NATURE|Vol 456|11 December 2008|doi:10.1038/nature07633

Reverse engineering the genotype–phenotype map with natural genetic variation

Matthew V. Rockman[1]

The genetic variation that occurs naturally in a population is a powerful resource for studying how genotype affects phenotype. Each allele is a perturbation of the biological system, and genetic crosses, through the processes of recombination and segregation, randomize the distribution of these alleles among the progeny of a cross. The randomized genetic perturbations affect traits directly and indirectly, and the similarities and differences between traits in their responses to common perturbations allow inferences about whether variation in a trait is a cause of a phenotype (such as disease) or whether the trait variation is, instead, an effect of that phenotype. It is then possible to use this information about causes and effects to build models of probabilistic 'causal networks'. These networks are beginning to define the outlines of the 'genotype–phenotype map'.

More than a decade into the genomic era, it remains easier to collect genomic data sets than to understand them. The research community has obtained vast quantities of data on genes, transcripts, proteins, metabolites and so on, but has discerned only faint outlines of the networks that connect these factors. Biologists are justifiably enthusiastic about the ability to describe networks of biological molecules that are co-expressed or co-localized, but a central goal of contemporary biology is to connect these observable patterns to form models that predict how biological networks operate as systems. The networks that matter in this context are networks of causal relationships, which can be uncovered by using experiments in which biological systems are perturbed[1–3]. The translation of genotype into phenotype depends solely on these causal relationships; many of the relationships are shaped by the co-expression of genes and physical interactions between cellular components, but many others are determined by intricate networks of cause and effect that are mediated by an organism's physiology, behaviour, and interactions with the environment[4] (Box 1). Inferring causal networks from observations is often called reverse engineering, because the goal is not merely to identify components that are functionally related or situated near to one another but to understand how the system works as an integrated whole.

The classic method for reverse engineering a system is to poke a component with a stick and then to characterize the effect of the perturbation. An alternative is to poke many components simultaneously and at random, repeating the experiment over many random sets of components. Ever since R. A. Fisher put forward his ideas[5] in the 1920s, statisticians have recognized such randomized multifactorial perturbation as the ideal experimental design for uncovering causation. Conveniently, the genetic variation that occurs naturally within a population is a source of multifactorial perturbation[6,7]. The use of natural genetic variation to probe the causal network that links genotype and phenotype has grown recently as large data sets have been generated for many experimental model species, crops and humans[8–10]. In this Review, I discuss recent progress in the application of natural genetic variation to reverse engineer the 'genotype–phenotype map'. After introducing the basic experimental approach, I describe its advantages over traditional genetic screens and show how the resultant data allow tentative inferences to be made about causation. Finally, I discuss the steps that are being taken to gain a mechanistic understanding of the network that connects genotype to phenotype, and I point out potential obstacles to this process, as well as potential shortcuts.

Quantitative genetics of transcript abundance

Genetically characterized populations are the central tool for uncovering the genetic variants that underlie phenotypic variation. A common approach is to cross two inbred lines, each homozygous at every locus, to yield a hybrid that is heterozygous at every locus that differs between the strains. In a typical cross, thousands to millions of genetic loci differ. The ordinary process of meiotic recombination rearranges these polymorphisms within the hybrid germ line, and segments of each of the initial genomes are passed on randomly to the progeny of the hybrids. By tracking the genomic segments with molecular markers, the regions of the genome that contain genetic variants that affect phenotypes (known as quantitative trait loci, QTLs) can be identified[11] (Fig. 1a–d).

An important recent advance in connecting the links between genotype and phenotype has been to measure the 'phenotypic states' of the links, most notably the abundance of the transcripts corresponding to each gene of interest[8–10,12]. Quantitative genetic analysis of genome-wide transcript abundance is sometimes called genetical genomics[6] or expression QTL mapping[9], and the results from this type of analysis — the correlations between genes and transcript phenotypes — can be represented by plotting the physical position in the genome of the gene corresponding to each transcript against the position of the loci associated with variation in transcript abundance (Fig. 1d). In a properly controlled cross, an association between genotype and phenotype implicates genetic variation as the cause of the phenotypic variation; as is the case for all claims based on empirical data, confidence in such a causal inference is defined statistically. The genetic analysis of genome-wide transcript abundance poses distinct technical, computational and analytical challenges that necessitate careful avoidance of potential artefacts[13,14] and that drive innovation in statistical genetics methods[15–19].

Analyses typically show many linkages between the abundances of transcripts and the regions in the genomes where the structural genes for those transcripts reside. Such genes contain QTLs for their

[1]Center for Genomics and Systems Biology, Department of Biology, New York University, 100 Washington Square East, New York, New York 10003, USA.

own transcript abundance. Consequently, the data points align on the diagonal in Fig. 1d. Most of these local linkages will be attributable to *cis*-acting regulatory polymorphisms, although there will be some contribution from polymorphisms that affect the transcript abundance by acting *in trans*[10,20,21].

Data points aligning in vertical bands in Fig. 1d indicate linkage hotspots: that is, regions of the genome at which variation alters the abundance of a large number of transcripts. The loci responsible for these large phenotypic effects might be highly influential regulatory loci[22], or they might be loci at which variation has a marked but non-specific effect on transcript abundances. Such pleiotropic alleles might alter cellular homeostasis in such a way as to shift the steady state for a large number of traits, without this shift being a regulatory effect[10,23].

The quantitative genetics approach described earlier, using natural genetic variation as a source of perturbations, has striking advantages over the classic (one gene at a time) approach[7]. First, the quantitative genetics approach involves massive hidden replication. The effect of each of the alleles present in the initial cross is measured repeatedly because each allele is present in a large number of the phenotyped progeny. A small phenotyping panel of 100 individuals represents on average a 50-fold replication in studying the effect of every allele; similar amounts of replication are not feasible for a one-at-a-time approach.

Second, the presence of simultaneous variation at multiple loci allows the interactions between perturbations (that is, genetic variants) to be uncovered. In the context of gene-expression genetics, such interactions are likely to be common[24,25], and observations have shown that the inheritance of many transcript-abundance traits involves interactions between genes[16,26,27]. A prime example of this gene-interaction phenomenon is genetic redundancy; only by varying the redundant loci simultaneously can their effects be detected.

Last, the simultaneous perturbation of a large number of factors results in the exploration of a larger 'space' of variation than when carrying out single perturbations. Genetically complex traits, which are shaped by variation at many loci, often show transgressive segregation: that is, the random assignment of alleles to the progeny of hybrids results in individuals with unusual collections of alleles, which yield extreme phenotypes. Transgressive segregation characterizes the majority of the transcript-abundance traits in crosses in which its frequency has been examined[26,28]. The large space of phenotypes covered by genetically segregating populations increases the ability to detect relationships between traits. Transcript-abundance data from such populations, for example, are unusually successful at predicting functional relationships between genes[29–33].

Causal ordering

A QTL can affect some traits directly and can affect others indirectly through the effects of intermediate traits. Of particular interest is whether variation in an organismal phenotype, such as a disease state or a behaviour, is an effect of transcript-abundance variation or a cause[34]. Only if the transcript-abundance trait is a cause (not an effect) can it be a target for perturbations that shape variation in the organismal trait, whether in the laboratory, the clinic or an evolving population.

Under a broad set of assumptions, causality shapes correlations in a recognizable way, and its signature is conditional independence. This can be shown by considering three causally related traits: *A*, *B* and *C*. Under standard Markov assumptions, if variation in *A* causes variation in *B*, which in turn causes variation in *C* (this can be written as $A \rightarrow B \rightarrow C$), then when the distribution of *B* is known, *A* provides no additional information about *C*. *A* and *C* are independent conditional on *B*. This statement of conditional independence would not hold if the causal ordering were $B \rightarrow C \rightarrow A$, for example. Conditional independence is by itself insufficient to order causal links uniquely: $A \leftarrow B \rightarrow C$ yields the same conditional independence statement as $A \rightarrow B \rightarrow C$. Nevertheless, conditional independence is a powerful tool for distinguishing direct links between traits from indirect links, even when causal ordering is not possible[1,35,36].

The key advantage of studying genetic perturbations is that many causal orderings are prohibited by the central dogma that genotypic variation can cause phenotypic variation, but, at least within an individual, phenotype does not feed back to affect genotype. Therefore, in a properly controlled cross, genetic perturbations are causally upstream of phenotypes and provide a terra firma into which causal networks can be rooted[34,37–40] (Box 2).

There are several methods for causally ordering pairs of phenotypes measured in segregating populations, including approaches that simultaneously map QTLs and fit causal models[18], approaches that apply formal statistical tests to identify direct causal links[39], and approaches that fit various causal models to triplets that comprise two traits and a QTL and then compare the fit of the models using information-theory criteria[34].

Analysis of the correlations between multiple traits that share an underlying QTL can also help to identify the causal gene within a QTL interval[33,37,41–44] — the major challenge in quantitative genetics today. Many such analyses incorporate sources of information in addition to the correlations, including data on the binding sites of transcription factors,

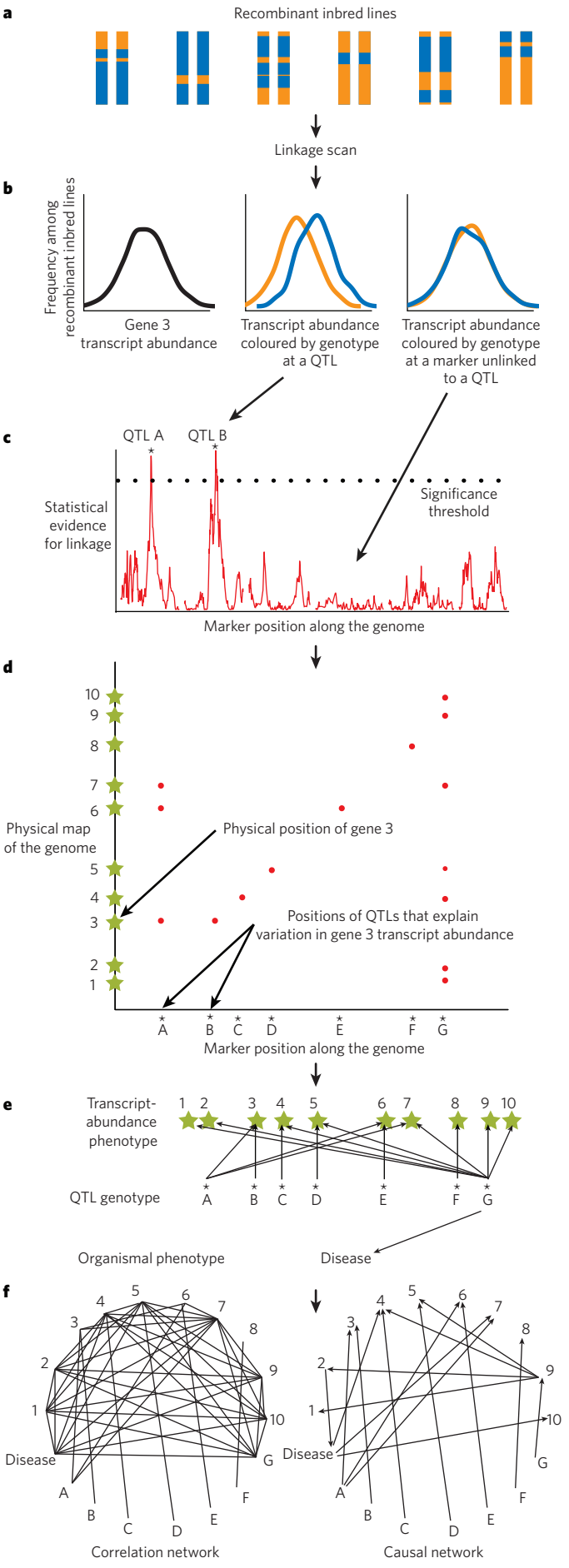
Box 1 | Do causal networks exist?

The concepts of causality and networks are controversial. Many biologists are sceptical about inferring causality from statistics and about how general and useful network models might be, so it is valuable to consider how causality and networks can be interpreted in the context of the relationship between genotype and phenotype.

When carrying out a biological experiment, most people are content with the everyday theory of causation that dictates that one fact precedes a second and alters its probability (setting aside questions about the ontological status of probabilities). Empirical claims about links between causes and effects rely on assumptions (often implicit) and on statistical measures of confidence. Even when testing a simple single-gene perturbation, such as in a gene-knockout organism, it is assumed that inductive reasoning will lead to knowledge, that genotypes are fixed and that they causally precede phenotypes, and that all variables are fully controlled (for example, by comparing the experimental organisms with wild-type full siblings, by blinding observers to differences in treatment, and by carrying out randomized replications of the entire experiment). A researcher's confidence that the wild-type organism and the mutant organism show real differences is influenced by statistical tests (for example, the *P* value from a *t*-test), which are themselves typically laden with assumptions.

In inferring probabilistic causal networks, the assumptions are more numerous, but the conceptual framework is the same. At the end of an analysis, the outcome is a claim about cause and effect, and this claim is accepted to an extent that is defined by the researcher's comfort with the assumptions and the statistics. For genetics experiments, the limits of comfort for most researchers lie not far beyond the single-gene perturbation experiment, and making inferences about a causal network is typically seen as a technique for nominating candidate genes for follow-up experiments.

Whether networks exist is a popular topic at biology department happy hours, but it is not necessary to subscribe to the reality of a Platonic Network, an ideal form independent of the material world, in order to embrace the idea that there is a many-to-many relationship between causes and effects in biology. The more pressing question is which components are needed to represent such a network in a predictive model: molecules, interactions, dynamics, all of these, or more? Conveniently, the set of variables in a causal network is entirely circumscribed by the set of things that vary. A molecule or an event can be required for a biological process, but if it does not vary, then it cannot be a cause. In that sense, causal networks in genetics are analogous to the geneticist's concept of heritability, which describes not the dependence of a trait on inherited genes but the proportion of the trait's variation that can be explained by genetic variation in the observed sample. A causal network that is inferred from a genetically segregating population will therefore depend on the genetic variation that is present in the population and on the distribution of variation in the environment across the sampled individuals. Within that framework, a causal network constitutes a predictive model of the consequences of perturbing the represented variables.



the interactions between proteins, and the presence of polymorphisms in the sequences of each gene in the QTL. As a starting point, a gene that is found at the same location as a QTL for its own abundance is a strong candidate for the causal gene underlying variation in other traits that link to the same location, with the gene's transcript abundance as a candidate causal trait[29].

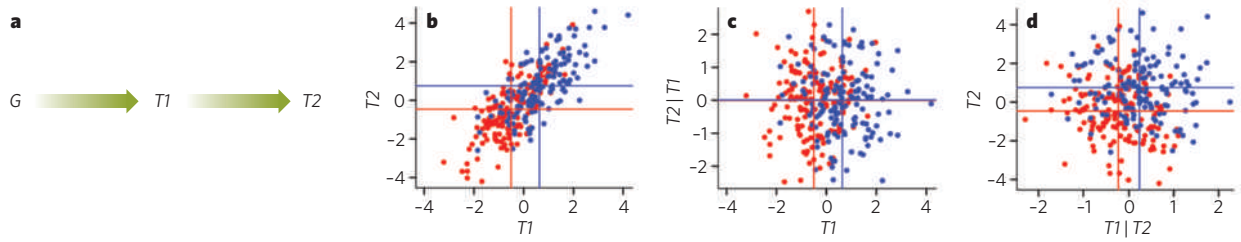
The use of phenotypic correlations to identify causal transcripts is especially promising in the context of genome-wide association studies. These studies have recently uncovered a wealth of high-confidence, replicated associations between genetic variants and diseases in humans (see page 728), but the disease-associated variants are often in non-coding regions with unknown function[45]. The mechanisms that link genotype and disease in these cases can be identified by taking advantage of the structure of the correlations among transcript-abundance traits and disease states in human populations[32,46–48]. In population-based association mapping, however, correlations between genotype and phenotype can arise from external causes that are common to the correlated variables, for example from the stratification of populations by age or ethnicity. Consequently, the causal 'anchor' provided by genotype in these cases is less secure than in the experimental setting of inbred line crosses. But studies in animal models can be used to corroborate findings, providing reassurance[31,49].

Causal networks

To gain a predictive, systems-level understanding of biological causation, researchers need to integrate the entire ensemble of genetic variants and phenotypic traits (and not just to causally order trait pairs). Several approaches aim at this more general goal, and Bayesian networks provide the most popular framework[1]. A Bayesian network is a graph of random variables, each representing a phenotype in this case, that are connected by directed edges. A set of probability distributions describes the state of each variable conditional on the variables with edges leading to it. The graph and probability distributions define a conditional probability statement. One problem with using Bayesian network graphs is that several directed graphs can be described by a

Figure 1 | From genetic randomization to causal network. A genetically randomized population, such as a panel of recombinant inbred lines whose chromosomes carry random segments of genome from two progenitor strains (depicted in orange and blue) (**a**), is a starting point for linkage analysis of a phenotype. In this case, the phenotype is the abundance of transcripts corresponding to a gene denoted as gene 3. The abundance of gene 3 transcripts varies between the recombinant lines (**b**, left). Each point along the genome is tested to see whether it affects the abundance of gene 3 transcripts (**b**, centre and right), and statistical evidence is uncovered for the linkage of gene 3 with two regions (indicated by asterisks) (**c**). These regions are called QTLs. If a similar experiment is carried out for many transcript-abundance phenotypes (not just for gene 3, but for genes 1–10), the positions of the QTLs that affect transcript abundance (asterisks) can be plotted against the physical positions of the gene corresponding to each transcript (green stars) (**d**). In such a plot, the data (red dots) along the diagonal line represent local linkages, typically due to *cis*-acting regulatory polymorphisms. Vertical alignments in the plot indicate linkage hotspots. The plot depicted implies a high-level causal network (shown in **e**), in which QTL variation is the cause of variation in transcript-abundance phenotype. Transcript-abundance QTLs can co-localize with QTLs for organismal phenotypes such as a disease (not shown); for illustrative purposes, disease is shown linked to QTL G. A goal of the reverse engineering of causal networks is to include phenotypes as variables, for example to determine whether the transcript abundances that are affected by QTL G are causes or effects of the disease. Although the traits are densely connected by correlations — as is evident from the hypothetical correlation network that is depicted (**f**), which connects all traits that share perturbations — a causal network (**f**) reveals that QTL G acts directly on gene 9, the transcript abundance of which affects genes 1, 2, 4 and 5. The transcript abundance of gene 2 is a cause of disease, which in turn alters the transcript abundances of genes 4, 7 and 10. Many transcripts are correlated with disease, but only perturbations of genes 2 and 9 will affect disease outcome.

Box 2 | Causal ordering yields conditional independence



The basis for causal inference can be shown graphically. Consider a population of haploid individuals with a single causal locus (*G*) that has two alleles. The allelic state at this locus causes variation in the abundance of the corresponding transcript (*T1*), and additional sources of variation (genetic, environmental and stochastic) also influence *T1*. Data can be simulated with the allelic effect modelled as β_1 and the additional variation modelled as normally distributed noise, ε. Thus, $T1 = \beta_1 G + \varepsilon$, where *G* is in a indicator variable for genotype. Variation in the abundance of *T1* causes variation in a downstream trait, *T2*, which is also affected by other sources of variation, so $T2 = \beta_2 T1 + \gamma$, where γ is an additional noise term. The causal ordering is shown as a scheme in panel **a** of the figure, and simulated data are plotted in panel **b** of the figure.

Data were simulated to represent 300 individuals. Each data point is coloured according to genotype (blue for one allele of *G* and red for the other), and the mean values for each trait are indicated by a coloured line for each genotype. (For this simulation, genotypes were assigned indicator variables: red was assigned –1, and blue was assigned 1. The parameters used were $\beta_1 = 0.5$ and $\varepsilon \sim N(0, 1)$, yielding a zero-mean trait *T1*. For *T2*, $\beta_2 = 1$ and $\gamma \sim N(0, 1)$, so *T2* is simply *T1* with additional noise.)

The causal links mean that the traits, *T1* and *T2*, are correlated with one another and that both are correlated with genotype (figure, **b**). Nevertheless, the relationship between *T2* and *G* is entirely mediated by *T1*. *T2* conditional on *T1* is independent of genotype; the mean phenotype for each genotype is the same (figure, **c**). Conversely, the distribution of *T1* conditional on *T2* remains dependent on genotype (figure, **d**). It is the noise component of *T1* variation, ε, propagated through *T2* that makes this mode of analysis possible, and it is the causal anchor of the genotype that gives it direction.

single conditional probability statement (as discussed earlier), so observations of the random variables (the transcript abundances in this case) cannot uniquely identify the directed network that underlies the graph. Moreover, a second problem is that Bayesian network graphs are acyclic and therefore cannot model feedback regulation; the consequences of this limitation for the utility of Bayesian network are unclear[1,31,37]. A third problem is that the space of possible network graphs is large, making causal-network inference a computationally intractable problem[1]. These difficulties notwithstanding, transcript-abundance measurements from genetically segregating populations are uniquely suited to uncovering directed Bayesian networks for two main reasons[29,37]. First, a trait that is caused by another trait should share an underlying genetic perturbation: a QTL. This simple filter excludes a huge proportion of the space of possible networks, making the problem tractable. Second, genetic perturbations anchor causal networks (as discussed earlier), giving direction to the edges. Although large-scale causal-network inference remains challenging, the incorporation of genetic data clearly improves the quality of the predictions over those derived solely from trait correlations[50].

In parallel with Bayesian network models, structural equation models have been applied to transcript-abundance data from segregating populations[38,51]. These models involve systems of linear equations organized into a network structure; a linear model is fitted with variables that simultaneously function as predictors and responses. Although structural equations, unlike Bayesian networks, have the advantage of allowing feedback cycles to be modelled, they require the standard assumptions of linear modelling. Therefore, when nonlinear causal dynamics underlie transcript abundances, problems can arise. Bayesian networks typically deal with nonlinearity incidentally, by classifying all of the data into simple discrete categories (for example, upregulated, downregulated and unchanged), a simplification that has its own drawbacks[50].

An alternative approach is to generate a simple network from pairwise trait correlations and then to trim this network by testing for conditional dependence relationships[35]. The resultant undirected graph can then be directed by anchoring the edges in QTLs[40]. There are clear computational advantages to starting with a network that is derived from pairwise correlations. Because correlations between genes typically show modularity — with clusters of highly correlated genes being largely uncorrelated with other such clusters — pairwise analyses can break intractably large problems into problems that focus on individual modules[29,52,53].

Two recent studies on disease phenotypes in humans and mice used this shortcut of partitioning the transcript-abundance data into modules of correlated traits[31,32]. Breaking the problem down further, the authors compared causal orderings for pairs of transcript abundances and disease phenotypes, to see whether each transcript could be placed causally upstream of the disease state. A single module, evident in both mouse and human data, was significantly enriched for putatively causal traits; subsequent experimental manipulations corroborated these inferences. Although this is far from a complete reverse engineering of the genotype–phenotype map, these empirical successes (reducing genomic data to the two-trait ordering problem) point to a coming age in which prediction will be a common tool.

Quantitative genetics is evolutionary genetics

The central dogma that genes are causes of phenotypes within an individual aids in the anchoring of directed networks. But among individuals, phenotypes feed back by selection to shape genes. Natural variation therefore samples a biased subset of possible genetic perturbations, a subset that is enriched for those variants that are not strongly deleterious. Under the classic infinitesimal model of the genotype–phenotype map, variation derives from mutations of small effect with limited pleiotropy[54]. If this model holds, the modularity of networks inferred from natural variation[29,31,32,37] might be an epiphenomenon of natural genetic perturbations, the effects of which are less systemic than those of random mutations.

The effect of selection is evident in the numbers and types of QTL detected in typical studies. James Ronald and Joshua Akey found that the proportion of genes showing local QTLs in a yeast cross is smaller than expected under neutrality, implying that negative selection keeps certain perturbations at low frequency[55]. Similarly, genes that are crucial regulators of essential processes are likely to be under-represented among genetically variable genes. Genes that encode transcription factors, for example, are clearly candidates when looking for the genes involved in varying gene expression, but these genes are largely absent from expression QTLs[18,56]. Nevertheless, links between transcription factors and target genes can be detected by causal inference approaches, even when the transcription-factor locus contains no genetic variation. The only requirement is that some phenotypic measure of the transcription factor's activity (such as the abundance of the corresponding transcript) is causally intermediate between the genotype and the target gene's phenotype[31].

The effect of selection on the filtering of genetic perturbations varies according to the type of experimental population studied. Inbred line crosses often involve genetically divergent lines, chosen to maximize phenotypic or genotypic differences. The alleles that contribute to divergence are likely to differ in their allelic effects from those that contribute to standing variation (the ordinary genetic diversity present within a population)[57–59]. Crosses between divergent inbred lines are more likely to uncover rare large-effect mutations (that is, linkage hotspots) than are samples of individuals from large populations[32,55,60]. Such crosses might also be biased towards a subset of perturbations with large effect, not individually but in combinations, as a result of coadaptation[61]. Pervasive genetic interaction means that causal links between pairs of genes will be poorly modelled, although there has been progress towards solving this problem recently[16].

The suite of naturally occurring perturbations is also shaped by population genetics phenomena that are unrelated to the fitness effects of the perturbations themselves. Genetic variants tightly linked to other variants that are the target of selection are evolutionarily coupled to them, yielding a correlation between local recombination rate and levels of genetic variation[62,63], and mutagenic recombination can produce the same pattern[64]. Genes present in regions that undergo recombination at a low rate are less likely to contribute to the pool of genetic perturbations than genes in regions where recombination occurs frequently. If gene location is non-random with respect to recombination rate, then there might be fewer perturbations belonging to particular functional classes[61].

Prospects for genetical systems biology

Causal network inference faces many difficulties, and its application to gene expression in segregating populations introduces additional challenges. Despite many published reports of empirical successes, it is important to consider the pitfalls that unpublished studies might have encountered. Many of these pitfalls are now well recognized, and there are clear paths around them, on both the experimental front and the analytical front, towards a richer understanding of the genotype–phenotype map.

One of the most problematic assumptions that is made when drawing causal inferences from gene-expression data is that measurement errors are similarly distributed across traits. If a causal trait is poorly measured and the trait it affects is well measured, then the measurements of the 'effect trait' might report the true values of the causal trait more accurately than measurements made directly on the causal trait itself[34,39,51]. In such cases, conditional correlation approaches can yield inverted inferences (Fig. 2). There are good biological reasons to be concerned about this situation. For example, regulatory molecules are often present at low abundances in a cell, but their effects are amplified by the dynamics of gene regulation. Thus, variation in the number of transcripts encoding a low-abundance transcription factor might be the cause of variation in the number of transcripts encoding a high-abundance structural protein. The difficulty arises in that the high-abundance transcripts might be easy to measure with great precision, whereas the low-abundance transcripts might be present at the threshold of detection. One possible solution is to return to the early designs of microarray experiments, when technical replicates were routine. Taking multiple independent measurements of transcript abundances from a single biological sample would generate empirical parameters for gene-specific error models. The potential to be misled when making causal inferences underscores the point that a causal inference yields an assumption-laden probability statement, and stronger claims about causality need to be experimentally validated[3].

Another concern is that data for transcript-abundance mapping are derived from mixed populations of cells. Consequently, the measurements describe cellular mixtures, the characteristics of which are determined by developmental and cellular demographics[31,32,65]. This is as true for yeast, which has been studied by measuring unsynchronized cultures (which contain cells at different stages of the cell cycle), as it is for animals and plants, in which tissues or whole organisms are assessed. A study of mice that varied genetically in the cell-cycle timing of their haematopoietic stem cells took advantage of the issue of mixed cell populations to identify which genes were expressed differentially in the different cell populations[65].

A related issue is that networks inferred from segregating populations are static representations. Without time-series data, there can be no account of dynamics. Inferred causal links correspond to perturbations that alter the steady state. In a system with feedback, which is likely to include almost all biological steady states, the causal links will describe only the causes that 'win out' over others in shifting the phenotypic balance[31]. For practical considerations of whether it is possible to predict how novel perturbations will affect steady states, these low-dimension projections of the network might be adequate; however, for true reverse engineering, more-complete models are needed. Time-series data have proved exceptionally important for solving the general problem of causal-network inference[2], and it is clear that integrating such information into studies of genetic variation will improve the knowledge gained from these studies.

Many network-inference methods depend on the inclusion of all causal variables in the models[1], but such completeness is not plausible for most biological networks. Some phenotypic causes will be overlooked because they are not measured: for example, unannotated genes, genes present in structural polymorphisms that are absent from reference genomes, and transcripts of small RNAs. In addition, the abundances of metabolites are typically not considered, although much progress has been made in this area recently[40,66]. Regulatory events for which there are no transcriptional indications (for example, post-translational regulation) will also be missed, although again progress is occurring on

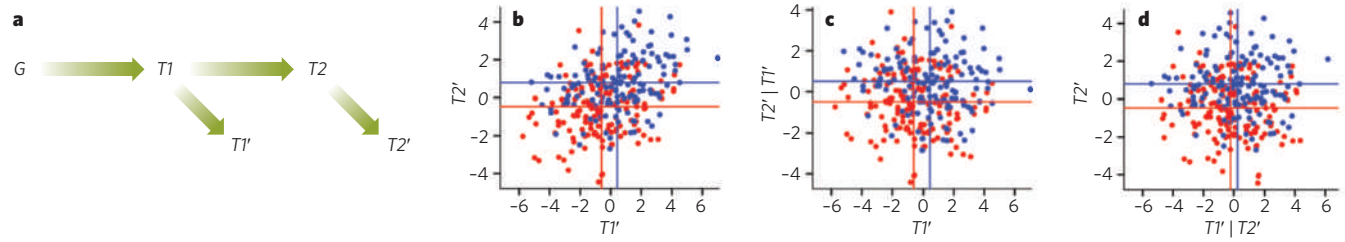


Figure 2 | Measurement error can confuse causal inference. The effect of errors in measurement on causal inference is depicted for the population, parameters and conditions set out in Box 2. In brief, a population of haploid individuals has a single causal locus (*G*) with two alleles, and the allelic state at this locus causes variation in the abundance of the corresponding transcript (*T1*), which subsequently affects the abundance of another transcript (*T2*). **a**, The measured values of *T1* and *T2* in this example were simulated as their true values from Box 2 plus normally distributed error, yielding *T1′* and *T2′*. For *T1′*, the error is normally distributed with a variance of 2, whereas the variance of *T2′* is tenfold lower. The causal ordering of this scheme is shown. **b**, *T1′* and *T2′* are correlated with one another and linked to the genotype, which is represented by the colours (blue for one allele of *G* and red for the other). However, the conditional correlations are now misleading with respect to the true causal network (shown in **a**). **c**, *T2′* remains dependent on genotype after taking *T1′* into account, which is unexpected given the causal ordering (**a**) (and given that *T2* conditional on *T1* does not depend on genotype; Box 2 figure, panel c). **d**, *T1′* is nearly independent of genotype after taking *T2′* into account, which is also unexpected given the causal ordering (**a**) (and given that *T1* has been shown to depend on genotype; Box 2 figure, panel c). In total, the effect of the differences in measurement error is to make *T2′* a better measure of the true *T1* than *T1′* itself.

this front[67,68]. Some missing causal links will arise from an organism's history, because genotypic variables act across an individual's lifespan. For example, the expression of a gene that acts early in an organism's life might show no correlation with the phenotypes it affected at the time of measurement, but the gene–trait relationship will remain[6]. The catalogue of genetic causes will also be incomplete. For large numbers of gene-expression traits, the underlying genetic variants that affect their expression will be undetected because of insufficient power[26]. A simple solution to this problem is to use larger sample sizes, and on this front progress is also being made[69].

Finally, the realm of biological causes is enormous[4], and most experiments limit exploration to a tiny controlled corner of this realm (Box 1). Studies of common lines in multiple environments are now providing the first steps towards integrating genetic perturbations and environmental perturbations into a single view of causal networks[70,71]. Analysing transcript abundances across multiple tissues and sexes is equally important for studying context-dependent causal networks, because each cell type provides a distinct environment for the genome[32,72–74]. Ultimately, as data collection becomes more rapid and less expensive, researchers will be able to study a broader range of conditions.

Natural genetic variation is the stuff of evolution and the cause of heritable susceptibility to diseases. Its properties are fundamentally important to a wide range of biological disciplines. It is therefore fortunate that natural genetic variation has the character of an ideal multifactorial perturbation, providing a natural experimental design that is helping researchers to uncover the mechanistic basis of the map that connects genotype to phenotype. As the molecular catalogues of genomics yield to the integrated models of systems biology, natural genetic variation will have an increasingly central role. ■

1. Friedman, N., Linial, M., Nachman, I. & Pe'er, D. Using Bayesian networks to analyze expression data. *J. Comput. Biol.* **7,** 601–620 (2000).
This paper provides a clear overview of Bayesian network formalisms, the main framework for causal network inference at present, and demonstrates how they can be applied to gene-expression data.
2. Bonneau, R. *et al.* A predictive model for transcriptional control of physiology in a free living cell. *Cell* **131,** 1354–1365 (2007).
3. Sieberts, S. K. & Schadt, E. E. Moving toward a system genetics view of disease. *Mamm. Genome* **18,** 389–401 (2007).
4. Oyama, S. *The Ontogeny of Information: Developmental Systems and Evolution* (Duke Univ. Press, 2000).
5. Fisher, R. A. The arrangement of field experiments. *J. Ministry Agric. Great Britain* **33,** 503–511 (1926).
6. Jansen, R. C. & Nap, J. P. Genetical genomics: the added value from segregation. *Trends Genet.* **17,** 388–391 (2001).
This was the first article in which the many advantages of using natural variation to probe gene-expression networks were articulated.
7. Jansen, R. C. Studying complex biological systems using multifactorial perturbation. *Nature Rev. Genet.* **4,** 145–151 (2003).
8. Brem, R. B., Yvert, G., Clinton, R. & Kruglyak, L. Genetic dissection of transcriptional regulation in budding yeast. *Science* **296,** 752–755 (2002).
9. Schadt, E. E. *et al.* Genetics of gene expression surveyed in maize, mouse and man. *Nature* **422,** 297–302 (2003).
References 8 and 9 provided the first empirical results showing the power of genetic analysis of genome-wide gene expression.
10. Rockman, M. V. & Kruglyak, L. Genetics of global gene expression. *Nature Rev. Genet.* **7,** 862–872 (2006).
11. Lynch, M. & Walsh, B. *Genetics and Analysis of Quantitative Traits* (Sinauer, 1998).
12. Stamatoyannopoulos, J. A. The genomics of gene expression. *Genomics* **84,** 449–457 (2004).
13. Perez-Enciso, M. *In silico* study of transcriptome genetic variation in outbred populations. *Genetics* **166,** 547–554 (2004).
14. Alberts, R. *et al.* A statistical multiprobe model for analyzing *cis* and *trans* genes in genetical genomics experiments with short-oligonucleotide arrays. *Genetics* **171,** 1437–1439 (2005).
15. Carlborg, O. *et al.* Methodological aspects of the genetic dissection of gene expression. *Bioinformatics* **21,** 2383–2393 (2005).
16. Storey, J. D., Akey, J. M. & Kruglyak, L. Multiple locus linkage analysis of genomewide expression in yeast. *PLoS Biol.* **3,** e267 (2005).
17. Kendziorski, C. M. *et al.* Statistical methods for expression quantitative trait loci (eQTL) mapping. *Biometrics* **62,** 19–27 (2006).
18. Kulp, D. C. & Jagalur, M. Causal inference of regulator–target pairs by gene mapping of expression phenotypes. *BMC Genomics* **7,** 125 (2006).
19. Jia, Z. & Xu, S. Mapping quantitative trait loci for expression abundance. *Genetics* **176,** 611–623 (2007).
20. Doss, S., Schadt, E. E., Drake, T. A. & Lusis, A. J. *Cis*-acting expression quantitative trait loci in mice. *Genome Res.* **15,** 681–691 (2005).
21. Ronald, J., Brem, R. B., Whittle, J. & Kruglyak, L. Local regulatory variation in *Saccharomyces cerevisiae*. *PLoS Genet.* **1,** e25 (2005).
22. Morley, M. *et al.* Genetic analysis of genome-wide variation in human gene expression. *Nature* **430,** 743–747 (2004).
23. Churchill, G. A. The genetics of gene expression. *Mamm. Genome* **17,** 465 (2006).
24. Omholt, S. W., Plahte, E., Øyehaug, L. & Xiang, K. Gene regulatory networks generating the phenomena of additivity, dominance and epistasis. *Genetics* **155,** 969–980 (2000).
25. Gjuvsland, A. B., Hayes, B. J., Omholt, S. W. & Carlborg, O. Statistical epistasis is a generic feature of gene regulatory networks. *Genetics* **175,** 411–420 (2007).
26. Brem, R. B. & Kruglyak, L. The landscape of genetic complexity across 5,700 gene expression traits in yeast. *Proc. Natl Acad. Sci. USA* **102,** 1572–1577 (2005).
27. Brem, R. B., Storey, J. D., Whittle, J. & Kruglyak, L. Genetic interactions between polymorphisms that affect gene expression in yeast. *Nature* **436,** 701–703 (2005).
28. West, M. A. *et al.* Global eQTL mapping reveals the complex genetic architecture of transcript-level variation in *Arabidopsis*. *Genetics* **175,** 1441–1450 (2007).
29. Lum, P. Y. *et al.* Elucidating the murine brain transcriptional network in a segregating mouse population to identify core functional modules for obesity and diabetes. *J. Neurochem.* **97** (suppl. 1), 50–62 (2006).
30. Huttenhower, C. *et al.* Nearest neighbor networks: clustering expression data based on gene neighborhoods. *BMC Bioinformatics* **8,** 250 (2007).
31. Chen, Y. *et al.* Variations in DNA elucidate molecular networks that cause disease. *Nature* **452,** 429–435 (2008).
32. Emilsson, V. *et al.* Genetics of gene expression and its effect on disease. *Nature* **452,** 423–428 (2008).
References 31 and 32 integrated association mapping in human populations and linkage mapping in mice to identify suites of functionally related genes that are causally implicated in disease.
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).
34. Schadt, E. E. *et al.* An integrative genomics approach to infer causal associations between gene expression and disease. *Nature Genet.* **37,** 710–717 (2005).
35. de la Fuente, A., Bing, N., Hoeschele, I. & Mendes, P. Discovery of meaningful associations in genomic data using partial correlation coefficients. *Bioinformatics* **20,** 3565–3574 (2004).
36. Magwene, P. M. & Kim, J. Estimating genomic coexpression networks using first-order conditional independence. *Genome Biol.* **5,** R100 (2004).
37. Zhu, J. *et al.* An integrative genomics approach to the reconstruction of gene networks in segregating populations. *Cytogenet. Genome Res.* **105,** 363–374 (2004).
This paper was the first to integrate expression QTL data and phenotypic correlation data into causal modelling, as well as to describe the crucial role of genetic perturbations in anchoring causal links in the Bayesian network context.
38. Li, R. *et al.* Structural model analysis of multiple quantitative traits. *PLoS Genet.* **2,** e114 (2006).
39. Chen, L. S., Emmert-Streib, F. & Storey, J. D. Harnessing naturally randomized transcription to infer regulatory relationships among genes. *Genome Biol.* **8,** R219 (2007).
This paper details a conservative analysis pipeline for uncovering high-confidence causal links with a well-defined false-discovery rate.
40. Ferrara, C. T. *et al.* Genetic networks of liver metabolism revealed by integration of metabolic and transcriptional profiling. *PLoS Genet.* **4,** e1000034 (2008).
41. Bing, N. & Hoeschele, I. Genetical genomics analysis of a yeast segregant population for transcription network inference. *Genetics* **170,** 533–542 (2005).
42. Li, H. *et al.* Inferring gene transcriptional modulatory relations: a genetical genomics approach. *Hum. Mol. Genet.* **14,** 1119–1125 (2005).
43. Tu, Z. *et al.* An integrative approach for causal gene identification and gene regulatory pathway inference. *Bioinformatics* **22,** e489–e496 (2006).
44. Suthram, S. *et al.* eQED: an efficient method for interpreting eQTL associations using protein networks. *Mol. Syst. Biol.* **4,** 162 (2008).
45. McCarthy, M. I. *et al.* Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Rev. Genet.* **9,** 356–369 (2008).
46. Dixon, A. L. *et al.* A genome-wide association study of global gene expression. *Nature Genet.* **39,** 1202–1207 (2007).
47. Goring, H. H. *et al.* Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes. *Nature Genet.* **39,** 1208–1216 (2007).
48. Stranger, B. E. *et al.* Population genomics of human gene expression. *Nature Genet.* **39,** 1217–1224 (2007).
49. Schadt, E. E. *et al.* Mapping the genetic architecture of gene expression in human liver. *PLoS Biol.* **6,** e107 (2008).
50. 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).
51. Liu, B., de la Fuente, A. & Hoeschele, I. Gene network inference via structural equation modeling in genetical genomics experiments. *Genetics* **178,** 1763–1776 (2008).
52. Ghazalpour, A. *et al.* Integrating genetic and network analysis to characterize genes related to mouse weight. *PLoS Genet.* **2,** e130 (2006).
53. Lee, S. I. *et al.* Identifying regulatory mechanisms using individual variation reveals key role for chromatin modification. *Proc. Natl Acad. Sci. USA* **103,** 14062–14067 (2006).
54. Fisher, R. A. *The Genetical Theory of Natural Selection* (Oxford Univ. Press, 1930).
55. Ronald, J. & Akey, J. M. The evolution of gene expression QTL in *Saccharomyces cerevisiae*. *PLoS ONE* **2,** e678 (2007).
This paper is a founding contribution to the field of functional population genomics; it addresses the genomic basis of phenotypic evolution from the perspective of the functional alleles segregating in populations.
56. Yvert, G. *et al. Trans*-acting regulatory variation in *Saccharomyces cerevisiae* and the role of transcription factors. *Nature Genet.* **35,** 57–64 (2003).
57. Barton, N. H. & Keightley, P. D. Understanding quantitative genetic variation. *Nature Rev. Genet.* **3,** 11–21 (2002).
58. Ohta, T. Origin of the neutral and nearly neutral theories of evolution. *J. Biosci.* **28,** 371–377 (2003).
59. Wittkopp, P. J., Haerum, B. K. & Clark, A. G. Regulatory changes underlying expression differences within and between *Drosophila* species. *Nature Genet.* **40,** 346–350 (2008).

60. Schliekelman, P. Statistical power of expression quantitative trait loci for mapping of complex trait loci in natural populations. *Genetics* **178,** 2201–2216 (2008).
61. Petkov, P. M. *et al.* Evidence of a large-scale functional organization of mammalian chromosomes. *PLoS Genet.* **1,** e33 (2005).
62. Begun, D. J. & Aquadro, C. F. Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. *Nature* **356,** 519–520 (1992).
63. Charlesworth, B., Morgan, M. T. & Charlesworth, D. The effect of deleterious mutations on neutral molecular variation. *Genetics* **134,** 1289–1303 (1993).
64. Kulathinal, R. J., Bennett, S. M., Fitzpatrick, C. L. & Noor, M. A. Fine-scale mapping of recombination rate in *Drosophila* refines its correlation to diversity and divergence. *Proc. Natl Acad. Sci. USA* **105,** 10051–10056 (2008).
65. Bystrykh, L. *et al.* Uncovering regulatory pathways that affect hematopoietic stem cell function using 'genetical genomics'. *Nature Genet.* **37,** 225–232 (2005).
66. Wentzell, A. M. *et al.* Linking metabolic QTLs with network and *cis*-eQTLs controlling biosynthetic pathways. *PLoS Genet.* **3,** 1687–1701 (2007).
67. Foss, E. J. *et al.* Genetic basis of proteome variation in yeast. *Nature Genet.* **39,** 1369–1375 (2007).
68. Stylianou, I. M. *et al.* Applying gene expression, proteomics and single-nucleotide polymorphism analysis for complex trait gene identification. *Genetics* **178,** 1795–1805 (2008).
69. Churchill, G. A. *et al.* The Collaborative Cross, a community resource for the genetic analysis of complex traits. *Nature Genet.* **36,** 1133–1137 (2004).
70. Li, Y. *et al.* Mapping determinants of gene expression plasticity by genetical genomics in *C. elegans*. *PLoS Genet.* **2,** e222 (2006).
71. Smith, E. N. & Kruglyak, L. Gene–environment interaction in yeast gene expression. *PLoS Biol.* **6,** e83 (2008).
72. Hubner, N. *et al.* Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease. *Nature Genet.* **37,** 243–253 (2005).
73. Cotsapas, C. J. *et al.* Genetic dissection of gene regulation in multiple mouse tissues. *Mamm. Genome* **17,** 490–495 (2006).
74. Wang, S. *et al.* Genetic and genomic analysis of a fat mass trait with complex inheritance reveals marked sex specificity. *PLoS Genet.* **2,** e15 (2006).

Acknowledgements I thank the Jane Coffin Childs Memorial Fund for Medical Research and New York University for support, and L. Chen for discussion.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The author declares no competing financial interests. Correspondence should be addressed to the author (mrockman@nyu.edu).