Scanning Life's Matrix: Genes, Proteins, and Small Molecules (2002)
Lecture One—Reading Genes and Genomes
Eric S. Lander, Ph.D.

1. Start of Lecture One (00:14)

From the Howard Hughes Medical Institute... the 2002 Holiday Lectures on Science. This year's lectures, "Scanning Life's Matrix: Genes, Proteins, and Small Molecules," will be given by Dr. Stuart Schreiber, Howard Hughes Medical Institute investigator at Harvard University; and Dr. Eric Lander, director of the Whitehead Institute/MIT Center for Genome Research. The first lecture is titled "Reading Genes and Genomes." And now, to introduce our program, the president of the Howard Hughes Medical Institute, Dr. Thomas Cech.

2. Introduction by HHMI President Dr. Thomas Cech (01:02)

Welcome to the Howard Hughes Medical Institute and to the 2002 Holiday Lectures on Science. This is the tenth year in a row that we've been able to provide high points of biomedical research to the public. This presentation is being Webcast live, and in addition, in our audience here, we have high school students from throughout the greater Washington, D.C., area. Now, you've all heard about the Human Genome Project and read about it in the newspapers. A genome, of course, is the entire set of genes in an organism--all of the DNA that's responsible for our inherited characteristics. In the case of the human genome, there are roughly 30,000 of these genes. Howard Hughes Medical Institute supports more than 300 researchers and their groups throughout the United States. Many of them were directly involved in figuring out the human genome, but others who were not directly involved in that process still have found their work really transformed by the availability of this vast amount of information. We're going to encourage you to visit our Web site to learn more about our investigators. Now, we're going to have 4 lectures for you-- two today and two tomorrow-- that will open the window on this fast-moving area of genome science. And if I had to, in a very quick way, summarize for you what you're going to learn in these lectures from our two speakers, it will be that Eric Lander observes, and Stuart Schreiber perturbs. You'll have to listen carefully to figure out that somewhat cryptic message. Our lead-off speaker, Eric Lander, is a master at not only determining and organizing the determination of genome sequences, but also figuring out or extracting from this information the most that you can about human biology and about evolution. And of course, one of the powerful directions that all of this is going is to increase the understanding of how cells work so that we can intervene in disease processes. So there are a lot of medical treatment implications coming down the road. Now, Eric is also a very engaging individual and colleague-- the sort of guy who will come into your office and take his shoes off in the middle of a scientific conversation. We're going to have 40 minutes of hearing from Eric, and before that, we'd like to show you a video to introduce our speaker.

3. Introductory interview with Dr. Eric Lander (03:46)

My interest in genetics was a little bit accidental. I really wanted to learn how the brain works, and so to understand that, I had to understand cell biology, and to understand that, I really had to understand molecular biology, and to understand that, I had to learn genetics. So I just began working backwards, and... I got stuck in genetics, and I'm still here, trying to understand it. There are so many interesting things in genetics that-- I think, of course, with my training as a pure mathematician, and genetics being so logical, it just appealed to me tremendously. I've just never looked back. In the big picture, the Genome Center here is about trying to extract all the information in the human genome. The real purpose of having the information in the human genome is to be able to understand the mechanisms that really cause disease. In the 20th century, most of medicine was directed at treating symptoms, not causes. We don't actually know the cause of most hypertension or diabetes or heart disease or asthma. It would be

really good if we understood at a molecular level what was wrong. We might be able to intervene much earlier to prevent damage. There's a certain sense now of inevitability. We are going to be able to pin down the causes of disease over the next decade or two, and while that's not going to immediately cure disease, it's going to give us such a great leg up in fashioning therapies and cures over the century. What's exciting to me about the world of genomics is it needs almost every area of science. It needs chemistry, it needs some physics, it needs engineering, it needs mathematics, and the distinctions kind of melt away. You've landed on some incredible new continent, and no one's ever explored any of this stuff, so you have to be ready with every possible tool. My goal for the Holiday Lectures is really easy. It's that there is so much opportunity now in trying to figure out the information in the human genome and what it really means for medicine, and it's going to take thousands of people working on it. The generation that grew up with molecular biology and computing as part of their world-- that's the generation that's really going to deliver on the promise of the human genome. And all I want to do is make sure that everybody knows that it's now open season on understanding what's in our genome.

4. Geneticists are interested in human variation (06:17)

Well, welcome. Glad to have you. We'll get comfortable. It's a pleasure to be here. Congratulations to the hardy souls who made it through D.C.'s biggest snowstorm in a long time. There's a lot to talk about today. What I'd like to do in today's lecture is talk about the Human Genome Project and about what we learn by studying genomes. See, I'm a geneticist, and as a geneticist, what I do is...I study variation. I'm really interested-- all geneticists are really interested--in variation. And this picture here-- a picture by a very famous photographer, Annie Leibovitz-- is my absolutely favorite picture to capture the great variation in our species. You may recognize that there on the right, it's Wilt Chamberlain, famous basketball player; and on the left, Willie Shoemaker, one of the most famous jockeys of the century. And the picture just illustrates so wonderfully the great differences in our species in height and weight and skin color-- all the different things that make it so wonderful to look around at the diverse human species. But to me, as a geneticist, it's also emblematic of the many differences that you don't see with your eyes-- the differences in who has a high risk of getting heart disease or of cancer or hypertension or diabetes, all of these things being underlain by the action of multiple genes working together with the environment. And what we want to do is nothing less than tease apart all those genetic factors, figure out how they work, and also how they work together with the environment. It's a tall order, and in fact it's been something that, in some sense, folks have been interested in for many centuries.

5. Roots of genetics in the age of exploration (07:59)

Indeed, the story really starts back in around 1500--the age of discovery, the age of exploration--when European explorers went out into new worlds and brought back with them tales of remarkable places, as well as samples of plants and animals that they took back home, and people there began experimenting with them and breeding them and making different varieties of plants. And about that time, economics and transportation also improved around Europe, so that if someone was really smart, they could come up with a new variety of a plant that would be more productive or more nutritious, and be able, because of transportation, to sell it to lots more people. So there began to become economic incentives to develop better and better agriculture. And because of that, a number of different cities around Europe began to become very interested in improving the scientific understanding of heredity. They set up civic organizations, one of which was set up in Moravia in the city of Brunn, or Brno, in the center of the Austro-Hungarian Empire, right smack in the middle of the textile industry, because they were very interested in this economics. And one of the people who ran one of those societies had as his day job being the abbot of an Augustinian monastery. And so what he began to do was look for young kids--boys-- who could become monks who had training in mathematics, physics, and other scientific things, and began to assign them relevant problems.

6. Gregor Mendel’s original study and its rediscovery in 1900 (09:27)

Well, as you may have guessed already, his best catch was this kid called Gregor Mendel, who he brought to the monastery and gave him the problem of understanding inheritance in peas, and Mendel went on to be quite a remarkable scientist. In fact, in 1865, he published this famous paper that I'm sure you've all learned about. These days, they probably teach you about this in third grade now or something. But Mendel is my hero, and I can't not mention Mendel, because his incredible clarity really organized the ideas of genetics for us, the idea of taking 2 pure-breeding strains-- in this case, green pea and yellow pea-- crossing them together, and observing and counting in the next generation and the next generation how many peas of each color you get. Noticing that in that second generation you get about 1/4 of yellow peas led Mendel to guess, to hypothesize, on the basis of nothing he could see, but just what he could think with his mind, that there must be discreet factors of inheritance. There must be a green form and a yellow form, which, when they come together, the yellow form's not visible in that first generation, but then it reappears when you get 2 copies of yellow. Now, you can imagine what people's reaction was to such an exciting theory. They completely ignored it. I mean, it sunk like a stone. Nobody really read the paper. Nobody--because, I mean, it's kind of abstract. What are these gene things, anyway? Mendel wasn't saying what they were. It was, after all, just something he thought up in his mind. And so, in fact, this really didn't have an impact in his own life, and, in fact, Mendel didn't go on to do much more in genetics. In fact, he became the abbot of the monastery and really mostly got caught up with administration, and that's pretty much the last we hear from Mendel scientifically. It wasn't for another 35 years until Mendel's ideas were picked up. In fact, in the opening weeks of the 20th century, Mendel's ideas were simultaneously rediscovered by 3 different groups around the world, and they published papers reporting this understanding of the basis of heredity. And, in fact, the first paper came out in January of 1900-- the opening weeks of the century. In some sense, the 20th century can be read as the eventual understanding of the idea of heredity, starting with the recognition that there were laws of heredity, but no idea what it was really about.

7. Review of genetic advances in 20th century (11:45)

The first quarter of the 20th century was largely devoted to figuring out that heredity resided in the cellular structures called the chromosomes. This was mostly because the way chromosomes were transmitted matched up so well with the way that Mendel's abstract particles were transmitted, and so folks said genes must somehow lie on chromosomes. But they had no idea what chromosomes were made out of. The next quarter of the 20th century was pretty much devoted to finding the molecular basis of this cellular structure, namely that the DNA molecule was the physical molecular basis of the information in these chromosomes, and that it had this famous double-helical structure. And by mid-century, James Watson and Francis Crick had worked out the double-helical nature of DNA, and understood how that explained the transmission of hereditary information, with corresponding sequences on the 2 strands allowing the DNA to be separated and used as a template for copying information to send to the daughters. Well, that was great. We knew now what the molecule was, but we had no idea how in fact information was written in this molecule. The next quarter of the 20th century was devoted to figuring out how it was that information was written. It was devoted to figuring out that DNA was read out into RNA; RNA translated into proteins by a genetic code, using a 3-letter lookup table; and also, quite remarkably, devoted to finding the laboratory techniques that would allow us to read out that information through recombinant DNA, through the cloning of particular molecules, and through the sequencing of their DNA basis. So that brought us 3/4 of the way through the century, with a molecular understanding and an informational understanding and the tools to begin to read. Well, the last quarter of the 20th century has been characterized by a voracious appetite to read as much of this information as we possibly can-- first individual genes, then collections of genes, then entire genomes of small organisms, medium-sized organisms, and then finally, in the closing weeks of the 20th century, the reading out of the nearly complete genetic information of the human being. It's not bad, as centuries go. If you realize that that's what got done in 100 years, just think ahead to what could possibly happen in the next 100 years, if this is

what a scientific community can do given a period of just 100 years.

8. The genome as nature's experimental notes (14:11)

Well, what it's done is it has given us all-- but particularly given you, as the students of the generation now coming into science-- the keys to the most remarkable library on our planet. I grew up in New York City, so to me, the picture I have in my mind of the most remarkable library is the new New York City Public Library here on Fifth Avenue in New York. You all may have your own image, but to me, the greatest library, I now realize, is the one in which evolution has been taking notes. Evolution's an experimentalist. It takes notes of its experiments, and it puts those notes in genomes. It does experiments, changing DNA sequences here and there, and seeing how it works. Sometimes it was a bad idea to change the DNA sequence, and the organism dies; sometimes it was good, and the organism is positively selected. And what we have now is the ability to read the results of 3.5 billion years of experimentation that evolution has been painstakingly involved in and very carefully taking notes on.

9. Reading DNA helps us understand disease (15:08)

Well, why read this information? Why read the DNA at all? What's so interesting about reading all this DNA sequence? Well, there are many reasons that people give, but to my mind, the single most important reason to do this is to be able to understand the basis of disease. For example, people have-- What do I mean by "understand the basis of disease"? I mean, for example, to be able to take a disease like sickle cell anemia and understand that what's wrong with it is a particular, single-based change in a DNA that leads to a single letter change in a protein here in the hemoglobin molecule; or to take a disease like phenylketonuria, in which individuals are unable to digest phenylalanine, and it builds up and poisons their brain, leading to mental retardation, and realize that there's a molecular basis to that, in a defect in one specific enzyme that breaks down phenylalanine. In fact, this is a very interesting picture here. This individual is Margaret Doll. Her uncle is the person who, motivated in fact by Margaret's own case of phenylketonuria, developed the incredibly powerful, simple, cheap blood test that every newborn in the United States is now subjected to in the first couple of days of life to pick up individuals with phenylketonuria so that in fact they can be put on a low-phenylalanine diet, and therefore don't become mentally retarded. So she, in fact, was the motivation for that. Well, that shows the power of molecular understanding, because now, for pennies a newborn, we can prevent those cases of PKU. So we want to understand molecular basis of disease, but the problem is, in both of these cases, what it took was being smart and lucky. People had to make inspired guesses as to what was wrong and what to do about it, and the problem with that is it's very hard to guarantee that you can be smart or lucky.

10. Finding disease genes by tracing lineage and chromosome walking (16:59)

What we'd really like to be is systematic. We'd like a guaranteed way to be able to find the basis of disease, no matter what it is. And so geneticists want a way to find disease genes, no matter what's going on. Well, here's how geneticists now do that. We look at a family in which a disease is being passed on. Here, Dad has passed on a disease to some, but not others of the children in the family. And we see here that Dad's got a defective copy of the gene. It's passed on to some of the offspring here, but not others, and we want to know, where does that lie on the chromosomes? I've drawn it as if it lies here, but, you know, we don't know that in advance. So what geneticists do is they just try a random spelling difference somewhere in the genome-- "C" and "A"-- and they see if that correlates with the pattern of the disease. No, it doesn't, so they try another spelling difference elsewhere in the genome, see if that correlates with the pattern of the disease. No, it doesn't. So they try another one somewhere in the genome and see if that correlates, and bingo--it does. The "T" goes along with the disease. So it must mean that whatever that disease gene is, it's nearby, but "nearby" still could be far away. So then begins a tedious process called "chromosome walking," where, having got some nearby markers that are close to the disease gene, we try to get the disease gene itself. We do that by taking the piece of DNA here, using it as a probe to collect a

larger piece of DNA from that region, using this in turn as a radioactive DNA probe to collect more pieces-- the next piece, the next piece, the next piece, the next piece... boy, this is boring... then use this as a probe to collect the next piece, the next piece... until you get to here. This is a very tedious, very long process, but eventually you get out the disease gene, and when you read it, you see something like this--

11. Genetic basis of cystic fibrosis (18:45)

lots of As and Ts and Cs and Gs-- and you can't possibly read that, right? But let me call your attention right over here to those 3 letters. These 3 letters, CTT, are the basis for cystic fibrosis. Most of the chromosomes in the population that predispose to cystic fibrosis have a deletion of exactly those 3 letters--CTT-- and that's the cause of the disease. In fact, knowing that, we could do a simple diagnostic test to find out who has that deletion and who is a carrier for that, such if they married another carrier, their children would be at risk. In fact, if everybody simply spit in a test tube, we could get enough DNA from the buccal cells that came off from inside your cheek that we could do the reaction just here and let you know by the end of the day. That's the power of it, but it's even more than that. You can take this boring sequence of letters, pop it into a computer, and ask the computer, "Ever seen any sequence of letters like that before?" And the computer will tell you, "Oh, yeah. "This looks an awful lot like a bunch of other genes "that have been characterized as cell transporters that transport things in and out of the cell." Congratulations--you've just discovered a transport gene. Well, that's fascinating. You didn't do any experiments. All you did was press "return," and you could find that out. That's way cool. Now, of course, you have to go verify that at the lab, but the power of connecting up all that knowledge is quite awesome.

12. The origin and goal of the Human Genome Project (20:07)

Well, the problem with this, I've got to say, is not conceptual. The scheme is great. The problem is that what I've just described to you took 5 years of work, it took 50 or 100 people, it took tens of millions of dollars, and it was just one of the many diseases we want to study. So what we wanted to do was short-circuit all that work so it could be done not by massive teams of researchers, but by one smart graduate student or postdoc. That was the origin of the Human Genome Project. The goal of the Human Genome Project was to put all this information at the disposal of individual scientists. It was first discussed in the 1980s. It was, in fact, a controversial proposal around 1985, 1986, and finally, after a number of discussions, by the late eighties it was decided that we should go for it-- we should try, in fact, to have a Human Genome Project that would be able to produce all of that information. By the end, over the course of about 13 years of work, 20 different laboratories around the world in 6 different countries-- the United States, the United Kingdom, in France, in Germany, in Japan, and in China-- participated in this project and led to the production of a huge amount of data.

13. Demonstration: The 80’s way of sequencing DNA (21:15)

Now, let me just tell you, back in 1985, when this was being discussed, the way to sequence DNA was to take a petri plate that had bacteria growing on it that had a piece of DNA in it... take the bacterial colony, inoculate some-- whoops, sorry about that-- inoculate some growth medium, grow it up overnight, when you were done growing up the bacterial colony-- the bacterial cells-- spin it down, pipette them out, put them into another tube here, add reagents to purify the DNA, add reagents to do a DNA sequencing reaction-- oops, all of this is radioactive, by the way; we should be doing this behind a shield-- and then, when we're done with all of those reactions, load one of these clunky gels one at a time with different samples, turn on the electricity, watch the DNA migrate, at some point turn off the electricity, take the radioactively labeled gel, put it up against a piece of X-ray film, stick it in the freezer, let it go overnight, take it out, develop the film, and then sit there with a Magic Marker and read off the lines that indicate the DNA sequence. This is why people were horrified by the idea of the Human Genome Project-- because they imagined legions of graduate students chained physically to the benches, doing this sort of work.

14. Video: Today's way of sequencing DNA (22:26)

But it didn't turn out like that, and the reason is because it all got much more automated, as shown on this video here. This is actually what goes on today. The toothpick-- this is now the toothpick, which we view as a very high-tech toothpick. The toothpick has[/I] a camera attached to it. It knows where to pick; it images. It's picking, picking, picking, and then it dunks it into the growth medium there. That's the computer analyzing everything it's been picking. Once the bacterial cells have grown up, they go onto this system, and instead of my standing there adding chemicals and cracking the cells open and making sequencing reactions and all that, they happen here automatically, and this particular system is processing about 110,000 of those tubes in the course of every day, setting up those reactions. And then finally, instead of these clunky gels here, they go back to a back room here with automatic sequencing machines. You put plastic trays with samples on it, and the machines run by themselves all day long, producing about 65 million letters of DNA every single day.

15. The human genome sequencing timeline (23:21)

That's what's allowed us to accelerate this process of understanding the human genome. And the timeline, once all this fell in, was quite remarkably fast. In 1999, the project got underway in terms of really large-scale sequencing, and the timeline over the course of the next couple years-- filling in first draft sequences, and then increasingly these red finished sequences with all the gaps closed and errors and things corrected. We're on target right now to produce a finished sequence of the human genome by next april 26. Why next april 26? Because that turns out to be the 50th anniversary of the day that Watson and Crick published their paper on the double helix, and we thought it would be kind of cool to be done then. And we're pretty much on target for that. There's still more work going on to do it, but looks like we'll be on target to have the finished sequence of the human genome so that--in fact, both Crick and Watson are alive and are planning to participate in those celebrations. Pretty cool. Uh, well... The sequence of the human genome, although not yet finished, is already in such a highly advanced draft form that investigators all over the world are making use of it. That sequence was published first in 19--in 2001, last February, in a paper that appeared in Nature, and got a whole lot of press at the time. You probably, if you were awake and watching anything, couldn't have missed all the celebrations about the sequence of the human genome. This was a high-quality draft sequence. It was the product of 13 years of work by a lot of people, myself included.

16. Sequencing the mouse genome (25:03)

What's very cool and what's very special about today is, the next step of the Human Genome Project, and a crucial step for interpreting this sequence, was to sequence the mouse genome and be able to line them up and compare them. Well, the mouse genome is almost as big as the human genome. They're both in the neighborhood of 3 billion bases. The mouse genome, it turns out, instead of taking 13 years to sequence, has taken 13 months or so to sequence. And in fact, an international consortium, including my own group, has written a paper on the sequence of the mouse genome that is appearing in the scientific journal Nature, and, as a matter of fact, I'm pretty excited today. Today is actually the official publication date of that paper. So this is, in fact... a pretty special day for me, scientifically, because we've been busy for[/I] the last 8 months working on this paper, and we have copies of this issue of Nature, which I think is a kind of historic issue of Nature, and we also have copies of the human paper, as well, for everybody who's here today. And for the people who aren't here today, all of this information is, in fact, freely available on Nature's Web site, and all of the data itself-- the URL is shown there-- and all of the data itself is freely available. You can download the entire sequence of the human, the entire sequence of the mouse, and play with it all you want-- no restrictions whatsoever. So there we are. What I'd like to do is turn to the question of what's in this human genome, what's in this mouse genome, and what did we learn? But before we do that, let me stop and take questions and see what's on your mind about that.

17. Q&A: Does “junk” DNA cause problems in interpreting the genome? (26:44)

Are there any questions? Yes. In our class, we learned that only about 5% of the genetic sequence actually codes for functional proteins. what sort of problems does this, like, make in your studies, knowing that a lot of this information that you get is just junk DNA? These are-- this is a great question. In fact, it's such a good question that what I think I'll do is devote the second half of my lecture to entirely your question. This is exactly what's on my mind, as well. You're right. It's a huge genome, and, you know, what's in it? How much of it is important stuff, junk stuff? And the truth is that many of the answers to those questions are things that I think we only know today, based on the comparison. And so I will indeed talk about it, but let's hold it a moment, because I have a whole bunch of slides for you on just that.

18. Q&A: Are most diseases caused by small changes in DNA? (27:31)

Yes. You said with cystic fibrosis that it was caused by a difference of about 3 letters. Exactly 3 letters-- CTT. With most diseases, is that the case? Is it a 1- or 2- or 3-letter difference, or are there some that there--it's more? So that's a really important question. Many simple Mendelian diseases in which there's one gene broken, or caused by one letter difference-- that's all it will take. You can, if you think about the genetic code, change one letter and turn something from the code for an amino acid into a stop codon, and that will prevent the protein from getting made properly. So one letter can do it, and for most single-gene disorders, a single letter suffices, although sometimes 3 letters, sometimes there will be a deletion of the entire gene. One of the things you learn as a geneticist is everything that can go wrong does go wrong, somehow and somewhere. All possible ways of screwing up a gene occur. What's also interesting, though, is, for many diseases, it's not a single gene. Like, when we talk about heart disease or hypertension or diabetes, it's not just one gene that's wrong. It's often a collaboration between variant forms of multiple genes, and that's much more complex, and we'll, in fact, turn to that tomorrow, 'cause that's the frontier right now, is clawing our way up from an understanding of single genes that are just broken and account totally for disease to polygenic disorders. That's a great question.

19. Q&A: Do genes involved in the same disease have the same promoter? (28:56)

Yes? You were talking about the... having different genes causing diseases in collaboration with one another. Are they usually genes that the synthesis of their mRNA is connected to one another, or are they separately created? Sorry. Different genes that are involved... Like, they have the same promoter. Are they usually genes that have the same-- Genes that are involved in the same disease? -Do they have the same promoter? -Right. Well, actually, it's very interesting. In bacteria, the way things often work is that related genes are made from a single promoter that makes a single, long mRNA that's translated into multiple proteins. That doesn't work, usually, in higher organisms, eukaryotic organisms. In us, even if we have 20 different genes that might together make proteins that function in a single molecular complex, they're usually made by 20 separate pieces of DNA, often living on totally different chromosomes with their own promoters. And so we have a much more difficult challenge of how do you coordinate those 20 genes working off their own promoters, whereas bacteria, I think, have solved it in a very simple, compact fashion, which is probably appropriate to the fact they have small genomes, have to replicate fast, and all that, but we're-- we, in fact--it doesn't all fit together like that. In fact, I just remembered, I have... We brought T-shirts from the Genome Center for all the people who are asking questions. One for you...and...one for you, and one for you. Whoops. There you go.

20. Q&A: Could you talk about gene therapy? (30:23)

Yes? Could you talk a little bit about gene therapy and maybe using viruses to maybe give a person with a genetic disease the proper, I guess, gene? Yeah. So, clearly, we'd like to be able to say, "If there's a broken gene, let's just put it into all the cells that need it." In some cases, that's a perfectly reasonable thing to do. When it's a problem of your blood system, and you might only need a modest number of

those cells churning out a particular protein, there are a number of researchers working on taking out blood cells from a patient, adding back genes through a virus, putting those back in the body so they can turn out, maybe, a clotting factor that a person might be missing. But when you're talking about a problem in a brain cell, or, in fact, the many brain cells, we have no way right now to deliver the correct gene in the correct place to all of those cells and make it function correctly, and so the best way to treat those diseases may be instead through some other route than a gene therapy. It may be through small molecules, or it may even be, like phenylketonuria, through diet, where someone, by avoiding phenylalanine in their diet, then doesn't get the particular form of mental retardation. So there are many different-- the key is understanding what is wrong at a molecular basis. Once you understand what's wrong, you can investigate different ways you might fix it, from gene therapy to small molecules to... early screening sometimes will suffice, all the way over to diet.

21. Q&A: What ethical problems do you run into? (31:49)

All right. Let me-- oh, last question here. What type of problems do you run into with ethics and people-- like, the example was parents choosing, by in vitro, the child that they want to be born, or whatever. What problems do you run into with ethics, and how do you deal with them? Well, you run into many complicated choices here, because who decides this? Is this something society should decide by passing a law that says, "Whenever we discover a child "who's going to have a serious genetic disorder, "it's up to society to make a decision "that the parents must have a child, can't have the child, whatever"? Is this up to the parents? I guess I personally come down pretty strongly that it's up to the parents. If I trust anybody, I trust individual parents to have the best interests of their children and their families at heart. But it can be a heart-wrenching decision to figure out what should be done in these cases, and I think it's some very personal ethics. I think the more we understand about it, the more we accept people's choices in that, provide the information and the support, the better we are in that. That's a really great question. I would love to talk more about the ethical issues. It's something I think about and write about a lot. Let me give you a T-shirt, as well. Thank you. -Thank you. -All right. Let's carry on and talk a little more about the human genome, what's in it, and then we'll get back to a bunch of questions. There will be lots of time for more questions at the end.

22. How big is the human genome? (33:06)

How big is this human genome? It's big, so it's a pretty big genome. I've been trying to work out exactly how big the human genome is, and, you know, people say it fits on one DVD. It's true, but that doesn't mean anything to me, because I have no idea how much information is on a DVD, right? But I can understand how much information there is in the New York Times, right? Can I borrow you for a second? Can you just stick this up here? That's good. There we go. And stick that there. And here. You get the tape, I get the papers, OK? Whoops. Here we go. Next one. Here we go. Why don't you take that? Two bits of tape for you. One... two. And... now we'll take the stock pages there. Lots of little writing there, like the genome. OK? Stick that there. Stick that there. All right. So is this more information or less information than the human genome? A lot less, right? How much information is this? A lot, lot less. This is about--if we count, in a typical page in the New York Times, about 70,000 characters, give or take. That is, like, puny compared to the 3 billion characters in the human genome. But how puny is it?

23. Video: New York's Fifth Avenue as the human genome (34:20)

Well, we wanted--You know, we're experimentalists. We wanted to find this out in an experimental fashion, so we undertook the following experiment to measure how long 3 billion letters really is. To do that, we used the following yardstick-- the island of Manhattan. If we start laying out New York Timeses[/I] 3 across like this, starting at the bottom of Fifth Avenue-- this is West Fourth Street in Washington Square Park, just at the top of Greenwich Village-- and we lay them out 6 across, how long will it go? So, rather than just sitting here doing calculations, we've undertaken this experimentally, if

you'll roll the video. All right, guys, listen up. our mission is, take the New York Times, some duct tape, and map the human genome. These are the only tools we got, aside from our own stamina. You guys ready to do this? Yeah. Here we go. Looking good. Oh, yeah. Spectacular. They haven't even gotten past the very beginning of chromosome one. Come on, guys. Keep going. OK. Now, you see, by the time they get up to 34th Street, the Empire State building, 700 million letters, but it's only the end of chromosome 3. Well, little bit further. One billion letters is Rockefeller Center. That's the big Christmas tree and the skating rink and all that-- Rockefeller Center. One billion letters. It's about the middle of chromosome 4. Let's check in on them and see how they're doing. OK, 10:30. Still looking good. Trim...pretty excited. A billion letters to go. Those are a billion letters so far. Two billion to go. Doing good, doing good. Keep going, guys. Yep, up to somewhere in chromosome 5 now. Keep going. Well, by the time we get up to FAO Schwarz, the toy store on 58th Street, we're in the middle of chromosome 6--not far from the histocompatibility locus, in fact. It's about 1.2 billion letters so far have been laid down across the length of Manhattan, but we still have a lot more to go. It's only chromosome 6 out of all of the chromosomes we've got to do-- 24 chromosomes, if we count both X and Y separately. When we get up to the Guggenheim Museum, 89th Street, it's 1.9 billion letters-- not quite 2/3 of the human genome, but actually only the end of chromosome 10, because all the big chromosomes come first. Let's see how the folks are doing here. Can we check in on them again? OK. 2:00. Oh, they're slowing down a bit. Two billion letters seems to take a bit of a toll on them. Hmm. Come on! Another billion letters to go. 14 more chromosomes. Come on, guys. You can do it. You can do it. It looks a little scraggly here. You can do it. Oh, boy. Here they go--not quite as chipper as they were before. That can do it to you. Well, Mount Sinai Hospital-- it's probably good. They may want to stop off at Mount Sinai Hospital here. It's about 2.2 billion letters into the project-- the end of chromosome 12. Marcus Garvey Park, up in North Manhattan, 124th Street-- 2.7 billion letters, end of chromosome 17. And finally, the tip of the end of the Y chromosome up on 142nd Street here, the end of the human genome-- 3 billion letters. Let's see what 3 billion letters did to our intrepid students. Oh, my God. This is what 3 billion letters will do to you. It's a pretty exhausting business, going through 3 billion letters' worth of information. Oh, my goodness. Well, they've done it. There, they've done it-- 3 billion letters, and we've measured the length of the human genome. The human genome pretty much stretches from one end of Manhattan to the other end of Manhattan. Next time you visit Manhattan, you don't need cross streets. You can just reference things to chromosomes. You now know, you know, roughly-- and New York is a pretty sophisticated city. Just ask any cabbie, "Take me to chromosome Fifth Avenue and chromosome 14," and you'll be able to do that, or at least if they watch the Howard Hughes videos, they'll be able to do that.

24. Finding genes in the genome (38:42)

So...all right. We've now got the genome laid out across Manhattan. Now we have to find the genes. So we have to trudge all the way back from 142nd Street back down to West Fourth, and then we have to get down on our hands and knees, and we have to start reading the text very carefully with a yellow highlighter and attempting to identify the genes. This is not easy to do. Genes are very hard to pick out, 'cause, as you've already told us, the genes constitute only a small minority of the entire sequence. In fact... we need very sophisticated computer programs to be able to pick out the genes from this sequence. When you get down to it, the human genome is mostly composed of a nongenic stuff. The human genome is mostly transposable elements, often somewhat derogatorily referred to as "junk DNA," although I have a lot of respect for this junk DNA, because, in fact, it constitutes transposons-- elements that can make copies of themselves and hop around your genome, duplicating themselves and inserting themselves, duplicating themselves and inserting yourselves, and half of your total DNA is made up of these transposable elements-- 4 types, ranging in copy number from millions of copies of some hundreds of thousands of copies of others. And from the transposons' point of view, you are largely a vehicle for propagating transposons. All the rest of the stuff in your DNA is probably incidental to them. From your point of view, you like to think that it's your genome, but on a majority basis, it's not. The genes themselves--the things that code for proteins, that encode all of the collagens and the keratins and the hemoglobins and the everything else in your body-- constitute only 1.5% of the whole coding sequence.

98.5% is noncoding for proteins. Half of it is recognizably transposons, which means we have a big problem of signal to noise-- how to pick out the genes from amongst all the noise. So people write computer programs. The computer programs run along and try to look for the signal in it that says, "I'm a gene," but we don't know that signal very well. But we have hints and things, and the programs are getting better and better and better, but it's still a tough thing. Well,

25. Distribution of genes in the human genome (40:59)

as we looked at the sequence in the human genome to try to interpret what's in it, we found some pretty remarkable things. First, it's a very lumpy, uneven distribution of genes. We found that there are many neighborhoods in the human genome that are chock-a-block full of genes, and then next-door neighborhoods that are really pretty devoid of genes, or very gene-poor, at least-- gene-dense, gene-poor-- and we don't really know why this is. It's fascinating, actually. The very gene-dense regions tend to correspond to the light bands when you look at chromosomes in the microscope, and the gene-poor regions tend to correspond to the dark bands. The reason for that? I have no idea, and neither does anybody else, but it's really kind of cool that the genome is so lumpy and such an uneven thing. And...

26. The human genome has surprisingly few genes (41:43)

well, probably the single most surprising thing was how many genes there were, or actually how few genes there were. I teach freshman biology at MIT, and I taught an entire decade's worth of students the official textbook line that there's 100,000 genes in the human genome. Well, there's not. We found out when we looked very closely at the human genome that in fact there are only about 30,000 protein-coding genes. We blew it by 70,000. This is very embarrassing, and I'm still writing notes to all my former students, trying to correct this misinformation. And it was really--you know, it was really quite a surprise. You know, there's been lots of discussions, lots of debates, it's been revised and reanalyzed, but the truth is it seems to really be settling down at about 30,000 protein-coding genes, not significantly more than that. Now, this is really disturbing; not just we had it wrong before, but because, in comparison to other organisms-- well, let me point out, uh... Your typical president of the Howard Hughes, for example-- a typical Nobel Laureate-- has 30,000 genes. But then again, the mustard weed, Arabidopsis thaliana, has 27,000 genes-- not significantly less than the average Nobel Laureate. What, then, is so special about the Nobel Laureate, as opposed to the weed? Well, we're not entirely sure, it has to be said.

27. Are vertebrate genes different from other species' genes? (43:09)

There's been a lot of effort to try to figure out, if we don't have a lot more genes, maybe our genes are better in some way. This is a very anthropocentric view, I confess. So we've looked a lot at the genes to try to understand what's so special about them, anyway, and the answer is, not that much, actually. We were hoping that when we looked, we'd have found that vertebrates had invented lots of new protein domains-- building blocks of proteins. But I'm afraid that that didn't pan out. 96% of all the recognizable protein domains-- the building blocks of proteins, the basic architectures-- were already present in flies and worms and, you know, vermin and things like that, and so we can't really claim to have been very innovative there. We do, it must be said, take those pre-existing domains and put them together in more mix-and-match combinations called architectures. We probably have about twice as many domain architectures. It's not that creative, but all right.

28. Many vertebrate genes arose from gene duplication followed by variation (44:02)

Actually, when you get down to it, what mostly characterizes a vertebrate genome is just the tremendous amount of genome gene duplication, followed by slight divergence of the genes. We actually have vast gene families, where a particular gene has given rise to 2 genes, 4 genes, 20 genes, and they diverge slightly to take up different functions. It's not very creative. You wouldn't get a lot of points for

innovation for this. But it seems to have worked. It's given rise to zillions of immunoglobulin genes involved in immune systems or intermediate filament proteins involved in epithelial surfaces. You own 111 keratin genes, as best we can count. Growth factors-- you own, we think, about 42 transforming growth-factor beta genes. Whereas flies and worms get by with just two related genes of that sort, you have 42 such genes, and the most extreme case of all--smell receptors. You may not realize, but out of the 30,000 or so genes in your genome... 1,000 of them are smell receptors. You have 1,000 smell receptors in your genome, indicating that somewhere in your evolutionary past, some ancestor was very interested in smell as an important modality. In fact, all mammals have about 1,000 smell receptors, because that was the principal sense. Now, disappointingly, I've got to tell you that your more recent hominid ancestors have lost interest in this sense. Most of your smell receptor genes--2/3 of them-- are actually broken and nonfunctional. It's not true in most species, but hominids have lost lots of their smell receptor genes, probably 'cause we got much more interested in sight. But nocturnal animals and other animals have kept their smell receptor genes in better working order. So in any case, this is a lot of what we see in the human genome-- a tremendous expansion of families tweaking with things, and in some sense, that's what we've got to settle for for our innovation.

29. Human and mouse comparisons: The mouse as a model for humans (45:53)

Now, what can we learn by comparing the genomes of human and mouse? Can we learn more? Let me turn to that for the last topic today. It's a very timely day to do that, because with the availability today of a human sequence, and so excitingly for me-- I mean, this is really cool. You guys put a lot of work in projects and things like that, and you're happy when they're done. I'm happy when we finally get to turn in our homework, too. We've got a mouse. We've got a human. What can we learn from it? Well, why is mouse so helpful? What's so useful about the mouse, anyway? Well, the mouse often can mimic for us human diseases in remarkable ways. This is a really cool picture. It shows a baby who has a disorder where the pigment cells that migrate during development from the back to around the front to the midline don't close, and there's big patches here on the belly and on the forehead of unpigmented cells, due to a mutation in a particular gene called KIT. This mouse, which has the same white patch on its belly and exactly the same white patch on its forehead, has a mutation in the same gene--KIT. So, in fact, here the mouse is an eerie model for the human, and this is true for many, many cases. So at the level of phenotypes and traits, we can observe the similarity between mouse and man.

30. Human and mouse comparisons: Genomic similarities (47:11)

What about at the level of the genome itself—the information? Well, if we line up the DNA of the human and the DNA of the mouse here-- in fact, reference to Manhattan-- and we start looking for particular bits of the chromosomes-- let's say, we'll look at particular bits here-- we find that all the genes in this bit of the chromosome of the human match up to genes in this bit of the mouse. The genes here match up to genes here. The genes, let's say, over here, match up to here, and the genes here match up to there. In fact, it's basically the same book, but the chapters have been shuffled in a different order. In fact, this happens because our common ancestor[/I] had a genome in a particular order that has been shuffled over the course of time. We have, in fact, a complete lookup table that tells us what parts of the genome, in fact, match up with what parts of, say, the mouse genome; here, mouse chromosome number 2 matching up here with bits of human chromosome number 20, for example. We have a complete lookup table, and there's about 350 chapters relating it.

31. Human and mouse comparisons: Finding conserved sequences (48:16)

Now, what happens if we dive deep into those chapters? Well, if we line up the sentences within those chapters, we sometimes can see even more. We can find the hidden messages that evolution has bothered to preserve over 75 million years. Can you see this hidden message there? "This is hidden." Well, evolution doesn't write that there, but what evolution does do is this... What it does is it shows us, when

we line up the sequences of-- oops, there we go-- the sequences of human and mouse and rat and baboon, we find all sorts of hidden messages there. Some of them we recognize as protein-coding sequences, genes, but some of them are clearly not. They are clearly noncoding genes. This is one particular region of the genome where we see about as much genic and nongenic stuff, and, in fact, this is completely typical for the genome. As we run across the human and the mouse genomes, we find about half a million positions in which we find a high degree of evolutionary conservation, but remarkably--this is one of the cool things in today's paper-- only about half of them are protein-coding. The other half are a stuff, and we don't know what it is-- maybe regulatory sequences or things that code for RNAs or structural features or we don't know, but as of this morning, there's a quarter of a million such things in the human genome that we know we had been missing, that we know we didn't see when the human genome got published, and we only see by virtue of comparing it to mouse. And it becomes the challenge to your generation to figure out how these control instructions are working, what they're doing. But as of today, at least, we know they're there.

32. Challenges of today and challenges of the future (50:01)

Now, we'll know even more soon as we have sequences of chimp next year, dog, probably, by the end of next year, cow, cat, chickens, and other things, and line this up and use all of the wonderful work that evolution has been doing over the last 100 million years as experimental lab notebooks for ourselves. So people talk about the dog as man's best friend. Actually, at least today, let me argue to you mouse is our best friend. It's our best friend because we share a genome with it that we can now, today, study, line up, and learn things from. It's our best friend 'cause it's such a good biomedical model, and it's not really so different from us. The genome's a little bit smaller, but it ain't for any lack of genes, actually. It's actually for lack of other junk in there. And so, to my mind, this is the fun of the coming century, is all this information is laid bare, and you guys, as the generation that grows up not making the distinction between wet biology and informational biology and just take it all for granted, I think you're going to have a field day. Let me stop and take questions.

33. Q&A: What is currently being done to figure out noncoding DNA? (51:03)

What questions have we got? I'm out of T-shirts at the moment, but we'll get you some afterwards. Yes? You were talking about the regulatory DNA and the junk DNA, and how we're going to be the generation that's going to figure out how that works. You bet. I'm counting on it. Yeah. Are there any current methods of how to figure that out, and how do those work? Well, there are. There are probably, oh... out of that quarter of a million, I bet there are 80 that are well-studied. In fact, in this paper, we were able to collect 80 examples that have been really well-characterized, and study them here. That's a relatively small fraction out of a quarter of a million, and most of those 80s represent the work of 5 or 6 whole laboratories with large numbers of people studying. I can point to 2 or 3 of them that have been the work of probably 30 people to understand how that works. They take that piece of DNA. They see in a test tube what proteins might bind to it. They try to figure out how it changes the structure of the chromatin-- the proteins decorating the chromosomes-- how it helps promote the transcription of the gene, but it's hand-to-hand combat. You know, it reminds me a lot of human genetics before the Genome Project. I think what we need over the next couple of years is new blood coming in that says, "We're not going to put up with doing this one at a time. "We're going to come up with methods "that will allow us to take those pieces of DNA, "grab them, grab the proteins, and fly them on mass specs, identify what's sitting there, et cetera." And the answer is, we don't know how to do it, but it's pretty cool. I hope you'll get interested.

34. Q&A: How do you prevent the symptoms of the disease PKU (52:36)

Other questions? Yes? When they're taking out PKU, the disease, how exactly do they take that out? To "take it out," meaning to prevent the symptoms of the disease? Yeah. Turns out to be extremely straightforward. What you do is, at birth or a couple days after birth, you take a little stab of blood from

the heel of a baby-- a heel stick. You paint it on a little card, and the card gets sent off for analysis to a laboratory. It takes a couple pennies to analyze, and a couple days, and for most kids, the result comes back "no PKU," but for about one in a million, it comes back "PKU." If so, the kids are put on a special, low-phenylalanine diet for life. And you even are aware of this, 'cause the next time you pick up a Diet Coke can, look on the side. It says on a Diet Coke can, "Warning to phenylketonurics: contains phenylalanine." It's discussing genetics on a Diet Coke can. It's telling people that Nutrasweet contains phenylalanine, and therefore, people with PKU shouldn't drink it.

35. Q&A: How would you identify nondisease traits? (53:37)

Yes? Other questions? Yes? Up until now, you've been discussing genetics so that we can look at what causes disease and why we get disease. Well, what about other traits? How would you identify which genes cause other traits? Well, what sort of traits interest you? Well, IQ has gotten a lot of press lately, but... So, yeah. Well, IQ, to me, isn't a "trait." It's very--I mean, I don't know how to assay that one. But...other personality traits-- Like, how would we know what causes things that don't, you know, cause us to have a certain hair color or cause us to have a certain illness? So things, you mean, beyond hair color or illness? Yes. Are there roles, for example, for genes in...manic depression, for example? I think, almost certainly, the answer is yes. And, in fact, there are a number of groups around the world who are working on techniques now that take things that aren't simple genes-- single gene disorders, and can't just be traced so simply in a family, but nonetheless correlate genetic differences with those traits, like manic depression or schizophrenia. And, in fact, if you'd like, I will give a lecture on that sort of stuff tomorrow.

36. Q&A: Is it possible to treat genetic diseases with proteins? (54:50)

That's good. Yes? I was wondering, in, like, diseases, is it possible to treat the symptoms with proteins, rather than with, like-- like, changing your diet, couldn't you instead receive an enzyme that digests whatever is causing the illness? And is it possible to put these... actual DNA--the codes for these enzymes--into vectors that you can then take and be made into part of you? Sometimes, yes. These are way-cool ideas. There are, in fact, genetic diseases where people lack an enzyme, and the treatment is to give them the enzyme. In fact, there are people who have trouble digesting lactose, and they take an enzyme lactase, often, that helps, and that's an example of that. But there are much more serious disorders where people do the same thing. The idea of putting the DNA in to have it make the missing enzyme is certainly something people are experimenting, but that's gene therapy and requires a great deal of careful clinical research to get right. That's the--see, I've told you about the generic end. We now have a completely generic way, I think, to find molecular basis of disease. We don't have a generic way to find the molecular therapy and cure for a disease. Right now, it still requires inspired, bright young scientists figuring out what's going to be the best therapy for this and for this and for this and for this, but I wouldn't be surprised if some number of years from now--a decade or two from now-- there's a pretty good menu, and that menu is enough to be able to figure out what to do with most disorders. This has been a lot of fun today. I'm looking forward to tomorrow. Thank you so much, and more to come.

37. Closing remarks by HHMI President Dr. Thomas Cech (56:34)

Thank you, Eric, for a scintillating lecture. I'm glad you remembered to take your shoes when you were done. We're now going to take a break for about a half an hour, and after that, we'll hear from Stuart Schreiber, a founder of the field of chemical genetics. Stuart's going to take us beyond simply observing the genome to perturbing it, or actually perturbing the protein products of the genome. This might also be a good opportunity to visit the Institute's Web site.