PERSPECTIVE

Learning to share

Genomics can provide powerful tools against cancer — but only once clinical information can be made broadly available, says **John Quackenbush**.

The unveiling of the draft sequence of the human genome in 2000 was met with enthusiastic predictions about how genomics would dramatically change the treatment of diseases such as cancer. The years since have brought a 100,000-fold drop in the cost of sequencing a human genome (to just a few thousand US dollars), and the time needed to sequence it has been cut from months to little more than a day. Researchers can therefore now generate unprecedented quantities of data to help in the battle against cancer (see page S66).

So far, however, our expanded data-generation capacity has not transformed medicine or our understanding of the disease to the degree that some expected. A major contributor to this disappointing outcome has been the failure to deal effectively with the problem of capturing and sharing appropriate clinical data on large collections of samples.

The ultimate goal of cancer researchers is to deliver actionable point-of-care information to doctors treating patients. This means, for example, producing easy-to-read reports that detail the associations between a patient's disease state and their probable response to available therapeutics — associations that are defined by a variety of clinical and genomic attributes and that should be supported by a large, well-curated knowledge base. This information can then help doctors to make rapid decisions about which course of therapy is likely to work best for each patient.

PUBLICLY AVAILABLE DATA SETS **RARELY INCLUDE** THE RIGHT CLINICAL INFORMATION TO **DEFINE** APPROPRIATE COHORTS OR TEST THE **RELEVANCE** OF A GENOMIC SIGNATURE.

Research has already established associations between a few gene variants or gene-expression profiles and clinical endpoints such as drug response. But given the ability to generate large-scale genomic profiling data, they have identified many fewer variants than might have been expected. This shortfall can be attributed to failings in current clinical-research paradigms.

The fundamental design of most clinical and translational-research studies involves comparisons between well-defined patient cohorts. Researchers may divide patients into groups on the basis of outcome — for example, response to a therapy — and ask whether there are genomic features such as mutations or patterns of gene expression that can robustly distinguish between responders and non-responders. Or they can define patient groups according to genomic status and then ask whether there are meaningful differences in some relevant endpoint, such as survival. Cancer research has produced thousands of such genomic studies, with data on hundreds of thousands of patients. But very few of the published studies have been thoroughly validated and fewer still have proved clinically useful.

Although researchers have rushed to generate genomic data, that alone is not sufficient to advance the field. One challenge is to develop analytical methods that are effective for huge amounts of genomic data. In particular, better methods are needed to 'normalize' the data generated by different technologies or at different sites, so that results can be compared across studies — a problem that may seem trivial but that nevertheless has defied a general solution. Methods are also needed to synthesize different types of information more effectively to make predictions, including ways to model the complex interacting networks of factors that drive disease. And standards must be developed to support reproducible research, facilitating validation of the results of any single study in the context of a collective body of data.

But the greatest barrier to the use of 'big data' in biomedical research is not one of methodology. It is, rather, the lack of uniform, anonymized clinical data about the patients whose samples are being analysed. Without such data, even defining experimental cohorts is difficult, and there is a risk of missing potentially obvious confounding factors. Unfortunately, nearly every published study lacks the clinical data to address fundamental research questions fully or to allow the findings of one study to be validated in others.

The first step towards solving this problem is to develop more flexible patient-consent procedures so as to allow the broad use of anonymized clinical data in research. This is particularly important because, at the start of a study, researchers may not know which variables could be important for defining a relevant cohort or could turn out to be confounding an analysis.

The second step is to develop hospital and laboratory computational-infrastructure and data-security protocols to improve the sharing, access and fair use of clinical data. A major barrier to reproducing results is that publicly available data sets rarely include the right clinical information to define appropriate cohorts or test the relevance of a genomic signature.

And, finally, the culture of data sharing must change. Although the publication of the results of genomics studies generally requires the sharing of genomic data, the sharing of clinical data is frequently limited to a bare minimum: just the details described in a manuscript. Even common clinical variables, such as patients' sex, treatment history, smoking history, ethnicity or even standard disease subtype are often not provided. The absence of such key information again makes it difficult to reproduce the results of an analysis or to validate other published data sets.

Big data has tremendous potential to provide fresh insights into diseases such as cancer. But that potential will be realized only by tackling how best to share the clinical information necessary to interpret it. And developing a more complete understanding is essential if we are ultimately to create the knowledge base necessary to provide clear, concise, reliable and actionable information to doctors and their patients. ■

John Quackenbush *is director of the Center for Cancer Computational Biology at the Dana-Farber Cancer Institute in Boston, Massachusetts.*
e-mail: johnq@jimmy.harvard.edu