Mapping the Cancer Genome

Pinpointing the genes involved in cancer will help chart a new course across the complex landscape of human malignancies

"If we wish to learn more about cancer, we must now concentrate on the cellular genome." Nobel laureate Renato Dulbecco penned those words more than 20 years ago in one of the earliest public calls for what would become the Human Genome Project. "We are at a turning point," Dulbecco, a pioneering cancer researcher, declared in 1986 in the journal Science. Discoveries in preceding years had made clear that much of the deranged behavior of cancer cells stemmed from damage to their genes and alterations in their functioning. "We have two options," he wrote. "Either try to discover the genes important in malignancy by a piecemeal approach, or . . . sequence the whole genome."

Dulbecco and others in the scientific community grasped that sequencing the human genome, though a monumental achievement itself, would mark just the first step of the quest to fully understand the biology of cancer. With the complete sequence of nucleotide bases in normal human DNA in hand, scientists would then need to classify the wide array of human genes according to their function--which in turn could reveal their roles in cancer. Over the span of two decades Dulbecco's vision has moved from pipe dream to reality. Less than three years after the Human Genome Project's completion, the National Institutes of Health has officially launched the pilot stage of an effort to create a comprehensive catalogue of the genomic changes involved in cancer: The Cancer Genome Atlas (TCGA).

The main reason to pursue this next ambitious venture in large-scale biology with great urgency is cancer's terrible toll on humankind. Every day more than 1,500 Americans die from cancer--about one person every minute. As the U.S. population ages, this rate is expected to rise significantly in the years ahead unless investigators find ways to accelerate the identification of new vulnerabilities within cancerous cells and develop novel strategies for attacking those targets.

Still, however noble the intent, it takes more than a desire to ease human suffering to justify a research enterprise of this magnitude. When applied to the 50 most common types of cancer, this effort could ultimately prove to be the equivalent of more than 10,000 Human Genome Projects in terms of the sheer volume of DNA to be sequenced. The dream must therefore be matched with an ambitious but realistic assessment of the emerging scientific opportunities for waging a smarter war against cancer.

A Disease of Genes

THE IDEA THAT ALTERATIONS to the cellular genome lie at the heart of all forms of cancer is not new. Since the first identification in 1981 of a cancer-promoting version of a human gene, known as an oncogene, scientists have increasingly come to understand that cancer is caused primarily by mutations in specific genes. The damage can be incurred through exposure to toxins or radiation, by faulty DNA repair processes or by errors that occur when DNA is copied prior to cell division. In relatively rare cases, a cancer-predisposing mutation is carried within a gene variant inherited from one's ancestors.

Whatever their origin, these mutations disrupt biological pathways in ways that result in the uncontrolled cell replication, or growth, that is characteristic of cancer as well as other hallmarks of malignancy, such as the ability to invade neighboring tissues and to spread to sites throughout the body. Some mutations may disable genes that normally protect against abnormal cell behavior, whereas others increase the activity of disruptive genes. Most cells must acquire at least several of these alterations before they become transformed into cancer cells--a process that can take years.

Over the past two decades many individual research groups have used groundbreaking molecular biology techniques to search for mutations in genes that are likely candidates for wreaking havoc on normal patterns of cell growth and behavior. This approach has identified about 350 cancer-related genes and yielded many significant insights into this diabolical disease. A database of these changes, called the catalogue of somatic mutations in cancer, or COSMIC, is maintained by Michael Stratton's group at the Wellcome Trust Sanger Institute in Cambridge, England. But no one imagines that it is the complete list.

So does it make sense to continue exploring the genomic basis of cancer at cottage-industry scale when we now possess the means to vastly increase the scope and speed of discovery? In recent years a number of ideas, tools and technologies have emerged and, more importantly, converged in a manner that has convinced many leading minds in the cancer and molecular biology communities that it is time for a systematic, collaborative and comprehensive exploration of the genomics of cancer. The Human Genome Project laid a solid foundation for TCGA by creating a standardized reference sequence of the three billion DNA base pairs in the genome of normal human tissues. Now another initiative is needed to compare the DNA sequences and other physical characteristics of the genomes of normal cells with those of cancerous cells, to identify the major genetic changes that drive the hallmark features of cancer [see box above]. The importance of international partnerships in large-scale

biology to pool resources and speed scientific discovery was also demonstrated by the Human Genome Project, and TCGA is exploring similar collaborations.
Finally, the Human Genome Project spurred significant advances in the technologies used to sequence and analyze genomes. At the start of that project in 1990, for example, the cost of DNA sequencing was more than $10 per "finished" nucleotide base. Today the cost is less than a penny per base and is expected to drop still further with the emergence of innovative sequencing methods [see "Genomes for All," by George M. Church; SCIENTIFIC AMERICAN, January 2006]. Because of these and other technological developments, the large-scale approach embodied in TCGA--unthinkable even a few years ago--has emerged as perhaps the most efficient and cost-effective way to identify the wide array of genomic factors involved in cancer.

Proofs of Concept

PILES OF DATA are, of course, not worth much without evidence that comprehensive knowledge of cancer's molecular origins can actually make a difference in the care of people. Several recent developments have provided proofs of concept that identifying specific genetic changes in cancer cells can indeed point to better ways to diagnose, treat and prevent the disease. They offer encouraging glimpses of what is to come and also demonstrate why the steps toward those rewards are complex, time-consuming and expensive.

In 2001, when the Wellcome Trust Sanger Institute began its own effort to use genomic technologies to explore cancer, the project's immediate goal was to optimize robotics and information management systems in test runs that involved sequencing 20 genes in 378 cancer samples. But the group hit pay dirt a year later when they found that a gene called BRAF was mutated in about 70 percent of the malignant melanoma cases they examined. A variety of researchers swiftly set their sights on this potential new therapeutic target in the most deadly form of skin cancer. They tested multiple approaches--from classic chemical drugs to small interfering ribonucleic acids--in cell lines and in mice, to see if these interventions could block or reduce the activity of B-RAF or inhibit a protein called MEK that is overproduced as a result of B-RAF mutations. Just five years later the most promising of these therapies are being tested in clinical trials.

Other research groups have already zeroed in on genetic mutations linked to certain types of breast cancer, colon cancer, leukemia, lymphoma and additional cancers to develop molecular diagnostics, as well as prognostic tests that can point to an agent in the current arsenal of chemotherapies to which a particular patient is most likely to

respond. Cancer genomics has also helped to directly shape the development and use of some of the newest treatments.

The drug Gleevec, for example, was designed to inhibit an enzyme produced by a mutant fused version of two genes, called BCR-ABL, that causes chronic myelogenous leukemia. Gleevec is proving dramatically effective against that disease and showing value in the treatment of more genetically complex malignancies, such as gastrointestinal stromal tumor and several other relatively rare cancers that involve similar enzymes. Herceptin, an agent that targets a cellular signal-receiving protein called HER2, is successful against breast cancers with an abnormal multiplication of the HER2 gene that causes overproduction of the receptor protein.

Strategies for selecting treatments based on specific gene mutations in a patient's cancer are also being tested in studies of the drugs Iressa and Tarceva for lung cancer, as well as Avastin for lung, colon and other cancers. The performance of these new gene-based diagnostics, prognostics and therapeutics is certainly good news, although the list of such interventions remains far shorter than it would be if researchers in academia and the private sector had ready access to the entire atlas of genomic changes that occur in cancer.

A recent study led by investigators at Johns Hopkins University illustrates both the power of large-scale genomics applied to the discovery of cancer genes and the tremendous undertaking a comprehensive cancer genome atlas will be. The group sequenced about 13,000 genes in tumor tissues taken from 11 colorectal cancer patients and 11 breast cancer patients and reported finding potentially significant mutations in nearly 200 different genes. Interestingly, only about a dozen genes had previously been linked to these two types of cancer, and most scientists had generally expected to find just a few more.

Among the major challenges encountered by researchers sequencing cancer cell genomes is the difficulty of distinguishing meaningless mutations in the tumor samples from those that are cancer-related. Somewhat surprisingly, early sequencing studies have also found very little overlap among the genetic mutations present in different types of cancer and even substantial variation in the pattern of genetic mutations among tumor samples from patients with the same type of cancer. Such findings underscore the idea that many different possible combinations of mutations can transform a normal cell into a cancer cell. Therefore, even among patients with cancers of the same body organ or tissue, the genetic profile of each individual's tumor can differ greatly.

To grasp the full scope of what TCGA hopes to achieve, one must consider the complexities identified in such early efforts and imagine extending the work to more than 100 types of cancer. It is enough to give even veterans of the Human Genome Project and seasoned cancer biologists pause. Yet TCGA participants and other scientific pioneers from around the world are forging ahead, because we are convinced that amid the intricacies of the cancer genome may lie the greatest promise for saving the lives of patients.

Although researchers will probably take many years to complete a comprehensive catalogue of all the genomic mutations that cause normal cells to become malignant, findings with the potential to revolutionize cancer treatment are likely to appear well before this compendium is finished, as the proofs of concept have shown. As each new type of cancer is studied and added to TCGA, investigators will gain another rich new set of genomic targets and profiles that can be used to develop more tailored therapies.

Compiling a Colossal Atlas

A PHASED - IN STRATEGY that proved successful at the beginning of the Human Genome Project was to test protocols and technology before scaling up to full DNA sequence "production." Similarly, TCGA is beginning with a pilot project to develop and test the scientific framework needed to ultimately map all the genomic abnormalities involved in cancer.

In 2006 the National Cancer Institute and National Human Genome Research Institute selected the scientific teams and facilities that will participate in this pilot project, along with the cancer types they will begin examining. Over the next three years these two institutes will devote $100 million to compiling an atlas of genomic changes in three tumor types: glioblastomas of the brain, lung cancer and ovarian cancer. These particular cancers were chosen for several reasons, including their value in gauging the feasibility of scaling up this project to a much larger number of cancer types. Indeed, only if this pilot phase achieves its goals will the NIH move forward with a full-fledged project to develop a complete cancer atlas.

The three malignancies that we selected for the pilot collectively account for more than 210,000 cancer cases in the U.S. every year and caused an estimated 191,000 deaths in this country in 2006 alone. Moreover, tumor specimen collections meeting the project's strict scientific, technical and ethical requirements exist for these cancer types. Last September our institutes announced the selection of three biorepositories to provide such specimens, along with new tumor samples as needed, and normal tissue from the same patients for comparison. Those facilities will deliver materials to a central

Biospecimen Core Resource, one of four major structural components in TCGA's pilot project.

Cancer Genome Characterization Centers, Genome Sequencing Centers and a Data Coordinating Center constitute the project's other three main elements [see illustration at right], and all these groups will collaborate and exchange data openly. Specifically, the seven Cancer Genome Characterization Centers will use a variety of technologies to examine the activity levels of genes within tumor samples and to uncover and catalogue so-called large-scale genomic changes that contribute to the development and progression of cancer. Such alterations include chromosome rearrangements, changes in gene copy numbers and epigenetic changes, which are chemical modifications of the DNA strand that can turn gene activity on or off without actually altering the DNA sequence.

Genes and other chromosomal areas of interest identified by the Cancer Genome Characterization Centers will become targets for sequencing by the three Genome Sequencing Centers. In addition, families of genes suspected to be important in cancer, such as those encoding enzymes involved in cell-cycle control known as tyrosine kinases and phosphatases, will be sequenced to identify genetic mutations or other small-scale changes in their DNA code. At present, we estimate that some 2,000 genes--in each of perhaps 1,500 tumor samples--will be sequenced during this pilot project. The exact numbers will, of course, depend on the samples obtained and what is discovered about them by the Cancer Genome Characterization Centers.

Both the sequencing and genome characterization groups, many of which were participants in the Human Genome Project, can expect to encounter a far greater level of complexity than that in the DNA of normal cells. Once cells become cancerous, they are prone to an even greater rate of mutation as their self-control and repair mechanisms fail. The genomic makeup of individual cells can therefore vary dramatically within a single tumor, and the integrated teams will need to develop robust methods for efficiently distinguishing the "signal" of a potentially biologically significant mutation from the "noise" of the high background rate of mutations seen in many tumors. Furthermore, tumors almost always harbor some nonmalignant cells, which can dilute the sample. If the tumor DNA to be sequenced is too heterogeneous, some important mutations may be missed.

Following the lead of the Human Genome Project and other recent medical genomics efforts, all these data will be made swiftly and freely available to the worldwide research community. To further enhance its usefulness to both basic and clinical researchers and, ultimately, health care professionals, TCGA will link its sequence data and genome

analyses with information about observable characteristics of the original tumors and the clinical outcomes of the sample donors. Developing the bioinformatic tools to gather, integrate and analyze those massive amounts of data, while safeguarding the confidentiality of patient information, is therefore another hurdle that must be cleared to turn our vision into reality.

Uncharted Territory

THE ROAD AHEAD is fraught with scientific, technological and policy challenges--some of which are known and others as yet unknown. Among the uncertainties to be resolved: Will new sequencing technologies deliver on their early promise in time to make this effort economically feasible? How quickly can we improve and expand our toolbox for systematically detecting epigenetic changes and other large-scale genomic alterations involved in cancer, especially those associated with metastasis? How can we harness the power of computational biology to create data portals that prove useful to basic biologists, clinical researchers and, eventually, health care professionals on the front lines? How can we balance intellectual-property rights in a way that promotes both basic research and the development of therapies? When will Congress finally pass genetic nondiscrimination legislation so that knowledge gained through TCGA will have the maximum positive influence on Americans' health? The list goes on.
To avoid raising false expectations, we also must be clear about the questions this project will not attempt to answer. Although it will serve as a resource for a broad range of biological exploration, TCGA is only a foundation for the future of cancer research and certainly not the entire house. And we face the sobering issue of time--something that is in short supply for many cancer patients and their families. As we survey the considerable empty spaces that exist in our current map of genomic knowledge about cancer, the prospect of filling those gaps is both exhilarating and daunting. Scientists and the public need to know up front that this unprecedented foray into molecular cartography is going to take years of hard work and creative problem solving by thousands of researchers from many different disciplines.

Where all this work will lead can only be dimly glimpsed today. In this sense, our position is similar to that of the early 19th-century explorers Meriwether Lewis and William Clark. As they ventured up the Missouri River into the largely uncharted Northwest Territory in 1804, their orders from President Thomas Jefferson were to "take observations of latitude and longitude at all remarkable points. . . . Your observations are to be taken with great pains and accuracy; to be entered distinctly and intelligibly for others, as well as yourself."

Although Lewis and Clark did not find the much-longed-for water route across the continent, their detailed maps proved valuable to their fledgling nation in myriad ways that Jefferson could never have imagined. For the sake of all those whose lives have and will be touched by cancer, we can only hope our 21st-century expedition into cancer biology exceeds even Renato Dulbecco's grandest dreams.

~~~~~~~~

By Francis S. Collins and Anna D. Barker
Francis S. Collins and Anna D. Barker are leaders of The Cancer Genome Atlas initiative in their positions as, respectively, director of the National Human Genome Research Institute and deputy director for Advanced Technologies and Strategic Partnerships of the National Cancer Institute. Collins led the Human Genome Project to its completion of the human DNA sequence, and Barker has headed drug development and biotechnology research efforts in the public and private sectors, with a particular focus on fighting cancer.
~~~~~~~~