Published in final edited form as:
Science. 2008 April 25; 320(5875): 495–497. doi:10.1126/science.1153716.

From Genotype to Phenotype: Systems Biology Meets Natural Variation

Philip N. Benfey[1,2,*] and **Thomas Mitchell-Olds**[1]

[1]Department of Biology, Duke University, Durham, NC 27708, USA.

[2]Institute for Genome Sciences and Policy–Center for Systems Biology, Duke University, Durham, NC 27708, USA.

Abstract

The promise that came with genome sequencing was that we would soon know what genes do, particularly genes involved in human diseases and those of importance to agriculture. We now have the full genomic sequence of human, chimpanzee, mouse, chicken, dog, worm, fly, rice, and cress, as well as those for a wide variety of other species, and yet we still have a lot of trouble figuring out what genes do. Mapping genes to their function is called the “genotype-to-phenotype problem,” where phenotype is whatever is changed in the organism when a gene’s function is altered.

Substantial progress in identifying gene function has been made. Studying the effects of modifying individual genes in model organisms such as *Drosophila*, *Caenorhabditis*, and *Arabidopsis* has allowed several thousand genes to be associated with phenotypes. Through similarities in the encoded protein sequence, we have also managed to identify the general function of many genes, classifying them as enzymes, receptors, transcription factors, and so forth. Another informative approach has been to compare genes descended from the same ancestor across many different organisms. In bacteria, this comparative genomics approach has been used to map genes shared among organisms that have similar phenotypes, resulting in the assignment of putative function to these genes (1). And yet we still do not know the function of a large number of the genes in either plants or animals, and we still cannot predict with any accuracy what the effect will be of modifying the activity of an uncharacterized gene, even when it has been assigned to a functional class. (Indeed, natural selection may act on effects, which are too subtle to be identified by experimental manipulations; hence, it may be impossible to determine the function of some genes.) Equally daunting is starting with a phenotypic variant and trying to predict what genes are likely to be involved. The problem is complicated by the fact that most phenotypes of medical or agricultural interest are “complex,” which means that more than one gene, in addition to environmental factors, contributes to expression of the phenotype. Not that single-gene traits are necessarily uninteresting for medicine or agriculture, but these were easier for geneticists to decipher. Now we are left with multigene traits that are harder to work out.

The difficulty in mapping genotype to phenotype can be traced to several causes, including inadequate description of phenotypes, too little data on genotypes, and the underlying complexity of the networks that regulate cellular functions. Recent technical advances for acquiring genome-wide data hold promise for improvements in genotyping and phenotyping. It is particularly exciting to contemplate the application of these advances to the myriad of interesting phenotypes found in nature. This natural variation is generated by additive and

[*]To whom correspondence should be addressed. philip.benfey@duke.edu.

epistatic effects of alleles across multiple genes, resulting in many individuals with phenotypes near the population mean, and a minority showing extreme phenotypes. Some combinations result in enhanced traits, whereas other combinations are deleterious to fitness in specific environments. Phenotypic alterations are usually in matters of amount, rather than in the presence or absence of a trait. The field of statistical genetics has developed sophisticated tools to map such quantitative traits to regions of chromosomes. The chromosomal regions are known as quantitative trait loci (QTLs) and are described in terms of the percentage of the variation of a trait that can be attributed to each region.

What has been generally missing is the context in which to place these percentages associated with QTLs. What does it mean, at the cellular or molecular level, that a particular allelic polymorphism has a large or small effect on a trait? This is where the complexity of the underlying cellular networks comes into play. Until recently, most molecular processes occurring within cells were described in terms of linear pathways. A signal received by a cell would be transmitted by a linear series of molecular interactions, ultimately resulting in a response such as a change in gene expression. The field of systems biology is expanding this view, replacing the linear pathways with interconnected networks. These networks frequently look like the “hub-and-spoke” configurations of airline routes. When viewed from the perspective of a network in which there are preferred and alternative routes, the magnitudes associated with quantitative trait loci take on new meaning. Because of the hub-and-spoke organization of the major airline routes, a snowstorm in Chicago can result in disruption of 35% of transcontinental air traffic, whereas a snowstorm in Des Moines might only cause a 2% change.

This analogy illustrates another way in which systems biology is changing the way we think about biological processes. The relative importance of the different cities is a function of the dynamics of transcontinental air traffic, not of the cities’ intrinsic size or location. A city that is central for one airline’s network is frequently peripheral for another airline’s network. Although the dynamics of metabolic networks have been studied for some time, it is only recently that the dynamics of signaling and transcriptional networks have come under scrutiny. To study the dynamics of a system requires perturbing it and then observing how it reacts to the perturbation. One way of perturbing a biological system is by changing the external stimuli that it perceives. A culture of bacteria can be given a new source of carbon, or a plant can be transferred from dark to light conditions. Alternatively, the genome can be altered and the effects observed. In traditional genetics, a primary goal is to knock out the activity of individual genes and assess the effects on the organism. From a network perspective, the major disadvantage of this approach is that it is often difficult to infer the normal functioning of the system from disruptions that completely remove a gene. Although perturbations with less drastic effects can be identified through traditional genetics (2), they are the norm among the alleles that contribute to natural variation. In the past, this has been considered a disadvantage of natural variation: that the genetic variation occurs at multiple loci, each making only a small contribution to the complex trait. However, for understanding the dynamics of a system, these smaller dispersed effects can become a major advantage. Linking genetic changes to small perturbations in the network may allow us to understand how tuning of the network can produce different outcomes (Fig. 1).

It sounds great in theory, but there are issues to be reckoned with before we can bring the insights of systems biology to bear on natural variation and vice versa. In outbred populations such as humans or many wild plants, the variance attributable to each polymorphic locus is influenced by two factors: how frequently the allele appears in the population, and what the allele does to an individual. Also, the genetic loci that contribute variation to a given phenotypic trait may vary from one population to another; hence, quantitative genetic analyses always are specific to a given reference population. Finally, when experiments have limited statistical power to detect QTLs, then the loci that achieve statistical significance will vary from one experiment to another, on a random basis. Ultimately, however, identifying and analyzing the loci that interact to give rise to natural variation will allow us to better understand how networks give rise to phenotypes.

Technologies either exist or are visible on the horizon that are likely to make natural variation accessible to systems biology approaches. Methods for sequencing DNA have become much faster and cheaper. A billion bases of DNA (about a third of the human genome) can now be sequenced for under $10,000 and in a matter of days. The goal for human diagnostics is to get to the $1000 genome, and this seems to be attainable within 5 years. For the study of natural variation, inexpensive and rapid DNA sequencing means that we will soon be able to have complete sequence information for all of the genotypes in a population. Hopefully, this does not mean that we will need to sequence every individual in the population; rather, sampling methods will be developed to determine the extent of variation within the population, and then informative genomes can be fully sequenced.

The other major transformation is the use of a host of technologies to improve the precision and breadth of phenotyping. Several studies have shown the value of precise phenotyping in QTL analyses. For example, to identify genes involved in asthma, researchers required highly specific diagnostic guidelines rather than more general ones (3). The technologies that are beginning to be applied to natural variation include fine-scale, real-time microscopic and macroscopic imaging (4), as well as genome-wide RNA, protein, and metabolite profiling. Together, these technologies are likely to redefine what we call a phenotype. In the past, a phenotype was generally a one-dimensional property: the height of a pea plant, the eye color of a fruit fly, or the glucose level of a human. In the future, phenotype will be a “high-dimensional” entity: the combination of morphological, transcriptional, protein, and metabolic readouts associated with a particular combination of alleles.

How do network models of genotype-phenotype relationships compare with quantitative genetic models of trait variation? Several groups have begun to combine known information on metabolic and regulatory networks with whole-genome expression data, with the goal of predicting organismal phenotypes (5–7). These mathematical models quantify a causal relationship from genes to gene products to phenotypes, and they can model causal influences of genes that are either monomorphic or polymorphic. In contrast, traditional quantitative genetic models deal only with segregating genetic variation, and thus causal effects of monomorphic loci are invisible to quantitative genetic analysis. There is great potential for a synthesis of network and quantitative genetic modeling to include network topologies and whole-genome expression data.

Can information from genetic networks predict which genes contribute to complex trait variation? Although available information on the genes that underlie QTLs is limited, several studies have examined a related question: the factors that influence rates of protein evolution among species. The best single predictor of the rate of amino acid change in proteins is the level of protein expression (8), with highly expressed genes evolving more slowly. Several studies find that genes on the network periphery are more likely to contribute to disease or show patterns of rapid or adaptive evolution (9–13), suggesting that peripheral proteins may be more likely to influence complex trait variation. However, not all studies support this conclusion (14), and there is substantial variation around this trend.

Now we can return to the hub-and-spoke analogy of natural variation in networks. Natural knockouts of hub proteins often will be lethal, whereas small changes in the function of hub proteins may have pleiotropic effects on multiple traits. For our airline analogy, intermittent thunderstorms in Chicago have a different effect on air traffic than that of a snowstorm that shuts down the airport entirely. Similarly, weakly deleterious alleles may segregate at low frequency in populations and contribute to disease and inbreeding depression (15). In contrast, genetic networks may tolerate substantial mutations in peripheral proteins, so natural allelic series at these genes may span a broad range: from small to large effects on phenotypes. Advances in phenotyping technology will increasingly enable saturating QTL screens to identify the natural allelic series that influence network function and that modulate the normal range of network functions we seek to understand in human health and agricultural production.

Analyses of natural variation have great potential to dissect the genetic networks controlling important biological processes. QTL approaches begin with functional polymorphisms influencing complex traits, which can be identified and manipulated via high-throughput techniques. Because these QTLs segregate within extant populations, many of them may be ecologically advantageous in nature. Breeders of both plants and animals long ago discovered that crossing individuals with advantageous traits sometimes results in far greater improvement than predicted (hybrid vigor) but frequently produces the opposite effect: offspring that are not as fit as either parent. This latter outcome has been attributed to epistasis, which traditionally has been interpreted as the effect of genes in a pathway in which the modification of one gene overrides any effect of the modification of a second gene. It is now clear that epistatic interactions among loci play a central role in complex trait variation (16,17) and indeed can arise from a broad range of network architectures, with or without feedback mechanisms (18). As with induced mutations, epistatic interactions involving natural variants may illuminate the function of biological networks (19). However, one disadvantage of using natural genetic variation is that we are limited to variants that are polymorphic in the studied populations. As high-throughput technologies advance, future saturating QTL studies of natural populations may be able to reveal most of the ways that network function can be modified.

Although much of the focus on natural variation has been on human diseases, model plant systems are likely to play an important role in the future synthesis of systems biology and quantitative genetics (Fig. 2). Unlike humans and most other mammals, many plants are experimentally tractable and allow easy control and quantification of environmental influences. Plants are amenable to quantitative phenotyping and to dissection of complex phenotypes into their physiological components, which tend to be more robust to experimental manipulations because their physiologies can tolerate more variation than animals. Available natural variation can be augmented by direct crosses, recombinant inbred lines, association panels, and completely sequenced genotypes, which provide a genome-wide catalog of polymorphic alleles.

Between the methods of systems biology and the resources inherent in natural variation, we expect to see insights into the networks that control biological processes such as growth and development. Out of this should come major progress in mapping genotypes to phenotypes. However, much remains to be done. Current methods rarely provide genome-wide analyses at the level of individual cell types or tissues, thus diluting or even losing critical information. This is particularly true when developmentally sensitive phenotypes are analyzed with whole-genome methods such as microarrays. Genes that are expressed highly in a few cell types are not detected when an entire organ or organism is the starting point for these analyses. Nevertheless, the integration of systems biology with quantitative genetic studies of natural variation may fulfill at least part of the promise of genomics to let us know what genes do.

References and Notes

1. Slonim N, Elemento O, Tavazoie S. Mol. Syst. Biol 2006;2
2. Friedman A, Perrimon N. Cell 2007;128:225. [PubMed: 17254958]

3. Van Eerdewegh P, et al. Nature 2002;418:426. [PubMed: 12110844]
4. Megason SG, Fraser SE. Cell 2007;130:784. [PubMed: 17803903]
5. Welch SM, Dong ZS, Roe JL, Das S. Aust. J. Agric. Res 2005;56:919.
6. Jonsson H, et al. Bioinformatics 2005;21:i232. [PubMed: 15961462]
7. Sieberts SK, Schadt EE. Mamm. Genome 2007;18:389. [PubMed: 17653589]
8. Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH. Proc. Natl. Acad. Sci. U.S.A 2005;102:14338. [PubMed: 16176987]
9. Vitkup D, Kharchenko P, Wagner A. Genome Biol 2006;7:R39. [PubMed: 16684370]
10. Kim PM, Lu LJ, Xia Y, Gerstein MB. Science 2006;314:1938. [PubMed: 17185604]
11. Makino T, Gojobori T. Mol. Biol. Evol 2006;23:784. [PubMed: 16407461]
12. Kim PM, Korbel JO, Gerstein MB. Proc. Natl. Acad. Sci. U.S.A 2007;104:20274. [PubMed: 18077332]
13. Goh K-I, et al. Proc. Natl. Acad. Sci. U.S.A 2007;104:8685. [PubMed: 17502601]
14. Batada NN, Hurst LD, Tyers M. PLoS Comput. Biol 2006;2:e88. [PubMed: 16839197]
15. Mitchell-Olds T, Willis JH, Goldstein DB. Nat. Rev. Genet 2007;8:845. [PubMed: 17943192]
16. Kroymann J, Mitchell-Olds T. Nature 2005;435:95. [PubMed: 15875023]
17. Kusterer B, et al. Genetics 2007;177:1839. [PubMed: 18039885]
18. Gjuvsland AB, Hayes BJ, Omholt SW, Carlborg Ö. Genetics 2007;175:411. [PubMed: 17028346]
19. Dworkin I, Palsson A, Birdsall K, Gibson G. Curr. Biol 2003;13:1888. [PubMed: 14588245]

20. We thank A. Hartemink, G. Wray, M. Noor, P. Magwene, J. Willis, and members of the Benfey and Mitchell-Olds labs for thoughtful comments on the manuscript, as well as J. Modliszewski and J. Willis for the images of *Mimulus* flowers. Work in the Mitchell-Olds lab on natural variation is funded by a grant from the NSF. Work in the Benfey lab on systems biology is funded by grants from NIH, NSF, and the Defense Advanced Research Projects Agency.

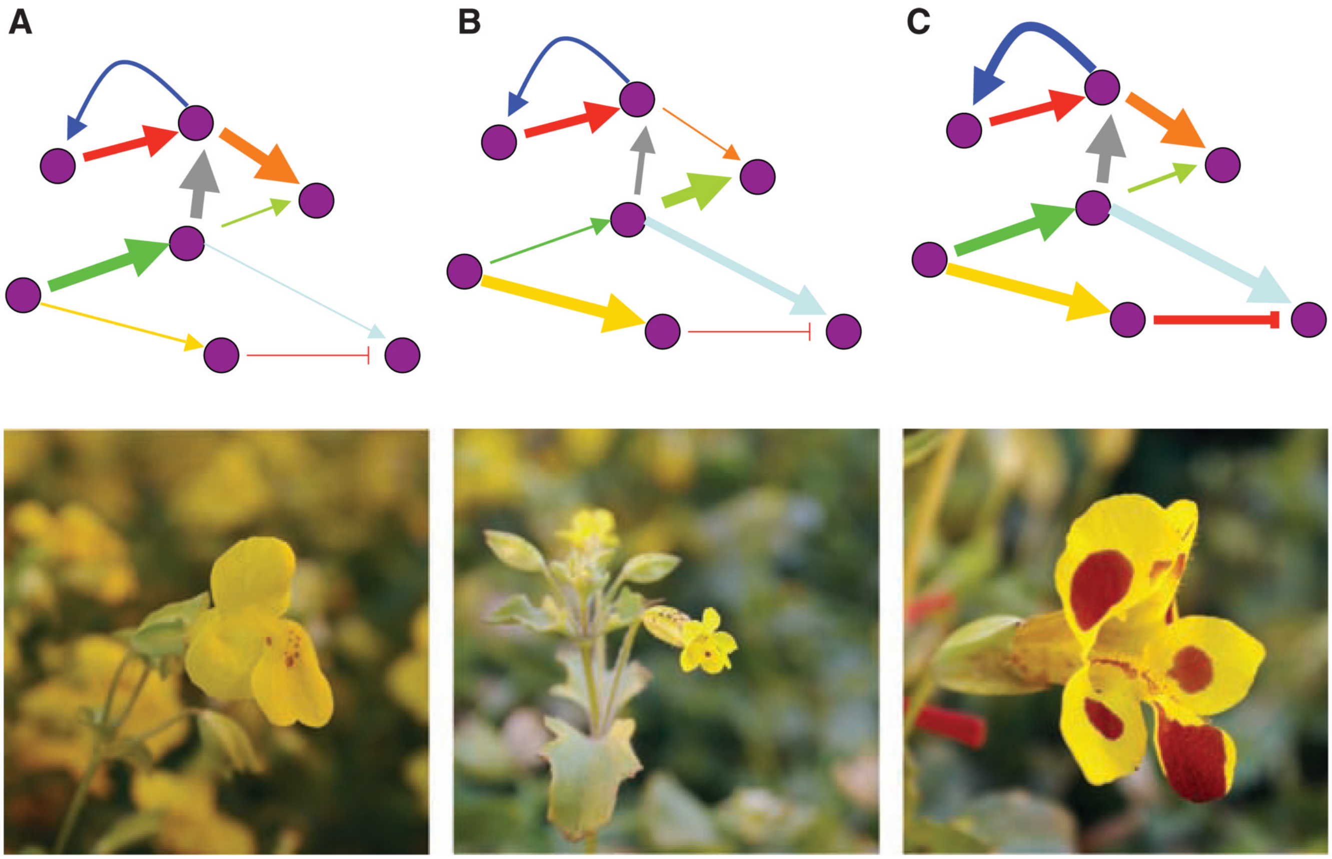

Fig. 1.
Ways in which a hypothetical network could control flower form and color among *Mimulus* species. The widespread species *M. guttatus* (**A**) has large, yellow flowers. In contrast, the flowers of *M. laciniatus* (**B**) are typically 75% smaller than those of *M. guttatus*. Other species show elevated expression of red anthocyanin pigments (**C**), as in this hybrid between subspecies of *M. luteus*. Changes at various points in the network (represented by differing widths of the connections [arrows] between network nodes [circles]) could be responsible for this natural variation. [Photos by J. Modliszewski]

Fig. 2.

Systems biology approaches can be applied to natural variation in wild relatives of *Arabidopsis*, such as this *Boechera* population on the continental divide in Montana (USA).