Using gene expression to investigate the genetic basis of complex disorders

Alexandra C. Nica and Emmanouil T. Dermitzakis*

The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge CB10 1HH, UK

Received August 8, 2008; Revised and Accepted September 7, 2008

The identification of complex disease susceptibility loci through genome-wide association studies (GWAS) has recently become possible and is now a method of choice for investigating the genetic basis of complex traits. The number of results from such studies is constantly increasing but the challenge lying forward is to identify the biological context in which these statistically significant candidate variants act. Regulatory variation plays an important role in shaping phenotypic differences among individuals and thus is very likely to also influence disease susceptibility. As such, integrating gene expression data and other disease relevant intermediate phenotypes with GWAS results could potentially help prioritize fine-mapping efforts and provide a shortcut to disease biology. Combining these different levels of information in a meaningful way is however not trivial. In the present review, we outline the several approaches that have been explored so far in this sense and their achievements. We also discuss the limitations of the methods and how upcoming technological developments could help circumvent these limitations. Overall, such efforts will be very helpful in understanding initially regulatory effects on disease and disease etiology in general.

INTRODUCTION

The ability of genome-wide association studies (GWAS) to help understand the genetic basis of complex disorders has recently become apparent. Well-documented common human genetic variation maps (e.g. HapMap project) (1), large patient samples with accurately recorded phenotypic information as well as appropriate statistical methods to assess significance (2) and account for potential biases, have all contributed to the current outburst of successful GWAS. Numerous susceptibility variants for a large number of complex diseases have been reported and effectively replicated. A present catalog of published GWAS (http://www.genome.gov/26525384) includes single nucleotide polymorphisms (SNPs) not only associated with major common disorders [Crohn’s disease (3), type 2 diabetes (4), lung cancer (5) etc.] but also with disease-relevant or anthropomorphic quantitative traits [e.g. body mass index (6) or height (7)].

What has not kept the pace however with the capacity to design and perform successful GWAS is our ability to understand how variants discovered via this hypothesis-free approach influence complex traits, In fact, few of the association studies go beyond reporting the most statistically significant hits and if they do, the suggested functionality is typically speculative, based on available annotation of genes in the vicinity of the variants. Since many of the discovered susceptibility polymorphisms fall in non-coding regions and with an increasing number of regulatory variants already implicated in a series of common disorders (8), one conventional approach has been to interrogate disease associated SNPs for associations with differential gene expression. Moffatt et al. (9) found that the same most significant SNPs associated with childhood asthma risk also explain 29.5% of the variance in ORMDL3 transcript levels, measured in lymphoblastoid cell lines. While an interesting observation, this still cannot be regarded as convincing evidence for a causal relationship between ORMDL3 and asthma onset. The concurrent progress towards uncovering the genetic basis of regulatory variation (10) has revealed an abundance of expression quantitative trait loci (eQTLs) in the human genome, making an accidental overlap between these and disease signals very likely. Thus, while gene expression is a very informative and immediate ‘DNA phenotype’, integrating expression data and disease studies genetics for an ultimate understanding of disease etiology is not straightforward.

*To whom correspondence should be addressed at: The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1HH, UK. Tel: +44 1223 4923222; Fax: +44 1223 494919; Email: md4@sanger.ac.uk

© The Author 2008. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org
ADVANCES AND CURRENT ISSUES IN EXPRESSION AND DISEASE STUDIES

Power of current eQTL studies

Natural variation in human gene expression has been recently quantified on a genome-wide scale using microarray technologies. Linkage and association studies coupling expression with genetic variability data have started to reveal the genetics underlying part of this variation, including complex allele-specific interactions (11) and its relatively high level of heritability (12–15). Most of the variants discovered with these approaches (a field also called *Genetical Genomics*) explain variance in transcript levels of nearby genes (so called *cis* eQTLs) but a few distal acting regulators have also been reported (*trans* associations). The sample sizes of genome-wide expression association studies have been fairly small though, meaning that the discoveries made so far represent generally large genetic effects [Stranger *et al.* (12) report an $R^2$ coefficient of determination ranging from 0.27 to almost 1 for the SNP–gene associations detected in the 270 HapMap individuals]. The magnitude of the discovered effects drops when pooling populations together with appropriate corrections, a direct consequence of the increased statistical power due to the larger sample sizes. The importance of appropriate statistical power has been extensively demonstrated in complex disease GWAS, where samples of a few thousand paired cases and controls have become a prerequisite (16). The main reason for this requirement is the fact that the individual contribution of genetic variants towards complex trait determination is known to be small. In fact, all susceptibility alleles discovered so far explain only a small fraction of disease risk, with odds ratios typically in the range of 1.2–1.5 (16,17). Given the marked difference between the magnitudes of detected genetic effects on expression variation and disease predisposition, respectively, it is not surprising that only few instances of overlapping signals have been observed, even when expression in a disease relevant tissue was considered. Small genetic effects on expression variation or complex interactions between regulatory variants with moderate or large effects could become decisive on a permissive environmental background. Current expression analyses are underpowered with respect to these kinds of discoveries; hence whole-genome expression association studies on larger samples would be very desirable. Such efforts are on the way, including the quantification of expression levels in blood cells of 820 HapMap III individuals from eight populations (Barbara Stranger, Stephen Montgomery and Emmanouil Dermitzakis, personal communication). Combined with SNP genotyping data, this resource will give insight into the level of expression differences among populations and generate many additional eQTLs with more subtle effects, some of them potentially related to disease. Also, many diseases manifest their phenotype in certain tissues exclusively. For these reasons, overlaying expression and disease signals is informative only if expression measurements are carried out in tissue types relevant to disease (Fig. 1). Particularly because our notion of relevance is still subjective in this case, identifying regulatory regions in multiple tissues is imperative for both a better understanding of the regulatory mechanisms in general and also the extent to which they are shared across tissues. Promising advancement in this direction can already be found in the literature. For the first time, genetic variants influencing normal human cortical expression have been reported (18). Myers *et al.* find little overlap between their set of eQTLs and results from previous studies on blood-derived cells. While differences in genotyping platforms and power can also partially explain this discrepancy, it is very likely that variants discovered here underlie brain-specific control of gene expression. These findings in conjunction with results from GWAS may help uncover the genetic basis of neurological disorders.

A more comprehensive view on the extent of shared eQTLs among tissues has been presented in a recent study by Schadt *et al.* (19). The authors have analyzed 400 human liver samples and identified more than 6000 SNP–gene associations. Many of these genes had already been linked to a variety of complex diseases, a fact to be expected given that liver is known to be essential for many human metabolic processes. In addition to data from other disease GWAS, these eQTLs were integrated with gene expression and clinical data from segregating mouse populations in order to build probabilistic gene networks (20,21). This approach has led to the prioritization of susceptibility genes for coronary artery disease, LDL cholesterol levels and type I diabetes (T1D). More importantly, the same expression platform employed for the human liver cohort here had been also used on a set of human blood and adipose tissues in another study (22). This allowed a direct evaluation of the *cis* eQTL overlap in the three tissues, which amounted to ~30%. While still a rough estimate, this amount of shared regulatory control makes a good case for interrogating disease susceptibility variants for expression associations in any available data set. Nevertheless, the remaining unshared fraction reflects the expected tissue-specific regulatory activity. These biological processes will have to be studied in appropriate tissues before they become useful for understanding tissue-specific complex disorders. It is of course still unclear what the pattern of diminishing returns is across human tissues and what set of tissues could serve as highly informative in large cohort collections.

**Finding the causal variants in expression and disease GWAS**

Even once expression data become available for a large variety of human tissues, the ability to pinpoint causal variants predisposing to disease via a regulatory mechanism will still depend on the density at which such variants have been interrogated. Currently, all genotyping platforms survey only a subset of the human sequence variation (23). Typing all common SNPs in the genome (currently estimated around 10 million) has been replaced by the inference of tag SNPs, a subset of approximately 300 000–600 000 markers, depending
on the population. This reduction exploits the pervasiveness of linkage disequilibrium (LD) in the human genome (24) and captures most of the common genetic variation in a region of interest. Still, this means that reported susceptibility variants are most probably only tagging the real functional variants and are not causal themselves. Hence, initial discoveries must be followed by fine mapping of the regions harboring the most significant statistical signals (25). So far this has not been thoroughly attempted, primarily because in the absence of other prior biological information, such tasks are financially prohibitive. Next generation sequencing will allow transcriptomic studies that will go deep into the transcript structure and abundance and will resolve issues of alternative splicing. In terms of in-depth interrogation of human genetic variation, a new resource (http://www.1000genomes.org/) promises to facilitate these efforts by generating a human genetic variation map at unprecedented resolution. The 1000 Genomes Project will sequence more than 1000 individuals with the final goal of cataloging almost all variants found at minor allele >1% frequency in human populations. Within genes, sequencing will go even deeper, down to 0.5% frequency. As a result, much like common variation documented by the HapMap has been used so far to perform successful GWAS, this project will further support disease studies aiming at those relatively rare variants. Importantly, it will allow the assessment of the individual impact of rare variants on complex traits, which has not been understood so far.

APPROACHES OF COMBINING EXPRESSION AND DISEASE DATA

Comparison of expression profiles in cases and controls

Gene expression signatures in relation to complex diseases have been investigated well before the era of GWAS. The area of cancer genetics is one such example. Expression profiles of different tumors have been broadly analyzed in order
to categorize cancer subtypes and even identify the genes that determine the subtype identification (26). Higher resolution diagnoses followed, as well as the hope that candidate cancer genes could be more easily identified. An obvious approach seemed to be the comparison of gene expression levels in the same tissue from patients and non-diseased individuals and from the differences observed to infer potentially causal pathways and biochemical functions. Given the substantial expression profile changes it causes in the cell, the complexity of genome rearrangements and the somatic nature of some of the effects, cancer is likely to be an exceptionally difficult case, and requires a slightly different treatment so we do not discuss it in this review. Even in other diseases, the daunting task in this type of experiment remains distinguishing causal from reactive effects. Changes at the cellular level will be produced as a result of the disease, and discriminating these from causal changes is extremely challenging. In addition, not all individuals have the same causal pathways so it is not necessarily true that disease and non-disease individuals have clear and well-defined differences in gene expression when compared. Finally, it is essential to derive the genetic dimension of such differences and not simply describe the gene expression differences. For all these reasons, it becomes evident that expression information derived from a reference set of samples rather than patients would be more useful for determining the disease predisposing genetic factors, while diagnosis purposes can still be served by quantifying expression in patient-derived samples. Patient samples and cellular resources will also be very useful in the context of cell-based studies that may investigate the impact of challenging these cells with various agents and candidate drugs.

Individual effect of disease variants on transcription levels

Expression variation in the general population is being actively researched and its genetic determinants uncovered. Without the burden of chasing a reactive effect, another common approach has been to use this data for determining whether disease variants from GWAS are also responsible for natural variation in human transcript levels. The underlying principle is the following: if one allele is more frequent in cases than controls and at the same time it is causal for gene expression effects of a nearby gene, which is itself important for the disease, then it is likely that causality can be established. Few genes have been prioritized by this method (19), but several caveats make it a difficult task. For example, the same best candidate SNP could be significantly associated with discernible expression differences in several nearby genes. Deciding which one among those is the best candidate would not be feasible. Moreover, absence of evidence is not evidence of absence: not finding expression associations with some genes flanking the variant of interest (provided that expression data is available for all of those genes and there is an effect on expression that eventually induces disease) does not mean that they are not potential disease contributors. Plenty of reasons from reduced power in the expression study to interrogation of a cell type irrelevant to disease could explain that.

Lessons from gene networks

The most successful integration of disease and expression data, circumventing most of the previously listed caveats has been achieved by Chen et al. (27). The authors exploit for the first time the complexity of interactions that lead to disease by modeling it as a system. Gene expression networks from liver and adipose tissues of segregating mouse populations were constructed and subsequently, network disruptions caused by DNA variations were observed. This allowed the identification of a particular network (macrophage-enriched metabolic network) enriched for genes correlated to obesity-related phenotypes. Experimental validation of the top candidate genes further supports the validity of this method. A similarly constructed network was also identified in a human blood and adipose tissue cohort (22), overlapping significantly with the mouse derived one. In addition to this discovery, gene network theory applied here demonstrated once more the naivety with which some of the links between expression and disease could be made. Despite the fact that variants associated with obesity traits resided in a region with known Apoa2 eQTLs, a gene involved in metabolic activity, it was other genes in the region that exhibited a significant linkage pattern with the phenotypes of interest. This argues for an independent relationship between Apoa2 expression and the metabolic traits under study, even though the same locus is associated with both. Clearly, the ability to distinguish between such coincidental and actual causal effects is essential. Studies based on the use of networks rather than sets of eQTLs show some theoretical promise since they account for additional complexity, but until we have a good sense of the complexity framework in which we can embed the impact of genetic variation they will lag behind.

Statistical challenges in combining GWAS and eQTL studies

Finally, the ever-increasing technological advancement supporting both expression and disease GWAS in humans, calls for the development of proper statistical tools to combine the two directly. Local genomic regions of interest with association data available from both types of analyses could provide gene prioritization clues, if explored properly. Superimposing patterns of association at commonly tested loci could provide gene prioritization clues, if explored properly. Superimposing patterns of association at commonly tested loci in expression and disease could distinguish between multiple genes affected in cis by the same variants (Fig. 2). This kind of discrimination would not be possible by solely examining the most significant disease variant for effects on expression. However, the same property that makes such correlations sound—LD—also forces the need for proper statistical assessment of the results. All variants in a genomic interval that have been historically co-segregating with the causal variant will be significantly associated with the two phenotypes by chance. On one hand, this allows association studies to be performed on just subsets of SNPs, but on the other hand, regions of high LD will exhibit tighter relationships between expression and disease patterns. This is not biologically meaningful in terms of disease and as such, correcting out the LD effect is necessary. In addition, as already pointed out, eQTLs are ubiquitously distributed throughout the genome.
human phenotypic variation, gene expression has become an obvious candidate as informative intermediate ‘DNA phenotype’. How to best integrate the two levels of complexity is nonetheless far from obvious. Currently, both expression and disease GWAS suffer from critical limitations. Expression studies have been performed on relatively small samples and on a restricted variety of tissue. Also, the transcriptome’s diversity such studies explore is bound to be incomplete, with only few probes presently designed per gene. In terms of the genetic variety underlying phenotypic differences, the picture is also far from complete. Genotyping platforms survey only a subset of the known human genetic variability, so the candidates found are most probably not causal, but just correlated with the real functional variants because of LD. All these caveats explain why a clear link between the genetics of disease and that of gene expression has not been established yet, despite its existence. Fortunately, the incredibly high rate at which technology is advancing will soon eliminate some of these limitations. Next generation sequencing has already made efforts like the 1000 Genomes Project possible, which would have seemed unthinkable few years ago. On the same note, we envisage in the near future a public resource providing complete expression profiles from a multitude of healthy human tissues derived from large samples of individuals. In conjunction with next generation sequencing data which will also greatly improve the resolution of disease GWAS, this resource would be the ideal solution to the present-day problems. In the meantime, alternative methods of integrating genotypic and phenotypic information towards a better understanding of human disease will have to be developed.

FUNDING

A.C.N. and E.T.D. are supported by Wellcome Trust funding.

ACKNOWLEDGEMENTS

We would like to thank Barbara Stranger, Stephen Montgomery, Antigone Dimas, Simon Tavare, Panos Deloukas and Gillean McVean for helpful discussions.

Conflict of Interest statement. None declared.

REFERENCES


