NEWS

Ideas Fly at Gene-Finding Jamboree

Computer experts and fruit fly geneticists worked side by side in an unusual jamboree to make sense of the new *Drosophila* genome

Call it an idea frenzy, a discovery lek, a Woodstock for science nerds. It's that moment every scientist lives for—a time when discoveries come fast and furious, prodded by the collective resourcefulness and creativity of researchers so caught up in their work that eating and sleeping are unwanted interruptions. In November 1999, about 45 bioinformatics experts, protein specialists, and fruit fly biologists experienced just such a moment when they gathered in Rockville, Maryland, to take a first look at the newly sequenced *Drosophila* genome and, more important, to see whether they could make sense of it. "It was some of the most exciting science I've done in a long time," raves William Gelbart, a developmental geneticist at Harvard University.

At stake was the reputation of an upstart sequencing company, Celera Genomics of Rockville, Maryland, and, in some ways, the future of genome sequencing. In 1998, the company's founder and president, J. Craig Venter, shocked the scientific community when he announced that his new company, formed in partnership with PE Corp., intended to sequence the entire 3-billion-base human genome in just 2 years—well ahead of the publicly funded Human Genome Project. What raised the eyebrows and the ire of the sequencing community was Venter's claim that he could accomplish this gargantuan feat with a sequencing strategy previously thought useful only for small microbial genomes (*Science*, 20 October 1995, p. 397; 3 September 1999, p. 1558). In contrast to the deliberate, chromosome-by-chromosome approach being pursued by the Human Genome Project, Venter planned to break the entire genome into small pieces, sequence them in one fell swoop with a phalanx of very fast and very expensive new PE sequencing machines, and use some of the world's most powerful supercomputers to assemble the sequenced fragments in the correct order.

As a dry run—and to prove to the world that this so-called shotgun strategy would work—Celera first took on the 180-million-base genome of the fruit fly *Drosophila melanogaster*. Venter set up a collaboration with the Berkeley Drosophila Genome Project (BDGP) and its European counterpart to help guide the effort and interpret the data, and he set his new sequencers to work on the fly DNA in May 1999 (*Science*, 5 February 1999, p. 767). By late fall, the sequencing was finished and the computers had assembled the pieces together. That's where the November meeting came in.

Sequencing and assembly were just the first steps. The tough task was to pinpoint the genes and begin to figure out what they do, a process called "annotation" in the jargon of genomics. Venter, Celera's Mark Adams, and Gerald Rubin, director of the BDGP, hit upon an annotation strategy that was as bold as the sequencing venture that preceded it: Just as Celera had sequenced and assembled the fly genome all at once, they would interpret the entire thing in an intense annotation "jamboree." They would essentially lock biologists and computer scientists in the same room to get the job done. Fly geneticists were eager to participate, if only so they could get a first look at the long-awaited *Drosophila* sequence. And Celera sweetened the deal by picking up the tab.

By all accounts, this slam-dunk approach, which took 11 days, worked even better than expected. The results of this effort were announced in February and are published in the following series of papers in this issue. Although the participants agree that the current descriptions and classifications of *Drosophila* genes represent a "first pass," it is still "a pretty good job," notes J. Michael Cherry, a bioinformaticist at Stanford—especially since the entire process of sequencing and annotation took less than a year. What's more, the workshop yielded a plethora of insights into the fly.

Although the fly sequence still has about 1000 small gaps, the results provide confidence that shotgun sequencing will work for other complex genomes—indeed, researchers involved in the publicly funded mouse genome project recently decided to adopt the approach (*Science*, 18 February, p. 1179). And Venter has erased most people's doubts that he will complete the human sequence later this year.

Jammin'. Experts from many disciplines gathered at Celera Genomics for a first look at the newly sequenced *Drosophila* genome.

CREDITS: (LEFT TO RIGHT) OXFORD SCIENTIFIC FILMS/ANIMALS ANIMALS; CELERA

A shaky start

This positive outcome seemed far from assured when the fly biologists first arrived at Celera last November. The company's timetable had slipped a few weeks, so the software specialists had barely a week to run the sequence data through the gene-finding programs to identify the beginnings, ends, and coding sections of what some thought would be about 20,000 genes. When Tom Brody, a fruit fly geneticist at the National Institute of Neurological Disorders and Stroke, arrived 3 days into the jamboree, "everything was in [a] shambles," he recalls. The researchers had found less than 4000 genes and had yet to run the computer program that compares selected sequences from other organisms to the entire fly genome to look for matches and thus hints about what genes and proteins have been conserved through time. The visiting scientists were frustrated because they had submitted their favorite sequences beforehand and expected to be able to begin analyzing the results as soon as they arrived. "They really weren't ready for us," says Brody.

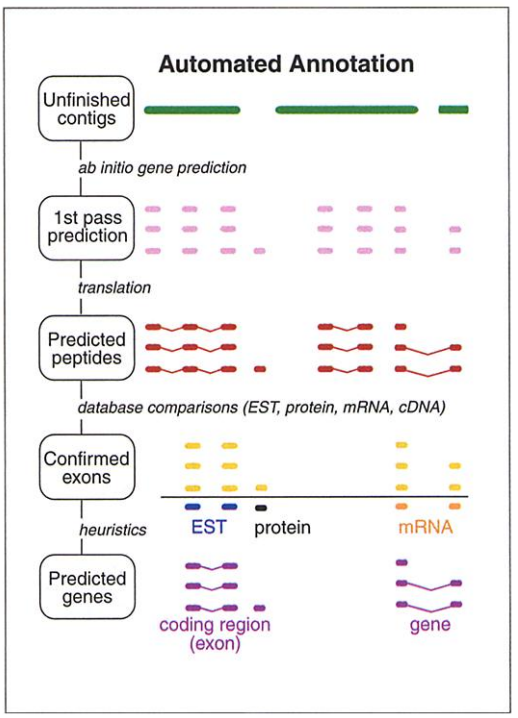
Within days, however, the group rallied. "It was as close as I've come to being in a small start-up where everyone is working 20 hours a day," Cherry recalls. "We were really

Are Sequencers Ready to 'Annotate' the Human Genome?

Imagine trying to put together a car engine when all you have is a parts list by numbers, no name or description. That's about how the rough draft of a genome looks to a biologist. To be useful, a genome must be annotated—that is, documented to provide at a minimum the putative start, stop, and structure of each gene. Biologists benefit more if information is included about the predicted gene's product and about similarities to other known or predicted genes and proteins. Only then can a researcher begin to piece together how genes and proteins interact to make life possible.

A bare-bones "parts list" for humans should be available later this spring: A rough-draft sequence is being assembled by a consortium of researchers funded by the U.S. government and Britain's Wellcome Trust, and Celera Genomics Corp. of Rockville, Maryland, has promised to produce its sequence sometime this year. Now, the genomics community is scrambling to figure out the best way to annotate those sequences. Celera will likely rely largely on its in-house team of experts and fast computers, while the public consortium is likely to put together a more disperse effort.

The challenge will be even more daunting than the one faced by fruit fly biologists in analyzing the *Drosophila* genome (see main text). "Annotation of the human genome is intellectually a very hard task, harder still than [was] *Drosophila*," says Michael Ashburner, a fruit fly geneticist turned bioinformaticist at the EMBL-European Bioinformatics Institute (EBI) in Cambridge, United Kingdom. For one, the fruit fly genome was completely sequenced when it was analyzed, but the rough draft of the human genome will be full of gaps, as well as sequence fragments that are out of place. As more sequencing is done, the publicly funded draft will evolve until it is "finished" in 2003. Until then, annotators have to learn how to work around these limitations to get the most from the current drafts.



Sequence to gene. Automated annotation by a program called Ensembl helps put new human genome sequence into a biological context.

Consistency is also lacking in the annotation that has been done on human sequence so far. For the fruit fly, researchers had set up a centralized database, called Flybase, long before the sequence was complete. Flybase contains an authoritative gene list that helped in the annotation effort. But there's no single database or list for the human.

At a January meeting at the National Institutes of Health, 10 scientists—both cell and molecular biologists and bioinformaticists, including one from EBI—met with staff from the National Human Genome Research Institute (NHGRI) and the National Center for Biotechnology Information (NCBI), which maintains the public U.S. sequence database. They agreed that they needed a standardized vocabulary for describing the 80,000 genes expected in the human sequence, as well as their functions. A single international gene index, similar to the one that exists for the fly, is crucial, some pointed out. "We need an index of human genes so we're not in a tower of Babel," says NHGRI director Francis Collins. EBI and NCBI are now collaborating on this index, as well as on integrating their two annotation strategies.

But the annotation experts could not decide how and when to bring biologists into the process. In November, a team of sequencers jump-started the annotation process for the fruit fly by holding a 2-week jamboree, during which the bioinformaticists writing the annotation software worked side by side with fly biologists who evaluated the computed results. Some see this as a model for the human annotation, but the question is by no means decided.

EBI's Tim Hubbard, for instance, is pushing for a distributed annotation system, a plan proposed by Lincoln D. Stein at Cold Spring Harbor Laboratory in New York and his colleagues (stein.cshl.org/das). In this model, NCBI and EBI would generate a minimally annotated sequence backbone—such as that now produced jointly by the Sanger Centre and EBI through an effort called Ensembl (www.ensembl.org)—to which other researchers could link their findings about a particular gene or protein. If places such as NCBI and EBI set up the computer infrastructure to do this, then a larger proportion of the biological community can get involved in this model, says Hubbard, than with a jamboree.

Many worry, however, that biologists will be less eager to contribute their specific knowledge and expertise to interpreting human sequence data than were the fly biologists. "With the *Drosophila* community, there's a certain unity. People were really gung ho," explains Steve Henikoff, a geneticist at the Fred Hutchinson Cancer Research Center in Seattle. "My impression is we don't have that kind of unity" among human biologists. The rivalries between laboratories can be keener, particularly when there is a lot of money at stake.

Others say contributing annotation may simply be a low priority for overworked biologists. "If you are back at your own desk, swamped with your own job, these [annotation needs] might not have the same importance," Stanford bioinformaticist J. Michael Cherry points out. Even though a group process, or jamboree, for the human genome would likely require many more people or many more weeks than did the fly event, "this would be one way for the human [biology] community to really get into it," he adds.

Indeed, getting biologists "into it" has proven difficult with other organisms. For example, biologists have not pitched in to annotate the genome of the soil bacterium *Pseudomonas* as readily as did *Drosophila* experts. "It's not the computing tools that are the issue, it's getting all those people to work together," says Maynard Olson of the University of Washington, Seattle. "It's a different way of doing science."

How this all unfolds over the next year will determine how useful the rough draft ultimately is, and potentially how expensive it will be for the average biologist to use. If biologists don't step up to the plate, warns Ashburner, then "private companies will be able to sell [their annotation] to the public domain for vast amounts of money."

–E.P.

SOURCE: T. HUBBARD/THE SANGER CENTRE

productive, doing stuff no one thought was possible." Each day, biologists and programmers spent hours at the computer screen, occasionally looking over each other's shoulders and discussing their findings with whomever was in the adjacent cubicle. A gong called them to the conference room for takeout lunches and dinners and for impromptu mid-afternoon seminars to discuss the day's finds.

As expected, *Drosophila* genes were at first harder to find than genes in either the nematode *Caenorhabditis elegans* or yeast —the two largest genomes sequenced until now. Instead of being simple stretches of sequence, *Drosophila* genes have more interruptions, called introns, in the coding regions. In addition, many genes can be expressed in different ways and thus have several "start" sequences, or alternative splice sites, within them. Finally, there's just more DNA between the genes to contend with than there has been in other organisms sequenced to date. "It's more of a hunt to piece these things together," says Cherry.

While some of the bioinformaticists worked on finding the genes, BDGP's Suzanna Lewis and Celera's Mark Yandell and Jennifer Wortman guided the rest of the computer experts' effort to use programs to translate those genes into proteins and check to see whether they matched known proteins in yeast, nematode, or human. One new database of protein domains, InterPro (www.ebi.ac.uk/interpro), just developed by Rolf Apweiler of the European Bioinformatics Institute and his colleagues, analyzed the proteins predicted from the fly sequences. Based on similarities with known proteins, InterPro decided whether each was, say, a protease enzyme that chews up other proteins or a membrane protein; it then classified the proteins into one of some 2000 possible families.

The biologists, many of them experts in particular protein families, pored over the results. They could spot when InterPro and other programs predicted a gene that was actually two genes, or lumped a protein into the wrong family. Feedback to the bioinformaticists led to almost instant improvements in the computer programs. Over the course of the day, the biologists came up with new ways to analyze or portray data. When they went back to their hotels for beer and some sleep, the programmers wrote the new code and ran the analyses, much to the amazement of their slumbering colleagues. "It was incredible," says Brody. "Every day they had a new way to look at genes and a new way to annotate them."

Discovery frenzy

At one point, the biologists were concerned that the sequence might be incomplete, as the fly seemed to have fewer genes than the nematode. To find out, Cherry checked whether the sequence contained the 2500 genes already known to exist in the fruit fly. It did. Cherry found all but 18 of the previously identified genes in the main scaffolds of the genome, sections where there was a lot of overlapping sequence and few gaps. He found another 12 in pieces of sequence that hadn't yet been fitted into these larger sections. When Cherry reported his findings that afternoon and wrote the six missing genes on the board, two fruit fly veterans stood up and said the first and likely another were known experimental artifacts. With further investigation into the remaining five, "we got down to one we didn't know about," meaning that they couldn't find it in their data, recalls Celera's Adams. It proved to be contaminating sequence from the polymerase chain reaction (PCR) probe used to isolate the gene, not fruit fly DNA at all. "As far as we know, we're not missing anything," he adds. The number of genes finally topped out at 13,600.

Ringleaders. BDGP's Gerry Rubin (left) discusses the jamboree's progress with Celera's Mark Adams (middle) and J. Craig Venter (right).

That small number of genes was just one of several surprises. Fly biologists expected more genes because *C. elegans* has 18,000 or so (*Science*, 11 December 1998, p. 1972), even though it consists of about 1000 cells whereas the fruit fly has 10 times as many cells. It turns out that these two multicellular organisms also differ in the number of proteins they use to carry out critical functions. Researchers had expected to find larger protein families in the more complex species. Instead, "a handful of families are greatly expanded in *C. elegans*, but in *Drosophila* those families are more modest in size," Adams notes. Certain receptors involved in development and nerve cell signaling are one example. The nematode has 1100 of these, but *Drosophila* has a mere 160, the big difference being in olfactory receptors. "We expected to find many more," Brody notes. The fly also has far fewer hormone receptors than either the worm or vertebrates. On the other hand, *Drosophila* has 199 trypsinlike peptidases—which are involved in signaling in digestion, development, and the immune system—compared to just seven in the worm and one in yeast. Fruit fly biologists will now have the challenge of figuring out why these proteins are so prominent in this organism.

With discoveries like these popping up almost hourly, any remaining skeptics soon came to appreciate the value of obtaining a complete genome sequence of the fly. Such comparisons among species "give you the potential for so much more insight into how these organisms work," says David Coates, an expert on proteases at the University of Leeds in the United Kingdom. They can also indicate which model organism is best suited for various studies. The nematode, for instance, seems to be a better model for studying certain kinds of membrane proteins, whereas the fruit fly is probably better for experiments involving certain proteases. "Everyone was very excited when I said that *C. elegans* wasn't the be-all and end-all," he adds, alluding to the long-running debate between fly and worm biologists over which organism is superior.

The fly's merits as a model for studying human biology and disease were bolstered when jamboree participants compared fly genes with all known human genes. As the researchers report on page 2204, the fly has counterparts for 177 of 289 genes known to be involved in human disease, including the tumor suppressor *p53* gene, as well as many genes involved with the insulin pathway.

These comparisons have whetted researchers' appetites for the human genome. But the human genome will be more complicated than that of the fly, and concerns are mounting that the sequencing community will not be ready to annotate it well (see sidebar). "A lot of people have spent a lot of time on sequencing, and they haven't spent a lot of time on annotating," complains BDGP's Martin Reese. "There's a big gap" (*Science*, 15 October 1999, p. 447).

But as critical as annotation is to interpreting a sequence, it is by no means the end. Annotation provides just a "nice approximation of the truth," says Apweiler. Not until biochemists, physiologists, and other wet-lab researchers have verified those predictions experimentally will anyone know for sure that they are right. Determining such truths will likely occupy biologists for much of the next century. **–ELIZABETH PENNISI**

CREDIT: CELERA