

A new paradigm for drug discovery: integrating clinical, genetic, genomic and molecular phenotype data to identify drug targets

E.E. Schadt^{*1}, S.A. Monks^{*†} and S.H. Friend^{*‡1}

^{*}Rosetta Inpharmatics LLC, 12040 115th Avenue N.E., Kirkland, WA 98034, U.S.A., [†]Department of Biostatistics, University of Washington, Seattle, WA 98195, U.S.A., and [‡]Merck Research Laboratories, W42-213 Sumneytown Pike, POB 4, Westpoint, PA 19846, U.S.A.

Abstract

Application of statistical genetics approaches to variations in mRNA transcript abundances in segregating populations can be used to identify genes and pathways associated with common human diseases. The combination of this genetic information with gene expression and clinical trait data can also be used to identify subtypes of a disease and the genetic loci specific to each subtype. Here we highlight results from some of our recent work in this area and further explore the many possibilities that exist in employing a more comprehensive genetics and functional genomics approach to the functional annotation of genomes, and in applying such methods to the validation of targets for complex traits in the drug discovery process.

Introduction

The identification of genes and pathways associated with complex traits that underlie common human diseases is fundamental to drug-discovery programs. High-throughput gene- and protein-expression technologies, combined with the completion of sequencing of the human, mouse and other genomes, have revolutionized our ability to monitor biological systems in a more comprehensive way. In addition, it has made possible the direct identification of genes and associated pathways underlying common human diseases. With the sequencing of the human and other genomes comes the availability of high-density single-nucleotide-polymorphism (SNP) maps that span entire genomes. This has led to the rapid evolution of the field of pharmacogenetics to the more general field of pharmacogenomics. One of the long-term visions in the pharmacogenomics field is identifying genetic profiles that explain the individual variation in drug-treatment response, with the aim of designing pharmaceutical agents that are optimized for individual response and which minimize toxicity for a given genotype constellation associated with drug response [1]. Unfortunately, while pharmacogenomics has caught the imagination of many, it is an area that still has little to offer regarding specific examples or defined strategies in the pharmaceutical industry. Success in this area will largely depend on the ability to fully incorporate the multivariate nature of disease and drug response through the use of genetic, mRNA expression, clinical, epidemiological and, if possible, proteomic and related molecular phenotype data.

Approaches that combine gene expression, genotype and clinical trait data to identify genes and pathways associated with clinical traits have historically been pursued in gene-specific ways [2–5]. An example of this is the treatment of mRNA transcript abundance as a quantitative trait to identify gene expression quantitative trait loci (eQTL) controlling the variation in mRNA expression through linkage analysis. Associations between transcript abundances and classic traits have been used to identify susceptibility loci for complex diseases such as diabetes and allergic asthma, by screening transcript abundances for thousands of genes using gene-expression microarrays [2–5]. More recently, others have reported the use of protein-expression levels in mouse brain as a quantitative trait and have mapped QTL for a moderate number of proteins polymorphic in the European collaborative interspecific backcross [6]. Brem et al. [7] recently completed the first ever comprehensive dissection of transcriptional regulation in budding yeast, giving a comprehensive glimpse of a genome-wide survey of the genetics of gene expression. Schadt et al. [8] also recently applied statistical genetics approaches to variations in mRNA transcript abundances in segregating populations to uncover the strength of genetic signature in mouse, plant and human populations. Unlike classic quantitative traits, which often represent gross clinical measurements that may be far removed from the biological processes giving rise to them, the genetic linkages associated with transcript abundances afford a closer look at the biochemical processes at play at the cellular level. Since transcript abundances are themselves potentially related to clinical traits of interest, this allows for the elucidation of the genetics of complex traits at a more refined level.

Schadt et al. [8] demonstrated the potential clinical applications and other possibilities that exist in employing a more comprehensive genetics and functional genomics approach

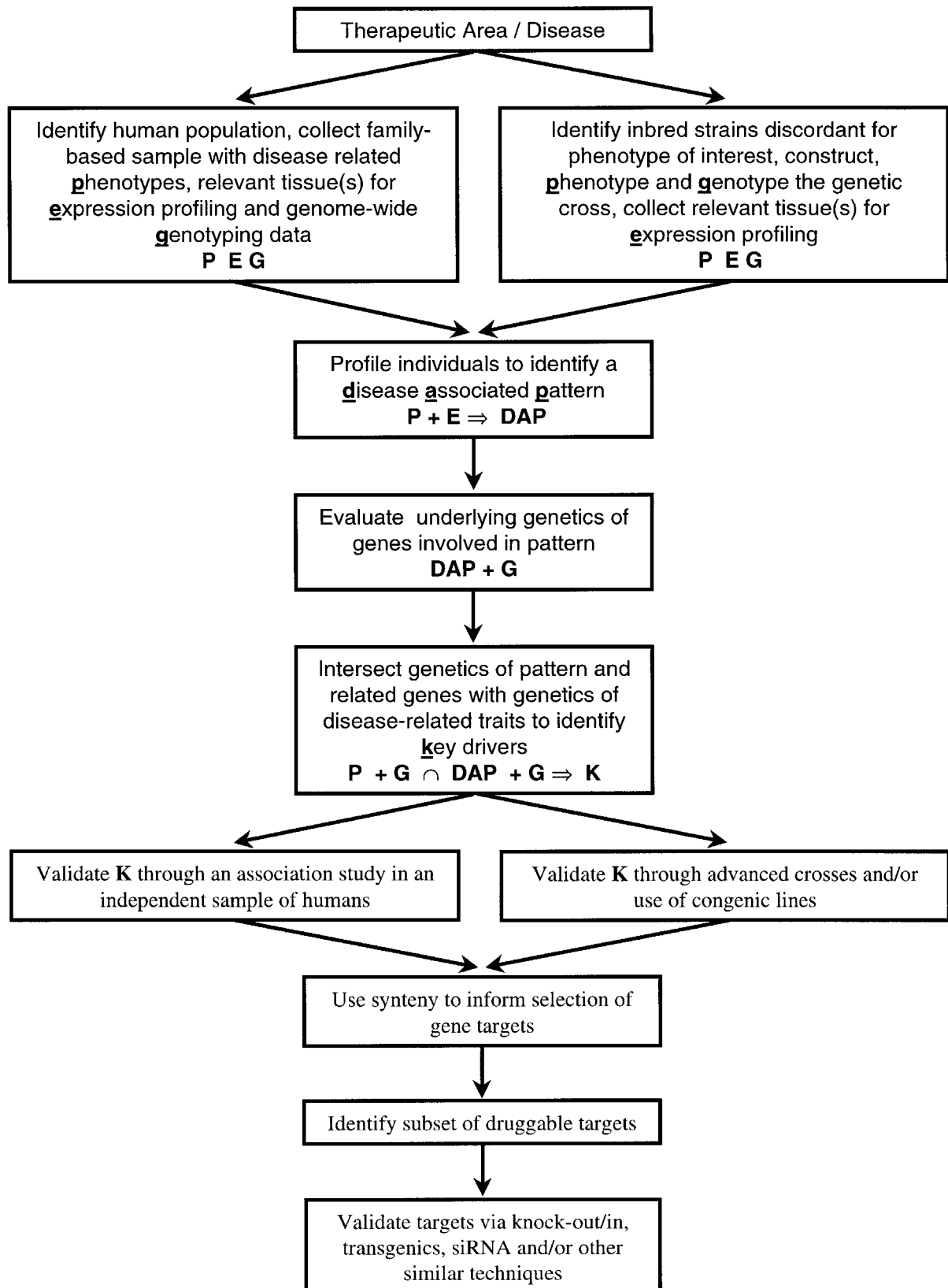
Key words: clinical, genetic linkage, mRNA expression, QTL (quantitative trait loci), sub-phenotype.

Abbreviations used: SNP, single nucleotide polymorphism; eQTL, expression quantitative trait loci; cQTL, clinical trait QTL; FPM, fat-pad mass; MC3R, melanocortin 3 receptor.

¹To whom correspondence should be addressed (e-mail eric.schadt@merck.com or stephen.friend@merck.com).

Scheme 1 | A new discovery paradigm

The new paradigm involves the combination of genetic, functional genomic and clinical data in mouse and human populations, and mapping between mouse and human genomes to identify genes and pathways underlying complex diseases of interest.



to the study of complex diseases. This included demonstrating the ability to refine the definition of a complex phenotype, identifying subtypes within a given complex phenotype that were heterogeneous with respect to underlying genetic causes, and uncovering pathways associated with the complex phenotype. Through such an approach, the potential exists to impact the more significant rate-limiting steps in the drug-discovery process: objectively classifying individuals according to disease subtypes and identifying the drivers of the pathways, the causal factors, underlying those disease subtypes. Here we demonstrate this potential with a more detailed discussion of the primary mouse example from Schadt et al. [8] In addition, we detail how results from this example can be utilized to significantly impact target discovery and target validation, as well as improve prioritization of targets for entry into the validation and lead development pipeline (see Scheme 1 for a pictorial representation). Finally, we will argue that in the past the human health and economic value derived from the use of pharmacogenomic information has been primarily restricted to the early steps of the drug-discovery process, involving target validation, or to the later steps of this process, where such information has served primarily to define non-responders with respect to a given experimental treatment. However, to realize the full economic and human health benefits from pharmacogenomic information, we need to define more precisely the currently undefined broad groups of patients that will have a high probability of response to a particular drug treatment.

Refining the definition of clinical traits using microarrays

Schadt et al. [8] describe an experiment in which livers from a population of female F₂ mice, constructed from a DBA/2J×C57BL/6 cross, were profiled using microarrays after the mice had been kept on a high-fat, atherogenic diet for 4 months. As described by Drake et al. [9], these mice model the spectrum of disease in a natural population, with many mice developing atherosclerotic lesions, and others having significantly higher fat-pad masses (FPMs), higher cholesterol levels and higher bone densities than others in the same population. Associating patterns of expression with a clinical trait and dissecting those patterns by associating them with susceptibility loci represents a potentially powerful way of dissecting complex diseases. The example here focuses on an application to an obesity-related risk trait, subcutaneous FPM.

Sub-phenotyping based on mRNA expression in phenotypic extremes

For mice comprising the upper and lower 25th percentiles of the subcutaneous FPM trait, Schadt et al. [8] identified a set of 280 genes (the FPM gene set) that represented the most differentially expressed set of genes in the extreme FPM mice. This FPM set of genes can be considered the most

transcriptionally active set of genes for the mice falling into the tails of the FPM trait distribution. The selection of genes in the FPM set was not biased by selecting genes based on (i) eQTL linkage information, (ii) their ability to discriminate between the FPM trait extremes or (iii) their correlation to genes identified by eQTL and/or trait-discrimination properties. This is noteworthy since clustering the extreme FPM mice over this gene set almost perfectly separated high FPM and low FPM mice. In addition, there appeared to be two distinct expression patterns for mice in the high FPM group, indicating some degree of heterogeneity among the high FPM mice.

Combining genetic linkage for expression and clinical phenotype

The patterns of expression resulting from clustering the extreme FPM animals over the FPM gene set serve to refine the definition of the FPM trait, as described by Schadt et al. [8]. In fact, the patterns of expression associated with different high FPM animals refine the definition of the FPM trait beyond what would be possible without the expression data. Heterogeneity of expression patterns associated with a clinical trait almost certainly points to heterogeneity in the clinical trait itself. Drake et al. [9] had previously performed a genome-wide scan to map QTL for the FPM trait. This scan revealed four clinical trait QTL (cQTL) with LOD scores greater than 2.0, and taken together these cQTL explained slightly less than 50% of the variation in the FPM trait values. To further elucidate this clinical trait, Schadt et al. [8] classified the F₂ animals into one of the three groups defined by the expression data: high FPM group 1, high FPM group 2 or low FPM group. The animals were then genetically analysed using QTL methods applied to the different high FPM groups, each combined with the low FPM group for the analysis. As an example of how genetic heterogeneity can be identified in such a setting, Schadt et al. [8] focused on the most significant FPM QTL, which fell on the distal end of chromosome 2. This QTL completely vanishes when considering one of the high FPM groups of mice, but then increases by almost two LOD units over the original LOD score when considering the other high FPM group of mice. In addition, another interesting locus was discovered on chromosome 19 that had been completely missed when all mice were considered simultaneously. In this instance, the high FPM group of mice that was not under the influence of the chromosome 2 QTL gave rise to a QTL with a significant LOD score, while the other high FPM group had a LOD score that was less significant than that obtained for the full set.

Such results provide the first-ever evidence that gene expression patterns can be used to refine the definition of a clinical trait into subtypes that are under the control of different genetic loci. The implications for drug discovery are significant and speak directly to the difficulty in dissecting complex diseases. In the situation just described, developing a compound that targeted only the gene underlying

the FPM chromosome 2 QTL would be likely to be ineffective for those in the high FPM group 1 (since they are not controlled by this locus), but would be quite effective for those in the high FPM group 2 (since they are controlled by this locus). Treating all obese individuals together in one group would result in a much less efficacious treatment than could otherwise be achieved by identifying those that would respond to the treatment. By defining the subpopulation of obese patients most likely to respond to a given drug treatment, the drug development and diagnostic components of the pharmaceutical industry will achieve greater productivity. Such productivity will naturally result from stratifying populations according to treatment groups at the earliest possible stages of drug development. This progressive strategy will more intimately link the two classically independent worlds of drug development and diagnostics. Similar arguments can be made for studying toxicity, since adverse response to a drug is also a complex trait that can be dissected in a fashion similar to that described above.

From mapping QTL for clinical traits to identification of causative genes

Identifying patterns of expression associated with a clinical trait, using the patterns of expression to identify subtypes of that clinical trait, and then identifying genetic loci that control for the different trait subtypes, does not in and of itself lead to the identification of genes underlying the QTL of interest. However, because genes underlying QTL controlling for a clinical trait may cause variation in the trait through polymorphic transcription due to DNA polymorphisms in the gene itself, it is possible to directly identify causative genes by the methods described by Schadt et al. [8]. Such direct identification of causal genes has been demonstrated in other studies. For instance, Karp et al. [2] identified a gene for airway hyper-responsiveness in a mouse model for allergic asthma by identifying genes that physically resided close to the major QTL controlling for that trait, and then identified the gene from this set by identifying that gene whose expression was most significantly associated with the clinical trait. Schadt et al. [8] have developed a more general approach to this problem by combining genetics and gene expression.

The Schadt et al. [8] approach involves the identification of eQTL that co-localize with cQTL and with the physical location of the gene whose transcription gives rise to the eQTL. In cases where the gene underlying a QTL for a clinical trait controls the variation of that trait through variation in transcription associated with DNA polymorphisms in the gene itself, the expression of that gene treated as a quantitative trait should give rise to an eQTL coincident with the cQTL. Depending on the degree of heritability of the clinical and expression traits, and the percentage of variation of the trait explained by the cQTL, we would not necessarily expect the clinical trait values and expression trait values to be

significantly correlated, even if variation in transcription of the gene causes variation in the clinical trait (for example see [10]). However, we would expect a significant genetic correlation between the clinical and gene expression traits in such cases, and so, by testing for interaction between the cQTL and gene eQTL, we can identify candidate genes underlying the cQTL for the clinical trait of interest. Schadt et al. [8] provide an example of this procedure that results in the identification of two candidate genes for the FPM trait. One of the key advantages in the application of this procedure is that the candidate genes are identified in a completely objective manner, by making full use of the genotype, expression and clinical trait data.

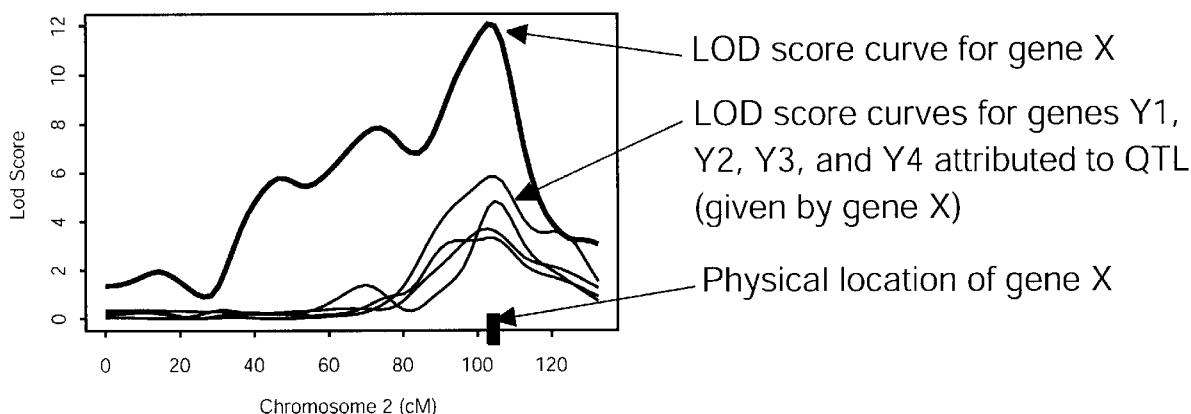
This approach serves to significantly reduce the number of genes that must be considered in identifying genes for complex traits. The QTL analysis alone reduces the number of genes to consider from all genes in the genome to those genes residing in QTL support intervals (so we go from considering tens of thousands to hundreds of genes). The gene expression/genetics combination further reduces the number of genes to consider by requiring those genes that physically reside in the QTL support interval to (i) be under the control of a *cis*-acting eQTL (so that the eQTL co-localizes with the cQTL) and (ii) have significant interaction between the eQTL and cQTL. Candidate genes identified in this way can be further validated as discussed below. While this approach is restricted to those cQTL that are associated with polymorphic transcription in the gene underlying the QTL, we expect that most complex traits under the control of many loci (say, under the control of greater than five loci) will have at least one QTL that is controlled by polymorphic transcription. Further, even in cases where the DNA polymorphism driving the cQTL leads to a functional change in the protein, we may still observe polymorphic transcription behaviour, as was the case for the susceptibility locus identified for allergic asthma [2].

Use of mouse-human synteny to identify high-quality targets, refine linkage regions and identify genes for follow-up

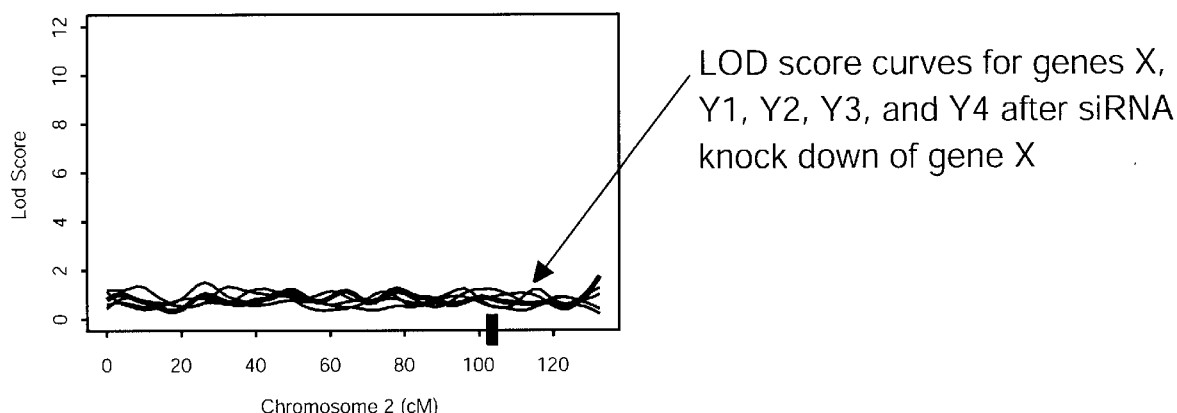
The discussion thus far has focused on using mouse models for common human diseases to elucidate the complexity of those diseases. However, information on human populations for traits that are analogous to those under study in the mouse can be useful in refining regions associated with the clinical traits in mouse or human populations. The same techniques discussed above for mouse can be applied to human populations. Further, putative candidates identified in mouse can be mapped to human orthologues, and one can determine whether the region in humans containing the orthologue has been associated with the clinical trait in human studies. The completion of the sequencing of the human [11,12] and mouse [13] genomes and the comparative maps that now exist between these two species provides for such cross-species comparisons to be a standard part of complex trait analysis.

Figure 1 | By employing *in vivo* siRNA strategies, genes can be validated as a QTL for a complex trait

LOD Score Plots with Wildtype Expression of Gene X



LOD Score Plots with siRNA Knock-Down of Gene X



This strategy was employed by Schadt et al. [8] to increase confidence in the mouse chromosome 2 QTL discussed above. The region supporting the chromosome 2 locus is homologous with human chromosome 20q12–q13.12, a region that has previously been linked to human obesity-related phenotypes. The human orthologues for the candidates identified in the mouse also reside in the human chromosome 20 region. While other genes such as melanocortin 3 receptor (MC3R) have been suggested as possible candidates for obesity at this locus, the data do not support MC3R as a candidate, and instead suggest two genes that have not been previously associated with obesity. Unlike MC3R, these two genes are not only significantly linked to the murine chromosome 2 locus, but also interact genetically with several of the FPM traits that are also linked to the chromosome 2 locus. Since the expression levels for MC3R are not linked to the chromosome 2 locus, and since there are no SNPs annotated in the exons or introns of this gene between the C57/BL6 and DBA/2J strains (as provided in the Celera RefSNP database), this provides evidence that MC3R may not be the gene underlying the

chromosome 2 linkage in this particular system. Of course, it is possible that MC3R is only expressed in the brain and that polymorphic expression of MC3R in the brain leads to changes of expression in the liver. However, it is worth noting that since there are no DNA polymorphisms in MC3R that lead to codon changes or that are likely to lead to *cis*-acting alternative splicing polymorphisms, then if it is the causative gene, it would most likely be acting through transcriptional regulation. More comprehensive gene expression profiling of different brain parts or other relevant tissues in these mice would provide more definitive evidence in this case.

Validation of high-quality targets

The small number of putative targets identified in the manner discussed above still requires an additional validation step to identify the gene underlying the QTL. Traditional methods such as gene knock-out or knock-in mice or transgenic mice can be employed for such validation. However, more recently developed methods such as RNA interference (RNAi) through the use of small interfering RNA (siRNA) offer exciting alternatives that are worthy of further investigation.

No matter the method, the aim is to identify an expression signature associated with the clinical trait, identify the causative loci driving the expression pattern and then perturb the expression of the candidate causative genes to determine whether genes associated with the expression of the causative gene are changed in a like manner.

Figure 1 provides a hypothetical example of this validation strategy. In this example, we take Y_1 , Y_2 , Y_3 and Y_4 to be genes that are part of an expression pattern associated with a complex trait of interest. The upper panel in Figure 1 plots the LOD score curves of the four genes for a particular chromosome, where the cluster of eQTL depicted here is coincident with a cQTL for the complex trait. By examining genes that physically reside in the QTL support interval, we can identify those genes that have *cis*-acting eQTL that are significantly genetically interacting with the other eQTL/cQTL. These genes represent the potential causative genes underlying the cQTL/eQTL. Gene X in Figure 1 highlights one such example. By knocking gene X out using *in vivo* siRNA methods, we can profile the siRNA-knockout animals and examine the genetic signatures of the original genes making up the eQTL cluster. The lower panel in Figure 1 highlights what we would expect if gene X were in fact driving the eQTL cluster shown in the upper panel. That is, the disappearance of the eQTL cluster would validate gene X's role as the causal factor underlying the expression pattern associated with the complex trait and would thus solidify its role as a key driver for the corresponding complex trait. If the complex trait were a disease like obesity, then validating a gene for the obesity trait directly would require the construction of, say, a knock-out animal for that gene, which is a lengthy process. However, by defining the complex trait in terms of expression patterns, we can perturb the candidate gene in more specialized ways and observe the effects on the expression pattern, which can happen in a much shorter time frame.

Finally, we note that even before a putative target is biologically validated in mice, association studies can be carried out in human populations to provide a source of validation in humans. Associating a gene in a human population with a clinical trait, where the gene in mouse (i) was physically co-localized with a cQTL for the corresponding clinical trait in a segregating mouse population, (ii) gave rise to a *cis*-acting QTL with respect to its transcription and (iii) was significantly genetically interacting with the cQTL, is itself a very powerful validation of a gene's role in the complex trait of interest. The combined validation in mouse and human provides all that is necessary to move a target forward in a discovery programme. Even in cases where the causal gene is not itself druggable, druggable targets driven by the causal gene can be identified by examining those targets that have eQTL that co-localize and are interacting with eQTL for the causative gene. This speaks to the more general use of the combined genetics/gene expression approach to reconstruct genetic networks, as discussed by Schadt et al. [8] and Jansen and Nap [14].

Discussion

For the last century genetics has been used to identify regions in the genome that 'cause' variation in a given trait. For the past decade gene expression has been used to identify those genes that are co-regulated over a range of conditions, presenting patterns of expression that help to elucidate those genes involved in complex traits. The two combined approaches have the power to refine the definition of complex phenotypes, identify subtypes within a given phenotype and uncover pathways associated with the phenotype in an unprecedented manner. The potential exists to impact the more significant rate-limiting steps in the drug-discovery process: objectively classifying individuals according to disease subtypes and identifying the key drivers of the pathways, i.e. the causal factors, underlying those disease subtypes. In the past, dissecting complex traits using genetics has met with limited success, and up to now, gene expression has served as an indirect marker for complex traits, causing many researchers to settle for functional uncertainty by restricting attention to the use of DNA markers in identifying the causal factors for complex traits. Here, we have discussed the combination of gene expression and genetics data and its potential to overcome these barriers. The addition of gene-expression data can be used to refine the disease phenotype, directly implicate pathways and genes comprising those pathways associated with the disease phenotype, and identify the key drivers of the pathways underlying the disease phenotype. Key pathway drivers can potentially be identified even in cases where these drivers are not expressed in the tissues profiled, since such key genes may be expressed in one tissue, yet drive patterns of expressions in different tissues. In such cases, transcript abundances of those genes comprising the expression patterns in the profiled tissues will be genetically linked to the physical location of the gene driving their expression. Further, while the causative genes themselves may not be druggable, the druggable genes controlled by causative genes will potentially have expression patterns that are genetically linked to the causative gene, so identification and prioritization of such genes are straightforward tasks, given the kinds of methods discussed above.

Success in elucidating complex human diseases will more and more come to depend on the ability to fully incorporate the multivariate nature of disease and drug response through the use of genetic, mRNA expression, clinical, epidemiological and, if possible, proteomic and related molecular phenotype data. It should be noted that the central dogma dictates such an approach. This type of all-encompassing approach, which we refer to as 'molecular profiling', will rely on technologies associated with SNPs and whole-genome monitoring of transcript and protein abundances in target tissues. Implementation of multiple new technologies will probably be necessary to incorporate proteomics and to facilitate the scaling-up of existing technologies to give them higher throughput. As discussed, the use of natural variation of gene expression, observed in segregating populations, can

be used to determine the effect of changes in the expression of one gene on other genes and in piecing together cellular signalling pathways. In addition, work done in yeast [15] has shown that the definitive analysis of cellular signalling pathways will often require the ability to modify protein expression and then analyse the resultant expression changes. In addition to the genetic methods described here, new methods using technologies to modify genes in mammalian cells such as siRNA [4] will be critical in delineating the key roles suggested by genomic and proteomic approaches.

In addition to these technologies, robust analytical methods will be needed to integrate the many orthogonal components of data to more optimally identify gene sets and pathways that reflect the characteristic set of altered interactions within each disease subtype. Such methods and advanced experimental designs should be a priority of the scientific community. Similarly, in order for genomics approaches to prove worthwhile in the drug-discovery process, they will need to demonstrate that they can (i) increase the probability of success for identifying targets and associated compounds, (ii) minimize costs of clinical trials and (iii) provide patients with drugs that are highly effective for specific diseases that are today loosely, or even incorrectly, grouped according to gross clinical symptoms such as depression and obesity. To achieve these benefits, hospitals, academic centres, diagnostic companies, regulatory agencies, patient advocacy groups and therapeutics-based companies will need to establish working relationships that link these groups together throughout the drug-discovery process. Such relationships will ultimately maximize the synergies between these groups in ways untested during the 20th century.

References

- 1 McLeod, H.L. and Evans, W.E. (2001) *Annu. Rev. Pharmacol. Toxicol.* **41**, 101–121
- 2 Karp, C.L., Grupe, A., Schadt, E., Ewart, S.L., Keane-Moore, M., Cuomo, P.J., Kohl, J., Wahl, L., Kuperman, D., Germer, S. et al. (2000) *Nat. Immunol.* **1**, 221–226
- 3 Liao, F., Andalibi, A., Qiao, J.H., Allayee, H., Fogelman, A.M. and Lusic, A.J. (1994) *J. Clin. Invest.* **94**, 877–884
- 4 Cohen, R.D., Castellani, L.W., Qiao, J.H., Van Lenten, B.J., Lusic, A.J. and Reue, K. (1997) *J. Clin. Invest.* **99**, 1906–1916
- 5 Eaves, I.A., Wicker, L.S., Ghandour, G., Lyons, P.A., Peterson, L.B., Todd, J.A. and Glynne, R.J. (2002) *Genome Res.* **12**, 232–243
- 6 Klose, J., Nock, C., Herrmann, M., Stuhler, K., Marcus, K., Bluggel, M., Krause, E., Schalkwyk, L.C., Rastan, S., Brown, S.D. et al. (2002) *Nat. Genet.* **30**, 385–393
- 7 Brem, R.B., Yvert, G., Clinton, R. and Kruglyak, L. (2002) *Science* **296**, 752–755
- 8 Schadt, E.E., Monks, S.A., Drake, T.A., Lusic, A.J., Che, N., Colinayo, V., Ruff, T.G., Milligan, S.B., Lamb, J.R., Cavet, G. et al. (2003) *Nature* (London), in the press
- 9 Drake, T.A., Schadt, E., Hannani, K., Kabo, J.M., Krass, K., Colinayo, V., Greaser, III, L.E., Goldin, J. and Lusic, A.J. (2001) *Physiol. Genomics* **5**, 205–215
- 10 Jiang, C. and Zeng, Z.B. (1995) *Genetics* **140**, 1111–1127
- 11 Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W. et al. (2001) *Nature* (London) **409**, 860–921
- 12 Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A. et al. (2001) *Science* **291**, 1304–1351
- 13 Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P. et al. (2002) *Nature* (London) **420**, 520–562
- 14 Jansen, R.C. and Nap, J.P. (2001) *Trends Genet.* **17**, 388–391
- 15 Roberts, C.J., Nelson, B., Marton, M.J., Stoughton, R., Meyer, M.R., Bennett, H.A., He, Y.D., Dai, H., Walker, W.L., Hughes, T.R. et al. (2000) *Science* **287**, 873–880

Received 22 December 2002