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 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.

**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 for 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.

**Abbreviations**

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