Published in final edited form as:
Nat Rev Genet. 2009 September ; 10(9): 595–604. doi:10.1038/nrg2630.

Genetics of human gene expression: mapping DNA variants that influence gene expression

Vivian G. Cheung[*,‡,§] and **Richard S. Spielman**[§,‖]

[*] Howard Hughes Medical Institute

[‡] Department of Pediatrics, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA

[§] Department of Genetics, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA

Abstract

There is extensive natural variation in human gene expression. As quantitative phenotypes, expression levels of genes are heritable. Genetic linkage and association mapping have identified *cis*- and *trans*-acting DNA variants that influence expression levels of human genes. New insights into human gene regulation are emerging from genetic analyses of gene expression in cells at rest and following exposure to stimuli. The integration of these genetic mapping results with data from co-expression networks is leading to a better understanding of how expression levels of individual genes are regulated and how genes interact with each other. These findings are important for basic understanding of gene regulation and of diseases that result from disruption of normal gene regulation.

Gene expression underlies cellular phenotypes; however, despite its importance, expression levels of many human genes differ among individuals. To understand how gene expression regulates key biological processes, early studies focused on identifying regulators, such as transcription factors, and their regulatory mechanisms. These studies improved our understanding of how gene expression is regulated in human cells and how its disruption can lead to developmental disorders and other human diseases. Although such studies shed light on regulatory mechanisms, they did not address normal variation in gene expression. In fact, for experimental studies of molecular mechanisms, highly variable observations are an unwanted complication. However, it has become clear that gene expression levels vary among individuals and can be analysed like other quantitative phenotypes such as height and serum glucose level[1–3]. The genetics of gene expression (referred to here as GOGE, pronounced ‘go-gee’) is the study of the genetic basis of variation in gene expression. GOGE studies (also known as expression QTL (eQTL)[4] studies or genetical genomics[5]) take advantage of this natural variation, enabling the study of gene expression. The results have already uncovered interesting and unexpected aspects of gene regulation[4,6–9].

Correspondence to V.G.C., vcheung@mail.med.upenn.edu.

‖This Review is dedicated to the memory of Richard Spielman, who passed away in April 2009 during the preparation of this article.

DATABASES

Entrez Gene: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene

ATP5J2 | *BCL11A* | *CPNE1* | *EPHX1* | *Lgr5* | *LTA4H* | *ORMDL3* | *PARK7* | *PDCD10* | *RB1* | *VDR*

FURTHER INFORMATION

The Cheung laboratory: http://genomics.med.upenn.edu/vcheung

CEPH: http://www.cephb.fr

International HapMap Project: http://www.hapmap.org

ALL LINKS ARE ACTIVE IN THE ONLINE PDF

Technical developments such as microarrays[10,11], which changed the scale of how gene expression can be measured, were important advances. They allowed measurement of the expression levels of thousands of genes in large numbers of individuals. Early microarray-based studies of gene expression provided a detailed map of expressed genes in various tissues and diseases, and the large volume of gene expression data revealed that the expression levels of many genes differ among individuals. With the ability to measure thousands of transcripts simultaneously, it was inevitable that some genetic studies began to shift from more traditional hypothesis-driven science to data-driven science. Identifying the extent of normal variation in human gene expression stimulated a fruitful merger of human genetics and genomics. GOGE studies have led to the identification of regulatory regions and DNA sequence variants that influence expression levels of genes in a range of organisms. For example, genome-wide GOGE studies have made it possible to evaluate the relative influence of *cis* and *trans* regulation on gene expression. In the last few years, several reviews of this field have been published[5,12–17]. Here, we focus specifically on GOGE studies in human cells. Because of the size and complexity of the human genome, and the fact that humans are not experimental organisms, the genetic analysis of human phenotypes and diseases carries a unique set of problems. The genetic analysis of gene expression as a human phenotype is no exception.

In this Review we discuss some of the early results from GOGE studies, the current challenges and the future developments. We start with an overview of how GOGE studies are carried out, and then we review the current understanding of the regulatory landscapes in cells under normal (baseline) conditions and of the variation between populations. We end by discussing new studies that use GOGE to understand genetic networks, and how studying cells after exposure to perturbation can reveal different perspectives on gene regulation.

Why study gene expression phenotypes?

The main goal of GOGE studies in humans is to identify the DNA variants (polymorphisms) that influence the expression levels of genes — that is, the gene expression phenotype. The significance of such findings is at least threefold. First, the studies connect variation at the DNA sequence level to that at the RNA level. There are over 3 million SNPs[18,19] and other sequence variants such as copy number polymorphism[20] in the human genome. Although most of these variants are presumably neutral, some are functional. However, identifying the functional variants has been challenging. GOGE studies narrow the field by pointing to regions and ultimately variants that regulate gene expression. Some of these regulatory variants have already been shown to be susceptibility alleles for human diseases such as asthma[21,22]. For further discussion on how the results of GOGE studies apply to the understanding of human diseases, see the recent review by Cookson and colleagues[23].

Second, in identifying variants that influence gene expression (or closely associated variants), GOGE studies scan the genome for regulators without the need for prior knowledge of the regulatory mechanisms. This allows GOGE studies to identify unknown regulators of gene expression. Third, unlike traditional molecular analyses, GOGE studies allow simultaneous investigation of many gene expression phenotypes. Thus, regulators for many phenotypes can be identified in parallel. The resulting regulator–target gene relationships facilitate the characterization of the gene expression regulatory landscape in human cells. This is a major advance from earlier gene expression profiling studies. In those earlier studies, one could identify genes that are activated or repressed in different cellular or disease states, and study the correlations among those genes. However, although gene correlations can imply co-regulation or a regulatory relationship, they do not indicate which genes are regulated and which are the regulators. GOGE mapping studies provide such information. When a gene expression phenotype maps to a particular region, the phenotype must be the target and the specified region must contain the regulator. Thus, by combining results from GOGE studies

with correlation analysis, one can improve gene co-expression networks from so-called undirected to directed graphs. This aspect of GOGE studies is described in more detail later.

How to carry out GOGE studies

Before we discuss results of GOGE studies, we describe how GOGE studies are done. We begin with the definition of expression levels of genes as phenotypes, and then discuss the human cells that have been used, and finally describe the genetic mapping approaches.

Phenotypic variation and heritability

It has only recently become clear that, within the same cell type and developmental stage, there is extensive individual variability in gene expression. FIGURE 1 illustrates the expression levels of 12 genes in 50 unrelated individuals measured in the same cell types and in the same microarray experiments: although the expression levels of two genes — *PARK7* and *ATP5J2* — show little variability among these individuals, other gene expression phenotypes showed extensive individual variation. This experiment was designed so that the non-genetic sources of variation that contribute to inter-individual differences were the same for all the genes[3,24]; the observed differences in variability among the genes are therefore best explained by underlying differences in the contribution from genetic variation, which is equivalent to the heritability of the phenotype. The variability among related individuals is less than that among unrelated individuals[3,24], thus indicating a genetic component to variation in human gene expression. More formal estimates of heritability in a variety of human cells — including lymphocytes and cells from immortalized cell lines, adipose tissue and brain tissue — have also shown a genetic contribution to variation in gene expression[24–27].

In the first years of GOGE studies in humans, it seemed that demonstrating heritability was a prerequisite to beginning genetic analyses, such as mapping by linkage and by association. When the degree of heritability is in doubt, it is still of interest to show that heritable variation contributes to gene expression variation. However, as with estimating heritability for other traits, various assumptions need to be made when calculating heritability for gene expression. Therefore, because in many cases DNA variants that influence expression levels of some genes have already been identified (that is, a heritable component of gene expression variation has been established), it is more practical to proceed directly to mapping, and find additional DNA polymorphisms that influence gene expression.

What cell types have been used for GOGE analyses?

Among the first questions in designing GOGE experiments is what type (or types) of cells to study. One of the challenges of studying human gene expression is availability of cells. However, as the central questions concern individual variation in gene expression, the studies require cells from a large number of individuals. In the late 1980s, Dausset and colleagues at the Centre for the Study of Human Polymorphisms (CEPH) in Paris, France, collected blood samples from large multigeneration families, and immortalized the B cells (to make lymphoblastoid cells) as a DNA source for genotyping, in order to construct genetic maps. Several GOGE studies have used cells from these CEPH pedigrees as an RNA source for studying gene expression[6,25]. As cell lines, they can be grown under uniform conditions, thus allowing one to minimize the environmental variables. However, a recent study suggests that other variables, such as titres of the Epstein–Barr virus used for immortalization of these cells, should be taken into account when designing experiments[28]. As these samples were used for the construction of several generations of genetic maps, many genotypes are available to verify that these cells have normal chromosomal content and show expected Mendelian inheritance of genetic markers. In addition to the immortalized B cells of the CEPH pedigrees, samples from other human populations collected by the International HapMap Project[18,19] and those

collected by Cookson and colleagues for an asthma study[21,22] have also been used for GOGE studies[7–9,29,30]. Results from GOGE studies of immortalized B cells are highly concordant, even though cells were grown independently in different laboratories and various platforms were used to measure gene expression[22,27,31].

The GOGE studies of immortalized B cells were followed by studies of other cell types. These studies analysed gene expression in cells from blood and subcutaneous adipose tissues from Icelandic populations[26], cells from tissues from brain banks[32], lymphocytes from a large-scale study of heart disease[27], and cells from liver samples from surgical resections and cadavers[33]. Although these samples were collected for gene expression analysis, many include health information and other biological data about the donors. The additional information will allow more extensive analyses, such as correlations of gene expression with clinical parameters.

Determining what cell types to use for GOGE studies depends on sample availability and the goals of the project. Primary cells from human subjects have the advantage that they have not been experimentally manipulated; however, it is difficult to control for the exposures (such as diet or medication) of the donors. These exposures (environmental factors) can have a significant influence on gene expression, and therefore can dampen the genetic influence on gene expression[34]. One of the most accessible human tissues is blood, but blood is not homogeneous and its composition differs between individuals. For example, some subjects have higher neutrophil counts and others have higher lymphocyte counts. If blood cells are used for studying variation in gene expression, it is important that these differential cell counts are taken into consideration. By contrast, cultured cells such as immortalized B cells are less natural, but they are from the same cell type — B lymphocytes — and can be grown under controlled conditions to minimize the environmental influence on gene expression. Although selection of the appropriate cell type is important in experimental design, it is reassuring that the regulatory variants found in immortalized B cells regulate the same target genes in other cell types (discussed further in a later section).

Given the difficulty of collecting human samples, one may wonder why model organisms are not studied instead. Studies in model organisms have provided valuable general insights into the genetic basis of variation in gene expression, but studying human cells is necessary as some components of gene regulation in humans are not captured by model organisms. In addition, humans are heterozygous at many loci and it is difficult to reconstruct heterozygosity at a large number of loci in inbred experimental organisms[17]. Thus, even though it is difficult to collect human samples, future studies of gene expression will need to continue to identify ways to analyse human tissues.

Genetic mapping to locate determinants of gene expression phenotypes

As expression phenotypes are intermediate phenotypes that are related to DNA sequence variants, they are more amenable to genetic studies than other human quantitative phenotypes, such as height and weight. This has been demonstrated by the successful identification of regulatory regions that influence gene expression phenotypes in multiple human tissues in genetic linkage analysis[6,26,27] and association studies[7–9,32,33] (BOX 1). However, it is challenging to identify the precise causal sequence variants. In experimental organisms and plants, studies have identified QTLs and, in some cases, even the causal nucleotide[35,36]. Although technological and methodological advances have improved QTL mapping in humans, mapping of quantitative human traits remains difficult[37].

Box 1

Methods in genetics of gene expression studies

Genetic linkage and association

Two loci (for example, a marker and a trait) that do not segregate independently of each other at meiosis are linked, implying that they are located near each other on the same chromosome. In linkage analysis, a large sample of families, ideally with large siblingships, is genotyped for a few thousand markers (SNPs) of known location throughout the genome. Each marker is tested for linkage with the phenotype of interest. The evidence for linkage is provided as a LOD score (base 10 logarithm of the odds, or 'log-odds') or as the corresponding *p* value. The results of this genome-wide linkage scan are usually presented graphically (FIG. 3a).

The underlying principles for association testing are different. The analysis is based on a large sample of unrelated individuals. These may be patients and unaffected individuals, as in a case–control study, or simply unrelated individuals who vary for a quantitative trait, such as a gene expression phenotype (for example, the expression level of copine I, *CPNE1*, in FIG. 3). For variation in a gene expression phenotype, association studies determine if the level of gene expression differs depending on SNP genotype. If it does differ then there is association between the gene expression level and the alleles (or genotypes) of that SNP. In a genome-wide association study this is done for a large number of SNPs (500,000 to 1 million) with locations spread through the genome. For each SNP location, the level of significance is estimated, and the results are presented graphically (FIG. 3b). In a genetics of gene expression (GOGE) study, this plot specifically shows candidate locations for determinants of variation in gene expression.

Transmission/disequilibrium test

The classical linkage test does not involve allelic association, and the association test does not make use of segregation in families. Is it possible to capture the strengths of both in one test? The transmission/disequilibrium test (TDT)[72] does exactly that by counting the number of transmissions of a specific marker allele from heterozygous parents to affected offspring. The TDT was originally designed for qualitative traits, but several methods and computer programs are available for extending the TDT to quantitative traits[73–75]. One of these, the quantitative TDT (QTDT)[73,74], has been used for GOGE studies.

Genome-wide analysis and the issue of multiple testing

Most classical statistical test procedures were developed to test one statistical hypothesis at a time. However, in all the approaches described above, genome-wide analysis is the goal and thousands of hypotheses may be tested — for instance, for many genes (for example, for gene expression levels) or for genetic markers. The investigator then gives most attention to the most significant test. As more tests are carried out, the chance increases of finding one or more statistical false positives that are significant by chance. To limit this effect, several statistical procedures have been developed. The two most often used are the Bonferroni procedure and the false discovery rate method. As these are solutions to technical statistical problems, we do not describe them here, but summaries can be found in a recent review by Rao and colleagues[76].

Regulation of baseline gene expression

It is well established that gene expression levels are controlled by a combination of *cis*- and *trans*-acting regulators: for example, the binding of *trans*-acting factors such as transcription factors to *cis*-acting regulatory target sequences. GOGE studies do not identify all the *cis*- and

trans-acting regulators but aim to find polymorphic variants that contribute to individual variation in human gene expression (FIG. 2). If the variants reside on a chromosome different to that of the target gene, the regulation has to be in *trans*. Variants that are close to the target genes (within a few kilobases of the target gene, for example) are usually considered to be *cis* determinants. Defining these determinants as *cis* only refers to the fact that they are close to the target genes; there can be polymorphism either in *cis*-regulatory sites or in *trans*-acting regulators that are close to the target genes. Unless the functional variants are identified, the *cis* or *trans* designation only implies the distance of the genetic signal relative to the target gene, it has no functional significance. For this reason, Kruglyak and colleagues have cautioned against using terms that imply functions, such as *cis* and *trans*; instead they suggest using 'local' and 'distant'.[13]

An illustration of how mapping results can identify a regulatory polymorphism is shown in FIG. 3. In this case, both linkage and association analyses identified a region close to the target gene, *CPNE1* (copine I), as the candidate regulatory region. SNPs in the gene showed differential allelic expression; individuals with the TT genotype for SNP rs3787165 have higher expression levels of *CPNE1* than those with the CC genotype.

Contribution of *cis*-acting variants

One expects to learn something about the relative contribution of *cis*- and *trans*-determinants of variation in human gene expression from GOGE studies. Unfortunately, interpretation of the data is not straightforward, partly because *cis*- and *trans*-acting determinants influence gene expression in different ways. To date, some GOGE studies[26,27,33] have found more determinants that map in *cis* than in *trans*, whereas others[6,32] found more *trans*-acting determinants. The differences in findings are probably due to differences in sample sizes and thresholds for statistical significance. When there is a polymorphic *cis*-acting variant, its effect on the expression level of the target gene is often large; therefore, they are easier to detect than *trans*acting variants. As it is difficult to obtain human tissues for gene expression studies, most studies have relatively small sample sizes and, therefore, have identified mostly *cis* determinants of gene expression.

Another approach to assess the proportion of *cis*-acting determinants that influence gene expression is to measure the relative expression of allelic forms of genes by differential allelic expression (DAE) studies[38–42]. In these analyses, one measures the relative expression levels of each allele at a heterozygous site in a transcribed (usually exonic) region of a gene[38–42]. As the two alleles are expected to be exposed to the same *trans*-acting factors, DAE studies allow a relatively direct assessment of the contributions of *cis*-acting determinants. Results of these DAE studies for expression phenotypes show that ~30–50% of the genes show differential allelic expression.

Price *et al.*[43] have estimated the proportions of *cis*- and *trans*-acting determinants by a different method that uses expression data from the admixed African American population. The key feature of the analysis is that the effect of allelic variation is estimated directly from the relationship between gene expression levels and marker allele frequencies in the admixed population, not from separate tests of each expressed gene. The resulting estimates for the contribution to variation in gene expression from *cis*- and *trans*-acting regulation are 0.05 and 0.38, respectively. The fraction that is due to *cis* effects is therefore calculated as 0.12 (0.05/0.43; with a standard error of 0.3%). Unlike almost all previous estimates, this method does not depend on choice of a threshold for *p* values.

Based on data from these various approaches, we estimate that ~20% of expression phenotypes at baseline (that is, in cells under normal, unstimulated growth conditions) are regulated by *cis* variants. Studies with larger sample sizes and other technologies such as RNA-Seq[44] that

provide alternative methods for measuring gene expression will allow more accurate estimates of the contribution of *cis*-acting determinants (see concluding remarks).

Mechanisms of polymorphic *cis* regulation

Cis variants can influence the expression levels of target genes in different ways, such as by affecting the transcription level or stability of the message. Generally, the mechanisms by which polymorphic *cis* variants influence gene expression are still being examined. A key challenge is that, although genetic mapping can be carried out on many phenotypes in parallel, methods to identify the molecular mechanisms of regulation are not amenable to such high-throughput analyses. So far, the mechanisms of how polymorphic variants affect gene expression have been worked out for only a small number of genes.

Some insights are offered by fine association mapping (BOX 1), which can identify more precisely where the regulatory variants are relative to the target genes. For example, in our analysis of 133 gene expression phenotypes, association mapping results showed that the regulatory sites are found in approximately the same proportion at the 5′ (27%) and 3′ (34%) ends of the genes, and within the target genes (25%)[7]. For 14% of the phenotypes, linkage disequilibrium was so strong that we were not able to narrow the region of *cis* association. The variants in the 5′ ends of genes may affect RNA polymerase II and transcription factor binding[7,45,46], those in the 3′ ends may affect stability of the transcripts[47,48], and variants in genes can also affect binding of transcription factors[27].

Trans-acting variants

Trans-acting variants are more difficult to identify because, unlike *cis* variants, they can be anywhere in the genome relative to the target gene, and genetic mapping results suggest that their effects on gene expression are smaller than the effects of *cis*-acting variants. This is probably because genes are usually influenced by several *trans*-acting regulators and, therefore, the effect of each *trans*-acting regulator on expression of its target gene is small, whereas there is usually one or only a few *cis*-acting regulators. However, to understand gene regulation, it is crucial to identify *trans*-acting regulators.

Although *trans*-acting regulatory regions have been identified through linkage analysis[6,26,27] and association studies[32], only a few *trans*-acting determinants of baseline gene expression have been identified. In linkage analysis, the candidate regulatory regions are often megabases in size and include several candidate regulators. FIGURE 4 illustrates how *trans*-acting regulatory regions can be found by linkage analysis: for the expression level of *PDCD10* (programmed cell death 10, located on chromosome 3), two significant linkage peaks were found — one on chromosome 4 and another on chromosome 19. The peaks on both chromosomes are several megabases in size. These regions contain the polymorphic *trans*-acting regulators that influence expression of *PDCD10*; fine mapping of the regions is needed to identify the regulatory variants. Despite the challenge of identifying *trans*-acting regulators, some examples of polymorphic *trans*-acting regulators of gene expression are beginning to emerge. Examples of genes in which regulatory variants exert a *trans*-acting effect include the epoxide hydrolase 1 gene (*EPHX1*), which regulates expression of *ORMDL3* (REF. [31]), and *BCL11A* (encoding a zinc finger protein), which influences γ-globin gene expression[49]. *EPHX1* was identified in a genome-wide association analysis of gene expression, and regulatory variants in *BCL11A* were identified in a search for regulators that influence individual variation in fetal haemoglobin level.

Even though only a few *trans*-acting regulators of gene expression have been identified, and many *trans*-acting regulatory regions are large, analyses of these regions in the human genome are leading to a better understanding of gene regulation. These analyses suggest that *trans*-

acting regulators are not enriched for known regulators of gene expression such as transcription factors or signalling molecules; instead, the polymorphic *trans*-acting regulators belong to diverse groups of genes, from cell surface receptor genes to structural genes. Similar findings were reported by Kruglyak and colleagues in their analysis of gene expression variation in yeast[50]. Despite the relative lack of progress in identifying *trans*-acting regulators of baseline gene expression, we discuss in a later section how polymorphic *trans*-acting regulators have been identified in studies of cells exposed to external stimuli.

Regulatory landscapes among different cell types

Unlike studies in model organisms such as yeast and *Caenorhabditis elegans*, studies of human gene expression cannot be carried out on whole organisms; instead, they are mostly restricted to specific cell types. As mentioned above, GOGE studies in humans have been carried out in various cell types, including lymphocytes, immortalized B cells, brain cells and liver cells. Even though some gene expression patterns are cell type specific, a large fraction of GOGE findings seem to be shared across different types of cells. For example, a comparison of results from a study of immortalized B cells with those from primary lymphocytes showed that seven of eight *cis*-linked phenotypes were shared among the cells[27]. Of course, B cells are a subset of lymphocytes so the shared regulation is not surprising. However, even between different cell types, such as adipose tissue and blood, ~30–50% of the *cis*-regulated phenotypes are shared[26,33]. Too few *trans*-acting regulatory variants have been identified to date for similar comparisons.

Population differences in gene expression

Several studies have shown that the average expression levels of many genes differ among populations[29–31,51]. The studies were carried out using samples from the International HapMap Project[52]. In our study of 60 CEU individuals (northern and western European ancestry) and 82 Asians (42 Han Chinese of Beijing, CHB, and 42 Japanese of Tokyo, JPT), 1,097 of 3,197 genes differ significantly ($p < 10^{-5}$) between the two groups[51]. With the same threshold, only 27 genes differ significantly between the CHB and JPT samples. Similar findings were reported by Dolan and colleagues[29].

We[51] and others[29,31] have investigated whether differences in these average expression phenotype levels are related to specific allele frequency differences. For ~12 of the phenotypes so far studied in detail, the population differences in gene expression are mostly accounted for by differences in allele frequencies of regulators that are *cis* linked to the gene[51]. This situation is revealed by SNPs that show strong linkage disequilibrium (association) with the expression level. Thus, the population differences in these cases are not due to regulatory mechanisms that are fundamentally different between the populations, but to different genotype frequencies for the same regulatory alleles. Further studies are needed to determine what proportion of population differences in expression level will be accounted for by allele and genotype frequency differences of this kind.

These studies of population differences in gene expression have recently been extended to examine the genetic basis of population differences in response to therapeutics. Dolan and colleagues studied the response of cells from CEU and Yoruba in Ibadan, Nigeria (YRI) individuals to cytarabine arabinoside (a chemotherapeutic agent) in order to understand the population differences in outcomes and toxicities among patients with acute myeloid leukemia. They found that different SNPs account for variability in sensitivity to cytarabine arabinoside in the two populations. Some of the differences can be also accounted for by differences in allele frequencies of the associated SNPs in the two populations[53].

More complex gene interactions and regulation

We have so far focused on the identification of genetic variants that influence expression of individual genes. Of course in cells the regulation is much more complex. Most *trans*-acting regulators influence multiple target genes, and genes interact with each other to carry out various functions. The same normal variation in gene expression that allows GOGE studies to be performed lends itself to the study of gene interactions.

Hot spots

Hot spots in GOGE studies are regions that contain DNA variants that influence the expression of multiple genes. They have also been termed master regulatory regions. As Rockman and Kruglyak point out[13], these variants can influence gene expression indirectly by affecting cellular function (in the extreme, cell death). Thus, it is more appropriate to call them hot spots rather than master regulatory regions.

Studies in yeast and other organisms have identified hot spots that contain genetic variants that influence multiple expression phenotypes[1,50,54–56]. Human studies have yielded mixed results; some studies report hot spots[6,33] and others do not[25,27]. As the genetic variants in hot spots act in *trans*, it is likely that the differences among studies are partly because of differences in power to detect *trans*-acting variants. Based on results from studies that did identify hot spots in the human genome, we can make some general remarks on how hot spots might influence human gene expression. The target genes with phenotypes that map to the same hot spots often share similar functions or reside close to each other[6]. As genes that share functions are often co-regulated, their polymorphic regulators would appear in GOGE studies as hot spots. The expression levels of co-regulated genes frequently show significant correlations. Although this correlation is often biologically important, it can also lead to an overestimation of the number of phenotypes mapping to a hot spot[57]. Besides shared function, some target genes of a hot spot are close to each other on a chromosome. This is perhaps not unexpected as it is not unusual to find members of a gene family that cluster in a chromosomal region, and these members are often co-regulated. In addition, nearby genes can share common enhancers; therefore, variants in those enhancers or in polymorphic transcription factors that bind to those enhancers can affect expression of several genes. Chromatin modulators can also affect expression of nearby genes by influencing the chromatin structure of a region.

Variation in gene expression and gene networks

Variation in gene expression not only allows genetic dissection of gene expression phenotypes but also facilitates studies of how genes interact with each other in networks[58]. Correlation analysis of gene expression underlies many co-expression network studies[59–61]: based on these correlations in gene expression, connections (so-called edges) can be drawn among genes. The resulting diagram of connectivity allows one to examine whole groups of correlated genes rather than focusing on only pairwise relationships. It also provides information on how each gene is connected to others in the network, and identifies genes that are more connected than others.

As gene expression underlies cellular phenotypes, studies of gene networks can facilitate the understanding of complex phenotypes. Recent studies that take advantage of natural variation in gene expression in *Drosophila melanogaster* found co-expressed modules that are associated with complex organismal phenotypes, such as duration of sleep[62,63]. These results suggest that the DNA variants that influence gene expression can also affect more complex phenotypes.

Gene correlation alone can only provide suggestions on biological relatedness. However, when information from GOGE studies is superimposed on these networks, it identifies the regulators

and targets in the network and therefore provides information on causal rather than just correlative relationships[64–66]. The integration of network and GOGE studies has been used to identify genes that affect complex phenotypes. Earlier studies identified genes in metabolic pathways that contribute to obesity[26,65]. Those results were recently validated by knockout studies in mice[67]. In addition, by combining results from network analysis and genetic mapping in mice, Balmain and colleagues[68] recently identified DNA variants in the G protein-coupled receptor gene *Lgr5* as determinants of the expression levels of 62 highly correlated genes in hair follicle cells. In addition, they found that DNA polymorphisms in the vitamin D receptor gene (*VDR*) influence expression levels of a network of genes that play a part in the inflammatory response.

GOGE in cells after perturbation

GOGE studies are not limited to the study of cells at baseline; they also allow study of 'stimulated' cells that have been exposed to various perturbations. Early examples of these studies include human cells that have been exposed to drugs[53], endoplasmic reticulum stress and toxins such as ionizing radiation[69]. These studies provide a platform for studying individual variation in response to various stresses. For example, individuals differ in response to many toxins and yet the genetics of sensitivity to toxins is poorly understood. As we cannot expose humans to stress or toxins for experimental purposes, there is a lack of well-defined sensitivity phenotypes from related individuals for genetic studies. GOGE studies of stimulated human cells provide some solutions. Cells from many individuals, including related individuals, can be exposed to stresses in a controlled environment and their responses (both gene expression and cellular phenotype) analysed. This allows genetic analysis of individual variation in response to the perturbation. Studies of stimulated cells can, in addition to improving our understanding of the response to specific stimuli, expand our knowledge of the general mechanisms by which expression levels of genes are regulated. By perturbing cells, we expect to uncover regulatory pathways that are difficult to examine in unstimulated, baseline cells. This type of analysis might provide insight into disease susceptibility pathways.

We will use results from our recent study of irradiated cells[69] to illustrate some early lessons from perturbation studies. We exposed cells from individuals in large families to ionizing radiation and measured gene expression and cellular phenotypes, such as cell death, in the irradiated cells. We then carried out genetic studies to map the DNA variants that influence responses to radiation exposure. We found significant linkage for expression levels of over 1,200 radiation-responsive genes. These results revealed a regulatory landscape that differs from that of cells at baseline. Using similar numbers of families for genetic mapping we found that, although 20% or more of genes at baseline are regulated in *cis*[6], following radiation exposure less than 1% of genes are *cis* regulated. In irradiated cells, >99% of the polymorphic regulators act in *trans* to the target genes. In *C. elegans*[70] and yeast[71], *trans*-acting regulators are also found to play a key part in regulating the gene expression response to stress. Unlike *cis* regulation, *trans*-acting regulators can affect the expression of many genes, thus allowing a coordinated gene expression response. In addition, most genes probably have several *trans*-acting regulators. This provides cells with different ways to regulate gene expression in order to deal with various stimuli.

In addition to revealing a different regulatory landscape, results from analyses of irradiated cells also allowed us to uncover genes that were not known to have a role in the response to radiation exposure. The polymorphic *trans*-acting regulators that we identified include transcription factors such as retinoblastoma 1 (*RB1*) and *VDR*, which were already known to play a part in regulating gene expression. However, we also identified genes such as leukotriene A4 hydrolase (*LTA4H*) that were not known to regulate gene expression. These results will facilitate the identification of unknown pathways involved in radiation response. As the

functions of many human genes remain unknown it seems likely that GOGE studies might help to identify those that regulate gene expression.

A third finding from studying responses of irradiated cells is the identification of DNA variants that influence individual variation in the gene expression and cellular responses to radiation. From the baseline studies, we expected that we would find DNA polymorphisms that influence expression levels of genes. Surprisingly, with our sample size of only 15 families, we found significant linkage for more than 1,200 (30%) of the radiation-induced expression phenotypes. For a subset of these candidate regions, we were able to identify polymorphic regulators by association mapping. As most individuals are not exposed to a significant amount of ionizing radiation, those regulators that influence radiation response are not under selective pressure and their frequencies remain high, unlike disease susceptibility variants. This may account for why it is relatively easy for us to identify these polymorphic regulators.

These are results from early studies of stimulated cells, but it is promising to see that gene expression responses to perturbation are easily mapped and that the polymorphisms that influence these gene expression responses also affect cellular phenotypes. We expect that additional studies will allow the development of genetic predictors of cellular response to toxins. A better understanding of how human cells deal with toxin exposure or other cellular stresses will facilitate the development of drugs that influence the sensitivity of cells to toxins.

Concluding remarks

As with many human phenotypes, expression levels of genes are highly variable and are genetically regulated. Genetic studies of gene expression as a phenotype have identified regulators that influence the expression levels of individual genes. Most of the regulatory variants that have been identified are close to the target (regulated) gene. Next, we need to identify the variants that act in *trans* to influence gene expression, and to understand the molecular mechanisms of how *cis*- and *trans*-acting regulatory variants influence the expression levels of genes. The mapping of *trans*-acting regulatory variants can be achieved by increasing the sample sizes used in GOGE studies, by obtaining more accurate phenotypes and by identifying regulators in candidate regions. Initial GOGE studies were carried out as proof-of-principle studies; therefore, the sample sizes were modest. Future studies with larger sample sizes and different human cell types will result in a more detailed map of the regulatory variants that influence human gene expression. The availability of high-throughput sequencing will enable more accurate determination of gene expression through RNA-Seq studies[44] and will also identify genes that reside in candidate regulatory regions. To date, most GOGE studies used gene expression from quantitative reverse transcription PCR or microarrays, these hybridization-based methods are invariably affected by noise from the nonspecific binding of RNA to the probes. The digital nature of RNA-Seq should provide more accurate gene expression phenotypes and allele-specific gene expression. However, better ways to map the short-read sequences need to be developed in order to achieve the most accurate measurement of gene expression. In addition, the cost of RNA-Seq also needs to be reduced in order to enable studies with large sample sizes and the accurate measurement of transcripts that are expressed at low levels. The ability to identify genes expressed at low levels is important for GOGE studies as most known regulators such as transcription factors are expressed at low levels. Therefore, a detailed catalogue of expressed genes that can potentially act as gene expression regulators will facilitate GOGE studies.

In addition to identifying regulators of individual genes, we need to expand the scope of these studies to understand the broader regulatory network. The strength of GOGE studies is their ability to survey the genome for regulatory variants. The identification of *trans*-acting variants is likely to uncover novel regulatory mechanisms and will allow us to assign new roles to

known genes. By expanding the study to understand regulatory relationships as networks we will learn how genes interact with each other, and why changes in expression of some genes have little biological consequence but changes in other genes cause major disruptions of cellular processes.

Acknowledgments

We thank the members of our laboratories for comments and discussions. I (V.G.C.) thank C. Gunter for support and encouragement in finishing this Review. This work is supported by the National Institutes of Health and the Howard Hughes Medical Institute.

Glossary

Gene expression phenotype	The expression level of a gene in an individual as determined by his or her genotype and the cellular environments in which the gene is expressed
Co-expression network	Groups of interconnected genes that are linked by the correlations in their expression levels
Heritability	The proportion of total phenotypic variation that is due to genetic variation
Regulatory polymorphism	DNA sequence variants that regulate cellular processes such as gene expression
Differential allelic expression	Polymorphic forms (different sequences) of a gene have different expression levels
Admixed	An admixed population contains offspring of individuals originating from genetically divergent parental populations
RNA-Seq	Sequence analysis of RNA (for example, after conversion into cDNA):the results can be used for various analyses, including study of gene expression, identification of coding SNPs and determination of allele-specific gene expression

References

1. Brem RB, Yvert G, Clinton R, Kruglyak L. Genetic dissection of transcriptional regulation in budding yeast. Science 2002;296:752–755. [PubMed: 11923494]
2. Cheung VG, Spielman RS. The genetics of variation in gene expression. Nature Genet 2002;32:522–525. [PubMed: 12454648]
3. Cheung VG, et al. Natural variation in human gene expression assessed in lymphoblastoid cells. Nature Genet 2003;33:422–425. [PubMed: 12567189]
4. Schadt EE, et al. Genetics of gene expression surveyed in maize, mouse and man. Nature 2003;422:297–302. [PubMed: 12646919]
5. Jansen RC, Nap JP. Genetical genomics: the added value from segregation. Trends Genet 2001;17:388–391. [PubMed: 11418218]
6. Morley M, et al. Genetic analysis of genome-wide variation in human gene expression. Nature 2004;430:743–747. [PubMed: 15269782]
7. Cheung VG, et al. Mapping determinants of human gene expression by regional and genome-wide association. Nature 2005;437:1365–1369. [PubMed: 16251966]
8. Stranger BE, et al. Genome-wide associations of gene expression variation in humans. PLoS Genet 2005;1:e78. [PubMed: 16362079]
9. Stranger BE, et al. Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science 2007;315:848–853. [PubMed: 17289997]

10. DeRisi J, et al. Use of a cDNA microarray to analyse gene expression patterns in human cancer. Nature Genet 1996;14:457–460. [PubMed: 8944026]
11. Fodor SP, et al. Multiplexed biochemical assays with biological chips. Nature 1993;364:555–556. [PubMed: 7687751]
12. Farrall M. Quantitative genetic variation: a post-modern view. Hum Mol Genet 2004;13:R1–R7. [PubMed: 14962979]
13. Rockman MV, Kruglyak L. Genetics of global gene expression. Nature Rev Genet 2006;7:862–872. [PubMed: 17047685]
14. Li J, Burmeister M. Genetical genomics: combining genetics with gene expression analysis. Hum Mol Genet 2005;14:R163–R169. [PubMed: 16244315]
15. Nica AC, Dermitzakis ET. Using gene expression to investigate the genetic basis of complex disorders. Hum Mol Genet 2008;17:R129–R134. [PubMed: 18852201]
16. Stranger BE, Dermitzakis ET. The genetics of regulatory variation in the human genome. Hum Genomics 2005;2:126–131. [PubMed: 16004727]
17. Gilad Y, Rifkin SA, Pritchard JK. Revealing the architecture of gene regulation: the promise of eQTL studies. Trends Genet 2008;24:408–415. [PubMed: 18597885]
18. International HapMap Consortium. The International HapMap Project. Nature 2003;426:789–796. [PubMed: 14685227]
19. International HapMap Consortium. A haplotype map of the human genome. Nature 2005;437:1299–1320. [PubMed: 16255080]
20. Redon R, et al. Global variation in copy number in the human genome. Nature 2006;444:444–454. [PubMed: 17122850]
21. Moffatt MF, et al. Genetic variants regulating *ORMDL3* expression contribute to the risk of childhood asthma. Nature 2007;448:470–473. Demonstrates that regulatory variants of expression of *ORMDL3* influence an individual's susceptibility to asthma. [PubMed: 17611496]
22. Dixon AL, et al. A genome-wide association study of global gene expression. Nature Genet 2007;39:1202–1207. [PubMed: 17873877]
23. Cookson W, Liang L, Abecasis G, Moffatt M, Lathrop M. Mapping complex disease traits with global gene expression. Nature Rev Genet 2009;10:184–194. [PubMed: 19223927]
24. Cheung VG, et al. Genetics of quantitative variation in human gene expression. Cold Spring Harbor Symp Quant Biol 2003;68:403–407. [PubMed: 15338642]
25. Monks SA, et al. Genetic inheritance of gene expression in human cell lines. Am J Hum Genet 2004;75:1094–1105. [PubMed: 15514893]
26. Emilsson V, et al. Genetics of gene expression and its effect on disease. Nature 2008;452:423–428. [PubMed: 18344981]
27. Goring HH, et al. Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes. Nature Genet 2007;39:1208–1216. [PubMed: 17873875]
28. Choy E, et al. Genetic analysis of human traits *in vitro*: drug response and gene expression in lymphoblastoid cell lines. PLoS Genet 2008;4:e1000287. [PubMed: 19043577]
29. Zhang W, et al. Evaluation of genetic variation contributing to differences in gene expression between populations. Am J Hum Genet 2008;82:631–640. [PubMed: 18313023]
30. Storey JD, et al. Gene-expression variation within and among human populations. Am J Hum Genet 2007;80:502–509. [PubMed: 17273971]
31. Duan S, et al. Genetic architecture of transcript-level variation in humans. Am J Hum Genet 2008;82:1101–1113. [PubMed: 18439551]
32. Myers AJ, et al. A survey of genetic human cortical gene expression. Nature Genet 2007;39:1494–1499. [PubMed: 17982457]
33. Schadt EE, et al. Mapping the genetic architecture of gene expression in human liver. PLoS Biol 2008;6:e107. [PubMed: 18462017]
34. Whitney AR, et al. Individuality and variation in gene expression patterns in human blood. Proc Natl Acad Sci USA 2003;100:1896–1901. [PubMed: 12578971]
35. Krattinger SG, et al. A putative ABC transporter confers durable resistance to multiple fungal pathogens in wheat. Science 2009;323:1360–1363. [PubMed: 19229000]

36. Grisart B, et al. Genetic and functional confirmation of the causality of the *DGAT1 K232A* quantitative trait nucleotide in affecting milk yield and composition. Proc Natl Acad Sci USA 2004;101:2398–2403. [PubMed: 14983021]
37. Ioannidis JP, Thomas G, Daly MJ. Validating, augmenting and refining genome-wide association signals. Nature Rev Genet 2009;10:318–329. [PubMed: 19373277]
38. Cheung VG, et al. Monozygotic twins reveal germline contribution to allelic expression differences. Am J Hum Genet 2008;82:1357–1360. [PubMed: 18513681]
39. Pant PV, et al. Analysis of allelic differential expression in human white blood cells. Genome Res 2006;16:331–339. A thorough study of differential allelic expression of human genes on a genome-wide scale. [PubMed: 16467561]
40. Pastinen T, et al. A survey of genetic and epigenetic variation affecting human gene expression. Physiol Genomics 2004;16:184–193. [PubMed: 14583597]
41. Pastinen T, Ge B, Hudson TJ. Influence of human genome polymorphism on gene expression. Hum Mol Genet 2006;15:R9–R16. [PubMed: 16651375]
42. Lo HS, et al. Allelic variation in gene expression is common in the human genome. Genome Res 2003;13:1855–1862. [PubMed: 12902379]
43. Price AL, et al. Effects of *cis* and *trans* genetic ancestry on gene expression in African Americans. PLoS Genet 2008;4:e1000294. [PubMed: 19057673]
44. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature Methods 2008;5:621–628. [PubMed: 18516045]
45. Knight JC, Keating BJ, Rockett KA, Kwiatkowski DP. *In vivo* characterization of regulatory polymorphisms by allele-specific quantification of RNA polymerase loading. Nature Genet 2003;33:469–475. Description of a molecular method that assesses whether *cis*-regulatory variants influence gene expression by differential allelic binding of RNA polymerase to promoter complexes. [PubMed: 12627232]
46. Liu X, et al. Expression-based discovery of variation in the human glutathione S-transferase M3 promoter and functional analysis in a glioma cell line using allele-specific chromatin immunoprecipitation. Cancer Res 2005;65:99–104. [PubMed: 15665284]
47. Fritsche LG, et al. Age-related macular degeneration is associated with an unstable *ARMS2* (*LOC387715*) mRNA. Nature Genet 2008;40:892–896. [PubMed: 18511946]
48. Mio F, et al. A functional polymorphism in *COL11A1*, which encodes the α1 chain of type XI collagen, is associated with susceptibility to lumbar disc herniation. Am J Hum Genet 2007;81:1271–1277. [PubMed: 17999364]
49. Sankaran VG, et al. Human fetal hemoglobin expression is regulated by the developmental stage-specific repressor *BCL11A*. Science 2008;322:1839–42. Illustrates that *BCL11A* is a *trans*-acting regulator of fetal haemoglobin expression. [PubMed: 19056937]
50. Yvert G, et al. *Trans*-acting regulatory variation in *Saccharomyces cerevisiae* and the role of transcription factors. Nature Genet 2003;3:57–64. [PubMed: 12897782]
51. Spielman RS, et al. Common genetic variants account for differences in gene expression among ethnic groups. Nature Genet 2007;39:226–230. [PubMed: 17206142]
52. The International Hapmap Consortium. A haplotype map of the human genome. Nature 2005;437:1299–1320. [PubMed: 16255080]
53. Hartford CM, et al. Population-specific genetic variants important in susceptibility to cytarabine arabinoside cytotoxicity. Blood 2009;113:2145–2153. [PubMed: 19109566]
54. Chesler EJ, et al. Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function. Nature Genet 2005;37:233–242. [PubMed: 15711545]
55. DeCook R, Lall S, Nettleton D, Howell SH. Genetic regulation of gene expression during shoot development in Arabidopsis. Genetics 2006;172:1155–1164. [PubMed: 15956669]
56. Hubner N, et al. Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease. Nature Genet 2005;37:243–253. [PubMed: 15711544]
57. Breitling R, et al. Genetical genomics: spotlight on QTL hotspots. PLoS Genet 2008;4:e1000232. [PubMed: 18949031]

58. Benfey PN, Mitchell-Olds T. From genotype to phenotype: systems biology meets natural variation. Science 2008;320:495–497. A thought-provoking review of how natural variation in gene expression can be used for network and other systems analysis. Although the focus is on plants, the ideas can be translated to all organisms. [PubMed: 18436781]
59. Jordan IK, Marino-Ramirez L, Wolf YI, Koonin EV. Conservation and coevolution in the scale-free human gene coexpression network. Mol Biol Evol 2004;21:2058–2070. [PubMed: 15282333]
60. Lee HK, Hsu AK, Sajdak J, Qin J, Pavlidis P. Coexpression analysis of human genes across many microarray data sets. Genome Res 2004;14:1085–1094. [PubMed: 15173114]
61. Gargalovic PS, et al. Identification of inflammatory gene modules based on variations of human endothelial cell responses to oxidized lipids. Proc Natl Acad Sci USA 2006;103:12741–12746. [PubMed: 16912112]
62. Harbison ST, et al. Co-regulated transcriptional networks contribute to natural genetic variation in *Drosophila* sleep. Nature Genet 2009;41:371–375. [PubMed: 19234472]
63. Ayroles JF, et al. Systems genetics of complex traits in *Drosophila melanogaster*. Nature Genet 2009;41:299–307. [PubMed: 19234471]
64. Aten JE, Fuller TF, Lusis AJ, Horvath S. Using genetic markers to orient the edges in quantitative trait networks: the NEO software. BMC Syst Biol 2008;2:34. [PubMed: 18412962]
65. Chen Y, et al. Variations in DNA elucidate molecular networks that cause disease. Nature 2008;452:429–435. [PubMed: 18344982]
66. Ghazalpour A, et al. Integrating genetic and network analysis to characterize genes related to mouse weight. PLoS Genet 2006;2:e130. [PubMed: 16934000]
67. Yang X, et al. Validation of candidate causal genes for obesity that affect shared metabolic pathways and networks. Nature Genet 2009;41:415–423. [PubMed: 19270708]
68. Quigley DA, et al. Genetic architecture of mouse skin inflammation and tumour susceptibility. Nature 2009;457:505–508. [PubMed: 19136944]
69. Smirnov DA, Morley M, Shin E, Spielman RS, Cheung VG. Genetic analysis of radiation-induced changes in human gene expression. Nature 2009;459:587–591. [PubMed: 19349959]
70. Li Y, et al. Mapping determinants of gene expression plasticity by genetical genomics in *C. elegans*. PLoS Genet 2006;2:e222. [PubMed: 17196041]
71. Smith EN, Kruglyak L. Gene–environment interaction in yeast gene expression. PLoS Biol 2008;6:e83. [PubMed: 18416601]
72. Spielman RS, McGinnis RE, Ewens WJ. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am J Hum Genet 1993;52:506–516. [PubMed: 8447318]
73. Abecasis GR, Cardon LR, Cookson WO. A general test of association for quantitative traits in nuclear families. Am J Hum Genet 2000;66:279–292. [PubMed: 10631157]
74. Abecasis GR, Cookson WO, Cardon LR. Pedigree tests of transmission disequilibrium. Eur J Hum Genet 2000;8:545–551. [PubMed: 10909856]
75. Lunetta KL, Faraone SV, Biederman J, Laird NM. Family-based tests of association and linkage that use unaffected sibs, covariates, and interactions. Am J Hum Genet 2000;66:605–614. [PubMed: 10677320]
76. Rice TK, Schork NJ, Rao DC. Methods for handling multiple testing. Adv Genet 2008;60:293–308. [PubMed: 18358325]

Figure 1. Inter-individual variation in gene expression levels

Shown are the expression levels of 12 genes in unrelated individuals. Each circle represents the expression level in one individual. The expression levels of two genes, *PARK7* and *ATP5J2*, are less variable than the other 10 genes, even though the 12 genes were measured using the same cells and methods.

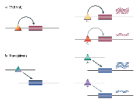

Figure 2. Effect of *cis*- and *trans*-acting DNA variants on expression levels of genes

Polymorphic forms of regulators that act in *cis* (**a**) or in *trans* (**b**) to the target genes (also called local and distal regulators, respectively) result in variation in expression levels of the target genes. *Cis*-acting variants are found close to the target genes and *trans*-acting variants are located far from the target genes, often on another chromosome. Different allelic forms of the *cis*- and *trans*-acting variants have different influence on gene expression. In this example, individuals with the G variant of the *cis* regulator have a higher expression level of the target gene than individuals with the C variant of the regulator. Similarly, individuals with the A variant of the *trans* regulator have a higher expression level of the target gene than those with the T variant of the regulator.

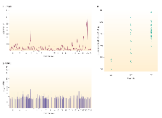

Figure 3. The expression level of copine I (*CPNE1*) is *cis* regulated

Results from linkage (**a**) and association (**b**) studies show that the polymorphic regulator of expression level of *CPNE1* is found on chromosome 20, close to the target gene (*CPNE1* is located on chromosome 20). A linear regression analysis of the expression level of *CPNE1* on the genotypes of the SNP rs3787165 in *CPNE1* (**c**) uncovered marked association between the SNP genotypes and expression levels of *CPNE1*; TT is associated with higher expression.

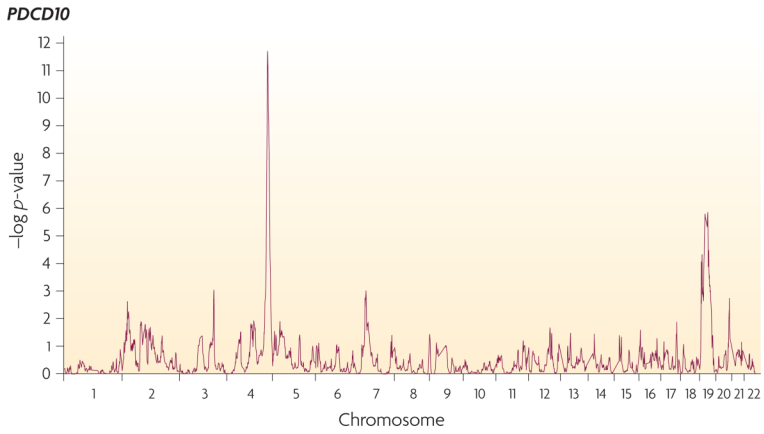

Figure 4. The expression level of programmed cell death 10 (*PDCD10*) is *trans* regulated

PDCD10 is located on chromosome 3. Linkage results show that regulators of this gene are located on chromosomes 4 and 19.