Genome Shortcut Leads to Problems

With the development of a system capable of cloning long stretches of human DNA, gene mappers' dreams appeared to have come true—but it has also brought a few nightmares

On the morning of 7 December 1992, a Federal Express jet touched down at Boston's Logan Airport with a Christmas present for U.S. genome researchers. Packed in dry ice was the final shipment of the much publicized French "megaYAC" library—the entire human genome stored in the extra-large yeast artificial chromosomes (YACs) developed at the Centre d'Etude du Polymorphism Humain (CEPH) in Paris. The keenness with which U.S. laboratories had anticipated these megaYACs was evident from the elaborate plans they had made to distribute them. A team at the Massachusetts Institute of Technology's Whitehead Institute reproduced the library within 14 days, shipping a copy to the Los Alamos National Laboratory and the Salk Institute, both of which then quickly copied it again and shipped it to four other labs each for further copying and distribution. With this chain-letter allocation scheme, most major U.S. gene mapping labs got the megaYACs within 6 weeks of their arrival in the United States. Everybody had a fair start on mapping the genome with these remarkable new tools.

STEVE MUREZ/BLACK STAR

MegaYAC pioneer. Daniel Cohen of CEPH.

But now, 3 months later, much of that initial enthusiasm is gone. The distribution system worked fine, but for many researchers the megaYACs have not. The great appeal of the new YACs is that they can carry gene sequences up to 1.4 million bases long—twice as long as the best traditional YACs, meaning that it should take many fewer YACs to make a rough physical map covering the entire human genome. But for many of the groups that had hoped to start right in using them, these past months have been filled with one frustration after another, as the limitations of megaYACs become clear. Most of the problems have a familiar ring—they're the well-known problems of any YAC writ large (see box below). But the number of researchers who have now taken to trashing megaYACs suggests that many had expected better of the new technology, if only because it was so mightily touted.

This is bad news for the U.S. Human Genome Project, which has sunk millions into mapping with megaYACs and was hoping to jump from today's physical maps to an aggressive sequencing program. And it's bad news for researchers who had been counting on megaYACs as a quick road to specific disease genes. The University of Michigan's Francis Collins, for example, says problems with the technology slowed him down in his efforts to find a breast cancer susceptibility gene (*Science*, 29 January, p. 624).

Collins is not alone. Some researchers who are focusing on particular regions of the genome are finding that the megaYACs in those regions are internally scrambled or missing large chunks, a phenomenon known as a "deletion." Other groups complain that the megaYACs seem to have an unexpectedly high degree of "chimerism"—unrelated areas of the genome spliced together into a single stretch of DNA. Last month, at a meeting sponsored by the Department of Energy (DOE) in Santa Fe, New Mexico, the grumbling finally went public, as one scientist after another reported troubles in working with the library. "A lot of researchers just threw up their hands," says David Galas, the director of DOE's genome program.

None of this comes as a surprise to Daniel Cohen, the codirector of CEPH, founder of the Généthon gene mapping center, and the leading proponent of the use of megaYACs for large-scale genome mapping. Having recently led the team that published the first map of chromosome 21 (*Nature*, 1 October 1992, p. 359), he knows first hand the limitations of the technology. MegaYACs, he says, "are very difficult to work with." But he ar-

YAC Troubles Run Deep

Most of the snags coming to light with megaYACS, researchers now suspect, are rooted in the basic YAC technology. YACs, or yeast artificial chromosomes, were first developed in 1987 by a group headed by Maynard Olson, then of Washington University in St. Louis, and the technology was quickly taken up by other researchers because it offered a way to clone long stretches of DNA. YACs can carry and reproduce up to 600 kilobases of foreign DNA—15 times the amount that can be handled by previous cloning systems. And the new megaYACs offered a similar improvement over YACs.

But some limitations were clear from the start. Olson originally predicted, for example, that about 10% of the clones in YAC libraries would consist of DNA fragments from different parts of the genome that had been spliced together. As researchers learned to spot such chimerism better, however, they found more of it, first 20%, then 30% and more. In a much-discussed footnote in the paper reporting the map of the Y chromosome published last year (*Science*, 2 October 1992, p. 60), David Page of the Whitehead Institute revealed that chimerism had plagued 59% of his YACs.

These problems with the YAC technology probably stem from the very attributes that make yeast a good medium for growing up large stretches of DNA: its ability to splice foreign DNA into its own genome. This splicing mechanism, part of yeast's capacity for repairing DNA damage, is apparently so effective that it can result in chimeras by stitching together fragments of foreign DNA. It may also be responsible for the problem of deletions in the cloned DNA. Regions of the human genome with many repeated sequences tend to set off the yeast's internal warning signals of DNA damage, triggering its repair mechanism. Other regions appear to be toxic to yeast or too fragile to be cloned without breaking, says Mel Simon of Caltech.

–C.A.

Genome Databases Worry About Yeast (and Other) Infections

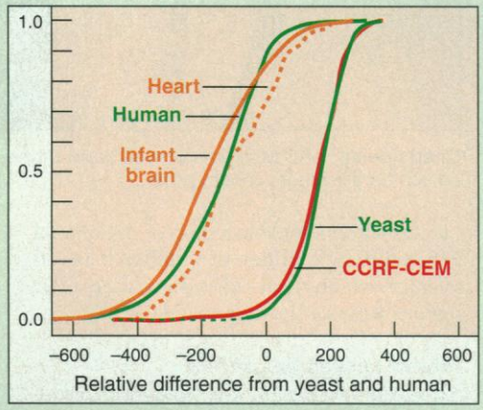
It was a ceremony that some participants would probably just as soon forget. On 18 October last year in Paris, a group from Généthon, the French gene-mapping laboratory, formally presented to the United Nations Environmental, Scientific, and Cultural Organization (UNESCO) a listing of some 2000 DNA sequences—a gift, they said, of "human genes to humanity." The Généthon group announced that the sequences had been deposited in a public database at the European Molecular Biology Laboratory (EMBL) in Heidelberg for all to use, free of any restrictions. But, as with many grand public gestures, this one has turned out to be something less than it first appeared. Most of the sequences, new analyses have revealed, were not what the Généthon group thought they were. In fact, most may not even be human DNA.

Computer analysis by two independent groups, one of which reports its results in a letter on page 1677, has revealed that more than half—and perhaps as much as 85%—of the Généthon sequences appear to be from yeast and several unidentified bacteria. The complementary DNA (cDNA) sequences had apparently been inadvertently cloned from DNA contaminants introduced either by the French group, the company that had supplied it with a cDNA library made from a human cell line, or perhaps some earlier source. Beyond being an embarrassment (genome pundits were quick to joke that Généthon had "given UNESCO a yeast infection"), the inadvertent release of mislabeled sequences reveals a serious weakness in the current method for distributing genetic data: Thousands of cDNA sequences are now being "published" by being directly submitted to electronic databases—without peer review or serious error-checking. Since contamination happens in even the best genome labs, database managers are getting worried that without better quality control, this case could be a harbinger of future problems. And there's good reason to worry. "A sequence with the wrong organism attached is not only useless, it's dangerous," says Graham Cameron, head of the EMBL data library. "Some researcher might waste months trying to follow it up."

The two groups—one headed by Chris Fields of the The Institute for Genomic Research (TIGR), based in Maryland, and the other led by Babis Savakis of the Institute of Molecular Biology and Biotechnology (IMBB) in Greece—used different approaches to reach the same sobering conclusion. In part to answer criticism that some of its own published cDNA sequences were actually nonhuman contaminants, Owen White of the TIGR group developed a computer algorithm to check unknown sequences for their species of origin. The program goes through a sequence six bases at a time, developing a statistical profile of the frequency of all the different combinations of "A"s, "C"s, "T"s, and "G"s in the entire sequence. Fields' group has found that sequences from different species have different statistical fingerprints, and the algorithm uses these fingerprints to match a sequence with its species of origin. Based on this statistical analysis, the TIGR group reported at a genome conference in Santa Fe, New Mexico, last month that the Généthon cDNA library may be up to 85% nonhuman.

As a double check, the TIGR group took a more straightforward approach—the same one the IMBB group employed independently. The groups simply compared the sequences in the suspect Généthon library against all other sequences in several nucleotide and protein databases, regardless of species. Both groups found matches to a small fraction of the Généthon sequences. But only 16% of the matches in the TIGR search were to human sequences, and less than one-third of the IMBB group's matches appear to be to human DNA.

Charles Auffray, the director of the Généthon cDNA sequencing program, says his laboratory became aware of problems with contamination with nonhuman DNA late last year. He notes that Généthon researchers sent out an e-mail message last November acknowledging "very significant sequence similarities" with yeast. When his group originally searched the EMBL library to determine if the Généthon sequences were unique, researchers did get hits on nonhuman sequences, he says, "but we didn't see 85%, or we wouldn't have submitted the data to the library." Although Généthon researchers have not confirmed the high contamination rate seen by the TIGR and IMBB groups, they are now attempting to verify all sequences by hybridizing them against DNAs from common contaminants and conducting computer searches before submitting them to public databases.



Not human. A statistical profile of Généthon's cDNA library (CCRF-CEM) shows that it is more like yeast than human DNA.

Auffray blames most of the yeast contamination on a cDNA clone library that Généthon purchased from a commercial supplier. Indeed, that supplier, Clontech Laboratories Inc. of Palo Alto, California, says it halted distribution of the library 6 months ago after it started getting complaints. But Kenneth Fong, Clontech's director of custom synthesis, says the company can account for only a small fraction of the yeast contamination in its own preparations.

This episode may have been an extreme case, but contamination is far from rare in genome research. And that raises the issue of how to keep incorrectly identified sequences out of the databases. Officials at EMBL and its U.S. equivalent, the Genbank database at the National Library of Medicine's National Center for Biotechnology Information (NCBI), agree that better screening and quality control are necessary, both by the databases and the researchers themselves. Both databases are looking for better error-checking tools, and Fields says he intends to make the TIGR algorithm freely available to the community.

But as journals become increasingly reluctant to publish and peer review reams of gene sequences, the databases are going to have to go even further in catching mistakes, says NCBI director David Lipman. "We're depending on the author to get [the data] to us in as good shape as possible," he says. "But there's no excuse for us not to do a better job on picking up the obvious inconsistencies."

Indeed, some databases are considering upping their standards. NCBI, for example, has put together a blue-ribbon panel to hash out an "editorial policy" for Genbank. "Should we do some refereeing?" asks Lipman. "Should we have an editorial board?" The answers aren't yet clear. But it won't take many more cases of gross contamination, he says, to shift the balance toward more quality control, even if it comes at the expense of some delay.

–C.A.

gues that there is no superior alternative; it is better to build somewhat flawed maps with megaYACs today than to wait for some better alternative to come along eventually. Anyone who expected more of megaYACs than that, he says, was "naive."

Perhaps. But the problems that have been cropping up took even the most sophisticated gene mappers by surprise. Take Collins, for example, who will become director of the Human Genome Project later this year. When his team analyzed the megaYACs in the region of chromosome 17 where the breast cancer gene is expected to lie, they were shocked by how few of the clones corresponded to what was already known about that region. "Only 20% of the megaYACs we've pulled out are reasonably representative of what's going on in chromosome 17," says Collins. Glen Evans of the Salk Institute says, "We had always anticipated that the larger the YAC, the more the chimerism." But now that his team is working with them, "it's turning out much worse," he says. "We're finding 70% to 80% chimerism."

"Chimerism for the big YACs is a major problem," agrees DOE's Galas. "But internal rearrangements and deletions are an even bigger problem, and that seems to be especially true with the megaYACs." Some deletions appear to be random, which suggests a simple, if time-consuming, solution: use enough YACs to get an accurate map eventually, even if many are scrambled. But other deletions may be regions that are simply uncloneable in YACs of any kind. Collins points out that one such deletion is "right in the middle of the Huntington's region." Another is in the breast cancer region, and Cohen says chromosome 15 seems to have problems as well. In general, Collins predicts there may be an uncloneable region of the human genome every 2 million to 3 million bases on average, which means that more than 1000 regions will not show up in YACs.

PAULINE LUBENS

Costly delay. YAC problems slowed down Francis Collins' search for the breast cancer gene.

Project backed megaYACs

This would be just another scientific debate if the genome project hadn't essentially declared megaYACs the mapping tool of the future. Earlier this year, the National Institutes of Health (NIH) gave its largest-ever genome grant—$24 million over 5 years—to geneticist Eric Lander's group at the Whitehead Institute to set up a major center to create a low-resolution map of the whole genome. Lander and Cohen are collaborators on the center, and they intend to base most of the map on megaYACs.

Lander has no real concern that his team will be able to finish the job with the tools at hand. His center, he points out, is making what genome researchers call a low-resolution STS content map—a representation showing the order of unique points called "sequence tagged sites" or STSs. "The YACs themselves hardly figure in an STS content map," says Lander. "They're simply a tool for showing that two points are nearby. YACs may be sloppy and horrible for some purposes, but they're perfectly adequate for the job of connecting points over big distances." Once you've done that, Lander concedes, "you want to get out of YACs as soon as possible" and move to some other system, such as cosmids, small loops of genetic material inserted into bacteria that carry less DNA, more reliably. Only then would one consider higher resolution mapping and sequencing of the genome, he says.

Some groups worry, however, that the genome project may have bet the farm on a technology that may not take them where they want to go. An STS content map is a perfect starting point for sequencing as long as you have reliable sources of DNA to go from one STS to the next. But megaYACs just aren't up to the task, and better strains of YACs that might be are still a year or more off. "MegaYACs may be too inaccurate for finding disease genes or sequencing—and, as far as I'm concerned, those are the primary purposes of the genome project," says Evans.

Chromosome 21

The chromosome 21 map, in fact, is a good example of how the community divides on this issue. When 21 was published last year, it was hailed as proof that the entire genome map was just around the corner. But since then, researchers have found problems with the chromosome 21 map that indicate how difficult it may be to go from a map based on megaYAC technology to one with finer resolution that would be the starting point for sequencing.

The original *Nature* article by Cohen and 35 coauthors acknowledges that the researchers could not place seven genome markers known to be on chromosome 21 and it said that 12% of the megaYACs they used contained deletions. But Jeffrey Gingrich of the Lawrence Berkeley National Laboratory's human genome center says he is finding that

BAC to the Future?

The emerging problems with YACs may have been a blow to many genome researchers (see main story), but some scientists have long been advocating more funding for alternative DNA carriers. As the Human Genome Project focused on YACs in recent years, however, support for the development of non-YAC systems has been hard to come by. But as the community comes to grips with the limitations of YACs, researchers are starting to take another look at the alternatives.

One of the most promising is the newest—the Bacteria Artificial Chromosome (BAC). Like a YAC, it is formed by inserting foreign DNA into a microorganism—*E. coli* in the case of the BAC—and letting the organism's genetic reproduction machinery take over to make copies. Mel Simon and his group at Caltech developed the BAC in 1990 specifically to try to get around some of the problems with YACs, and Simon says they have been at least partially successful. BACs don't appear to result in unrelated pieces of DNA being spliced together and cloned, for example, and very preliminary analysis suggests that deletion of stretches of DNA will not be a significant problem. Best of all, he says, those regions of the human genome that are uncloneable in YACs will probably clone just fine in BACs, and vice-versa.

BACs do, however, have one significant disadvantage: They can carry stretches of only 200,000 to 300,000 bases of foreign DNA—less than one-quarter as long as the best YACs. And other possible alternatives to YACs are limited to even shorter fragments. Another bacteria-based carrier known as a P1 can carry DNA insets of about 100,000 bases, and cosmids, which have been around the longest, are shorter yet—about 40,000 bases. While none of these systems is going to challenge the YAC for the size of their genome fragments, Simon says, they should provide a more precise complement to the YACs' brute-force capabilities.

–C.A.

as many as half the megaYAC clones on the 21 map either fall short of their reported length, contain misplaced STS markers, or have deletions of unknown length. "If you look at the distribution of every clone versus the STSs, they were wrong on many, many points," he says. "I hope this isn't the case with every other chromosome."

Gingrich places most of the blame for the problems in the chromosome 21 map on hurried mapping techniques that didn't compensate for known deficiencies in the megaYAC technology. But, adds Evans, "it's obvious from what they published on 21 that there are some really unanticipated problems. It represents a very low resolution map with a lot of mistakes." As a result, he says, "you've got to do it all again and correct all the errors."

That's not the same as starting from scratch, of course. "It's somewhat easier [the second time] because you're basing it on some [known] structure," says Mary Kay McCormick of the Los Alamos National Laboratory's human genome center. But that doesn't mean it will be easy, either. "There was a lot of grumbling at the Santa Fe meeting by people who had tried to construct maps using [megaYACs] as a starting point," she says. "They're not going to be able to construct the physical maps that they had hoped."

This sort of criticism infuriates Cohen. He points out that the *Nature* paper is full of caveats and discussion of problems with the megaYAC libraries and his team's mapping technique. And he has no time for what he describes as after-the-fact nit-picking. "There is no map today that does not need to be refined," he says. "Of course, with this one, the more you study it, the more errors you will find." But that, he says, is the way of science: breakthroughs, followed by refinements.

The search for solutions

No one is suggesting that YACs do not have their place in the genome project. Instead, the problems have forced researchers to find ways to improve the technology. Evans' group, among others, is developing YACs from a hybrid pool of human and nonhuman DNA. That may reduce chimerism, because the yeast is less likely to splice together genetic material of unrelated species. Lander's group is creating special recombination-deficient strains of yeast that may cause fewer deletions. "I'm convinced that by the end of the year, we'll have a good host strain that will solve most of the problems with YACs," says Caltech's Mel Simon. But it is late in the game to be fixing the basic tools, he points out.

The sobering of the enthusiasm for YACs has also spurred a search for YAC alternatives (see sidebar on page 1686). As Collins puts it, the emerging deficiencies of YACs "say that we have not yet discovered the perfect vector for building whole genome maps." But he also emphasizes that the years spent assembling the maps to date were not wasted. "Next year, when some terrific new vector comes along that doesn't have problems with deletions and chimeras, we'll be ready" for high-resolution mapping and sequencing. Just because some 5% of the genome is YAC-unfriendly, he says, "I don't think we should hold off [mapping] the 95%" of the genome that YACs can handle. And the University of Washington's Maynard Olson, who invented YACs in 1987, says he is "more impressed with the megaYACs than with the grumbling about them."

"As the megaYACs trickle down," warns Lander, "there may be people who don't know what they're getting into." Yet just as megaYACs weren't the Second Coming, neither are they the genome Apocalypse. "My view is that this is still a new technology with lots of promise," says David Botstein, chairman of genetics at Stanford. "Like all new technologies, it's going to have drawbacks." If you know what they can and can't do, he says, you won't get burned. Perhaps the best lesson from the megaYAC's troubled U.S. introduction is the oldest: Caveat emptor.

–Christopher Anderson

SPACE PROGRAM

NASA Puts the Squeeze on the Station

Space Station Freedom is beginning its latest painful metamorphosis—by some counts, the third in less than a decade. Under orders from the Clinton Administration to come up with a drastically scaled-back plan by 1 June, National Aeronautics and Space Administration (NASA) planners have begun putting agency administrator Daniel Goldin's "smaller, faster, cheaper" slogan to its toughest test yet. How they will go about it is still far from clear, but last week a few outlines of their plan began to emerge.

Administration officials will not reveal how drastic a cut they are seeking in the cost of the station, but Goldin announced at a meeting of the American Astronautical Society in Crystal City on 10 March that the revised plan may halve the station's current $30 billion development cost and $100 billion, 30-year operating cost. To do so, says space station deputy director Marty Kress, the "redesign team," headed by NASA assistant deputy administrator and Massachusetts Institute of Technology aeronautics professor Joseph Shea, may retreat from the original goal of a full-time manned station to a partially manned, or "man-tended" project. Additional savings might come, said Goldin, from a reduction in the station's planned lifespan from 30 years to 10 or 15 years.

The retrenching should make it possible to build the station with only half of the 17 to 20 shuttle flights planned earlier, Goldin told the Astronautical Society. He also noted that NASA planners are considering the use of other types of rockets to assemble the station, rather than relying only on the shuttle. All of which should make the station cheaper—and also quicker to build, Goldin noted, helping NASA keep promises made before the redesign to get it off the ground by 1997.

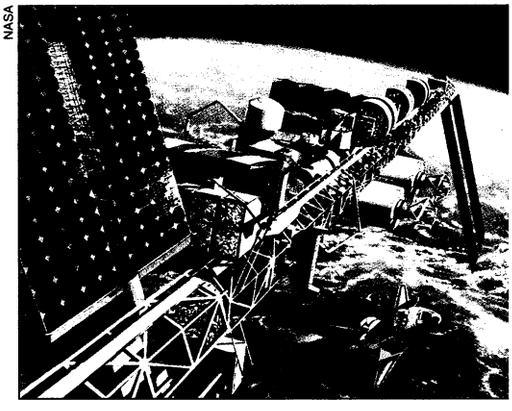
So far, he and other NASA officials aren't talking about what sacrifices, if any, the scaled-back station will entail. In a publicly circulated letter to NASA officials Goldin stated, vaguely, that the redesigned station will still "satisfy high-priority research goals in materials and life sciences." But those priorities are still being sorted out, judging by the contradictions in NASA statements last week. In the letter, Goldin said one of the goals of the redesign is to "support long-duration research (but not necessarily permanently manned)." But at a press conference he stressed that NASA is still aiming toward a permanently manned station. "I believe we will have a permanent human presence," he said.

One reason for the ambiguity may be the pressure to keep the project fully manned that is coming from a variety of sources: collaborators in Japan, Europe, and Canada, and members of Congress whose districts are benefiting from space station contracts. Another reason for sticking to the earlier goal is to protect the $8.5 billion NASA has already poured into the current station plans. Contractors and NASA centers have already designed the station's power system, notes NASA's Kress. How much of that work will survive the redesign is not clear, he adds. "That becomes the issue for all of us."

The new slimmed-down design will stand or fall on an even more basic issue, however: Is it still too big for Congress to swallow?

–Faye Flam

NASA



Last year's model. The space station that was.