# Idiomatic (gene) expressions

**Matthew V. Rockman**

**Summary**

**Hidden among the myriad nucleotide variants that constitute each species' gene pool are a few variants that contribute to phenotypic variation. Many of these differences that make a difference are non-coding *cis*-regulatory variants, which, unlike coding variants, can only be identified through laborious experimental analysis. Recently, Cowles et al.[1] described a screening method that does an end-run around this problem by searching for genes whose *cis* regulation varies without having to find the polymorphic nucleotides that influence transcription. While we will continue to require a diverse arsenal of experimental methods, this versatile method will speed the identification of functional genetic variation. *BioEssays* 25:421–424, 2003.**


## Introduction

There are only a few major kinds of genetic variation that can influence phenotypic variation: amino acid replacements that alter protein function, gene duplications and deletions, and *cis*-regulatory variants that influence gene expression—its induction, level, developmental timing and spatial pattern. While variation in the first two categories is easily detected by examining DNA sequences, variation in the third category, *cis*-regulatory variation, has proved refractory to bioinformatic discovery; the regulatory needles are buried in a haystack of functionless variants. Although *cis*-regulatory variation has been forwarded as the major genetic basis for phenotypic variation and evolution,[2–6] we know remarkably little about its distribution in nature.

Traditionally, the way to verify the functional consequences of a *cis*-regulatory variant is to clone the *cis*-regulatory DNA from each allele into otherwise identical reporter constructs, which are then transfected into cultured cells. Differences in reporter activity point to differences in *cis*-regulation. This approach has been remarkably successful at validating suspected functional variants; it has verified the effects of *cis*-regulatory polymorphisms on the transcription of more than 100 human genes.[6] But the technique has severe drawbacks. First, it requires that candidate functional variants be identified. Because *cis*-regulatory DNA can extend over tens or hundreds of kilobases from the start of transcription, there are often too many variants to test. Second, the effects of *cis*-regulatory variation are notoriously context dependent. Because transcriptional initiation involves dozens of protein–DNA and protein–protein interactions, failure to include particular DNA elements in a reporter construct may result in false negatives, suggesting that variants are functionless when in fact they are missing required functional partners. Third, the use of plasmid DNA transfected into transformed cell lines raises the possibility of misleading artifacts—due to supercoiling of the plasmids or the high concentration of plasmid relative to transcription factors, for example. Finally, generating and evaluating allelic reporter constructs is an expensive and time-consuming method ill-suited to a systematic search for regulatory variants.

Department of Biology, Duke University, Box 90338, Durham, NC 27708, USA. E-mail: mrockman@duke.edu
DOI 10.1002/bies.10279
Published online in Wiley InterScience (www.interscience.wiley.com).

## The method

Cowles et al.[1] do away with the need to identify and experimentally analyze regulatory variants by ignoring the variants; instead they focus on finding the genes whose regulation is influenced by the variants. The method is simple: allele-specific quantitative RT-PCR.

Cowles et al.[1] crossed isogenic strains of mice to produce $F_1$ mice bearing one allele at each locus from each parental strain (Fig. 1). They then estimated the relative abundance of the RNA transcribed from each allele at each locus (69 genes in their initial report); differences in the abundance of allelic mRNAs point to *cis*-regulatory differences between the parental strains.

The method eliminates potential sources of artifact and controls for the usual confounding variables. The problems of artifact associated with generating reporter constructs and introducing them into cell lines are solved by comparing expression from two alleles in their native chromosomal context in vivo. Moreover, the two alleles compete against one another in a common cellular environment; genetic background and environmental variation are therefore eliminated from the equation. Any allelic difference in transcript abundance is necessarily related to genetic differences in *cis*.

The basic requirement for allele-specific quantitative RT-PCR is a means of identifying the allelic source of the transcripts. Cowles et al.[1] use single nucleotide polymorphisms (SNPs) in the transcripts, which lend themselves to easy measurement in single base extension assays (Fig. 1). The necessity for allele-specific markers is perhaps the largest impediment to wider application of the method; comparisons

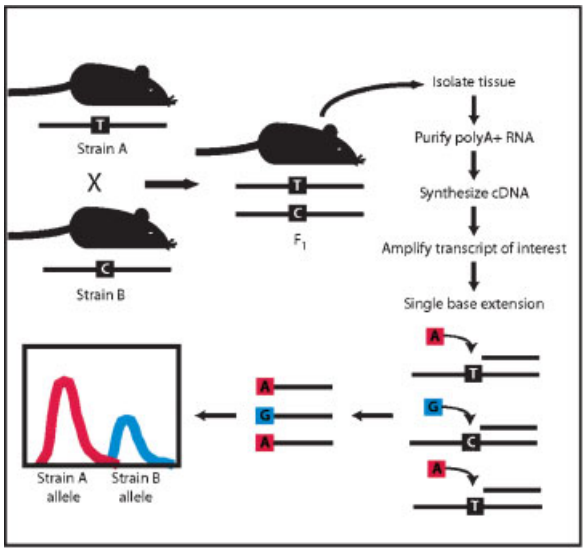


**Figure 1.** Cowles et al.[1] crossed mice from strains that differ at a single nucleotide polymorphism within a transcript of interest. The $F_1$ mice express both transcripts, but the abundance of each is a function of regulatory variants linked in *cis* with the marker SNP, which here is a C/T polymorphism. After the transcript is amplified from a tissue of interest, the relative abundance of the two allelic versions of the transcript is estimated by single base extension: the target strand of the cDNA is incubated with fluorescently labeled nucleotides and primers that terminate at the base adjacent to the SNP; in effect, single base extension is a one-base-long sequencing reaction. The results of single base extension can be read on an automated sequencer; here, the transcript from the strain A chromosome, with the T allele of the marker SNP, is more abundant than the transcript from the strain B chromosome. Cowles et al.[1] use DNA from the tissue of interest, and pooled DNA from the parental strains, as standards to calibrate the sequencer reads.

among strains bearing the same marker allele are impossible, and many genes lack common exonic variants. Cowles et al.[1] could perform only 41% of the possible pairwise comparisons among four strains for the 69 genes; for the other comparisons, the parental strains did not differ at the marker SNP.

## How much variation?

Four of the 69 genes examined showed allelic differences in gene expression of at least 1.5-fold among the strains. This result points to a large amount of segregating *cis*-regulatory variation genome-wide, but, as Cowles et al.[1] are careful to note, their numbers are certainly underestimates, for several reasons.

First, most genes were only compared among two or three of the four strains; consequently, their result is less like a survey of variation in a population and more like an estimate of regulatory heterozygosity. By comparison, human nucleotide heterozygosity is less than one in a thousand, but in a larger population survey, nucleotide variants typically occur every several hundred bases.[7]

Second, the results are based on expression in only three tissues: spleen, liver and brain from adult females. Every additional tissue or developmental stage examined provides another chance to find allelic differences in gene expression. Fortunately, the versatile method of Cowles et al.[1] can be extended to any number of tissues from a single mouse, and tissues from $F_1$ sib mice can be harvested from different developmental stages. Sib mice can also be raised under different environmental conditions; genotype-by-environment interaction is known to be a common phenomenon in gene expression variation.[6] In the future, it will also be valuable to examine both males and females; sex-specific genetic variation in gene expression may be quite common, as suggested by experiments in *Drosophila*.[8]

Ultimately, all methods of finding regulatory variation face a major false-negative problem. Gene expression is sensitive to a vast multidimensional array of variables, including not just tissue type, developmental stage and sex, but also temperature, presence of inductive cues, intracellular ion concentration, and so on. Until heritable variation in gene expression can be assessed over the full state space of these variables, all estimates of its magnitude will fall short of the true value.

Nevertheless, Cowles et al.'s[1] results are quite informative. Of the four genes confirmed to show allelic differences in expression, one is expressed in one tissue only, one is differentially expressed in all three tissues, and two show allelic differences in one tissue, but not the others. These data point to a question of growing importance in quantitative, developmental and evolutionary genetics: to what extent is gene expression variation correlated among expression domains? This question is at the heart of such ideas as modularity, evolvability and coordinate pleiotropy.[9–12] Cowles et al.'s four data points show that expression variation can be, but need not be, correlated among tissues.

## Genetic backgrounds

Allele-specific quantitative RT-PCR controls perfectly for genetic background, but that does not mean that genetic background can be ignored. One peculiarity of measuring expression in the genetic background of an $F_1$ hybrid is the high potential to discover allelic variation in *cis*-regulation that is not present in either parental strain. The simplest scenario involves one strain with a transcriptional activator but no *cis*-regulatory site for it to bind, and the other strain with a binding site but no transcription factor. In this case, the two alleles will exhibit differential expression in the $F_1$, where the transcription factor from one strain can bind the *cis*-regulatory allele from the other, but the two parental strains will not differ from one another. The *cis*-regulatory polymorphism is real, but the $F_1$

results are misinformative about expression differences between the parents.

Should such epistasis, or interaction among loci, be a major concern? Theoretical arguments and empirical evidence suggest that the scenario described above may be common.

First, transcriptional regulation is notoriously polygenic, so there are many potential epistatic actors, all of which (to a first approximation) will be expressed in the $F_1$. In the simplest case, the *trans*-acting element is a transcription factor, but it could equally be a DNA acetylase, a transcriptional cofactor, a kinase that activates the cofactor, a diffusable signaling molecule that leads to the activation of the kinase, or even a calcium channel that alters the intracellular ion concentration. Quantitative models of variation in gene networks have consistently found that epistasis is a common property of such systems.[13–16]

Second, the limited empirical evidence points to rampant epistasis in the genetic basis of variation in gene expression. In a classic study, Damerval et al.[17] used two-dimensional gels to measure the abundance of 72 proteins in an $F_2$ progeny from a cross of maize lines. They then treated each protein as a quantitative trait and mapped quantitative trait loci (QTL) underlying the observed variation. Their analysis revealed that interlocus interactions, epistasis, contributed to variation in abundance of ten of the proteins, 14% of the total. As Damerval et al.[17] note, such pervasive epistasis in not commonly encountered in QTL mapping of agronomic traits; the proximity of gene expression phenotypes to their genetic bases may make the underlying intermolecular interactions more important contributors to observed variation.

Further empirical evidence comes from microarray comparisons of gene expression among mouse strains, which suggest that a substantial fraction of genes are differentially expressed among strains; each of these genes is available to act epistatically with *cis*-regulatory variants. For example, Karp et al.[18] found that 739 of 2718 genes—27%—are differentially expressed in the lungs of two laboratory strains of mice.

## Complementary approaches

One time-consuming but powerful way around the problem of a hybrid genetic background is to use repeated backcrosses into each parental strain to produce mice congenic for each *cis*-regulatory allele in each genetic background. Congenic mice have been used successfully to discover *cis*-regulatory variation since early work on *β-glucuronidase*.[19] More recently, Rozzo et al.[20] used microarrays to measure gene expression from mice congenic for an interval on chromosome 1 and observed that only genes on that interval showed expression differences between the congenic and parental strain. The *cis*-regulatory basis of the difference was confirmed by allele-specific quantitative RT-PCR on backcross mice heterozygous only at the interval of interest.

Another approach is to map the genetic basis of gene expression variation. Expression of thousands of genes can be measured at once on microarrays, and the expression level of each treated as a quantitative trait. Brem et al.[21] mapped QTL for expression variation for 570 yeast genes. For about 36% of these genes, expression variation mapped to the gene locus itself, indicating *cis*-acting differences between the parental strains.

QTL mapping has advantages and disadvantages relative to the Cowles et al.[1] method. On the one hand, quantitative genetics requires a vastly larger number of crosses and measurements. On the other hand, QTL mapping with microarrays does away with the need to develop and test gene-specific assays. Moreover, QTL mapping can explicitly accommodate epistatic interactions between *cis*- and *trans*-acting variants. The numbers are also telling: the Brem et al.[17] analysis of yeast identified about 200 genes with *cis*-regulatory variation, while Cowles et al.[20] screened only 69 genes, and found variation at only four.

## Elaborations on the allele-specific method

There is no best way to find regulatory variation, but the Cowles et al.[1] method adds a potent tool to compare alleles in vivo, and the method has a proven track record in human genetics. Though the inability to cross isogenic lines has prohibited its use as a general screening method in humans, allele-specific quantitative RT-PCR has been used to validate the effect of suspected functional variants. The approach is to measure gene expression in individuals heterozygous for the transcribed marker variant when that variant is in demonstrated linkage disequilibrium with suspected functional variants. To date, the method has confirmed allelic differences in transcription at *HLA-DQB1, INS*, and *UCP2*.[22–24] The human work also points to potential extensions of the allele-specific quantitation method. In a study of the *COL1A1* locus, Mann et al.[25] used allele-specific quantitative RT-PCR on total RNA rather than the poly(A) fraction; this allowed them to use a marker SNP in an intron. Such an approach allows the method to be applied to the many human genes lacking common exonic SNPs.

Protein-based allele-specific methods have also been developed. Klose et al.[26] applied allele-specific quantitation to mouse brain proteins using two-dimensional gel analysis of a cross between a laboratory mouse strain and *Mus spretus*. Of 293 proteins whose allelic source could be discriminated on the gels (i.e., those that differ in mass or charge between the two parental strains), 221 also showed differences in abundance. Of these, 134 were distinct enough to allow further analysis in additional crosses. In these crosses, cosegregation of the protein migration pattern and protein abundance points to the role of *cis*-acting differences between the parental strains. Some 98 proteins, or 73% of those measured, proved to have *cis*-regulatory differences between the strains.

Damerval et al.'s[17] earlier study of maize produced similar (though fewer) data: of 15 loci with distinguishable alleles, ten showed abundance differences, and for seven of these the abundance differences cosegregated with protein migration pattern.

## Conclusion

Allele-specific methods, of which the Cowles et al.[1] approach is the most systematic, promise to rapidly identify genes whose expression varies among individuals due to *cis*-regulatory polymorphism. In conjunction with QTL mapping, we now have the tools to find *cis*-regulatory variation efficiently. Ultimately, however, we will want to know which individual nucleotides contribute to the variation that we find, and we will want to know the mechanisms by which the variants act. While we will therefore need to retain reporter constructs in our experimental arsenal, Cowles et al.'s method will allow us to focus our efforts.

## References


1. Cowles CR, Hirschhorn JN, Altshuler D, Lander ES. Detection of regulatory variation in mouse genes. Nat Genet 2002;32:432–437.
2. Britten RJ, Davidson EH. Gene regulation for higher cells: a theory. Science 1969;165:349–357.
3. King MC, Wilson AC. Evolution at two levels in humans and chimpanzees. Science 1975;188:107–116.
4. Stern DL. Evolutionary developmental biology and the problem of variation. Evolution 2000;54:1079–1091.
5. Carroll SB, Grenier JK, Weatherbee SD. From DNA to diversity: molecular genetics and the evolution of animal design. London: Blackwell Science. 2001.
6. Rockman MV, Wray GA. Abundant raw material for *cis*-regulatory evolution in humans. Mol Biol Evol 2002;19:1991–2004.
7. Przeworski M, Hudson RR, Di Rienzo A. Adjusting the focus on human variation. Trends Genet 2000;16:296–302.
8. Jin W, Riley RM, Wolfinger RD, White KP, Passador-Gurgel G, Gibson G. The contributions of sex, genotype and age to transcriptional variance in *Drosophila* melanogaster. Nat Genet 2001;29:389–395.
9. Cheverud JM, Rutledge JJ, Atchley WR. Quantitative genetics of development: genetic correlations among age-specific trait values and the evolution of ontogeny. Evolution 1983;37:895–905.
10. Kirchhamer CV, Yuh CH, Davidson EH. Modular *cis*-regulatory organization of developmentally expressed genes: two genes transcribed territorially in the sea urchin embryo, and additional examples. Proc Natl Acad Sci USA 1996;93:9322–9328.
11. Raff R. The shape of life: genes, development, and the evolution of animal form. Chicago: University of Chicago Press. 1996.
12. Beldade P, Koops K, Brakefield PM. Modularity, individuality, and evo-devo in butterfly wings. Proc Natl Acad Sci USA 2002;99:14262–14267.
13. Gibson G. Epistasis and pleiotropy as natural properties of transcriptional regulation. Theor Pop Biol 1996;49:58–89.
14. Nijhout HF, Paulsen SM. Developmental models and polygenic characters. Am Nat 1997;149:394–405.
15. Omholt SW, Plahte E, Øyehaug L, Xiang K. Gene regulatory networks generating the phenomena of additivity, dominance, and epistasis. Genetics 2000;155:969–980.
16. de Vienne D, Bost B, Fiêvet J, Zivy M, Dillmann C. Genetic variability of proteome expression and metabolic control. Plant Physiol Biochem 2001;39:271–283.
17. Damerval C, Maurice A, Josse JM, de Vienne D. Quantitative trait loci underlying gene product variation: a novel perspective for analyzing regulation of genome expression. Genetics 1994;137:289–301.
18. Karp CL, et al. Identification of complement factor 5 as a susceptibility locus for experimental allergic asthma. Nat Immun 2000;1:221–226.
19. Paigen K. Experimental approaches to the study of regulatory evolution. Am Nat 1989;134:440–458.
20. Rozzo SJ, Allard JD, Choubey D, Vyse TJ, Izui S, Peltz G, Kotzin BL. Evidence for an interferon-inducible gene, *Ifi202*, in the susceptibility to systemic lupus. Immunity 2001;15:435–443.
21. Brem RB, Yvert G, Clinton R, Kruglyak L. Genetic dissection of transcriptional regulation in budding yeast. Science 2002;296:752–755.
22. Bennett ST, et al. Susceptibility to human type 1 diabetes at *IDDM2* is determined by tandem repeat variation at the insulin gene minisatellite locus. Nat Genet 1995;9:284–292.
23. Beaty JS, West KA, Nepom GT. Functional effects of a natural polymorphism in the transcriptional regulatory sequence of *HLA-DQB1*. Mol Cell Biol 1995;15:4771–4782.
24. Esterbauer H, et al. A common polymorphism in the promoter of *UCP2* is associated with decreased risk of obesity in middle-aged humans. Nat Genet 2001;28:178–183.
25. Mann V, Hobson EE, Li B, Stewart TL, Grant SFA, Robins SP, Aspden RM, Ralston SH. A *COL1A1* Sp1 binding site polymorphism predisposes to osteoporotic fracture by affecting bone density and quality. J Clin Invest 2001;107:899–907.
26. Klose J, et al. Genetic analysis of the mouse brain proteome. Nat Genet 2002;30:385–393.