Mapping complex disease traits with global gene expression

William Cookson, Liming Liang‡, Gonçalo Abecasis‡, Miriam Moffatt* and Mark Lathrop§*

Abstract | Variation in gene expression is an important mechanism underlying susceptibility to complex disease. The simultaneous genome-wide assay of gene expression and genetic variation allows the mapping of the genetic factors that underpin individual differences in quantitative levels of expression (expression QTLs; eQTLs). The availability of systematically generated eQTL information could provide immediate insight into a biological basis for disease associations identified through genome-wide association (GWA) studies, and can help to identify networks of genes involved in disease pathogenesis. Although there are limitations to current eQTL maps, understanding of disease will be enhanced with novel technologies and international efforts that extend to a wide range of new samples and tissues.

Genome-wide association study
(GWA study). An examination of common genetic variation across the genome designed to identify associations with traits such as common diseases. Typically, several hundred thousand SNPs are interrogated using microarray or bead chip technologies.

**National Heart and Lung Institute, Imperial College London, London SW3 6LY, UK.*
‡Center for Statistical Genetics, Department of Biostatistics, SPH II, Ann Arbor, Michigan 48109-2029, USA.
§CEA / Centre National de Genotypage, 91057 Evry, France.
Correspondence to W.C., L.L., G.A., M.M. or M.L.
e-mails:
w.cookson@imperial.ac.uk;
lianglim@umich.edu;
goncalo@umich.edu;
m.moffatt@imperial.ac.uk;
mark@cng.fr
doi:10.1038/nrg2537

Genome-wide association (GWA) studies of common complex or multifactorial diseases have been spectacularly successful in the last 2 years, with many new loci identified with levels of probability that were once thought unattainable. However, the extraordinary levels of significance of the association signals have yet to be translated into a full understanding of the genes or genetic elements that are mediating disease susceptibility at particular loci.

The functional effects of DNA polymorphism on multifactorial disease can be mediated through several mechanisms. Polymorphisms that alter protein function can have very important effects, such as *NOD2* (nucleotide-binding oligomerization domain-containing 2; also known as *CARD15*) mutations in inflammatory bowel disease[1] and *FLG* (filaggrin) mutations in eczema (atopic dermatitis)[2]. However, systematic study of complex diseases with known non-synonymous SNPs has not yielded many highly significant results[3], and variation in gene expression is probably a more important mechanism underlying susceptibility to complex disease. The abundance of a gene transcript is directly modified by polymorphism in regulatory elements. Consequently, transcript abundance might be considered as a quantitative trait that can be mapped with considerable power. These have been named expression QTLs (eQTLs)[4,5].

There is a substantial gap between SNP associations from a GWA study and understanding how the locus contributes to disease. Further genotyping and statistical analyses are often necessary to identify causal variants, which are then functionally investigated. This Review explores the value of systematic identification of eQTLs as one means of characterizing the function of loci underlying complex disease traits. The combination of whole-genome genetic association studies and the measurement of global gene expression allows the systematic identification of eQTLs. By assaying gene expression and genetic variation simultaneously on a genome-wide basis in a large number of individuals, statistical genetic methods can be used to map the genetic factors that underpin individual differences in quantitative levels of expression of many thousands of transcripts.

The resulting comprehensive eQTL maps provide an important reference source for categorizing both *cis* and *trans* effects of disease-associated SNPs on gene expression. In addition to providing information about the biological control of gene expression, such data aid in interpreting the results of GWA studies. Once the statistical evidence for association of genetic markers to a disease trait has been established, genome-wide eQTL mapping data can be examined to see if the same genetic markers are also associated with quantitative transcript levels of one or more genes — such markers are known as eSNPs. The availability of systematically generated eQTL information provides immediate insight into a probable biological basis for the disease associations, and can help to identify networks of genes involved in disease pathogenesis.

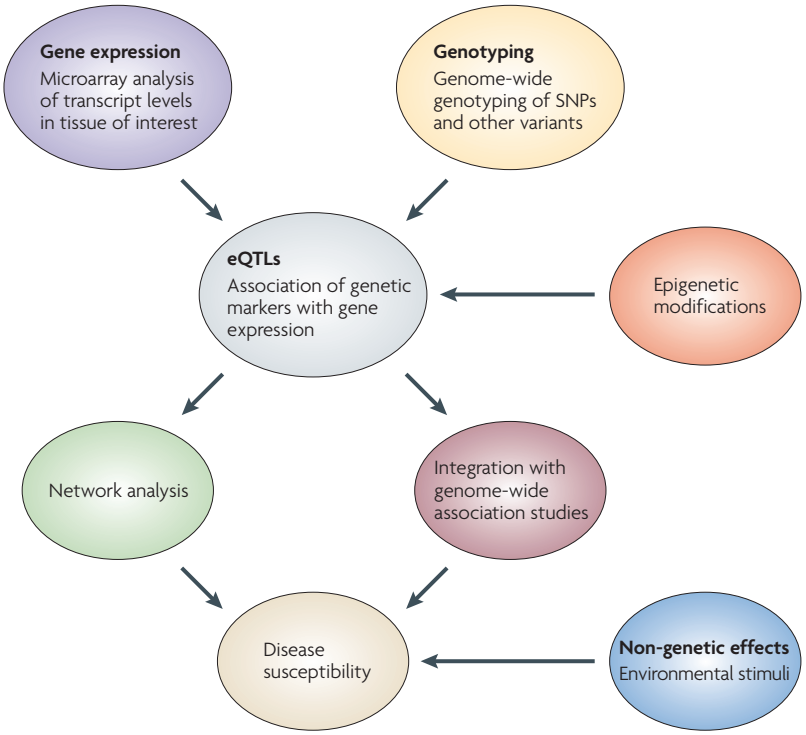

Figure 1 | **eQTL mapping.** Expression QTL (eQTL) mapping begins with the measurement of gene expression in a target cell or tissue from multiple individuals. This information is the substrate for investigating the effects of DNA polymorphism (of any type) on the expression of individual genes. Other factors that can alter transcription, such as epigenetic CpG methylation, may also be mapped. Network analyses build upon strong correlations that are present between transcripts, and allow the identification of modules of genes that mediate complex functions. This information can then be made available and used to interpret genetic associations and mapping information from the study of complex disease.

The potential of genome-wide eQTL identification was shown originally in the yeast *Saccharomyces cerevisiae*[6] and then in humans, animals and plants[4,7]. The history of eQTL mapping has been comprehensively reviewed[7–9], and will not be described in detail here. This Review will show how the combination of genetics and global gene expression can be a powerful tool for systematically unravelling the effects of variation in transcription on disease. First, we briefly introduce the principles and current methods of eQTL mapping and describe the basis of eQTLs. We then explore the relevance of these results to disease gene identification. The limits of current eQTL mapping data are discussed, as are the expected impact of new technologies, international efforts to extend results to new samples and tissues, and how cell lines might be tested with stimuli that are relevant to disease.

eQTL mapping

In practical terms, the starting point for eQTL mapping is the measurement of gene expression in a target cell or tissue from multiple individuals (FIG. 1). This information is the substrate for investigating the effects of DNA polymorphism (of any type) on the expression of individual genes. The use of microarray technology to measure gene expression from many thousand of genes simultaneously has been a principle driving force for systematic mapping of eQTLs[7]. The field is benefiting from progressively more sophisticated platforms for such studies, which are described in the later sections of this Review. Procedures for eQTL mapping are based on the insight that expression levels can be analysed with genetic approaches in the same manner as any other quantitative trait phenotype, such as body weight or blood lipids. In particular, study designs and statistical methods that are used traditionally to map QTLs can be successfully applied to the identification of eQTLs[10–12]. Interpretation of eQTL data can then be developed further by the incorporation of additional biological information, such as epigenetic modifications and analysis of regulatory networks, which are discussed below.

eQTLs are influenced not only by genetic polymorphisms, but also by a range of other biological factors. These can be dissected systematically, starting with the measurement of heritability (H^2).

Heritability. Family studies have shown that many human eQTLs are highly heritable[13,14]. The linkage approach, in which family members are studied, has been valuable in demonstrating that genetic factors have widespread and identifiable influences on eQTLs in humans, and such studies have provided broad localization for some of the underlying genetic factors[15,16]. GWA mapping of common genetic variants that underlie eQTLs has recently become possible owing to the wide availability of high-throughput and low-cost SNP genotyping. These results are particularly relevant to disease mapping that is also focused on common SNPs characterized with similar SNP arrays. Moreover, the interpretation of these eQTL data relies strongly on methodologies that have been developed for disease GWA[13]. For example, a family study of lymphoblastoid cell lines (LCLs) identified nearly 15,000 traits (each corresponding to an individual Affymetrix probe) with an estimated $H^2 > 0.3$, indicating that genetic influences on gene expression seem to be widespread[13]. Other studies have similarly described a high H^2 of many eQTLs in LCLs and other tissues[4,17,18].

Genetic factors (with both *cis*-acting and *trans*-acting effects; see below), are often identified for eQTLs that have high H^2. For example, in the LCL study mentioned above[13], eQTLs for 81% of traits with $H^2 > 0.8$ could be mapped to one or more SNPs at genome-wide significance. However, the SNP map on average accounted for less than 20% of the estimated trait H^2, consistent with results obtained by other studies[16]. This indicates the presence of genetic or other factors affecting familial clustering on transcription that are not detectable in these genetic associations. Factors other than SNPs that might affect H^2 are discussed in more detail below. Further understanding of disease phenotypes can also be gained from analysing whether particular types of genes have more heritable variation in expression level[13,19] (BOX 1).

Cis *and* trans *effects.* Statistical analyses of eQTLs need to take into account that the loci identified can influence gene expression either in *cis* or in *trans*. The definition of

Epigenetic
A mitotically stable change in gene expression that depends not on a change in DNA sequence, but on covalent modifications of DNA or chromatin proteins such as histones.

Heritability
(H^2). The heritability of an individual trait is estimated by the ratio of genetic variance to total trait variance, so that 0 indicates no genetic effects on trait variance and 1 indicates that all variance is under genetic control.

Major histocompatibility complex
(MHC). A complex locus on chromosome 6p that comprises numerous genes, including the human leukocyte antigen genes, which are involved in the immune response.

Gene Ontology
(GO). A widely used classification system of gene functions and other gene attributes that uses a standardized vocabulary. The system uses a hierarchical organization of concepts (an ontology) with three organizing principles: molecular functions (the tasks done by individual gene products), biological processes (for example, mitosis) and cellular components (examples include the nucleus and the telomere).

a *cis* effect is somewhat arbitrary, but *cis*-acting eQTLs are typically considered to include SNPs within 100 kb upstream and downstream of the gene that is affected by that eQTL. This definition becomes more problematic in regions of extended linkage disequilibrium, such as the major histocompatibility complex (MHC) locus.

Detailed analysis of the position of mapped *cis*-acting eQTL effects have shown that these are enriched around transcription start sites and within 250 bp upstream of transcription end sites, and they rarely reside more than 20 kb away from the gene[20]. *Cis*-acting variants also seem to occur more often in exonic SNPs[20]. *Trans* effects are usually weaker than *cis* effects in humans[4,5] and in rats[21], but they are more numerous.

It is not known if *trans* effects are mostly mediated through transcription factor variants or through other mechanisms. 'Master regulators' are *trans*-acting factors with multiple effects on gene expression that have been identified in *S. cerevisiae*[22], in rat tissues[21] and in the human genome[5]. It is of interest that, at least in yeast, master regulators are not enriched for transcription factors, and *trans*-regulatory variation seems to be broadly dispersed across classes of genes with different molecular functions[22].

Other types of variant. The function of DNA can be altered by many mechanisms in addition to SNPs. Transcription can also be modified by copy number variants (CNVs), insertions and deletions, short tandem repeats and single amino acid repeats[23]. A systematic investigation of the effects of CNVs in individuals who are part of the International HapMap project showed that SNPs and CNVs captured 84% and 18%, respectively, of the total detected genetic variation in gene expression but the signals from the two types of variation had little overlap[24]. It has been shown that CNVs in regulatory hot spots in the malaria parasite genome dictate transcriptional variation[25]. It has also been observed that small-scale copy number variation (that is, a single or few copies) can lead to multiple orders of magnitude change in gene expression and, in some cases, switches in deterministic control[26].

Box 1 | **Gene ontology analyses**

Many expression QTLs (eQTLs) are highly heritable. Therefore, Gene Ontology (GO) analyses can be applied to eQTL databases to identify the types of gene that show the most inherited variation in their levels of expression (at least in the cell type studied, usually lymphoblastoid cell lines; LCLs). The most highly heritable GO biological process for eQTLs in LCLs in one study was, unexpectedly, 'response to unfolded proteins', a group containing numerous chaperonins and heat shock proteins. The individual variation in response to unfolded proteins may be an evolutionary response to cellular stress, and these genes could be candidates in the study of neurodegenerative diseases and the ageing processes. Genes that regulate RNA processing, DNA repair and progression through the cell cycle were also exceptionally heritable. The evolutionary advantage of individual variation in these genes is unclear.

As expected, genes with significant heritability are also enriched in GO categories of immune response[13,19]. These highly heritable immune genes may be of particular value for the study of infectious and inflammatory diseases. The most heritable traits can be considered as candidate genes for effects on particular disease traits, but they could also be studied in large population samples, such as those contained in national biobanks, to investigate their actions on unexpected phenotypes.

Epigenetic factors. In addition to DNA sequence variants, gene transcription is also modulated by epigenetic modifications (discussed further in the 'Limitations of mapping studies' section below). For example, non-germ line epigenetic methylation of CpG residues that regulate gene expression is common in the human genome[27]. In a limited study of three chromosomes, 17% of genes can be differentially methylated in their 5′ UTRs and approximately one-third of the differentially methylated 5′ UTRs are inversely correlated with transcription[27]. A further level of complexity comes from post-translational modifications of histones that modulate DNA accessibility and chromatin stability to provide an enormous variety of alternative interaction surfaces for *trans*-acting factors (reviewed in REF. 28).

eQTLs and disease gene mapping

Combining eQTL and GWA studies. One of the most important consequences of eQTL mapping is the link that it provides between genetic markers of disease identified in GWA studies and the expression of a specific gene or genes. In particular, the power of these studies depends upon the identification of specific genetic markers that are simultaneously associated with disease and eQTLs, whereas simply comparing differences in gene expression in cases and controls might not provide sufficient power to detect important differences with the available sample sizes. The value of this is illustrated by several recent investigations in which eQTL analysis was incorporated directly as a component of the GWA study design (included in TABLE 1). The number of GWA studies continues to rise rapidly. In GWA studies to date, 10–15% of the top hits have affected a known eQTL in a public data set (TABLE 1). We will therefore discuss selected instances of these to show the value of the method.

For example, a recent study generated genome-wide transcriptional profiles of lymphocyte samples from participants in the San Antonio Family Heart Study, and showed that high-density lipoprotein cholesterol concentration was influenced by the *cis*-regulated vanin 1 (*VNN1*) gene[15]. Similarly, a study of post-mortem brain tissue identified eQTLs affecting the *MAPT* (microtubule-associated protein tau) and *APOE* (apolipoprotein E) genes, which play an important part in Alzheimer's disease[29].

At the same time as the San Antonio study the results of a GWA study of asthma[13,30] identified a series of SNPs in strong linkage disequilibrium and spanning more than 200 kb of chromosome 17q23. The study showed that these SNPs were strongly associated with the risk of asthma[30]. The region of association contains 19 genes, none of which is an obvious candidate for disease. Examination of eQTL data derived from Affymetrix HU133A arrays[13,30] on the same families showed that the disease-associated SNPs had highly significant ($p < 10^{-22}$) effects in *cis* on the expression of one the genes: *ORMDL3* (ORM1-like 3).

This locus illustrates the utility of combining eQTL and disease mapping studies. Despite the highly significant association with both expression and disease, the

Table 1 | **Disease-linked associations with significant expression QTLs from the literature and public databases**

Study	Trait	Region	Candidate gene(s)	Transcript affected by SNP	Transcript region	Logarithm of odds (LOD) score
Gudbjartsson *et al.*[102]*	Height	7p22	*GNA12*	*GNA12*	7p22	13
		11q13.2	Intergenic	*CCND1*	11q13	7.4
		7q21.3	*LMTK2*	*C17orf37*	17q21	6.0
				HSD17B8	6	6.4
				NDUFS8	11	6.1
		3p14.3	*PXK*	*RPP14*	3	9.2
Göring *et al.*[15]	High-density lipoprotein cholesterol levels	6q21	*VNN1*	*HDL* (serum)	Multiple sites	8.0
Kathiresan *et al.*[40]	Polygenic dyslipidaemia	20q13	*PLTP*	*PLTP*	20q13	16
		15q22	*LIPC*	*LIPC*	15q22	17
		11q12	*FADS1*, *FADS2*, *FADS3*	*FADS1*	11q12	35
				FADS3	11q12	8.0
		9p22	*TTC39B*	*TTC39B*	9p22	7.0
		1p13	*CELSR2*, *PSRC1*, *SORT1*	*SORT1*	1p13	270
				PSRC1	1p13	249
				CELSR2	1p13	80
		12q24	*MMAB*, *MVK*	*MMAB*	12q24	43
		1p31	*ANGPLT3*	*DOCK7*	1p31	27
				ANGPLT3	1p31	11
Libioulle *et al.*[37]	Crohn's disease	5p13	Intergenic	*PTGER4*	5p13	3.0
Barrett *et al.*[36]	Crohn's disease	5q31	*OCTN1*, *SLC22A4*, *SLC22A5*	*SLC22A5*	5q31	Unknown
Hom *et al.*[103]*	Systemic lupus erythematosus	8p23.1	*C8orf13*, *BLK*	*BLK*	8p23.1	20
				C8orf13	8p23.1	28
Harkonason *et al.*[104]*	Type 1 diabetes	12q13	*RAB5B*, *SUOX*, *IKZF4*	*RPS26*	12q13	33
		1p31.3	*ANGPTL3*	*DOCK7*	1p31.3	16
Wellcome Trust Case Control Consortium[105]*	Type 1 diabetes	12q13.2	*ERBB3*	*RPS26*	12q13.2	43.2
Todd *et al.*[106]*	Type 1 diabetes	12q13.2	*ERBB3*	*RPS26*	12q13.2	30.3
Plenge *et al.*[107]*	Rheumatoid arthritis	9q34	*TRAF1-C5*	*LOC253039*	9q34	6.3
Thein *et al.*[108]	Fetal haemoglobin F production	6q23.3	Intergenic	*HBS1L*	6q23.3	6.0
Moffatt *et al.*[30]	Childhood asthma	17q21	Intergenic	*ORMDL3*	17	14
Wellcome Trust Case Control Consortium[105]*	Bipolar disorder	16p12	*PALB2*, *NDUFAB1*, *DCTN5*	*DCTN5*	16p12	9.2
		6p21	*NR*	*HLA-DQB1*	6p21	8.9
				HLA-DRB4	6p21	11
Di Bernardo *et al.*[109]*	Chronic lymphatic leukaemia	2q37	*SP140*	*SP140*	2q37	8.8

*Identified through comparison of the National Human Genome Research Institute's Catalog of Published Genome-Wide Association Studies and the mRNA by SNP Browser v 1.0.1.

predicted expression differences in cases and controls, which was averaged over all genotypes, was not expected to be significant given the sample size: this was in agreement with the observed results[30].

In these data, borderline significant effects were also observed in the expression of the gene neighbouring *ORMDL3*, *GSDML* (gasdermin-like)[30]. Subsequent eQTL studies with the Illumina platform and RT-PCR experiments confirmed that the same SNPs determine eQTLs with both genes. These results focus attention on one or both of these genes as probable candidates for a role in disease pathology. Many additional studies are now underway to investigate the biological functions of these two genes and their relationship to asthma[31,32–35].

Using eQTLs to interpret GWA studies. Such findings have encouraged the use of these eQTL data as a general tool for interpreting results from GWA studies.

Recent analyses of Crohn's disease (CD) illustrate this approach[36,37]. Initially, markers on chromosome 5 were shown to be strongly associated with CD in one GWA scan, but their biological effects could not be readily deduced as they reside in a 1.25 Mb gene desert. Examination of the LCL eQTLs database showed that one or more of these polymorphisms act as a long-range *cis*-acting factor influencing expression of *PTGER4* (prostaglandin E receptor 4), a gene that resides approximately 270 kb proximal to the association region[37]. The homologue of this gene has been implicated in phenotypes similar to CD in the mouse[37,38]. Thus, research is now focused on *PTGER4* as a primary candidate gene for this disease susceptibility locus.

Subsequently, the eQTL approach has been applied systematically in a meta-analysis of GWA studies of CD, and several other interesting results have been obtained[36]. For example, eQTLs were used to address an outstanding question in CD genetics related to the identification of the CD susceptibility gene or genes in the cytokine cluster on chromosome 5q31, where SNPs have an established association with disease[39]. The disease-associated SNPs in the meta-analysis of this region were all shown to be correlated with decreased *SLC22A5* (solute carrier family 22, member 5) mRNA expression levels.

Another CD locus identified in the meta-analysis coincided with the asthma risk locus on chromosome 17, in which the disease markers are also correlated with expression of *ORMDL3* and *GSDML*, as described above. Thus the same genetic variants contribute to susceptibility to both CD and asthma, possibly by perturbing expression of one or both of these genes. Several additional examples of eQTLs within CD susceptibility loci have also been reported[36]. These co-localizations greatly exceed the number that would be expected by chance, suggesting that many of them are indicative of underlying biological processes involved in disease susceptibility[36].

Public GWA study results are available at the National Human Genome Research Institute's Catalog of Published Genome-Wide Association Studies. Examination of these results identifies many other disease associations for which eQTL data provide similar insights (TABLE 1). For example, a recent large study of polygenic dyslipidaemia identified 30 loci with highly significant effects on blood lipid measurements[40]. Examination of gene expression in samples of liver from 957 subjects allowed highly significant eQTLs to be identified for 7 of the 30 loci[40] (TABLE 1). In some cases, the eQTL data give genetic evidence to support a candidate gene for which a role was previously suggested from location and biological hypotheses (such as *GNA12* for height on chromosome 7p22, and *BLK* and *C8orf13* for auto-immune systemic lupus erythematosis on 8p23.1). More often the gene expression data identifies different genes or suggests a particular gene from a number of candidates. Examples of this include the cluster of *trans*-acting genes from the height locus on chromosome 7q21.3, the *RPS26* gene from the type 1 diabetes locus on 12q13.2, and the *DCTN5* gene from the bipolar disorder locus on chromosome 16p12.1.

Not all examples of eQTL findings are straightforward, as exemplified by the association reported between the *SH2B1* (SH2B adaptor protein 1) locus and body mass index (BMI)[41]. In this study, a missense SNP in *SH2B1* was also associated with significant variation in transcript abundances of *EIF3C* (eukaryotic translation initiation factor 3, subunit C) and *TUFM* (Tu translation elongation factor, mitochondrial). When mutated, the homologue of *SH2B1* leads to extreme obesity in mice, apparently because of a failure in proper regulation of appetite. The authors speculate that the *SH2B1* variant has a causal role but is in linkage disequilibrium with a different variant that influences *EIF3C* and *TUFM* mRNA levels; alternatively, regulation of *EIF3C* or *TUFM* mRNA levels could have a causal role instead of, or in addition to, variation in *SH2B1* (REF. 41).

eQTL databases. mRNA by SNP Browser v 1.0.1 is a database of eQTLs from asthma studies[13,30] that allows searches by genes, chromosomal regions and SNPs, and is a good example of how data from this kind of research can be examined. VarySysDB is another public database and contains 190,000 extensively annotated mRNA transcripts from 36,000 loci. VarySysDB offers information encompassing published human genetic polymorphisms for each of these transcripts separately. In addition to SNP effects on transcription, VarySysDB includes deletion–insertion polymorphisms from dbSNP, CNVs from the Database of Genomic Variants, short tandem repeats and single amino acid repeats from H-InvDB and linkage disequilibrium regions from D-HaploDB[23].

MHC locus. Analysis of eQTLs in the MHC locus is of particular interest for studies of diseases in which infection and autoimmunity is a major component[42]. Intense study of the MHC locus over many years has revealed many genes that are duplicated or polymorphic, and DNA variants in the MHC locus have been associated with more diseases than any other region of the human genome[42]. Many disease associations have been attributed to selective binding of processed antigen to the antigen-presenting grooves of human leukocyte antigen (HLA) variants.

The results of eQTL studies on the MHC locus must be interpreted with caution. This is because the high degree of genetic variability and linkage disequilibrium across the MHC locus could introduce some spurious results owing to polymorphism in sequences corresponding to probes used for expression measurements[43] (see below). Nevertheless, global gene expression data has shown very strong effects of particular SNPs on the level of expression of the classical MHC antigens *HLA-A*, *HLA-C*, *HLA-DP*, *HLA-DQ* and *HLA-DR* ($p < 10^{-20}$ to $p = 10^{-30}$)[13]. This confirmed the effect of genetic variation on the level of *HLA-DQ* expression observed previously[44]. The strength of these effects suggests that associations of MHC class I and class II polymorphism might depend on the level of gene transcription as much as restriction of response to antigen[13]. A possible example is type I diabetes, in which the functional effects of the long-recognized association to the class II MHC genes[45] have not been elucidated, despite combined p values of less than 10^{-100}.

Human leukocyte antigen (HLA). A glycoprotein, encoded at the major histocompatibility complex locus, that is found on the surface of antigen-presenting cells and that present antigen for recognition by helper T cells.

Serial analysis of gene expression
(SAGE). A method for quantitative and simultaneous analysis of a large number of transcripts. Short sequence tags are isolated, concentrated and cloned; their sequencing reveals a gene expression pattern that is characteristic of the tissue or cell type from which the tags were isolated.

These results suggest that even in this intensively studied region, the investigation of eQTLs could add to our understanding of the many known genetic associations.

Additional biological interpretation and validation. A genome exerts its functions not through particular genes or proteins, but through highly complex networks that produce a range of responses[46]. As perturbations of such networks underlie the pathogenesis of many diseases[47,48], network analysis incorporating eQTL data has recently provided important novel insights into mechanisms underlying multifactorial diseases[16,17,49] (BOX 2).

Extensive investigations of human populations, animal models and cellular systems are required to provide biological validation of the relationship between specific genes and multifactorial disease traits, even when the relationship is identified through eQTL analysis. Given the substantial effort that is required for validation, careful selection of only the strongest candidates is essential. As shown in the above examples, the combination of GWA studies and eQTL analysis is a powerful way of identifying a small number of candidate genes and pathways. With the deployment of new technologies, such as exon arrays and RNA resequencing, and expansion of the tissue types covered, as described below, we expect future eQTL databases to be even more powerful tools for such identifications.

Box 2 | **Networks and other analytical tools**

Traditional genetics and cellular biology has rested on the assumption that a single stimulus (or DNA variant) when applied to a cell (or gene) will have a single outcome. The reality is that even a simple stimulus will induce changes in transcription in many genes that interact in complex networks, with an outcome that affects many different transcripts and processes.

The networks can be considered to be made up of multiple pathways that act at genetic, genomic, cellular, tissue and whole-organism levels[46]. The technology that is already available to gather global information on gene expression, proteins and metabolites is now allowing the systematic identification of the networks of genes that interact in disease processes[92,93]. Analysis of genetic variants that perturb networks through the effects of expression QTLs (eQTLs) has recently provided important novel insights into mechanisms underlying multifactorial diseases[16,17,49]. This type of analysis may also lead to the systematic identification of transcription modules[94] and the construction of regulatory networks[95]. The potential of genetic mapping approaches to identify networks of genes operating on hematopoietic stem cells[96] and immune responses[97] are amongst the examples that have been discussed in the literature.

The impact of combining eQTL analysis with an investigation of gene networks is shown by the recent detection of genetic variants associated with transcript abundance of a macrophage-enriched network and obesity-related traits in human subjects. Parallel studies in mice and humans identified a network module for obesity-related traits that was enriched for genes involved in the inflammatory and immune response. eQTL mapping was then used to identify *cis*-acting genetic variants associated with this network of genes. The authors characterized these genetic variants in a large cohort of individuals, and showed statistical enrichment for variants that were associated with obesity-related biometric traits[16]. This approach allowed the identification of genetic variants that had minor individual effects on the trait, but that can be identified as a group because of the overall perturbation of the network. Three genes in this network, lipoprotein lipase (*Lpl*), lactamase β (*Lactb*) and protein phosphatase 1-like (*Ppm1l*), were validated by gene knockouts, strengthening the association between this network and metabolic disease traits[49].

A bibliography and a range of statistical routines for network analysis can be found on the Weighted Gene Co-expression Network site.

Potential limitations and future directions

Despite the power of eQTL mapping to help identify the genetic basis of disease, there are many limitations to current methodologies and potential for considerable improvements as technologies develop. The best appreciated technical barriers to optimal eQTL mapping are in the use of microarrays to measure gene expression (BOX 3). Other problems and their potential solutions are discussed below.

Comparisons between microarray platforms. It was assumed that different microarray platforms give broadly comparable results[50]. However, numerous studies are now showing that the overlap in transcript detection between platforms is only ~30–40%, whether considered as presence or absence of detectable transcripts or the absolute level of transcript abundance[51–53]. The same level of discordance appears whether comparisons are made between Affymetrix arrays and serial analysis of gene expression (SAGE)[52], Affymetrix and Illumina arrays[50], Affymetrix and Applied Biosystems arrays[53], or across multiple platforms[51].

Some of this discrepancy may be because individual genes are commonly interrogated by different sequences on different platforms. The situation can be improved when matching of genes is sought using genomic sequence rather than sequences inferred from the UniGene database of transcripts[54]. Concordance between platforms is improved further when probes are compared only when they target overlapping transcript sequence regions on cDNA microarrays or gene chips[55].

These discrepancies may follow from the complex and unpredictable factors that determine hybridization of particular nucleic acids to complementary array-bound sequences[56,57]. In addition, the selection of sequences on microarrays has been strongly biased to the 3′ end of genes, simply because public cDNA databases were first populated with genes identified by 3′ tags.

A consistent conclusion of comparison studies has been that different platforms provide complementary results[51,52], probably because they are all sampling only a selected fraction of the total transcriptome from the cells or tissue under study. The use of multiple platforms to extract all the expression information from a cell or tissue is impractical.

New platforms for measuring gene expression. A more comprehensive measurement of gene expression comes from arrays that interrogate all known human exons. Affymetrix have produced global exon arrays[58], which show a high degree of correspondence in terms of fold changes with their pre-existing 'classical microarrays', suggesting that the additional probe sets on the exon arrays will provide reliable as well as more detailed coverage of the transcriptome[59]. The use of exon arrays allows the identification of tissue-specific alternative splicing events as well as significant expression outside of known exons and well-annotated genes[60]. Exon arrays on other platforms are likely to provide similarly robust results.

Box 3 | **Pitfalls with microarrays**

The use of microarrays to measure gene expression has led directly to the development of expression QTL (eQTL) analyses. However, the microarray approaches that underlie most eQTL studies to date provide only partial gene coverage and have a limited dynamic range for quantitative detection of expression. Specific problems inherent in the use of these microarrays include: the systematic bias that can be introduced during sample preparation, hybridization and measurement of expression; batch to batch variation in array manufacture; and day to day variation in laboratory conditions[98]. These types of effects are probably under-recognized, as exemplified by a report of large-scale differences in gene expression between ethnic groups[98,99]. In this case the highly significant differences in gene expression that the data had suggested between the groups[98] were found to be due to the separate processing of expression measurements in lymphoblastoid cell lines (LCLs) from subjects of European and Asian ancestry.

Cis-eQTL artefacts can also arise from the overlap of SNPs with transcript probes[100]. Alterations in hybridization efficiency owing to the SNP can give an erroneous impression of differences in transcript abundance attributable to the SNP (and to other DNA variants with which it is in linkage disequilibrium)[100]. It has been estimated that 15% of microarray probes for any given gene will overlap with SNPs that are polymorphic in the population under study[100]. However, most coding SNPs in the human genome are uncommon, and it also seems that measurements of abundances are robust against mismatches between the probe and RNA sequences[101]. Although evidence that these artefacts have an effect has been presented[43], it is reassuring that in a large study in humans Emilsson *et al.*[16] found no evidence of systematic or specific hybridization artefacts from SNPs in their eQTL data. Nevertheless, important findings from microarrays need confirmation by specific assays, such as quantitative PCR, that avoid polymorphic sequences. Statistical methodology to account for batch effects, polymorphism and other sources of artefact is discussed by Alberts *et al.*[102]

Most human studies of eQTLs have been performed in LCLs, primarily because LCLs were often created as a source of nucleic acids for genetic studies. However, LCLs can exhibit progressive genomic instability with multiple passages of storage and re-growth, with the resulting potential for artefacts.

Many of the problems that are inherent in the use of microarrays can be solved by massively parallel, ultra-high-throughput DNA sequencing systems (reviewed in REF. 61). These systems allow direct ultra-high-throughput sequencing of RNA, which can then be mapped back to the genome. Sequencing of RNA provides a generic tool that can support a family of assays for measuring the genome-wide profiles of mRNAs, small RNAs, transcription factor binding, chromatin structure, DNase hypersensitivity and DNA methylation status[61]. RNA splices may also be effectively mapped by sequence-based methods.

Despite the formidable promise, ultra-high-throughput sequencing is still not without problems. The machines can produce terabytes of data daily, and make profound demands on bioinformatics for data storage and assembly of reads. Short reads may pose severe problems for the interpretation of transcripts arising from gene families with high homology or repetitive regions of the genome. Nevertheless, it can be anticipated that within 2 years many studies will rely on this technology, and that alternative or complementary approaches, such as large-scale real-time PCR-based expression assays (for example, as described by Watson *et al.*[62] and developed by WaferGen), will continue to evolve.

Limitations of mapping studies. As discussed in the section on heritability, currently mapped loci account for only a portion of the estimated heritability of eQTLs. A similar degree of unattributed or 'dark' heritability has been observed in GWA studies of common complex traits and diseases. A large GWA meta-analysis, for example, recently identified 20 variants that are significantly associated with adult height. The combined effects of the 20 SNPs explained only 3% of height variation, taking into account such factors as age and population[63]. Similarly, a large GWA meta-analysis of Crohn's disease identified 32 loci that significantly affect the disease, which together explained only 10% of the overall variance in disease risk and 20% of the genetic risk[36].

A large portion of the unattributed heritability is expected to result from the effects of multiple loci that are too weak to detect using current sample sizes[18]. This explanation would be consistent with data in yeast, in which only 3% of highly heritable transcript abundances are explained by single-locus (monogenic) inheritance and 50% are consistent with more than five controlling loci of equal effect[64]. Although current SNP arrays provide relatively comprehensive coverage of the genome (more than 80%), some of the unattributed heritability will be due to genetic factors that reside in unmapped regions, or to variation that is not effectively tagged at present, such as CNVs. Dominance and interaction effects may also account for some of the unattributed heritability, as these may be confounded with additive genetic effects in the heritability estimates with some study designs.

A previously described global eQTL study was based on sibling pairs, allowing estimates of heritability for all the transcripts measured[13]. The study suggested that dominance had a minimal effect on gene transcription[13]. Interestingly it seemed that genetic interactions may have important influences on regulation of expression for some genes, but inclusion of interaction effects had a minimal impact on the overall attributable heritability[13].

Epigenetic modifications and other factors that affect transcript abundance might not be accounted for in SNP-based association studies (see 'The basis of eQTLs' section above). Genomic imprinting is a particular case of an epigenetic effect with a parent of origin-dependent pattern. Monoalleleic expression is established at imprinted loci, via epigenetic marks transmitted through the germ line. Several common complex diseases exhibit parent-of-origin effects that might indicate underlying imprinting, including asthma[65], type I diabetes[66,67], rheumatoid arthritis[68], psoriasis[69], inflammatory bowel disease[70] and selective immunoglobin A deficiency[71], but systematic analysis of parent-of-origin effects in eQTL data has not yet been reported.

Finally, transcript abundance is a function of transcript stability as well as transcript production. Many factors mediate transcript stability, particularly in *trans*, either through protein–RNA interaction or through mechanisms mediated by small interfering RNAs (siRNAs)[72]. It seems clear that future studies of disease susceptibility as well as eQTLs will need to take these mechanisms into account.

Additive genetic effects
A mechanism of quantitative inheritance such that the combined effects of genetic alleles at two or more gene loci are equal to the sum of their individual effects.

Gene expression in tissues. Although RNA for eQTL analyses would ideally be obtained from a wide variety of tissues, most human studies of eQTLs have been performed in LCLs, primarily because LCLs were often created as a renewable source of nucleic acids for genetic studies. Gene expression in LCLs, however, represents the particular circumstances of Epstein–Barr virus infection of B-cells and their subsequent uncontrolled growth. LCLs may also exhibit extreme clonality with random patterns of monoallelic expression in single clones[73].

Although only 60% of genes from any particular cell type will also be found in LCLs[4,13], it has been established that LCLs provide information about gene expression for some genes that do not function primarily in these cells[4,74–76]. In addition, a recent comparison of eQTLs derived from the analysis of blood and adipose tissue showed little difference in the number of eQTLs that could be mapped, and there was an approximately 50% overlap of mapped loci from the two RNA sources[16]. Similarly, comparison between four different tissues showed no statistically significant differences in the number of mapped transcripts in experiments involving mapped recombinant inbred strains of mice[18].

Despite the continued utility and convenience of LCL studies of gene expression, it is evident that many of the transcripts expressed in LCLs may be housekeeping genes, and transcripts that determine specialized cell functions and that modify disease may be more parsimoniously distributed. In addition, LCLs are removed from the stimuli that can induce disordered gene transcription in disease. This is exemplified by the differences that are observed in gene expression between LCLs derived from asthmatics and genes known to be expressed in asthmatic airways[30]. These factors all indicate that the direct examination of tissues that are involved in disease can provide much more information than the LCL alone.

Some eQTL studies of human tissue have already been carried out, notably of liver[17], adipose[16,49] and brain[29] tissue. These show that approximately 60% of the transcriptome is expressed in each tissue, and that eQTLs from these tissues may be a valuable source of information for genetic mapping. Data from animal models suggest that tissue samples can allow detection of *trans*-eQTLs that are important in determining the composition of individual tissues[18]. Tissue samples will also allow the use of network analyses to identify the complex interactions that may underlie disease[16,17,49] (BOX 2).

The costs of reagents and the limited availability of appropriate tissues have to date restricted studies in humans to several hundreds of subjects. Although a formal evaluation of optimal study sizes is difficult because of unknown trait heritability, we know empirically that studies with a few hundred subjects have consistently identified numerous eQTLs with vanishingly small *p* values[4,5,75]. It is also clear that subtle effects, particularly in *trans*, would be detected more reliably with larger samples.

It is therefore timely that the promise of eQTLs as a tool for disease genetics has been sufficiently exciting to prompt a National Institutes of Health (NIH) proposal for the ambitious Genotype-Tissue Expression (GTEx) project, a database that might include 1,000 samples from each of 30 different tissues. The GTEx project is currently running as a 2-year pilot study with the primary goal of testing the feasibility of collecting high-quality RNA and DNA from multiple tissues from approximately 160 donors identified through low post-mortem interval autopsy or organ transplant settings. If the pilot phase proves successful, the project will be scaled up to involve approximately 1,000 donors, with the eventual creation of a database to house existing and GTEx-generated eQTL data.

The use of tissues poses a number of problems that need to be resolved. Normal and diseased tissue samples can be difficult to access, and their use requires careful attention to ethical, legal and social issues. Samples taken at post-mortem from many tissues robustly retain their histological architecture and contain RNA that can be of sufficient quality for measurements of gene expression. However, the changes in gene expression that might accompany death or surgical resection have not yet been documented in any detail. Tissues typically consist of different cell types, and their composition can vary inconsistently in the presence of disease. Finally, tissue-specific DNA methylation profiles may affect 20% of genes[27], and are expected to be important in understanding tissue eQTLs.

Although some of these problems may be expected to degrade the information available from the study of any particular tissue, it should be appreciated that they will not systematically lead to false positives in eQTL analyses[17], emphasizing the robustness of the eQTL approach.

Exercising the genome. Tissue biopsies and other samples extend the 'expression space' that can be examined by eQTL studies. They nevertheless still have limitations for functional analyses (particularly in humans as opposed to model organisms) when compared with cells that can be grown freely in culture and manipulated by systematic knock downs.

Although the transcripts in a particular cell under particular conditions reflect only part of the function of a particular genome, the range of transcripts from a given cell type can be widened by stimulating the cell in a variety of ways. The experimental extension of the genome expression space has been called 'exercising the genome'[77], and this strategy can be used to learn much more about gene expression and integrated gene functions. Experimentally, evidence is already emerging that environmental actions on gene expression are profound in humans[78] and model organisms[79,80] (reviewed in REF. 81), and it is reasonable to assume that these components of gene expression can be fruitfully accessed through exposure to relevant stimuli. It is interesting that, in model organisms, environmentally induced changes in gene expression seem to act through prominent *trans* effects[79,80] that may not be present in unstressed cells and tissues.

It is therefore desirable that the genome of human LCLs or primary cells of particular interest be exercised by stimulating their gene expression in different ways. Model stimuli that could be tested in these systems include pro-inflammatory stresses, metabolic stresses (such as high or low glucose, or hypoxia), the response to radiation, the response to signalling molecules (such

Box 4 | **eQTLs and network analyses of cancer**

Mutations that disrupt cell growth control mechanisms are a feature of cancer. In addition, the unchecked cell division that is characteristic of cancer can in time result in many secondary mutations and progressive genomic disorganization[103]. Genetic studies of cancer tissue (so called somatic cell genetics) have been used to identify the most common mutations in various tumours. Global gene expression studies have also been used in many cancer types, typically to identify gene signatures that can predict the clinical outcome[104]. However, most signature-based outcome predictions have not been replicated by independent studies[104], perhaps owing to the innate heterogeneity of cancerous tissue and the problems of deriving statistically stringent results from the measurement of thousands of transcripts in limited numbers of samples. Expression QTL (eQTL) analyses are a powerful tool to identify the functional consequences of the numerous copy number variants (CNVs), deletions and epigenetic modifications that are a feature of neoplastic cells. eQTL mapping allows the identification not only of genes underlying malignant processes[105] but also of genes modifying disease progression[106] and genes modulating individual responses to chemotherapy[107]. Network analyses have not yet been widely applied to the study of cancer, but they have already led to interesting findings, such as the identification of the *ASPM* gene as a molecular target in patients with glioblastoma. The application of network analyses to cancer eQTLs may be expected to greatly alleviate problems with multiple comparisons and to lead to easier biological interpretation of results[108,109]. Direct comparison of the transcript network architecture of cancerous tissue against normal tissues may also allow much deeper understanding of cancer biology.

as neurotransmitters, hormones and peptides) and the response to therapeutic and chemotherapeutic agents.

Conclusions

It is now well established that transcript abundances of genes may be considered as quantitative traits that can be mapped with considerable power, and that assaying gene expression and genetic variation simultaneously on a genome-wide basis in a large number of individuals will provide valuable tools for identifying the function of previously mapped susceptibility alleles underlying common complex diseases.

Although eQTLs are shown to be effective in mapping complex traits, there are many levels of information that are inherent in the measurement of global gene expression that have yet to be accessed, such as the effects of transcript stability, epigenetic effects or environmental stimuli. In addition, larger studies involving thousands of subjects may be necessary to identify weak *trans* effects with the same precision as the more powerful effects that are often observed in *cis*. Although *trans* effects can be relatively weak, the genes they modify (the *trans*-transcriptome) are likely to contain master regulators with wide effects on key processes that might feature more strongly in tissues or in cells subjected to particular environmental stimuli. Many genes are only expressed in particular tissues or at specific times during development. Thus, although systematic studies of eQTLs are already being planned for a wide variety of tissues, other strategies will need to be formed to study particular cell types and tissues at specific stages of differentiation and development.

Understanding the genome of cancer cells and tissues is particularly challenging because the primary lesions that initially drive cellular proliferation are difficult to find when uncontrolled division results in progressive secondary damage to the genome and the transcriptome. eQTL analyses may be of particular value in malignant disease, because they allow a more integrated picture of what is happening in cancer cells (BOX 4).

Good progress is being made in cataloguing the SNPs and other polymorphisms that regulate transcription, and this could be the basis for a systematic listing of regulatory sequences and regulatory proteins. Identifying epigenetic effects is likely to be more difficult, particularly if they are mediated through histone modifications (which are difficult to detect on a large scale) rather than through differential CpG methylation.

The remarkable diversity of human transcriptional regulation raises new questions about the evolutionary value of unexpected variation in genes that mediate basic mechanisms, such as heat shock proteins or genes influencing the cell cycle and DNA repair. 'Inverse genetics' could be used to study the SNPs with the strongest effects on expression of such genes, and to investigate their actions on unexpected phenotypes measured in epidemiological samples.

New analytical techniques, particularly network analyses, promise rapid advances in reducing the complexity of expression data. Modules of co-expressed genes mediating complex functions may also be identified by time-series studies of the response of particular cell types to environmental stimuli[82]. In future, integration of eQTLs with data from large-scale approaches for genome resequencing, from proteomic and metabolomic analyses, from epigenomic studies and from functional screening of genes may provide a powerful set of tools for a systems biology approach to multifactorial disease, as well as a way to identify and biologically validate susceptibility genes[83].

In the future, complex disease geneticists will require integrated public databases. Existing databases include the asthma study database (mRNA by SNP Browser v 1.0.1) and VarySysDB. A more comprehensive database is planned as part of the NIH GTEx project, which will house existing as well as GTEx-generated eQTL data. Future databases should include eQTL maps with SNPs, epigenetic marks, *trans* and *cis* effects, as well as effects that are specific for particular cells, tissues and environmental stimuli. Ultimately, they will also allow browsing for networks, modules and comparisons with model organisms.

1. Hugot, J. P. *et al.* Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature* **411**, 599–603 (2001).
2. Palmer, C. N. *et al.* Common loss-of-function variants of the epidermal barrier protein filaggrin are a major predisposing factor for atopic dermatitis. *Nature Genet.* **38**, 441–446 (2006).
3. Burton, P. R. *et al.* Association scan of 14,500 non-synonymous SNPs in four diseases identifies autoimmunity variants. *Nature Genet.* **39**, 1329–1337 (2007).
4. Schadt, E. E. *et al.* Genetics of gene expression surveyed in maize, mouse and man. *Nature* **422**, 297–302 (2003). **This paper shows the power of eQTL analysis in humans.**
5. Morley, M. *et al.* Genetic analysis of genome-wide variation in human gene expression. *Nature* **430**, 743–747 (2004).
6. Brem, R. B., Yvert, G., Clinton, R. & Kruglyak, L. Genetic dissection of transcriptional regulation in budding yeast. *Science* **296**, 752–755 (2002).
7. Rockman, M. V. & Kruglyak, L. Genetics of global gene expression. *Nature Rev. Genet.* **7**, 862–872 (2006).
8. Gilad, Y., Rifkin, S. A. & Pritchard, J. K. Revealing the architecture of gene regulation: the promise of eQTL studies. *Trends Genet.* **24**, 408–415 (2008).
9. Jia, Z. & Xu, S. Mapping quantitative trait loci for expression abundance. *Genetics* **176**, 611–623 (2007).
10. Carlborg, O. *et al.* Methodological aspects of the genetic dissection of gene expression. *Bioinformatics* **21**, 2383–2393 (2005).

11. Kendziorski, C. M., Chen, M., Yuan, M., Lan, H. & Attie, A. D. Statistical methods for expression quantitative trait loci (eQTL) mapping. *Biometrics* **62**, 19–27 (2006).
12. Schliekelman, P. Statistical power of expression quantitative trait loci for mapping of complex trait loci in natural populations. *Genetics* **178**, 2201–2216 (2008).
13. Dixon, A. L. *et al.* A genome-wide association study of global gene expression. *Nature Genet.* **39**, 1202–1207 (2007).
14. Visscher, P. M., Hill, W. G. & Wray, N. R. Heritability in the genomics era — concepts and misconceptions. *Nature Rev. Genet.* **9**, 255–266 (2008).
15. Goring, H. H. *et al.* Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes. *Nature Genet.* **39**, 1208–1216 (2007).
16. Emilsson, V. *et al.* Genetics of gene expression and its effect on disease. *Nature* **452**, 423–428 (2008).
This paper illustrates the power of eQTL and network analysis in unravelling complex trait genetics.
17. Schadt, E. E. *et al.* Mapping the genetic architecture of gene expression in human liver. *PLoS Biol.* **6**, e107 (2008).
18. Petretto, E. *et al.* Heritability and tissue specificity of expression quantitative trait loci. *PLoS Genet.* **2**, e172 (2006).
19. Monks, S. A. *et al.* Genetic inheritance of gene expression in human cell lines. *Am. J. Hum. Genet.* **75**, 1094–1105 (2004).
20. Veyrieras, J. B. *et al.* High-resolution mapping of expression-QTLs yields insight into human gene regulation. *PLoS Genet.* **4**, e1000214 (2008).
21. Hubner, N. *et al.* Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease. *Nature Genet.* **37**, 243–253 (2005).
22. Yvert, G. *et al. Trans*-acting regulatory variation in *Saccharomyces cerevisiae* and the role of transcription factors. *Nature Genet.* **35**, 57–64 (2003).
23. Shimada, M. K. *et al.* VarySysDB: a human genetic polymorphism database based on all H-InvDB transcripts. *Nucleic Acids Res.* **37**, D810–D815 (2008).
24. Stranger, B. E. *et al.* Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* **315**, 848–853 (2007).
25. Gonzales, J. M. *et al.* Regulatory hotspots in the malaria parasite genome dictate transcriptional variation. *PLoS Biol.* **6**, e238 (2008).
26. Mileyko, Y., Joh, R. I. & Weitz, J. S. Small-scale copy number variation and large-scale changes in gene expression. *Proc. Natl Acad. Sci. USA* **105**, 16659–16664 (2008).
27. Eckhardt, F. *et al.* DNA methylation profiling of human chromosomes 6, 20 and 22. *Nature Genet.* **38**, 1378–1385 (2006).
This paper shows the extent and distribution of methylation in the human genome.
28. Krebs, J. E. Moving marks: dynamic histone modifications in yeast. *Mol. Biosyst.* **3**, 590–597 (2007).
29. Myers, A. J. *et al.* A survey of genetic human cortical gene expression. *Nature Genet.* **39**, 1494–1499 (2007).
30. Moffatt, M. F. *et al.* Genetic variants regulating *ORMDL3* expression contribute to the risk of childhood asthma. *Nature* **448**, 470–473 (2007).
31. Bouzigon, E. *et al.* Effect of 17q21 variants and smoking exposure in early-onset asthma. *N. Engl. J. Med.* **359**, 1985–1994 (2008).
32. Duan, S. *et al.* Genetic architecture of transcript-level variation in humans. *Am. J. Hum. Genet.* **82**, 1101–1113 (2008).
33. Galanter, J. *et al.* ORMDL3 gene is associated with asthma in three ethnically diverse populations. *Am. J. Respir. Crit. Care Med.* **177**, 1194–1200 (2008).
34. Sleiman, P. M. *et al. ORMDL3* variants associated with asthma susceptibility in North Americans of European ancestry. *J. Allergy Clin. Immunol.* **122**, 1225–1227 (2008).
35. Tavendale, R., Macgregor, D. F., Mukhopadhyay, S. & Palmer, C. N. A polymorphism controlling *ORMDL3* expression is associated with asthma that is poorly controlled by current medications. *J. Allergy Clin. Immunol.* **121**, 860–863 (2008).
36. Barrett, J. C. *et al.* Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nature Genet.* **40**, 955–962 (2008).
A substantial meta-analysis of susceptibility loci underlying Crohn's disease that illustrates the problem of unattributed heritability and the utility of eQTL data in understanding the function of disease-associated SNPs.
37. Libioulle, C. *et al.* Novel Crohn disease locus identified by genome-wide association maps to a gene desert on 5p13.1 and modulates expression of *PTGER4*. *PLoS Genet.* **3**, e58 (2007).
38. Kabashima, K. *et al.* The prostaglandin receptor EP4 suppresses colitis, mucosal damage and CD4 cell activation in the gut. *J. Clin. Invest.* **109**, 883–893 (2002).
39. Rioux, J. D. *et al.* Genetic variation in the 5q31 cytokine gene cluster confers susceptibility to Crohn disease. *Nature Genet.* **29**, 223–228 (2001).
40. Kathiresan, S. *et al.* Common variants at 30 loci contribute to polygenic dyslipidemia. *Nature Genet.* **41**, 56–65 (2009).
41. Willer, C. J. *et al.* Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. *Nature Genet.* **41**, 25–34 (2009).
42. Horton, R. *et al.* Gene map of the extended human MHC. *Nature Rev. Genet.* **5**, 889–899 (2004).
43. Alberts, R. *et al.* Sequence polymorphisms cause many false *cis* eQTLs. *PLoS ONE* **2**, e622 (2007).
44. Beaty, J. S., West, K. A. & Nepom, G. T. Functional effects of a natural polymorphism in the transcriptional regulatory sequence of HLA-DQB1. *Mol. Cell. Biol.* **15**, 4771–4782 (1995).
45. Nejentsev, S. *et al.* Localization of type 1 diabetes susceptibility to the MHC class I genes *HLA-B* and *HLA-A*. *Nature* **450**, 887–892 (2007).
46. Sieberts, S. K. & Schadt, E. E. Moving toward a system genetics view of disease. *Mamm. Genome* **18**, 389–401 (2007).
47. Lamb, J. *et al.* The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* **313**, 1929–1935 (2006).
48. Goh, K. I. *et al.* The human disease network. *Proc. Natl Acad. Sci. USA* **104**, 8685–8690 (2007).
49. Chen, Y. *et al.* Variations in DNA elucidate molecular networks that cause disease. *Nature* **452**, 429–435 (2008).
50. Barnes, M., Freudenberg, J., Thompson, S., Aronow, B. & Pavlidis, P. Experimental comparison and cross-validation of the Affymetrix and Illumina gene expression analysis platforms. *Nucleic Acids Res.* **33**, 5914–5923 (2005).
51. Pedotti, P. *et al.* Can subtle changes in gene expression be consistently detected with different microarray platforms? *BMC Genomics* **9**, 124 (2008).
52. van Ruissen, F. *et al.* Evaluation of the similarity of gene expression data estimated with SAGE and Affymetrix GeneChips. *BMC Genomics* **6**, 91 (2005).
53. Bosotti, R. *et al.* Cross platform microarray analysis for robust identification of differentially expressed genes. *BMC Bioinformatics* **8** (Suppl. 1), S5 (2007).
54. Ji, Y. *et al.* RefSeq refinements of UniGene-based gene matching improve the correlation of expression measurements between two microarray platforms. *Appl. Bioinformatics* **5**, 89–98 (2006).
55. Carter, S. L., Eklund, A. C., Mecham, B. H., Kohane, I. S. & Szallasi, Z. Redefinition of Affymetrix probe sets by sequence overlap with cDNA microarray probes reduces cross-platform inconsistencies in cancer-associated gene expression measurements. *BMC Bioinformatics* **6**, 107 (2005).
56. Sohail, M., Akhtar, S. & Southern, E. M. The folding of large RNAs studied by hybridization to arrays of complementary oligonucleotides. *RNA* **5**, 646–655 (1999).
57. Southern, E., Mir, K. & Shchepinov, M. Molecular interactions on microarrays. *Nature Genet.* **21**, 5–9 (1999).
This review, by the inventor of DNA microarrays, highlights the complexity and unpredictability of the interactions between nucleic acids in solution and target sequences on solid supports.
58. Kapur, K., Xing, Y., Ouyang, Z. & Wong, W. H. Exon arrays provide accurate assessments of gene expression. *Genome Biol.* **8**, R82 (2007).
59. Okoniewski, M. J., Hey, Y., Pepper, S. D. & Miller, C. J. High correspondence between Affymetrix exon and standard expression arrays. *Biotechniques* **42**, 181–185 (2007).
60. Clark, T. A. *et al.* Discovery of tissue-specific exons using comprehensive human exon microarrays. *Genome Biol.* **8**, R64 (2007).
61. Wold, B. & Myers, R. M. Sequence census methods for functional genomics. *Nature Methods* **5**, 19–21 (2008).
62. Watson, R. M., Griaznova, O. I., Long, C. M. & Holland, M. J. Increased sample capacity for genotyping and expression profiling by kinetic polymerase chain reaction. *Anal. Biochem.* **329**, 58–67 (2004).
63. Weedon, M. N. *et al.* Genome-wide association analysis identifies 20 loci that influence adult height. *Nature Genet.* **40**, 575–583 (2008).
64. Brem, R. B. & Kruglyak, L. The landscape of genetic complexity across 5,700 gene expression traits in yeast. *Proc. Natl Acad. Sci. USA* **102**, 1572–1577 (2005).
65. Moffatt, M. & Cookson, W. The genetics of asthma. Maternal effects in atopic disease. *Clin. Exp. Allergy* **28** (Suppl. 1), 56–61 (1998).
66. Bennett, S. & Todd, J. Human type 1 diabetes and the insulin gene: principles of mapping polygenes. *Annu. Rev. Genet.* **30**, 343–370 (1996).
67. Warram, J. H., Krolewski, A. S., Gottlieb, M. S. & Kahn, C. R. Differences in risk of insulin-dependent diabetes in offspring of diabetic mothers and diabetic fathers. *N. Engl. J. Med.* **311**, 149–152 (1984).
68. Koumantaki, Y. *et al.* Family history as a risk factor for rheumatoid arthritis: a case–control study. *J. Rheumatol.* **24**, 1522–1526 (1997).
69. Burden, A. *et al.* Genetics of psoriasis: paternal inheritance and a locus on chromosome 6p. *J. Invest. Dermatol.* **110**, 958–960 (1998); comment **112**, 514–516 (1999).
70. Akolkar, P. N. *et al.* Differences in risk of Crohn's disease in offspring of mothers and fathers with inflammatory bowel disease. *Am. J. Gastroenterol.* **92**, 2241–2244 (1997).
71. Vorechovsky, I., Webster, A. D., Plebani, A. & Hammarstrom, L. Genetic linkage of IgA deficiency to the major histocompatibility complex: evidence for allele segregation distortion, parent-of-origin penetrance differences, and the role of anti-IgA antibodies in disease predisposition. *Am. J. Hum. Genet.* **64**, 1096–1109 (1999).
72. Grosshans, H. & Filipowicz, W. Molecular biology: the expanding world of small RNAs. *Nature* **451**, 414–416 (2008).
73. Plagnol, V. *et al.* Extreme clonality in lymphoblastoid cell lines with implications for allele specific expression analyses. *PLoS ONE* **3**, e2966 (2008).
74. Yan, H., Yuan, W., Velculescu, V. E., Vogelstein, B. & Kinzler, K. W. Allelic variation in human gene expression. *Science* **297**, 1143 (2002).
75. Cheung, V. G. *et al.* Natural variation in human gene expression assessed in lymphoblastoid cells. *Nature Genet.* **33**, 422–425 (2003).
76. Gretarsdottir, S. *et al.* The gene encoding phosphodiesterase 4D confers risk of ischemic stroke. *Nature Genet.* **35**, 131–138 (2003).
77. Kohane, I. S., Kho, A. T. & Butte, A. J. *Microarrays for an Integrative Genomics* (MIT Press, Cambridge, Massachusetts, 2002).
78. Idaghdour, Y., Storey, J. D., Jadallah, S. J. & Gibson, G. A genome-wide gene expression signature of environmental geography in leukocytes of Moroccan Amazighs. *PLoS Genet.* **4**, e1000052 (2008).
Although this paper describes a small study, it shows the profound effects of different environments on gene expression in peripheral blood lymphocytes.
79. Li, Y. *et al.* Mapping determinants of gene expression plasticity by genetical genomics in *C. elegans*. *PLoS Genet.* **2**, e222 (2006).
80. Smith, E. N. & Kruglyak, L. Gene–environment interaction in yeast gene expression. *PLoS Biol.* **6**, e83 (2008).
81. Gibson, G. The environmental contribution to gene expression profiles. *Nature Rev. Genet.* **9**, 575–581 (2008).
82. Reis, B. Y., Butte, A. S. & Kohane, I. S. Extracting knowledge from dynamics in gene expression. *J. Biomed. Inform.* **34**, 15–27 (2001).
This paper shows the utility of using time-series measurements of gene expression to identify co-regulated modules of genes.

83. Schadt, E. E. & Lum, P. Y. Thematic review series: systems biology approaches to metabolic and cardiovascular disorders. Reverse engineering gene networks to identify key drivers of complex disease phenotypes. *J. Lipid Res.* **47**, 2601–2613 (2006).
84. Gudbjartsson, D. F. *et al.* Many sequence variants affecting diversity of adult human height. *Nature Genet.* **40**, 609–615 (2008).
85. Hom, G. *et al.* Association of systemic lupus erythematosus with *C8orf13–BLK* and *ITGAM–ITGAX*. *N. Engl. J. Med.* **358**, 900–909 (2008).
86. Hakonarson, H. *et al.* A novel susceptibility locus for type 1 diabetes on Chr12q13 identified by a genome-wide association study. *Diabetes* **57**, 1143–1146 (2008).
87. The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).
88. Todd, J. A. *et al.* Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. *Nature Genet.* **39**, 857–864 (2007).
89. Plenge, R. M. *et al. TRAF1–C5* as a risk locus for rheumatoid arthritis — a genomewide study. *N. Engl. J. Med.* **357**, 1199–1209 (2007).
90. Thein, S. L. *et al.* Intergenic variants of *HBS1L–MYB* are responsible for a major QTL on chromosome 6q23 influencing HbF levels in adults. *Proc. Natl Acad. Sci. USA* (in the press).
91. Di Bernardo, M. C. *et al.* A genome-wide association study identifies six susceptibility loci for chronic lymphocytic leukemia. *Nature Genet.* **40**, 1204–1210 (2008).
92. Gardner, T. S., di Bernardo, D., Lorenz, D. & Collins, J. J. Inferring genetic networks and identifying compound mode of action via expression profiling. *Science* **301**, 102–105 (2003).
93. Sontag, E., Kiyatkin, A. & Kholodenko, B. N. Inferring dynamic architecture of cellular networks using time series of gene expression, protein and metabolite data. *Bioinformatics* **20**, 1877–1886 (2004).
94. Li, H. *et al.* Integrative genetic analysis of transcription modules: towards filling the gap between genetic loci and inherited traits. *Hum. Mol. Genet.* **15**, 481–492 (2006).
95. Keurentjes, J. J. *et al.* Regulatory network construction in *Arabidopsis* by using genome-wide gene expression quantitative trait loci. *Proc. Natl Acad. Sci. USA* **104**, 1708–1713 (2007).
96. Gerrits, A., Dykstra, B., Otten, M., Bystrykh, L. & de Haan, G. Combining transcriptional profiling and genetic linkage analysis to uncover gene networks operating in hematopoietic stem cells and their progeny. *Immunogenetics* **60**, 411–422 (2008).
97. de Koning, D. J., Carlborg, O. & Haley, C. S. The genetic dissection of immune response using gene-expression studies and genome mapping. *Vet. Immunol. Immunopathol.* **105**, 343–352 (2005).
98. Akey, J. M., Biswas, S., Leek, J. T. & Storey, J. D. On the design and analysis of gene expression studies in human populations. *Nature Genet.* **39**, 807–808; author reply 808–809 (2007).
99. Spielman, R. S. *et al.* Common genetic variants account for differences in gene expression among ethnic groups. *Nature Genet.* **39**, 226–231 (2007).
100. Doss, S., Schadt, E. E., Drake, T. A. & Lusis, A. J. Cis-acting expression quantitative trait loci in mice. *Genome Res.* **15**, 681–691 (2005).
101. Hughes, T. R. *et al.* Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nature Biotechnol.* **19**, 342–347 (2001).
102. Alberts, R., Terpstra, P., Bystrykh, L. V., de Haan, G. & Jansen, R. C. A statistical multiprobe model for analyzing *cis* and *trans* genes in genetical genomics experiments with short-oligonucleotide arrays. *Genetics* **171**, 1437–1439 (2005).
103. Halazonetis, T. D., Gorgoulis, V. G. & Bartek, J. An oncogene-induced DNA damage model for cancer development. *Science* **319**, 1352–1355 (2008).
104. Sun, Z., Wigle, D. A. & Yang, P. Non-overlapping and non-cell-type-specific gene expression signatures predict lung cancer survival. *J. Clin. Oncol.* **26**, 877–883 (2008).
105. Walker, B. A. *et al.* Integration of global SNP-based mapping and expression arrays reveals key regions, mechanisms, and genes important in the pathogenesis of multiple myeloma. *Blood* **108**, 1733–1743 (2006).
106. Lastowska, M. *et al.* Identification of candidate genes involved in neuroblastoma progression by combining genomic and expression microarrays with survival data. *Oncogene* **26**, 7432–7444 (2007).
107. Huang, R. S. *et al.* A genome-wide approach to identify genetic variants that contribute to etoposide-induced cytotoxicity. *Proc. Natl Acad. Sci. USA* **104**, 9758–9763 (2007).
108. Zhang, B. & Horvath, S. A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.* **4**, Article17 (2005). **This paper describes a statistical approach to network analyses and provides a set of software tools for their implementation.**
109. Horvath, S. *et al.* Analysis of oncogenic signaling networks in glioblastoma identifies ASPM as a molecular target. *Proc. Natl Acad. Sci. USA* **103**, 17402–17407 (2006).

Acknowledgements

The work was supported by the Wellcome Trust and the EC funded GABRIEL project, the French Ministry of Research and Higher Education and by grants from the National Institutes of Health.

FURTHER INFORMATION

Liming Liang's homepage: http://www.sph.umich.edu/csg/liang
Abecasis laboratory homepage (contains programs for genome-scale data analysis): http://www.sph.umich.edu/csg/abecasis
Catalog of Published Genome-Wide Association Studies: http://www.genome.gov/gwastudies
Database of Genomic Variants: http://projects.tcag.ca/variation
dbSNP: http://www.ncbi.nlm.nih.gov/projects/SNP
D-HaploDB: http://orca.gen.kyushu-u.ac.jp
Genotype-Tissue Expression (GTEx): http://nihroadmap.nih.gov/GTEx
H-InvDB: http://www.h-invitational.jp
mRNA by SNP Browser v 1.0.1: http://www.sph.umich.edu/csg/liang/asthma
UniGene: http://www.ncbi.nlm.nih.gov/unigene
VarySysDB: http://www.h-invitational.jp/varygene/home.htm
WaferGen: http://www.wafergen.com

ALL LINKS ARE ACTIVE IN THE ONLINE PDF