

Review

## Integrating post-genomic approaches as a strategy to advance our understanding of health and disease

Jing Tang\*, Chong Yew Tan<sup>†</sup>, Matej Oresic\* and Antonio Vidal-Puig<sup>†</sup>

Addresses: \*VTT Technical Research Centre of Finland, Tietotie 2, PO Box 1000, FIN-02044, Espoo, Finland. <sup>†</sup>Metabolic Research Laboratories, Level 4, Institute of Metabolic Science, Box 289, Addenbrooke's Hospital, Cambridge CB2 0QQ, UK.

Correspondence: Matej Oresic. Email: matej.oresic@vtt.fi; Antonio Vidal-Puig. Email: ajv22@cam.ac.uk

Published: 30 March 2009

*Genome Medicine* 2009, **1**:N (doi:10.1186/gm35)

The electronic version of this article is the complete one and can be found online at <http://genomemedicine.com/content/1/3/35>

© 2009 BioMed Central Ltd

### Abstract

Following the publication of the complete human genomic sequence, the post-genomic era is driven by the need to extract useful information from genomic data. Genomics, transcriptomics, proteomics, metabolomics, epidemiological data and microbial data provide different angles to our understanding of gene-environment interactions and the determinants of disease and health. Our goal and our challenge are to integrate these very different types of data and perspectives of disease into a global model suitable for dissecting the mechanisms of disease and for predicting novel therapeutic strategies. This review aims to highlight the need for and problems with complex data integration, and proposes a framework for data integration. While there are many obstacles to overcome, biological models based upon multiple datasets will probably become the basis that drives future biomedical research.

### Genetic analysis in the post-genomic era

In 1990, the human genome project was established to sequence the human genome [1], with the aim of applying the acquired genomic data to improve disease diagnosis and determine genetic susceptibility [2]. The publication of the first draft sequence of the human genome in 2001 [3] was thus followed by a rapid growth of different approaches to extract useful information from the genomic sequence. These approaches included, but were not limited to, the analysis of genetic variation (genomics), gene expression (transcriptomics), and gene products (proteomics) and their metabolic effects (metabolomics).

Each of these post-genomic approaches has already contributed to our understanding of specific aspects of the disease process and the development of diagnostic/prognostic clinical applications. Cardiovascular disease [4,5], obesity [6-8], diabetes [9-11], autoimmune disease

[12,13] and neurodegenerative disorders [14,15] are some of the disease areas that have benefited from these types of data. Taking the metabolic syndrome as an example, our knowledge on all aspects of the disease has grown. The metabolic syndrome is the result of a complex bioenergetic problem characterized by disturbances in lipid, carbohydrate and energy metabolism and blood pressure. In combination, these metabolic factors contribute to an increased susceptibility to cardiovascular disease, morbidity and mortality [16]. Genome-wide association (GWA) studies have identified possible genes involved in each aspect of the syndrome: namely type 2 diabetes [11], obesity [17] and hyperlipidaemia [18]. The findings have confirmed the role of certain candidate genes as well as the polygenetic nature of the syndrome. Not surprisingly, replicate GWA studies of type 2 diabetes revealed that the genes associated with disease, among others, are involved in beta-cell function and adipocyte biology [11,17,19]. In contrast, genes found to be

associated with obesity appear to be those that are predominantly involved in central appetite regulation [20-22] as key contributors to positive energy balance.

Genetic association studies in epidemiology have highlighted a number of issues. Firstly, many common disease states are related to either many genetic polymorphisms of small effect or, in selected cases, to a few of large effect. The involvement of multiple genes with unequal contributions to disease hints of complex gene-gene and gene-environment interactions. The understanding of such interactions becomes a daunting task when other modulating factors remain unknown. Secondly, some common diseases such as type 2 diabetes [12] appear to be relatively less genetically determined compared to diseases such as rheumatoid arthritis [12] and obesity [23]. In these situations, our understanding of pathophysiology requires additional data outside of genomic information. Thirdly, the initial failures to find robust replicable associations between most of the identified genetic variants and common complex diseases suggest that genomic analysis alone will not account for all of the heritability and phenotypic variation [9,24]. For this reason, there is a growing need to incorporate information derived from environmental studies and post-genomic data into genetic analysis.

#### **Advantages of combining multiple types of data**

It is clear that the genetic approach captures only one layer of the complexity inherent within human biology. There is thus a need to integrate multiple 'omics' datasets when aiming to unravel the molecular networks underlying common human disease traits [25]. Attempts have been made to combine two datasets in relation to the clinical phenotype, and this is reflected in the combination of terms found in the literature, for example metagenomics, pharmacogenomics and epigenetics. Many of the post-genomic approaches linking the genetic association data with other 'omics' layers focus on the use of 'omics'-derived phenotypic data as quantitative traits. The utility of such approaches has been previously applied, by combining genetics and metabolomics, in plant functional genomics [26]. More recently, such approaches have also been applied to human datasets. For example, Papassotiropoulos and colleagues [15] identified clusters of cholesterol-associated susceptibility genes for Alzheimer's disease by combining genetics with sterol profiling, while Gieger and colleagues [27] used ratios of metabolites to identify the function of putative genes. In another study, proteomics was linked to quantitative trait loci (QTL) in an attempt to identify changes in function rather than quantity of the protein [28].

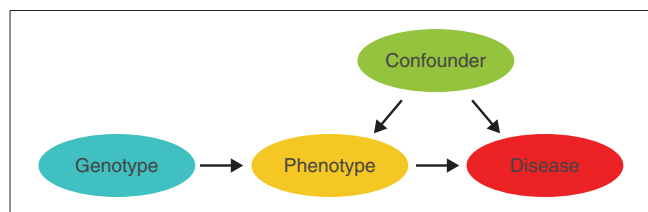
By combining multiple types of techniques, including genetics, transcriptomics, proteomics and metabolomics, we are expecting a shift toward 'environmentome' research, where all available information from periconception to

disease onset, using both longitudinal and cross-sectional experimental designs, can be obtained [9]. The measurement of traits that are modulated but not encoded by the DNA sequence, commonly referred to as intermediate phenotypes, is of particular interest. These intermediate phenotypes include not only biochemical (metabolites) and genomic (gene expression) traits, but also an individual's microbial (gut microflora) [29,30] and social traits. It is conceivable that by comprehensively examining an individual's 'environmentome', we would be able not only to understand both the genetic and environmental determinants of disease, but also to develop 'feasible' personalized medicine, that is, tailor specific personalized interventions to the individual's own environmental profile. As a pioneering example of this kind, Orešič and colleagues [10] investigated metabolic profiles of children between birth and type 1 diabetes onset in a large birth cohort, and established that specific metabolic phenotypes, not dependent on human leukocyte antigen (HLA)-associated genetic risk, precede the first autoimmune response. The excitement of this research is the expectation that these early metabolic phenotypes may be validated as specific diagnostic and prognostic markers of disease, with therapeutic implications.

#### **Establishing disease causality as a framework for data integration**

The goal of inferring disease causality and disease mechanisms from integrated data is complicated by the fact that measuring more variables may provide a better characterization of the process but still does not contribute directly to our understanding of cause and effect. In fact, given the progressively increasing number of variables that we can measure, the odds of finding spurious associations that do not reflect true causality are much higher. Confounding and reverse causality are among the main sources of bias for failures to replicate apparently robust associations between risk factors and diseases [31]. Confounding specifically refers to a spurious causal effect inferred from the association between a risk factor and a disease due to the existence of some common causes, that is, confounding factors to both of them. This type of spurious causal effect can be removed if we have enough knowledge about the most likely confounding factor candidates. However, the truth is that for most epidemiological studies confounding factors are unknown and difficult to measure, especially in case-control studies. Reverse causality, the second source of bias, refers to an alternative explanation for the observed association between a risk factor and disease, which states that the 'risk factor' is a result of the disease, rather than vice versa. The problem of reverse causality is particularly prevalent in retrospective case-control studies.

One example of a potential confounding association is the established epidemiological evidence of a strong link between obesity and insulin resistance. This association has



**Figure 1**

A causal model based upon Mendelian randomization. The model demonstrates the core assumptions for making a valid causal inference between a phenotype and disease. The three assumptions are: (1) genotype is independent of the confounder; (2) genotype is associated with phenotype; (3) genotype is independent of disease conditioning on phenotype and confounder. If these assumptions are valid, then an observed association between genotype and disease would imply the causality from phenotype to disease.

recently been brought into question from the identification of specific clinical settings where fat mass dissociates from insulin resistance [32,33]. This implies that adipose tissue expansion typically associated with obesity *per se* may not be the cause of metabolic complications. A potential alternative explanation may be related to an individual's ability to optimally store fat. In the presence of caloric excess, a person is likely to remain metabolically healthy despite obesity, provided their adipose tissue can continue to expand and safely store fat [34]. Therefore, while the epidemiological evidence associates the risk of metabolic complication with increased body weight, this relationship may not be direct and may not necessarily reflect a truly biologically relevant process.

A randomized control trial (RCT) is the golden standard for excluding the spurious association that arises from confounding and reverse causality. A RCT involves random allocation of risk factors to subjects, such that distribution of known and unknown confounders in the different groups is roughly equal, that is, the risk factors become disassociated from any confounders due to the randomization. Furthermore, since the initial randomization is done preceding the disease response, this renders reverse causality highly unlikely. However, the use of RCTs to determine causality is often not possible due to enormous ethical, financial or technical difficulties.

An alternative to RCTs could be Mendelian randomization, which has been proposed as a practical strategy to overcome the problem of experimental bias while significantly reducing the difficulties inherent to RCTs [35,36]. The experimental design of Mendelian randomization aims at providing a potential way to discern true causality from spurious associations, provided that several basic assumptions are valid (Figure 1). The idea of Mendelian randomization originated from Katan's letter to *The Lancet* [37], where the main objective was testing the hypothesis that low serum cholesterol increases the risk of cancer versus the

alternative one that the cancer induces a lowering of cholesterol, that is, a hypothesis testing against reverse causality. Using a language of graphical models [38], Mendelian randomization could be formulated in a triangulation representation as shown in Figure 1. The essence of Mendelian randomization is the use of a genetic variant as a proxy for the random assignment of a risk factor to subjects, given that the inheritance of the genetic variant in a population is also random according to Mendel's second law. Mendelian randomization may provide a rational approximation to RCTs that can be used to identify real causal factors contributing to diseases.

### Data integration based upon Mendelian randomization

We envisage that the potential of combining different post-genome approaches for discovering disease causality and mechanisms could be integrated within the framework of Mendelian randomization. In order to apply this idea to distinguish between association and causation, we need to first justify the three core assumptions that underlie the applicability of Mendelian randomization (Figure 1). Two of the three assumptions (1 and 3) depend on unobserved confounding factors and, therefore, cannot be formally tested from observable data. Therefore, the three associations that are needed in the Mendelian randomization model, that is, the genotype-phenotype association, the phenotype-disease association, and the genotype-disease association, require a certain degree of initial characterization. Clearly, these initial models will need to be continually refined as new data challenge the validity of the assumptions. The downstream impact of these assumptions is not trivial, as a failure to detect robust associations could invalidate the power of Mendelian randomization. While this may imply that Mendelian randomization requires our complete understanding of the biological system, in practice some apparent violations may not actually negate its biological implications [36,39]. Applied carefully, Mendelian randomization can become a useful framework for data integration.

In determining truly positive associations in the presence of a large number of variables and relatively few samples, one needs to resort to novel statistical techniques that can handle such complexity. Bayesian statistical methods can be seen as an alternative to conventional hypothesis testing and appear better able to deal with large post-genomics datasets. In contrast to conventional *P*-value-centered statistics, a Bayesian approach provides a measure of the probability of a hypothesis being true by taking all evidence in an explicit way. This is clearly a desirable feature as it allows different forms of data to be combined into a unified hypothetical model. Competing models are then entered into a selection framework such that the hypotheses that are most supported by data are favored. For example, using the language of a causal Bayesian network [40,41], Mendelian randomization can be explicitly represented in the graphical model as

shown in Figure 1; in which the directions of the arrows (or edges) between the nodes indicate non-reversible causal relationships and reflect the three core assumptions made. The plausibility of the graphical model can then be tested through Bayesian rules, with the evidence provided by all available 'omics' data from different studies. A pioneering example of using a Bayesian network to infer disease causality can be found in reference [42], where three possible model networks that characterize the relationships between QTLs, RNA levels and disease traits were evaluated. However, it should be noted that most of the current applications of Bayesian networks consider phenotypes and disease traits as discrete rather than continuous variables; this is due to the computational difficulties of model selection from an extremely large model space.

### Major methodological challenges with complex data integration

While the use of heterogeneous high-dimensional post-genomic data carries many potential benefits, several challenges exist in the areas of biological interpretation, computing and informatics, which will need to be addressed to take full advantage of the wealth of post-genomic data. See Box 1 for the key issues.

### Conclusions

Over the last few years, biomolecular research has progressed from the completion of the human genome project to functional genomics and the application of this knowledge to advance our understanding of health and disease. It is clear that genomic information alone, although crucial, is not sufficient to completely explain disease states, which involve the interaction between genome and environment. Post-genomic approaches attempt to contribute to our understanding of this interaction, with each approach capturing a different angle of the global picture. Intuitively, the next step forward is to integrate these datasets, an approach that, if successful, could be much more informative and predictive than working exclusively on a single platform.

Associating and correlating variables between datasets as a means of integrating the large datasets is wrought with issues such as extracting biological meaning (biology is not always linear and is often context dependent) and determining causality and spurious associations. We propose that data integration should be built upon a model, such as a Bayesian model, that takes into account the non-linearity and context-dependent nature of human biology. We further propose that a putative biological relationship between individual data points, identified through association studies, can be efficiently tested (and validated) using strategies, such as Mendelian randomization, that

#### Box 1.

- No model is perfect, and inevitably assumptions have to be made. It is likely that initial models built around Mendelian randomization will not accurately model f or epistasis, pleiotropy, copy number variants, gene-gene interactions or protein-protein interactions.
- Computational power is becoming a bottleneck when building complex models from heterogeneous high-dimensional data. For example, the inclusion of a single nucleotide polymorphism (SNP) into a model will require large computational power to correct for linkage disequilibrium. Similarly, the more genetic, mRNA, protein or metabolite data are included, the more permutations are present to be built into and cross-validate the models.
- It would be difficult for a single center to generate the complete spectrum of data required for such complex integration. Data from different experimental paradigms and from different populations are required for cross-validation and optimal model selection. As such, datasets generated from different centers need to be standardized in terms of nomenclature and structure. Efforts along these lines can be seen, for example, in transcriptomics [43], proteomics [44] and metabolomics [45].
- A new breed of scientist with a working knowledge of different post-genomic approaches, disease pathophysiology and mathematical modeling will be needed during the initial attempts at data integration. For example, experimental design and subject selections (such as appropriate controls) will need to be tailored to utilize the strengths of each profiling platform and optimize the final dataset for modeling. This needs to be followed by appropriate model interpretation that takes into account all the assumptions and limitations of the experimental and modeling processes. It is likely that such 'integrative' researchers will identify new insights and unexpected limitations during data integration, thus providing an additional element of 'quality control' over the final model.

approximate the design strengths of a RCT. While there are clearly obstacles that need to be overcome, biological models based upon multiple datasets are likely to become the basis that drives future research.

### Abbreviations

GWA, genome-wide association; HLA, human leukocyte antigen; QTL, quantitative trait loci; RCT, randomized controlled trial; SNP, single nucleotide polymorphism.

## Competing interests

The authors declare that they have no competing interests.

## Authors' contributions

All authors contributed equally to this work.

## Authors' information

JT is a postdoctoral researcher in MO's group, focusing on developing applications of Bayesian statistics to integration of heterogeneous genomic and post-genomic data. MO is research professor of systems biology and bioinformatics. His main research areas are metabolomics applications in biomedical research and integrative bioinformatics. CYT is a clinical research fellow in AVP's group, focusing on a systems-biology approach to studying obesity-related metabolic complications. AVP is a reader in metabolic medicine at Cambridge University.

## Acknowledgements

This project was supported by the ATHEROREMO project (FP7-HEALTH-2007-A contract number 201668) funding to MO, HEPADIP project (EU FP6 Contract LSHM-CT-2005-018734) funding to AVP and MO, and MRC-CORD funding to AVP.

## References

- Human Genome Project Information [www.ornl.gov/sci/techresources/Human\_Genome/home.shtml].
- Collins FS, Morgan M, Patrinos A: **The Human Genome Project: lessons from large-scale biology.** *Science* 2003, **300**:286-290.
- Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor Miklos GL, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, et al.: **The sequence of the human genome.** *Science* 2001, **291**:1304-1351.
- Edwards AV, White MY, Cordwell SJ: **The role of proteomics in clinical cardiovascular biomarker discovery.** *Mol Cell Proteomics* 2008, **7**:1824-1837.
- Giovane A, Balestrieri A, Napoli C: **New insights into cardiovascular and lipid metabolomics.** *J Cell Biochem* 2008, **105**:648-654.
- Blakemore AI, Froguel P: **Is obesity our genetic legacy?** *J Clin Endocrinol Metab* 2008, **93**(11 Suppl 1):S51-56.
- Pietilainen KH, Sysi-Aho M, Rissanen A, Seppänen-Laakso T, Yki-Jarvinen H, Kaprio J, Oresic M: **Acquired obesity is associated with changes in the serum lipidomic profile independent of genetic effects: a monozygotic twin study.** *PLoS ONE* 2007, **2**:e218.
- Chen X, Hess S: **Adipose proteome analysis: focus on mediators of insulin resistance.** *Expert Rev Proteomics* 2008, **5**:827-839.
- Bougueres P, Valleron AJ: **Causes of early-onset type 1 diabetes: toward data-driven environmental approaches.** *J Exp Med* 2008, **205**:2953-2957.
- Orešić M, Simell S, Sysi-Aho M, Nantö-Salonen K, Seppänen-Laakso T, Parikka V, Katajamaa M, Hekkala A, Mattila I, Keskinen P, Yetukuri L, Reinikainen A, Lähde J, Suortti T, Hakalax J, Simell T, Hyöty H, Veijola R, Ilonen J, Lahesmaa R, Knip M, Simell O: **Dysregulation of lipid and amino acid metabolism precedes islet autoimmunity in children who later progress to type 1 diabetes.** *J Exp Med* 2008, **205**:2975-2984.
- Frayling TM: **Genome-wide association studies provide new insights into type 2 diabetes aetiology.** *Nat Rev Genet* 2007, **8**:657-662.
- Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 2007, **447**:661-678.
- Su LF: **Updates on high-throughput molecular profiling for the study of rheumatoid arthritis.** *Isr Med Assoc J* 2008, **10**:307-309.
- Quintana FJ, Farez MF, Weiner HL: **Systems biology approaches for the study of multiple sclerosis.** *J Cell Mol Med* 2008, **12**:1087-1093.
- Papassotiropoulos A, Wollmer MA, Tsolaki M, Brunner F, Molyva D, Lutjohann D, Nitsch RM, Hock C: **A cluster of cholesterol-related genes confers susceptibility for Alzheimer's disease.** *J Clin Psychiatry* 2005, **66**:940-947.
- Cornier MA, Dabelea D, Hernandez TL, Lindstrom RC, Steig AJ, Stob NR, Van Pelt RE, Wang H, Eckel RH: **The metabolic syndrome.** *Endocr Rev* 2008, **29**:777-822.
- Lindgren CM, McCarthy MI: **Mechanisms of disease: genetic insights into the etiology of type 2 diabetes and obesity.** *Nat Clin Pract End Met* 2008, **4**:156-163.
- Hegele RA: **Plasma lipoproteins: genetic influences and clinical implications.** *Nat Rev Genet* 2009, **10**:109-121.
- Perry JR, Frayling TM: **New gene variants alter type 2 diabetes risk predominantly through reduced beta-cell function.** *Curr Opin Clin Nutr Metab Care* 2008, **11**:371-377.
- Li S, Loos RJ: **Progress in the genetics of common obesity: size matters.** *Curr Opin Lipidol* 2008, **19**:113-121.
- Loos RJ, Lindgren CM, Li S, Wheeler E, Zhao JH, Prokopenko I, Inouye M, Freathy RM, Attwood AP, Beckmann JS, Berndt SI, Prostate, Lung, Colorectal, and Ovarian (PLCO) Cancer Screening Trial, Jacobs KB, Chanock SJ, Hayes RB, Bergmann S, Bennett AJ, Bingham SA, Bochud M, Brown M, Cauchi S, Connell JM, Cooper C, Smith GD, Day I, Dina C, De S, Dermitzakis ET, Doney AS, Elliott KS, et al.: **Common variants near MC4R are associated with fat mass, weight and risk of obesity.** *Nat Genet* 2008, **40**:768-775.
- Scuteri A, Sanna S, Chen WM, Uda M, Albai G, Strait J, Najjar S, Nagaraja R, Orrù M, Usala G, Dei M, Lai S, Maschio A, Busonero F, Mulas A, Ehret GB, Fink AA, Weder AB, Cooper RS, Galan P, Chakravarti A, Schlessinger D, Cao A, Lakatta E, Abecasis GR: **Genome-wide association scan shows genetic variants in the FTO gene are associated with obesity-related traits.** *PLoS Genet* 2007, **3**:e115.
- Maes HH, Neale MC, Eaves LJ: **Genetic and environmental factors in relative body weight and human adiposity.** *Behav Genet* 1997, **27**:325-351.
- Maher B: **Personal genomes: The case of the missing heritability.** *Nature* 2008, **456**:18-21.
- Zhu J, Wiener MC, Zhang C, Fridman A, Minch E, Lum PY, Sachs JR, Schadt EE: **Increasing the power to detect causal associations by combining genotypic and expression data in segregating populations.** *PLoS Comput Biol* 2007, **3**:e69.
- Raamsdonk LM, Teusink B, Broadhurst D, Zhang N, Hayes A, Walsh MC, Berden JA, Brindle KM, Kell DB, Rowland JJ, Westerhoff HV, van Dam K, Oliver SG: **A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations.** *Nat Biotechnol* 2001, **19**:45-50.
- Gieger C, Geistlinger L, Altmaier E, Hrabé de Angelis M, Kronenberg F, Meitinger T, Mewes HW, Wichmann HE, Weinberger KM, Adamski J, Illig T, Suhre K: **Genetics meets metabolomics: a genome-wide association study of metabolite profiles in human serum.** *PLoS Genet* 2008, **4**:e1000282.
- Stylianou IM, Affourtit JP, Shockley KR, Wilpan RY, Abdi FA, Bhardwaj S, Rollins J, Churchill GA, Paigen B: **Applying gene expression, proteomics and single-nucleotide polymorphism analysis for complex trait gene identification.** *Genetics* 2008, **178**:1795-1805.
- Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, Gordon JL: **The human microbiome project.** *Nature* 2007, **449**:804-810.
- Turnbaugh PJ, Gordon JL: **An invitation to the marriage of metagenomics and metabolomics.** *Cell* 2008, **134**:708-713.
- Lawlor DA, Hart CL, Hole DJ, Davey Smith G: **Reverse causality and confounding and the associations of overweight and obesity with mortality.** *Obesity* 2006, **14**:2294-2304.
- Garg A: **Acquired and inherited lipodystrophies.** *N Engl J Med* 2004, **350**:1220-1234.
- Wildman RP, Muntner P, Reynolds K, McGinn AP, Rajpathak S, Wylie-Rosett J, Sowers MR: **The obese without cardiometabolic risk factor clustering and the normal weight with cardiometabolic risk factor clustering: prevalence and correlates of 2 phenotypes among the US population (NHANES 1999-2004).** *Arch Intern Med* 2008, **168**:1617-1624.
- Virtue S, Vidal-Puig A: **It's not how fat you are, it's what you do with it that counts.** *PLoS Biol* 2008, **6**:e237.

35. Ebrahim S, Davey Smith G: **Mendelian randomization: can genetic epidemiology help redress the failures of observational epidemiology?** *Hum Genet* 2008, **123**:15-33.
36. Sheehan NA, Didelez V, Burton PR, Tobin MD: **Mendelian randomization and causal inference in observational epidemiology.** *PLoS Med* 2008, **5**:e177.
37. Katan MB: **Apolipoprotein E isoforms, serum cholesterol, and cancer.** *Lancet* 1986, **1**:507-508.
38. Jordan M (ed.): *Learning in Graphical Models*. Cambridge, MA: The MIT Press; 1999.
39. Didelez V, Sheehan N: **Mendelian randomization as an instrumental variable approach to causal inference.** *Stat Methods Med Res* 2007, **16**:309-330.
40. Pearl J: *Causality: Models, Reasoning, and Inference*. Cambridge, UK: Cambridge University Press; 2000.
41. Williamson J: **Foundations for Bayesian networks.** In *Foundations of Bayesianism*. Edited by Corfield D, Williamson J. Dordrecht: Kluwer Academic Publishers; 2001:71-115.
42. Schadt EE, Lamb J, Yang X, Zhu J, Edwards S, Guhathakurta D, Sieberts SK, Monks S, Reitman M, Zhang C, Lum PY, Leonardson A, Thieringer R, Metzger JM, Yang L, Castle J, Zhu H, Kash SF, Drake TA, Sachs A, Lusk AJ: **An integrative genomics approach to infer causal associations between gene expression and disease.** *Nat Genet* 2005, **37**:710-717.
43. **Minimum Information About a Microarray Experiment - MIAME** [[www.mged.org/Workgroups/MIAME/miame.html](http://www.mged.org/Workgroups/MIAME/miame.html)].
44. **The HUPO Proteomics Standards Initiative** [[www.psdev.info/](http://www.psdev.info/)].
45. **The Metabolomics Standards Initiative (MSI)** [<http://msi-workgroups.sourceforge.net/>].