he identified two distinct subgroups—"diseases within a disease," Staudt says. One gene expression profile appears to carry a good chance of survival; the other does not. If such results hold up, genetic profiling could be useful in diagnosing and treating lymphoma.

Data overload

With such tools coming on line and interest in expression studies on the rise, the volume of data in this field is likely to grow exponentially in the next few years. Already, Brown and others have been talking about new ways of storing, sharing, and publishing these huge files. Each experiment produces a flood of data: Trent's melanoma expression data, for example, would produce a print-out about 10 meters long if printed at full length—too big to publish in a journal.

For the moment, Brown says, microarray users are storing results in their own Web-accessible files and opening them to the public when they publish a journal article. Personally, Brown would be happy to skip the journal-controlled part of this process and put the data right out on the Web. That's why he's enthusiastic about NIH's plan for online publishing, PubMed Central (*Science*, 3 September, p. 1466).

One problem—where to archive data—may be solved soon. At the Arizona microarray meeting in September, David Lipman, director of NIH's National Center for Biotechnology Information, announced that NCBI staffer Alex Lash is heading up the design of a new database for the field, to be called the Gene Expression Omnibus (GEO). It will connect sets of experiments that appear "relevant to each other," so that a user could quickly find all the experiments involving certain gene families and look for common themes. "We're working on fields and data structures and will load samples this fall," Lipman says. He hopes GEO will be running by spring.

Yet to be resolved, however, is how to make results comparable. GEO will ask researchers submitting the data to define the experimental "platforms" they use. That may be simple for people using arrays or array services such as those provided by Affymetrix and Incyte. But there are no standards for homemade devices, and small differences in experimental conditions may lead to discrepancies in results.

But Lipman isn't rushing to impose standards on the young field. Brown thinks that's the right course: It would be a mistake, he says, to try to impose rules on the field while investigators are still in exploratory mode, pointing their microarray telescopes at the universe of genes. Better to let standards evolve gradually, as the data start pouring into GEO in 2000. **–ELIOT MARSHALL**

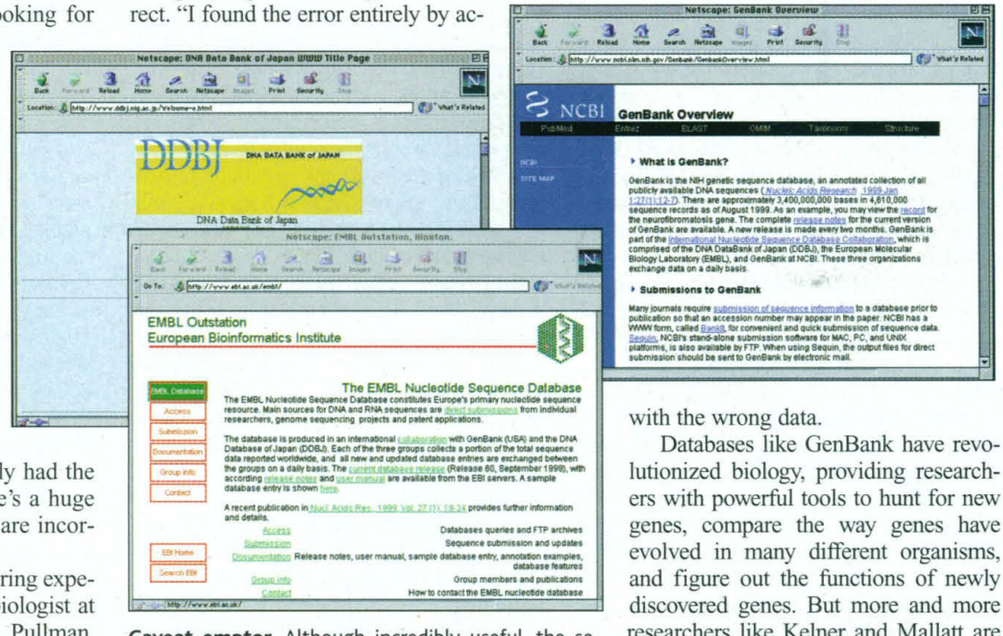
NEWS

Keeping Genome Databases Clean and Up to Date

As the size of GenBank and the number of other biological databases grows so does the need for ways to update and coordinate the information they contain

Last year, Michael Kelner thought he had finally gotten his hands on the elusive front end of a gene he'd been working on, on and off, for months. But when he searched GenBank, a public archive that contains every published DNA sequence, looking for similar genes that might hold clues to his gene's function, he knew something was wrong: He turned up more than 100 matches, or "hits," from a wide array of genes from many different organisms. Kelner, a molecular pathologist at the University of California, San Diego, soon realized that the sequence all these genes had in common was a contaminant, introduced by the commercial kit he had used to clone his gene. And the fact that it turned up in so many genes in GenBank suggests that many other scientists unknowingly had the same problem. As a result, "there's a huge number of public sequences that are incorrect," he says.

John Mallatt had a similar sobering experience recently. An evolutionary biologist at Washington State University in Pullman, Mallatt was trying to determine evolutionary relationships between various organisms by comparing the sequences of specific genes. After several months' work, he realized that the GenBank sequence he had been relying on for one of the ribosomal RNA genes of *Xenopus*, a species of frog, was incorrect. "I found the error entirely by accident as I stumbled on [a report in the scientific literature] with the corrected sequence," he recalls. "It took me about 10 hours, crawling through the correct and incorrect sequences base by base, to fix it and enter the correct sequence into my phylogenetic analysis." Mallatt subsequently discovered that GenBank contains both sequences, but there is no indication which is the correct one. Because *Xenopus* is one of the few amphibians whose genes have been sequenced, it is widely used for evolutionary studies, so it's likely that other researchers have completed and published phylogenies with the wrong data.



Caveat emptor. Although incredibly useful, the sequence data archived in DDBJ, EBI, and GenBank are not completely error-free.

Databases like GenBank have revolutionized biology, providing researchers with powerful tools to hunt for new genes, compare the way genes have evolved in many different organisms, and figure out the functions of newly discovered genes. But more and more researchers like Kelner and Mallatt are discovering that this mother lode of information contains some fools' gold that

can mislead the unwary biological prospector. Based on their surveys, genomics experts estimate that some 2% of GenBank's entries may contain DNA introduced by experimental procedures. In other entries, bases are missing or incorrect in stretches of supposedly finished sequence, or genes are even placed on the wrong chromosome. Even more problematic, some say, are inaccuracies in the labels and annotations that accompany many sequences. Hamster sequence is called human DNA, for example, and partial genes are misclassified as complete.

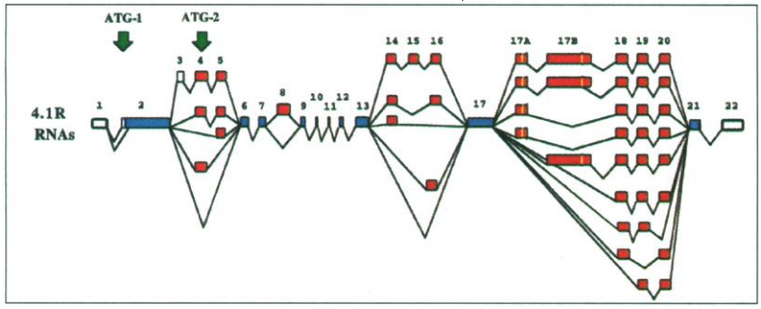
"GenBank is full of mistakes," says Michael Ashburner, who helps run a fruit fly database called FlyBase and other databases at the European Molecular Biology Laboratory (EMBL)–European Bioinformatics Institute (EBI) in Cambridge, U.K. GenBank's counterparts, the DNA Database of Japan (DDBJ) and EBI, and archival databases such as the Protein Data Bank (PDB), are also accumulating errors. Even databases whose entries are reviewed and updated, such as FlyBase or SwissProt, a long-established protein database based at EBI and at the Swiss Institute of Bioinformatics in Geneva and Lausanne, can have mistakes or missing data. And these problems are only going to intensify as labs around the world pour out sequence data from the human genome and other organisms.

Dozens of teams of bioinformaticists and biologists are trying to tackle the problems, but it's a daunting task. For one, "the databases are getting so large that getting people to correct the errors and knowing what the errors are is a major thing," laments Terry Attwood, a biophysicist at the University of Manchester in the U.K. Neither GenBank, nor EBI, nor DDBJ, discriminates between correct and incorrect data. Like PDB, they expect the discoverers and depositors of the data to update and correct the information they supply, yet many researchers find that task too burdensome. "There's no reward for it," says William Gelbart, a Harvard developmental geneticist who is one of the coordinators of FlyBase.

And a lack of funds is hampering more systematic approaches to the problem, such as having experts review incoming information or developing ways to link existing entries with new data. Until now, bioinformaticists have been able to "roll a solution together with bubble gum and bailing wire," says Owen White, a bioinformaticist at The Institute for Genomic Research (TIGR) in Rockville, Maryland, but that won't be enough. "We must have real money from the granting agencies, or we're basically going to have the [biological equivalent of the] Hubble [Space] Telescope and no way to look at the data." And it will also require a change of attitude on the part of many researchers. "We have to educate people about databases so [researchers] don't assume [the databases] are right," says Attwood.

Problematic sequences

Douglas Crawford knows only too well the need for a better way to look at the biological data in databases. Over the past 5 years, "the utility of GenBank has declined greatly," says Crawford, an evolutionary biologist at the University of Missouri, Kansas City. At one time, Crawford and his colleagues, who study genes for metabolic enzymes, were eager to look through GenBank for any matches to a new gene they isolated. Now, he says, such searches tend to turn up "hundreds of hits," including "a lot of sequences which by themselves are meaningless" because they are just pieces of genes, or worse, because they are slight variations on the same gene from the same species. It takes many hours of tracking down the primary literature to sort out if any of those matches are useful. Sometimes, after going through the trouble to find a gene that is supposedly complete enough to warrant further study, Crawford says he finds that the sequence is missing key bases at the beginning or end, or it may lack a coding region found in that same gene from other species, "and we don't know whether that [loss] is real or not."



One gene, many proteins. Depending on the coding regions used (red), this one sequence describes several proteins.

One annoying type of contamination is what led Kelner astray: the inclusion of a piece of DNA from an entirely unrelated organism in a stretch of sequence. Most such problems arise because vector DNA—bits of genome from the phage or bacterium used to clone the sequence under study—was not removed before the new sequence was submitted to the databases. By the end of last year, according to EBI's Rodrigo Lopez and his colleagues in Cambridge, U.K., some 219 published reports of contamination in the major sequence databases had been published, several of which noted dozens of different kinds of problems. Because it's up to the discoverers to make corrections, "there is very little that database curators can do to remove [the errors]," Lopez and his colleagues wrote in the September 1998 EMBnet newsletter.

Kelner's experience illustrates this problem. After he scored so many hits in GenBank, he quickly suspected the commercial cloning kit he had used. Sure enough, when he looked up the journal reports of a few of the sequences that matched his, he realized that all the researchers had used the same CLONTECH Marathon kit to pull out their gene's front end. Like many researchers before him, Kelner hadn't seen instructions buried in the appendices telling him to trim out the kit's DNA. Kelner caught the error before he deposited the sequence in GenBank, but many other researchers evidently did not, and the contaminant is now officially recorded as part of scores of genes. GenBank's Paul Kitts says he had not been aware of contamination with the CLONTECH sequence until Kelner brought it to his attention, but he is not surprised.

Sequences that contain errors or missing segments represent a more insidious problem that can trip up the unsophisticated database user. If a gap in a supposedly completed genome or gene region is large enough, or the region is studied by enough people, then it is likely to be caught fairly quickly, says Mark Boguski of the National Center for Biotechnology Information (NCBI). Such was the case with some six megabases of sequence from the nematode genome that was still missing from GenBank when it was published late last year (*Science*, 11 December 1998, pp. 1972 and 2012). Several months later, after other researchers complained to GenBank, the nematode sequencing teams made public the missing segment. But misassembled data, in which adjacent sequences don't really belong next to each other, or small gaps such as the loss of a single coding region in a gene with multiple coding regions, are more likely to go unnoticed, says Boguski. The same is true for incorrect sequences. And even when these problems are detected, there is no mechanism to flag them or to replace them with the correct data. For this reason, "there will always be an error legacy which will be very difficult to correct," says Attwood.

Missteps in translating form to function

Even if a gene's sequence is complete and accurate, an unwary biologist can still be led

CREDIT: SPENGLER, ZORN, AND CONBAY/LAWRENCE BERKELEY NATIONAL LABORATORY

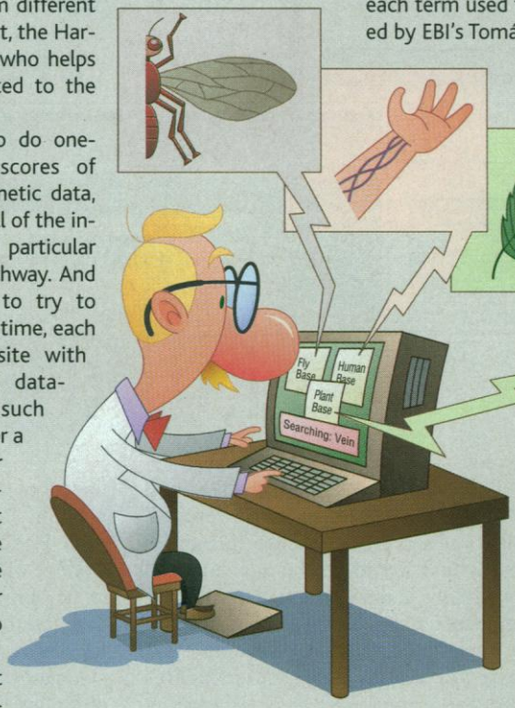
Seeking Common Language In a Tower of Babel

With gene sequences by the thousands pouring into databases, efforts are revving up to figure out what all those genes do—what proteins they make and how they fit into the workings of living things. Comparing genomic data from different organisms will be key to answering those questions. Already, researchers have found that organisms from microbes to the sequoia tree and whales to mushrooms have much more in common than was once appreciated, and those similarities are shedding light on the functions of unknown genes and their protein products.

But these efforts suffer a big handicap: Genetic information is stored in different ways in different databases, which makes it hard to compare their holdings. So, while computational biologists are trying to improve the quality of the databases (see main text), they are also working to build bridges between them. So far, they have had only limited success. "The main problem is interoperability—how to merge information from different databases," says William Gelbart, the Harvard developmental geneticist who helps run FlyBase, a database devoted to the fruit fly *Drosophila*.

Ideally, researchers want to do one-stop shopping among the scores of databases that now collect genetic data, conducting a single search for all of the information on record about a particular gene, protein, organism, or pathway. And bioinformaticists have begun to try to make this possible. In the meantime, each database has its own Web site with unique navigation tools and data-storage formats that make such searching difficult, by a person or a program. Users have to master the idiosyncrasies of each database's tools, and programs can't easily recognize data that are not stored in a uniform way. The lack of a common language for gene functions is also proving to be a serious problem.

Alternate spellings, different names for the same gene, or different uses for the same word can trip up even smart search programs. Take, for example, a search for genes involved in vein development: It is likely to pull up information related to the human circulatory system, a leaf, or a fruit fly wing. Remedying the problem, says Gelbart, is "a question of how to undo 100 years of [building] a tower of Babel." And the digital babble threatens to become deafening as new kinds of databases, such as those cataloging gene expression data (see p. 444) or data from large-scale protein identification experiments, come online.



Searching in vain. Computers can't distinguish when a word has different meanings, confounding data quests.

Even 2 years ago, when there were fewer databases and much less data, cross-searching was difficult, says Nathan Goodman, a bioinformaticist at Compaq Computer Corp. in Marlboro, Massachusetts. Goodman, who was then at The Jackson Laboratory in Bar Harbor, Maine, and Jackson Lab mouse geneticist John Macauley conducted a test to see how easily they could cross-search GenBank, a sequence archive, and smaller mouse and human genome databases, which contain other types of information about these genes, to pool information on particular genes. They reported in the August 1998 issue of *Bioinformatics* that they failed to identify counterparts for 26% of the mouse genome database entries that were known to have an equivalent human gene; the reverse was true for 17% of the genes in the human genome database.

Some promising efforts are under way to tackle these problems. One, called the gene ontology project spearheaded by Michael Ashburner at the European Bioinformatics Institute (EBI) in Cambridge, U.K., seeks to come up with a set of common, shared definitions for each term used to describe biological data. Another, now coordinated by EBI's Tomás Flores, is establishing standard ways of representing data. Both have a long way to go and do not yet have the full support of the community, however.

A few researchers, such as Goodman, argue that the best way to minimize incompatibility is to centralize the data collection and storage. If one "federation" oversees the various databases, they argue, then it is more likely that standards will be established and links between the databases will be kept up to date. "A centralized database might be much easier to maintain," Gelbart notes.

But that approach seems to be losing ground, as new data archives proliferate. At least three groups are independently coming up with their own way to store and display data from microarrays, for example. They include Stanford—where microarrays were invented—EBI, and the National Center for Genome Resources. Similarly, at least three databases for cataloging slight variations in genes called single nucleotide polymorphisms, or SNPs, have been set up in the past 2 years.

Some experts argue that such multiple efforts are healthy. "It allows people to experiment and come up with new ideas," says Ashburner. But others worry about conflicting standards and duplication of effort. Letting 100 flowers bloom "is expensive, and it doesn't scale well," says Owen White of The Institute for Genomic Research in Rockville, Maryland. "And the hazard you head toward is that everyone has a different way of representing their data." As White sees it, the way things are going now, "in the near future, people will want to ask simple questions and will find the databases inadequate." **–E.P.**

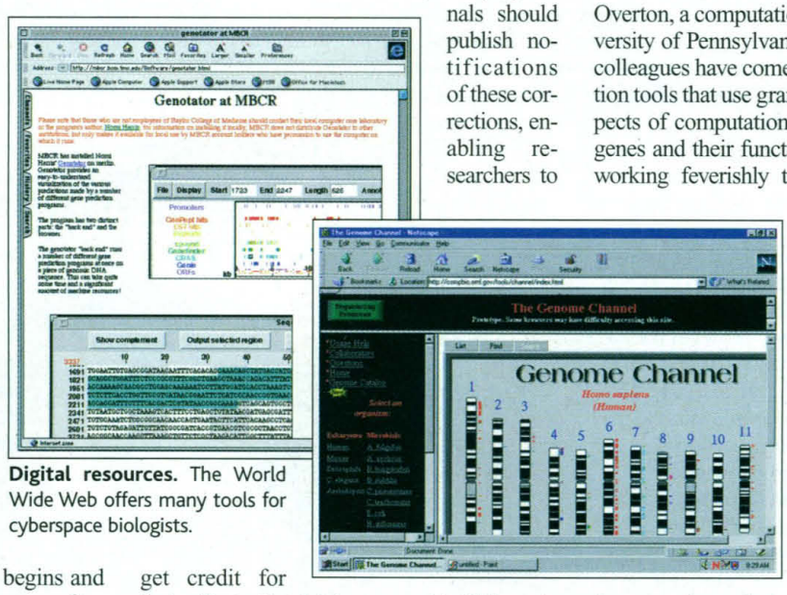
astray if information about the gene's function is incorrect. Indeed, Amos Bairoch, a biochemist who heads SwissProt, believes that errors in the annotation that accompanies sequence data are more worrisome than errors in the sequences themselves. Five years ago, he notes, geneticists mostly sequenced genes whose functions they already knew. Now the reverse is true. "The [genes] that have been characterized are a very small island in a flood of [sequence] data," says Bairoch.

Researchers take a first stab at figuring out an unknown gene's function by running its sequence through computer programs that suggest what type of protein it codes for, based on how closely its sequence resembles that of a known gene. But the computer programs can be tripped up, because most proteins consist of several parts, or domains, that have different roles. One may bind to DNA, say, while a second attracts another type of protein, and a third catalyzes some chemical reaction. Protein A may look like B

CREDIT: TERRY E. SMITH

because it has a similar catalytic domain and thus is assumed to have the same function, say as an alcohol dehydrogenase. Later, a new protein, C, looks like B, again because the sequences share a degree of similarity, and so the computer program assumes that C is also an alcohol dehydrogenase. But this time, the similarity might be between B and C's protein-binding domains, not their catalytic portions, and C may not be an alcohol dehydrogenase after all. Down the line, a fourth protein, D, that resembles C will also be called an alcohol dehydrogenase. Yet, in reality, "you now have no clue what the function is," White points out, and it becomes harder and harder to figure out where the assignment went wrong.

Peer Bork, a computational biologist at EMBL in Heidelberg, Germany, estimates that about 15% of the annotation in GenBank is either unverified or not up to date, and "the error propagation is explosive," he says. And these errors are showing up in print as well, Bairoch notes. At SwissProt, for example, a team of some 40 Ph.D. researchers, each with a year's training in interpreting sequence and annotation data, are finding an alarming number of published reports with problems similar to those seen in GenBank.



Digital resources. The World Wide Web offers many tools for cyberspace biologists.

Even seemingly simple information, such as where a gene begins and ends, can be wrong: Complex genes often code for more than one protein, depending on where the DNA-transcribing enzymes start and stop, and the different proteins often have different functions. "As many as a third of the genes are alternatively spliced," says Sylvia Spengler, a biophysicist at Lawrence Berkeley National Laboratory in California. Charting these alternative ways of reading a gene "is important," she adds, "but we don't yet know how to deal with that [in databases]." Both NCBI and Spengler's team have begun experimenting with ways to indicate when genes code for more than one protein, but it's still uncertain how clear these kinds of annotations will be.

Interpreting electronic "literature"

The jury is still out on the best way to tackle these problems. Automated screening systems can catch some of the errors, such as contaminant sequences. GenBank's Kitts, for example, has compiled and updated a list of possible contaminating sequences—more than 2000 have been identified to date—and perfected a computer program that flags these sequences in new submissions to the archive. EBI is taking similar steps. When researchers in Europe search for matches in EBI, a computer program automatically scans the data to screen out possible contaminants.

These efforts should cut down on the amount of contaminating sequence entering these databases and will help ensure that such sequences don't throw off genomics analyses. For other kinds of errors, there's no quick fix. Some experts, such as Nathan Goodman, a bioinformaticist at Compaq Computer Corp. in Marlboro, Massachusetts, would like individual scientists or groups of scientists to take responsibility for keeping the information about their particular genes of interest up to date. To foster this kind of care, he says, journals should publish notifications of these corrections, enabling researchers to get credit for their efforts. But databases would still have to develop ways for the corrections to be entered.

Others advocate new, intensively curated databases that would be the equivalent of review articles in the world of electronic literature, containing time-tested, authoritatively annotated entries. Several such databases are already being set up. NCBI's RefSeq, for example (*Science*, 30 April, p. 707), identifies one "best" sequence for each way the gene can be expressed, lists the gene's approved name and synonyms, and describes or links to functional information. Unlike GenBank, curators of this new database "may make editorial comments and corrections that the original authors don't agree with," NCBI computational biologist James Ostell points out, even though that may miff the original authors. Moreover, "the reference collection provides an armature on which we can put all sorts of information." Using a "Link Out" system, individuals or groups will eventually be able to insert pointers in GenBank that will direct users to sites with more, perhaps better, information about a particular sequence. Ostell thinks users will eventually go to RefSeq first, bypassing many of GenBank's shortcomings.

Other groups around the world are also setting up their own curated databases, some with specialized foci, others with the intent of annotating the human genome with their own customized tools. In this way, "there will be competing products; then the audience can judge which is most useful," says Boguski. TIGR, for example, downloads new sequence data daily from GenBank and has a software package that includes Glimmer, a program that trains itself to recognize genes in microbial genomes, and a program called AAT that matches new DNA data to sequences in protein and cDNA databases. Likewise, Chris Overton, a computational biologist at the University of Pennsylvania, Philadelphia, and his colleagues have come up with a set of annotation tools that use grammar rules and other aspects of computational linguistics to sort out genes and their function. They and others are working feverishly to make their sequence comparisons more sophisticated so as to improve the program's ability to predict function correctly.

Still others are positioning themselves to be go-betweens who can help researchers who don't use electronic resources very often to search GenBank data more intelligently and thoroughly, and they, too, have their own approaches to annotation. "The key for biologists," says John Bouck, a bioinformaticist at Baylor College of Medicine in Houston, Texas, "is to be able to understand what the limitations are but also to realize how much information is there." The Web-accessible Genotator, for example, combines several programs that find genes, look for matches between genes, or check for sequence that signals the beginnings and ends of genes, thereby enabling a gene hunter to explore a range of tools all at once. The GENOME CHANNEL, based at Oak Ridge National Laboratory in Tennessee, is a tool for a comprehensive sequence-based view of genomes. And courses and workshops are popping up to teach researchers what they can—and cannot—expect from their cyberspace explorations.

As ideas for fixing the databases proliferate and funding agencies step up support for bioinformatics (*Science*, 11 June, p. 1742), even the more skeptical researchers are expressing some optimism. "These [problems] are all solvable," says Goodman, "if there is enough will in the community to solve them."

–ELIZABETH PENNISI