NATURE|Vol 461|10 September 2009|doi:10.1038/nature08454

Molecular networks as sensors and drivers of common human diseases

Eric E. Schadt[1]

The molecular biology revolution led to an intense focus on the study of interactions between DNA, RNA and protein biosynthesis in order to develop a more comprehensive understanding of the cell. One consequence of this focus was a reduced attention to whole-system physiology, making it difficult to link molecular biology to clinical medicine. Equipped with the tools emerging from the genomics revolution, we are now in a position to link molecular states to physiological ones through the reverse engineering of molecular networks that sense DNA and environmental perturbations and, as a result, drive variations in physiological states associated with disease.

Our understanding of common human diseases and how best to treat them is hampered by the complexity of the human system in which they are manifested. Unlike simple Mendelian disorders, in which highly expressive, highly penetrant mutations make it possible to identify the causal genes within families in which traits associated with the disorders segregate[1], common human diseases originate from a more complex interplay between constellations of changes in DNA (both rare and common variations) and a broad range of factors such as diet, age, gender and exposure to environmental toxins.

These complex arrays of interacting factors are thought to affect entire network states that in turn increase or decrease the risk of disease or affect disease severity. In the context of common human diseases, the disease states can be considered emergent properties of molecular networks[2], as opposed to the core biological processes associated with a disease being driven by responses to changes in a small number of genes. Integrating large-scale, high-dimensional molecular and physiological data holds promise not only for defining the molecular networks that directly respond to genetic and environmental perturbations that associate with disease but also for causally associating such networks with the physiological states associated with disease. Given what must be considered a deluge of data of many different types flooding life sciences and biomedical research today, including genome-wide single-nucleotide polymorphism (SNP) genotyping data, whole-genome transcription data, next-generation DNA sequencing data, RNA sequencing data, chromatin immunoprecipitation (ChIP) sequencing data and image data, it is now time to begin addressing how these large-scale, high-dimensional data sets can be integrated to better understand the molecular networks underlying physiological states associated with disease. Here, I review the progress made over the past few years to integrate DNA variation, molecular profiling and clinical data collected in populations in order to construct causal probabilistic networks of disease, providing a more comprehensive view of disease than can be achieved by examining the different data dimensions on their own. Particular attention is paid to describing how the predictive networks produced from this type of integrative modelling can help link molecular states to physiological ones, providing an alternative path for understanding how molecular states drive complex disease processes.

GWAS provide insights into human diseases

Roughly three billion nucleotides make up the human genome, so the number of nucleotide changes that can affect the activities of genes is effectively infinite with respect to our ability to determine the effects of combinations of such changes experimentally. Therefore, exploiting naturally occurring DNA variation in human populations is among the most attractive approaches to inferring the constellation of genes that affect disease risk. For most diseases, changes in DNA that correlate with disease can be inferred as tagging or directly representing causal components of disease. Therefore, DNA variation directly elucidates disease aetiology and is extremely useful (Fig. 1a). Genome-wide association studies (GWAS) are now well proven to uncover genetic loci that affect disease risk or progression[3].

The emergence of technologies capable of characterizing DNA variation systematically over the entire genome and in whole populations has revolutionized our ability to apply GWAS approaches to many human diseases, with more than 200 loci now identified and highly replicated for Crohn's disease[4], type 2 diabetes[5], serum lipid levels[6,7], prostrate cancer[8,9], age-related macular degeneration[10,11], obesity[12] and more than 50 other human diseases[3]. By comparing the frequencies of genetic variants between individuals with and without disease, or by directly testing for correlations between a quantitative disease trait and genotypes at a given locus, GWAS can lead directly to the causal variants of disease or to variants that are in strong linkage disequilibrium with variants of disease. Therefore, the power of approaches such as GWAS lies in their ability to identify the genetic causes of disease, which can be used to predict disease risk and to elucidate signalling pathways associated with disease, information that is of use in drug discovery.

Integrative genomics and disease networks

GWAS have uncovered many genetic loci that associate with human diseases, but two fundamental limitations have hampered our ability to translate these results into clinically useful predictors of disease and drug targets. First, the genetic loci associated with disease generally explain very little of the disease risk. The odds of having a risk genotype at a particular disease locus given that you have the disease, divided by the odds of having a risk genotype given that you do not have the disease, are typically less than 1.5 (ref. 3). Second, the SNP–trait associations alone do not necessarily lead directly to the identification of the causal gene(s), much less elucidate the context in which the causal gene(s) operates[3,13,14]. Understanding the biological context in which a given causal gene for disease operates is a necessary step in identifying the best drug targets[15,16].

[1]Pacific Biosciences, 1505 Adams Drive, Menlo Park, California 94025, USA.

Interestingly, in the span of just a few years, the realization that tractable drug targets and clinically useful biomarkers of disease are not immediately apparent from GWAS data has, for some, reduced enthusiasm for the GWAS approach[17–19]. However, given that variations in DNA do not on their own directly impact on physiological states associated with disease, there is the potential to enhance our understanding of GWAS data by layering in a hierarchy of phenotypes that define the molecular and physiological states associated with disease[13,14,20–22]. Because variations in DNA more proximally (relative to disease states) induce changes in molecular states that in turn drive variations in physiological states associated with disease, incorporating such data can allow the identification of causal genes and the broader biological context in which they operate. Therefore, elucidating changes in molecular states that more directly respond to changes in DNA and that in turn influence disease has the potential to fill in the gaps left by GWAS.

In fact, the advances made in mapping DNA loci for diseases have occurred simultaneously with the mapping of DNA loci for molecular traits such as transcript abundances[13,14,20,22–24]. Identifying the RNAs that mediate the flow of information from DNA to disease is of particular interest in this context, given that, because it is transcribed directly from a DNA template, RNA is the most proximal non-DNA species of all molecular entities in the cell. In studies that seek to map genetic loci that affect RNA levels, SNP genotypes are tested for association with tens of thousands of RNA traits scored simultaneously in population samples. A number of such studies have demonstrated that the amount of variation in RNA levels explained by a given genetic locus can often be greater than 50% (refs 13, 14, 22 and 24). In addition, family-based studies of the genetics of RNA levels in multiple tissues have estimated that a majority of RNA traits on average have a genetic variance component of 30% (ref. 13). The mapping of genetic loci for molecular traits is not constrained only to RNA levels. Any molecular species that can be reasonably well measured (for example protein or metabolite levels) is amenable to genetic mapping and can complement genetic mapping for RNA traits[25]. Mapping studies involving RNA traits are not without significant analysis issues. The large number of RNA traits and markers that can be tested demands that significance levels for association be rigorously adjusted to control for false-discovery rates[14].

Molecular traits controlled by genetic loci associated with disease can be treated as intermediate phenotypes of disease and thus elucidate the molecular networks underlying disease. This can aid in the interpretation of GWAS data by identifying genes whose RNA levels associate with genetic loci that also associate with disease[6,13,14,20,26,27]. Furthermore, these data can be treated more formally to infer causal relationships between molecular traits and disease states[2,21,28,29], a process that has been shown to aid in the identification of genes or specific isoforms of genes corresponding to loci identified in the GWAS[14,20,30] (Fig. 1b). One of the central issues related to the use of RNA traits to enhance identification of genes in genomic regions associated with disease is assessing whether a given locus is jointly associated with disease and RNA levels, or whether two closely linked loci control the RNA levels and disease independently[14,21]. Formal statistical procedures that examine the joint probabilities for the genotype, RNA and disease data can be applied to establish whether RNA levels and disease are related in either an independent relationship or a causal or reactive relationship[2,21,28,29].

The introduction of molecular traits can enhance the interpretation of GWAS results by placing them in a broader biological context that may support the identification of disease-susceptibility genes and more generally elucidate networks (Box 1) that define the biological processes associated with disease[14]. One of the more intriguing examples of this approach was the identification of three candidate susceptibility genes (*SORT1*, *CELSR2* and *PSRC1*) for cardiovascular disease and lipid levels[2,7], where the disease-associated and lipid-associated SNPs were also significantly associated with the liver expression of the three candidate genes, which were physically located near the disease-associated SNP. These genes were also supported as causal for low-density-lipoprotein cholesterol levels in a previously described experimental mouse cross[2]. Furthermore, all three genes were found to be connected in liver gene networks that were constructed from mouse and human liver samples and in which the constituent genes were enriched for in a previously described macrophage-enriched metabolic network associated with a number of processes related to immune function and inflammation[2,13,14].

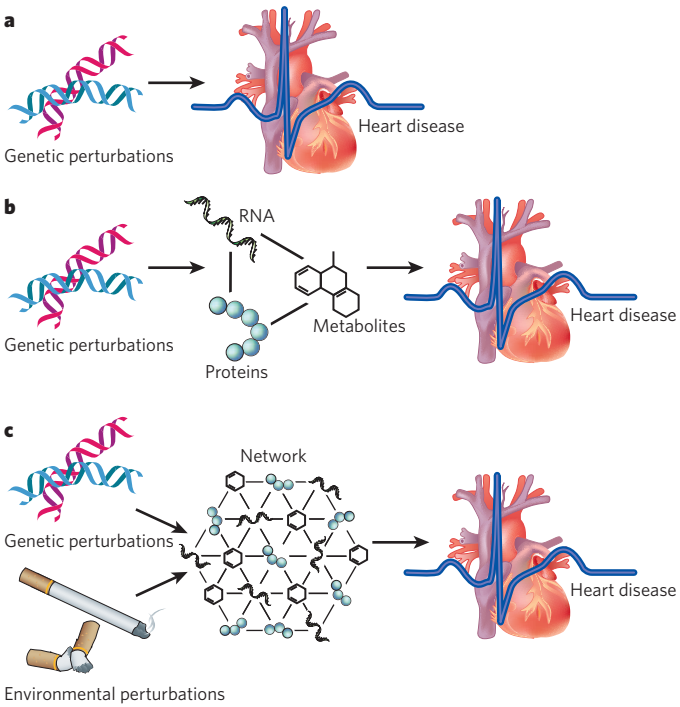

Figure 1 | Hierarchy of causal relationships. a, Classic genetic association approaches seek to identify variations in DNA that correlate with disease state or with quantitative traits associated with disease. The attraction of this approach is the identification of the genetic causes of disease. **b**, Changes in DNA on their own do not lead to disease but, instead, lead to changes in molecular traits that go on to affect disease risk. By layering in molecular phenotypes as intermediate phenotypes, causal relationships between genes and disease can be established directly. **c**, Disease gene networks sense constellations of genetic and environmental perturbations. Therefore, a more realistic model is one in which constellations of genetic and environmental perturbations affect molecular states of networks that in turn affect disease risk.

Disease networks respond to disease loci

Identifying genetic loci that associate with disease and intermediate molecular phenotypes that respond more proximally to these loci and in turn cause disease are excellent first steps to uncovering the drivers of disease. However, the view of disease becoming clear from the large-scale genomic studies is that common forms of disease are emergent properties of networks whose states are affected by a complex interaction of genetic and environmental factors. To understand the behaviour of any one gene in the context of human disease, individual genes must be understood in the context of molecular networks that define the disease states. In fact, several studies have now shown that for single diseases or traits such as height, tens or even hundreds of genes may be involved but may not be randomly distributed with respect to biological function.

For example, sequencing of DNA from tumour samples found scores of genes affected by rare variations that influence cancer risk and progression. The genes affected were shown to be significantly more likely to belong to pathways known to be involved in tumorigenesis or tumour progression than was the case for the set of all genes that were resequenced as part of this study[31,32]. In a separate study, my research group identified a macrophage-enriched metabolic network (MEMN) that in mice was strongly indicated to be causal for a number of metabolic-disease traits[2]. The same network was not only found to be associated with metabolic traits and conserved in human populations but also to be enriched for DNA variations near

these genes that are associated with obesity, suggesting that hundreds or thousands of genes may subtly affect obesity risk[13]. Constructing networks that underlie core biological processes associated with disease makes it possible to identify the functional units that respond to genetic perturbations and then in turn affect disease risk (Fig. 1c). In this way, any given gene can be studied in the context of many different networks to learn whether one or more of the networks in which a given gene operates influences physiological states associated with the disease. Such mappings not only allow the identification of causal relationships among genes and between genes and more complex traits such as disease[2,21,29] but also more generally allow the construction of predictive gene networks[2,33].

Before this can be achieved, however, we must integrate the diverse data necessary to construct the gene networks. There have been a number of recent advances in the construction of networks capable of predicting complex system behaviour. Examining the action of many genes simultaneously in populations segregating common disease traits has led to the identification of whole gene networks that both define disease at the molecular level and drive the onset and progression of disease[2,13,14,31–35]. The construction of these networks allows the identification of the functional units of the system underlying physiological states[2,29,34,36,37].

Networks generally provide a convenient framework for exploring the context within which single genes operate (Box 1). Networks are simply graphical models that comprise nodes and edges and are convenient for visualizing complex mathematical models that describe how variables of a system associate with one another in different contexts of interest. For gene networks associated with biological systems, the nodes in the network typically represent genes, gene products or other important molecular entities, and an edge between any two nodes indicates a relationship between the corresponding genes, gene products or other molecular entities. For example, an edge between two genes may indicate that the corresponding expression traits are correlated[33,38], that the corresponding proteins interact[39] or that changes in the activity of one gene lead to changes in the activity of the other[21]. Interaction, or association, networks, which have recently become widely used in the biological community, are formed by considering only pairwise relationships between genes, including protein interactions[40] and co-expression relationships[37,41].

Interaction networks allow the identification of subnetworks (coherent gene modules) corresponding to the functional units of a living system[2,29,36,37,42,43]. Increasing evidence suggests that these functional units are directly linked to physiological states, defining in humans the molecular states that lead to physiological states associated with disease. Genetic perturbations that associate with disease have been shown to act through these functional units by altering the corresponding network state. The networks therefore can serve as an organizing framework for causal perturbations that lead to disease. That is, networks sense variations in the genome, in the methylome and in the environment more generally, given that these different types of variation affect the function of the proteins or the expression levels of the genes or proteins constituting these networks, thus altering their states. In this way, the network more maximally captures, or senses, these different sources of variation and, as a result, induces changes in physiological states associated with disease (Fig. 1c).

Although there is now an extensive literature on the construction and application of interaction networks to elucidate the complexity of disease, these methods are typically applied to gene expression data alone and therefore do not strictly reflect causal relationships among gene expression traits or between expression traits and disease. Probabilistic causal networks represent an alternative approach capable of integrating multiple types of data and inferring from these data whether two or more genes are causally connected to each other or to disease traits. Bayesian network-reconstruction methods are one of the more common approaches of this sort. They provide an elegant way of incorporating diverse data pertaining to causal relationships, such as DNA variation, gene expression, protein interaction, DNA–protein binding, and proteomic and, more recently, metabolomic data. Recent work has demonstrated that by considering these types of data simultaneously, it is possible to construct networks that are able to predict future states of the representative system[33,44]. The construction of networks in which the relationships between genes can be understood from the standpoint of causal control is one of the ultimate aims in life sciences and biomedical research, as an understanding of predictive gene networks can lead directly to drug targets and biomarkers of disease[15,16,45].

The MEMN is an example of a causal network constructed by integrating different data types. The MEMN was identified from liver and adipose gene expression data generated in mouse and human populations segregating metabolic-disease phenotypes. From the resultant tissue gene networks, the MEMN was identified as strongly conserved between tissues, between sexes and between species, and was strongly associated with metabolic traits related to obesity, diabetes and heart disease[2,13]. It was also observed to respond to variations in DNA that are associated with disease traits[13]. A statistical procedure[21] was applied to infer whether the MEMN was responding to the DNA changes and causing variations in the metabolic traits as a result or whether it was responding to changes in the metabolic traits induced by the DNA changes. The MEMN was strongly indicated to be causal for all of the obesity, diabetes and heart-disease traits scored in an experimental mouse population.

Biological processes represented in the MEMN supported the idea of macrophages as a key driver of disease pathogenesis, consistent with recent evidence that chronic inflammation is a key feature of obesity[2,46]. Importantly, the mouse MEMN was highly conserved in humans, in

Box 1 | Gene networks

Cells comprise many tens of thousands of proteins, metabolites, RNAs and DNAs, all interacting in complex ways. In turn, complex biological systems comprise many types of cell operating within and between the many types of tissue that make up different organ systems, all of which interact in complex ways to give rise to a vast array of phenotypes that manifest themselves in living systems. Modelling the extent of such relationships between molecular entities, between cells, and between organ systems is a daunting task. Networks are a convenient framework in which to represent the relationships among these different variables. In the context of biological systems, a network can be viewed as a graphical model that represents relationships among DNAs, RNAs, proteins, metabolites and higher-order phenotypes such as disease state. In this way, networks provide a way to visualize extremely large-scale, complex relationships among molecular and higher-order phenotypes in any given context. In this Review, I am interested in networks that represent relationships among molecular entities in a living system, as determined empirically in populations of individuals.

In this context, biological networks comprise nodes, which represent molecular entities that are observed to vary in the population under study (for example DNA variations, RNA levels, protein states or metabolite levels). Edges between the nodes represent relationships between the molecular entities, and these edges can either be directed, indicating a cause–effect relationship, or undirected, indicating an association or interaction. For example, a DNA node in the network representing a given locus that varies in a population of interest may be connected to a transcript-abundance trait, indicating that changes at the particular DNA locus induce changes in the levels of the transcript. The potentially millions of such relationships represented in a network define the overall connectivity structure, or topology, of the network. Any realistic network topology will necessarily be complicated and nonlinear from the standpoint of the more classic biochemical pathway diagrams presented in text books and pathway databases such as the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway database[54]. The more classic pathway view represents molecular processes on an individual level, whereas networks represent global (population-level) metrics describing variations between individuals in a population of interest; these variations in turn define the coherent biological processes in the tissue or cells associated with the network.

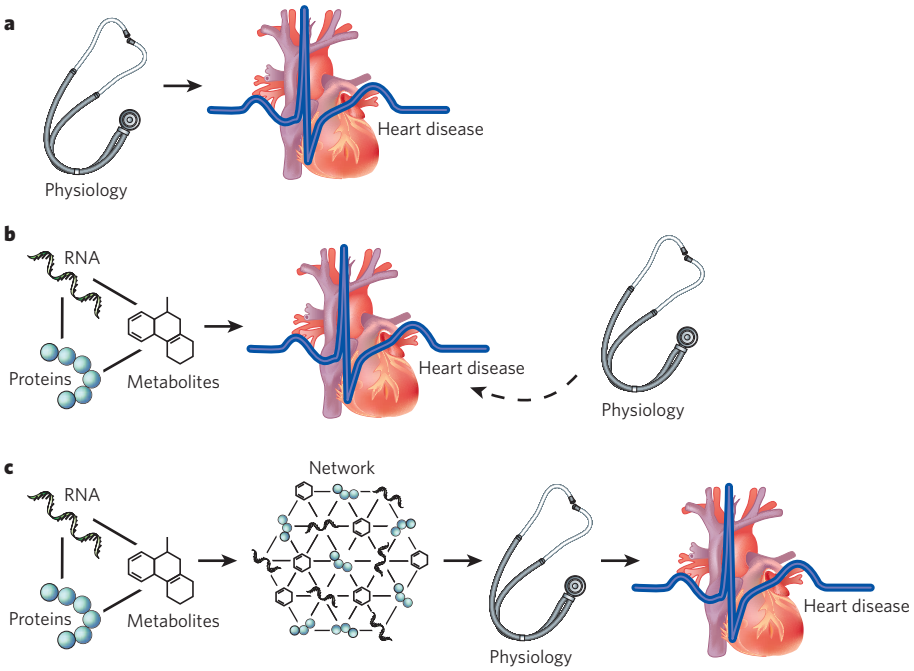

Figure 2 | Linking molecular biology to physiology through molecular networks. **a**, Before the molecular biology revolution, disease was studied primarily in the context of physiology. **b**, As a result of the molecular biology revolution, physiology has played a less prominent role in the study of the molecular bases of disease, given the reductionist push to associate molecular changes in a given gene (affecting protein levels, activity or function) directly with changes in disease states. **c**, The complexity of molecular biology — given the ability to monitor DNA variation, RNA variation, metabolite variation and protein variation in populations on a comprehensive scale — has driven a systems view of disease, in which networks of interacting molecular entities are constructed to define physiological states of the system associated with disease. In this way, the molecular networks allow a direct link between molecular biology and clinical medicine by connecting molecular biology to physiology.

whom it was also indicated to be causal for metabolic traits. A number of genes in the MEMN were predicted to be causal for metabolic-disease traits. This has now been experimentally verified, and the genes have been shown to be involved in complex feedback control, with many of them indicated and confirmed to be causal for each other[2,14,28,29].

Linking molecular and physiological states

The identification of the MEMN as a key driver of metabolic disease highlights several important features of the network approach to understanding disease that have implications for drug discovery: first, the network analyses revealed hundreds of disease-causing genes acting together in coherent networks; second, within a given network supported as being causal for disease, perturbing individual genes supported as being causal for disease affected the state of the network; and, third, DNA and other sources of variation in one species can be used to construct disease networks that are relevant in a second species and that act as sensors for many sources of variation (for example genetic, epigenetic and environmental sources) and in turn modulate physiological traits associated with disease. These features taken together suggest that networks such as the MEMN underlie or define the physiological states associated with disease. The data further suggest that highly efficacious treatments of diseases such as obesity might not be achieved by targeting single genes, at least not without taking into account the role of an individual gene in the network[15,16].

Core subnetworks associated with disease provide a path directly linking molecular biology to physiology, and it is this link that may ultimately lead to a more significant clinical impact (Fig. 2). Networks have now been modelled both within and between multiple tissues that are relevant to disease. The identification of subnetworks interacting between islet, adipose, liver, muscle and brain tissues has highlighted the importance of using a network framework directly to model physiological states associated with diabetes[34]. One of the most recent studies[42] in modelling cross-tissue networks highlighted coherent subnetworks that were not part of any of the single-tissue networks but, instead, specific to cross-tissue interactions, showing that modelling molecular interactions operating between tissues is critical if we hope to understand physiological states associated with disease.

Whereas classic molecular biology provided very narrow views connecting molecular entities to disease, today's technologies allow the generation of comprehensive snapshots of living systems, which in turn allows a more systems-level view of the molecular states underlying physiological states associated with disease. In single experiments, we can now generate terabytes of genotype, sequence, gene expression, physiological and imaging data. The degree to which any one of these different data types informs our view of disease may vary, but these data types provide complementary views that are useful individually and potentially exceptionally valuable when considered collectively.

Disease-associated networks such as the MEMN comprise hundreds of genes interacting in complex ways that collectively associate with physiological states such as fat mass, insulin levels and atherosclerotic-lesion size. Such networks may be indicated to cause variations in disease-associated traits and can also respond to (or sense) genetic and environmental variations that influence disease risk. For example, the MEMN was demonstrated to respond to a wide range of DNA variations in genes distributed throughout the genome and also responded to environmental perturbations such as changes in diet. For mice placed on a high-fat diet, more than 40% of the RNA traits that changed relative to those of mice on a normal, chow diet were concentrated in the MEMN (the probability of this overlap occurring by chance was computed to be $<10^{-200}$).

Perspectives

The disease-associated molecular networks that we can construct today are necessarily based on grossly incomplete sets of data. Even given the ability to assay DNA and RNA variation in whole populations in a comprehensive manner, the information is not complete, because we are far from completely characterizing rare variation, DNA variation other than SNP and copy number, variation in non-coding RNA levels and variation in the different isoforms of genes in any sample, much less in entire populations. Beyond DNA and RNA, it is not possible with existing technologies to measure all protein-associated traits or all the interactions between proteins and DNA/RNA, metabolite levels and other molecular entities important to the functioning of living systems. Furthermore, the types of high-dimensional data we are able to generate routinely today in populations represent only a snapshot at a single time point, which may allow the identification of the functional units of the system under study and how these units relate to one another but does not allow a complete understanding of how the functional units are put together or the mechanistic underpinnings of the complex set of functions carried out by individual cells, by entire organs and by whole systems comprising multiple organs.

Technological advances, however, allow the generation of increasingly higher dimensional data, so we continue to progress towards a

more complete understanding of human disease. The next-generation sequencing technologies are already having a major impact on DNA sequencing, identifying rare variations in tumour tissues associated with different cancer types[31,32]. In addition, subsequent generations of sequencing technologies are on the horizon and promise to deliver the sequence of entire human genomes in days and at a reasonable cost[47]. Sequencing technologies can also be used to identify patterns of methylation[48], to fully characterize the transcriptome[49] and to identify transcripts that are being actively translated[50]. The advances of the sequencing revolution therefore stand ready to provide unprecedented snapshots of complex systems that will allow a more accurate network view, which in turn will lead to models of disease that have greater predictive power.

One area in need of development regarding network-based approaches centres on the interpretation of high-dimensional data from which complex relationships and mathematical models are derived. The genomics field generally has been plagued by examples in which high-dimensional data have resulted in an unacceptably high rate of false positives. One striking example of this is a study that was undertaken to replicate published associations between 85 DNA variants and acute coronary syndromes. Of the 85 variants tested, only 1 gave rise to a nominally significant *P* value, highlighting a complete lack of support for the hypothesis that any of the variants previously reported in scores of publications as associating with acute coronary syndromes truly did so[51]. This problem is exacerbated when linking genotypes scored on hundreds of thousands of markers with tens of thousands of molecular phenotypes. Furthermore, understanding how to validate the accuracy of network models, how to compare networks across multiple conditions, species and methods, and, importantly, how to enable researchers to benefit from these models, which they may not fully understand, are among the most pressing problems to address if we are to move forwards. These issues are beginning to be addressed[44], and efforts such as the Dialogue for Reverse Engineering Assessments and Methods are making rapid progress in catalysing the type of interaction needed between experiment and theory to assess the accuracy of biological networks[52].

Ultimately, our ability to construct predictive disease models will depend on our mastering the large-scale information being collected on systems relevant to disease. To accomplish this, data sharing must be more open, not only within industry but also within academic communities, where strong incentives to restrict data distribution exist to maintain competitive advantages. In addition, the development of tools and software platforms that allow the integration of large-scale, diverse data sets into complex models that can then be operated upon and refined by experimentalists in an iterative fashion is perhaps the most critical milestone we must reach in the biological sciences if large-scale data and results are to impact on biological research routinely at all levels.

The primary aims of generating and mining large-scale biological data sets are to learn the fundamental rules that govern complex living systems and to derive, as a result, predictive models of their behaviour. Without sophisticated mathematical algorithms capable of appropriately integrating the large-scale data, and without high-performance computing environments in which to apply these algorithms, it will be difficult to build generally predictive models. Information-systems support services will become increasingly critical both for building predictive models and for representing complex states of knowledge and making such knowledge accessible to researchers so that they may refine and correct the models of disease. Recent successes in programming machines to mine complex data to derive the fundamental laws of motion[53] perhaps represent a glimpse into the future of biology, in which machines may be able to derive fundamental rules in complex living systems, given large-scale data sets. The complexity of disease mechanisms must be recognized with investments in research directed towards these types of approach, which take a more holistic view in identifying the molecular networks that underlie physiological states associated with disease. Although systems approaches are still in their infancy, as a matter of necessity they will be viewed more and more as a crucial step towards an understanding of complex biological processes such as disease. ■

1. McKusick, V. A. *Mendelian Inheritance in Man: A Catalog of Human Genes and Genetic Disorders* (Johns Hopkins Univ. Press, 1998).
2. Chen, Y. *et al.* Variations in DNA elucidate molecular networks that cause disease. *Nature* **452,** 429–435 (2008).
This paper was the first demonstration that coherent networks of genes respond to genetic and environmental perturbations and in turn influence disease-associated traits, directly showing that common forms of disease are probably emergent properties of networks rather than the result of single gene changes.
3. Altshuler, D., Daly, M. J. & Lander, E. S. Genetic mapping in human disease. *Science* **322,** 881–888 (2008).
4. Barrett, J. C. *et al.* Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nature Genet.* **40,** 955–962 (2008).
5. Zeggini, E. *et al.* Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nature Genet.* **40,** 638–645 (2008).
6. Kathiresan, S. *et al.* Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. *Nature Genet.* **40,** 189–197 (2008).
7. Willer, C. J. *et al.* Newly identified loci that influence lipid concentrations and risk of coronary artery disease. *Nature Genet.* **40,** 161–169 (2008).
8. Haiman, C. A. *et al.* A common genetic risk factor for colorectal and prostate cancer. *Nature Genet.* **39,** 954–956 (2007).
9. Haiman, C. A. *et al.* Multiple regions within 8q24 independently affect risk for prostate cancer. *Nature Genet.* **39,** 638–644 (2007).
10. Li, M. *et al. CFH* haplotypes without the Y402H coding variant show strong association with susceptibility to age-related macular degeneration. *Nature Genet.* **38,** 1049–1054 (2006).
11. Maller, J. *et al.* Common variation in three genes, including a noncoding variant in *CFH*, strongly influences risk of age-related macular degeneration. *Nature Genet.* **38,** 1055–1059 (2006).
12. Thorleifsson, G. *et al.* Genome-wide association yields new sequence variants at seven loci that associate with measures of obesity. *Nature Genet.* **41,** 18–24 (2009).
13. Emilsson, V. *et al.* Genetics of gene expression and its effect on disease. *Nature* **452,** 423–428 (2008).
This paper is a confirmation in a human population that common diseases like obesity are the result of complex molecular networks responding to genetic and environmental perturbations.
14. Schadt, E. E. *et al.* Mapping the genetic architecture of gene expression in human liver. *PLoS Biol.* **6,** e107 (2008).
15. Lum, P. Y., Derry, J. M. & Schadt, E. E. Integrative genomics and drug development. *Pharmacogenomics* **10,** 203–212 (2009).
16. Schadt, E. E., Friend, S. H. & Shaywitz, D. A. A network view of disease and compound screening. *Nature Rev. Drug Discov.* **8,** 286–295 (2009).
17. Goldstein, D. B. Common genetic variation and human traits. *N. Engl. J. Med.* **360,** 1696–1698 (2009).
18. Hardy, J. & Singleton, A. Genomewide association studies and human disease. *N. Engl. J. Med.* **360,** 1759–1768 (2009).
19. Kraft, P. & Hunter, D. J. Genetic risk prediction — are we there yet? *N. Engl. J. Med.* **360,** 1701–1703 (2009).
20. Moffatt, M. F. *et al.* Genetic variants regulating *ORMDL3* expression contribute to the risk of childhood asthma. *Nature* **448,** 470–473 (2007).
This was among the first studies to identify a disease-susceptibility gene by restricting attention to DNA variants that simultaneously associate with the disease and the expression levels of genes in the neighbourhood of the disease-associated variant.
21. Schadt, E. E. *et al.* An integrative genomics approach to infer causal associations between gene expression and disease. *Nature Genet.* **37,** 710–717 (2005).
This was the first study to demonstrate that causal relationships between molecular-profiling traits (such as gene expression) and disease traits could be systematically inferred by integrating these data with genotypic data in human and experimental populations.
22. Schadt, E. E. *et al.* Genetics of gene expression surveyed in maize, mouse and man. *Nature* **422,** 297–302 (2003).
23. Monks, S. A. *et al.* Genetic inheritance of gene expression in human cell lines. *Am. J. Hum. Genet.* **75,** 1094–1105 (2004).
24. Morley, M. *et al.* Genetic analysis of genome-wide variation in human gene expression. *Nature* **430,** 743–747 (2004).
25. Foss, E. J. *et al.* Genetic basis of proteome variation in yeast. *Nature Genet.* **39,** 1369–1375 (2007).
26. Fraser, H. B. & Xie, X. Common polymorphic transcript variation in human disease. *Genome Res.* **19,** 567–575 (2009).
27. Smirnov, D. A., Morley, M., Shin, E., Spielman, R. S. & Cheung, V. G. Genetic analysis of radiation-induced changes in human gene expression. *Nature* **459,** 587–591 (2009).
28. Mehrabian, M. *et al.* Integrating genotypic and expression data in a segregating mouse population to identify 5-lipoxygenase as a susceptibility gene for obesity and bone traits. *Nature Genet.* **37,** 1224–1233 (2005).
29. Yang, X. *et al.* Validation of candidate causal genes for abdominal obesity that affect shared metabolic pathways and networks. *Nature Genet.* **41,** 415–423 (2009).
30. Goldstein, D. B. Genomics and biology come together to fight HIV. *PLoS Biol.* **6,** e76 (2008).
31. Jones, S. *et al.* Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science* **321,** 1801–1806 (2008).
32. Parsons, D. W. *et al.* An integrated genomic analysis of human glioblastoma multiforme. *Science* **321,** 1807–1812 (2008).
33. Zhu, J. *et al.* Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. *Nature Genet.* **40,** 854–861 (2008).
This paper generalizes the early idea of integrating gene expression and DNA-variation data to infer causal relationships among gene expression traits and between gene

expression and disease traits by integrating diverse types of data, including genotype, gene expression, protein-interaction and DNA–protein-binding data.
34. Keller, M. P. *et al.* A gene expression network model of type 2 diabetes links cell cycle regulation in islets with diabetes susceptibility. *Genome Res.* **18,** 706–716 (2008).
35. Meng, H. *et al.* Identification of *Abcc6* as the major causal gene for dystrophic cardiac calcification in mice through integrative genomics. *Proc. Natl Acad. Sci. USA* **104,** 4530–4535 (2007).
36. Ghazalpour, A. *et al.* Genomic analysis of metabolic pathway gene expression in mice. *Genome Biol.* **6,** R59 (2005).
37. Ghazalpour, A. *et al.* Integrating genetic and network analysis to characterize genes related to mouse weight. *PLoS Genet.* **2,** e130 (2006).
38. Zhu, J. *et al.* An integrative genomics approach to the reconstruction of gene networks in segregating populations. *Cytogenet. Genome Res.* **105,** 363–374 (2004).
39. Rual, J. F. *et al.* Towards a proteome-scale map of the human protein–protein interaction network. *Nature* **437,** 1173–1178 (2005).
40. Han, J. D. *et al.* Evidence for dynamically organized modularity in the yeast protein–protein interaction network. *Nature* **430,** 88–93 (2004).
41. Gargalovic, P. S. *et al.* Identification of inflammatory gene modules based on variations of human endothelial cell responses to oxidized lipids. *Proc. Natl Acad. Sci. USA* **103,** 12741–12746 (2006).
42. Dobrin, R. *et al.* Multi-tissue coexpression networks reveal unexpected subnetworks associated with disease. *Genome Biol.* **10,** R55 (2009).
43. Pe'er, D., Regev, A., Elidan, G. & Friedman, N. Inferring subnetworks from perturbed expression profiles. *Bioinformatics* **17** (suppl. 1), S215–S224 (2001).
44. Zhu, J. *et al.* Increasing the power to detect causal associations by combining genotypic and expression data in segregating populations. *PLOS Comput. Biol.* **3,** e69 (2007).
45. Schadt, E. E., Sachs, A. & Friend, S. Embracing complexity, inching closer to reality. *Sci. STKE* **2005,** pe40 (2005).
46. Zeyda, M. & Stulnig, T. M. Adipose tissue macrophages. *Immunol. Lett.* **112,** 61–67 (2007).
47. Eid, J. *et al.* Real-time DNA sequencing from single polymerase molecules. *Science* **323,** 133–138 (2009).
48. Cokus, S. J. *et al.* Shotgun bisulphite sequencing of the *Arabidopsis* genome reveals DNA methylation patterning. *Nature* **452,** 215–219 (2008).
49. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Rev. Genet.* **10,** 57–63 (2009).
50. Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. & Weissman, J. S. Genome-wide analysis *in vivo* of translation with nucleotide resolution using ribosome profiling. *Science* **324,** 218–223 (2009).
51. Morgan, T. M., Krumholz, H. M., Lifton, R. P. & Spertus, J. A. Nonvalidation of reported genetic risk factors for acute coronary syndrome in a large-scale replication study. *J. Am. Med. Assoc.* **297,** 1551–1561 (2007).
52. Stolovitsky, G. & Califano, A. (eds). *Reverse Engineering Biological Networks: Opportunities and Challenges in Computational Methods for Pathway Inference* (Wiley, 2007).
53. Schmidt, M. & Lipson, H. Distilling free-form natural laws from experimental data. *Science* **324,** 81–85 (2009).
54. Bock, G. & Goode, J. A. (eds). 'In Silico' *Simulation of Biological Processes* 91–103; 119–128; 244–252 (Wiley, 2002).

Author Information Reprints and permissions information is available at www.nature.com/reprints. The author declares competing financial interests: details accompany the full-text HTML version of the paper at www.nature.com/nature. Correspondence should be addressed to E.E.S. (eschadt@pacificbiosciences.com).