From DNA to RNA to disease and back: The 'central dogma' of regulatory disease variation

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

Date received (in revised form): 25th April 2006

Abstract

Much of the focus of human disease genetics is directed towards identifying nucleotide variants that contribute to disease phenotypes. This is a complex problem, often involving contributions from multiple loci and their interactions, as well as effects due to environmental factors. Although some diseases with a genetic basis are caused by nucleotide changes that alter an amino acid sequence, in other cases, disease risk is associated with altered gene regulation. This paper focuses on how studies of gene expression variation might complement disease studies and provide crucial links between genotype and phenotype.

Keywords: *gene expression, human disease, linkage mapping, association mapping*

Introduction

Understanding the causes of human disease is one of the most fundamental goals of modern medicine. Individuals differ with respect to disease susceptibility, disease progression and effectiveness of treatment. Identifying the factors contributing to these differences, and elucidating their interactions as they contribute to aspects of disease phenotype, is a precursor to improved prevention, detection and treatment of disease.

Much of the understanding of human disease derives from the study of those diseases that segregate in families in a Mendelian fashion, where the causative variants and the genes in which they reside have been identified through classical family linkage approaches[1] and through studies in large pedigrees and in isolated populations based on founder effects.[2] The vast majority of common diseases exhibit a more complex mode of inheritance, however, aggregating in families but rarely exhibiting Mendelian inheritance. Examples of diseases of this type include diabetes, obesity, schizophrenia and asthma. Understanding of these 'complex' diseases is improving, although still limited, but it is clear that genetic variation plays an important role in susceptibility to disease, for example in autoimmune and infectious diseases.[3] Most complex disease is thought to be caused by the combined effect of genetic variants at a few loci or multiple loci, each with only modest functional effects on susceptibility. Additional roles are played by environmental factors and their interactions. Mapping the genomic regions contributing to disease creates new directions for disease research and is an important step towards improving human health.

Approaches to identifying the genes involved in complex disease can be generally grouped into two categories: candidate gene studies and linkage/association studies. Candidate gene studies use knowledge about the biology of a disease, and about genes in physiologically or biochemically relevant pathways, and attempt to correlate genetic variation at these 'candidate genes' with disease phenotype. Unfortunately, for most diseases, this type of information is not available or complete enough to prove widely useful, and including them in some of the analyses is more likely to increase the noise than it is to reduce the search space for the disease. Genome-wide linkage studies and association analyses serve as alternative approaches to surveying the contribution to disease of genetic variants located anywhere in the genome. The genome-wide aspect means that these studies do not require any *a priori* hypothesis that a particular region is involved, although predictions about the potential effect of specific variants (eg non-synonymous single nucleotide polymorphisms [SNPs]) can be incorporated in the models. In this respect, these approaches are unbiased. Family-based linkage studies entail identifying genetic variants in families that co-segregate with disease more often than would be expected by chance. In general, linkage studies have achieved limited success in identifying genomic regions involved in complex disease, in part because they are underpowered to detect moderate genetic effects. Furthermore, because identification of a region or

regions associated with the disease or trait requires identifying those alleles that segregate with the disease in families, which in turn depends on recombination within the families, it can be difficult to narrow a region exhibiting significant linkage. An alternative methodology is to perform association analysis, which looks for correlation of genetic variants with aspects of phenotype, but does not require a pedigree structure for the individuals. Association analyses are more powerful for the detection of common disease alleles with small to modest effects,[4,5] and increasingly are being used successfully in studies to identify genes contributing to disease.[6,7]

Although many phenotypic differences among individuals are attributable to variants in coding DNA,[8] variants in non-coding DNA can have profound effects on phenotypes, including disease phenotypes. For example, regulatory variants affecting transcription initiation, splicing, RNA stability and translational efficiency are known to play roles in conditions including autoimmune disease (*CTLA4*[9]), malaria (*DARC*[10]), various cancers (*SMYD3*[11]) and other examples (shown in Table 1). In studies of such diseases, gene expression may serve as an intermediate phenotype between disease phenotype and genotype.[30,31] Gene expression, or mRNA levels, can be modulated by variants in coding or non-coding DNA (eg transcription factors or binding sites within promoters). Whole-genome association studies of gene expression (expression quantitative trait loci [eQTL] mapping) may generate hypotheses for disease susceptibility by identifying those regions of the genome with functional effects on gene expression. These might then serve as candidate regions for evaluating for association with disease phenotypes, as is discussed below.

Table 1. Genes with non-coding variants affecting disease.

Gene	Disease	Reference
CAT	Hypertension	12
CCR5	HIV-1 progression and transmission	13,14
CD209	Dengue fever	15
CRP	Cardiovascular disease risk	16
CTLA4	Autoimmune disease	9
DARC	Malaria susceptibility	10
DAT1	Attention deficit hyperactivity disorder	17
FCRL3	Rheumatoid arthritis and autoimmune disease	18
GSK3B	Parkinson's disease	19
IGFBP3	Breast cancer	20
INS	Type I diabetes	21
IPF1	Type II diabetes	22
MMP-1	Breast cancer progression	23
MMP-3	Coronary artery disease	24,25
NFKB1	Inflammatory bowel disease	26
RET	Hirschsprung's disease	27,28
SMYD3	Colorectal cancer, breast cancer, hepatocellular carcinoma	11
TNF	Malaria	29

Resources for genome-wide analysis

An efficient approach to the study of human disease benefits from the use of shared resources. For example, in order to perform genome-wide linkage or association analyses, suitable DNA markers are required. The human genome is estimated to harbour more than 10 million SNPs, present at >1 per cent frequency,[32] and these SNPs are located throughout the genome in regions of coding and non-coding DNA. Publicly available databases of SNP alleles, assays and genotypes are accessible online (eg dbSNP[33] and HapMap[34]). High-throughput genotyping platforms and reductions in genotyping costs now make whole-genome genotyping feasible for large numbers of samples. Gene expression can also be quantified in a high-throughput manner using commercially available microarrays, permitting the detection of small differences in expression levels among samples.

The establishment of cell lines creates resources that can be used by multiple research groups from around the world to survey various cellular phenotypes. With respect to the study of gene expression, it is desirable to establish cell lines from different tissues because gene expression is highly dependent on developmental and cellular context and, indeed, some diseases manifest their phenotypes only in specific tissues. In addition, the cell perturbations that accompany the establishment of cell lines suggest the study of gene expression in primary tissues, although, clearly, the choice of sample depends on the purpose, stage and feasibility of a study, the sample size required and its availability. Despite some shortcomings of cell lines as perfect proxies for the complete set of human tissues, data can be collected on a large scale with respect to sample size and reproducibility and can provide candidates for further study in other samples. Currently, there are relatively few data on gene expression across the diversity of healthy human tissues or across multiple individuals from different populations. These data from healthy individuals will provide important information on the range of naturally occurring gene expression variation and will serve as a baseline against which to compare disease-associated molecular phenotypes.

Statistical issues in genome-wide analysis

Although genome-wide association studies are thought to have more power than family-based linkage studies, they present strong challenges in the form of statistical interpretation. For example, a simple genome-wide association study may test hundreds of thousands of SNPs for association to a phenotype (or, more typically, multiple phenotypes), and more complicated models allowing for SNP–SNP interactions vastly increase the already large number of statistical tests. With such a large number of tests, the significance threshold must be adjusted to control for the number of false-positive associations. Although procedures for multiple test correction exist — for example, Bonferroni correction, false discovery rate[35,36] and permutations of phenotypes relative to genotypes[37] — it remains unclear which is the best method to apply in this context. It is also not trivial to infer the biological significance of an association from statistical significance, because allele frequencies, variance of the phenotype, density of markers and linkage disequilibrium (LD) can have a tremendous impact on the statistical significance inferred.

Human genetic variation is structured into haplotypes, such that alleles at nearby loci often show strong statistical association with one another. Because of this association, known as LD, a large region may contain multiple SNPs exhibiting a significant association with a given phenotype. Although this structure of human genetic variation facilitates association mapping, it can complicate subsequent fine-scale mapping to narrow the associated region and locate the causal variant, as discussed below. Another concern in association studies is the potential for false associations caused by population stratification,[38] so care must be taken to reduce these effects through appropriate experimental design and data analysis.[39,40]

eQTL mapping

The interrogation of gene expression to facilitate the design and interpretation of disease association studies can be a powerful tool for the identification of biologically functional variants and the interpretation of biological effects.[41] By introducing a quantifiable and easily measurable biological outcome, it is possible to assess the relevance of statistical significance and eliminate some of the issues raised above. The use of gene expression variation in the context of disease mapping can be viewed in two ways (Figure 1). On the one hand, hypotheses can be generated by discovering functional variation using eQTL mapping and subsequently testing those functional variants in large case-control samples. On the other hand, following a disease association study that has identified several signals located within regions of non-coding DNA, eQTL mapping can be used to interpret and dissect the functional effect of the candidate disease variants. Below, the two directions will be explored in more detail.

Generating functional candidates using eQTL mapping

Regions with functional effects on gene expression can be localised through the use of association mapping. Gene expression, or mRNA level, is a quantitative phenotype that can be assayed in multiple individuals. When the same individuals are surveyed for genetic variation at marker loci, for example SNPs, association analysis tests whether variation at each SNP can explain the observed phenotypic variation. The rationale behind this analysis is that markers themselves are either the causal variant or are highly correlated (in LD) with the causal variant.

Association mapping of gene expression variation has been successful in many species, including human,[42–45] yeast,[46–48] mouse,[49–52] rat,[53] fish[54,55] and maize.[52] Together, these studies provide several striking observations related to the nature of functional variation influencing gene expression. First, variation in gene expression levels among individuals is common — and much of that phenotypic variation has a genetic basis. Much of the association signal is located *cis*- to the gene of interest,[45,52,56] although *trans*-acting variants have also been observed. Hotspots of gene regulation (ie regions of the genome influencing expression of several genes) have been observed in some,[44] but not all, studies.

There are several ways in which the study of the regulation of gene expression can enhance disease studies as well as narrow the choice of candidate regions for disease association studies. Where information exists about the contribution of particular genes to a disease phenotype or susceptibility, understanding the regulatory control of those genes may assist in elucidating the complete set of effects. In addition, understanding the regulation of categories of genes, or genes of a particular pathway, may provide targets for further follow-up in disease studies. It may prove more time-efficient and cost-effective to have a list of many potential functional variants located throughout the genome, however, and test them against a large number of diseases. Whole-genome eQTL studies can provide a list of regions of the genome with functional effects on the expression of known genes (Figure 1a). SNPs located within these regions can then serve as candidates for disease association studies, much in the same way that non-synonymous SNPs are often considered because of their potential functional effect. There are several advantages to this type of targeted approach over a whole-genome scan. First, because the number of SNPs to be genotyped in each individual is reduced, many more individuals can be surveyed in a disease study without vastly increasing costs. The reduction in the number of markers tested can eliminate some of the problems of multiple test correction, more sensible thresholds can be used and smaller effect variants can be

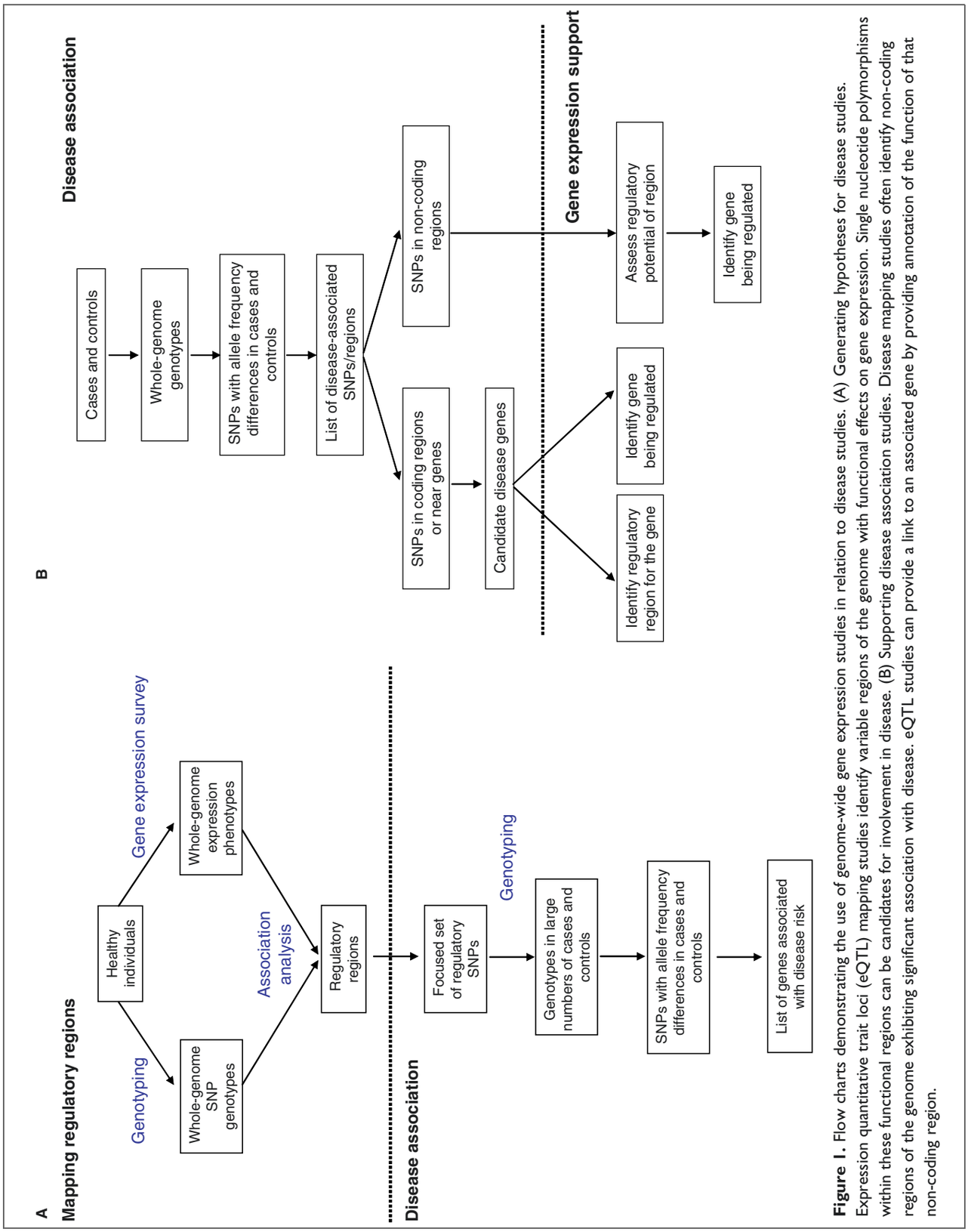

Figure 1. Flow charts demonstrating the use of genome-wide gene expression studies in relation to disease studies. (A) Generating hypotheses for disease studies. Expression quantitative trait loci (eQTL) mapping studies identify variable regions of the genome with functional effects on gene expression. Single nucleotide polymorphisms within these functional regions can be candidates for involvement in disease. (B) Supporting disease association studies. Disease mapping studies often identify non-coding regions of the genome exhibiting significant association with disease. eQTL studies can provide a link to an associated gene by providing annotation of the function of that non-coding region.

detected. Secondly, any significant associations detected between SNP and disease phenotype provide both a mechanism (gene regulation) and the identity of the affected gene. Finally, the fact that potential causal regulatory variants were initially discovered in healthy individuals and subsequently have been associated with disease means that such variants are common and are likely to contribute significantly to the disease risk of the population.

The methodology above carries the risk of focusing only on certain types of genomic variants, while it is known that much of genome function is still missing. A way to circumvent this problem is to enhance disease studies by incorporating the data on functional regulatory regions while using commercially available whole-genome SNP genotyping chips in disease studies, in order to perform the association analysis using Bayesian methods that assign different prior probabilities to SNPs on the array. Under such a scenario, SNPs located in regions with known functional effects on expression of specific genes — as identified through eQTL studies — would be assigned a higher prior probability of being associated with a phenotype. In addition, one might assign a higher prior probability to SNPs in known promoters, enhancers or transcription factors. Thus, one could focus on the effects of candidate variants without missing other important signals. Another substantial advance of knowing regulatory variants before performing a genome-wide association study is that one can correlate phenotypes and regulatory networks and utilise such information in the statistical modelling of the disease.

Supporting genome-wide disease association studies: Narrowing on disease-associated non-coding signals

Genome-wide association studies are now increasing in frequency,[6,57] and although it would have been preferable to have identified all functional regulatory variants in advance, investigators will be faced with the challenge of interpreting some of the strongest association signals. Many of the association studies have a multi-phase design, wherein a fraction of the SNPs with the top statistical significance in the first phase are genotyped in a subsequent phase in a new set of individuals. The statistical exercise must eventually give way to biological interpretation, however, and the identification of the causal variant will be necessary. Although most of the confirmed disease-causing variants are located in coding regions, this observation is due to an ascertainment bias in the ability to predict the potential functional consequences of nucleotide variation. As the human genome is composed of only ~3–5 per cent coding DNA, and studies increasingly attribute function to non-coding DNA, it might be expected that much of the disease-causing variation will be non-coding and that many of the significant peaks in an association analysis will fall in regions devoid of genes.

Disease association studies often identify non-coding regions of the genome exhibiting significant association with disease. The exploration of those non-coding regions will benefit from the survey of gene expression variation and how it relates to genetic variation (eQTL mapping). For any disease-associated non-coding region (eg from a case-control study), it is possible to test whether the disease-associated SNPs and haplotypes are also associated with gene expression variation of nearby genes (as identified from eQTL studies; Figure 2). This enables conclusions to be drawn about the nature of the function of the causal variant. For instance, if the same haplotype that appears to increase the disease risk also appears to be associated with high expression of a nearby gene, it is possible to start making some connections between the biology of the affected gene and the disease itself. Moreover, one can hypothesise (and hopefully test) how levels of expression of a gene might affect disease risk. This simple connection between the two types of study could provide not only the identity of the gene that is linked to the disease, but also the consequence of genome variation that linked the gene with the disease. It may also provide some clues about other candidates (upstream transcriptional regulators, interacting proteins etc).

Several studies illustrate the utility and validity of using gene expression variation for disease fine mapping. Two of these studies have focused on identifying functional nucleotide variation by focusing primarily on the regions surrounding each of a set of genes (*cis*-), but also considering other regions located *trans*- to the genes.[42,45] These studies showed that a large fraction of genes (10–20 per cent) have significant variation that affect their gene expression in *cis*, and in some cases in *trans*. The regulatory variants that affect gene expression variation can be mapped with the same resolution as disease variants in genome-wide association studies because, in both cases, the resolution depends on the LD structure of the human populations. These studies, which allow the identification of regulatory haplotypes, need to be verified before functional experiments can be performed. The most appropriate way to perform a first-pass verification is to test whether allelic imbalance in expression is correlated with heterozygosity in the same SNPs as those that showed genotypic association with gene expression.

Even with this information, and the fact that the effect of a causal variant may be known to have an effect on gene regulation, it is still a long way from being able to identify the exact DNA variant that causes the regulatory effect and subsequent increased disease risk. This is a stage where things become complex for many reasons. For example, although the genome-wide distribution of LD is quite variable, average LD in the human genome extends over large regions, which makes it challenging to fine map a causal variant in many regions. In the best case scenario, associated SNPs would be identified in a region of very low LD, thus reducing the

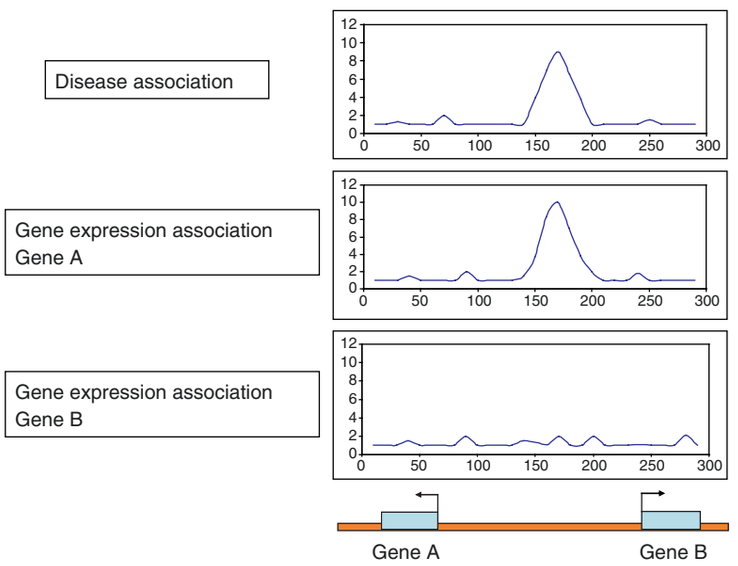

Figure 2. The three panels show the signal of association (as $-\log p$-value of individual single nucleotide polymorphism (SNP)–phenotype associations) of genomic regions with disease, gene expression of gene A and gene expression of gene B. From this plot, it is suggested that genetic variation that increases disease risk is also associated with gene expression variation of gene A (assuming that the associated SNPs and haplotypes are the same). This probably indicates that the disease risk is a regulatory effect and that the amount of transcript (or protein) of gene A is critical for the development of the disease.

number of potential causal variants to test subsequently. More often, an associated region of approximately 10–20 kilobases will be identified.[45] Although fine mapping in a population with reduced LD (eg Africans) might assist in identifying shorter associated regions, it is at this stage where extensive amounts of information about genome function are crucial. The diversity of methodologies for large-scale interrogation of the human genome for function is increasing; the resulting information will be very important for prioritising which of those associated DNA segments to focus on first.

Interpreting regulatory variation

The identification of the causal variant can benefit from incorporating information about genome function. Many studies to determine functionality within the human genome sequence are now in progress using high-throughput, genome-wide methodologies. The ENCyclopedia Of DNA Elements (ENCODE) project is the best example[58] of this type of study. The aim of this project is to attribute a functional identity to each nucleotide of the human genome. In its pilot phase, 1 per cent of the human genome (44 genomic regions) has been studied extensively for function, interspecies conservation and population genetic variation. The comprehensive analysis of these 44 regions will provide important clues for the pattern and structure of genome function and will allow predictions for the nature of variations behind complex disease and phenotypic variation. This, and other ongoing studies, will offer a first-pass annotation of functional elements in the human genome and will provide the framework for detailed characterisation of functional variation.

If an established and confirmed association of a region with disease and gene expression variation exists, and there is light annotation of the associated region for coding and non-coding elements, it is possible to apply brute force approaches to identify the specific DNA changes that are causal. A recommended strategy is to perform extensive resequencing of potentially functional segments of the region in high and low expressing individuals. The number of individuals required to be assayed depends on the magnitude of the functional effect and the predicted within-population frequency of the causal variant. This can be assisted by initial power calculations that allow prediction of what is likely to be identified, given the study design. The optimal approach seems to be to sample sequences from individuals at each of the two ends of the phenotypic (expression) distribution, and then proceed inwards.

As soon as a set of genomic segments have been resequenced, one should look for variants that appear to have equal or better correlation with the phenotype than that observed in the initial association study. This can be determined by genotyping all of the potentially functional variants (identified in the resequencing approach in a subset of the individuals) in the original complete sample. Depending on the strength of these correlations, and where the highly

correlated variants are located, the appropriate approach should be adopted for direct functional testing of causal haplotypes. Such approaches can include reporter constructs, binding assays, RNA stability assays and chromatin modification assays using all of the alternative haplotypes.

Summary

In this paper, some issues have been discussed that arise from the incorporation of gene expression variation data in disease studies. The overall message is that gene expression can greatly assist the discovery of disease variants, as well as the interpretation of the biological effects of causal variants. Further exploration of gene expression variation in more samples and more cell types will greatly enhance both our understanding of phenotypic variation in humans and also the nature of regulatory variation and its impact on complex disease.

References

1. Jimenez-Sanchez, G., Childs, B. and Valle, D. (2001), 'Human disease genes', *Nature* Vol. 409, pp. 853–855.
2. Gulcher, J.R., Kong, A. and Stefansson, K. (2001), 'The role of linkage studies for common diseases', *Curr. Opin. Genet. Dev.* Vol. 11, pp. 264–267.
3. Cooke, G.S. and Hill, A.V. (2001), 'Genetics of susceptibility to human infectious disease', *Nat. Rev. Genet.* Vol. 2, pp. 967–977.
4. Risch, N. and Merikangas, K. (1996), 'The future of genetic studies of complex human diseases', *Science* Vol. 273, pp. 1516–1517.
5. Hirschhorn, J.N. and Daly, M.J. (2005), 'Genome-wide association studies for common diseases and complex traits', *Nat. Rev. Genet.* Vol. 6, pp. 95–108.
6. Klein, R.J., Zeiss, C., Chew, E.Y. *et al.* (2005), 'Complement factor H polymorphism in age-related macular degeneration', *Science* Vol. 308, pp. 385–389.
7. Ozaki, K., Ohnishi, Y., Iida, A. *et al.* (2002), 'Functional SNPs in the lymphotoxin-alpha gene that are associated with susceptibility to myocardial infarction', *Nat. Genet.* Vol. 32, pp. 650–654.
8. Kasvosve, I., Delanghe, J.R., Gomo, Z.A. *et al.* (2000), 'Transferrin polymorphism influences iron status in blacks', *Clin. Chem.* Vol. 46, pp. 1535–1539.
9. Ueda, H., Howson, J.M., Esposito, L. *et al.* (2003), 'Association of the T-cell regulatory gene *CTLA4* with susceptibility to autoimmune disease', *Nature* Vol. 423, pp. 506–511.
10. Tournamille, C., Le Van Kim, C., Gane, P. *et al.* (1995), 'Molecular basis and PCR-DNA typing of the Fya/fyb blood group polymorphism', *Hum. Genet.* Vol. 95, pp. 407–410.
11. Tsuge, M., Hamamoto, R., Silva, F.P. *et al.* (2005), 'A variable number of tandem repeats polymorphism in an E2F-1 binding element in the 5′ flanking region of SMYD3 is a risk factor for human cancers', *Nat. Genet.* Vol. 37, pp. 1104–1107.
12. Zhou, X.F., Cui, J., DeStefano, A.L. *et al.* (2005), 'Polymorphisms in the promoter region of catalase gene and essential hypertension', *Dis. Markers* Vol. 21, pp. 3–7.
13. McDermott, D.H., Zimmerman, P.A., Guignard, F. *et al.* (1998), 'CCR5 promoter polymorphism and HIV-1 disease progression. Multicenter AIDS Cohort Study (MACS)', *Lancet* Vol. 352, pp. 866–870.
14. Kostrikis, L.G., Neumann, A.U., Thomson, B. *et al.* (1999), 'A polymorphism in the regulatory region of the CC-chemokine receptor 5 gene influences perinatal transmission of human immunodeficiency virus type 1 to African-American infants', *J. Virol.* Vol. 73, pp. 10264–10271.
15. Sakuntabhai, A., Turbpaiboon, C., Casademont, I. *et al.* (2005), 'A variant in the CD209 promoter is associated with severity of dengue disease', *Nat. Genet.* Vol. 37, pp. 507–513.
16. Carlson, C.S., Aldred, S.F., Lee, P.K. *et al.* (2005), 'Polymorphisms within the C-reactive protein (CRP) promoter region are associated with plasma CRP levels', *Am. J. Hum. Genet.* Vol. 77, pp. 64–77.
17. VanNess, S.H., Owens, M.J. and Kilts, C.D. (2005), 'The variable number of tandem repeats element in DAT1 regulates in vitro dopamine transporter density', *BMC Genet.* Vol. 6, p. 55.
18. Kochi, Y., Yamada, R., Suzuki, A. *et al.* (2005), 'A functional variant in FCRL3, encoding Fc receptor-like 3, is associated with rheumatoid arthritis and several autoimmunities', *Nat. Genet.* Vol. 37, pp. 478–485.
19. Kwok, J.B., Hallupp, M., Loy, C.T. *et al.* (2005), 'GSK3B polymorphisms alter transcription and splicing in Parkinson's disease', *Ann. Neurol.* Vol. 58, pp. 829–839.
20. Al-Zahrani, A., Sandhu, M.S., Luben, R.N. *et al.* (2006), 'IGF1 and IGFBP3 tagging polymorphisms are associated with circulating levels of IGF1, IGFBP3 and risk of breast cancer', *Hum. Mol. Genet.* Vol. 15, pp. 1–10.
21. Bennett, S.T., Lucassen, A.M., Gough, S.C. *et al.* (1995), 'Susceptibility to human type 1 diabetes at IDDM2 is determined by tandem repeat variation at the insulin gene minisatellite locus', *Nat. Genet.* Vol. 9, pp. 284–292.
22. Karim, M.A., Wang, X., Hale, T.C. and Elbein, S.C. (2005), 'Insulin promoter factor 1 variation is associated with type 2 diabetes in African Americans', *BMC Med. Genet.* Vol. 6, p. 37.
23. Przybylowska, K., Kluczna, A., Zadrozny, M. *et al.* (2006), 'Polymorphisms of the promoter regions of matrix metalloproteinases genes MMP-1 and MMP-9 in breast cancer', *Breast Cancer Res. Treat.* Vol. 95, pp. 65–72.
24. Humphries, S.E., Luong, L.A., Talmud, P.J. *et al.* (1998), 'The 5A/6A polymorphism in the promoter of the stromelysin-1 (MMP-3) gene predicts progression of angiographically determined coronary artery disease in men in the LOCAT gemfibrozil study Lopid Coronary Angiography Trial', *Atherosclerosis* Vol. 139, pp. 49–56.
25. Ye, S., Eriksson, P., Hamsten, A. *et al.* (1996), 'Progression of coronary atherosclerosis is associated with a common genetic variant of the human stromelysin-1 promoter which results in reduced gene expression', *J. Biol. Chem.* Vol. 271, pp. 13055–13060.
26. Borm, M.E., van Bodegraven, A.A., Mulder, C.J. *et al.* (2005), 'A NFKB1 promoter polymorphism is involved in susceptibility to ulcerative colitis', *Int. J. Immunogenet.* Vol. 32, pp. 401–405.
27. Emison, E.S., McCallion, A.S., Kashuk, C.S. *et al.* (2005), 'A common sex-dependent mutation in a RET enhancer underlies Hirschsprung disease risk', *Nature* Vol. 434, pp. 857–863.
28. Grice, E.A., Rochelle, E.S., Green, E.D. *et al.* (2005), 'Evaluation of the RET regulatory landscape reveals the biological relevance of a HSCR-implicated enhancer', *Hum. Mol. Genet.* Vol. 14, pp. 3837–3845.
29. Knight, J.C., Udalova, I., Hill, A.V. *et al.* (1999), 'A polymorphism that affects OCT-1 binding to the TNF promoter region is associated with severe malaria', *Nat. Genet.* Vol. 22, pp. 145–150.
30. 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.
31. 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.
32. Kruglyak, L. and Nickerson, D.A. (2001), 'Variation is the spice of life', *Nat. Genet.* Vol. 27, pp. 234–236.
33. http://www.ncbi.nlm.nih.gov/SNP/index.html
34. http://www.hapmap.org
35. Benjamini, Y. and Hochberg, Y. (1995), 'Controlling the false discovery rate — A practical approach to multiple testing', *J. R. Stat. Soc. Ser. B Methodol.* Vol. 57, pp. 289–300.
36. Storey, J.D. and Tibshirani, R. (2003), 'Statistical significance for genomewide studies', *Proc. Natl. Acad. Sci. USA* Vol. 100, pp. 9440–9445.
37. Doerge, R.W. and Churchill, G.A. (1996), 'Permutation tests for multiple loci affecting a quantitative character', *Genetics* Vol. 142, pp. 285–294.
38. Cardon, L.R. and Palmer, L.J. (2003), 'Population stratification and spurious allelic association', *Lancet* Vol. 361, pp. 598–604.

39. Pritchard, J.K., Stephens, M., Rosenberg, N.A. and Donnelly, P. (2000), 'Association mapping in structured populations', *Am. J. Hum. Genet.* Vol. 67, pp. 170–181.
40. Tang, H., Quertermous, T., Rodriguez, B. *et al.* (2005), 'Genetic structure, self-identified race/ethnicity, and confounding in case-control association studies', *Am J. Hum. Genet.* Vol. 76, pp. 268–275.
41. Stranger, B.E. and Dermitzakis, E.T. (2005), 'The genetics of regulatory variation in the human genome', *Hum. Genomics* Vol. 2, pp. 126–131.
42. Cheung, V.G., Spielman, R.S., Ewens, K.G. *et al.* (2005), 'Mapping determinants of human gene expression by regional and genome-wide association', *Nature* Vol. 437, pp. 1365–1369.
43. 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.
44. 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.
45. Stranger, B.E., Forrest, M.S., Clark, A.G. *et al.* (2005), 'Genome-wide associations of gene expression variation in humans', *PLoS Genet.* Vol. 1, p. e78.
46. Brem, R.B. and Kruglyak, L. (2005), 'The landscape of genetic complexity across 5,700 gene expression traits in yeast', *Proc. Natl. Acad. Sci. USA* Vol. 102, pp. 1572–1577.
47. 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.
48. Yvert, G., Brem, R.B., Whittle, J. *et al.* (2003), 'Trans-acting regulatory variation in *Saccharomyces cerevisiae* and the role of transcription factors', *Nat. Genet.* Vol. 35, pp. 57–64.
49. 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.
50. Doss, S., Schadt, E.E., Drake, T.A. and Lusis, A.J. (2005), 'Cis-acting expression quantitative trait loci in mice', *Genome Res.* Vol. 15, pp. 681–691.
51. 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.
52. 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.
53. 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.
54. 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.
55. 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.
56. Yan, H., Yuan, W., Velculescu, V.E. *et al.* (2002), 'Allelic variation in human gene expression', *Science* Vol. 297, pp. 1143.
57. Herbert, A., Gerry, N.P., McQueen, M.B. *et al.* (2004), 'A common genetic variant is associated with adult and childhood obesity', *Science* Vol. 312, pp. 279–283.
58. The ENCODE Project Consortium (2004), 'The ENCODE (ENCyclopedia Of DNA Elements) Project', *Science* Vol. 306, pp. 636–640.