The genetics of regulatory variation in the human genome

Barbara E. Stranger[*] *and Emmanouil T. Dermitzakis*[*]

The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK
* *Correspondence to*: Tel: +44 (0)1223 834244; Fax: +44 (0)1223 494919; E-mail: bes@sanger.ac.uk; md4@sanger.ac.uk

Date received (in revised form): 17th January 2005

Abstract

The regulation of gene expression plays an important role in complex phenotypes, including disease in humans. For some genes, the genetic mechanisms influencing gene expression are well elucidated; however, it is unclear how applicable these results are to gene expression on a genome-wide level. Studies in model organisms and humans have clearly documented gene expression variation among individuals and shown that a significant proportion of this variation has a genetic basis. Recent studies combine microarray surveys of gene expression for thousands of genes with dense marker maps, and are beginning to identify regions in the human genome that have functional effects on gene expression. This paper reviews recent developments and methodologies in this field, and discusses implications and future directions of this research in the context of understanding the influence of human genomic variation on the regulation of gene expression.

Keywords: *gene expression, regulatory variation, association mapping, linkage mapping*

Introduction

Gene expression in eukaryotic organisms is a complex trait influenced by genetic, environmental and epigenetic factors. Although there are many mechanisms of gene expression regulation (Figure 1), including chromatin condensation, alternative splicing, DNA methylation, transcription initiation, mRNA stability, translational controls, post-translational controls and protein degradation, transcription initiation is the most common point of control, with elements located both *cis* (proximal to the gene) and *trans* to the coding locus interacting to control initiation of transcription.[1] The simplest model of transcriptional regulation involves the binding of transcription factors (TFs) in a sequence-specific manner to TF binding sites, short stretches of DNA usually near a gene, thereby altering rates of transcription. Both the identity of TFs present and their binding affinities play an important role in transcription initiation. Genetic mutations that alter either the nucleotide sequence of a TF binding site, the nucleotide sequence of the transcript (eg affecting stability) or the transcription factor amino acid sequence are just some examples of types of mutations that can have substantial effects on mRNA transcript levels.

The control and maintenance of appropriate levels of transcription for each of the genes expressed in a given cell type are vital cellular processes. Many human disorders are caused by molecular changes that have an impact at the level of gene expression.[2] Often, alterations in gene expression are extreme and involve either many-fold overexpression (eg Burkitt's lymphoma[3]) or partial/complete loss of expression (eg alpha thalassaemia[4]). Typically, these are monogenic or rare disorders where the effect is strong and the mutations are present at low frequency. In other instances, however, such as Type 1 diabetes,[5] subtle changes in expression can have small phenotypic consequences that are conditional on the genetic background of the individuals. For example, Eaves *et al.*[5] and Karp *et al.*[6] used differential gene expression patterns to map susceptibility genes. It is hypothesised that complex disorders are likely to be associated with gene expression variation, since susceptibility alleles have mostly quantitative rather than qualitative differences between individuals.[7] The use of gene expression variation as an endophenotype[8] between nucleotide polymorphism and disease susceptibility can prove useful in identifying the underlying genetic basis of complex disorders and designing appropriate models to test it.[9] Gene expression levels can thus be used as genetic markers which can help in linking nucleotide variation with a disease phenotype. A comprehensive study of segregating gene expression variation will provide an important starting point for the utilisation of gene expression as an intermediate level of phenotypic attributes between a nucleotide polymorphism and a complex disorder.

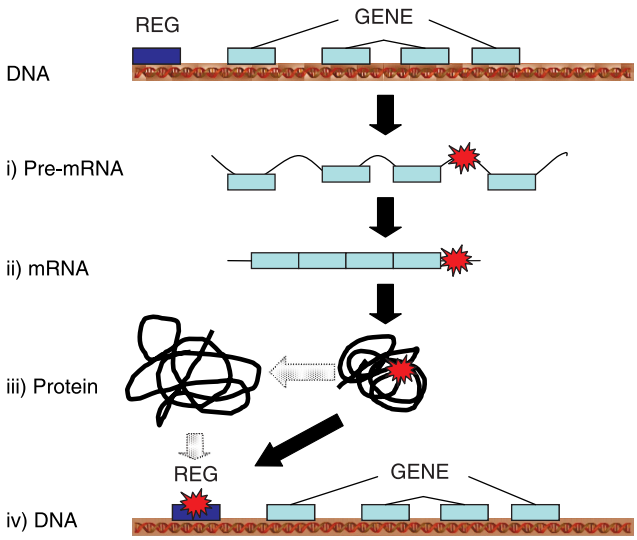

Figure 1. The different levels at which regulatory mutations can have an effect. 'Reg' indicates a regulatary region. Red stars indicate the position of a putative mutation. There can be mutations that: i) have an effect on the stability of the pre-mRNA; ii) have an effect on the stability of the mRNA or the rate of its translation (eg micro-RNA pairing); iii) have an effect on the structure of the protein and influence interactions with DNA or other proteins; iv) have an effect directly at the DNA level and abolish or modify sites of protein binding or affect DNA conformation.

Challenges and approaches for identifying regulatory regions

One of the major challenges ahead is to identify the DNA sequences that carry the *cis* signals for regulation of the spatial and temporal expression of all of the genes in the human genome.[10–12] This is a difficult task because little is known about the characteristics of the structure of regulatory regions. One of the main limitations is that little is known about the functional units that comprise a *cis* regulatory region, and current knowledge is restricted to transcription factor binding sites that may be only one part of a regulatory region organisation. Transcription factor binding sites are found throughout the genome but the organisation and abundance requirements of such sites within regulatory elements is not yet clear.[13–15] Given the redundancy and short length of binding sites, it is expected that specificity is achieved through a higher order code of organisation. Therefore, this code needs to be deciphered to be able reliably to identify them in the genome. Current knowledge of regulatory sequence organisation lacks uniform and easily interpretable sequence characteristics, such as the open reading frame (ORF) and the splicing signals in coding sequences. In other words, we are missing information about the code that the cell recognises in order to identify *cis* regulatory regions in the raw sequence data. Moreover, and the dimensionality of this code is currently unknown.

Experimental identification of variation in regulatory sequences currently relies on extensive use of model systems such as yeast, mouse and cultured human cell lines. Studies to detect differences in gene expression caused by known promoter polymorphisms have generally used reporter gene assays with allele-specific promoter constructs;[16,17] however, these experiments have several limitations that make them impractical for whole-genome analyses. First, they require knowledge of the promoter region and candidate functional variants to test; at present, there are experimentally validated promoters for less than 10 per cent of human genes,[18] and the properties of long-distance regulators remain unknown and untested. Secondly, these methods are indirect and performed either *in vitro* or in cellular conditions (tissue, developmental stage, environmental stimuli) distant from the tissue context.[19,20] Therefore, any inference about the potential regulatory role of a sequence relies on the assumption that the experimental system is similar to the *in vivo* conditions. Thirdly, experiments that target candidate regulatory regions based on computational predictions are performed without previous knowledge of the target gene — for example, distant enhancers and suppressors — and therefore the significance of the identified sequence for genome function cannot be fully revealed. Finally, experiments that make use of the proper genomic and cellular context are intensive and slow when intended for large-scale analysis,[21,22] so for practical reasons they cannot be applied to all of the genes of the human genome.

Detailed functional experiments to elucidate mechanisms of gene regulation have typically been carried out at the level of individual genes or sets of genes. Although the mechanism of gene expression is well-documented for some genes in great detail (HOX genes and the genes encoding β-globin, α-globin etc), it is unknown how transferable these results are to the whole genome. With the development of microarray technologies, it is now feasible to quantify the transcript abundance of thousands of genes simultaneously and efficiently in a single experiment. These technologies have medical applications, for example in identifying genes that are differentially expressed in a disease state versus non-disease controls,[23] to classify disease into subtypes[24] and to examine differences in transcript profile among different tissues or organs.[25] More recently, researchers have begun to use these technologies to quantify naturally-occurring variation in gene expression for many genes among multiple 'non-diseased' individuals of a species.

The Genetic basis of Gene Expression Variation

Large-scale surveys of gene expression variation in humans can provide important baseline information about 'normal' naturally-occurring variation among individuals. These data can be used to assess the significance of variation observed in

experimental studies where groups of individuals (eg disease versus non-disease controls) are compared. In addition to being useful to the medical community, these studies will fundamentally increase our understanding of the causes of naturally-occurring gene expression variation. The total phenotypic variance among individuals (V_P) for any trait can be broken down into a component of variance due to genotype (V_G), a component due to environment (V_E) and a component due to different genotypes in different environments (V_{GE}), according to the following equation:

$$V_P = V_G + V_E + V_{GE}.$$

The genetic component of phenotypic expression variation reflects interindividual genetic differences that result in inter-individual expression differences.

Little is known about the genetic basis of natural variation in gene expression. There are questions of fundamental biological importance, including, but not limited to:

- How much variation in gene expression (mRNA transcript levels) exists among individuals of a natural population of a single species?
- How much of this variation has a genetic component?
- How common are genetic polymorphisms in regulatory elements in natural populations, and what are the magnitudes of effect of these variants on mRNA levels?
- Are there 'regulatory hotspots', regions of the genome that affect transcription patterns of multiple genes?
- Is there interaction among loci, such that co-expressed gene complexes are observed?

Recent work in model organisms and humans has begun to address these and other questions.

Studies in model organism systems have documented significant, naturally-occurring variation in gene expression among individuals, including yeast,[26,27] *Drosophila*[28,29] and mouse,[30–33] although additional studies have made similar observations in fish,[34,35] maize,[33] primates[36] and humans.[37–42] As it has become accepted that naturally-occurring variation in gene expression among individuals is a common phenomenon, focus has shifted toward trying to quantify the contribution of genetic factors to that variation and to locate the responsible genomic regions.

Yan *et al.*[42] were among the first to demonstrate a genetic component of expression variation in humans. For six of 13 loci surveyed in 96 Centre d'Etude du Polymorphisme Humain (CEPH)[43] individuals, they observed significant differences in mRNA transcript abundance for the two alleles of heterozygous individuals (allelic imbalance). Furthermore, when families of individuals exhibiting allelic expression differences were examined; one-third of them showed expression patterns consistent with underlying Mendelian inheritance of functional variants. Other recent studies of allelic imbalance in humans and mice provided similar evidence for a functional genetic influence separate from that attributable to imprinting.[18,30,31] In a large-scale microarray study, Cheung *et al.*[38] provided further evidence of familial aggregation of expression profiles. The authors surveyed genome-wide patterns of gene expression in immortalised lymphoblastoid cells of humans and identified a set of genes whose transcript level varied greatly among 35 unrelated CEPH individuals. To determine whether the variation was influenced by genetic differences segregating among individuals, mRNA transcript levels of the most variable genes were quantified in several samples of individuals of different degrees of genetic inter-relatedness, including a sample of 49 unrelated CEPH individuals (the 35 individuals mentioned above plus an additional 14), offspring from five CEPH families and ten pairs of monozygotic twins. The authors observed that genes exhibited less variability in transcript abundance in more closely related individuals, suggesting a heritable component of gene expression variation among individuals.

Some studies have gone a step further and used large-scale studies to estimate the percentage of genes that exhibit significant heritability. In a study of gene expression in lymphoblastoid cell lines of CEPH pedigrees, Schadt *et al.*[33] reported extensive differences among 56 individuals of four CEPH families in mRNA transcript levels, and through heritability analyses were able to estimate that approximately 29 per cent of these genes had a genetic component influencing these levels. Monks *et al.*[39] followed up this study with a massive survey of expression of 23,499 genes in 167 individuals of 15 CEPH families. Of the detected genes, 31 per cent exhibited significant heritability (false discovery rate 0.05), with a median heritability of 0.34.

The above studies in human and other species demonstrate gene expression is an abundantly variable phenotype with a genetic component; thus, gene expression — or mRNA transcript level among individuals — can be considered as a quantitative trait. In general, quantitative traits exhibit continuous phenotypic variation among individuals, and the genetic component of that variation is often due to contributions of more than one locus. By combining microarray quantification of gene expression among individuals with marker genotype data (eg single nucleotide polymorphisms; SNPs) for the same individuals, it has become possible to map the genomic regions containing factors responsible for natural variation in human gene expression by performing association analyses. In these analyses, first referred to as 'genetical genomics',[44] transcript abundance of each of thousands of genes is treated as a quantitative phenotype[9] that is under genetic control. Association analyses are used to map functional regulatory regions by associating genotype at an individual marker locus with the expression of each gene (Figure 2A and 2B). These methods differ from family-based linkage analysis that traces genotypes and phenotypes of related individuals, looking for polymorphisms that co-segregate with the phenotype (Figure 2C).

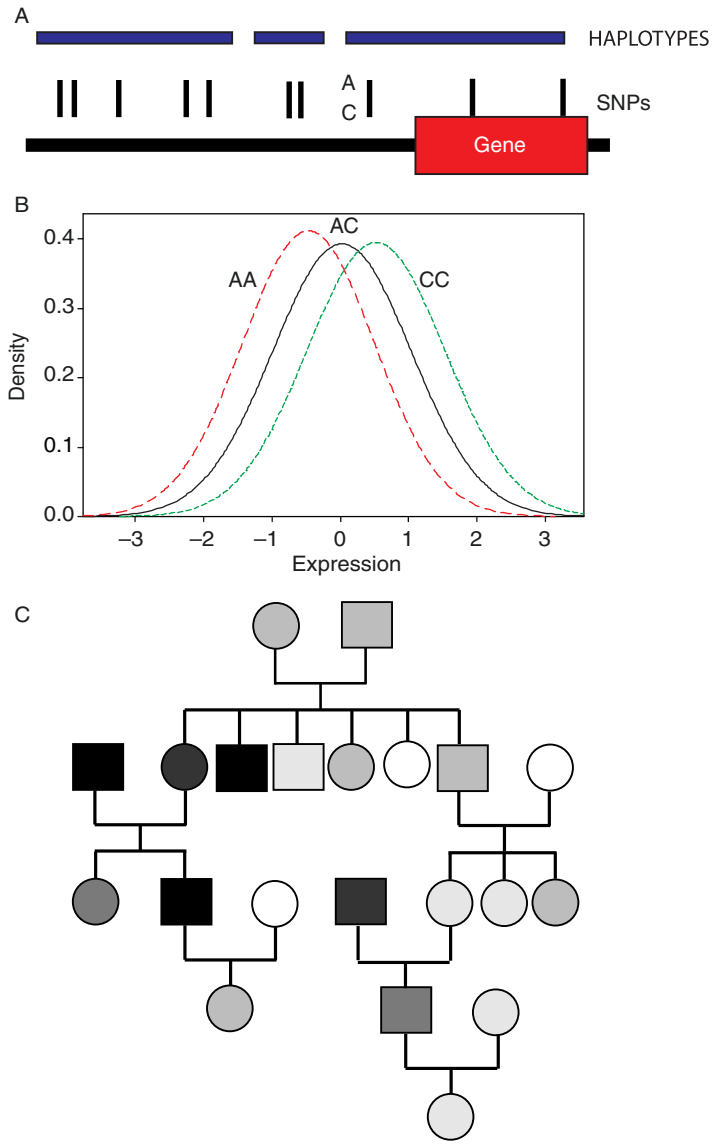

Figure 2. (A) Structure of a hypothetical gene and the haplotype organisation of single nucleotide polymorphisms (SNPs) in the region. Vertical lines represent the location of SNPs, with two nucleotides of a single SNP shown. Horizontal blue lines represent the arrangement of SNPs into haplotype blocks. (B) Relationship between a hypothetical phenotypic distribution and associated SNP genotypic frequencies in a population. In this example, the individuals with highest transcript abundance are almost exclusively genotype CC, and those with lowest transcript abundance are almost exclusively AA. The individuals with transcript abundance closest to the population mean are primarily AC. (C) A hypothetical family pedigree used to identify functional variants through family-based linkage analysis. The individuals are coloured in grey scale according to a two-locus genotype additive model, with black corresponding to AABB and white corresponding to aabb.

Progress through use of genome-wide association mapping

One of the advantages of a large-scale association approach is that it may be possible to identify functionally important regulatory variants without requiring any previous knowledge about specific *cis* or *trans* regulatory regions. Because these methods link expression variation of a particular gene to the genomic sequences directly or indirectly affecting it, there is a causal connection between phenotype and genotype. The identification of many significant associations between markers and individual gene expression phenotypes will allow researchers to address the issue of the relative proportion of *cis* or *trans* regulatory variation for each of many thousands

of genes. These studies also have the potential to identify sets of genes exhibiting correlated expression patterns and may identify clusters of regulators of multiple genes suggesting networks of co-regulated genes. Furthermore, because these methods look at the effects of naturally-occurring alleles, they may be able to identify regulatory regions that have subtle effects, as opposed to the large effects generally observed in knockout experiments.

Some recent work applying linkage and association analyses to expression variation in humans has led to the identification of regions of the genome influencing observed variation. A recent study using microarrays to measure gene expression variation[40] employed genome-wide linkage analysis to map regions influencing gene expression in immortalised B cells of 14 CEPH families (all parents and a mean of eight offspring per sibship). The authors performed linkage analyses for the expression phenotype of 3,554 genes (observed to be highly variable in a sample of 94 CEPH grandparents) and the genotypes of 2,519 SNP markers in the same individuals. They identified nearly 1,000 genes exhibiting significant linkage. Of the 142 genes with the strongest evidence for linkage, 110 (77.5 per cent) had only a *trans*-acting regulator, although 27 (19 per cent) had only a *cis*-acting transcriptional regulator (defined in this study as 5 megabases from the target gene). Interestingly, they identified regions that were hotspots of transcriptional regulation where there were clusters of SNPs with strong linkage to the expression phenotype of multiple genes. A quantitative transmission disequilibrium test performed on 17 of the 27 phenotypes displaying significant *cis* linkage identified 14 phenotypes exhibiting both significant linkage and association. A regression-based association analysis of the same 17 phenotypes in 94 CEPH grandparents confirmed significant association between the same 14 gene expression levels and an SNP located within or near the gene. Additional surveys in humans, mice and maize have confirmed that genetic variation located *cis* to the locus in question has functional effects on the transcript level of that gene.[18,30,33]

It is essential to point out that the distinction between *cis* and *trans* effects becomes less clear, and sometimes problematic, when one looks at genome-wide expression data. If one is taking a gene-centric view of the genome and is interested in the proportion of genes that have *cis* or *trans* regulatory variation and the relative contribution to genetic variance *per gene,* then it is appropriate to use this distinction because the view of the data remains gene-centric. If, however, one is interested in the overall contribution of genetic variation to gene expression variation as a whole-genome property, then the terms *cis* and *trans* become irrelevant, since all genetic variation of any nature (amino acid, transcript or *cis*-regulatory region; Figure 1) is mapped to unique locations in the genome. There is no such thing as a genetic variant *trans* to a whole genome because all genetic variation is encoded in the DNA.

Considerations

The massive amounts of data produced in microarray experiments require some significant statistical considerations. First, it is necessary to assess the quality of the measurements reliably and omit low-quality data from the primary analysis. Normalisation methods are then applied to the data to adjust for any sources of variability due to the experiment (different arrays, hybridisation efficiency differences, mRNA preparation differences etc) that may interfere with detecting those differences that reflect real biological variability. These methods are data transformations, and there are many procedures to choose from, some of which may be more relevant to certain microarray platforms and raw data distributions. The normalised microarray data and marker genotypes can then be subjected to association analyses, in which the genotype of the individuals is the primary classification variable and the response variable is the normalised transcript level of each gene (Figure 2B). Because these procedures test for association between each gene expression phenotype and many marker genotypes, the threshold for assessing significance must be adjusted to control for the massive amount of multiple testing inherent in testing each gene.

Can these methods lead to the identification of the individual nucleotides responsible for naturally-occurring variation in gene expression in humans? Regulatory variant identification in humans is complicated by the non-random association of alleles at different loci (linkage disequilibrium [LD]) in the human genome. In one of the available human cell line populations, the CEPH pedigrees, for example, on average the LD is high.[45] If LD is high among markers for a region showing a very strong association between genotype and expression phenotype, it can be difficult to pinpoint the causal functional variant, as multiple variants covering a large region might all exhibit the same strong association. In cases like these, it may be possible to narrow down the length of the responsible genomic region by generating a map with a high local marker density, effectively identifying markers that are not perfectly correlated with each other. Alternatively, fine-scale mapping may be facilitated by examining several populations that exhibit different patterns of LD. Finally, sequencing the region around a strongly associated marker may permit identification of the responsible regulatory variant that is in LD with the associated marker. At this stage, individual nucleotide variants linked to the marker are tested for association with the phenotype. Candidate regulatory regions (and variants) can be tested with experimental procedures to determine their potential to modulate gene expression.

In conclusion, the availability of genotyped (or nearly genotyped) human pedigrees (CEPH and other populations; Coriell's repositories; the International HapMap Consortium[45]), as well as more sensitive and less expensive microarray technologies for gene expression and genotyping, means that the time is right for carrying out large-scale genome-wide association

studies. This will contribute greatly to our understanding of the genetic basis of complex phenotypes in human populations, and may lead to novel diagnostics, preventative methods and therapeutics for human disease.

References

1. Wray, G.A. (2003), 'Transcriptional regulation and the evolution of development', *Int. J. Dev. Biol.* Vol. 47, pp. 675–684.
2. Watts, J.A., Morley, M., Burdick, J.T. *et al.* (2002), 'Gene expression phenotype in heterozygous carriers of ataxia telangiectasia', *Am. J. Hum. Genet.* Vol. 71, pp. 791–800.
3. Boxer, L.M. and Dang, C.V. (2001), 'Translocations involving c-myc and c-myc function', *Oncogene* Vol. 20, pp. 5595–5610.
4. Weatherall, D.J. (1998), 'Pathophysiology of thalassaemia', *Baillieres Clin. Haematol.* Vol. 11, pp. 127–146.
5. Eaves, I.A., Wicker, L.S., Ghandour, G. *et al.* (2002), 'Combining mouse congenic strains and microarray gene expression analyses to study a complex trait: The NOD model of type 1 diabetes', *Genome Res.* Vol. 12, pp. 232–243.
6. Karp, C.L., Grupe, A., Schadt, E. *et al.* (2000), 'Identification of complement factor 5 as a susceptibility locus for experimental allergic asthma', *Nat. Immunol.* Vol. 1, pp. 221–226.
7. Knight, J.C. (2005), 'Regulatory polymorphisms underlying complex disease traits', *J. Mol. Med.* Vol. 83, pp. 97–109.
8. Gottesman, I.I. and Gould, T.D. (2003), 'The endophenotype concept in psychiatry: Etymology and strategic intentions', *Am. J. Psychiatry* Vol. 160, pp. 636–645.
9. Cheung, V.G. and Spielman, R.S. (2002), 'The genetics of variation in gene expression', *Nat. Genet.* Vol. 32, pp. 522–525.
10. Collins, F.S., Green, E.D., Guttmacher, A.E. and Guyer, M.S. (2003), 'A vision for the future of genomics research', *Nature* Vol. 422, pp. 835–847.
11. Hardison, R.C. (2000), 'Conserved noncoding sequences are reliable guides to regulatory elements', *Trends Genet.* Vol. 16, pp. 369–372.
12. Hardison, R.C., Oeltjen, J. and Miller, W. (1997), 'Long human-mouse sequence alignments reveal novel regulatory elements: A reason to sequence the mouse genome', *Genome Res.* Vol. 7, pp. 959–966.
13. Dermitzakis, E.T., Bergman, C.M. and Clark, A.G. (2003), 'Tracing the evolutionary history of *Drosophila* regulatory regions with models that identify transcription factor binding sites', *Mol. Biol. Evol.* Vol. 20, pp. 703–714.
14. Dermitzakis, E.T. and Clark, A.G. (2002), 'Evolution of transcription factor binding sites in mammalian gene regulatory regions: Conservation and turnover', *Mol. Biol. Evol.* Vol. 19, pp. 1114–1121.
15. Wasserman, W.W., Palumbo, M., Thompson, W. *et al.* (2000), 'Human–mouse genome comparisons to locate regulatory sites', *Nat. Genet.* Vol. 26, pp. 225–228.
16. Buckland, P.R., Hoogendoorn, B., Guy, C.A. *et al.* (2004), 'A high proportion of polymorphisms in the promoters of brain expressed genes influences transcriptional activity', *Biochim. Biophys. Acta* Vol. 1690, pp. 238–249.
17. Hoogendoorn, B., Coleman, S.L., Guy, C.A. *et al.* (2004), 'Functional analysis of polymorphisms in the promoter regions of genes on 22q11', *Hum. Mutat.* Vol. 24, pp. 35–42.
18. Pastinen, T. and Hudson, T.J. (2004), 'Cis-acting regulatory variation in the human genome', *Science* Vol. 306, pp. 647–650.
19. Cui, C., Wani, M.A., Wight, D. *et al.* (1994), 'Reporter genes in transgenic mice', *Transgenic Res.* Vol. 3, pp. 182–194.
20. Trinklein, N.D., Aldred, S.J., Saldanha, A.J. and Myers, R.M. (2003), 'Identification and functional analysis of human transcriptional promoters', *Genome Res.* Vol. 13, pp. 308–312.
21. Loots, G.G., Locksley, R.M., Blankespoor, C.M. *et al.* (2000), 'Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons', *Science* Vol. 288, pp. 136–140.
22. Nobrega, M.A., Ovcharenko, I., Afzal, V. and Rubin, E.M. (2003), 'Scanning human gene deserts for long-range enhancers', *Science* Vol. 302, pp. 413.
23. Welsh, J.B., Zarrinkar, P.P., Sapinoso, L.M. *et al.* (2001), 'Analysis of gene expression profiles in normal and neoplastic ovarian tissue samples identifies candidate molecular markers of epithelial ovarian cancer', *Proc. Natl. Acad. Sci., USA* Vol. 98, pp. 1176–1181.
24. Martinez-Delgado, B., Melendez, B., Cuadros, M. *et al.* (2004), 'Expression profiling of T-cell lymphomas differentiates peripheral and lymphoblastic lymphomas and defines survival related genes', *Clin. Cancer Res.* Vol. 10, pp. 4971–4982.
25. Walker, J.R., Su, A.I., Self, D.W. *et al.* (2004), 'Applications of a rat multiple tissue gene expression data set', *Genome Res.* Vol. 14, pp. 742–749.
26. Brem, R.B., Yvert, G., Clinton, R. and Kruglyak, L. (2002), 'Genetic dissection of transcriptional regulation in budding yeast', *Science* Vol. 296, pp. 752–755.
27. Steinmetz, L.M., Sinha, H., Richards, D.R. *et al.* (2002), 'Dissecting the architecture of a quantitative trait locus in yeast', *Nature* Vol. 416, pp. 326–330.
28. Jin, W., Riley, R.M., Wolfinger, R.D. *et al.* (2001), 'The contributions of sex, genotype and age to transcriptional variance in *Drosophila melanogaster*', *Nat. Genet.* Vol. 29, pp. 389–395.
29. Wittkopp, P.J., Haerum, B.K. and Clark, A.G. (2004), 'Evolutionary changes in cis and trans gene regulation', *Nature* Vol. 430, pp. 85–88.
30. Cowles, C.R., Hirschhorn, J.N., Altshuler, D. and Lander, E.S. (2002), 'Detection of regulatory variation in mouse genes', *Nat. Genet.* Vol. 32, pp. 432–437.
31. Lo, H.S., Wang, Z., Hu, Y. *et al.* (2003), 'Allelic variation in gene expression is common in the human genome', *Genome Res.* Vol. 13, pp. 1855–1862.
32. Sandberg, R., Yasuda, R., Pankratz, D.G. *et al.* (2000), 'Regional and strain-specific gene expression mapping in the adult mouse brain', *Proc. Natl. Acad. Sci., USA* Vol. 97, pp. 11038–11043.
33. Schadt, E.E., Monks, S.A., Drake, T.A. *et al.* (2003), 'Genetics of gene expression surveyed in maize, mouse and man', *Nature* Vol. 422, pp. 297–302.
34. Oleksiak, M.F., Churchill, G.A. and Crawford, D.L. (2002), 'Variation in gene expression within and among natural populations', *Nat. Genet.* Vol. 32, pp. 261–266.
35. Oleksiak, M.F., Roach, J.L. and Crawford, D.L. (2005), 'Natural variation in cardiac metabolism and gene expression in *Fundulus heteroclitus*', *Nat. Genet.* Vol. 37, pp. 67–72.
36. Enard, W., Khaitovich, P., Klose, J. *et al.* (2002), 'Intra- and interspecific variation in primate gene expression patterns', *Science* Vol. 296, pp. 340–343.
37. Bray, N.J., Buckland, P.R., Owen, M.J. and O'Donovan, M.C. (2003), '*Cis*-acting variation in the expression of a high proportion of genes in human brain', *Hum. Genet.* Vol. 113, pp. 149–153.
38. Cheung, V.G., Conlin, L.K., Weber, T.M. *et al.* (2003), 'Natural variation in human gene expression assessed in lymphoblastoid cells', *Nat. Genet.* Vol. 33, pp. 422–425.
39. Monks, S.A., Leonardson, A., Zhu, H. *et al.* (2004), 'Genetic inheritance of gene expression in human cell lines', *Am. J. Hum. Genet.* Vol. 75, pp. 1094–1105.
40. Morley, M., Molony, C.M., Weber, T.M. *et al.* (2004), 'Genetic analysis of genome-wide variation in human gene expression', *Nature* Vol. 430, pp. 743–747.
41. Pastinen, T., Sladek, R., Gurd, S. *et al.* (2004), 'A survey of genetic and epigenetic variation affecting human gene expression', *Physiol. Genomics* Vol. 16, pp. 184–193.
42. Yan, H., Yuan, W., Velculescu, V.E. *et al.* (2002), 'Allelic variation in human gene expression', *Science* Vol. 297, pp. 1143.
43. Dausset, J., Cann, H., Cohen, D. *et al.* (1990), 'Centre d'étude du polymorphisme humain (CEPH): Collaborative genetic mapping of the human genome', *Genomics* Vol. 6, pp. 575–577.
44. Jansen, R.C. and Nap, J.P (2001), 'Genetical genemics: The added value from segregation', *Trends in Genetics*, Vol.17, pp. 388–391.
45. The International HapMap Consortium(2003), 'The International HapMap Project', *Nature* Vol. 426, pp. 789–796.