As the public effort to sequence the human genome comes into the final stretch, its methods and goals are increasingly being challenged by private firms

In the Crossfire: Collins on Genomes, Patents, and 'Rivalry'

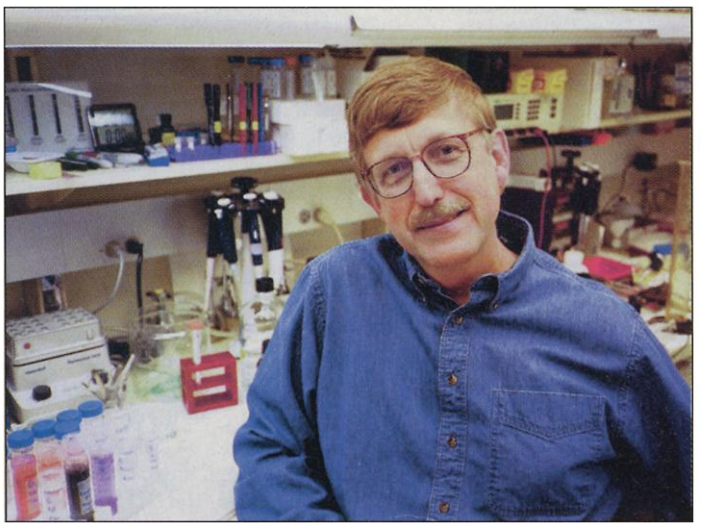
Francis Collins, head of the National Human Genome Research Institute (NHGRI), is in the hot seat. As the leader of the public effort to sequence all 3 billion bases of the human genome, he's helping steer the most ambitious and visible project ever undertaken in biology. And he is facing a huge challenge on several levels from a privately funded team led by J. Craig Venter, president of Celera Genomics in Rockville, Maryland. Both teams are racing to complete a draft of the human genome in the next few months, and both are engaged in a vigorous, and sometimes noisy, competition.

Just in the past few weeks, any hope of collaboration between the rival teams broke down, with both sides accusing the other of bad faith. At issue are the terms of data release: The leaders of the public project insist it be immediate and unrestricted; Celera wants those who use its data to agree not to redistribute them to others. On 14 March, the disagreement ratcheted up a notch when President Bill Clinton and British Prime Minister Tony Blair applauded the policy of instant data release. But the statement sent biotech stocks—including Celera's—into a nose dive.

Controversy is not new to the genome project. Indeed, 7 years ago when Collins, an M.D.-Ph.D., took the helm, many scientists were arguing that the project was too ambitious for its own good. It was folly, they argued, to promise the public that the entire human genome could be sequenced by 2005, the initial target date. Biologists bemoaned the entry into their field of "big science." Today, such concerns seem remote.

In 1995, NHGRI began to accelerate the effort, funding six pilot projects in high-volume sequencing. A turning point came in 1998 when Robert Waterston at Washington University in St. Louis, who is funded by NHGRI, and his collaborator John Sulston of the Sanger Centre near Cambridge, U.K., who is funded by the Wellcome Trust, announced that they had deciphered the complete genome (97 million bases) of the nematode, *Caenorhabditis elegans*.

Meanwhile, at The Institute for Genomic Research, a nonprofit in Rockville, Maryland, Venter was perfecting a faster "whole-genome shotgun" approach. He wowed the community in 1995 by producing the complete genome of the bacterium *Haemophilus influenzae* (1.8 million bases long) at record speed. In May 1998, Venter dropped a bombshell: Backed by PE Corp. of Norwalk, Connecticut, he launched Celera and announced that it would sequence the entire human genome by 2001 using the whole-genome shotgun method.



Collins and members of the Human Genome Project (HGP) consortium responded by speeding up their own timetable. In September 1998, they announced that they would produce a "working draft" by 2001, covering 90% of the human genome with one error per 100 bases. By 2003, the public consortium promised to deliver a 99.99% complete human genome.

In a meeting at his National Institutes of Health (NIH) office on 14 March, *Science* asked Collins to discuss his views on this controversy, the rivalry with Celera, and future priorities of the HGP. The following is a transcript of the interview edited for brevity.

Q: *Are you on target for finishing the draft human genome? Has your goal for completeness and accuracy changed?*

A: We are on target and expect to reach that goal this spring. In fact, we just passed the 2 billion base pair mark, which is about 70%. It took 4 years to obtain the first billion and 4 months to get the second billion. We're adding 10% of the genome to GenBank each month now. The goal for completing the working draft has not changed since it was first announced: 90% coverage of the euchromatic [informative] portion of the human genome sequence.

Q: *What do you think of the fruit fly sequence, recently published by Celera Genomics in collaboration with the Berkeley Drosophila Genome Project?*

A: It is a remarkable accomplishment—an extremely exciting opportunity to look at the complete instruction book of an organism that has occupied the minds of geneticists for 100 years. It will require another year of cleanup to close the gaps. But there is no question that as it stands now, it is of tremendous value to the community.

Q: *Does this prove that Celera's whole-genome shotgun strategy works?*

A: The paper tells you that, at least for *Drosophila*, with its 120 megabases of euchromatic DNA, the whole-genome shotgun method works quite well. Not perfectly. There are certainly places where the algorithm could not assemble the sequence; there are more than 1000 gaps. But it is an excellent proof of principle, though it also indicates that there's a lot more refining to do to make it practical for larger genomes.

Q: *Has the rivalry between Celera and the Human Genome Project gone beyond healthy competition?*

A: There are substantive issues about data access at the heart of the situation: Will the sequence of the human genome be freely

CREDIT: MARTY KATZ

accessible without restrictions of any sort to researchers in the private and public sectors, or will it not? Regrettably, relatively little of the press attention has focused on those bedrock issues. Far too much has been written about the personalities and the "rivalry."

Q: *Have companies complained to Congress or the Administration about your handling of data release?*

A: We have no way of tracking such complaints. But many companies have repeatedly expressed to NHGRI, the Congress, and the Administration their strong support of the Human Genome Project's data release policy. Congress remains strongly supportive of the HGP's data access policy, as evidenced by the recent remarks of Congressman John Porter [R–IL] at the NIH Hearing on 8 March, and Congressman David Obey [D–WI] 2 weeks earlier. As shown by the recent Clinton-Blair statement, the Administration's support for the HGP and its position on immediate data release could hardly be more enthusiastic.

Q: *You and others wrote to Celera in February that it would be unethical for them to publish the human genome without the approval of scientists who contributed to the public database. Why?*

A: From the point of view of publishing ethics, I think the sequencing centers that have done the labor ought to have the chance to say what they did for themselves—and to have the whole body of work go through peer review—before somebody else does it for them. For instance, it would have been outside the normal boundaries of publishing ethics for someone not in one of the labs that actually produced the *Drosophila* data to massage information from the public databases and rush a paper about the total genome sequence into print.

The other principle is that if you publish a paper, it traditionally means you've looked at the raw data and you can vouch for it. The raw data for sequencing is sequence traces that don't find their way into GenBank. (We wouldn't know where to put them all.) So if you are downloading a very large amount of data from a public database and publishing it when it hasn't previously been published, you're treading a bit upon that particular principle.

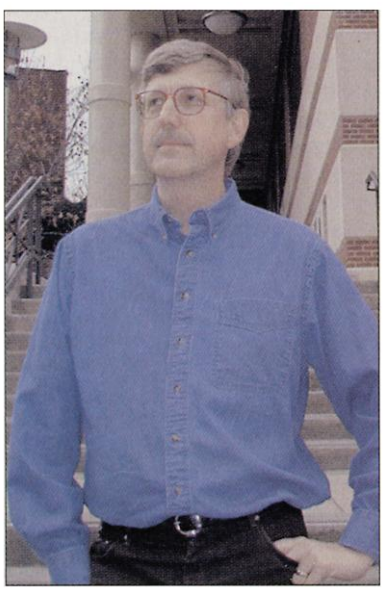
Q: *How do you respond to biotech investors who think the Clinton-Blair statement urging the release of genome data hurt them? What was the statement's purpose?*

A: In retrospect, most analysts agree there was nothing in the text of the Clinton-Blair statement that justified the market reaction. This was an exhortation to take genomic data—the raw fundamental stuff—and make it publicly accessible without restrictions. Part of the exhortation was to say, 'Let's get all the sequence into the public domain so that if patents have not already been filed, it will be less attractive to do so.' We've probably got enough patents filed already. ... Thousands of patent applications are sitting in the Patent and Trademark Office [PTO] right now waiting to be decided upon. ... Patents covering large numbers of genes could turn out to be quite deadly to the future of genomics if licenses are negotiated in an exclusive way.

Q: *You've said you support responsible patenting. What is that?*

A: I think the Patent Office deserves credit for moving toward a stronger requirement for utility. Several years ago, it looked as if any DNA sequence would be considered useful because it could be used as a probe. Now, in their proposed new guidelines, PTO says such a claim would not be specific enough to demonstrate utility.

The Patent Office is seeing fewer of what they call "generation one" patents, where there's just a sequence and no clue as to what it does. PTO intends to reject those. They are seeing a reasonable number of "generation two" applications, where there's a sequence, and homology suggests a function. NIH views such applications as problematic, since homology often provides only a sketchy view of function. Increasingly, PTO is seeing more in the "generation three" category, which I think most people would agree is more appropriate for patent protection. These are gene sequences for which you have biochemical, or cell biological, or genetic data describing function. So we are seeing a shift in the sensible direction.



CREDIT: MARTY KATZ

Q: *Will sequencing grants soon become a lower priority for the program?*

A: Sequencing will continue to have a critical place in the armamentarium of genome activities for at least the next 5 years. Understanding what the sequence means will require us to make multiple comparisons. For that reason, we are already plunging into the mouse genome, which is every bit as hard as the human. We'll do it differently in terms of the strategy, but it will still require around 60 million reads to get it done. That's a lot of work, and there's no way around that now. The arguments are quite strong for sequencing other mammals besides human and mouse. Having two genomes to compare will be useful, but having three would be really useful, particularly if you're looking for smaller conserved elements that are involved in regulation of gene expression. We will clearly not want to stop the vertebrate genome list after the mouse. The zebrafish and the rat genome will be highly useful to sequence. There will be strong arguments for doing the pig, the dog, or the cow—and for doing another primate. ... Whether other vertebrate genomes will need a full finished genome sequence, or whether most of the value could be derived from a draft, is still being discussed.

Q: *Why not offer contracts to sequence these genomes efficiently?*

A: The contract model, one might argue, is the most efficient way to get sequencing done, if we are really moving into a production mode. But a compelling argument can be made for doing large-scale sequencing at academic institutions, where it can have lots of useful spin-offs. ... If you had this activity segregated off in an industrial atmosphere, you would lose something both in terms of training and in terms of other research ideas. Here the Genome Institute will need to be heavily guided by what the biological community says the priorities ought to be.

Q: *How do you plan to encourage new ideas for genomics research?*

A: A major new initiative, approved by our council last month, is to establish Centers of Excellence in Genomic Science. We believe that ideas about technology development, computational approaches, population genetics, expression analysis, and proteomics are most likely to bubble up in academic environments where multidisciplinary teams of several investigators are focused around a common theme. We would like to see a lot of our research funding shift into that mode. ... The annual budget cap will be in the neighborhood of $4 million to $5 million, with an initial grant period of 5 years. If these centers are to have the kind of stability that encourages people to take risks, we have to give them a longer lifetime than the normal 3-year genome grant. Center grants would be renewable for one cycle, but after

roughly 10 years, they would need to find other sources of funding.

Q: *Where does the genome program most need support right now?*

A: I would say in bioinformatics. That poses a real challenge, given the paucity of trained individuals who are expert in both computational methods and biology. And, boy, do we need them. The talent pool that does exist has migrated heavily to the private sector, and that has injured the ability to train the next generation. I think the scientific community is really revved up to solve this problem. When I talk in academic institutions, undergraduates or beginning graduate students come up to me and ask how they can get into computational biology. They can see this coming. We just have to be sure that we're providing them with superb training experiences. One of the intentions of these centers, besides doing great science, is to provide such great training opportunities.

Q: *People expect the genome project to yield benefits soon. Are you disappointed with the clinical payoffs to date?*

A: No, I am not at all disappointed, but I am impatient. Anybody who has thought about the path from gene discovery to clinical use knows that it includes complicated and unpredictable steps. ... The Herceptin story is a good example of how molecular understanding of breast cancer has led to a therapy that has significant clinical benefit. But it's hard to point to a long list where we have a home run. ... We have to be realistic that the full flowering of the medical benefits of understanding the human genome probably lies 15 to 20 years away. The message we have been trying to convey is that this is the greatest revolution medicine has experienced since the introduction of antibiotics, but it's not a revolution that happens overnight.

Q: *Was gene therapy hurt by the death of a patient reported last fall?*

A: Everybody is quite shaken. ... A young man has lost his life, and this tragedy has rocked the field down to its toes. It is sobering to consider that this approach not only hasn't led to cures in the past 10 years but is actually capable of harm. ... It has caused everybody to tighten up a lot. I think we'll get past this. Gene therapy will find a very important niche for the treatment of a number of diseases in 10 or 15 years.

Collins was interviewed by Eliot Marshall, Elizabeth Pennisi, and Leslie Roberts.

EPIDEMIOLOGY

When an Entire Country Is a Cohort

Denmark has gathered more data on its citizens than any other country. Now scientists are pushing to make this vast array of statistics even more useful

For years, any woman who got an abortion had to accept more than the loss of her fetus: For some unknown reason, she also faced an elevated risk for breast cancer. At least that was what several small case-control studies had suggested before Mads Melbye, an epidemiologist at the Statens Serum Institute in Copenhagen, undertook the largest effort ever to explore the link. He and his colleagues obtained records on 400,000 women in Denmark's national Abortion Register, then checked how many of the same women were listed in the Danish Cancer Register. Their foray into the two databases led to a surprising result: As they reported in *The New England Journal of Medicine* in 1997, there appears to be no connection between abortion and breast cancer.

Their success underscores the value of a trove of data the Danish government has accumulated on its citizenry, which today totals about 5 million people. Other Scandinavian countries have created powerful database systems, but Denmark has earned a preeminent reputation for possessing the most complete and interwoven collection of statistics touching on almost every aspect of life. The Danish government has compiled nearly 200 databases, some begun in the 1930s, on everything from medical records to socioeconomic data on jobs and salaries. What makes the databases a plum research tool is the fact that they can all be linked by a 10-digit personal identification number, called the CPR, that follows each Dane from cradle to grave. According to Melbye, "our registers allow for instant, large cohort studies that are impossible in most countries."

Beauty in numbers. These Danish twins starred in a variety show at the turn of the 20th century; now it's their medical records, part of a database, that are in demand.

CREDIT: PHOTO COURTESY OF THE DANISH TWIN REGISTER, UNIVERSITY OF SOUTHERN DENMARK

But Melbye and other scientists think they can extract even more from this data gold mine. They argue that not enough money is being spent on maintaining and expanding existing databases, and they say that red tape is hampering studies that require correlation of health and demographic data. The problem is that, while they have unfettered access to more than 80 medical databases maintained by the Danish Board of Health and public hospitals, their use of 120 demographic databases overseen by the agency Statistics Denmark is tightly restricted. Statistics Denmark won't allow researchers to remove from its premises data coded by CPR, and the procedures for accessing information at all are unwieldy and expensive.

Statistics Denmark officials are reluctant to release data tied to CPRs, citing privacy concerns. "The public should have confidence that information identifying them as individuals does not reside outside of this institution," says the agency's Otto Andersen. Last month, Danish research minister Birte Weiss formed a committee to break the impasse. Denmark's databases are "a resource which can be used more optimally," she told *Science*. "This should be a scientific flagship."

Working the health databases can yield powerful results. For years the U.S. National Institutes of Health has supported a study following twins, hoping to tease out the relative contributions of genes and lifestyle to aging. Led by University of Southern Denmark gerontologist Kaare Christensen, the project has tapped the Danish Twin Register, which includes 110,000 pairs of twins born since 1870. After following more than 2000 pairs of twins aged 70 or older, Christensen's group has so far tied to genes about a quarter of the variation in human longevity. "The project is made possible by the unmatched age and completeness of the Danish Twin Register," he says.

The health databases have proven invaluable for probing contradictions raised by smaller studies and following disease pro-