

The genetics of variation in gene expression

Vivian G. Cheung^{1–3} & Richard S. Spielman²

doi:10.1038/ng1036

The genetic basis of variation in gene expression lends itself to investigation by microarrays. For genetic analysis, we view the expression level of a gene as a quantitative or ‘complex’ trait, analogous to an individual’s height or cholesterol level, and, therefore, as an inherited phenotype. Several genetic analyses of ‘gene expression phenotypes’ have been carried out in experimental organisms, and initial steps have been taken toward similar studies in humans—although these present challenging technical and statistical problems. Further advances in the genetic analysis of variation in gene expression will contribute to our understanding of transcriptional regulation and will provide models for studying other quantitative and complex traits.

In human studies, microarrays have been used to examine both variation in gene expression and variation in DNA sequence, but typically not in the same study. These two approaches can be combined to address new questions in genetics (Fig. 1). Microarrays provide measurements of many genes for large numbers of individuals, so the results can be used to identify those genes whose expression is most variable in the population. The expression level of a highly variable gene in an individual is considered as the ‘phenotype’, which is possibly influenced by genetic determinants. Genetic analysis can therefore be used to map and to identify the genes and/or regulatory regions that control expression phenotypes. In this review, we describe several recent studies that have used microarrays to obtain data on gene expression phenotypes, on genotypes, or on both. Although most of these studies were carried out in experimental organisms^{1–7}, they illustrate how microarrays also can be used to characterize and to map complex phenotypes in humans.

Lessons from model organisms

Suppose that the expression level of gene *X* is the phenotype of interest. The genetic determinants of variation in this phenotype might map near gene *X*, but it is also possible that they are not in gene *X* or even closely linked to it. Recent genetic analyses of gene expression in yeast provide examples of these possibilities. Brem *et al.*¹ compared the expression profiles of two strains of the yeast *Saccharomyces cerevisiae* using cDNA microarrays that contained more than 6,000 open reading frames from the yeast genome. They found that 1,528 genes were expressed differentially ($P < 0.005$) between the two strains instead of the 23 genes expected by chance. The expression levels of these genes were tested further in 40 haploid segregants from a cross between the two parental strains. After confirming earlier findings by Cavalieri *et al.*² that established expression phenotypes as highly heritable traits, Brem *et al.*¹ used a genome-wide genetic linkage approach to map the determinants of variation in gene expression. By testing for linkage with 3,312 markers in the yeast genome, they found that only 308 genes (20%) of the 1,528 showed linkage to

one or more loci. A simulation experiment indicated that if a single locus controlled expression variation, then many more genes (97%) would have shown linkage. These results suggest that the expression of most genes is affected by more than one locus. Thus, it seems that the control of gene expression is highly complex even in a relatively simple organism such as yeast.

A detailed study of a single phenotype in yeast further illustrates the complexity of mapping quantitative traits. Steinmetz *et al.*³ looked for genes (quantitative trait loci; QTLs) whose allelic variants affect the ability of yeast to grow at high temperatures. A strain with the high-temperature phenotype (Htg⁺) was crossed with one lacking this phenotype, and 19 of the Htg⁺ segregants were selected. These segregants and the parental strains were genotyped on Affymetrix oligonucleotide arrays containing more than 3,000 markers that spanned the yeast genome. Fine mapping showed that a QTL region detected on chromosome XIV consists of three genes, *MKT1*, *END3* and *RHO2*, that greatly influence the ability of yeast to grow at high temperature. The resolution of the QTL into three distinct loci was unexpected, but it suggests that similar situations might complicate QTL studies in other species, especially in non-experimental organisms such as humans.

In addition to studies in yeast, studies in *Drosophila* and mice have also demonstrated genotypic contributions to variation in the expression of genes^{4,5}. Jin *et al.*⁴ showed that significant differences in expression levels of genes could be attributed to sex and to genotypic differences among strains of *Drosophila*. Sandberg *et al.*⁵ compared the brain expression profiles of two mouse strains and found regional and strain-specific differences. Some of the genes that showed differences in expression mapped to chromosomal regions that have been linked in other studies to complex quantitative phenotypes, such as alcohol drinking preference and susceptibility to seizures induced by methyl β -carboline-3-carboxylate.

Extension to human genetics

The studies in model organisms illustrate the genome-wide view of gene expression levels as heritable phenotypes. They also sug-

¹Department of Pediatrics and ²Department of Genetics, University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania 19104, USA.

³The Children’s Hospital of Philadelphia, Philadelphia, Pennsylvania 19104, USA. Correspondence should be addressed to V.G.C. (e-mail: vcheung@mail.med.upenn.edu)

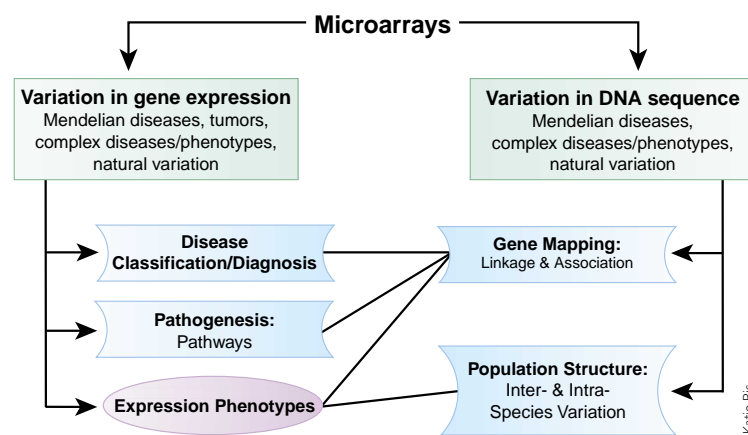


Fig. 1 Microarray analysis of genetic variation. Current studies focus on variation in either gene expression or DNA sequence. Microarray studies that merge the two types of variation will yield better understanding of the genetic basis of complex human traits and diseases.

might lead us to identify sets of genes that play a role in complex human diseases such as diabetes and hypertension. Of course, it is likely that only a small part of variation in expression phenotypes is due to heterozygosity for genes involved in recessive diseases. Therefore, instead of starting with heterozygous carriers for known mutations, we would like to identify highly variable expression phenotypes in the general population and then find the genetic determinants by linkage mapping and other genetic analyses.

gest that a similar approach can be used to identify gene expression phenotypes in humans and to map their determinants. Such studies will increase our understanding of the genetics of complex traits and diseases.

Microarrays are used in various ways to identify DNA variants and to analyze gene expression patterns in human studies (Fig. 1). It is already possible to couple the growing body of information about germline variation with high-throughput methods for assaying differences in gene expression between individuals. The results from integrating data on sequence variation and on expression variation will lead to further developments in human genetics. We comment first on some implications for carrier identification, and then more speculatively on the genetic analysis of mRNA levels as a model for analyzing complex phenotypes and diseases.

Heterozygous carriers of disease. The parents of an individual with a rare recessive genetic disease are unaffected, but they are obligate heterozygotes. They therefore provide an opportunity to look for a possible subtle expression phenotype resulting from germline differences. Identification of such a phenotype would allow carrier detection and might provide basic information about the control of gene expression.

In a recent study, we compared the gene expression levels of about 3,000 genes in the lymphoblastoid cells of carriers of ataxia telangiectasia, a typical autosomal recessive disease, with those of normal controls⁸. Whereas individuals affected with ataxia telangiectasia are rare, heterozygous carriers of the disease are fairly common (roughly 1 in 100) and cannot be detected reliably by physical examination or existing medical tests. The possibility of some phenotype in heterozygotes is supported by several studies, including one in which mice engineered to carry one mutated human allele for ataxia telangiectasia developed significantly more solid tumors than did wildtype mice^{9,10}.

Our study identified 71 genes whose expression was significantly different ($P < 0.01$) between carriers of ataxia telangiectasia and normal controls. But in combination the expression levels of just four genes (*LIM*, *CDKN2D*, *TFRC* and *ARF6*) allowed the correct classification of 19 out of 20 individuals tested. We suspect that carriers of other diseases can be identified by similar methods. Our result also illustrates that even for a recessive disease, heterozygous carriers may have a distinctive phenotype.

Most individuals are likely to be heterozygous for several recessive diseases, and some combinations of recessive mutations might confer susceptibility to common diseases. For example, a person heterozygous for mutations in several genes in the DNA repair pathways may have an increased risk for several cancers. Thus, the expression phenotypes of recessive diseases

Genetics of natural variation in gene expression. Studies of genome-wide expression in humans have not yet considered inherited variation, except for assessing the effects of single-gene mutations. How might the genetics of expression phenotypes actually be approached? We can imagine an extension of the study of the ataxia telangiectasia carriers⁸ in a framework similar to that of Brem *et al.*¹. This involves broadening the viewpoint of the ataxia telangiectasia studies in two ways. First, expression phenotypes would be studied in a random sample of subjects (and ideally also in family members), not only in patients or carriers of a known genetic disease. Such a genome-wide approach would provide an estimate of natural variation in gene expression in humans as a starting point. Second, the goal would be to determine how variation in expression phenotypes can be accounted for by considering DNA sequence variation anywhere in the genome, not only in one 'disease' gene.

Classically, genome scans are done to find genes for complex phenotypes such as cholesterol level, body mass index or diabetes. In other words, the phenotypes are measured directly in individuals, and, ideally, whole families are studied. Similarly, when the expression levels of genes are defined as phenotypes, genetic analysis can be done to map, identify and characterize the genetic determinants that are responsible. Such studies will advance our knowledge of the genetic basis of gene expression and also provide simpler models for the situation in complex traits and diseases. As illustrated by the findings for the high-temperature growth phenotype in yeast, current methods and expectations will probably need to be refined as they are used to examine complex traits in humans³. However, gene expression phenotypes in humans are probably simpler than complex traits and disease phenotypes, and they may therefore represent an intermediate level of complexity between the phenotype of mendelian diseases and that of complex genetic disease. Thus, genetic analysis of gene expression will also advance the tools and methods necessary to study other complex phenotypes in humans.

Variation in DNA sequence

Dissecting the genetics of variation in gene expression requires both the phenotyping described above and the mapping of genetic determinants. Below we describe briefly how variation in DNA sequence makes the genetic mapping possible.

Unlike most mendelian diseases, complex diseases are common, and the genes that are responsible for them are likely to be polymorphic; that is, to show high-frequency allelic variation. This idea has been called the 'common disease, common variant' hypothesis^{11,12}. Of course, the actual variants that contribute to most complex diseases are not known, but the most common

form of variation in DNA sequence is the single-nucleotide polymorphism (SNP); SNPs are present at a frequency of about 1 in 1,000 nucleotides in humans¹³. Thus, an essential step in mapping complex traits is to determine which SNPs, among the large number in the genome, influence disease risk.

SNP variants that are closely linked do not occur independently of each other. Instead, there is marked non-random association, known as 'linkage disequilibrium' (LD), between neighboring SNPs. These groups of correlated SNPs typically span about 40 kb of DNA but can sometimes extend over 1,000 kb (refs 14,15). The resulting conserved sequences of DNA are termed 'haplotype blocks' and contain SNP sites in very strong LD with each other¹⁶. Several studies^{14–17} are under way to catalogue all of the haplotype blocks in the human genome. The resulting haplotype maps or 'HapMaps'¹⁸ of LD will provide new materials for many population genetic and disease-mapping studies.

The great abundance of SNPs, and the finding that they occur as clusters of sites in strong LD, have focused attention on the use of association methods, rather than classical linkage mapping, to locate disease genes^{19,20}. The premise is that the DNA variants that contribute to complex diseases will often be embedded in haplotype blocks, thereby forming 'disease haplotypes.' As a result of the LD, all of the SNPs in the block will show association with disease. Thus, to reveal the location of the disease gene, it may be possible to type one SNP per block and use it as a proxy for the others. If the haplotype blocks in the region are small, for example less than 50 kb (as is common), the finding of disease association implies the presence of a predisposing allele at a very nearby gene. By contrast, classical linkage methods applied to complex diseases can rarely narrow the candidate region to less than a few megabases. This potential power of association studies for locating genes has led to several associated-based statistical methods^{21–24}.

Classical linkage mapping relies on recombinants in families, and the frequency of recombination in humans permits a set of 350 markers (at roughly 10 Mb intervals) to extract most of the available information for linkage. By contrast, association-based tests require markers that can capture LD between blocks and disease genes, and so the intervals must be much smaller, perhaps 40–100 kb. This density can be achieved with SNPs as markers but requires the typing of 30,000 to 100,000 SNPs per individual. Thus, the shift from linkage to association methods demands techniques for typing many more polymorphic markers.

Various high-throughput genotyping methods^{25–32}, including several that use microarrays, are being developed to meet this need. The next generation of SNP typing microarrays will need to provide genome-wide or region-specific SNP typing for association, presumably by using SNPs that capture the variation in haplotype blocks. As the SNP collections and the microarrays for genotyping improve further, association studies based on haplotype blocks may become the methods of choice for the genetic analysis of complex human diseases.

Although SNP-based genome-wide association studies are conceptually straightforward, they still face technical and statistical obstacles. First, reliable and efficient techniques are needed to type the large number of markers required. Whereas expression-profiling technologies have improved greatly in the past several years, SNP genotyping technologies have lagged behind. Compared with mRNA, genomic DNA is highly complex and repetitive, so it is more difficult to analyze markers in genomic DNA by hybridization-based array technologies. Nevertheless, microarrays are already available to genotype a few thousand SNP markers that span the human genome (see Affymetrix: <http://www.affymetrix.com>; and Illumina: <http://www.illumina.com>).

Second, although it is not practical to type all of the known SNP markers, it is also not clear which are the most informative markers or what density will be needed for association studies. The HapMap project is expected to provide some of this information¹⁸. Third, the genotype data must be analyzed with methods that can detect small contributions from several genes, while also dealing with the errors that arise when so many markers are tested. Last, existing molecular and statistical methods usually do not take into account gene–gene or gene–environment interactions, which are likely to have a key role in susceptibility to diseases. As new methods that surmount these obstacles are invented, our ability to understand the genetics of complex human phenotypes and diseases will improve greatly.

The Chipping Forecast in human genetics and genomics

The goal of human genetics is to understand how the phenotypic variation seen in normal and clinical contexts is related to underlying sequence variation in the genome. The development of microarrays has made it possible to expand the phenotype to include another form of variation: genome-wide gene expression levels. At the same time, the expanding catalogue of SNPs has provided new information about the underlying variation in the genome.

Figure 1 summarizes our view of various aspects of phenotypic and sequence variation. Our discussion of work in this area has focused on understanding how sequence and expression variation determines complex traits and diseases in individual organisms. However, parallel investigation of expression variation among populations⁶ and species⁷ have also been carried out. Of course, it is unlikely that complex human traits can be understood by studying the expression of transcripts alone; ultimately, it will be necessary to extend the studies to include variation in proteins. At present, however, microarrays have made it possible to analyze phenotype at the transcript level. The analysis of DNA sequence variants that contribute to expression and other phenotypic differences promises to provide the field of genetics with new ways of understanding human variation.

Acknowledgments

We thank H. Kazazian for comments on the manuscript, and A. Downend for assistance in manuscript preparation. This work is supported by grants from the US National Institutes of Health.

- Brem, R.B., Yvert, G., Clinton, R. & Kruglyak, L. Genetic dissection of transcriptional regulation in budding yeast. *Science* **296**, 752–755 (2002).
- Cavallieri, D., Townsend, J.P. & Hartl, D.L. Manifold anomalies in gene expression in a vineyard isolate of *Saccharomyces cerevisiae* revealed by DNA microarray analysis. *Proc. Natl Acad. Sci. USA* **97**, 12369–12374 (2000).
- Steinmetz, L.M. et al. Dissecting the architecture of a quantitative trait locus in yeast. *Nature* **416**, 326–330 (2002).
- Jin, W. et al. The contributions of sex, genotype and age to transcriptional variance in *Drosophila melanogaster*. *Nature Genet.* **29**, 389–395 (2001).
- Sandberg, R. et al. Regional and strain-specific gene expression mapping in the adult mouse brain. *Proc. Natl Acad. Sci. USA* **97**, 11038–11043 (2000).
- Oleksiak, M.F., Churchill, G.A. & Crawford, D.L. Variation in gene expression within and among natural populations. *Nature Genet.* **32**, 261–266 (2002).
- Enard, W. et al. Intra- and interspecific variation in primate gene expression patterns. *Science* **296**, 340–343 (2002).
- Watts, J.A. et al. Gene expression phenotype in heterozygous carriers of ataxia telangiectasia. *Am. J. Hum. Genet.* **71**, 791–800 (2002).
- Concannon, P. ATM heterozygosity and cancer risk. *Nature Genet.* **32**, 89–90 (2002).
- Spring, K. et al. Mice heterozygous for mutation in *Atm*, the gene involved in ataxia-telangiectasia, have heightened susceptibility to cancer. *Nature Genet.* **32**, 185–190 (2002).
- Cargill, M. et al. Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nature Genet.* **22**, 231–238 (1999).
- Halushka, M.K. et al. Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. *Nature Genet.* **22**, 239–247 (1999).
- Kruglyak, L. & Nickerson, D. Variation is the spice of life. *Nature Genet.* **27**, 234–236 (2001).
- Gabriel, S.B. et al. The structure of haplotype blocks in the human genome. *Science* **296**, 2225–2229 (2002).
- Dawson, E. et al. First-generation linkage disequilibrium map of human chromosome 22. *Nature* **418**, 544–548 (2002).

16. Daly, M. J., Rioux, J.D., Schaffner, S. F., Hudson, T.J. & Lander, E. S. High-resolution haplotype structure in the human genome. *Nature Genet.* **29**, 229–232 (2001).
17. Patil, N. *et al.* Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* **294**, 1719–1723 (2001).
18. Couzin, J. New mapping project splits the community. *Science* **296**, 1391–1392 (2002).
19. Ewens, W.J. & Spielman, R.S. Locating genes by linkage and association. *Theor. Pop. Biol.* **60**, 135–139 (2001).
20. Jorde, L.B. Linkage disequilibrium and the search for complex disease genes. *Genome Res.* **10**, 1435–1444 (2000).
21. Spielman, R.S., McGinnis, R.E. & Ewens, W.J. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus. *Am. J. Hum. Genet.* **52**, 506–516 (1993).
22. Thomson, G. Mapping disease genes: family-based association studies. *Am. J. Hum. Genet.* **57**, 487–498 (1995).
23. Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics* **55**, 997–1004 (1999).
24. Pritchard, J.K., Stephens, M., Rosenberg, N.A. & Donnelly, P. Association mapping in structured populations. *Am. J. Hum. Genet.* **67**, 170–181 (2000).
25. Chen, X. & Kwok, P.Y. Template-directed dye-terminator incorporation (TDI) assay: a homogeneous DNA diagnostic method based on fluorescence resonance energy-transfer. *Nucleic Acid Res.* **25**, 347–353 (1997).
26. Wang, D.G. *et al.* Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* **280**, 1077–1082 (1998).
27. Chen, X., Levine, L. & Kwok, P.Y. Fluorescence polarization in homogeneous nucleic acid analysis. *Genome Res.* **9**, 492–498 (1999).
28. Howell, W.M., Jobs, M., Gyllensten, U. & Brookes, A.J. Dynamic allele-specific hybridization. A new method for scoring single nucleotide polymorphisms. *Nat. Biotechnol.* **17**, 87–88 (1999).
29. Griffin, T.J., Hall, J.G., Prudent, J.R. & Smith, L.M. Direct genetic analysis by matrix-assisted laser desorption/ionization mass spectrometry. *Proc. Natl Acad. Sci. USA* **96**, 6301–6306 (1999).
30. Lindblad-Toh, K. *et al.* Large-scale discovery and genotyping of single-nucleotide polymorphisms in the mouse. *Nature Genet.* **24**, 381–386 (2000).
31. Fan, J.B. *et al.* Parallel genotyping of human SNPs using generic high-density oligonucleotide tag arrays. *Genome Res.* **10**, 853–860 (2000).
32. Oliphant, A., Barker, D.L., Stuelpnagel, J.R. & Chee, M.S. BeadArray Technology: enabling an accurate, cost-effective approach to high-throughput genotyping. *Biotechniques* **32**, S56–S61 (2002).