Genetics of global gene expression

Matthew V. Rockman and Leonid Kruglyak

Abstract | A new field of genetic analysis of global gene expression has emerged in recent years, driven by the realization that traditional techniques of linkage and association analysis can be applied to thousands of transcript levels measured by microarrays. Genetic dissection of transcript abundance has shed light on the architecture of quantitative traits, provided a new approach for connecting DNA sequence variation with phenotypic variation, and improved our understanding of transcriptional regulation and regulatory variation.

Complex and quantitative traits
Phenotypes that are shaped by multiple and possibly interacting genetic and environmental factors. Quantitative traits (as distinguished from discrete traits) are measured on continuous scales.

Effect size
The magnitude of contribution of a locus to variation in a phenotype.

Lewis-Sigler Institute for Integrative Genomics and Department of Ecology and Evolutionary Biology, Princeton University, Princeton, New Jersey 08544, USA.
Correspondence to L.K.
e-mail: leonid@genomics.princeton.edu
doi:10.1038/nrg1964

More than a century after the rediscovery of Mendel, the genetic basis of complex and quantitative traits resists generalization. Basic questions remain unanswered, including the number of loci that underlie variation in heritable phenotypes, the distribution of their effect sizes, their molecular natures and mechanisms of action and interaction, and their dependence on environmental variables. These questions are at the centre of pressing issues in medical and agricultural genetics, as well as in basic evolutionary biology, in which the outstanding unresolved question concerns the forces that create, maintain and sort heritable phenotypic variation. Now, an emerging approach, genetic mapping of genome-wide gene expression (BOX 1), is beginning to provide the requisite empirical data to address these questions. Since the first empirical linkage study of global transcript levels was published in 2002 (REF. 1), many general principles have been established and represent solid ground on which further work can build.

Although small-scale studies of the genetics of gene expression have a long and rich history (BOX 2), modern large-scale studies owe their existence to the development of microarray technology in the mid-1990s. Microarrays were first applied to the study of genetic variation in 2000. They revealed that gene expression differs between strains in both yeast and mice[2,3] and that such differences segregate in crosses[4,5]. Subsequent studies documented abundant heritable variation in gene expression in *Drosophila melanogaster*[6] and killifish[7]. By the time Jansen and Nap[8] proposed genetic mapping of genome-wide gene expression, such work was well underway in several research groups, and the first empirical study mapping global gene expression in a yeast cross appeared early the following year[1]. Since then, further studies have documented heritable variation in genome-wide gene expression in more than a dozen species and have mapped the loci for many expression traits in yeast, mice, maize, humans, rats, *Eucalyptus* and *Arabidopsis thaliana*[9–21]. This diversity of model systems promises to reveal important connections between genome-wide gene expression and features of population biology — population sizes, breeding systems, demographic histories and patterns of natural selection. Differences can already be seen among species: most species show ubiquitous heritable variation in expression, whereas the malarial parasite *Plasmodium falciparum* shows remarkably little[22].

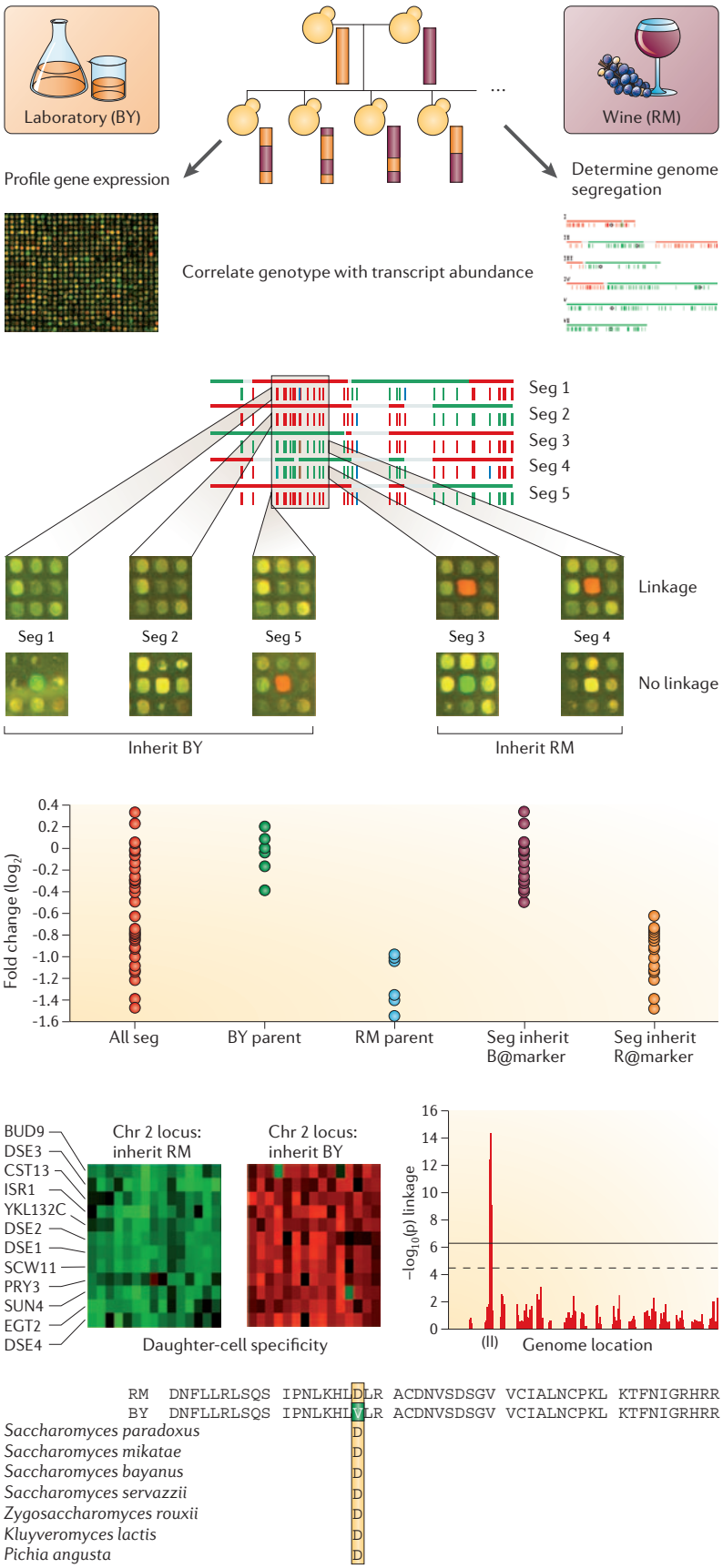
The abundance of a transcript is a quantitative trait and, like all such traits, its inheritance can be described using the classical methods of biometrical genetics and its genetic basis can be discovered using linkage and association mapping. However, transcript abundance is in many ways an extraordinary phenotype, with special attributes that confer particular importance on an understanding of its genetics. The primary transformative potential of genome-wide gene expression genetics is the sheer number of traits — thousands — that can be assayed simultaneously. Whereas studies of one or a few traits offer only anecdotal examples of the underlying genetic architectures, studying thousands of traits allows a detailed description of the distribution over the landscape of all possible architectures. Individual traits are typically preselected on the basis of their phenotypic divergence or biological interest, whereas genome-wide expression studies provide data on a large and unbiased set of traits. The radical increase in the number of traits accessible to study has raised new challenges to analysis and interpretation, and genome-wide genetic mapping of gene expression has consequently become a central proving ground for new statistical genetics techniques[23].

Another special feature of transcript abundance as a phenotype is that it represents the phenotype most immediately connected to DNA sequence variation

Box 1 | **Genetics of global gene expression: a primer**

A study of the genetics of global gene expression begins with a mapping population. Several choices are available and have been used: progeny from a cross between two parent strains, recombinant inbred lines, collections of pedigrees and samples of unrelated individuals. The study population is then genotyped for a set of polymorphic markers that cover the genome. RNA is extracted from each individual or strain, and the abundance of each transcript is measured, typically by hybridization to microarrays (panel **a** illustrates the experimental design for a cross between two yeast strains[1]). High levels of measurement replication are built into the design, because each allele at each QTL will be present in a large number of samples; the effect of the QTL on gene expression will therefore be measured many times. The resulting data set is then analysed to find genetic loci that affect transcript abundance. The simple approach is to treat the abundance of each transcript as a separate quantitative trait, and to carry out conventional linkage or association analysis with all the markers tested either individually or using multipoint methods. Panel **b** shows how, at a given genomic location, the samples are separated according to the inherited marker alleles, and linkage (or association) is declared if the groups differ significantly in expression level (in this case of the gene at the centre of the small array section). An example of an actual linkage from a yeast cross is shown in panel **c**. Because millions of statistical tests are carried out (thousands of transcripts are each tested against hundreds or thousands of markers in linkage studies and up to a million markers in association studies), careful control of false positives through multiple testing correction is essential. Empirical significance levels obtained through permutation tests[40,96] should be used whenever possible to account for the complex correlations in the data. Approaches that are based on the false-discovery rate (FDR) are very useful[97]. Methods are available to search for multiple loci that affect a given trait either additively or through interactions, but even greater statistical care is required[23,38,39,98]. Methods that use dimensional reduction (for example, clustering or principal component analysis) can lower the number of tests and can potentially improve mapping power by combining multiple transcripts that behave similarly into single traits[9,99].

Depending on the study design and sample size, loci are identified for anywhere from a few transcripts to thousands of transcripts, with up to half of all transcripts showing linkage in some studies[29]. Following locus identification, the studies take several paths. The genetic architectures of the traits can be assessed by examining heritabilities, detection rates, numbers of loci identified and their effect sizes[29]. Loci can be classified according to whether they affect transcripts encoded at the same genomic location or elsewhere in the genome, and whether they affect many transcripts or few[1,10]. Bioinformatic approaches can be used to identify functional relationships among the transcripts affected by common loci and to investigate the structure of the underlying regulatory networks[1,13,76,100,101]. Finally, the causative polymorphisms responsible for variation in transcript abundance can be pursued using molecular genetics tools (panel **d**). (B@marker, R@marker, the allele carried by the segregant at the marker is the BY allele (B) or the RM allele (R), respectively; chr, chromosome; seg, segregants.)



Panel b and c reproduced with permission from REF. 1 © (2002) American Association for the Advancement of Sciences.

Box 2 | **Foundational work in the genetics of natural variation in gene expression**

Dramatic recent progress in understanding genetic variation in gene expression builds on a rich history, dating at least to Haldane's prescient treatment of variation in the timing of gene activity[102]. Decades before the discovery of the molecular gene, Haldane recognized that genetic variation in a gene's activity could be due to variation in the gene itself or to variation at an unlinked locus. Thinking in terms of physiological genetics, he pointed to membrane-spanning ion channels as candidate sites for mutations that could act on other genes in *trans* by altering cellular pH.

The earliest genetic mapping of natural variation in gene expression appeared immediately after Jacob and Monod[103] introduced a mechanism for gene regulation. Schwartz[104] showed that variation in the time of activity of an esterase gene in maize depends on variation that is tightly linked to the structural locus. Linkage disequilibrium between the regulatory locus and electrophoretically distinguishable alleles of the enzyme showed that the regulatory variation acted in *cis*.

Jacob and Monod's model inspired much speculation about variation in gene regulation and its importance to evolution[105,106], but empirical research really took off only after King and Wilson's analysis refocused attention on the subject[24]. Using amylase genes as models, drosophilists documented distant linking, *trans*-acting regulatory variation affecting spatial regulation[107,108], heritable variation in enzyme inducibility[109] and fitness differences attributable to regulatory variation[110]. Parallel results for other enzymes and taxa rapidly accumulated[111–113].

Genetic analysis of gene expression variation entered a new era with the precocious introduction of genomic approaches by Damerval *et al.*[61] Scoring spot intensities on two-dimensional gels as measures of abundance of 72 proteins in the F2 progeny of a maize line cross, Damerval *et al.* mapped QTLs that underlie the observed variation, including epistatic interactions among the loci. Because the two-dimensional gels allow allele-specific protein quantification, the study could distinguish between *cis*-acting and *trans*-acting variation. In the past several years, genome-wide genetic analyses of gene expression have far surpassed Damerval *et al.* in the number of traits examined and in the number and precision of the QTLs discovered, but much of the conceptual model for these studies is present in their classic paper.

— the road from genotype to phenotype runs through gene expression. Regulatory sequence variation, including both variation in a regulatory region of a gene that affects its own expression and variation in the coding region of a gene that affects expression of other genes, is probably the main mediator of phenotypic divergence in evolution[24–27]. Moreover, the intermediate position of gene expression between genotype and organismal phenotype makes it ideally suited to serve as a bridge between the two in mapping studies. Genetic correlations between expression phenotypes and organismal phenotypes point to the molecular pathways that underlie the organismal phenotypes, whereas colocalization of QTLs for expression and organismal phenotypes speeds up the identification of causal mutations. Gene expression also provides a universal subphenotype for complex and heterogeneous organismal phenotypes[10].

In this review, we first discuss what the genetics of global gene expression has taught us about the genetic architecture of quantitative traits. We then describe the features of the two types of regulatory sequence variation that underlie differences in gene expression: local variation, which maps close to the physical location of the affected gene, and distant variation, which maps elsewhere in the genome. We quantify the prevalence of each type, and delineate the difference between the local and distant distinction, which is based on location, and the mechanistic distinctions (such as *cis*-acting and *trans*-acting, or *cis*-regulatory and protein-coding), which are based on the function altered by the variants. Having discussed the insights provided by studies of global gene expression, we look at the future of the genetics of global molecular phenotypes.

Genetic complexity of transcript levels

Historically, most quantitative phenotypes have proved to be genetically complex, explicable only by multiple underlying loci and possibly interactions among the loci and with environmental variables. A key finding from multiple studies of the genetics of gene expression in different species is that complex inheritance is also consistently observed for the thousands of transcript-level traits. Despite their close connection to DNA sequence, transcript abundances exhibit substantial genetic complexity.

QTL number and effect size. Two important and seemingly simple questions are: how many QTLs underlie a quantitative trait, and how much of the heritable variation in the trait does each QTL explain? These questions turn out to be surprisingly difficult to answer because of the following methodological problem. Unless a study samples a very large number of individuals, which has not been practical when the phenotype is genome-wide gene expression, only QTLs with the largest effect on the trait can be detected. Therefore, the observed number of loci is usually an extreme underestimate of the actual number, and the observed effect sizes represent the high end of the overall distribution of effect sizes. Moreover, a common statistical artefact, known as the Beavis effect, causes the overestimation of effect sizes of the detected loci[28]. Nevertheless, useful estimates can be made.

The genetic architecture of most expression traits involves multiple QTLs, and most of these QTLs explain a minority of trait variation. Because all mapping studies until now have detected only a single locus for most traits, this conclusion is based primarily on what has not been detected. The argument runs as follows. For a trait with measurable heritability, it is straightforward to calculate the probability that a QTL that explains a certain fraction of trait variation will be detected in a study with a given sample size. For a single trait, a QTL is either detected or not, but when many traits are examined simultaneously, as in the case of genome-wide expression, we can ask whether the fraction with detected QTLs is as expected from the detection probability. If the detected fraction is

Recombinant inbred lines
Panels of genetically mosaic but homozygous strains generated by crossing parental strains and inbreeding the progeny.

False-discovery rate
The fraction of results declared significant at a given threshold that are expected to be false positives.

Dimensional reduction
A class of mathematical techniques for summarizing the main characteristics of multivariate data with fewer variables.

QTL
Quantitative trait locus; a region of the genome that contributes to variation in a quantitative trait.

Beavis effect
A statistical artefact that is due to the deviation of estimates from true values by random error. In a mapping experiment, the loci that are deemed significant are enriched for those in which the estimated effects benefit from random error that happens to fall in the right direction. Therefore, significant QTLs are disproportionately those in which the effect sizes are inflated by chance.

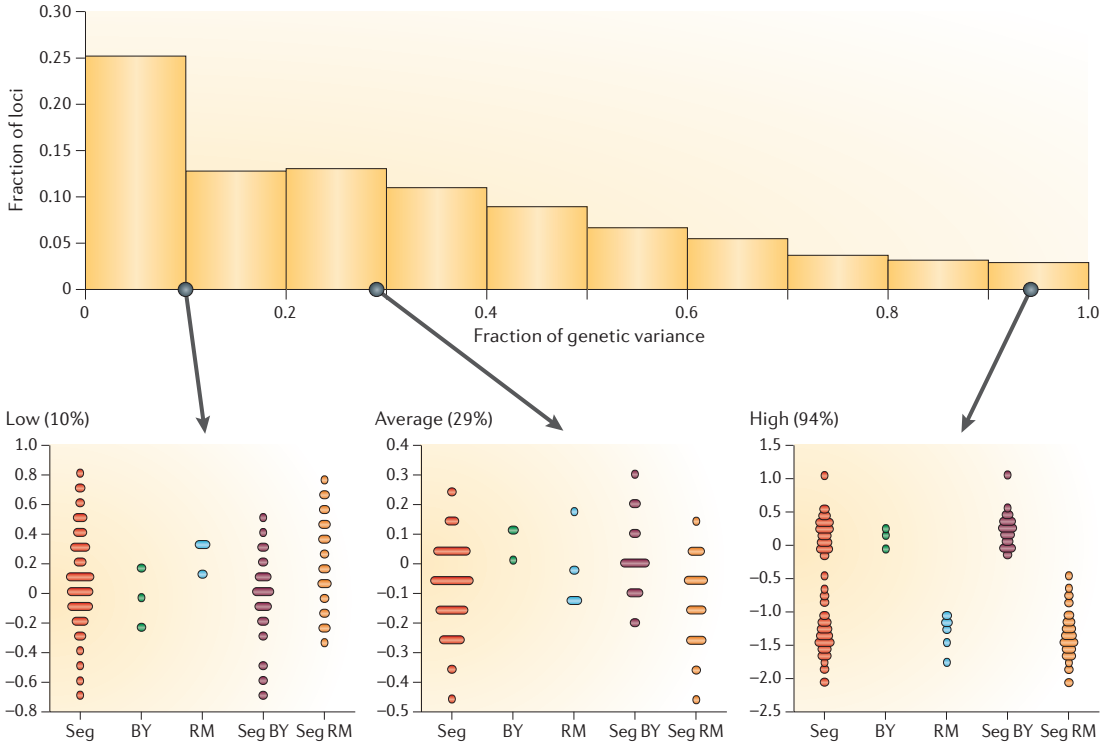


Figure 1 | **Most gene expression traits are affected by multiple loci.** Each bar represents the fraction of QTLs that explain a percentage of genetic variance in the range on the *x* axis. For each trait with significant linkage(s), only the single most significant QTL is included. Data are derived from the first table in REF. 29. The panels below the plot show examples of QTLs that explain, from left to right, low, average and high percentages of genetic variance. In each panel, the left-most column shows the relative expression of the corresponding gene in all 112 segregants (seg), the next two columns show the expression in replicates of the two parent strains, and the last two columns show the expression in the segregants that inherit the QTL allele from the first and second parent strains.

lower, typical QTLs must be weaker than assumed in the calculation, and because no individual QTL can explain most of the genetic variation, there must be multiple QTLs. Such arguments have been used to estimate that in a yeast cross only 3% of expression traits are consistent with single-locus inheritance, that most traits require more than two additive QTLs, and that segregation of many traits can only be explained by very complex genetics[29]. In an F2 cross of 111 mice, QTLs were detected for only 27% of genes with significant genetic differences in expression, implying considerable genetic complexity given the study's high power to detect QTLs for traits with relatively simple inheritance[10]. A combination of genetic complexity and low statistical power probably explains the low detection rates in the human studies carried out so far[16–19].

Direct evidence of genetic complexity comes from detecting multiple QTLs for at least some expression traits. Moreover, even the detected QTLs typically explain only a minority of trait variation. In yeast, the median phenotypic effect of a detected QTL was 27% of genetic (heritable) variance explained, and only 23% of traits had a QTL that explained >50% of genetic variance[29] (FIG. 1). Similarly, in mice[10], mapped QTLs explained on average 25% of the variance in expression of the corresponding genes. In humans, effect-size estimates averaging 27–29% have been reported for loci near the affected genes[17,19]. Therefore, even the strongest QTLs that underlie variation in gene expression typically explain a quarter or less of the variation. Nevertheless, routine observation of QTLs that explain a substantial fraction of phenotypic variance contradicts the infinitesimal theory of Fisher[30], in which quantitative traits are determined by a very large number of loci with very small effects. QTLs with appreciable effects are also reported in studies of non-expression phenotypes[31] and cannot all be explained away by the Beavis effect. The empirical results are better described by the exponential model developed by Orr[32], in which large-effect mutations represent the expected initial steps in adaptation, as discussed by Farrall[31] and by Barton and Keightley[33]; one implication is that strong QTLs might be observed because of natural selection, and not despite it.

The many forms of genetic complexity. The correspondence between genes, alleles and environments on the one hand, and phenotypes on the other, can be complex in many ways, with a parallel proliferation in terminology. Studies of the genetics of global gene expression have illuminated the prevalence of each type of complexity (FIG. 2). In yeast, most heritable transcripts show transgressive segregation[29], which is consistent with the accumulation

Heritability
The fraction of total phenotypic variance that is attributable to additive genetic effects. Estimators with different technical definitions and biological meanings abound. This is not an inherent property of a trait; heritability depends on the nature of the genetic sample (for example, intercross, inbred lines, twins and random populations) and the space of environments surveyed.

Transgressive segregation
A distribution of trait values for a segregating population that extends significantly beyond the range defined by the progenitor strains.

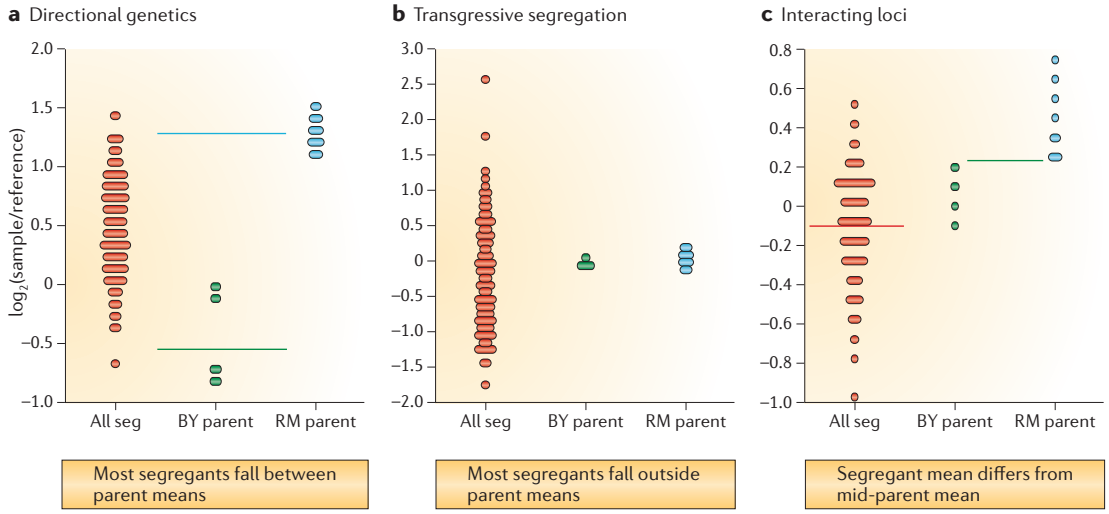


Figure 2 | **Types of complex inheritance of transcript levels. a** | Directional genetics with most segregants (seg) showing expression between the two parent values. **b** | Transgressive segregation with most segregants showing expression outside the two parent values. **c** | Genetic interaction with the segregant average expression differing from the mid-parent mean expression. In each panel, the first column shows the relative expression of the corresponding gene in all 112 segregants, and the next two columns show the expression in replicates of the two parent strains. Modified with permission from REF. 29 © (2005) National Academy of Sciences, USA.

in the parental strains of a large number of alleles of modest effect. A smaller fraction of traits show directional genetics[29], which is suggestive of directional selection that drives the phenotypic divergence of the parental strains. Several studies[34–37] have investigated the prevalence of non-additivity, where gene expression in F1 heterozygotes differs from the mid-value of the homozygous parental strains. Questions of additivity depend on the measurement scale — values can be additive on some scales but non-additive on others. Moreover, microarrays yield linear measures of gene expression only within certain ranges. Despite these caveats, experimental follow-up has validated the finding from microarrays that non-additivity is common in *D. melanogaster*, *A. thaliana* and maize, and that its extreme forms, overdominance and underdominance, are not rare[34–37].

Genetic interactions have been observed in several studies, and a systematic scan for interacting QTLs found non-additive interactions among loci for roughly half of all transcripts[38]. The detection of interacting QTLs will be aided by methodological advances[39], as well as by future studies with much larger sample sizes[40]. The detection of population association at only a minority of loci previously identified by linkage in families[18] indicates that many individual QTLs which underlie expression traits display considerable allelic heterogeneity. The effects of genetic variation on gene expression are condition-dependent, and such gene-by-environment interactions have been documented in comparisons of inbred strains across conditions[6,41–43]. In multicellular organisms, the local conditions differ in each tissue, and genetic variation with a cell-type-dependent influence on gene expression represents a special case of gene-by-environment interaction. Studies of gene expression in mouse brain[13], haematopoietic stem cells[12], fat[44] and liver[10,44], and in rat kidney and fat[11], have found that the genetic basis of variation in a gene's expression is sometimes shared between different tissues but is often unique to each tissue[45]. Studies in flies and mice have also shown extensive sex dependence of gene expression[6,21]. Finally, the existence of loci that affect the expression of many genes (see below), or individual genes across many conditions, establishes pleiotropy as a common feature of the genetics of gene expression.

Local versus distant QTLs

Transcript levels differ from other phenotypes in that each transcript has a corresponding encoding gene with a known position in the genome. Therefore, because mapping studies reveal the locations of QTLs, an expression QTL can be immediately classified as 'local' (near the genomic location of the gene encoding the transcript) or 'distant' (elsewhere in the genome). Because mapping studies do not reveal the underlying molecular nature of QTLs, we prefer the strictly positional terms local and distant to the commonly used terms *cis*- and *trans*-linking, which have implicit mechanistic connotations. In fact, both local and distant QTLs can include polymorphisms in *cis*-acting and *trans*-acting factors, under the classic definition of the terms[46–48]. The casual conflation of different usages of *cis* and *trans* has resulted in a significant amount of confusion. Some uses describe the pattern of co-inheritance of trait and locus (-linking), whereas others describe the mechanism of action of a locus with respect to a trait (-acting). Neither distinction captures any information about the molecular nature of the loci, that is, whether the mutation alters a protein or a functional RNA or acts at the level of DNA in *cis*-regulation.

Directional genetics
A distribution of trait values for a segregating population that is significantly concentrated within the range defined by the progenitor strains.

Non-additivity
A property of alleles at a locus, such that the trait value of heterozygous individuals is not the average of the trait values of homozygotes for each allele.

Genetic interaction
A property of alleles at different loci, such that their combined effect on a phenotype deviates from the sum of their individual effects (this is often called epistasis).

Allelic heterogeneity
The phenomenon in which a genetically diverse population harbours many different alleles at a QTL.

Gene-by-environment interaction
The effect of a locus on a trait depends on the environment, and the effect of the environment on the trait depends on the locus.

Pleiotropy
The capacity of a single mutation to affect multiple traits.

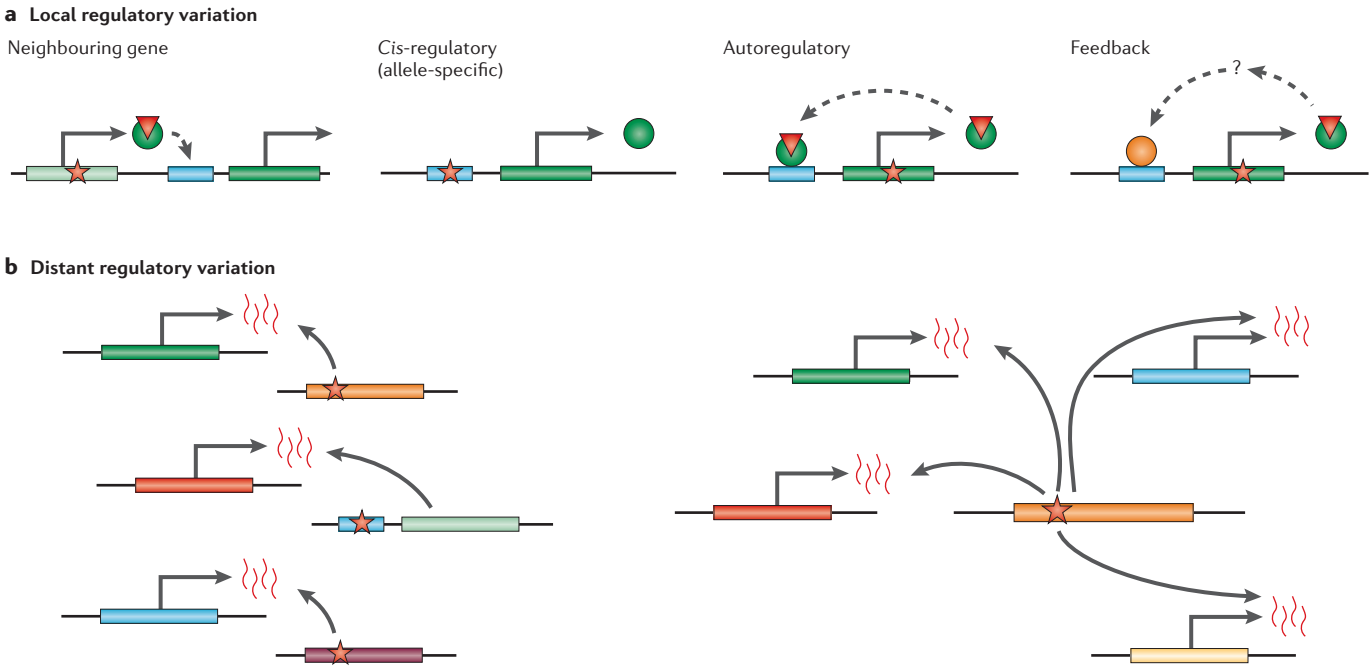


Figure 3 | **Local and distant regulatory variation. a** | Local regulatory variation. From left to right, regulation by a neighbouring gene, *cis*-regulatory variation, autoregulatory variation and feedback variation. The red star denotes the regulatory variant, the bar with arrow denotes the coding region of the gene (dark green), the expression of which is affected by the variant, and the circle denotes the protein product of the gene. The blue rectangle to the left of the coding region represents an upstream regulatory element. **b** | Distant regulatory variation. Arrows are drawn from a gene with regulatory variation to the gene or genes it regulates. On the left, each regulatory variant affects expression of a different gene. On the right, a regulatory variant affects expression of many genes. Such a variant would show up as a linkage hot spot (FIG. 4).

Some arbitrariness is introduced by the definition of a marker being local or distant. Local is most commonly defined as being within some physical distance of a chosen point in the gene. Statistical arguments can be used to choose this distance so that the probability that a linked marker will fall that close to the gene by chance is small[1]. More sophisticated definitions can include considerations of whether the gene is close to the linkage peak: for example, does it fall within the confidence region for the linkage localization[49]?

Local linkage can arise as a result of several scenarios (FIG. 3a). First, the linkage might be due to a polymorphism in a nearby gene that regulates the gene for which expression is being measured. This can occur either by chance or owing to a nonrandom close location in the genome of regulators and their targets. Statistical arguments can be used to ensure that there is only a small probability of a chance occurrence. Second, and more typically, local linkage will be due to a polymorphism (or polymorphisms) in the gene itself. Such polymorphisms might act in *cis* by altering classic *cis*-acting regulatory elements and consequently changing transcription; they could also act in *cis* post-transcriptionally, by altering message stability or by altering sites that are targets for messenger RNA processing and decay. Polymorphisms in the gene might also act in *trans* (that is, affecting the expression of both alleles in a heterozygous diploid) by triggering feedback loops either directly through changes in the coding sequence of an autoregulatory gene or indirectly by changing the coding sequence or message levels that are sensed and responded to by the cell. Classic *cis*/*trans* tests of allele-specific expression in a diploid hybrid can be used to distinguish among these possibilities[49,50].

Distant regulatory variation typically acts in *trans* through the downstream effects of coding or *cis*-regulatory polymorphisms in different types of genes, with transcription factors being the most obvious example. But distant regulation can occur with many degrees of indirectness, including non-cell-autonomous effects and beyond (for example, a polymorphism that influences diet choice would affect the expression of genes that respond to specific nutrients and exogenous molecules). Moreover, the existence of regulatory elements that are located far from the genes they regulate[51–53] means that distant loci can act in *cis*.

Even more complications arise from DNA elements that act directly to regulate distant genes through physical contact; interchromosomal interactions can result in '*cis*-regulatory' DNA that acts on a different chromosome[54]. A genetically diffuse version of this phenomenon is due to variation in the total genome-wide number of binding motifs for a transcription factor; the binding of the factor to non-functional '*cis*-regulatory' DNA titrates the proteins out of the nucleoplasm, decreasing their availability with effects in *trans*[55].

Box 3 | Units of measure influence contributions to variation in gene expression

An explicit statistical model is central to any attempt to address the relative contributions of different kinds of QTLs to phenotypic variation. But apart from the problem of what qualifies as significant is the problem of what units are being counted. We might want to count the number of gene expression traits that show each kind of genetic basis, or we might want to count the number of genetic loci that underlie the inferred linkages; a major source of confusion is that both traits and loci are often called 'genes'. In other contexts, the quantity of interest might be the number of linkages, in which case a single genetic locus could count multiple times if it pleiotropically affects multiple gene expression traits. Because linkages are often called QTLs, counts of linkages and loci are readily confused. Yet another unit of measure is the fraction of phenotypic variance that is attributable to a particular type of genetic variation. Most studies describe results for only some of these units, hindering comparisons among studies. Interpretation also depends on whether we are measuring trait-by-trait averages or genome-wide global quantities. For example, for each gene expression trait, most loci will be distant, but on a genome-wide scale, the majority of loci might be local[9], owing to the pleiotropic effects of distant loci.

Another variable among studies is the phenotypic space over which the questions are asked. The sampling properties of networks are underexplored[114] but relevant to claims about the genetic architecture of gene expression, as counts of traits, loci, linkages and variances have different ranges and scales. For example, the number of loci segregating in a cross is finite, whereas the number of possible traits (and therefore linkages) is technically infinite[115]. The expression of each gene is a suite of traits distributed over a space of environments, continuous in some dimensions, such as temperature, and discontinuous in others, such as cell type. At present we know very little about how the counts of loci (or linkages or variances) scale with the extent of phenotypic state space explored. However, in the case of local versus distant loci there is a clear bias. The maximum number of local loci discovered is limited by the space of traits examined because current genome-wide linkage methods are unable to resolve multiple locally linking loci for a single trait. As the number of traits examined increases, the estimated genome-wide number of local loci might increase linearly, whereas the number of distant loci will increase less than linearly as particular pleiotropic loci are discovered again and again. One consequence is that estimated ratios of local versus distant loci are biased by the number of sampled traits. For example, Morley *et al.*[17] sampled just 3,554 human gene expression traits; if these traits were downstream of the major pleiotropic *trans*-acting loci, then most of the genome's distant QTLs will have been identified, whereas, at best, less than 20% of transcripts (assuming 20,000 genes) were examined for local linkage.

A final variable is the nature of the genetic sample, that is, the genotypic space surveyed. So far, most studies have involved simple line crosses in which two alleles at most are segregating at each locus. Several more recent studies have used association methods in larger population samples, where more alleles could be segregating. If the loci that underlie different kinds of linkages have different allele frequency distributions[116], different kinds of sample will yield different answers to questions about their prevalences and contributions to variation[117].

Empirical results: local regulatory variation. As many as 25% of all gene expression traits in a yeast cross are affected by local regulatory variation[49]. Local linkages have been observed to account for anywhere from 25% (for example, see REF. 9) to 100% (for example, see REF. 19) of detected loci. This range is primarily a function of sample size, as local linkages on average explain more trait variance than distant linkages, and are therefore more likely to be detected in smaller studies[10,56]. In addition, the power to detect local linkages is higher than it is for distant linkages because the multiple testing problem is less severe. In both cases thousands of transcripts are tested for linkage, but in the case of distant linkages many markers that cover the entire genome are tested for linkage to each transcript, whereas in the case of local linkage only markers near the gene encoding the transcript need to be tested. The difference is especially acute for association studies, where up to a million markers are used to cover the genome[18,19].

Several types of polymorphism in different locations, in or near a gene, and acting through different molecular mechanisms, can cause local variation. It seems that most but not all local regulatory variation acts in *cis*, with perhaps a quarter to a third acting in *trans*[49,50]. *Cis*-acting local variation can result from differential transcription, splicing, mRNA decay or even gene copy number; each phenomenon was observed in a mouse cross[10]. A variant in the *AMN1* gene in *Saccharomyces cerevisiae* provides an example of a local but *trans*-acting effect of a coding polymorphism on its corresponding transcript[49].

Because transcript levels with local linkages are most likely affected by regulatory variation in the corresponding gene, each local linkage comes with an immediate candidate gene, removing the need for cumbersome positional approaches to gene identification. However, finding functional polymorphisms is still not straightforward, as genes are likely to contain multiple polymorphisms, and both coding and regulatory regions need to be examined. Linkage and linkage disequilibrium between nearby polymorphisms further complicate the identification of functional polymorphisms, and the effect on expression might be due to haplotypes that combine multiple alleles. Therefore, the path from local QTLs to QTNs (quantitative trait nucleotides) remains tortuous and dependent on targeted experiments.

A recent and unexpected finding from *A. thaliana* is the 'neighbourhood effect', whereby a large number of physically clustered genes show linkage to a local QTL[14]. These hot spots of local linkage might be due to tight clustering of functionally related genes, or to coincidental colocalization of a *trans*-acting variant and its target genes, but a plausible alternative is that the QTLs represent mutations that alter the regional structure of chromatin, acting in *cis* to influence a large number of genes.

Multiple testing problem
The number of false-positive results increases when multiple statistical tests are carried out, requiring more stringent thresholds to reach the same level of significance.

Linkage disequilibrium
The nonrandom association of alleles at different loci in a population.

QTN
Quantitative trait nucleotide; the actual sequence polymorphism responsible for variation in a quantitative trait.

In parallel with mapping studies, allele-specific measurement techniques, which pinpoint genes that have variable transcription due to *cis*-acting allelic variation, have been addressing the same questions. Five assay techniques — allele-specific quantitative PCR[49,50,57–59], measurement of polymerase loading[60], quantitative two-dimensional protein gels[61,62], allele-specific expression arrays[63–65] and experimental reporter assays[66–69] — have all documented abundant *cis*-acting genetic variation (see REF. 70 for a review of these techniques). Only the array-based methods approach genome-wide scope, and the results are striking and consistent. In each case a large fraction of variable traits exhibit *cis*-acting variation. Pant *et al.*[65], by measuring allele-specific expression of 1,389 genes in human white blood cells, found that more than half of the genes exhibited *cis*-acting heterozygosity in a sample of 12 individuals, and that on average individuals exhibit *cis*-acting heterozygosity at 25% of the genes. This high number is compatible with an earlier extrapolation from a survey of reporter assays[69]. Reporter assays have allowed the dissection of *cis*-regulatory haplotypes, and have established that the single local QTL that is found for a gene by linkage or association might represent haplotypes that differ by many variants with individual and interacting effects on expression[71,72].

Empirical results: distant regulatory variation. In studies with larger sample sizes, most transcripts link to loci distant from the genomic locations of the genes that encode the corresponding transcripts. For example, Yvert *et al.*[9] found that 578 of 2,294 expression traits show linkage to the genomic regions at which they are transcribed, whereas 1,716 link to distant loci. An estimated 100–200 loci accounted for these distant linkages. Because a particular locus can influence many distant genes (FIG. 3b), the number of distant linkages is often much higher than the number of local linkages, even though, as in this case, the number of loci that account for local linkages is much higher than the number of distant loci (BOX 3).

Because the resolution of the linkages tends to be relatively low, it is difficult to estimate the number of loci that affect gene expression of distant transcripts in any given study, as well as the number of transcripts affected by a typical locus. However, one common feature observed in multiple studies is the presence of hot spots: individual loci that affect large numbers of transcripts[1,10]. Hot spots are usually defined as those loci for which the number of linked or associated transcripts statistically significantly exceeds that expected if such transcripts were randomly distributed along the genetic map[1] (FIG. 4). Care must be taken to correct for the fact that many loci are examined, and for correlation in expression levels of different genes that might cause apparent hot spots in the absence of underlying common polymorphisms[73]; in other words, a false-positive QTL for one trait will be a false positive for all correlated traits.

Although hot-spot QTLs are often called 'master regulators'[17], the QTLs themselves are simply mutations that segregate in populations, and their hot-spot status comes from the pleiotropic effects of the mutations rather than from any necessary regulatory function of the genes. A mutation in an essential structural protein with no regulatory role is the ultimate hot spot: a locus that acts in *trans* to reduce the expression of every gene to zero (by means of lethality). Of course, lethality is an extreme example, and in general mutations in genes at the top of regulatory hierarchies are promising candidates for hot-spot QTLs.

Finding the genes that underlie distant loci is more challenging than in the case of local loci. Several approaches might prove to be helpful. Some polymorphisms also affect the expression of the gene that contains them, with the effects on other genes being either pleiotropic or a direct consequence of the change in expression. Therefore, looking for a local linkage among the transcripts linking to a hot spot might lead directly to the causative gene. In mouse, for example, a local QTL influencing expression of the transcription factor gene *Runx1* acts as a distant QTL for genes known to be targets of *Runx1* regulation[12]. Combining the transcripts affected by a hot spot into a single phenotype (for example, by clustering) can increase resolution and make the phenotype simpler (more monogenic) and therefore more amenable to recombinational fine mapping[9]. More sophisticated approaches that leverage the correlation structure among phenotypes — for example, by finding a function of multiple traits that maximizes support for a pleiotropic QTL — also hold promise[74–77].

Because many regulatory and biochemical networks are well characterized, the identity and annotation of the transcripts that link to a hot spot can greatly assist candidate gene identification[1]. Studying gene expression and growth rates in *Eucalyptus*, Kirst *et al.* identified a suite of genes of which the expression explained much of the variance in tree growth[15]. This set of genes included nearly the entire lignin biosynthesis network, and their expression traits share a set of common QTLs. Although

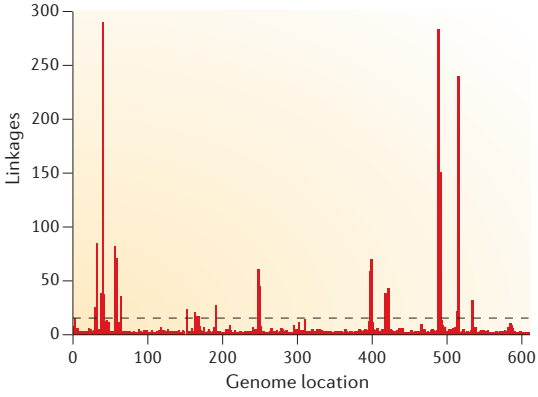


Figure 4 | **Hot spots of distant regulatory variation.** The number of linkages is plotted against genome location. The yeast genome has been divided into 611 bins of 20 kb, shown in chromosomal order on the *x* axis. The number of distinct transcripts linked to markers in each bin is shown by bar height. Local linkages are excluded. The dashed line shows the maximum number of linkages that would be expected to fall into any one bin by chance with a probability of >5%, corrected for the number of bins. Bins with bar heights above this line represent hot spots.

a *Eucalyptus* genome sequence is unavailable, Kirst *et al.* mapped some of the lignin biosynthesis genes and found that one, *S-adenosylmethionine synthase*, coincides with the major QTL for both growth and the suite of lignin biosynthesis genes.

A natural assumption is that distant linkages that underlie gene expression variation are due to polymorphisms in transcription factors, the archetypal *trans*-acting transcriptional regulators. The hypothesis is difficult to test without mapping QTLs to the resolution of single genes, but genome-wide expression mapping provides sufficient data to allow confident inferences. In a yeast cross, known yeast transcription factors were not over-represented near the markers that defined QTLs for 1,716 distant linkages, and an analysis of Gene Ontology categories indicated that QTLs are not enriched for any particular class of molecular function[9] (these questions are revisited in REF. 77).

The future

Much has been learned about the genetics of global gene expression in the past few years. However, the field is young and much remains to be discovered. So far, most of the conclusions are based on the identification of genetic loci that have not been resolved into individual genes and polymorphisms that affect expression. Although a few examples of detailed molecular characterization of expression QTLs exist (for example, see REF. 9), many more are needed for a clear understanding of the types of allele that are responsible for genetic differences in gene expression, and of the location and nature of the causative polymorphisms. Such high-throughput identification and characterization of polymorphisms that affect expression remains a challenge. Promising research directions include greatly increasing sample sizes of studies by taking advantage of high-throughput techniques for genotyping[78,79] and expression profiling, as well as developing techniques for rapid and comprehensive mutation detection[80], and for exchanging alleles between strains and directly testing the effects on expression[81,82].

Two other major directions involve expanding the genetic samples and the set of global molecular phenotypes. So far, most studies have used linkage analysis in experimental crosses, recombinant inbred lines or reference human pedigrees. It is important to extend studies to population samples in order to ask questions about allele frequencies and selective forces[83]. In those cases in which gene expression variation has been mapped to the level of nucleotides, population genetic analysis has revealed an important role for natural selection in shaping and maintaining variation[84–93]. Population studies have begun to be carried out, but until now have been limited to humans and have not yet used global expression assays[18,19]. As global expression analysis merges with population genomics, we anticipate a fuller understanding of the causes of heritable variation[33,83].

The expansion of global molecular phenotyping includes measuring gene expression in multiple environments, developmental stages, and cell and tissue types. Emerging technologies also open up the ability to examine other molecular components of cells, including proteins[62] and metabolites[94]. Global phenotyping can also include characterizing cellular properties and responses to large classes of external perturbations, such as small molecules[95]. Where possible, multiple classes of global phenotypes should be collected for the same study samples, so that connections might be drawn between the multiple levels of phenotypic information, with the goal of achieving an understanding of how changes at the DNA sequence level are translated into changes at the level of organismal phenotypes through changes in the intermediates. Clearly, exciting times lie ahead.

1. Brem, R. B., Yvert, G., Clinton, R. & Kruglyak, L. Genetic dissection of transcriptional regulation in budding yeast. *Science* **296**, 752–755 (2002).
The first study to report genetic mapping of global gene expression measured with microarrays. The authors described local and distant linkages and expression QTL hot spots, and showed that the inheritance of transcript levels is usually complex.
2. Sandberg, R. *et al.* Regional and strain-specific gene expression mapping in the adult mouse brain. *Proc. Natl Acad. Sci. USA* **97**, 11038–11043 (2000).
3. Primig, M. *et al.* The core meiotic transcriptome in budding yeasts. *Nature Genet.* **26**, 415–423 (2000).
4. Cavalieri, D., Townsend, J. P. & Hartl, D. L. Manifold anomalies in gene expression in a vineyard isolate of *Saccharomyces cerevisiae* revealed by DNA microarray analysis. *Proc. Natl Acad. Sci. USA* **97**, 12369–12374 (2000).
5. Karp, C. L. *et al.* Identification of complement factor 5 as a susceptibility locus for experimental allergic asthma. *Nature Immunol.* **1**, 221–226 (2000).
6. Jin, W. *et al.* The contributions of sex, genotype and age to transcriptional variance in *Drosophila melanogaster*. *Nature Genet.* **29**, 389–395 (2001).
7. Oleksiak, M. F., Churchill, G. A. & Crawford, D. L. Variation in gene expression within and among natural populations. *Nature Genet.* **32**, 261–266 (2002).
8. Jansen, R. C. & Nap, J. P. Genetical genomics: the added value from segregation. *Trends Genet.* **17**, 388–391 (2001).
A theoretical proposal to combine linkage analysis with expression profiling.
9. Yvert, G. *et al.* *Trans*-acting regulatory variation in *Saccharomyces cerevisiae* and the role of transcription factors. *Nature Genet.* **35**, 57–64 (2003).
This detailed analysis of distant regulatory variation identified two polymorphisms that are responsible for expression QTL hot spots and showed that distant regulatory loci are not enriched for transcription factors.
10. Schadt, E. E. *et al.* Genetics of gene expression surveyed in maize, mouse and man. *Nature* **422**, 297–302 (2003).
This first study of the genetics of global gene expression in a mammal focused on a mouse cross and showed how expression traits can be connected with a physiological phenotype such as obesity.
11. Hubner, N. *et al.* Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease. *Nature Genet.* **37**, 243–253 (2005).
12. Bystrykh, L. *et al.* Uncovering regulatory pathways that affect hematopoietic stem cell function using 'genetical genomics'. *Nature Genet.* **37**, 225–232 (2005).
13. Chesler, E. J. *et al.* Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function. *Nature Genet.* **37**, 233–242 (2005).
References 11–13 describe extensive studies of the genetics of global gene expression in recombinant inbred lines of rodents.
14. DeCook, R., Lall, S., Nettleton, D. & Howell, S. H. Genetic regulation of gene expression during shoot development in *Arabidopsis*. *Genetics* **172**, 1155–1164 (2006).
15. Kirst, M. *et al.* Coordinated genetic regulation of growth and lignin revealed by quantitative trait locus analysis of cDNA microarray data in an interspecific backcross of eucalyptus. *Plant Physiol.* **135**, 2368–2378 (2004).
16. Monks, S. A. *et al.* Genetic inheritance of gene expression in human cell lines. *Am. J. Hum. Genet.* **75**, 1094–1105 (2004).
17. Morley, M. *et al.* Genetic analysis of genome-wide variation in human gene expression. *Nature* **430**, 743–747 (2004).
References 16 and 17 extended linkage studies of global gene expression to humans by carrying out expression profiling in cell lines from previously genotyped CEPH reference families.
18. Cheung, V. G. *et al.* Mapping determinants of human gene expression by regional and genome-wide association. *Nature* **437**, 1365–1369 (2005).
19. Stranger, B. E. *et al.* Genome-wide associations of gene expression variation in humans. *PLoS Genet.* **1**, e78 (2005).
References 18 and 19 reported the first population association studies of expression of a subset of human genes measured in cell lines from CEPH individuals previously densely genotyped at over a million SNPs by the International HapMap Project.

20. 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).
21. 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).
22. Llinas, M., Bozdech, Z., Wong, E. D., Adai, A. T. & DeRisi, J. L. Comparative whole genome transcriptome analysis of three *Plasmodium falciparum* strains. *Nucleic Acids Res.* **34**, 1166–1173 (2006).
23. Kendziorski, C. & Wang, P. A review of statistical methods for expression quantitative trait loci mapping. *Mamm. Genome* **17**, 509–517 (2006).
24. King, M. C. & Wilson, A. C. Evolution at two levels in humans and chimpanzees. *Science* **188**, 107–116 (1975).
This classic paper proposed that regulatory mutations that alter gene expression have a key role in biological differences between species.
25. Stern, D. L. Evolutionary developmental biology and the problem of variation. *Evolution* **54**, 1079–1091 (2000).
26. Wray, G. A. *et al.* The evolution of transcriptional regulation in eukaryotes. *Mol. Biol. Evol.* **20**, 1377–1419 (2003).
27. Carroll, S. B., Grenier, J. K. & Weatherbee, S. D. *From DNA to Diversity: Molecular Genetics and the Evolution of Animal Design* (Blackwell Science, London, 2001).
28. Beavis, W. D. in *Molecular Dissection of Complex Traits* (ed. Patterson, A. H.) 145–162 (CRC, New York, 1998).
29. 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).
A detailed analysis of complex genetic architectures of expression traits.
30. Fisher, R. A. *The Genetical Theory of Natural Selection* (Oxford Univ. Press, Oxford, 1930).
31. Farrall, M. Quantitative genetic variation: a post-modern view. *Hum. Mol. Genet.* **13**, R1–R7 (2004).
32. Orr, H. A. The population genetics of adaptation: the distribution of factors fixed during adaptive evolution. *Evolution* **52**, 935–949 (1998).
33. Barton, N. H. & Keightley, P. D. Understanding quantitative genetic variation. *Nature Rev. Genet.* **3**, 11–21 (2002).
34. Gibson, G. *et al.* Extensive sex-specific nonadditivity of gene expression in *Drosophila melanogaster*. *Genetics* **167**, 1791–1799 (2004).
35. Swanson-Wagner, R. A. *et al.* All possible modes of gene action are observed in a global comparison of gene expression in a maize F1 hybrid and its inbred parents. *Proc. Natl Acad. Sci. USA* **103**, 6805–6810 (2006).
36. Wayne, M. L., Pan, Y. J., Nuzhdin, S. V. & McIntyre, L. M. Additivity and *trans*-acting effects on gene expression in male *Drosophila simulans*. *Genetics* **168**, 1413–1420 (2004).
37. Vuylsteke, M., van Eeuwijk, F., Van Hummelen, P., Kuiper, M. & Zabeau, M. Genetic analysis of variation in gene expression in *Arabidopsis thaliana*. *Genetics* **171**, 1267–1275 (2005).
38. 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).
39. Storey, J. D., Akey, J. M. & Kruglyak, L. Multiple locus linkage analysis of genomewide expression in yeast. *PLoS Biol.* **3**, e267 (2005).
40. Carlborg, O. *et al.* Methodological aspects of the genetic dissection of gene expression. *Bioinformatics* **21**, 2383–2393 (2005).
41. Landry, C. R., Oh, J., Hartl, D. L. & Cavalieri, D. Genome-wide scan reveals that genetic variation for transcriptional plasticity in yeast is biased towards multi-copy and dispensable genes. *Gene* **366**, 343–351 (2006).
42. Chen, W. J. *et al.* Contribution of transcriptional regulation to natural variations in *Arabidopsis*. *Genome Biol.* **6**, R32 (2005).
43. Whitehead, A. & Crawford, D. L. Variation in tissue-specific gene expression among natural populations. *Genome Biol.* **6**, R13 (2005).
44. Yang, X. *et al.* Tissue-specific expression and regulation of sexually dimorphic genes in mice. *Genome Res.* **16**, 995–1004 (2006).
45. Cotsapas, C. J. *et al.* Genetic dissection of gene regulation in multiple mouse tissues. *Mamm. Genome* **17**, 490–495 (2006).
46. Benzer, S. Fine structure of a genetic region in *Bacteriophage*. *Proc. Natl Acad. Sci. USA* **41**, 344–354 (1955).
47. Lewis, E. B. Pseudoallelism and gene evolution. *Cold Spring Harb. Symp. Quant. Biol.* **16**, 159–174 (1951).
48. Pontecorvo, G. Genetic formulation of gene structure and gene action. *Adv. Enzymol. Relat. Subj. Biochem.* **13**, 121–149 (1952).
49. Ronald, J., Brem, R. B., Whittle, J. & Kruglyak, L. Local regulatory variation in *Saccharomyces cerevisiae*. *PLoS Genet.* **1**, e25 (2005).
50. 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).
References 49 and 50 use classic *cis*/*trans* tests to determine the contribution of *cis*-acting polymorphisms to local regulatory variation.
51. Lettice, L. A. *et al.* Disruption of a long-range *cis*-acting regulator for *Shh* causes preaxial polydactyly. *Proc. Natl Acad. Sci. USA* **99**, 7548–7553 (2002).
52. Loots, G. G. *et al.* Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. *Science* **288**, 136–140 (2000).
53. Nobrega, M. A., Ovcharenko, I., Afzal, V. & Rubin, E. M. Scanning human gene deserts for long-range enhancers. *Science* **302**, 413 (2003).
54. Spilianakis, C. G., Lalioti, M. D., Town, T., Lee, G. R. & Flavell, R. A. Interchromosomal associations between alternatively expressed loci. *Nature* **435**, 637–645 (2005).
55. Hahn, M. W., Stajich, J. E. & Wray, G. A. The effects of selection against spurious transcription factor binding sites. *Mol. Biol. Evol.* **20**, 901–906 (2003).
56. Schadt, E. E. Exploiting naturally occurring DNA variation and molecular profiling data to dissect disease and drug response traits. *Curr. Opin. Biotechnol.* **16**, 647–654 (2005).
57. Yan, H., Yuan, W., Velculescu, V. E., Vogelstein, B. & Kinzler, K. W. Allelic variation in human gene expression. *Science* **297**, 1143 (2002).
58. Cowles, C. R., Hirschhorn, J. N., Altshuler, D. & Lander, E. S. Detection of regulatory variation in mouse genes. *Nature Genet.* **32**, 432–437 (2002).
59. Pastinen, T. *et al.* A survey of genetic and epigenetic variation affecting human gene expression. *Physiol. Genomics* **16**, 184–193 (2004).
60. Knight, J. C., Keating, B. J. & Kwiatkowski, D. P. Allele-specific repression of lymphotoxin-α by activated B cell factor-1. *Nature Genet.* **36**, 394–349 (2004).
61. Damerval, C., Maurice, A., Josse, J. M. & de Vienne, D. Quantitative trait loci underlying gene product variation: a novel perspective for analyzing regulation of genome expression. *Genetics* **137**, 289–301 (1994).
62. Klose, J. *et al.* Genetic analysis of the mouse brain proteome. *Nature Genet.* **30**, 385–393 (2002).
References 61 and 62 describe classic and modern studies (respectively) of the genetics of protein abundance measured with two-dimensional gels.
63. Lo, H. S. *et al.* Allelic variation in gene expression is common in the human genome. *Genome Res.* **13**, 1855–1862 (2003).
64. Ronald, J. *et al.* Simultaneous genotyping, gene-expression measurement, and detection of allele-specific expression with oligonucleotide arrays. *Genome Res.* **15**, 284–291 (2005).
65. Pant, P. V. *et al.* Analysis of allelic differential expression in human white blood cells. *Genome Res.* **16**, 331–339 (2006).
66. Buckland, P. R. *et al.* A high proportion of polymorphisms in the promoters of brain expressed genes influences transcriptional activity. *Biochim. Biophys. Acta* **1690**, 238–249 (2004).
67. Hoogendoorn, B. *et al.* Functional analysis of human promoter polymorphisms. *Hum. Mol. Genet.* **12**, 2249–2254 (2003).
68. Buckland, P. R. *et al.* A high proportion of chromosome 21 promoter polymorphisms influence transcriptional activity. *Gene Expr.* **11**, 233–239 (2004).
69. Rockman, M. V. & Wray, G. A. Abundant raw material for *cis*-regulatory evolution in humans. *Mol. Biol. Evol.* **19**, 1991–2004 (2002).
70. Knight, J. C. Allele-specific gene expression uncovered. *Trends Genet.* **20**, 113–116 (2004).
71. Horan, M. *et al.* Human growth hormone 1 (*GH1*) gene expression: complex haplotype-dependent influence of polymorphic variation in the proximal promoter and locus control region. *Hum. Mutat.* **21**, 408–423 (2003).
72. Tao, H., Cox, D. R. & Frazer, K. A. Allele-specific *KRT1* expression is a complex trait. *PLoS Genet.* **2**, e93 (2006).
73. Perez-Enciso, M. *In silico* study of transcriptome genetic variation in outbred populations. *Genetics* **166**, 547–554 (2004).
74. Allison, D. B. *et al.* Multiple phenotype modeling in gene-mapping studies of quantitative traits: power advantages. *Am. J. Hum. Genet.* **63**, 1190–1201 (1998).
75. Jiang, C. & Zeng, Z. B. Multiple trait analysis of genetic mapping for quantitative trait loci. *Genetics* **140**, 1111–1127 (1995).
76. Bing, N. & Hoeschele, I. Genetical genomics analysis of a yeast segregant population for transcription network inference. *Genetics* **170**, 533–542 (2005).
77. Kulp, D. C. & Jagalur, M. Causal inference of regulator-target pairs by gene mapping of expression phenotypes. *BMC Genomics* **7**, 125 (2006).
78. Gunderson, K. L. *et al.* Whole-genome genotyping of haplotype tag single nucleotide polymorphisms. *Pharmacogenomics* **7**, 641–648 (2006).
79. Matsuzaki, H. *et al.* Parallel genotyping of over 10,000 SNPs using a one-primer assay on a high-density oligonucleotide array. *Genome Res.* **14**, 414–425 (2004).
80. Gresham, D. *et al.* Genome-wide detection of polymorphisms at nucleotide resolution with a single DNA microarray. *Science* **311**, 1932–1936 (2006).
81. Deutschbauer, A. M. & Davis, R. W. Quantitative trait loci mapped to single-nucleotide resolution in yeast. *Nature Genet.* **37**, 1333–1340 (2005).
82. Steinmetz, L. M. *et al.* Dissecting the architecture of a quantitative trait locus in yeast. *Nature* **416**, 326–330 (2002).
83. Gibson, G. & Weir, B. The quantitative genetics of transcription. *Trends Genet.* **21**, 616–623 (2005).
84. Hamblin, M. T., Thompson, E. E. & Di Rienzo, A. Complex signatures of natural selection at the *Duffy* blood group locus. *Am. J. Hum. Genet.* **70**, 369–383 (2002).
85. Bersaglieri, T. *et al.* Genetic signatures of strong recent positive selection at the lactase gene. *Am. J. Hum. Genet.* **74**, 1111–1120 (2004).
86. Crawford, D. L., Segal, J. A. & Barnett, J. L. Evolutionary analysis of TATA-less proximal promoter function. *Mol. Biol. Evol.* **16**, 194–207 (1999).
87. Hahn, M. W., Rockman, M. V., Soranzo, N., Goldstein, D. B. & Wray, G. A. Population genetic and phylogenetic evidence for positive selection on regulatory mutations at the factor VII locus in humans. *Genetics* **167**, 867–877 (2004).
88. Rockman, M. V., Hahn, M. W., Soranzo, N., Goldstein, D. B. & Wray, G. A. Positive selection on a human-specific transcription factor binding site regulating IL4 expression. *Curr. Biol.* **13**, 2118–2123 (2003).
89. Rockman, M. V. *et al.* Positive selection on MMP3 regulation has shaped heart disease risk. *Curr. Biol.* **14**, 1531–1539 (2004).
90. Rockman, M. V. *et al.* Ancient and recent positive selection transformed opioid *cis*-regulation in humans. *PLoS Biol.* **3**, e387 (2005).
91. Bamshad, M. J. *et al.* A strong signature of balancing selection in the 5′ *cis*-regulatory region of *CCR5*. *Proc. Natl Acad. Sci. USA* **99**, 10539–10544 (2002).
92. Emison, E. S. *et al.* A common sex-dependent mutation in a *RET* enhancer underlies Hirschsprung disease risk. *Nature* **434**, 857–863 (2005).
93. Takahashi, A., Tsaur, S. C., Coyne, J. A. & Wu, C. I. The nucleotide changes governing cuticular hydrocarbon variation and their evolution in *Drosophila melanogaster*. *Proc. Natl Acad. Sci. USA* **98**, 3920–3925 (2001).
94. Keurentjes, J. J. *et al.* The genetics of plant metabolism. *Nature Genet.* **38**, 842–849 (2006).
95. Perlstein, E. O. *et al.* Revealing complex traits with small molecules and naturally recombinant yeast strains. *Chem. Biol.* **13**, 319–327 (2006).
96. Churchill, G. A. & Doerge, R. W. Empirical threshold values for quantitative trait mapping. *Genetics* **138**, 963–971 (1994).
97. Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl Acad. Sci. USA* **100**, 9440–9445 (2003).
98. Perez-Enciso, M. Multiple association analysis via simulated annealing (MASSA). *Bioinformatics* **22**, 573–580 (2006).

99. Lan, H. *et al.* Dimension reduction for mapping mRNA abundance as quantitative traits. *Genetics* **164**, 1607–1614 (2003).
100. Schadt, E. E. *et al.* An integrative genomics approach to infer causal associations between gene expression and disease. *Nature Genet.* **37**, 710–717 (2005).
101. Li, H. Q. *et al.* Inferring gene transcriptional modulatory relations: a genetical genomics approach. *Hum. Mol. Genet.* **14**, 1119–1125 (2005).
102. Haldane, J. B. S. The time of action of genes, and its bearing on some evolutionary problems. *Am. Nat.* **66**, 5–24 (1932).
103. Jacob, F. & Monod, J. Genetic regulatory mechanisms in the synthesis of proteins. *J. Mol. Biol.* **3**, 318–356 (1961).
104. Schwartz, D. Genetic studies on mutant enzymes in maize. III. Contol of gene action in the synthesis of pH 7.5 esterase. *Genetics* **47**, 1609–1615 (1962).
105. Wallace, B. Genetic diversity, genetic uniformity, and heterosis. *Can. J. Genet. Cytol.* **5**, 239–253 (1963).
106. Zuckerkandl, E. in *Classification and Human Evolution* (ed. Washburn, S. L.) (Aldine, Chicago, 1963).
107. Abraham, I. & Doane, W. W. Genetic regulation of tissue-specific expression of amylase structural genes in *Drosophila melanogaster*. *Proc. Natl Acad. Sci. USA* **75**, 4446–4450 (1978).
108. Powell, J. R. Population genetics of *Drosophila* amylase. II. Geographic patterns in *D. pseudoobscura*. *Genetics* **92**, 613–622 (1979).
109. Matsuo, Y. & Yamazaki, T. Genetic analysis of natural populations of *Drosophila melanogaster* in Japan. IV. Natural selection on the inducibility, but not on the structural genes, of amylase loci. *Genetics* **108**, 879–896 (1984).
110. Powell, J. R. & Andjelkovic, M. Population genetics of *Drosophila amylase*. IV. Selection in laboratory populations maintained on different carbohydrates. *Genetics* **103**, 675–89 (1983).
111. Ayala, F. J. & McDonald, J. F. Continuous variation: possible role of regulatory genes. *Genetica* **52/53**, 1–15 (1980).
112. Allendorf, F. W., Knudsen, K. L. & Phelps, S. R. Identification of a gene regulating the tissue expression of a phosphoglucomutase locus in rainbow trout. *Genetics* **102**, 259–268 (1982).
113. Laurie-Ahlberg, C. C., Maroni, G., Bewley, G. C., Lucchesi, J. C. & Weir, B. S. Quantitative genetic variation of enzyme activities in natural populations of *Drosophila melanogaster*. *Proc. Natl Acad. Sci. USA* **77**, 1073–1077 (1980).
114. Stumpf, M. P., Wiuf, C. & May, R. M. Subnets of scale-free networks are not scale-free: sampling properties of networks. *Proc. Natl Acad. Sci. USA* **102**, 4221–4224 (2005).
115. Gomulkiewicz, R. & Kirkpatrick, M. Quantitative genetics and the evolution of reaction norms. *Evolution* **46**, 390–411 (1992).
116. Ohta, T. Origin of the neutral and nearly neutral theories of evolution. *J. Biosci.* **28**, 371–377 (2003).
117. Wittkopp, P. J., Haerum, B. K. & Clark, A. G. Evolutionary changes in *cis* and *trans* gene regulation. *Nature* **430**, 85–88 (2004).

Acknowledgements
We regret that space constraints prevented us from citing additional work in the field. We thank past and current members of the Kruglyak laboratory and our collaborators for discussions. Work was supported in part by a grant from the National Institute of Mental Health and a James S. McDonnell Foundation Centennial Fellowship to L.K. M.V.R. is supported by The Jane Coffin Childs Memorial Fund for Medical Research.

Competing interests statement
The authors declare no competing financial interests.

DATABASES
The following terms in this article are linked online to:
Entrez Gene:
http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene
AMN1 | *Runx1*

FURTHER INFORMATION
Gene Ontology: http://www.geneontology.org/
Leonid Kruglyak's homepage: http://www.molbio2.princeton.edu/index.php?option=content&task=view&id=217
Access to this links box is available online.