THE CANCER LETTER INTERACTIVE

PO Box 9905 Washington DC 20016 Telephone 202-362-1809

Vol. 27 No. 7
Feb. 16, 2001

On "Day One Of Post-Genome World," Eric Lander Recounts 15-Year Journey

The day after the announcement of the publication of the human genome, Eric Lander faced the National Cancer Advisory Board.

"I'm very happy to be here to celebrate Day One of the post-genome world with you," Lander, head of the Whitehead Institute Center for Genome Research, said Feb. 13, as he recounted the 15-year history of the Human Genome Project.

Lander's NCAB appearance wrapped up a frenetic weekend dénouement to the pre-genome world. In two days in Washington, Lander had ushered in the birth of the post-genome world with media interviews, a press conference, a scientific symposium, and a "genome party" at the National Building Museum. At the party, a band of NIH scientists called The Directors, led by National Human Genome Research Institute Director Francis Collins and NCI Director Richard Klausner, played guitars and sang their own post-genomic lyrics to mid-20th-century rock 'n' roll.

So, if Lander, the lead author on the lead paper published in Nature this week, still wanted to celebrate on Tuesday morning, no one in NIH Building 31 Conference Room 10 intended to stop him.

Just as a participant in a revolution may emerge unscathed, though permanently altered, with an urge to tell the story to anyone who will listen, Lander giddily told the story, tangents and all.

Following is an edited transcript of Lander's NCAB presentation:

Not just one paper, but 20 different papers came out yesterday reporting, first, the primary sequence of the human genome, analyses of that sequence, and, in some very real sense, declaring the beginning of the post-genome world. In Nature, David Baltimore does indeed declare the start of the post-genome world....

I expected nobody would show up at the Capital Hilton for the press conference [Monday], because the embargo had been blown on this story Saturday night in London by a reporter for the London Observer, who had gotten some quotes at a meeting in Lyon and decided that he could run with the story. Saturday night was filled with phone calls flying back and forth, and one of the journals [Science] decided it would unilaterally lift its embargo, so Nature had to lift its embargo. We found ourselves confronted with stories in the Sunday morning papers. So who's going to show up on Monday morning?

The room was packed. They couldn't get enough of this....

Then, we got a cab and came to the NIH for an even more important unveiling, an historic symposium. It started after a spectacular audio-visual

(Continued to page 2)

Special Report: The Human Genome Project

Web Links:

DNA sequence of the human genome is accessible from the National Center for Biotechnology Information, National Library of Medicine: http://www.ncbi.nlm.nih.gov/genome/guide/human

Nature: http://www.nature.com/genomics/human/

Science: http://www.sciencemag.org/genome2001/

Feb. 12 press conference may be viewed at: http://videocast.nih.gov

Scientific lecture series, Insights from the DNA Sequence of the Human Genome: http://www.nhgri.nih.gov/CONF/genome_insights.html

Lander Celebrates "Day One" With Presentation To NCAB

(Continued from page 1)

look-back at the last 15 years of the Human Genome Project, with talks by Francis Crick and James Watson. Crick, who can't travel at the moment, nonetheless videotaped the remarkable talk in which he described what they were thinking would happen in 1953. He said he had no idea how fast all this would happen. He thought even working out the basic details of messenger RNA and how proteins are made would be the work of the second half of the 20th century, not the work of just the decade or so that followed him. No conception that one would have the sequence of the human genome by the end of the century....

Then we decamped to the National Building Museum to the party where we partied all night to celebrate what was for me, 15 years of work that began in a couple of places, but one of them was right here in this room in 1986. I was quite a kid then, barely knew biology, and I was asked to come to a meeting in this very room, which was the first NIH meeting about the Human Genome Project.

It was really good that nobody in Congress asked, "Do you know what you're doing? Do you know how you're going to do this?" This was very analogous, in terms of technical ability, to going to the moon, when [President] Kennedy said we're going to get there by the end of the decade. If you pressed anybody as to exactly how we were going to get there by the end of the decade, they had no clue, but they had great confidence that, if everybody pooled their resources, we'd somehow get there.

It was the same thing. No one had a clue how this was going to happen, back in the mid-1980's, but through the real power of science, scientific cooperation, scientific organization, it was broken down into steps, biological, technological, computer science, and the problem was deconstructed, solved, put back together, and—though by no means are we done—it is no longer a rate-limiting step in human genetics. That barrier has now been leveled.

The Book of Life—Free!

It's a very exciting thing. It's going to make a difference scientifically, and it was really satisfying to find out last night that it already is making a difference in public consciousness.

As we were trying to go over to the Building Museum, we had to catch a cab. It was not easy, the Metro was down, it was raining. My daughter finally managed to get us a cab. We got in. We were clearly racing, we were late.

The cabbie looked at us all dressed up, and said, "Are you going to a party? What party?"

"It's a genome party," we said.

"Oh! The 23 chromosomes! The book of life! Information is power!"

My jaw dropped. Then, it knocked my socks off when he turned around and said, "Are you the guys who gave it away for free?"

I said, "Yeah."

That felt so-o-o good.

1985-99: Building A Foundation

There have been a lot of systematic components to trace disease genes by tracing their inheritance in families. Once you localize the disease gene, you've got to then narrow down the disease gene to a particular small piece of DNA, which in the old days meant marching along the chromosome in a tedious fashion. It could take years and years to do that.

People had dreamed that we should be able to go to the shelf and pull down those pieces of DNA. That's a physical map, and those began to be completed in the mid-1990's.... The real dream was, we wouldn't have to analyze the region, sequence the region, find the genes in the region. In fact, what we ought to be able to do is to double-click on a region. That is the sequencing phase of this.

THE CANCER LETTER

Member, Newsletter and Electronic Publishers Association

World Wide Web: http://www.cancerletter.com

Editor & Publisher: Kirsten Boyd Goldberg
Editor: Paul Goldberg
Editorial Assistant: Shelley Whitmore Wolfe

Editorial: 202-362-1809 Fax: 202-318-4030
PO Box 9905, Washington DC 20016
E-mail: news@cancerletter.com

Customer Service: 800-513-7042
PO Box 40724, Nashville TN 37204-0724
E-mail: info@cancerletter.com

Subscription $295 per year worldwide. ISSN 0096-3917. Published 46 times a year by The Cancer Letter Inc. Founded Dec. 21, 1973, by Jerry D. Boyd

That began as a pilot project in 1996. That three-year pilot project was meant to lay the groundwork of tools and ideas. It finished in 1999. On the strength of that, a large scale-up was initiated at NIH and other organizations around the world. It was done in cooperation with groups in England, France, Germany, Japan, and a late entrant, but every bit as important, China. The genome center in Beijing joined in the last year and a half and contributed 1.5 percent of the genome. We're delighted, because this is very much the heritage of all of humanity.

It meant over the course of that time, a massive change in the way we do biology. When we started, we used single-channel pipetters, moving around one reaction at a time. The big advance was eight-channel pipetters, then 12-channel pipetters. Then lab robots that would move around and do things for you.

At the Whitehead Institute, we have six large robots going around picking bacterial clones. They have 120,000 clones a day to analyze. They have to be purified, and sequencing reactions have to be set up. Instead of the armies of hundreds of people who might do that, the work is done at the center in what resembles a factory production floor. It consists of conveyer belts, moving around microtiter plates, adding solutions. This is run by about eight people. [The data] gets put onto commercial sequence detectors, and they produce about 65 million letters of DNA sequence per day. It's a very different world from the one we had 15 years ago.

When we scaled up in 1999, it was a remarkable year. In 12 months, the amount of DNA sequence skyrocketed. Why? Because a firm foundation had been laid during the three-year pilot project.

1999: "A Very Hectic Year"

Ah, there is nothing like being left alone to do your science. And this was nothing like being left alone to do your science. It was a very hectic year. We found ourselves waking up every morning to stories about the Human Genome Project.

As you may or may not have heard, there was a company [Celera Genomics] that decided it was going to give us a go for this thing. There was a conclusion that there had been a miscalculation. The fact that we were a pilot project meant that the public effort was slow and pokey, rather than working out its methods. On the strength of the notion that somehow the public project was moving slowly, the company declared that they were going to do the whole thing by some other method and get it done very, very, very quickly.

There's no doubt that that certainly spurred folks into action more quickly. I think it ended up accelerating the time table by six to nine months. We are here six to nine months sooner than we would have been without that acceleration. But it was also a rather stressful period. There were press releases flying around that were very hard to deal with, about what had or hadn't been done.

Yesterday was a very satisfying day, because we could find out finally what had been done. There was a paper published by the private group [Celera] which came out in the magazine Science yesterday, and I will say, and I am absolutely delighted to finally have the data to say it, the assembly of that paper is largely an assembly of the public data. Sixty percent of the data in that paper comes from the public, all of the mapping information comes from the public, the strategy that eventually worked was the public strategy. I think the NIH deserves tremendous credit for having stuck to its guns in the face of criticism, and soldiered on, stayed the course on the project, because in the end, it was the public data that led to both papers. I'm very proud of that. It was not an easy time, but we made the right decision.

Last spring was a tough period. So a brilliant solution was hit upon. We would just declare victory. Everybody decided that there had been so much battling going on that the thing to do—I think this was Sen. [George] Aiken's strategy with respect to the Vietnam war—was to declare victory and withdraw. We were going to declare victory, but we didn't intend to withdraw. So there was a huge celebration at the White House last June 26, when it was announced in banner headlines that the human genome had been completely sequenced.

Of course, the human genome hadn't been completely sequenced, it was a draft sequence. There were still holes, there was still work. But there was something fundamentally right about those announcements, which was if you clicked on the Web, the vast majority of information was totally, freely available from this public consortium, if anybody wanted to download it. It was worth celebrating.

About 900 Gaps Left

We didn't have a clue what it said. We had a huge pile of letters sitting over in the corner and by statistical tests we knew we were kind of done, but we were breathless by the end of this period. We took several weeks to catch our breath and sit down with this text of 3 billion letters and flip through….

Having had six months of work with the most remarkable set of colleagues you could ever hope for, the Genome Analysis Group, 50 of the world's best computational biologists who came together in a completely free-flowing consortium over the course of these six months, exchanging ideas in over 5,500 emails, in conference calls twice a week, two hours each, we managed to put this thing together and learn a little about some of the cool stories in this anthology.

The genome was assembled chapter by chapter. Individual chapters were sequenced in different parts of the world. The sequences of those chapters were pasted together into units of the book. There are still gaps. There are gaps where we're missing a couple hundred letters. The major gaps that we care about are those where we don't already have in our freezer the DNA needed to close that gap. There are now about 900 of those....

The punch list is about 900 spots in the genome. Basically on the order of about 40 spots for a typical chromosome, where we're going to have to go in and fish out the new piece of DNA we don't already have. That's not trivial. My guess is that a couple hundred of those are going to be really hard….

The final assembly order seems to be quite good at the large scale…. The data are not perfect, but they're not bad: 91 percent of the data in the databases are accurate at the 99.99 percent level. Every letter in the sequence has an accuracy attached to it. 96 percent of the letters are accurate at 99.9 percent….

Then we kept annotating the sequence, adding on layers and layers of information, coverage of genes and RNAs and similarities to other organisms like puffer fishes, laying on information about the repeated sequences in the genome and how different parts correspond to the chromosomes. All of this can be obtained from freely available browsers on the Web. There is a browser at University of California, Santa Cruz, where you can download all this information. There is a browser in Europe at the European Bioinformatics Institute, where you can click on this information. For anybody with really good eyesight, in your issue of Nature there is a very long fold-out that has the human genome with all these features marked, but I warn you, it is at the scale of 3.3 million bases per centimeter. So, it's meant artistically for your wall. You need the browser if you want to do anything about interpreting the genome.

A Lumpy Genome, Lots Of Interesting "Stuff"

As we begin to look at the genome landscape, it is incredible. We have the lumpiest genome that's ever been seen before. It's not uniform in any way. There are vast tracts which are enriched for G's and C's. On average, we're about 41 percent GC content, but it varies widely. We have huge areas where it goes up and huge areas where it comes down. The ups, the GC regions, are very gene-rich. Then there are deserts that are gene-poor. We have repeat elements all over our genome and they, too, have preferences as to what parts of the genome they want to live on, whether it's the mountain-tops or the valleys.

Even in a microscope, when you look at chromosomes and you see light bands and dark bands? It turns out that the light bands correspond to the gene-rich, GC-rich regions. In one of the accompanying papers that appears in Nature, there is the product of an international consortium that has taken 700 or 800 of the clones that we sequenced and mapped them onto to the cytogenetic map of the chromosomes. [See http://www.nature.com/genomics/human/papers/maps.html]

I may not have to tell you how incredibly important that is for cancer, because in cancer there are all sorts of chromosomal rearrangements, and people might spend years working out how a particular rearrangement seen in the microscope corresponds to a change seen in the sequence. But now that we've mapped those banding patterns at a sequence level, we have much higher resolution to be able to zoom in on the relevant regions in such rearrangements.

I'll tell you about some of the cool findings:

—Most of your genome is not genes. In fact, only about one to one-and-a-half percent of your genome turns out to code for proteins. The other 98 percent of your genome is not coding, and I bet only about 3 percent of your genome is either coding regions of genes or regulatory regions of genes. The rest of it is stuff.

This stuff consists primarily of ancient repeat elements, DNA that has moved around the chromosomes over the course of more than a billion years, and accounts for most of our sequence. It's sometimes referred to as junk DNA, as if it's uninteresting, but, in fact, it tells amazing scientific stories. We are extremely rich in repeat DNA. More than 50 percent of our DNA is clearly traceable to these transposable elements, and I bet much of the rest is, too, but it has degenerated too far by mutation to pick it up readily.

By contrast, the mustard plant, the worm C. elegans, the fruit fly, all have much smaller bits of

repetitive DNA.

—Your repeat sequences come in four flavors. They're called LINEs, SINEs, LTR retroposons, and DNA transposons. They each know how to move around the genome.

LINEs make copies of themselves into RNA. The RNAs float off, get used to make proteins. The proteins grab the RNAs, take them back to the nucleus, cut the chromosome and copy them back in. Very clever self-sufficient elements that have been around for a billion and a half years. They are probably the most successful invention in the whole eukaryotic world. The little SINE elements make copies into RNA, but encode nothing. The only thing they seem to be good at was getting the line proteins to move them. So they are parasites on parasites.

Then these LTR things, they are thought to be the origins of retroviruses. These elements encode a 'gag' and 'pol' as found in retroviruses, but not a cellular envelope gene. It's thought that these elements first worked out how to transpose themselves around using gag and pall, then picked up a little coat, and once it puts its little coat on, it can move between cells and organ systems. This is probably the origin of retroviruses.

Then these DNA transposons, they don't move through RNA. They move through making a protein that cuts out their DNA and moves them around.

—Well, here's the cool fact. We now found the fossil record. Everyone of these things that hopped around the genome hopped on a particular day, maybe 35 million years ago, and came from a particular active element, with a particular sequence. But all the elements that came from that active element, we can recognize because they started with identical sequences, and then began to degenerate. We can see that they are cousins. We can build a whole family tree out of these elements. So we've taken all 3 million repeat elements across your chromosomes and built a family tree out of them, so we know which ones are related to which ones, and can figure out when they hopped. Which ones hopped in our common ancestors with fish, which ones hopped in our common ancestors with mice, which ones hopped in our common ancestors with chimpanzees, and which ones hopped more recently.

It turns out our repeats are very old. More than half of them hopped before our divergence from mice, but some amazing findings come out. For one thing, there is a dramatic decrease in the hopping rate lately. In the last 30 or 40 million years, the rate at which these things are moving has plummeted. One of these four types has gone extinct. Another one of these four types on the brink of its extinction. We can find at most three possible active elements for the LTR family and they may not even really be active.

There is a serious ecological problem: transposable elements seem to be dying out in our genomes. We don't know why. They aren't dying out in the mouse genome. This takes some pretty fancy explaining. We don't quite know the answers. I can wildly speculate for you about why the hominid and rodent lineages should be so different, but since I have no proof, I won't inflict that on you today, but it has to do with population genetics.

—There's another weirdness I have to tell you about. The repeats all have preferences where they want to live. If you were a repeat that was trying to work out a détente with the organism you were parasitizing, if the genes live over here in the GC-rich regions, it might be sensible for you to go to the AT-rich regions, because then you wouldn't interfere with the genes, and you'd cause less harm to the organism. So LINEs do that, they all go to the AT-rich regions. The LTR retroposons, they do that. The DNA transposons, they do that.

But the SINE elements don't. They pile up near the genes. That's weird. Why would they want to do that? It's weird, because how in the world can they do that, they have any ability to transpose on their own. They get moved by the same proteins that move the LINEs. How do they end up in a different place? Are they so smart that they can reprogram the protein to put them somewhere else? Or is it that they actually land in the same place LINEs do, and evolution actually likes to retain those that land near genes?

We know the answer. We can just look at the sequence and the sequence tells us, because we can sort them by how old they are. We can look at newborn ones, middle-aged ones, old ones, and if the newborn ones look like they are in the AT-rich regions, and as they get older and older, they switch to the GC-rich regions, evolution must be favoring those.

Darned if that isn't exactly what we see. Evolution is reshaping the distribution of these apparently useless, parasitic elements, sufficiently so, 13-fold over 25 million years, that I think it's really tough to call these things useless, parasitic elements. I believe that they are actually symbiots in our genome who earn their keep in our DNA by serving a useful function. This function is probably related to regulation of protein translation.

Fun With Your Genome

—Here's another cheap thing you can do just by reading the history, the paleontology of own chromosomes. If repeat elements hop, and some hop up to the X chromosome, and some hop up to the Y chromosome, we can follow their fates. It's epidemiology, isn't it? It's a cohort study. We look at a cohort that hopped onto either the Y or the X chromosome and we see what happened to the cohort.

Well, it turns out that landing on the Y chromosome is bad for the element in that it picks up a lot more mutations than if it had landed on the X chromosome. The Y chromosome has a much higher mutation rate. Why? Y chromosomes always pass through the male. X chromosomes, two-thirds of the time, pass through the female. This tells us clearly that sperm are twice as mutagenic as eggs. Two-thirds of all mutations happen in men, rather than women.

This has led to wonderful discussion, because the women are saying males are responsible for two-thirds of all the genetic defects that arise, and the males I know are saying that males are responsible for two-thirds of all evolutionary progress. I remain a conscientious abstainer.

We can tell all sorts of things. We can tell the mutation rate is different across the genome and we have interesting ideas about why that's so. We can tell that about 50 genes in our own genome really come from transposons passing through that left a gene for us. We can find very recent elements that have hopped in our genome that are so recent that they're still polymorphic, variable in the population, and they provide incredible markers for tracing human migrations, because these are completely unique events. A transposon lands, and all the people who have it have to share a common ancestor for that region. With a thousand of these events now identified in the human genome sequence, we can reconstruct population migrations in a much more powerful way.

Fewer Genes Than Predicted? The True Story

We also did a careful analysis of the genes in the genome. We did a careful analysis, but the results are difficult to interpret, because the human turns to have only about one and a half percent of the genome encoded. So the state of the gene collection is a little ragged. It is not perfect by any means, but nonetheless, we have a very good handle on how many genes are in the genome. It will probably be about a year or two before a really clean gene collection is available. But we have a good handle, and it's an interesting surprise.... We find only about 35,000 genes.

First off, this is about three-fold fewer genes than I'd been teaching my students in Introductory Biology for the last 10 years. We had been advertising about 100,000 genes in the genome. We were very concerned about where in the world this discrepancy came from. I can happily report that we tracked down the source of this.

The scientist Wally Gilbert at Harvard was responsible for that number of 100,000. I called Wally and got the story. Wally said, "Typical gene is 30,000 letters. Genome is 3 billion letters. Divide one into the other and you get 100,000. Seems about right."

The number was so—round—that it stuck.

Wally is a former physicist and this is exactly what a physicist does. It's a back-of-the-envelope estimate. I told Wally, "You know, it really turns out to be 35,000." He said: "So we got it right!"

As Wally points out, to a physicist, that's half an order of magnitude. That's just fine! So in point of fact, it's just that everybody else misunderstood that, in fact, this was only meant to be good within a factor of three. Nonetheless, this is the big news you're reading about. The reporters don't really have the full story, but it's a great story.

It is very interesting, because it's actually less than twice as many genes as the lowly nematode soil worm C. elegans and maybe only two and a half times as many genes as a fruit fly. This certainly does seem like an affront to human dignity, that we should only have about twice as many genes, and we've been scrambling to try to explain ourselves.

We can explain our apparently greater complexity by some observations. It looks like a human gene probably has about two or three times as many alternatively spliced forms as a fly or worm gene does. That is, we make more products out of the genes we have. Most of these splices occur in protein-encoded parts, so that they probably result in more proteins.

Relatively little innovation has occurred in invertebrates with respect to the invention of new protein domains. That's not a surprise. Back in the primordial ooze, evolution had time to sit around and fashion new genes. They didn't have to work so well, because what was the competition, anyway?

But once life gets better and better at doing this, evolutionary competition picks up, and genes have to work from the get-go. So the only way to make genes is out of existing, previously known domains. It's analogous to why it's possible to start a new

automobile manufacturer at the beginning of the 20th century, but we can't start a new manufacturer at the beginning of the 21st century, because the niche is already filled with reasonably active competition. So the only way to make a new auto manufacturer today is to recombine pieces of existing automobile manufacturers, which is what is done.

We find a lot of new parts are put together in a lot of new ways. We have twice as many architectures as flies and worms. Our innovation, while it's derivative—it is after all, just recombining old pieces in new ways—appears to be quite powerful. Intracellular signaling mechanisms and cell-surface molecules seem to be made through this diversity of architectural combinations that we're beginning to understand.

The simplest way to make more genes is to simply copy existing genes and let them slightly diverge. If we look in our genome, there's a big pile of smell receptor genes. About a thousand smell receptor genes we stumble upon in the genome, indicating that our ancestors were really into smell. But then, disappointingly, we've discarded about two-thirds of them. Two-thirds are broken genes. They have mutations in them, indicating that while the vertebrate spent a lot of time building up his repertoire of smell receptor genes, in the hominid lineage, we've decided to discard most of our smell receptor genes, probably in favor of sight....

You can tell stories and stories about what's going on in our genomes by looking at the nature of our genes. We can also find out that about 223 of our genes we didn't invent at all, but got from bacteria. They probably came by horizontal transfer, by a bacteria injecting DNA back when we were fishes....

Human Variation

We've got to experiment and test all the things I've said, but everything I've told you comes from just peering at the sequence, learning how to read the book.... What's happened is a very different way of doing science. First you look at the huge amount of data and see what it's kind of telling you.

One place you can do that is population genetics. Another paper in this issue of Nature was looking at human variation. We're reporting just under one and a half million variants in the human genome. We want to understand these variants. What's the nature of human genetic variation? Which of these differences cause disease, and how are we going to make that connection?

We have a few common variants. That reflects the fact that we are descendants of tiny population of maybe 10,000-20,000 individuals, from Africa, about 5,000 generations ago. We are all very close cousins by comparison to most species. We are a tiny population grown large in the blink of an eye. We are 99.9 percent identical. If you take two random people on this globe, they are more close genetically than two random chimpanzees in Africa are to each other. We are a very close, very recent species.

That means that each gene has only a finite number of flavors it comes in. That has tremendous implications for medicine. It means that we can identify the flavors that account for 98 percent of the copies of genes….

Suppose we knew all the variants, three variants, let's say, for each of the 35,000 genes. Suppose we test each of those in people with diabetes, arthritis, stroke. It may boil down to a very simple matrix like that. We're not that far off from having all those variants. We can test them directly and indirectly. We're learning a lot about how to do this based on how big ancestral segments are. They are bigger than we thought, which is good news, because we are going to need fewer markers. They differ somewhat between populations, but they have nice, quantum, chunky structures that I think will lead in the next two years to the development of a linkage disequilibrium map, a map of ancestral segments that will allow us to trace disease in a powerful way.

Obviously, knowing all the components leads to a tremendous advance in cancer genomics. This is a major initiative of NCI. If we know all the components, we can start asking broad molecular questions about tumors....

Importance Of Building Infrastructure

This was the first project where biology said, "We need large-scale infrastructure-building to be able to propel our science forward, and we will have to get together in new forms of organization to build that infrastructure."

There's now a considerable list of infrastructure-building projects that must follow: Finding the rest of the rest of the genetic variations; sequencing other organisms to line them up and see these regulatory elements; producing the full-length cDNAs, which is an initiative led by NCI; being able with those full-length cDNAs to express all the proteins in the body, so that when a drug company wants to study a small molecule and figure out what its target is, you have

access to all those proteins. When it has a side effect, you can figure out all the things it binds to so you can figure out what might be causing the side effect.

Our goal in medicine is not to be shooting in the dark anymore. The only way not to shoot in the dark is to have the lights turned on really, really well. We have turned on the lights with respect to the sequence of the human....

In the early days of the genome project, there were worries. Is big science going to distort little science? To the extent that this is big science, this is big science in the service of small science. This is building tools so that everyone who has a smart idea is not limited by anything other than their imagination.

There's a lot of work to do, and though NCI is not the lead agency for the genome, there is no doubt that it has been the second lead agency with respect to the genome, because all the other initiatives that have propelled the other key components of infrastructure in the last couple of years, the cDNA collections, the CGAP program, NCI has shouldered that burden. So while you guys aren't officially responsible for the genome, you are, in fact, responsible for much of the genome, and you have a lot of responsibility going forward as we think about how we deliver on the promise....

Q & A: Private Sector Role To Add Value

LARRY NORTON (NCAB member): I read somewhere I can buy about 120,000 genes. Where can I buy them?

LANDER: You're referring to this advertisement that appears frequently in Nature Genetics or elsewhere that says, 'Lifeseek Database, 120,000 genes, including 60,000 not contained in anybody else's database.'

NORTON: Yeah, where do I buy them?

LANDER: As they say where I come from, *ver vaist*? I don't know. I've had disagreements with Craig Venter and Celera over methodology, but when we got up here yesterday at the press conference, we got no disagreement on the count of the genes. We can't find any more than that! Now, I'll say, technically, as a scientist, it's conceivable that there are genes that are completely unrelated to any gene you've ever seen, highly novel things, and are never expressed in the EST databases. So they are genes that look like nothing and they're not turned on except on Tuesday in the left toe or something, so we call them dark matter. I can't exclude the possibility of dark matter.

Bill Haseltine at Human Genome Sciences also has, as I understand it, 150,000 genes. A couple of companies have called the Sanger Center with respect to chromosome 22 and said, "We have twice as many genes as you do on chromosome 22." The ones I know about where they have done a co-analysis of it, have added only one or two genes. Most of those go away upon close analysis....

Give them a region of the genome, pick your favorite 1 percent of the genome and ask them to disclose all the genes in that 1 percent of the genome, and let's do a test.

IVOR ROYSTON (NCAB member): The stories in the New York Times yesterday talk about the ease by which you can navigate the Celera database. Can you comment about that?

LANDER: If you pay the money, it's much easier. This is exactly what industry should be doing. Industry should be producing value-added tools that allow easier navigation of databases. I am all for the value addition to the public sequence by companies like Celera. The world will have layered needs for access to data....

I feel very strongly that the underlying data and the ability to annotate it and add value, either in the academic or commercial sector, must be available in a free and unfettered way. As for the specifics about whether it's better or worse, you have to plunk down your money in order to find out.

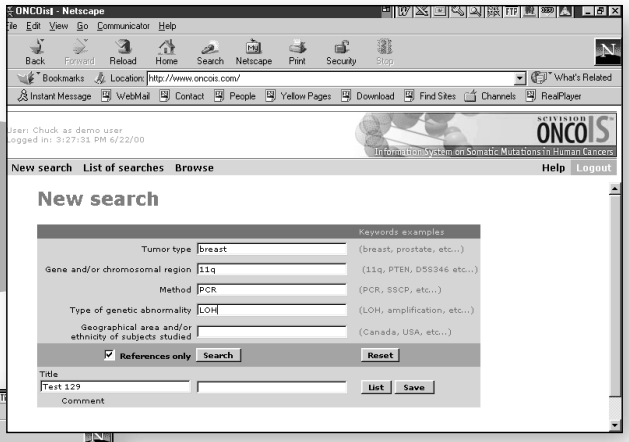
PHILLIP SHARP (NCAB chairman): Is there enough investment in annotating and making this sequence available to the users?

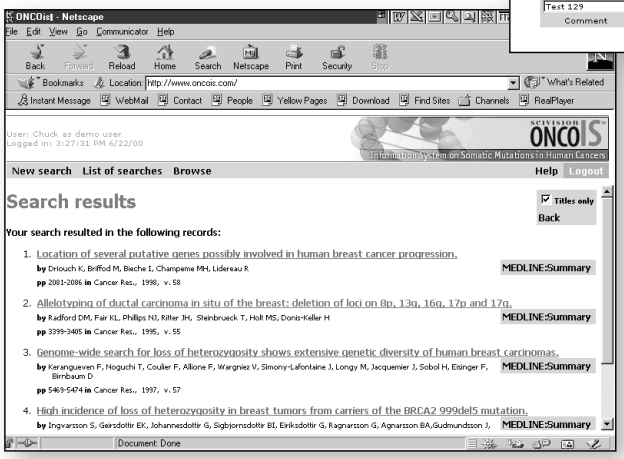
LANDER: I don't mean to dismiss the obligation on the part of the public sector, particularly NIH, through the NCBI to provide many great tools. I think they are doing a great job of providing very good tools. Today you can zoom in on any scale and all the features anyone has ever described in the public literature, right there on the site. My own sense is, that is pretty powerful stuff. Are we doing enough of a job? No. We've put the simplest features there. We don't have a robust effort to think of more features. NCBI has to be saluted for getting the job done with modest funding compared to what the needs are, and we've got to put a lot more money behind it....

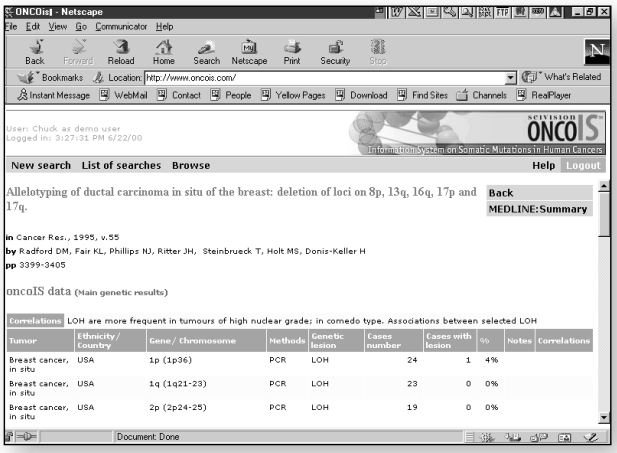
I believe a prudent strategy would be to have a tremendously robust investment in bioinformatics in the public sector, and to encourage, through contacts with the private sector, standardized nomenclatures and languages so many companies can build modules that can be interoperable with each other. We have a role to play to encourage the private sector to supply us with tools.

Mutations in Cancers…

FREE
24-hour trial at
www.oncois.com







SCIVISION ONCOIS™ is an Internet Information System providing current, concise, structured access to a vast amount of heterogeneous information about somatic mutations in human cancers. This growing knowledge-base contains qualitative and quantitative data, expertly extracted from published literature by molecular oncologists and updated quarterly.

ONCOIS™ is highly informative, as relevant results are extracted by experts from numerous, original publications and presented in the form of tables, including data concerning:

- **Tumor types studied**
- **Geographical area and/or ethnicity** of the patients
- **Genes and/or chromosomal regions** examined
- **Laboratory Methods**
- **Number of cases studied**
- **Types and frequencies of genetic lesions**
- **Genetic markers used,** including the frequencies of heterozygosity
- **Details on micromutations**
- **Clinical and other correlations**

SCIVISION
making information work for you
www.scivision.com

Visit **www.oncois.com** for detailed information or call, **1-800-861-6274** or **781-272-4949.**

SciVision is a Division of Academic Press
A Harcourt Science and Technology Company
Harcourt

Copying Policy for The Cancer Letter Interactive

The software that comes with your issue allows you to make a printout, intended for your own personal use. Because we cannot control what you do with the printout, we would like to remind you that routine cover-to-cover photocopying of The Cancer Letter Interactive is theft of intellectual property and is a crime under U.S. and international law.

Here are guidelines we advise our subscribers to follow regarding photocopying or distribution of the copyrighted material in The Cancer Letter Inc. publications in compliance with the U.S. Copyright Act:

What you can do:

--Route the printout of the newsletter to anyone in your office.

--Copy, on an occasional basis, a single story or article and send it to colleagues.

--Consider purchasing multiple subscriptions. Contact us for information on multiple subscription discounts.

What you can't do without prior permission:

--Make copies of an entire issue of the newsletter. The law forbids cover-to-cover photocopying.

--Routinely copy and distribute portions of the newsletter.

--Republish or repackage the contents of the newsletter.

We can provide reprints for nominal fees. If you have any questions or comments regarding photocopying, please contact Publisher Kirsten Boyd Goldberg, phone: 202-362-1809, email: kirsten@cancerletter.com

We welcome the opportunity to speak to you regarding your information needs.

Click Here for
Photocopying Guidelines