So Many Choices, So Little Money

With the human genome almost finished, leaders of the project are trying to decide where next to focus their energies. They're debating a plethora of options, from more sequencing to proteomics—some of which could break the bank

Just months after they pulled off one of the greatest feats in biology—deciphering the human genome—the leaders of that international effort find themselves in a quandary: What to do next?

When biologists mounted a megaproject to sequence the human genome a dozen or so years ago, the goal was clear, even if how to reach it was not. As James Watson of Cold Spring Harbor Laboratory in New York state, who led the U.S. Human Genome Project (HGP) in its early days, put it, that goal was to find out what makes us human. The effort promised to transform 21st century biology, opening up avenues of research previously impenetrable. The allure was undeniable, and the public—and Congress with its deep pockets—embraced the idea.

But now that most of the 3 billion bases in our 24 chromosomes are in order, the goal is considerably more amorphous—and harder to sell. Should the huge centers created for the task now go on to sequence the genomes of other organisms? Devote themselves to discerning the functions of the estimated 45,000 genes? Plunge into a "Proteome Project" that would identify and characterize all proteins? Develop new technologies to speed analysis? Or, when the researchers finish the genome, should they pack up their bags and go home, bequeathing their fantastic new tool—the sequence—to the biological community?

Packing up their bags is not an option. A massive infrastructure is already in place. The National Institutes of Health (NIH), for instance, created a new institute specifically for the job, funded to the tune of $1.4 billion over the past 10 years. The Wellcome Trust, a British biomedical charity, put up more than $300 million just for the HGP, and numerous other countries contributed in kind, helping make it the biggest biology project ever. Equally compelling, with the genome sequence almost in hand, a wealth of scientific questions are waiting to be answered. But no one is quite sure about the best way to tackle them.

"I can tell you what we plan for the next couple of years. But further out, it is more difficult to predict," says Michael Morgan of the Wellcome Trust. "A lot of the ideas are out there, and a lot of people are discussing them in a lot of different venues," says Gerald Rubin, science director of the Howard Hughes Medical Institute in Bethesda, Maryland.

And, for the most part, leaders of the public project are continuing to think big. Indeed, a number of the proposed projects—such as analyzing hundreds, even thousands, of genes, proteins, or protein interactions at once—potentially rival and perhaps even surpass the HGP in scope. And although big-ticket pilot projects have been started, none has the immediacy or finiteness of sequencing the human genome. "There's a big issue of how we continue to capture the public's imagination," concedes William Gelbart, a developmental geneticist at Harvard University. If the public loses interest, Gelbart and other academic biologists fear, too much of the work will take place behind the closed doors of industry.

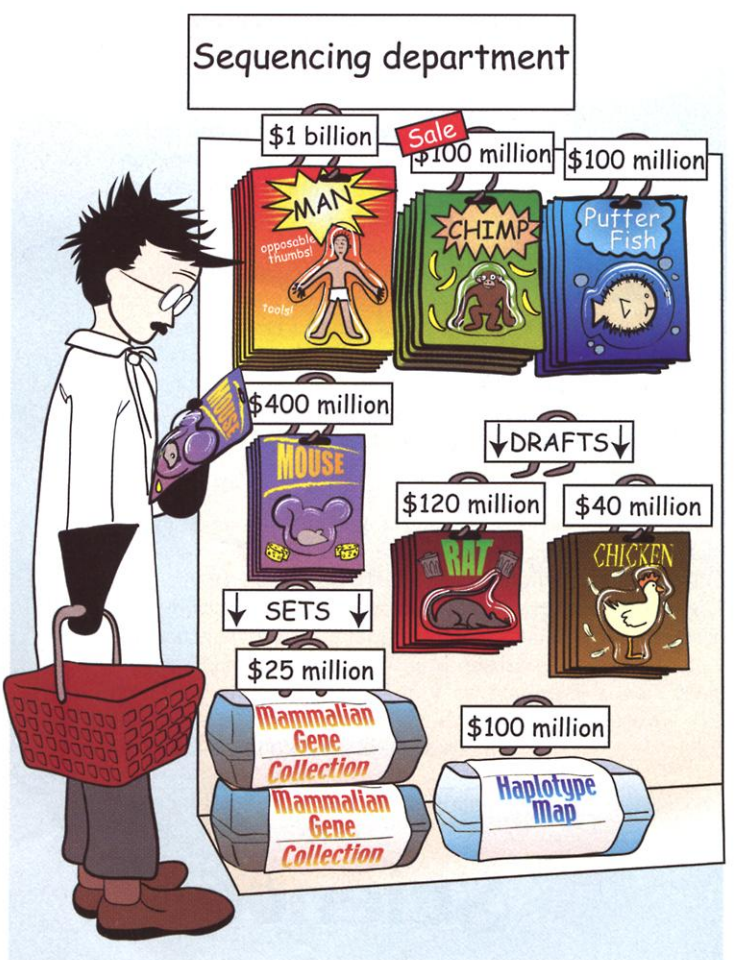


ILLUSTRATION: C. SLAYDEN

Job one: Sequence

Everyone agrees that the first job for the international HGP is to finish at least one vertebrate genome—preferably the human genome. The goal is to complete this task by 2003.

Also high on the list is the sequencing of other genomes. Comparing those new genomes to the human genome will help researchers pinpoint hard-to-find genes and DNA regulatory regions and reveal secrets of evolutionary history. For that reason, "large-scale sequencing will continue to be a major activity for the next five, if not 10, years," says Francis Collins, director of the National Human Genome Research Institute (NHGRI), which funds a large portion of the U.S. share of the HGP.

The mouse, rat, zebrafish, and puffer fish genomes are already under way, although none is near completion. The mouse should be sequenced in great detail by 2005 at the latest. For the others, the research community and the funding agencies are debating how thoroughly to sequence them. Will two or three passes over the genome do? Or is more in-depth coverage needed to make sense of the sequence and render it reliable?

The Sanger Centre in Hinxton, U.K., plans to expand its sequencing efforts beyond the human, mouse, zebrafish, and microbes that it's working on now. At the 10th International Strategy Meeting on Human Genome Sequencing in Hangzhou, China, in September, Chinese and European researchers agreed that they should set up a joint project to sequence the chicken genome (*Science*, 7 September, p. 1745), with the Sanger

Centre possibly taking a lead role. China and other European collaborators have begun work on the pig genome, while Japanese and German collaborators are pushing forward on the chimp. "Having a zoo of genomes is going to make a tremendous difference," says David Haussler, a computer scientist at the University of California, Santa Cruz.

Moving beyond the half-dozen organisms already in the pipeline, other researchers are clamoring to have their favorite sequenced. Setting priorities will be key, says Collins, because sequencing funds and capacity cannot accommodate too many more species unless sequencing costs decrease substantially. At a July workshop, NHGRI-invited experts set ground rules for selection in the United States, at least. Advocates of deciphering, say, the cow genome will need to submit a description of the size of its genome, the organism's place on the evolutionary tree, the ease of performing genetic analyses on it, the size of the community that studies it, and the ways the sequence will aid the interpretation of the human genome.

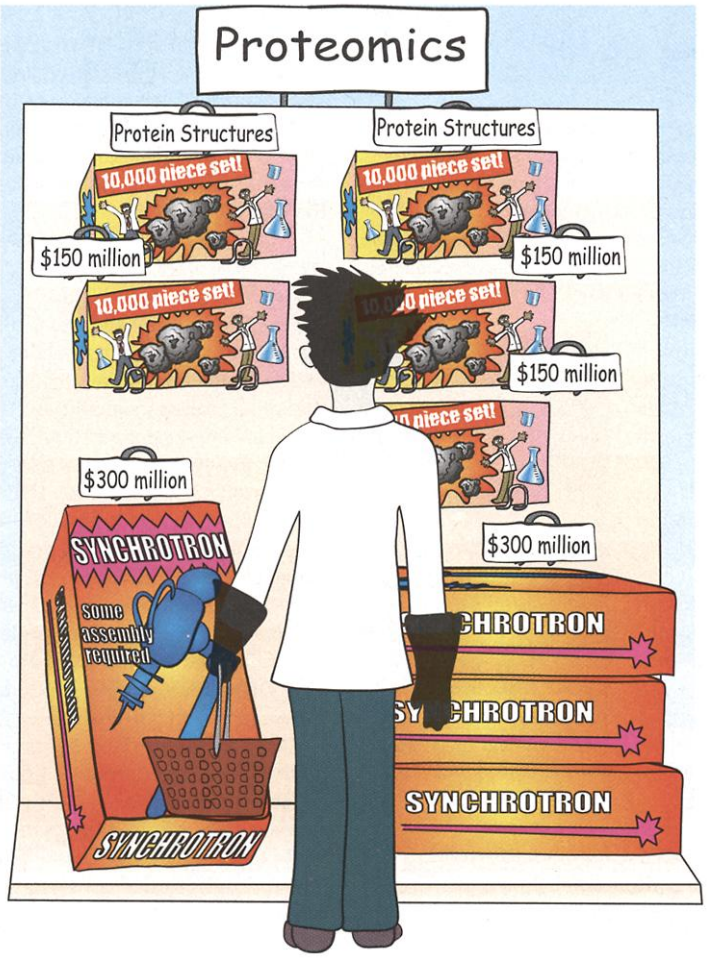


ILLUSTRATION: C. SLAYDEN

Gold-standard genes

In terms of mining the human genome, the top priority is to get "the gold-standard set of human genes," says Haussler, a sentiment shared by both academic and corporate scientists. A number of companies, such as Invitrogen in Carlsbad, California, and several publicly funded groups have projects well under way. But a straightforward task—particularly if there are less than 45,000 genes as predicted—it is not. Already researchers have realized that the one gene–one protein dogma is wrong—and that one gene, by having different combinations of its coding regions converted into what is called an expressed transcript, might specify several proteins. This process of generating multiple transcripts is called alternative splicing.

"There's a huge amount of work that needs to be done to understand the full complement of expressed transcripts," Haussler notes. "It will be tremendously difficult." The first iteration of the gold-standard set of genes will likely contain just the straightforward sequences, or transcripts, from those 40,000-plus genes, with the alternatively spliced transcripts coming much later. Researchers in Japan, Germany, and the United States have independently begun defining the genes, by generating what are called full-length complementary DNAs (cDNAs) that represent the transcript sequence.

Since 1999, NIH has put $25 million into the Mammalian Gene Collection. In this effort, researchers first define the sequence of the cDNA and then warehouse each in bacterial clones for use by researchers. As of September, the collection contained 19,000 putative and 6700 confirmed human cDNAs.

Japanese scientists are well along in generating both mouse and human cDNAs. They started in the mid-1990s developing efficient ways to produce these cDNAs, and by last year they had produced a comprehensive collection of 20,000 mouse cDNAs that are now available to the public. More recently, they have applied that expertise to human genes (see sidebar).

Germany is investing $5 million a year in a European cDNA effort, headquartered at the German Cancer Research Center in Heidelberg. With that money, says group leader Stefan Wiemann, the team will not only generate cDNAs but also pinpoint where the proteins specified by those cDNAs work in the cell. The German project has placed data for several thousand putative cDNAs on the Web.

These groups are now discussing whether they can merge their efforts to get the job done faster, because many cDNAs remain to be discovered. The sticking point, at the moment, is how quickly their work will become available to the broader scientific community—and in what form. "We've had several meetings on coordination issues," notes Wiemann.

Wiemann hopes that his group will get some of the $175 million that Germany committed in March in an effort to boost Germany's profile in genomics. But that is not guaranteed. And whereas Collins says the Mammalian Gene Collection is a "meat-and-potatoes endeavor" that will continue, others worry that it could lose out to costly, glitzier projects. "We need to make the commitment to [finishing] this or we won't have this [essential] gene set," Haussler insists. Without it, adds Robert Strausberg, director of the Cancer Genomics Office at the National Cancer Institute, studies to determine the function of these genes will be undermined: "For the world of genes and proteins, [the collection] is a platform for going forward in proteomics and functional genomics."

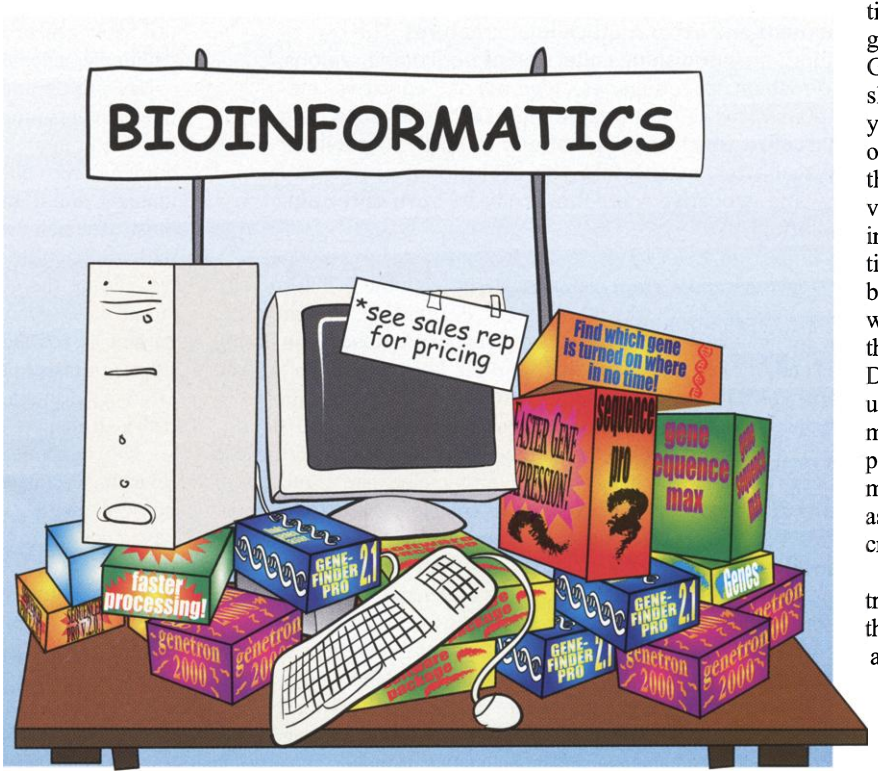
NHGRI and the Wellcome Trust are talking about collaborating on another project—this one to take the next step in studying genetic variation. Even before a draft human genome sequence was finished, a public-private partnership called the SNP Consortium began cataloging single-base differences, called single-nucleotide polymorphisms (SNPs), that pepper the genome. These simple variations, in which one person has a different base at a particular location than someone else has, can help researchers pinpoint particular genes and variations involved in disease. Yet even with 3 million SNPs in the public databases, geneticists have had difficulty using them effectively (*Science*, 27 July, p. 593).

Earlier this year, several groups found that particular sets of SNPs tend to occur in blocks along the genome. By mapping these blocks, or haplotypes, researchers now think they can simplify their searches for genetic defects (*Science*, 27 July, p. 583). "We not only want to know the SNPs individually, but we want to know their neighborhoods [in the

genome]," explains NHGRI's Collins. Although building a haplotype map could cost $100 million over the next several years, enthusiasm is mounting.

Technology development

But other projects with high price tags may be slower to gain momentum, in part because public purses aren't large enough to fund them all, at least at their current costs. Researchers are eager to pursue three broad areas: functional genomics, proteomics, and bioinformatics. Functional genomics aims to understand how genes are regulated and what they do, largely through massive parallel studies of gene expression over time and in a variety of tissues. Proteomics promises to make the identity, function, and structure of each protein known and to elucidate protein-protein interactions. New developments in bioinformatics would enhance the ability of researchers to manipulate, collect, and analyze data more quickly and in new ways. Experts predict that more biologists will do their work in silico, using the computer to synthesize, analyze, and interpret the many terabytes of data now being generated.



None of this comes cheap, however, and in some instances the experiments are not yet possible. "The whole area of proteomics is in a muddle," Collins points out, noting that the goals and even definition of proteomics are unclear. Collins worries that proteomics and the databases and bioinformatics it requires could become "the part of the budget that eats the rest." And although many researchers are incorporating microarrays and DNA chips into efforts to learn when and where genes are turned on, for instance, they lack the more sophisticated, automated approaches needed to study all the proteins those genes encode. "We don't know how to make the measurements [of function and interactions] that are really critical in a high-throughput manner," notes Leroy Hood, director of the Institute for Systems Biology in Seattle.

Until researchers do, embarking on massive endeavors, say, to describe all protein-protein interactions or to characterize the many ways in which a gene can be expressed—so-called alternative splicing—might be foolish. "I think technology development is going to be a key part of the future," Hood says, echoing the arguments he made at the outset of the HGP, when he pushed for faster and cheaper sequencing technology.

The sequencing technology that emerged from the HGP, says Harvard genome researcher George Church, is as important as the sequence itself. He thinks technology development should be a big part of new grants, because that could lead to reduced costs, not only for sequencing but also for efforts in proteomics and functional genomics: "The cost determines what questions you ask." Both Hood and Church are particularly interested, for example, in technology that can monitor and pinpoint multiple protein-protein and protein-gene interactions in living cells. They are looking to NHGRI, as well as other NIH institutes and the Department of Energy (DOE), to help stimulate these pursuits. Nascent strategic plans at NHGRI will likely reflect the importance of technology, says Collins.

Another technology that needs an infusion of funds is scientific computing, says Gene Myers, a bioinformatics expert at Celera Genomics in Rockville, Maryland. "This area is now pushing the envelope of what we are able to do," says Myers. The problem, he says, is not so much computer power but the lack of efficient methods to move large amounts of information around. Genome data can fill thousands of CD-ROMs, and no one knows how to send that much data from one computer to another quickly, or even how to move it around within a single machine, he says. New algorithms for interpreting and analyzing these data are also needed, he and others say, as are ways to visualize and present genomic information to researchers. "The sequence is not in a state where someone like me, a geneticist, can manipulate it," says Mary-Claire King of the University of Washington, Seattle.

DOE is primed to take the lead in scientific computing for biologists. As part of its Genomes to Life program, slated to start in fiscal year (FY) 2002, DOE biologists will team up with their colleagues in advanced scientific computing to build a computational infrastructure combining hardware and software development. With this infrastructure in hand, DOE plans to explore regulatory networks, determine how protein complexes—life's molecular machines—operate, and assess the function of microbial communities.

"Computation is central to what we do, and this is a strength we have at DOE," says Ari Patrinos, who directs biological and environmental research at the agency. Overall, he expects to have $20 million in FY 2002, which will help support "fairly large teams of people working on complex problems."

Even though DOE wants to put large teams to work on these problems, not all genome research and technology development needs to be done that way. "We may not need big science," says Church, who argues that postsequencing goals might be better met with "many labs in which there would be more diversity of machines and more clever engineering—some of which could drive costs down." So, while NHGRI, the Sanger Centre, and funding agencies across the globe are debating what the next focus should be, they are also trying to assess the right mix of "mega" and little science. In the United States, NHGRI will devote the next year to polling biologists from many fields about these questions. The answers it comes up with could shape genomics for decades to come, not just in the United States but around the world.

–ELIZABETH PENNISI

With reporting by Evelyn Strauss, Michael Balter, Robert Koenig, and Dennis Normile.

ILLUSTRATION BY C. SLAYDEN