

What's Next for the Genome Centers?

Like a company that has just performed the first opera in Wagner's four-part epic *Ring Cycle*, the international public sequencing consortium is now hoping it can stay together long enough to complete the opus—while some of its performers go off to stage less gargantuan works.

It is an impressive company, fully capable of putting on multiple productions. Collectively, the entire cast can churn out close to 2000 bases every second—that's 7.2 million per hour, 172 million per day, every day, 365 days a year—counting all the sequencing machines in all the labs responsible for the human genome effort. "What was a month's work a year ago is a week's work now," says Richard Gibbs, director of the sequencing center at Baylor College of Medicine in Houston. That capacity could translate into a first pass at sequencing about two big mammalian genomes a year, if all players turned their machines on the same species.

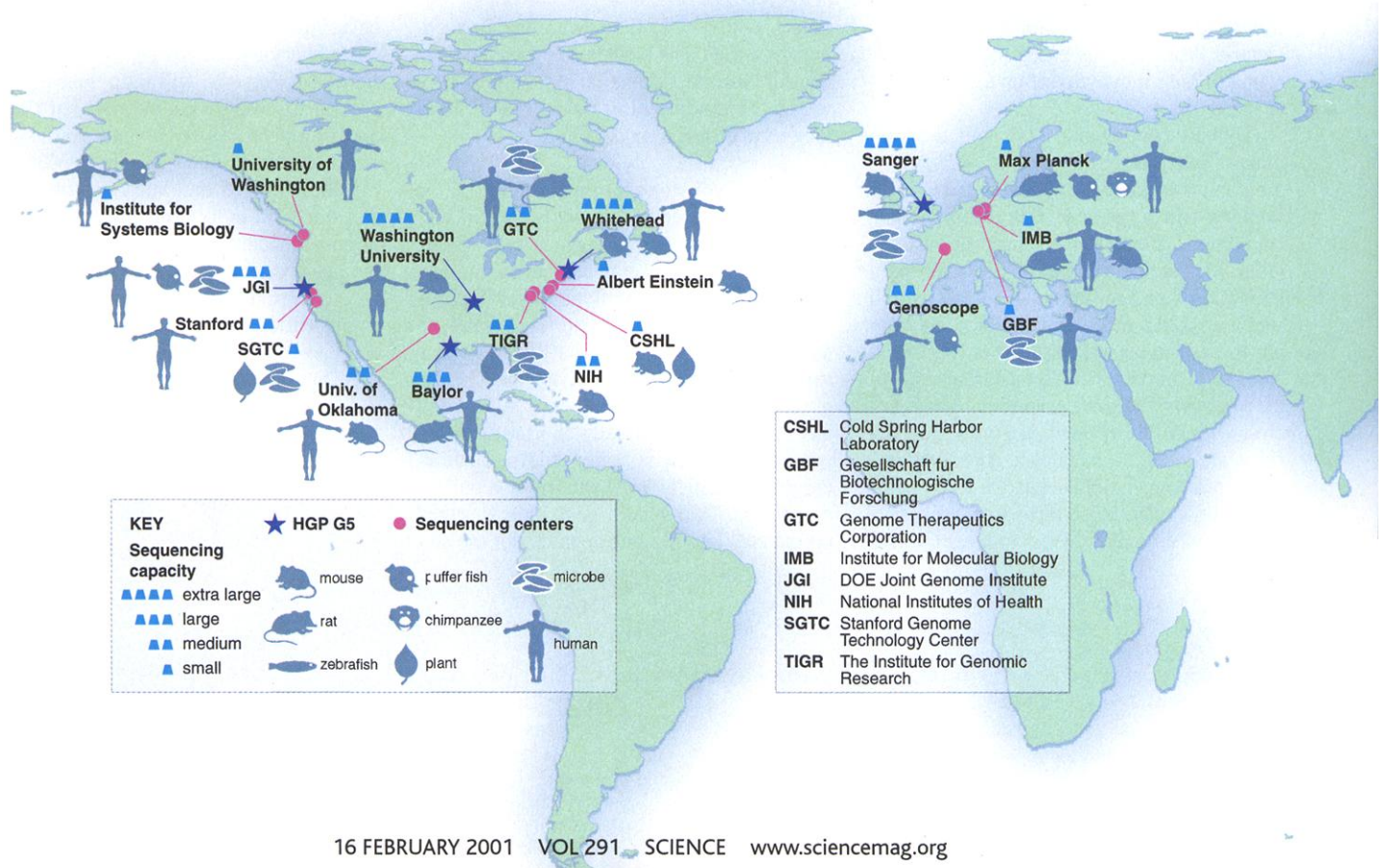
Now that the curtain has closed on the first production, the next order of business is to finish the entire work. Over the next 2 years, most of the G16—the 16 principal players involved in generating the bulk of the public's working draft of the human genome—will finish sequencing their allotted pieces, while their computer-savvy colleagues process and polish that sequence into gleaming, "final" form.

Yet already, several of the G16 have begun to move on, sometimes scaling back their involvement with human DNA and often leaving high-volume genomic sequencing to labs with more sequencing machines and bigger budgets. For many, the goal now is to develop the tools to make sense of all this sequence data and to move into the study of gene function. Meanwhile, mammoth operations such as the Whitehead/MIT Genome Center in Cambridge, Massachusetts, and the Sanger Centre in Hinxton, U.K., continue to

grow, adding more machines to tackle more and more large genomes and expanding the global sequencing capacity to levels unthinkable 5 years ago. The expectation is that each new genome will help bring the human genome into clearer focus. As a result, "the appetite for sequencing continues to go up," notes Francis Collins, director of the National Human Genome Research Institute (NHGRI), which funded much of the U.S. human genome effort.

Job one: The polished draft

As pressure mounted over the past year or so to complete a working draft of the genome, debate intensified about how best to do the next step: finishing the job. Finishing is the "dotting the i's" stage of sequencing, the process that kicks in once each section of DNA is sequenced 8 to 10 times over. That "8x" or "10x" coverage should be more than enough to identify each base with confidence. During finishing, the many stretches of sequence are put in order until they make up one long, virtually continuous series of A's, T's, G's, and C's representing a chromosome from one end, or telomere, to the hard-to-sequence centromere in the middle. A second stretch of sequence covers the rest of the chromosome, starting from the other side of the centromere and ending at the other telomere. Barring a few impossible-to-determine gaps, chromosomes 21 and 22 are now at that



stage in the public draft.

As the adage goes, the last 5% is often as hard as the first 95%. Finishing is so labor-intensive that it's unclear how well the economies of scale realized by the bigger, highly automated sequencing centers would apply to this job. So, while the megacenters are shouldering ever more of the world's sequencing and will do the lion's share of finishing, the public consortium has decided to spread that polishing around. Two smaller groups have made finishing their primary activity. Maynard Olson's team at the University of Washington, Seattle, for example, is trying to automate this phase. Richard Myers's group at the Stanford Human Genome Center has essentially become a finishing arm for the Department of Energy (DOE), taking on chromosomes 5, 19, and part of 16.

Even with these specialists, though, "we really are relying on everybody kicking in and doing this," says Myers. Most groups are completing the sections they started. Like many others, he worries that as the end nears, more and more groups will spin off into other genomic ventures, leaving the human genome incomplete. "It needs to be finished," warns Allan Bradley, director of the Sanger Centre. Even though the overall draft might seem complete, he says, most gene hunters will find just a piece of what they are looking for in GenBank and other databases. The genome centers are "better prepared" than in-

dividuals to fill in the blanks, he adds.

To keep everyone focused on the goal, the consortium is trying to maintain its tight-knit collaboration, with regular conference calls and frequent e-mails among the partners. The finishing experts met at the recent Marco Island genome meeting in February, and in May, the group will huddle at the annual genome meeting at Cold Spring Harbor Laboratory in New York. Collins expects the full shotgun phase of the sequencing to be done possibly as soon as July. As for dotting all the i's, "we'd like to have it all finished to the same standard as chromosomes 21 and 22 by April 25, 2003," the 50th anniversary of the publication of James Watson and Francis Crick's seminal paper on the double helical structure of DNA.

Calling more genomes

Making a finished genome meaningful requires "annotation," in which genes and other features of the genomic landscape are located and described (see p. 1177). Because deciphering the genomes of other species should speed and improve annotation, the big players in the human genome effort have taken on new genomes as well. At the top of their list: the lab mouse, which Celera Genomics, the Rockville, Maryland, company that sequenced the human genome separately from the public effort, has already sequenced to some degree.

As for the public mouse effort, "it's going great guns," Collins says. After NHGRI awarded 10 mouse sequencing and mapping grants 15 months ago, the groups involved were jockeying over how best to tackle the task and how to divide up the work, as it was rapidly becoming possible for a single center to take on the entire genome on its own. But now, an infusion of more than \$30 million in October 2000 from several companies and Britain's Wellcome Trust has turbo-charged the work. With the new money, the Washington University, Whitehead, and Sanger centers bought more high-powered capillary sequencers; they expect to generate a working draft of the mouse genome, covering 95% of the 3 billion bases 2.5 to 3 times over, by March.

In addition to that draft, several smaller centers, such as the Institute of Molecular Biotechnology in Jena, Germany, and the University of Oklahoma, Norman, are producing finished data for biologically important regions of the mouse genome. "The mouse data will be gloriously useful," Collins says. "Ninety-five percent of the human

[genes] can be found if we have the mouse."

A close second for sequencing is the rat genome, already under way at a modest level with NHGRI funds but about to receive a big boost from the U.S. National Heart, Lung, and Blood Institute. The rat is a key model for many diseases, because it's slightly larger than the mouse and its physiology is more amenable to study. Sequencing is under way at Baylor College of Medicine and at Genome Therapeutics Corp. in Waltham, Massachusetts.

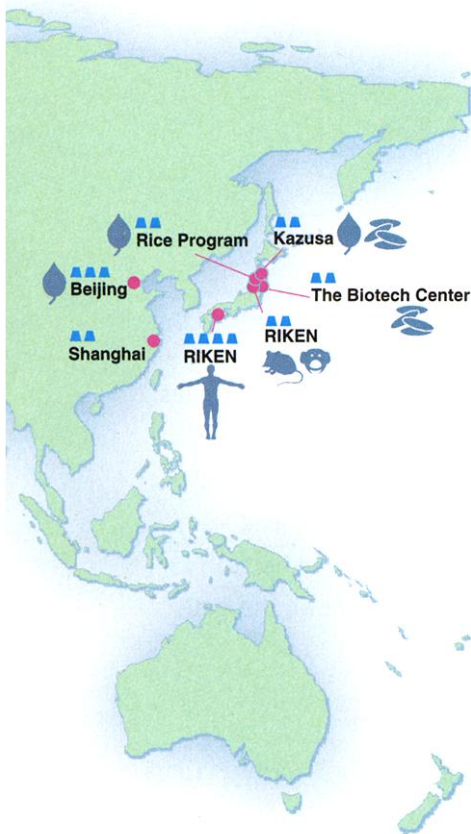
The list continues. China has its eyes on



Principal players. Leaders of the G16 sequencing centers, which did most of the sequencing for the public consortium.

the pig (*Science*, 3 November 2000, p. 913). DOE's Joint Genome Institute (JGI) plans to start on a sea squirt in March, after polishing off a rough draft of the Japanese puffer fish genome, which has about the same complement of genes but much less junk DNA. The whole genome is only 400 million bases, compared to the human and mouse at 3 billion apiece. Several other groups are dogging the genomes of the freshwater puffer fish, and the Max Planck Institute for Molecular Genetics in Berlin has even sequenced a small fraction of the chimpanzee genome in collaboration with RIKEN's Genomic Sciences Center in Japan. "There are many other nominees raising their hands saying, 'Sequence me, please,'" says Collins, who notes that the National Institutes of Health will host a workshop this spring to sort out which to do first. The lobbying is already intense.

These sequencing efforts are the first steps toward what many biologists see as the best way to learn about the human genome: through comparative genomics. Each new genome helps biologists develop a clearer picture of what's important in DNA, as key regions, such as protein-coding exons or binding sites for transcription factors, are conserved to varying degrees among species. Genomicist Eric Green of NHGRI has picked a few key regions, including



CREDIT: (PHOTO) WELLCOME TRUST

those containing the cystic fibrosis gene and the gene for Williams syndrome, to sequence in dog, cat, horse, cow, baboon, and five other vertebrates. Eric Lander's group at Whitehead is sequencing several fungi of known evolutionary distance from bakers' yeast to find out how evolutionary distance affects the comparisons. It turns out two closely related species can be too similar to reveal certain key conserved features.

In all likelihood, the big centers will tackle large genomes on their own in a few years. And that prospect is exciting not only to geneticists and biomedical researchers but also to a broad range of biologists. For instance, evolutionary and developmental biologists would like to decipher the genomes of some 100 species distributed across the evolutionary tree; they have already asked the National Science Foundation, which supported the first sequencing of a plant, for planning support. It won't be long, says Oklahoma's Bruce Roe, before "we have the technology to answer very broad-based questions on how organisms evolved."

21st century biology

As valuable as these genome sequences will be, the sequence by itself doesn't tell researchers what genes do—and that's rapidly becoming the focus of a number of centers, both large and small. National genome budgets are beginning to reflect this emphasis. Already, about 90% of Germany's human genome project budget goes toward functional genomics, says Jörg Wadzack, a molecular biologist for the German Human Genome Project. NHGRI is evaluating proposals for centers of excellence that will also push the U.S. human genome effort in a functional direction. Stanford, for example, is one of several centers helping to build the Mammalian Gene Collection, a set of 25,000 full-length complementary DNAs (cDNAs) for mouse and human. (cDNA includes all the coding regions of a gene.) Already, the Japanese have collected 20,000 full-length mouse cDNAs that are part of a mouse gene encyclopedia (*Science*, 9 February, p. 963). These cDNAs will "help annotate the genome and find genes," says Stanford's Myers. Adds Collins: "This [collection] will be one of the durable goods of the genome project."

At the Sanger Centre as well, "the big emphasis is going to be understanding how

NEW SCIENCE: Hunting for Collaborators Of Killer Toxins

Many gene hunters track sequences that inevitably lead to disease. Environmental health researchers seek a different quarry: variations in genes that by themselves might be harmless, but, when a person is exposed to environmental toxins, can amplify that person's risk of illness. Studies of these genes, which often code for enzymes that metabolize toxins or repair DNA damage from carcinogens, could lead to a better understanding of how the genes make people vulnerable and which individuals are at risk.

The growing list of diseases linked to these environmental susceptibility

genes includes asthma, diabetes, lead poisoning, and a lung disease caused by the metal beryllium. Many disorders involve more than one gene: Variants of two genes, for example, boost the risk of bladder cancer 10-fold in people who smoke, according to



recent studies at the National Institute of Environmental Health Sciences in Research Triangle Park, North Carolina.

To help find these genes and explore how they operate, toxicologists are using

DNA arrays—glass chips dotted with gene sequences (*Science*, 28 July 2000, p. 536). For instance, Leona Samson's team at Harvard University has found that DNA-damaging chemicals known as alkylating agents turn on or off at least 400 genes in yeast cells, including "totally unexpected" genes involved in novel repair pathways, says Samson. The complete sequence of the human genome will enable scientists to do such toxicogenomic studies with cells from various human organs, notes Michael Gallo of the University of Medicine and Dentistry of New Jersey—Robert Wood Johnson Medical School in Piscataway. "That's the real beauty of the genome. We can now start to dissect toxic responses at the molecular level," says Gallo.

—JOCELYN KAISER

genes work," says bioinformaticist Richard Durbin. "We see ourselves expanding our biological programs." The Whitehead Institute is already well along that road: Since 1997 it has worked with Bristol-Myers Squibb Co., Affymetrix Inc., and Millennium Pharmaceuticals Inc. to develop microarrays for monitoring gene expression. Lander expects to increase his group's emphasis on the genetics of disease traits. "We got into the genome project 15 years ago because our interest was in studying complex traits," he says. "The genome was part of the necessary infrastructure [for] studying those traits."

While these researchers are expanding into functional and comparative genomics, they are also venturing into new territory. "We are sitting on a [sequencing] capacity that can really change our thinking," says Trevor Hawkins, director of JGI. Hawkins thinks it's now practical to sequence, say, the same coding region from 100 people to begin to understand the effect of variation on a

particular disease or trait. Sanger's Bradley agrees: "A lot of sequencing capacity will be used for resequencing, looking at sequence variation, and looking for disease genes," he predicts.

As Collins, his predecessor James Watson, and others predicted at the outset, the sequence of the human genome is turning out to be a tool to enable an astounding new array of biological studies. And Sanger's Durbin agrees: "It's going to provide a kick-off for a whole lot of interesting science for a very broad set of scientists." Far from signaling the end of the genomic opera, this week's publication is merely the close of the first act.

—ELIZABETH PENNISI

With reporting by Dennis Normile and Robert Koenig.



UNSUNG HERO: ALAN COULSON

After developing sequencing technology with Fred Sanger and producing physical maps of *Caenorhabditis elegans*, Coulson headed up the sequencing effort with John Sulston at the Sanger Centre in Hinxton, U.K. As Sanger scaled up to tackle the human genome, Coulson "quietly rolled up his sleeves," says Sulston, to run the team that produced and selected clones for sequencing.