Scanning Life's Matrix: Genes, Proteins, and Small Molecules (2002)
Lecture Three—Human Genetics: A New Guide for Medicine
Eric S. Lander, Ph.D.

1. Start of Lecture Three (00:15) *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 third lecture is titled "Human Genomics: "A New Guide For Medicine." And now to introduce our program, the vice president for Grants and Special Programs of the Howard Hughes Medical Institute: Dr. Peter Bruns.*

2. Introduction by HHMI Vice President Dr. Peter Bruns (01:06)

Welcome to day 2 of the 2002 Holiday Lectures On Science. Welcome back. Our speakers Eric Lander and Stuart Schreiber are here to continue the story of understanding the human genome and using it in the diagnosis and treatment of human disease. You know, yesterday there was a lot of talk about the various other kinds of disciplines that play a role in modern biology and that the path to becoming a contemporary biology researcher is not always a direct one. You should know that one of our speakers--Dr. Lander-- has a degree not in biology but in mathematics. And in fact, his first teaching job was teaching economics at the Harvard Business School. Look at him now. Tom Cech yesterday spoke a bit about the activities of the Institute in basic science. We're also very active in science education and international research. The Holiday Lectures is one of the things we do in education. Another is a web site that we've created to bring modern biology to your desktop. It's called biointeractive. You can look it up. You might write this down-- www.biointeractive.org. On that, we've put 3-d animations, interactive activities, experiments, and a thing called virtual labs, and this holiday lecture and all the others we've done. And it's also a site you can go to to order free DVDs of the Holiday Lectures. Just as biointeractive uses technology for science education, Dr. Lander certainly uses technology to organize genetics as a very large data set in science. He is going to continue now with the third lecture. His lecture's on human genomics: A New Guide For Medicine. And first we'll have a short video to say more about Dr. Lander.

3. Introductory interview with Dr. Eric Lander (03:08)

What I like best about my job is, I get to hang out with incredibly smart young scientists. There's nothing more exciting I can think of doing on this planet than being part of a revolution in understanding what life's about, what medicine's about. Science is a really wonderful activity in that sense because it's a community activity. Each person has a deep individual stake in it, but it only makes progress by everybody testing each other's ideas. Science is the most playful activity I know. The folks who really succeed in science are fundamentally playful. They love talking about ideas, and it's not like they're so smart. There's no way to go in and be so smart. You can't outsmart nature. What you can do is play with a problem long enough that you stumble across some really interesting fact, and then you have the self-confidence to follow up on it. Right now, we're changing the objective. The 2 things I'd say to a young person interested in science today is: 1, get a really good, firm foundation in basics. It really pays to learn some physics and some chemistry and some mathematics well and deeply, because these are just like your basic skills for going out into the wilderness, and you never know what tool you're gonna need in your pocket. But at the same time, pick something and just get passionately into it. And so if you combine that-- the really solid learning of basics and the experience of what it's like to chase something, those 2 things will be the pieces out of which you can weave a life. One of the major, major accomplishments of the human genome project is telling us how little we really know. Before we had a list of all the 30,000 components, we didn't realize how totally ignorant we were. Now that we have them in front of us, in some sense, that's the starting gun for the 21st century of science. Now we can start asking questions. And

so I think, you know, this is the best time to go into science right now, because you finally can make major, major progress. And you'll look back on the 20th century as prehistoric times.

4. Observing what nature has already perturbed (05:38)

Well, welcome back for the second day of the Holiday Lectures. I'm glad to have even more of you here today. And I want to continue our theme that we introduced in the first day, about how there's a great duality between observing and perturbing. The question of observing... well, that's what I'm gonna mostly focus on. And what Stuart, in the second lecture, is gonna mostly focus on today is perturbing. But they go hand in hand. I'm interested in situations where nature has already perturbed something-- perturbed our genes or ourselves in some way to give rise to a disease-- and I want to observe what's going on, to be able to figure out what the mechanism is. And then once one has observed the results of nature's perturbation, one is gonna want to then go and perturb it yourself to see if you really understand it, and Stuart's gonna talk about that.

5. How similar are the two copies of your DNA? (06:40)

Well, there are 2 kinds of variation that I want to observe: 1, the variation in DNA; the other, variations in RNA. I'm gonna start by talking about DNA variations. So everybody has, as we talked about yesterday, 3 billion bases of inherited information that comes from their mom and 3 billion bases of inherited information that comes from their dad. How similar are those 2 copies-- the one you got from mom and the one you got from dad? Well, here's some DNA sequence here. You don't need to write that down. If I were to sequence this DNA in you and check the sequence of each and every one of those bases, the positions at which you were heterozygous-- you had a different spelling on the chromosome you got from mom compared to the chromosome you got from dad-- would be right here. In case you can't spot that, let me circle it for you. There you go. That's it. The copy you got from mom and the copy you got from dad differ by only about 1 letter in 1,000. Those positions, 1 single letter out of 1,000-- it's actually a little bit less than that. Our best guess is about 1 in 1,200 or 1 in 1,300. One in 1,000 is round enough for this. Those things are called single nucleotide polymorphisms, or just single letter differences of the nucleotides-- or just the letters of DNA-- and "polymorphism" means variations. But we call them single nucleotide polymorphisms. And because that's a mouthful, we often refer to them as just SNPs or "snips."

6. How similar is DNA from two people? (08:22)

So we might find a SNP once every 1,000 bases in your DNA. Well, OK. What about if I take 2 different people in this audience and I compare their DNA sequences? How similar will they be? It turns out that if I take 2 random students in Maryland or anywhere in the D.C. area, the answer is also 1 in 1,000. In fact, actually, it has nothing to do with the D.C. area. If I take 2 random people anywhere on this planet, no matter where they come from, they will differ by 1 letter in 1,000 in their DNA, and no more than that. That is how similar we all are-- about 99.9% identical. Whether we're talking about presidents of the United States or someone living in a village in South America or someone living in Nepal, the genomes are that similar. Now, that's actually more similar than you might realize-- surprisingly similar, because if I were to take 2 chimpanzees in Africa and compare them, they would differ by about 2-3 times as much as any 2 humans anywhere on this planet. And if I were to take 2 orangutans in Southeast Asia and sequence their DNA, they would differ by 8-10 times as much as any 2 humans on this planet. You see, you think the orangutans all look the same. Well, they think we all look the same, and they're right. See, this is-- We are a very closely related species, as species go. Most species out there have much more variation. Fruit flies have hundredsfold more variation than we do.

7. Human origins and why we have little genetic variation (10:04)

Well, how is it possible that we have so little genetic variation as these things go? Well, the answer is, it's a fact that goes back to our own history. The human population, it turns out, all traces back a pretty short time ago-- a mere 3,000 or so generations-- to a founding population in Africa. All of the current human population spread out from Africa perhaps 60,000, 70,000 years ago-- about 3,000 generations or so. And that small founding population in Africa had perhaps 10,000 individuals or so. And population geneticists know that the amount of variation that there is in a population depends mostly on the size of the population and the mutation rate in the population. Well, we know the mutation rate. It's about 2 times 10 to the minus eighth. It's a very small number-- 2 times 10 to the minus eighth per generation. And the population was pretty small, and you can work out by theory that a population of that size should have about 1 polymorphism in every 1,000 bases. But then we left Africa, spread around the entire world. We're now distributed everywhere, and we now are 6 billion people. Why don't we have lots more variation? Well, the point is, 3,000 generations is too short. It's the blink of an eye for evolution. If the mutations are accumulating at a rate of only 2 times 10 to the minus eighth per generation, after 3,000 generations, we still all largely have the variation that we had in Africa. Some new variations accumulated, but most of the variation that we'll find anywhere in the world is the variation that we as a species had when we were all together in Africa. Sample a village in South America, sample a village in northern Japan, sample a village in Mongolia. Eighty-five percent of the variation that you will find in any of those places is variation that you'll find in all of those places. Only a small fraction of our variation is local. And so, although we may think of ourselves as a very big species, we're really just a very small species grown large in the blink of an eye.

8. Tracing human migrations by looking at genetic variation (12:22)

So we've spread out, we've carried with us largely the variation that we had in Africa, built up a little bit of additional variation. Now, in those 3,000 generations, some new variants have arisen, and they can be very interesting to look at. So for example, suppose in the migration that led to the peopling of the Americas that occurred, we think, about 13,000 years ago, some tribe that was migrating this way over the land bridge that connected northern Russia to Alaska over the Bering Strait here-- suppose that in that tribe, some random mutation happened on some particular Tuesday that changed a "C" to an "A"-- some particular letter in the human genome. Then some people in that tribe would have the genetic difference-- not all, not most, but some. Well, that would carry with them as they migrated, and you would find that that single letter difference, that relatively rare single nucleotide change, would only be found along the track of this migration. And you'd be able to tell that you-- it would be like leaving breadcrumbs across this migration area. In fact, geneticists are doing that today. They're looking for relatively rare spelling differences that let us trace the history of the movements of people around the world, and it's actually really quite amazing. You can trace all sorts of things about where peoples have gone.

9. What differences do genetic variations make? (13:48)

But even more interesting than that, to my mind, is to trace the impact of the variation that's common to every group in the world; the spelling differences that we find in almost all populations. So what difference do those spelling differences make? I said you have 1 in 1,000 letters that vary. Most of them aren't in the regions that code for proteins. We talked yesterday about the fact that only about 1 1/2% of the human genome codes for proteins. That's a very small portion of it. And we talked about the fact that maybe 5% of the total genome is evolutionarily well-conserved sequence. So most of the single nucleotide polymorphisms-- most of these SNPs are just scattered around DNA that probably doesn't have a strong function. But enough of them land in functionally important DNA-- and some of them in protein-coding DNA-- that they could make a difference, and sometimes, we know they do.

10. Single nucleotide polymorphism (SNP) can affect Alzheimer disease (14:44)

Here I show a particular DNA polymorphism SNP. It's actually 2 SNPs: a position here where people in the population might have a "T" or a "C" and where people here might have a "T" or a "C." They happen to both lie in the sequence of this gene called apolipoprotein E that lives on chromosome number 19. So the APOE gene, because it has these 2 sites of spelling difference, happens to have 3 alternative spellings: t-t, t-c, or c-c. As it happens, c-t is not found with any appreciable frequency. It turns out that if you happen to be homozygous for the E4 spelling-- that is, you got the e4 spelling from mom and the e4 spelling from dad-- you have about a 60% or 70% lifetime risk of Alzheimer's disease. About 3% of people are homozygous for E4. And we could, in fact, if you wanted to-- it's a big if you wanted to-- test you and let you know what your spelling was for APOE4 and let you know if you're homozygous for that. That wouldn't tell you with certainty whether you were gonna get Alzheimer's disease, but it sure would change the odds from the general rate in the population of 5% or 6% to perhaps 60% or so. You might want to ask yourself, do you want to know? Well, I could do this for myself in the lab, but I haven't done it 'cause I don't want to know. Why don't I want to know? Because today there's nothing you can do about it. However, several pharmaceutical companies are taking this information and, now knowing that this apolipoprotein-e gene must be involved in the biological process that produces Alzheimer's disease, have been tracking down how it's doing that and trying to make small-molecule drugs that will inhibit and slow down that process so that, for example, we might be able to delay alzheimer's disease from your 60s or 70s or 80s into maybe your 120s or 130s. That would be OK if you just put it off till then. By a factor of 2, slowing it down would be just fine. As soon as they've got those molecules, I'll sign up for the test to find out if I should be taking those particular drugs to try to delay it. Well,

11. Other examples of variations affect diseases (17:00)

we in fact know more examples than just that one. We know examples of a particular genetic variant in your genome that makes you at higher risk for deep-vein blood clots-- deep venous thrombosis. About 5% of us have a particular spelling difference in our factor v gene. The particular spelling difference gets the name the factor V Leiden mutation. And those individuals account for about half of all admissions to emergency rooms for deep venous thrombosis. And in particular, individuals who have that genotype are at higher risk of blood clots if they should be taking birth control drugs. So that's an important thing to know. There's PPAR-gamma, another gene on that list. Here's a very interesting example. There's a variant in PPAR-gamma that makes you about 30% more likely to have diabetes-- just 30% more likely. It's not huge, but it's a noticeable risk. But it turns out that 85% of you have that variant. Eighty-five percent of you have the risk variant. Fifteen percent of you have the protective variant. And so although it's a relatively small increase, a mere 30% increase, since most people have the risk form, it actually accounts for about 25% of the total risk of diabetes. That is, if magically everyone had the other genotype, the risk of diabetes would be about 25% lower. We can't magically make everyone have the other genotype, but you can look at the particular protein that that makes and ask, "How does it do that?" and "Can we inhibit it as it happens?" One of the well-used drugs against adult diabetes of the troglitazone class is actually directed against the protein encoded by that particular gene. Well, there's a case where most of us have the variant that predisposes to disease.

12. Some mutations can protect against AIDS (19:01)

Now, here's another case, but even more remarkable: CCR5. Let's take a look at CCR5. That's a particular gene where 10% of the copies in the human genome have a deletion that make it nonfunctional. Do you want to sign up to have that deletion? Sounds bad. It's a defective gene. One percent of you are homozygous for that defective gene. Sounds even worse. It turns out, though, that those of you who are homozygous for that defective gene fail, as a result of that, to make a particular protein on the surface of certain of your immune cells, and that protein is necessary for the HIV virus-- the AIDS-causing virus-- to get into your cells. And so those of you who are homozygous for that genetic "defect" are in fact immune to infection by the AIDS virus. This is a case where that supposed genetic defect-- indeed, if we're just reading the human genome, we'd say, "Ooh, big problem. "We got a big mutation in that gene. "Gotta do

something about that"-- is in fact a great advantage. Now, of course, this only protects about 1% of the population against AIDS, but the same story. Pharmaceutical companies now are now realizing that if it protects them by preventing the AIDS virus from interacting with that cell surface receptor, why not make a medication, why not make a small molecule that blocks that interaction so that everybody can block the entry of the AIDS virus? That hasn't been done yet, but people are in fact trying to develop such drugs. Indeed, I know that some are intended to come to clinical trials.

13. Cataloging all human variations (20:47)

Well, we'd like to find this out not just for the 5 or 6 cases I gave you, but we'd like to find this out by screening all possible variants. In the human population, the common variants that we had in Africa and spread around the world with-- there are about 8 or 10 million common variants in the human population. Why not just write all of them down? In a way, it would be the complementary project to the human genome project. The human genome project was to get 1 reference sequence for all the human DNA. Why not make another project to collect all of the possible common variations across the human population in that DNA and then correlate it with disease? Well, that seems kind of crazy. How are we gonna get all the variation in the human population? Well, the point is, because we are such a small species, and it is such a limited variation; only 8 or 10 million variants-- 8 or 10 million isn't that big-- we could do it. And in fact, that program is pretty far underway. When you were in fifth grade, the number of common variants that had been identified in the entire human genome was about 100 out of the 8 or 10 million. When you were in seventh grade-- I'm assuming you're seniors now; I know some of you aren't-- about 4,000 common variants had been identified in the first large-scale project that had been conducted as a parallel project to the human genome. By 2000, the year 2000, with the time of the publication of the first draft sequence of the human genome-- actually a little before that-- about a million variants had been identified: a huge increase in the course of just 3 years. With an explosive work in sequencing and in variant identification, by today, the count is 3 million. Possibly even 4 million common variants have been identified, and I feel pretty comfortable that by the time you guys graduate from college, we will have a nearly comprehensive list of all the human variants.

14. Filling in life's matrix: Genes, phenotypes, and SNPs (22:43)

So then what's left to do to be able to extend this? Well, all we have to do is fill out life's matrix. So we could imagine having GenBank. GenBank is actually the name for the computer database in which the sequences of all the genes in the genome is stored. But I mean to extend the concept of GenBank to also include all the variations in the letters on those sequences. So GenBank will have the whole human sequence, and every position that's a site of variation will be marked with the particular variants that occur there. So then how do we study human disease? Well, in a sense, it could be kind of simple. We take 1 very big matrix, 1 very big excel spreadsheet on our computer, and we start filling out. For the first gene, we'll write down all the different variants in that gene; for the second gene or region of the genome, all the different variants there; for the third, for the fourth, for the fifth. And then all we have to do in principle is examine a large collection of patients with diabetes and say, "Which variants are enriched? "Aha! There's a variant here in gene number 2. "There's a variant here in gene number 4 "and another one in gene number 4 "which occur far more often "in patients with diabetes. "Bingo! This gene must be involved." Next, hypertension. Same deal-- here's the variant. Next, height. We could try height. Same deal-- a variant, a variant. And in some sense, the whole idea of tracing out the connection between genetic variation-- inherited DNA variation-- and phenotypic variation could kind of become pretty straightforward. Now, of course, you'd have to feel comfortable testing 8 million variants or so, but if you could do that-- and I think we will be able to do that-- it becomes a very doable task: to use the fact that we're a very small population, in effect, to probe the basis of medical conditions.

15. Examples of genetic bases of human phenotypic variations (24:42)

How far can you go in that? Well, let me talk about some nonstandard kind of conditions: the ability to digest milk. This is not life-threatening, but most people on this planet can't digest milk as adults. They're lactose-intolerant. In fact, only a minority of folks can really digest milk into adulthood. Everybody can do it as a baby, because you drink mother's milk, but the enzyme that digests the lactose--called lactase-- turns off in adulthood usually. But in some people, it persists through life, and those people can drink milk. As it happens, the natural, original human state was to have that particular gene off in adulthood, but some mutations happened thousands of years ago that let it stay on. And when it stayed on, it was useful and favorable for people who were engaged in agriculture, and that state-- in fact, the gene has just been cloned in the last year. That variant has been identified, and people now know that that's a new mutation favorable in some populations, but other people have the native human state. Digesting alcohol: same thing. Some populations, particularly enriched in Asian populations, individuals have a particular mutation that makes an enzyme-- aldehyde dehydrogenase-- relatively inactive. And when they take a drink of alcohol, they get a very unpleasant facial flush. May actually be a good thing. They may become less likely to be alcoholic because of that, so it's actually perhaps even a useful thing. But again, it's an example of a general human difference in the population.

16. Can "Olympic gold medalist" be a phenotype with genetic basis? (26:26)

Let me take 1 really extreme case that I've put at the bottom of that slide: Olympic gold medals. I'll take another minute or 2 on this. Olympic gold medals. You might think that that's an example of something that's not genetic, and of course you're right, but not completely right. See, in 1964, a Finnish cross-country skier won 3 cross-country skiing medals in the Olympics. His name was Eero Mantyranta. Well, Eero was accused, by some people, of illegal blood doping. Back in 1964, they didn't have drugs for-- but what they thought Eero was doing was adding extra red blood cells to his circulation, because they found that he had 15% more red blood cells than the typical person and that that gave him the stamina to win these cross-country races. Well, there was no evidence to really support that charge against him. He kept his gold medals, and, you know, that was that until 1993, when a friend of mine-- Albert de la Chapelle-- in Helsinki, finally worked out how it was that Eero had 15% more red blood cells. Well, it turns out there's a hormone called erythropoietin that stimulates the production of red blood cells. It's illegal nowadays, by the way, to shoot yourself up with erythropoietin before the Olympics because it'll stimulate the production of red blood cells. Eero, it turns out, had a mutation in his erythropoietin receptor such that it always behaved as if it was active. He naturally behaved as if he had been taking these illegal injections of erythropoietin, but he wasn't taking illegal injections of erythropoietin. So, in fact, this was a natural example of a rare mutation that I think clearly helped him win a gold medal. But, we should say, is there any chance he would have won this gold medal without a tremendous amount of practice, without a tremendous amount of devotion to it? No, I think not. So, in fact, it's very much like, perhaps, the genetic variations that cause some people to be exceedingly tall and therefore have an advantage in the NBA. There are many tall people who still can't shoot a basket worth beans. It takes a great deal of practice as well as perhaps taking advantage of whatever genetic advantages and disadvantages you may have with respect to any activity. Well, let's stop there on genetic variation, and let me turn to the audience and take some questions. Then we'll come back and talk about RNA variation.

17. Q&A: Are mutation rates different in different species? (29:02)

Are there any questions? Yes? You said the probability of mutation in 1 human generation is about 2 times 10 to the negative eighth? Per base per generation. Two times 10 to the minus eighth, that's correct. For only humans, or for other species as well? That's very interesting. It varies a little bit. In the mouse, the probability of a mutation per generation is actually lower. But because they have many more generations per year than we do, the probability of a mutation per year in the mouse population is higher. So mice are actually mutating about 4 or 5 times faster per year but maybe 4 or 5 times slower per generation. But all things considered, that basic number is not so far off. Something like-- Well, I mean, let's put it in very concrete terms. Two times 10 to the minus eighth; you have a genome of 3 billion

letters. That means when you were born, you had 60 new mutations that weren't present in your father and 60 new mutations that weren't present in your mother. Sixty out of 3 billion. It's not a lot, but it's actually enough to provide the substrate for evolution to select upon over the course of long periods of time. It's pretty faithful, but not too faithful. In fact, it probably would be a bad idea for organisms to ever ratchet that down by another 2 orders of magnitude, because there wouldn't be any variation to select on. Great question. In fact, can I give you a T-shirt for your great question? There you go. Good catch. Yes?

18. Q&A: Has the mutation rate increased with a larger human population? (30:33)

Hi. I was wondering if the mutation factor-- has that increased since the population has gotten a lot larger? Or was--back in the day in Africa, when there was only 10 million... The rate of mutation? The rate, yeah. Has that increased, or is that staying the same because it's just, like, a general... No. The rate of mutation per person per generation seems to be determined by our biochemistry: by the accuracy of our DNA copying machinery, by the accuracy of the machinery that proofreads that DNA. And we have a pretty good idea that that has stayed constant at least for 40 or 50 million years, because we have some ways of reading out the sequence of the human genome and dating when various things occurred. So we can, in fact, by looking at some of these so-called junk DNA sequences that hop around-- and we know when they hopped-- see how much mutation they've accumulated. So we can actually get not just the rate currently for mutation, but we have a plot of the rate of mutation going back about 40, 50 million years, and it's pretty much a straight line. That's great.

19. Q&A: Would alcohol digestion problems affect alcoholism? (31:36)

Let me take 1 last question back there. You said that Asians can't digest alcohol easily. Would that make them more likely to be alcoholics if they did? No, less likely. OK, less likely. So, in fact-- and this isn't all Asians, and it's not all Europeans, but it's the case that some people have a particular mutation in a gene called aldehyde dehydrogenase-2 which causes them to be less able to digest alcohol. Because of that, acetaldehyde builds up, and they get a very uncomfortable sensation and a flush in their face. That's more common amongst Asians than it is amongst Europeans, but by no means absolutely universal. Yeah. That's just examples of many of the fascinating differences. So, in fact, you'll find that in Europe, you'll find that in Asia, but you'll find different frequencies there. Well, great. Great questions. There you go. Whoa! Almost.

20. Measuring variations in the levels of all RNA expressions (32:37)

Let me turn now to a different kind of variation. We've been talking up to now about variation in DNA sequence. That's 1 way in which nature has perturbed our system, and we're observing the consequences of that in terms of individuals' risks of getting diabetes or individuals' risks of Alzheimer's disease. But what I'd like to do now is look at another way in which we can read out information from life's matrix, and that's in terms of RNA variation. So when genes are turned on, they make an RNA message. And in the vast majority of cases, that RNA message is used to then make a protein which goes off and does a function. But if we could somehow peer into the cell and measure all of the RNA messages that were going on inside of the cell, boy, would we get a rich description of the biology of each cell and each disease. It would be in a way like just popping the hood on a car and looking under the hood and saying, "Wow. I can see what's going on, "and I can really diagnose what's wrong with the car" as compared to, say, just putting your ear to the hood of the car and trying to figure out what the problem was without being able to pop the hood.

21. Can differences in leukemias be detected by microscopy? (33:55)

Well, I want to describe a case in which we can really see the power of being able to look at that kind of variation. The case has to do with leukemias. You'll see here 2 pictures of a blood cancer called leukemia.

I'd like you to all look very closely, OK? Study the leukemia on the left. Now study the leukemia on the right. Who can spot the difference? What's the difference? They're closer together. What other differences do you see? Yes? Less little ones. Yes? More formation. Yes? They're clumping together more. Yes? ...on the left are more circular. More circular. The white ones have decreased in size. Wow. This is fascinating. You guys have incredible observational powers. Most clinicians who have looked closely at this would agree that there's no difference at all between these 2 particular kinds of leukemias; that, in fact, microscopically, you can't tell these apart. Now, of course, what you've seen is 1 field, and what you've said is absolutely right. In this field and in that field, it's a little more clumped, a little rounder, whatever, but those just happen to be what those 2 fields look like. If you actually looked at a larger sample from those 2 patients, you'd find yourself very hard-pressed to distinguish those. This, by the way, is a very good point about blind controls and things like that, because if you just were to hand it to people and say, "What's the difference?" people would not be shy about telling you the difference that they saw, but you'd have to really test to see if that was a meaningful difference that held up.

22. The discovery of two kinds of leukemia: AML and ALL (35:46)

Well, in fact, for many years, doctors taking patients with leukemia really couldn't see any difference between the different types of patients and their leukemias. And yet they observed that some of them did better on treatments than others. The treatments they had worked better for some than others. And 1 doctor in particular-- Sidney Farber-- working in Boston in the 1950s, decided to really pursue this observation that some patients seemed to do better on the therapies they had available. And he thought that it wasn't an accident; that there probably were 2 different kinds of leukemias but that we just didn't really have the way to see them, so he did what you were doing. He peered on the microscope a very long time trying to recognize this and trying to correlate it with the patient outcomes. And what he found was, he sort of could convince himself, like what you were doing, that the nuclei from 1 class of patients looked a little more granular than the nuclei from another group of patients, and that this first group of patients had a different outcome-- slightly-- than this other group of patients. I've gotta say, the treatments back then were pretty terrible. They weren't that successful. But he convinced himself. He didn't convince that many other people, because it really required you to say, "See? It kind of looks bumpy and grainy" and all that. Well, by the 1960s, enzyme tests came along, and he and his colleagues were able to show that, in fact, you could do enzymatic tests. And really, in fact, the leukemias on the left side behaved differently with regard to certain enzyme tests than the one on the right side. And then people developed immunohistochemical markers to look for things on the cell surface, and they began to distinguish that there were different molecules on the cell surface. And then they began to look at cytology, the structure of the chromosomes, and see that there were different chromosomal breaks. And so you were, in fact, correct. They are different kinds of leukemias. Today, the one on the left would be called AML, the one on the right ALL.

23. Limitations of conventional methods for diagnosing leukemia (37:43)

But it took 40 years of hard work, clawing up, to establish that as a rock-solid distinction in medicine. It took first convincing yourself with your eyes, then convincing yourself with enzymes, with surface markers, with chromosomal arrangements... and it turns out to be crucial. When a patient comes in today with acute leukemia, it's crucial to know whether they have AML or ALL because by now, 2 different treatments have been worked out-- 1 of which is pretty efficacious for the patients with ALL, another of which is pretty efficacious for the patients with AML. But if you give a patient the wrong treatment, their chance of survival is much lower than if you give them the right treatment. So it's a very heroic and very fun story-- a little more reminiscent of the story I told yesterday about the huge amount of work necessary to find a single gene. The only problem with the story is, it took 40 years. Could we do this a lot faster? Could we do this in... 10 minutes? That would be cool, because if we could do this in 10 minutes, we could try it for more cancers and more cancers and more cancers. So how could we do that? Well, we somehow have to pop the hood on the car. We've gotta look under the hood and see a much richer

description of the cell than what you're gonna see in the microscope or even a much richer description of the cell than you'll see by guessing with this or that enzyme. How can we do that? We'd like to actually measure the activity of every single gene in the cell. Well, it turns out that with genomics, it's possible to do that.

24.How to make a DNA Microarray (39:17)

The way you can do it is using one of these interesting devices. These are DNA microarrays. They come in different flavors and forms. They're made by somewhat different technologies and by different companies and academic laboratories. I've picked one here that comes in a very nice package. This has a little sliver of glass, a little square of glass in which, in those little squares that you see on the slide there, each square has a different DNA sequence. There's a specific 25-letter DNA sequence in that square and a different 25-letter DNA sequence in that square and a different one in that square and a different one in that square. Every one of these has whatever DNA sequence you would like to specify. You could type them in, and someone could make a DNA array that had different 25-letter sequences in it. How in the world could you do that? Well, I suppose you could go to the chemical laboratory and synthesize the first one and then come and stick it down, then synthesize the next one and stick it down and the next one and stick it down. But actually, the way they do this is to actually do it in parallel. They do it the same way that people make microprocessor chips in Silicon Valley. They have a mask, they shine a light on the glass. Where the light shines, the surface is deprotected, and you can wash on one of the DNA letters. You then reprotect the surface, shine a light through another mask, deprotect certain spots, and wash on the next letter. Shine a light through a mask, wash on another letter. After 100 such masks, you could build up an average of about 25 specific letters in each spot. Depending whether on each mask you had black or clear, you could either activate or not activate each spot and build up a specific sequence of DNA letters in each spot.

25. Using microarrays to detect the activities of all genes in a tumor (41:05)

So any sequences you want. Well, which one should you pick? Well, you could pick that the first spot would have the complementary sequence to the first gene in the human genome. The second spot could have the complementary Crick-Watson partner to the second gene and the third and the fourth, so that every spot could be a detector for its own gene. And then what you could do is take RNA from a cell... in fact, take a tumor. Grind up the tumor, prepare the RNA from the tumor, label it with something-- maybe a fluorescent chemical that we'll be able to follow-- take the RNA and inject it into this chip. There's a little hole back there. We'll inject it into the chip, swish it around, and each RNA will stick to its own detector by Crick and Watson double-helical base pair. Stick that in a scanner, the scanner will raster across there and read out the intensity of each spot and therefore tell you how much each gene's turned on and off. Way cool. These are-- You know, it's quite remarkable. You get a lot of information. In fact, every one of those reads... every one of those chips converts the tumor into a gene sequence-- gene 1, gene 2, gene 3-- saying which ones are low, low being blue. Ah, there's one that's high. There's one that's low. It becomes a long string of data. So what used to be just a picture in the microscope is now a huge string of data-- 1 read for every of the 30,000 genes. Here, have a chip. Yeah. Toss out some chips for folks. Give some more out. Chips, chips, chips... We don't have that many chips, but we'll just throw some chips around. There we go. Oh, back there? A few chips? All right.

26. Using microarrays to differentiate AML and ALL (42:56)

OK. So, more chips later. We've got all sorts of chips for people. So, now the question is, what can we do with that information? Well, it turns out that if we now go in and take a look at our AMLs and ALLs... if we look closely, the ALL tumors here all have certain genes high. Each column here represents a gene. Certain genes are high, and certain genes-- these are other genes here, these columns-- are low for the ALL tumors. For the AML tumors, this set of genes in these columns here are low, indicated by blue.

This set of genes in these columns here are high for AML. In other words-- Now, what I haven't done is shown you all 30,000 possible columns. They go off there. I've picked the columns that do a good job of distinguishing between ALLs and AMLs. If I now gave you a new tumor-- a patient comes in to the doctor, and we take their tumor, we put it on this chip, and we get a readout. I'd like you to diagnose it for me. So would you diagnose the following tumor? What would you say that is: AML or ALL? AML. AML. Looks like AML. How about this one? AML. This one? All. One hundred percent right. Could you write a computer program to do that? Sure could. We did. Gets it 100% right. So in fact, the computer can now accurately assign these samples to AML or ALL. The current way to do it is, a pathologist, you know, does all sorts of tests, but in fact, there's more than enough information in the RNA variation there to assign it to one class or the other. Well, Sidney Farber would be excited about that, but what would he really say to us? He'd say, "Well, that's good, "but, like, what did we have to do-- "How did we find this distinction "in the first place, right? "I mean, it took 40 years to find the distinction. "It's great that, given the distinction, "you're able to now diagnose patients, "but how about finding the distinction?" See, we're cheating here. We use the distinction between ALL and AML to get some known samples to figure out which genes to look at.

27. Using computers to sort RNA expression data (45:16)

Well... suppose we didn't already know the distinction between AML and ALL, and we just took a whole bunch of leukemia patients. We're back in 1950, before Sidney Farber's told anything apart, but we still have our DNA chips. Here's the data. We see genes going down the columns here. We see tumors going across here. Can you spot the difference? Same question I showed you before with those 2 tumors by eye. I'm now giving you a whole bunch of tumors here. Each line is a tumor. Can you see that they fall into 2 categories? How come? Tell me what you're doing. Yes? Some of the rows are red on the left side. Some of the rows are redder. And there are some that are blue all the way across; some that are blue only halfway; but they all would fall into 1 of those categories. Indeed, you could hand this to a computer and say, "Hey, computer, "could you in a completely unbiased, blind fashion, "divide up the samples into 2 types "and discover that there are 2 kinds?" And in fact, we wrote a computer program to do that, and we didn't tell the computer what was AML and what was ALL. We just gave it all the samples. And the computer sorted through and said, "Oh, yeah, I would say there's one type and... "there's another type." And the computer nailed it almost perfectly. The difference was that it took the computer about 10 seconds to spot the distinction. Now, I mean, we shouldn't be surprised. The tumor, after all, knows what it is. We just hadn't been asking up to now. But clearly there's a very rich difference between these 2 kinds of tumors.

28. Discovering a novel type of leukemia (47:07)

Well, you can imagine what leukemia doctors said about this. They said, "Sounds nice, but we already knew this. "Tell us something we don't know." So we tried to do that. We took ALLs and tried to ask, "Are there, in fact, really further subtypes of ALL "that folks have been missing up to now?" In fact, there was a hint that there might be. Some patients with ALL have a particular mutation in the gene called MLL, and it happens that those patients tend to have a poor prognosis and to tend to be somewhat more-- it's enriched for infants. And so the question was, "Are there really "2 different types?" So what do you do? Get a whole bunch of patients now with just ALL. Prepare their tumors, prepare RNA from their tumors, put them on gene chips, read out the data, and then see if the computer can sort them. And bingo, the computer sorts them beautifully. There's actually 2 flavors of ALL: ALL 1 and ALL 2. So as of the last year, we really think that there should not be a distinction of 2, but really 3 types of these acute leukemias. Moreover, if we look at the genes carefully, some of these genes are very interesting. One of them in particular, right here, encodes a gene called the FLT3 kinase, whatever that is. What was interesting for us was that there happens to be, for other purposes, a drug against the activity of the FLT3 kinase-- a small molecule. Lo and behold, we can take that drug, apply it to patients' leukemia cells in vitro, and it turns out to kill those ALLs that fall into class 1 but not class 2. And so in fact, people are

now gonna try a clinical trial with this already existing drug to see if it might even provide a treatment here for this subtype of ALL.

29. Building taxonomies for tumors and other biological functions (49:03)

Well, you can imagine how to take this further. People are building global cancer maps. They're taking zillions of different tumors and trying to get the whole expression patterns of the RNA variation in all of those tumors and sort them out-- lung tumors, breast tumors, prostate tumors-- and be able to figure out, "How are those different cancers different from each other? "How, in fact, can they be split even further?" And to try to get a real molecular taxonomy. It is, in fact, the case of observing, in great detail, what is going on. By observing the RNA variation or the DNA variation, we can begin to put together nature's own taxonomy. Now, in fact, this observation is occurring because nature has already perturbed. And as you hear, once we've observed these differences, to convince ourselves that we really understand, the only way to do that is to be able to go back and perturb the system again. So observation and perturbation. Let's take some questions. Yes?

30. Q&A: How accessible is microarray technology to doctors? (50:08)

How accessible is this technology to most doctors? Are doctors working side by side with research scientists? Oh, in terms of real clinical applications today, I would say this is not broadly being used-- either the DNA variation or the RNA variation-- because most of what we're talking about today is stuff that's only been appearing in research papers in the last 3, 4, 5 years. And so these distinctions, for example, between the ALL 1 and ALL 2 is about a year or so old. There are similar distinctions for lymphomas and breast cancers and other things, but they're still at the research level. We've gotta first get multiple groups around the world to confirm those tests, and then reduce it to a kind of laboratory test that could be used at the bedside. I think there's a lot of work to do in the next 5 or 6 years to be able to take both the DNA variation tests and the RNA variation tests and package them up in such a way that they're reliable to use. We're even gonna need much larger population samples to be able to-- I mean, it's all well and good to have you diagnose the ALL and the AML here based on that, but the right way to do this is run a clinical trial where we look at hundreds of samples and figure out, "How often do we occasionally get that wrong?" and "What else might it be?" And so we're at the point now where I think this is extremely exciting science, where we can just look over the horizon and see this being deployed in the clinic, but not quite yet. I think it'll take some years to be able to do, but not so many years. Small numbers. Thank you for a great question.

31. Q&A: Are some mutations inherently bad and selected against? (51:39)

Say you come across a certain DNA mutation, and I was thinking that this could possibly run into an ethical dilemma. Because if it's a detrimental mutation, this is sort of nature's way of saying, "Hey, I don't want this "to become a part of the population. "This isn't a gene "that's advantageous to this organism. "They shouldn't be able to reproduce "and continue making that detrimental mutation "in the population." Is that somewhat a way of, you know, kind of continuing this disease to move on into the population? So you're saying, "Are some DNA variants inherently bad things "that should die out?" Maybe there are some things that are really extreme, but I'm a little careful about that because things aren't good or bad by themselves. They're good or bad in a context, usually. We have a whole bunch of DNA variants that, um... Well, put it this way. None of us are adapted to live in the middle of the sort of snowstorms we live in here. We should all be dying out in this environment. But in fact, we've changed our environment so that the genotype we have does just fine in this environment. So I don't know that it's any different than that. To say that something is a disease-causing mutation, when I can change your environment by changing your diet or giving you a drug or something like that so that you live perfectly fine with it, I'm not sure it's a disease anymore if it's treatable in that fashion. And therefore I wouldn't attach a normative value to "That's a variant that should die out." I think that's a variant where we should fix the environment

so it's not a problem. But these are very important questions you're asking. Here's a T-shirt for you. A question there.

32. Q&A: Wouldn't DNA variation confound the microarray detector mechanism? (53:20)

You were talking today about how 1 SNP can change the way that an entire protein fold when it's made, and then you were showing us those, uh, those, I guess, gene chip-type data things. And you sort of said that, oh, it's pink, or it's red, so it hybridized pretty well, and so those ones are all in 1 category. But if it comes out red in 1, and it comes out pink in the other, that's because there was 1 or 2 bases that didn't quite hybridize perfectly, and those make all the difference. So how can you, uh, sort of put all the red and pink ones together? You are very insightful. What a spectacular question. You're worried about the fact that the DNA sequence of the gene might differ at the DNA level and therefore affect my detector at the RNA level. So you've put together the 2 types of variation we've been talking about. A DNA sequence difference might interfere with that RNA detector we built. You're right. So maybe what we'd better do is put down not 1 detector for 1 particular 25-base pair sequence, but 2 or 3 or 20. And if I'd really given you the proper details, I'd tell you that in fact, there were 20 independent probes for each of the genes down there to deal with exactly the question you've raised. Great question. Let me take 1 last question.

33. Q&A: Are DNA microchips reusable? (54:44)

You've got the last question. For the DNA microchips, after you use it once, is there any way to remove the nucleotide...base pairs and use the chip again? Ha ha ha ha! What a great question. Is this disposable or reusable? As you might imagine, the company that makes this and sells it for some fair amount of money prefers that this be disposable. Researchers out there would sure love to find ways to wash it out and reuse it. And this is-- the constant thing that the salesmen are always asked is, "How often can I reuse this chip?" You're thinking like a great lab scientist already who is husbanding his or her budget to make it go as far as possible. Well, at the moment, I think they're not so safe to reuse, because we want to have absolutely rock-solid results since we want to have it absolutely reproducible. But my own sense is that we have to look to a world, not so many years from now, where these things become so cheap-- you know, a couple bucks rather than a couple hundred bucks, which they are today-- here, have a chip for a couple hundred bucks. And, you know, where we can be doing this on any old experiment we want to. Now, I don't have these chips in complete supply for everybody-- and these aren't actually ones I would advise you to put any DNA on-- but the chips are made in another variety. Including these glass wafers here with all of the nucleotides, chips are also made by the same company in chocolate, and so I have chocolate chips here. And to thank you all for being a spectacular audience, we have, for everybody, chocolate chips. Thanks very much. Here we go. Have a chocolate...

34. Closing remarks by HHMI Vice President Dr. Peter Bruns (56:28)

Well, thank you, Eric, for that terrific talk and, students, for the wonderful questions. I'll point out you can keep asking questions-- not only today but in the future if you're watching this video next year on DVD-- because again, at our web site, called biointeractive, we have a site called "Ask a Scientist," where our group of volunteer scientists will answer your question and will get back to you. You can e-mail your question, and they'll get back to you. In fact, if it's a question that's really good, we'll put the question and answer on the web site. We're going to take a short break now, and then we'll return for the final lecture by Dr. Schreiber, who is going to continue his discussion about chemical genetics and talk about his activity to build a thing called ChemBank, which is a large, large database cataloguing the biological activity of small molecules. So please join us for Lecture 4.