

Genetics of Quantitative Variation in Human Gene Expression

V.G. CHEUNG,*† K.-Y. JEN,* T. WEBER,* M. MORLEY,* J.L. DEVLIN,†
K.G. EWENS,† AND R.S. SPIELMAN†

Departments of *Pediatrics and †Genetics, University of Pennsylvania,
Philadelphia, Pennsylvania 19104

The extent of variation among individuals at the DNA sequence level has been well characterized. The goal of many genetic studies is to determine the consequences of these sequence variants, for both normal and disease phenotypes. We have extended the study of genome variation from the sequence to mRNA transcript level, with the goal of understanding natural variation in gene expression in humans. We began by measuring the quantitative differences in expression levels of genes among normal individuals and determining whether there is an inherited component to this variation. We found a set of genes whose expression levels are highly variable in lymphoblastoid cells prepared from white blood cells of normal individuals. For these genes, we observed that genetically related individuals tend to have more similar transcript levels than unrelated individuals. This suggests that there is a genetic component in gene expression phenotype. Next, we are identifying the sequence differences that control variation in gene expression phenotype in a *cis*- or *trans*-acting manner. Like other quantitative traits, baseline variation in gene expression levels is likely to be regulated by a variety of genetic determinants, as well as environmental effects.

GENOME VARIATION

The study of genetic polymorphisms in the human genome has evolved from analysis of variation in proteins to DNA sequence and now mRNA. Early characterization of the extent of variation was performed on blood proteins using electrophoretic techniques. During the 1960s, polymorphic variants of many proteins were discovered (Harris 1966, 1969). Subsequently, as methods for analysis of DNA became available, the frequency of DNA sequence variants was estimated for specific regions of the genome and then extrapolated to genome-wide estimates (Jeffreys 1979; Ewens et al. 1981). These studies provided estimates of the extent of natural variation in DNA sequence and in proteins in humans. Information about the frequency of DNA sequence variants (about 1 per 1000 bp) has been important for understanding population structure and also for disease gene mapping.

Recent advances in microarray technology have allowed extension of the study of variation at the mRNA level to a genomic scale. The extent of intra- and inter-species variation in gene expression has been assessed in primates (Enard et al. 2002), and various approaches have shown that there is appreciable variation in gene expres-

sion in other species, including mice, fish, and yeast (Cowles et al. 2002; Enard et al. 2002; Oleksiak et al. 2002; Steinmetz et al. 2002; Townsend et al. 2003). The genetic *control* of variation in gene expression has been explored in various organisms from yeast to man (Cowles et al. 2002; Enard et al. 2002; Yan et al. 2002; Cheung et al. 2003; Schadt et al. 2003; Yvert et al. 2003), mainly by focusing on preferential expression of one allele in heterozygous individuals (Cowles et al. 2002; Yan et al. 2002; Lo et al. 2003). The systematic study of natural variation in human gene expression is still in its infancy.

HUMAN GENE EXPRESSION AS A COMPLEX TRAIT

There are several reasons for focusing our interest on *natural* variation in gene expression in humans. First, with development of high-throughput tools such as microarrays and serial analysis of gene expression, there have been many studies that compared expression profiles of normal versus diseased cells. However, few studies have analyzed natural variation in unaffected control individuals. This baseline information is important for assessing the significance of the gene expression in disease. Second, expression level of genes is a phenotype that can be measured quite precisely in a large number of unrelated and related individuals. Therefore, the "expression phenotype" can be analyzed genetically as a quantitative trait in order to identify the determinants of the variation in gene expression. The mechanisms that control transcription remain largely unknown. Globally, some control mechanisms are known, such as regulation (1) at the synthesis step, by modulating transcription initiation and elongation and (2) at the decay step, by changing the stability of transcripts. However, for most individual genes, the specific regulatory mechanisms are unknown. Linkage-based methods allow us to map genetically variable transcriptional control elements in the genome without having to know in advance whether regulation occurs via a *cis*- or *trans*-acting mechanism.

Finally, expression levels of genes are intermediate phenotypes that will be useful in developing better methods for analyzing quantitative traits. Studies of the genetic basis of monogenic (qualitative) conditions have been very successful. However, the genetic basis of most common complex traits remains poorly understood, in part because of the difficulties they pose for statistical ge-

netic analysis. Expression phenotypes are good models for developing molecular and statistical tools for analyzing quantitative traits more generally. In comparison to other human phenotypes, it is relatively easy to measure a large number of expression phenotypes at once in the same person. There are various mechanisms that might regulate gene expression in humans. In some cases, the expression level of a gene is regulated only by a *cis*-acting element. In others, the expression level is regulated by several genes that act in *trans* or a combination of regulatory elements. These levels of complexity make expression phenotypes good models for the many components underlying complex traits and diseases in humans.

ANALYSIS OF NATURAL VARIATION IN HUMAN GENE EXPRESSION

Our goals are (1) to define the extent of variation in gene expression in normal individuals and (2) to determine the genetic basis of that variation. We analyzed the expression levels of genes in lymphoblastoid cells from 50 unrelated individuals from the Utah pedigrees of the Centre d'Etude du Polymorphisme Humain (CEPH) using microarrays (Affymetrix Human Genome Focus Chip) that contain about 8500 human genes. Expression profiling for each sample was performed in duplicate. As a measure of variation in gene expression, for each gene, we calculated the variance ratio (variance among individuals divided by mean variance between microarray replicates on same individual). This allows us to characterize variation among individuals relative to measurement noise (Cheung et al. 2003). We ranked the genes by variance ratio and focused on those that are the most variable (i.e., with the highest variance ratio) since they are likely to be more amenable to genetic dissection. Then, we measured the expression levels of some of these "variable" genes in individuals in large families and also in sets of monozygotic twin pairs.

LYMPHOBLASTOID CELLS AS RNA SOURCE FOR GENE EXPRESSION ANALYSIS

We have chosen to use lymphoblastoid cells in our study for several reasons. First, we need an RNA source that can be obtained from a large number of normal individuals in large pedigrees. Immortalized lymphocytes (transformed by EBV) are available from all the members of the CEPH pedigrees. These are exceptionally large three-generation families that have been studied extensively. Genotypes for genetic mapping are available for many of these families, which will facilitate our effort to map the genetic determinants of variation in gene expression. Second, lymphoblastoid cells from our study subjects can be grown under near-identical conditions. It has been shown that large differences are found in expression levels of genes that are studied on different occasions in fresh blood samples from the same individual (Whitney et al. 2003). With the lymphoblastoid cells, we can control the growth conditions in order to minimize environmental variation.

We were concerned about how transformation may affect gene expression. To examine this issue, we compared

the differences in expression levels of 15 highly variable genes among unrelated individuals to those between monozygotic (MZ) twins. We found that in these transformed cells, the variance within MZ twin pairs, as a fraction of variance among unrelated individuals, ranged from 0.002 to 0.57 (mean 0.19, median 0.19). These findings indicated that the expression levels of genes are highly correlated among monozygotic twins compared to unrelated individuals. Differences in expression levels of genes in lymphoblastoid cells reflect germ-line genetic differences, despite the transformation process.

VARIABLE GENES

We analyzed data for ~3800 (45%) of the 8500 genes, restricting attention to those that were expressed in at least 10 out of the 50 unrelated individuals who were studied. Among these genes, the variance ratios ranged from 0.2 to 48.9 (mean 2.6, median 1.6). For most of the genes, the variance of expression levels between individuals is higher than the variance between microarray replicates (Fig. 1). This shows the reproducibility of the microarray data. In our data, some genes have variance ratios that are approximately 1; for these genes, there is no evidence for meaningful variation among individuals. Therefore, in order to maximize the chance to detect genetic differences that account for variation in gene expression, we focus on the genes with variance ratio appreciably greater than 1. We expect that these are the genes where variation is most likely to be biologically meaningful, and therefore potentially due to genetic differences. The points in red in Figure 1 represent the genes (top 5%) with the highest variance ratios.

We examined the genomic location of these "variable" genes (Fig. 2). We found that they were not clustered in the genome; instead, they mapped to many sites across the genome. Figure 2 shows the chromosomal locations of the 50 most variable genes. For these genes, the range

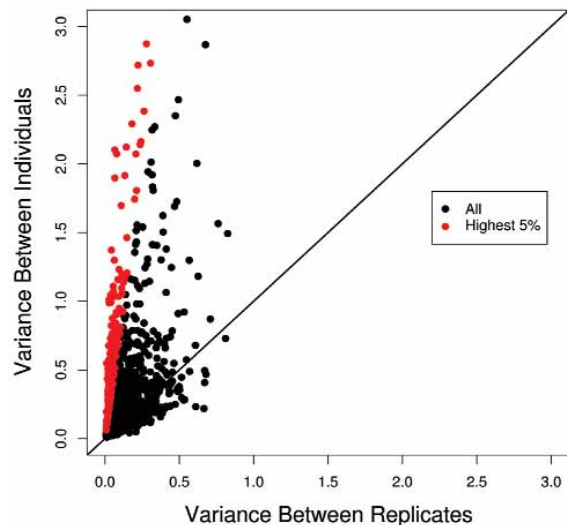


Figure 1. Scatter plot of variance in expression levels between individuals and between microarray replicates for ~3800 genes. The genes with the highest variance ratios (top 5%) are highlighted in red. The solid line indicates a variance ratio of 1.0.

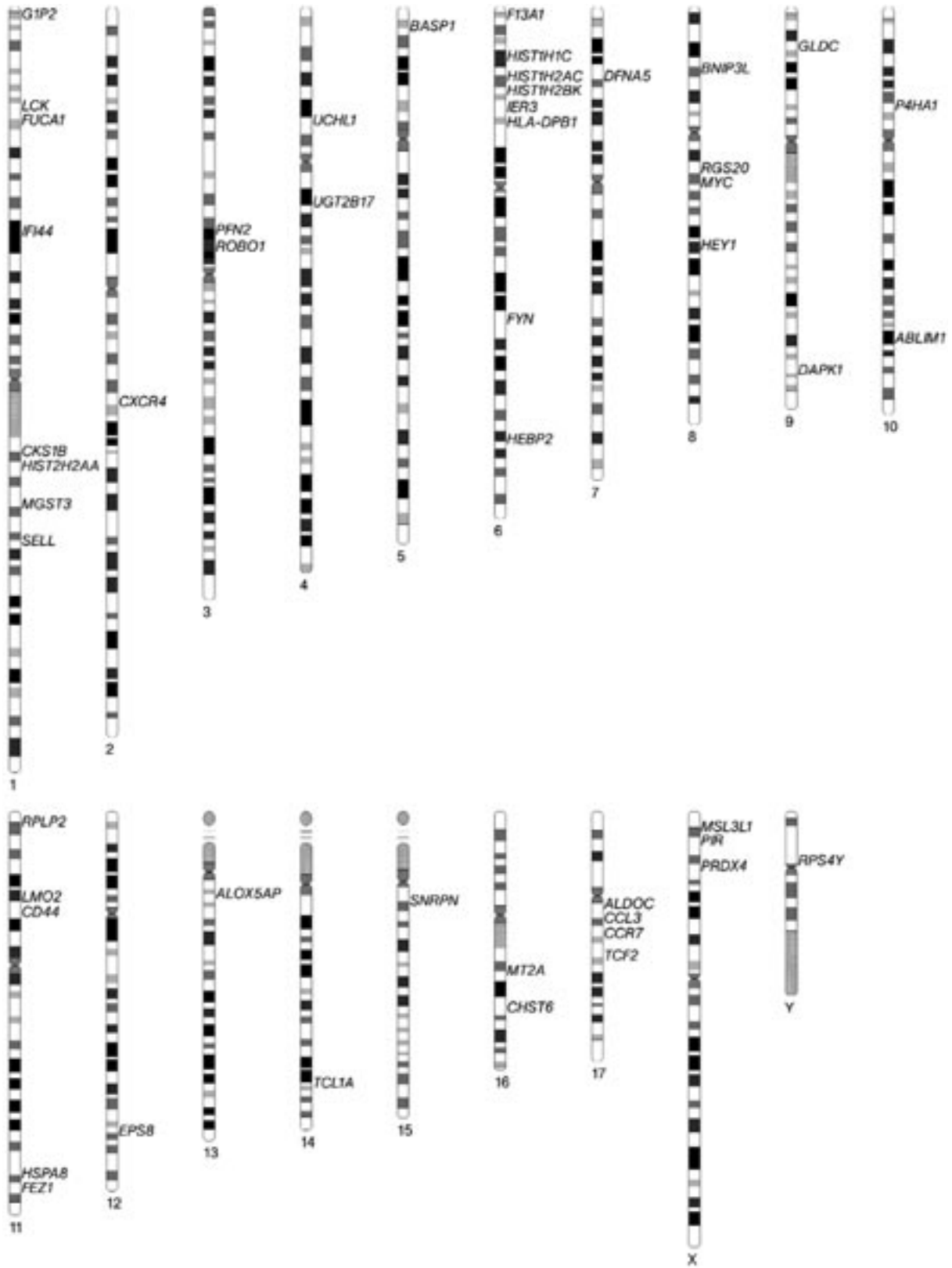


Figure 2. Genomic locations of 50 most variable genes. Chromosomes that do not contain any of these genes are not included in the figure.

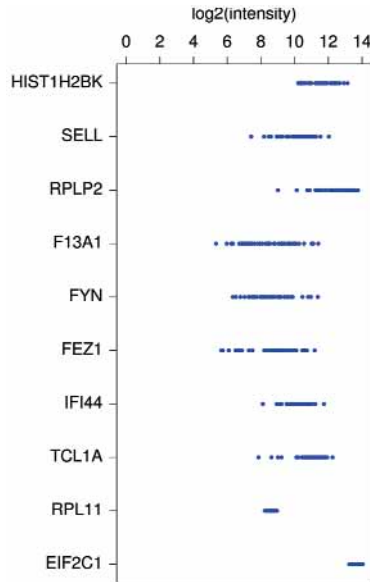


Figure 3. The expression levels measured by microarrays for eight highly variable and two relatively nonvariable genes in 50 individuals. Each point represents the expression level for an individual.

of expression levels among the individuals studied was from 14- to 48-fold. The expression levels of 8 of the highly variable genes and 2 relatively nonvariable genes for the 50 unrelated subjects are shown in Figure 3.

EVIDENCE FOR HERITABILITY

We assessed the heritability of expression variation in two ways. In the first method, we used data from sets of individuals with different degrees of relatedness, and in the second, we used the resemblance between offspring and parent. Figure 4 shows the results from the first approach and compares the variance in expression level among three groups of subjects: 50 unrelated members of CEPH families (parents of the large CEPH sibships), 10 sets of siblings (also CEPH sibships), and 10 pairs of MZ twins. We found that the variance is largest among the unrelated individuals (a sample from the European popula-

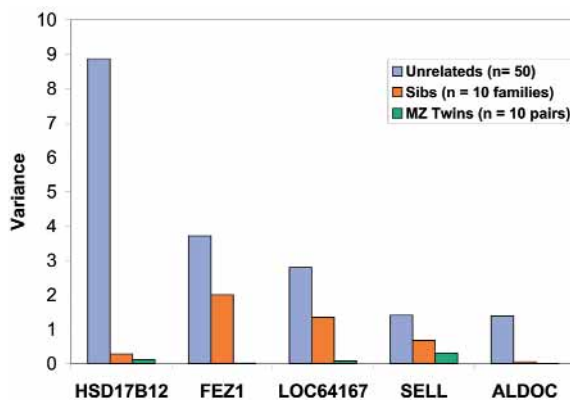


Figure 4. Variance in expression level for five genes. Quantitative RT-PCR data for 50 unrelated individuals, 89 offspring in 10 CEPH sibships, and 10 sets of monozygotic twins.

tion) and smallest between members of the MZ twin pairs. For all 5 genes shown in Figure 4 (and for 10 others that we have tested), we found the following pattern: as the degree of relatedness increases, the variance in expression levels decreases, suggesting that genetic (germ-line) differences contribute to variation in gene expression.

Although this kind of comparison of variances provides a valid initial assessment of evidence for heritability in some sense, other, more standard methods have the advantage of providing numerical estimates of the conventional measure, the so-called “narrow-sense” heritability. In addition to data for the CEPH parents, we have obtained gene expression data for their parents (CEPH grandparents). It is known that the regression of offspring on mid-parent value (in standard linear regression) provides an estimate of the desired heritability (Falconer and MacKay 1996), so we used the expression data for the unrelated CEPH parents and *their* parents to estimate this regression coefficient. A striking example is shown for glutathione *S*-transferase M2, *GSTM2*, in Figure 5. We have estimated the regression of parent on mid-grandparent this way for all ~3800 genes on the microarray that were expressed in the lymphoblastoid cells. Approximately 50% of these values are negative, and thus give no evidence for a heritable component. Viewing the collection as a whole, if no genes showed evidence for heritability, we would expect that the distribution of regression estimates would have as many negative as positive values. Instead, we find an excess of large positive values. The regression coefficient (b) is greater than 0.5 for 94 genes, but $b < -0.5$ for only 48. Similarly, we find $b > 0.75$ for 9 genes, but $b < -0.75$ for only 4 genes.

Of course, we consider these estimates at best crude indicators of heritability, especially since we have looked at so many genes. However, our interest in the estimate of heritability is not mainly as an end in itself; we use it in conjunction with the variance ratio to select genes for further study. In the follow-up studies, usually by RT-PCR, we will carry out group comparisons (sibships, twins, etc.) like those shown in Figure 4, tests of closely linked

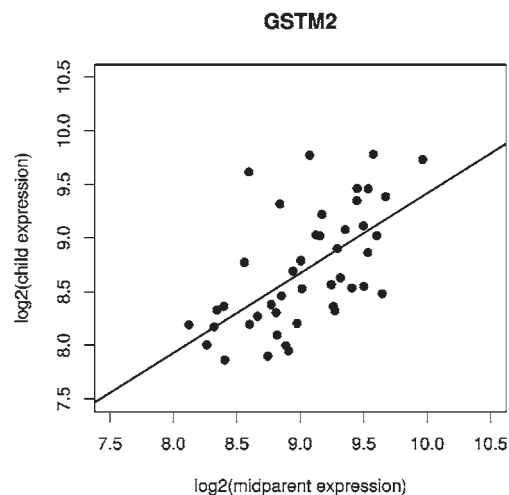


Figure 5. Regression plot of expression level of *GSTM2* for 50 offspring–midparent pairs. The slope (represented by *solid line*) for the regression is 0.75 with a standard error of 0.45.

SNPs for association with expression level, and genome scans to identify unlinked genetic determinants that influence expression.

CONCLUSIONS

In this study, we have assessed the extent of natural variation in gene expression in humans. Our results suggest that there is a genetic component to this variation. Next, we will identify the genetic determinants for this variation. Identification and characterization of these determinants, called “expression control elements (ECEs)” by Cheung et al. (2003), will lead to a better understanding of transcriptional control. Less than 10% of the genome represents coding regions. Comparative genomic studies have shown that a substantial portion of the non-coding sequence is conserved between species, so it is believed that these conserved noncoding regions play a role in regulating gene expression and function. By using genetic approaches such as genome scans and methods of quantitative trait locus (QTL) analysis to map the determinants of gene expression, we expect to identify new elements that regulate gene expression. Some of these ECEs will regulate transcription in a *cis*-acting manner, whereas others will act via *trans*-acting mechanisms. We expect that a combination of association studies and genome scans will allow us to discover both types of determinants.

So far, we have considered the expression level of each gene as a separate phenotype. However, it is also possible to consider the coordinated expression of correlated genes as a single complex phenotype. It is likely that *trans*-acting regulators influence expression of several to many genes, directly or indirectly. Thus, it is reasonable to expect to find these determinants by clustering the genes by their expression phenotypes, and using these clusters as “super-phenotypes” in QTL mapping. Clustering of correlated genes has been done in many studies to find coregulated genes (Eisen et al. 1998; Golub et al. 1999; Yvert et al. 2003), in most cases, the correlations between genes were useful for identification of common pathways that were defective in diseased cells, or in the case of tumor samples, for classification purposes. In our study, the coordinated expression should minimize the number of genome scans that need to be performed, since the genes can be analyzed as groups rather than as singletons.

Observations in experimental organisms including plants, yeast, fish, and mice (Cavalieri et al. 2000; Jansen and Nap 2001; Brem et al. 2002; Oleksiak et al. 2002; Yvert et al. 2003) reveal that levels of gene expression, like the genes themselves, show abundant natural variation. This variation is viewed here as an expression phenotype which is itself under genetic control. By combining the power of microarray and classical genetics, we expect to identify the determinants for natural variation in human gene expression.

ACKNOWLEDGMENTS

We are grateful to Warren J. Ewens for advice throughout this project. This work was supported by grants from

the National Institutes of Health (HG-02386, HG-01880) and from the W.W. Smith Endowed Chair (to V.G.C.).

REFERENCES

- Brem R.B., Yvert G., Clinton R., and Kruglyak L. 2002. Genetic dissection of transcriptional regulation in budding yeast. *Science* **296**: 752.
- Cavalieri D., Townsend J.P., and Hartl D.L. 2000. Manifold anomalies in gene expression in a vineyard isolate of *Saccharomyces cerevisiae* revealed by DNA microarray analysis. *Proc. Natl. Acad. Sci.* **97**: 12369.
- Cheung V.G., Conlin L.K., Weber T.M., Arcaro M., Jen K.-Y., Morley M., and Spielman R.S. 2003. Natural variation in human gene expression assessed in lymphoblastoid cells. *Nat. Genet.* **33**: 422.
- Cowles C.R., Hirschhorn J.N., Altshuler D., and Lander E.S. 2002. Detection of regulatory variation in mouse genes. *Nat. Genet.* **32**: 432.
- Eisen M.B., Spellman P.T., Brown P.O., and Botstein D. 1998. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci.* **95**: 14863.
- Enard W., Khaitovich P., Klose J., Zollner S., Heissig F., Givalisco P., Nieselt-Struwe K., Muchmore E., Varki A., Ravid R., Doxiadis G.M., Bontrop R.E., and Paabo S. 2002. Intra- and interspecific variation in primate gene expression patterns. *Science* **296**: 340.
- Ewens W.J., Spielman R.S., and Harris H. 1981. Estimation of genetic variation at the DNA level from restriction endonuclease data. *Proc. Natl. Acad. Sci.* **78**: 3748.
- Falconer D.S. and MacKay T.F.C. 1996. *Introduction to quantitative genetics*, 4th edition, Longman, London.
- Golub T.R., Slonim D.K., Tamayo P., Huard C., Gaasenbeek M., Mesirov J.P., Coller H., Loh M.L., Downing J.R., Caligiuri M.A., Bloomfield C.D., and Lander E.S. 1999. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science* **286**: 531.
- Harris H. 1966. Enzyme polymorphisms in man. *Proc. R. Soc. Lond. B Biol. Sci.* **164**: 298.
- . 1969. Enzyme and protein polymorphism in human populations. *Br. Med. Bull.* **25**: 5.
- Jansen R.C. and Nap J.P. 2001. Genetical genomics: The added value from segregation. *Trends Genet.* **17**: 388.
- Jeffreys A.J. 1979. DNA sequence variants in the G gamma-, A gamma-, delta- and beta-globin genes of man. *Cell* **18**: 1.
- Lo H.S., Wang Z., Hu Y., Yang H.H., Gere S., Buetow K.H., and Lee M.P. 2003. Allelic variation in gene expression is common in the human genome. *Genome Res.* **13**: 1855.
- Oleksiak M.F., Churchill G.A., and Crawford D.L. 2002. Variation in gene expression within and among natural populations. *Nat. Genet.* **32**: 261.
- Schadt E.E., Monks S.A., Drake T.A., Lusk A.J., Che N., Colinao V., Ruff T.G., Milligan S.B., Lamb J.R., Cavet G., Linsley P.S., Mao M., Stoughton R.B., and Friend S.H. 2003. Genetics of gene expression surveyed in maize, mouse and man. *Nature* **422**: 297.
- Steinmetz L.M., Sinha H., Richards D.R., Spiegelman J.I., Oefner P.J., McCusker J.H., and Davis R.W. 2002. Dissecting the architecture of a quantitative trait locus in yeast. *Nature* **416**: 326.
- Townsend J.P., Cavalieri D., and Hartl D.L. 2003. Population genetic variation in genome-wide gene expression. *Mol. Biol. Evol.* **20**: 955.
- Whitney A.R., Diehn M., Popper S.J., Alizadeh A.A., Boldrick J.C., Relman D.A., and Brown P.O. 2003. Individuality and variation in gene expression patterns in human blood. *Proc. Natl. Acad. Sci.* **100**: 1896.
- Yan H., Yuan W., Velculescu V.E., Vogelstein B., and Kinzler K.W. 2002. Allelic variation in human gene expression. *Science* **297**: 1143.
- Yvert G., Brem R.B., Whittle J., Akey J.M., Foss E., Smith E.N., Mackelprang R., and Kruglyak L. 2003. *Trans-acting regulatory variation in Saccharomyces cerevisiae and the role of transcription factors.* *Nat. Genet.* **3**: 3.

