nature medicine

Genome privacy

In December, 2005, The Cancer Genome Atlas (TCGA) was launched to systematically identify the genetic, genomic and epigenetic alterations associated with cancer. With a joint commitment of US$50 million each from the US National Cancer Institute and the National Human Genome Research Institute to an initial pilot project, and an anticipated need of US$1 billion for a full-scale effort, the rewards of data-mining are expected to outweigh the enormous cost. But logistical issues abound, and money is a poor substitute for scissors when it comes to red tape.

The goals of the project—improved detection, treatment and prevention of cancer—are laudable. The three-year pilot will characterize mutations in genes in at least two types of tumors in order to assess the feasibility of the full-scale project, which would analyze all the major types of cancer. Whereas the technological hurdles of high-throughput sequencing and development of epigenetic and genomic analyses are high, they are straightforward in comparison to those pertaining to tissue acquisition, use and public dissemination of genetic information (see p. 747).

Ideally, for high-throughput analysis and cross-comparison among samples, tissue harvesting and processing should be standardized using methods compatible with genomic and proteomic protocols. The effect of collection methods and processing on assay variability has yet to be determined. But universal protocols for handling different tissues are not presently feasible, and sample preparation is usually determined by pathologists' needs. Assuming an investment of research into best methods of tissue preservation for different applications, future technological advances will allow greater use of archived material. But because of the variability in existing samples, TCGA pilot must be confined to tumors for which many samples at the same stage exist (250–500 per tumor), in the hopes of generating statistically robust mutation association data. A pre-pilot Tumor Sequencing Project is now underway at three sequencing centers to study lung adenocarcinomas and will provide a measure of the impact of sample heterogeneity.

But reliance on archived samples is problematic for reasons beyond sample quality and variability. In the US, the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule, implemented in 2003, regulates the use of health information in order to ensure patient privacy. It requires the individual's informed consent and authorization for the use of protected health information for specific research purposes. Violations of the rule carry fines and the possibility of criminal charges. And although no fines have been levied, as a federally sponsored initiative TCGA is not in a position to transgress the law. Yet many samples in tissue banks were collected with insufficient consent to allow their use without contacting the patients anew, an avenue that might be impossible in many instances and infeasible on the whole. Reconsenting individuals also introduces an unintentional bias against highly lethal tumors. Hopefully TCGA can obtain an exemption to HIPAA to permit use of coded samples from which all patient-identifying information has been removed, without requiring further authorization from individuals for the public release of their genetic information. Otherwise, the pilot project may be severely hampered.

With the appropriate authorization, deidentified data would allow researchers to follow patients prospectively, adding further value to the sequence information. But because the ultimate aim of TCGA is to generate entire genome sequences of individual tumors linked to personal medical information, coded data could eventually be traced to patients. HIPAA does not presently include genotype in its list of 18 identifiers of an individual and therefore does not explicitly regulate its use. Securing the privacy of health information must be weighed against the value of deidentified samples versus anonymized samples, which lack any code to link the sample to the individual and would prevent future outcome analyses. But in comparing the options, science is not an objective party and individuals must be informed of the potential risks of posting genetic data.

In the UK, the Sanger Institute's Cancer Genome Project (CGP) analyzes primary tumors and cell lines, as well as existing literature to identify small somatic mutations in genes that may have a role in cancer. The goal is to eventually screen all genes in the human genome, and progress is posted online each month, with 28,859 mutations reported thus far. As entire genome sequences are not reported, the information is less amenable to misuse. Moreover, by initially focusing on a small set of genes, the CGP may identify important mutations associated with cancer that TCGA will miss due to its significance cutoff level of five percent. Although overlap exists between the projects, the molecular profiles of tumors generated by the two approaches will provide an unparalleled resource for cancer researchers. But in order for TCGA to take the data one step further, the difficulties linking genotype to private medical information must be resolved.