, and Ucp2 in whole liver was observed with a significant cis-eQTL equals to LOD=18.1 at 115Mb on the same chromosome. Mc4r (melanocortin 4 receptor) was observed with two significant trans-eQTL on chromosomes 6 and 14 while it locates at 67Mb on chromosome 8. Linkage analysis was performed by interval mapping with WebQTL (<http://www.genenetwork.org>). The significant (p-value=0.05, pink horizontal line) and suggestive (p-value=0.63, light purple horizontal line) thresholds of QTL were defined by permutation test (n=1,000).

B. Example Patterns Employed by Causality Inference Model.

Several examples describing patterns employed in simplified causality inference model (single edge model, 1-3) and advanced causality network (4-6) were presented. Among highly correlated traits, genetic information can be leveraged to infer the direction of information flow, termed causality. (1) Causal model: a genetic variant leads to clinical

trait variability by influencing the expression of a gene. Accordingly, the gene is defined as the causal gene driving the phenotype variation in the model. (2) Reactive mode: gene expression is influenced by the clinical trait driven by a genetic factor. (3) Independent model: both gene expression and clinical trait are driven by a common genetic factor independently without influencing each other. (4) Multi-edge model: multiple genetic factors can influence the expression patterns of various genes simultaneously and lead to phenotype variability. (5) Cyclic pattern: albeit genetic effect upon transcript abundance of a gene, the interaction and tight co-regulation between various genes can mask the causal relationship between genetic factor and the gene truly regulated by genetic variants. This pattern is resolved by Chibub Neto and colleagues earlier this year by employing linkage analysis results (43), and a directed graph was reconstructed between phenotypes by analyzing metabolite profiles. (6) Directed gene network: a simplified case of causality network reconstructed based on causality inference process is shown here. The genetic markers directly impact the transcript abundance of primary causal genes that further lead to the primary phenotype variation. However, genes reactive to the primary phenotypes can lead to the variation of secondary phenotypes. Black arrows between G and primary E indicate the cis-acting manner while the blue ones indicate the trans-acting in terms of the proximity between gene and the possible regulatory mechanism encompassing genetic variants. As a cluster of variability of different phenotypes, complex diseases have a nature that requires integration of data collected from various dimension to reveal the mechanisms underlying them. Blue dots (G): genetic marker; green dots (E): gene expression patterns; orange dots (C): clinical trait variation. Light green dots (E) and pink dots (C): reactive or secondary gene expression patterns and clinical trait variations respectively.

Figure 3. Network Construction Approach Reveals Molecular Signatures Associated with Complex Diseases.

Association network comprising gene modules and edges was constructed by k-mean clustering method with gene expression profile in whole livers of a genetically randomized mouse population. Blue and orange edges represent positive and negative correlation between modules. Phenotypes related to metabolic and cardiovascular

diseases were selected in this figure to indicate that examining molecular signatures at transcript level can reveal trait interconnectedness. Purple and orange modules represent unique and common modules among relevant phenotypes in the network (Unpublished work).

Figure 1

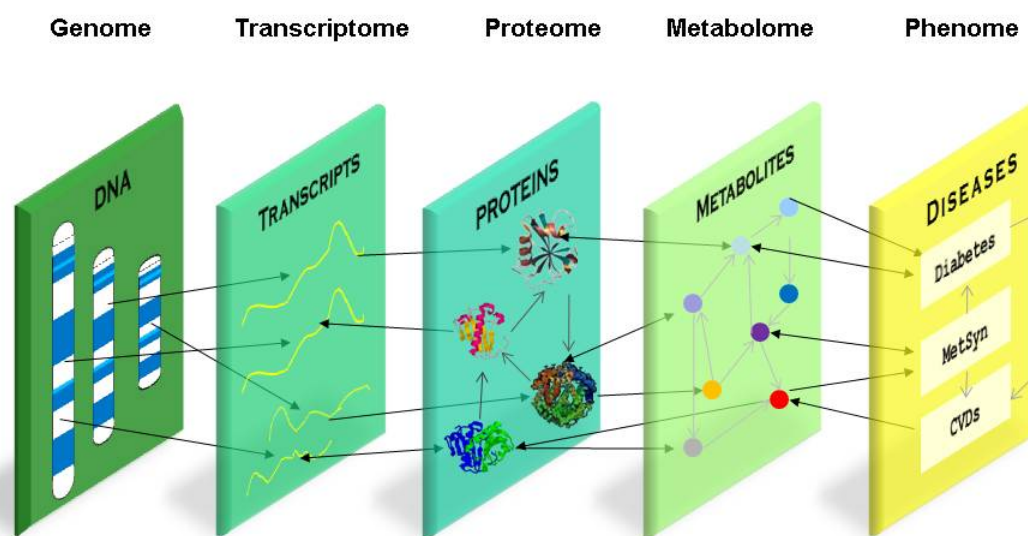


Figure 2A

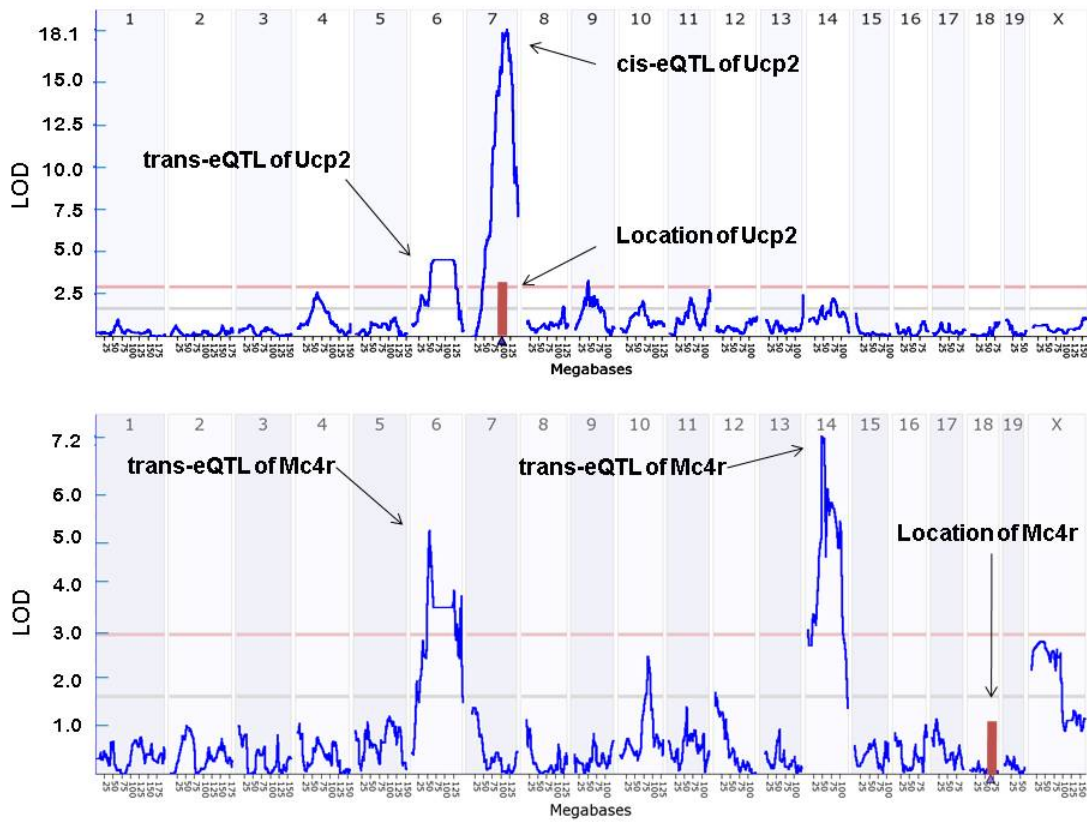


Figure 2B

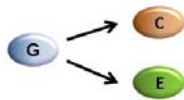
(1) Causal model



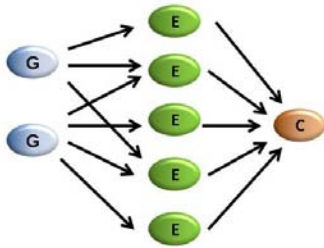
(2) Reactive model



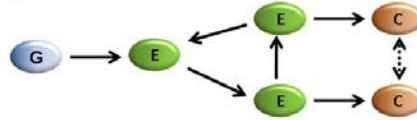
(3) Independent model



(4) Multi-edge model



(5) Cyclic Pattern



(6) Directed Gene Network

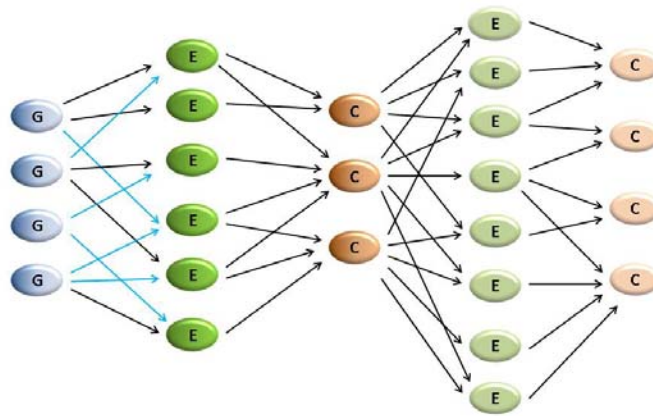


Figure 3

