Editorial

Recent technological advances have increased drastically the number of experiments or assays one can perform with a small team of researchers and a manageable set of instruments. The resulting reduction in time and cost makes it feasible to perform large-scale studies. While researchers can produce enormous amounts of data over a short period of time with funding from an average research grant, extracting useful information from the data is a challenge, the difficulty of which most large-scale studies underestimate. Usually understaffed, and with an inadequate budget, the analysis teams often lack the proper analytical tools, yet have the unenviable task of drawing conclusions from data without a full understanding of the conditions under which the data were obtained or the tools used today. Unless we pay more attention to data analysis, we will soon drown in a sea of conflicting data and miss the biological pattern and information hidden within them.

How to extract signal out of noise is a problem that physicists and engineers have been grappling with for a long time. The fact that a large group of people at the airport can communicate with specific individuals around the world by mobile phones and wireless internet connections is proof that the engineers have solved the problem of extracting good signal out of noise in the area of global communications. The biological world, however, especially in the study of humans, is a lot more complicated. Unlike the world of communications, where one is extracting the signal from a narrow band of frequencies in real time, human studies are mostly performed with data generated from subjects on one or, at most, a handful of occasions. To complicate matters even further, these 'snapshots' are usually taken without much consideration for the environment in which the subjects find themselves.

The seriousness of the data analysis problem is seen in all three of the areas on which this issue of *Human Genomics* focuses. In human genomics, the reference human genome sequence will take years to fully annotate. In fact, there are still a significant number of places in the reference sequence where mistakes in sequence assembly are found. These mistakes were made because the automated assembly software could not handle low copy duplications in the genome, especially when they were very close to each other. The misassembled sequences can only be corrected when experts carefully analyse the genome sequence in their regions of interest. Similarly, as single nucleotide polymorphisms (SNPs) in their hundreds of thousands are being genotyped for genetic association studies, confounding characteristics such as differences in population substructure between cases and controls must be taken into account, or spurious associations will result or true associations may be missed.

In proteomics, advances in protein separation and mass spectrometry make it possible to obtain protein profiles of biological materials in terms of both the specific proteins present and their relative abundance. Even with a good catalogue of proteins found in a specimen, however, the fact that most tissue specimens consist of a mixture of different types of cells, and that they are obtained under slightly different physiological conditions, creates a great deal of additional noise in the system. It is no surprise, therefore, that proteomic data can only be analysed in a qualitative and superficial way.

In gene expression profiling, probably the most mature form of large-scale studies in the genomics field, intriguing patterns of expressions are seen when one compares different tissues. Once again, heterogeneity in the tissues studied and differences in the conditions under which the samples are obtained affect the gene expression patterns in significant ways. Even so, in some cases, one can predict the prognosis of a patient by looking at the expression pattern of cancerous tissue. There is, however, still great uncertainty in the predictive value of gene expression profiling as a diagnostic tool. Much like the case of proteomics, these patterns will not lead to a deeper understanding of the biological pathways involved in a disease without careful cell biology studies.

Given these difficulties, how does one take advantage of the amazing technologies available in the fields of genomics and proteomics? I believe that the field must restrain from generating data for its own sake. Instead, one must face the problem head on and take the following actions.

First, one must define the specimens under study with a great deal more detail, so that one is comparing different specimens that are appropriately grouped. For example, instead of labelling DNA samples simply as being from patients having a certain disease or from a group of 'normal controls', one should define the specimens further with as much phenotypic and demographic information as possible, including, at very least, carefully defined clinical diagnoses, key laboratory findings, major clinical features, age at disease onset, sex and ethnic origins of the four grandparents (including their places of birth and self-described ancestry). Instead of labelling tissue samples for RNA or protein studies simply as 'tumour', 'tissue with active disease' or 'normal tissue', one should include not only the phenotypic and demographic information needed for DNA samples, but also information on the conditions under which the tissue was obtained. In some cases, it is important to note whether a specimen is obtained

under fasting conditions or shortly after a meal, in the morning or in the evening, while the person has been at rest or after a period of activity. As one carefully controls for the 'background' of the DNA or tissue specimens, noise is greatly reduced and the resultant signals are more easily identified.

Secondly, the accuracy of the data must be determined by periodic validation of the molecular methods and by inclusion of proper controls in the study. Whether one is performing DNA sequencing, genotyping of genetic markers, determining the global gene expression patterns of a tissue or profiling the protein content of a cell type, quality control must be done consistently throughout the course of a study. Knowing the degree of uncertainty in the data, and taking this into account during data analysis, enhances the power of the analysis and strengthens the conclusions derived from the results. Testing duplicate samples at various time points during the study, repeating a subset of experiments using a different platform and looking for consistency of the data based on family or other sample relationships are some of the ways one can determine the quality of the data.

Thirdly, it is essential to develop analytical tools that are appropriate for the data, so that they can be applied properly. In many cases, standard statistical methods cannot be used for genomic, gene expression or proteomic data obtained from a heterogeneous group of individuals. Because biological data are so complex, and one cannot control with great precision the conditions under which the samples are obtained, the assumptions of standard statistical tools regarding the data properties cannot be met. Without proper understanding of the conditions required for a statistical method to be valid, one can apply the wrong test for a dataset and obtain erroneous results. Biostatisticians are becoming more familiar with the explosion of genomic, gene expression and proteomic data being produced and, in due time, an appropriate set of analytical methods will be available for all to use.

Until these practices are standard in the field, one has to take the results and conclusions of large-scale studies with a healthy dose of scepticism. Journal reviewers must insist that those who want to publish the results of genomic, gene expression or proteomic studies address the questions of phenotypic, population and specimen heterogeneity, data quality and the rationale for their choice of analytical method for their data. When these issues are properly addressed, the quality of publications in our field will drastically improve, the number of studies showing conflicting results will decrease and the day when we will decipher the mysteries of the biological patterns contained in our genome and proteome will arrive much sooner.

P.Y. Kwok
Editorial Board Member
Human Genomics