COLUMBIA250 COLUMBIA UNIVERSITY DIGITAL KNOWLEDGE VENTURES

Genes and Genomes: Impact on Medicine and Society

Genes, Genomes and Evolution
October 16, 2003

Isidore S. Edelman, M.D., Columbia University
Gerald D. Fischbach, M.D., Executive Vice President for Health and Biomedical Sciences and Dean of the Faculty of Medicine, Columbia University

Welcome and Introduction

Welcome by Isidore S. Edelman

Isidore S. Edelman: We should begin. We're going to make an attempt to keep the morning symposium on track, on schedule. This is the symposium on genes' and genomes' impact on medicine and society. And it's part of Columbia University's 250th anniversary.

To initiate this program I'm calling on Dr. Gerald D. Fischbach, who is the executive vice president of the Health Sciences and dean of the Faculty of Medicine at Columbia University.

I'll tell you a little bit about Dr. Fischbach and then ask him to come to the podium. Dr. Fischbach earned his M.D. degree from Cornell University; after an internship he spent six years in neurobiological research at the NIH. His exceptional accomplishments led to endowed professorships at Washington University in St. Louis and at Harvard University, and he was elected to the United States National Academy of Sciences based on his research accomplishments in 1983.

Dr. Fischbach made seminal discoveries on induction of receptors in neuromuscular junctions. His current research is on synaptic trophic factors and nerve-cell survival. The latter topic is of compelling interest to everyone over the age of sixty. Dr. Fischbach chaired departments of neurobiology at Washington University and Harvard Medical School, also served as director of the National Institute of Neurological Disorders and Stroke at the NIH. Dr. Fischbach.

Introduction by Gerald D. Fischbach

Gerald D. Fischbach: Good morning. My job is to give you some introduction to the symposium and some commentary about Columbia, and those go well together because genes, genomes, and science is very appropriate to help begin Columbia's 250th anniversary, for at least two reasons that I can think of.

One, we are still in the midst—and we will be for far into the future—of enormous advances in genome science and genetics. Since the discovery of the structure of DNA fifty years ago, all the work in the previous fifty years has expanded enormously. As we understand how DNA replicates and how it can be modified, it touches every aspect of our lives, from basic science through social sciences and history and forensics, and you'll hear about that through the symposium.

The second reason is that Columbia has played an extraordinary role in this history, and I'm going to dwell a little bit on the T. H. Morgan and the Fly Room and give you a brief overview of what has happened at Columbia since then, and I will end up with something I know more about, the brain and how our knowledge of gene expression in the brain has enormous implications for neurodegenerative disorders.

I want to start by thanking the two people who really organized this symposium, Joanna Rubenstein and Tom Jessell, who in a very intense period of time planned the symposium, and then Joanna in particular followed through and made sure that everything worked well. And I think you're in for a treat during today's sessions and tomorrow morning's. So let me begin with my introduction, and I'm going to need help. How do I begin my slides? This is the problem with putting them on one computer.

T.H. Morgan and Colleagues at Columbia

This story begins with Thomas Hunt Morgan, who as a young man studied biology at Johns Hopkins University with an extraordinary group of people, his fellow students being Edward G. Conklin, Ross Harrison, and a number of others. They trained under a man named H. Newell Martin, who was very imbued with physiological thinking, having trained and been influenced by Michael Foster in England, who developed the first school of physiology, indeed transformed the field of physiology and England with it, University College, training John Langley and leading in a direct path to many of the greats of electrophysiology that form the basis of our field, Charles Sherrington and John Eccles included. But it was this physiological thinking and Morgan's mentor, I believe, at least what I've been able to read, that convinced Morgan that anatomy would not solve the problems he was interested in. And he wanted to manipulate embryos to study the developmental dynamics of how embryos adapted as they matured. He tried very hard after he moved to Bryn Mawr in 1891 to become chairman of the department at Bryn Mawr at the age of 25. Now you would think that was an onerous job, but there really was only one other member of the department at the time, and that was Jacques Loeb, who became a lifelong friend of Morgan's. He displaced E. B. Wilson, a wonderful cell biologist who in 1891 moved to Columbia, and this pair teamed up later on with Wilson, focusing on the cytology of chromosomes and how chromosomes may be involved in heritability and inheritance, and Morgan, as you will see, focusing on the physiology and the genetics. But here's Jacques Loeb, who he met when he moved to Bryn Mawr in 1891, and they remained lifelong friends.

This is another tie to Columbia: Jacques Loeb is the father of Robert Loeb, probably the most distinguished chairman of medicine in Columbia's history. But Jacques and Morgan did much of their work at Bryn Mawr together, but then they spent 63 summers at the marine biology lab in Woods Hole. And last night John Loeb, Jacques' grandson, told me a wonderful story that when they worked in the same building Jacques Loeb was quite annoyed by an organ grinder who appeared every afternoon outside of his window and played music, and he could not concentrate, and he was quite frustrated. But the organ grinder would never leave. And it was years later that he found out that Morgan, whose lab was on the third floor, kept throwing pennies out the window to keep the organ grinder in place, to annoy Loeb. It's not the only reason Morgan won the Nobel Prize, but it certainly counted for something.

Well, Morgan, as you must realize, became interested in evolution. And it's hard to imagine now, because our thinking is so ingrained in Morgan's thoughts, that when he came to Columbia in 1904 he did not believe in Mendel's laws, and he did not think that chromosomes had anything to do with heritability. Indeed he thought that heritable substances were present in the cytoplasm and that they were influenced by environmental factors. He just didn't believe the anatomy that E. B. Wilson and others were working on. But he felt that one way to approach it was to try and make changes in the chromosome structure of individual animals, and understand whether chromosomes could influence the outcome, the phenotype, of the animal. He tried for many years to work with vertebrate animals, but he realized very soon that he could never approach the level of analysis that he was interested in with these vertebrate animals. He basically wanted to test hypotheses put forward by DeVries and others, seen here with Morgan in front of his home—I don't think you can see it clearly, but the address on this building is 409, that's 409 West 117th Street, where Morgan lived for 28 years while he was at Columbia.

Morgan's Fruit-Fly Experiments

DeVries had a theory that species change by macroscopic mutations, huge mutations, along the lines of Darwin's sports; there was not a gradual series of changes but once in a while there would be a massive change and a new morphologically identifiable species would emerge. And Morgan set out to test that hypothesis. It was along the lines of his interest in adaptation: how do organisms adapt to the environment? And in 1907–1908 he began to work on fruit flies—the story of this is quite interesting how this all came about, involving putting bananas on the window ledge and letting flies migrate into the lab and congregate. But he did not have much space; he worked in a room, as you'll see these colleagues later, that was 16 by 23 feet, so he needed an animal that didn't take up much space and that reproduced rapidly. And he found that *Drosophila* could reproduce every 12 days, whereas the mouse, of course, was much, much longer, and the other animals he was working with even longer than that.

Out of that room, by the way, came a total—the people who worked in that room, and I'll show you some of them later—came a total of about 24 books, most of them transforming the field between Morgan, Dobzhansky, Sturtevant, and his other colleagues. And I'll touch on some of that later, one of the most productive groupings of scientists in the history of modern biology.

Well, Morgan labored for two years, trying to induce changes in the appearance of these flies to test the hypothesis that DeVries had settled on. And then one day one member, one student in the lab, noticed that in one bottle, one milk bottle of these flies which they had been growing for two years and injecting all sorts of toxic chemicals into, that one of these flies had white instead of brown eyes. And I won't go through the details of the genetics, but it was this observation that changed Morgan's way of thinking about chromosomes and about genes and their roles in inheritance. He bred these flies together, the white flies with their sibling red-fly parents, and then he bred them again, and at the end of these breeding experiment he realized that the white-fly trait seemed to correlate with a particular chromosome, and it was the chromosome that determined the sex of the fly, the X chromosome. And it was that basic observation that convinced Morgan that chromosomes did play a role in heritability and that genes, the term defined elsewhere, that genes must be located on these chromosomes and that heritability really did depend on the nucleus after all.

And he noticed many other things in those years after 1910. He noticed that certain genes, he discovered many other mutations, were inherited together, they seemed to be linked, but once in a while these genes that were inherited together became uncoupled, and he developed the whole notion of recombination between chromosomes, a concept that was further extended by individuals in the lab here. Here Morgan is in the Fly Room at Columbia in Schermerhorn Hall. And here's a picture of two of his young colleagues, Alfred Sturtevant on the left and Calvin Bridges on the right. And I'm showing you this to show you that these were kids; the people that came to Morgan's lab were students at Columbia. Sturtevant was in his third year, and I think when Bridges joined he was a freshman at Columbia. But the story goes, and I encourage you to read the book by Garland Allen about Thomas Hunt Morgan, the story goes that Sturtevant went home one night from the lab, understood Morgan's notion of recombination of chromosomes, and instead of doing his normal homework assignment developed the first map, genetic map, of a chromosome in the fly, came back in the morning and had many of the genes that the lab had discovered over the years mapped out, and that was the beginning of a technique of genetic mapping that has been extremely powerful right up to the time of the direct physical sequencing. These people stayed with Morgan throughout their career; they remained at Columbia until 1928, and they all left for Cal Tech at that time.

This period was extremely exciting; it something that Sidney Brenner has called "Morgan's Diversion"; that is, he was really interested in embryology and in regeneration and how organisms adapt, but this diversion into genetics created an extraordinary field and gave rise to molecular genetics rather directly from the people who came to Columbia, and those people who came to join him at Cal Tech, if not directly in his lab, came because of the atmosphere of genetic work that he described. They moved onto new organisms and they developed the field of molecular genetics.

These are his collaborators: Hermann Muller, who won his own Nobel Prize in 1942 for developing techniques for creating mutations by X-radiation and further studying recombination; Alfred Sturtevant, who I showed you; Calvin Bridges, I showed you; and Theodosius Dobzhansky, who joined him at Columbia, moved to Cal Tech, and then extended these studies to a series of brilliant investigations of population genetics.

The Relevance of Flies

This is a picture of the Fly Room, and you can see these guys sitting together. When you read accounts of this experimental group, the main thing that comes across is it was just fun, people enjoyed what they were doing, they challenged each other, and they were extremely productive over the years. And I have incorporated this into my talk to the faculty who ask me for more space every day. This is a lab of less than 400 square feet of space, and probably the most productive germinal zone in the history of Columbia, maybe in American science.

Now I want to deviate for a minute and tell you a bit about how flies are still relevant. My main interest at the NIH was in neurodegenerative disorders, and one of these disorders is Parkinson's Disease. And I'm going to tell you how flies have had a huge impact on the human disorder of Parkinson's Disease. I don't have a pointer here, but . . . There is a group of nerve cells in our brains—I'm reluctant to try this on this computer—is there a pointer here? —deep in the middle of the brain, which make the neurotransmitter dopamine. Eric, you're tall. A group of nerve cells in the middle of the brain, deep in the region called the midbrain, that make a neurotransmitter called dopamine, and I won't have time to tell you the details of this and how it was discovered. But these nerve cells send processes called axons throughout the brain and when they begin to degenerate, as they do in Parkinson's Disease, a critical region of the brain called the basal ganglia become depleted of dopamine and the hallmarks, the movement hallmarks, of Parkinson's Disease, a paucity of movement, a fine three-per-second tremor, and a very disabling rigidity of the limbs, which takes enormous energy and is draining, ensues. And there have been a number of attempts to understand this disorder from a genetic point of view. There is a form of the disease that runs in families, and several years ago a very large family was analyzed by a group at the NIH and in New Jersey, and they understood that a piece of abnormal DNA seemed to go along, be present in members of the family that had Parkinson's Disease. And analyzing that piece of abnormal DNA, using the molecular techniques developed by the children of Morgan, his scientific children, they identified an abnormal mutation in a gene encoding a protein that was quite well known called synuclein, and that has proved enormous helpful in our analysis of the mechanisms of Parkinson's Disease.

You can see here a section through that region of the brain as depleted of the neurotransmitter dopamine, shown here as depletion of these small stained cells, which were stained with an antibody against the enzyme that makes dopamine.

Well how does that occur? This has been analyzed in flies, in *Drosophila*. A group a few years ago at Harvard, Mel Feany, who was then just out of her M.D. / Ph.D. degree, introduced synuclein, the human gene for synuclein, into the fruit fly, Morgan's fruit flies, and this gene was expressed, the mutant gene, was expressed in every single nerve cell in the brain. But the remarkable thing is that only those nerve cells that contain the neurotransmitter dopamine degenerated. Here are the normal cells on the left, and here's the absence of those cells on the right. The fly duplicated one of the great mysteries of human neurodegenerative disease: in a gene that is expressed everywhere in the body, why are only a few cells vulnerable? And these flies developed a movement disorder, and if you are generous and stretch your imagination you can think of this as homologous to the movement disorder in human beings.

Here you can see the ability of these flies to climb out of a right circular cylinder, a water glass versus the age of the fly. And normal flies can do this over a period of time and then decline when they reach about 45 or 50 days. But the animals bearing mutations in synuclein, or even just overexpressing the human synuclein, lose this ability to climb out of the glass at an earlier age. So the age dependence of Parkinson's Disease and the selectivity for the neurons evident in humans is also evident in *Drosophila*. So the homologies extend beyond molecules to behavior. And one has the feeling with the genome sequence in the *Drosophila* that we will soon understand the factors that influence both the age dependence and the neuronal specificity.

Columbia's Human Genetics Research

I'm not going to be able to—we haven't time to extend this further, but I did want to trace for you a bit more of the history of Columbia, the creation of the department of human genetics that Paul Marks chaired initially in 1969 when he brought together people like Elvin Kabat and Sol Spiegelman, who helped transform the way we think about DNA, especially DNA-RNA hybridization. And I do want to mention the work, the sad work, of Erwin Chargaff because I think no story of Columbia science would be complete without that, and then I will stop.

Erwin Chargaff was very influenced by Avery's findings of DNA as the transforming principle in pneumococcus. Avery, by the way, who worked at Rockefeller, was a Columbia graduate as well. And Chargaff's very careful biochemistry, using one of the first ultraviolet spectrophotometers and paper chromatography, showed that the ratio of bases in DNA followed a certain rule, that the amount of adenine was equal to thiamine, and the amount of guanine was equal to cytosine, and he felt always that this pair ruling was influential in the Watson-Crick Model of the structure of DNA. And I have a quote from him which I find very moving, not because of the outcome of the science but because of a life in science and what it may mean to be disappointed. He did not share in the Nobel Prize and he felt he was never recognized properly, and his own personality probably complicated that measure. But he did say, "This was a time when I began to feel awfully alone, neither country nor profession, neither language nor society, and not even the tranquil and reverential inspection of nature seemed to offer a refuge. 'We all die in an armor of ice,' I used to say, but I was not yet 55. The orderly, loving and careful study of life had been replaced by a frantic and noisy search for breakthrough." And one can't help but empathize with this man at age 55. He lived to age 96 and never returned to the vigorous, vibrant science that he once conducted. So the hope is that we will all think about this during the symposium, think about the spirit and the fun that Morgan typified, and remember that this is for the science, not for the personal gain as we go forward in genetics. So I think I will end here. I would like to introduce the person who introduced me. Izzy Edelman, when he came to Columbia from San Francisco, began and continued to create what I believe is one of the most remarkable departments of science anywhere in the country, recruiting wonderful biochemists and molecular biologists. His own work on the sodium-potassium ATPase is renowned, but it's his leadership in science in that department and throughout the University here and before that in San Francisco, which has really been so extraordinary. Izzy.

Sydney Brenner, Ph.D., Salk Institute, La Jolla, CA
From Genes to Organisms

Introduction by Isidore S. Edelman

Isidore S. Edelman: Dr. Sydney Brenner will speak to us on a topic from genes to organisms. Dr. Brenner was born and achieved his baccalaureate in the Union of South Africa. He earned a doctor of philosophy at Oxford University in the UK, and soon thereafter joined the MRC Unit in Cambridge. His successes in the

MRC are quite remarkable, spanning cell biology, developmental biology and molecular genetics. He introduced and exploited one of the most powerful model systems, the worm, with the marvelous name of *Caenorhabditis elegans*. Dr. Brenner served as director of MRC Laboratories for a decade and a half. In the 1990s, he translocated to California, where he is now a distinguished professor in the Salk Institute in La Jolla. He played a particularly important role in the International Genome Initiative. He introduced the pufferfish as a model system for studies on gene expression, owing to its paucity of introns. Some of the difficulty in understanding molecular genetics has to do with the fact that the active genes are interspersed between regions which are not involved directly in gene function. Two of the model systems which were chosen for total genome sequencing, *C. elegans* and the pufferfish, Fugu, were championed by Dr. Brenner.

Time does not permit a listing of all of Dr. Brenner's honors and awards, but I will read his 2002 Nobel Prize citation, which he shared with Dr. Robert Horvitz and John Sulston. "For their discoveries concerning genetic regulation of organ development and programmed cell death" he was awarded the Nobel Prize, and his cane is right here.

The Natural Complexity of Genes

Sydney Brenner: Well, I'm very pleased to be attending yet another celebration at Columbia. The last one I came to was the 200th anniversary of Columbia medical school, not of the University. That was in 1968 and we were recalling that it had been held in the Armory and during the course of the lectures people started to drive trucks around behind the screen. I think it was only Paul Marks who put on a uniform and ordered them out of the lecture hall. But of course there are other anniversaries that we are celebrating this year, and that is the fiftieth anniversary of the discovery of DNA. And I understand that next year we will be celebrating the fifty-first anniversary of the discovery of DNA.

So I've had quite a load of celebration, and it's given me a lot of time to think about what it is we actually did. And when I saw that I was asked to speak on a topic called "Genes Are History" I didn't know how to interpret that, whether it was going to be the history of the genes themselves or whether it was going to be a history of genetics, a talk about genes. And in fact then I decided I would give one of my standard titles, because I discovered you only need two titles to talk about the whole of genetics. The first lecture is called "From Organisms to Genes," that deals with the past, most of the present, and "From Genes to Organisms" deals with the future. So, if you like, this is a history of the future. I am still waiting for a misprint in my title where the letters NI are left out. You can work that out. The solution of that conundrum is left to the reader.

So I think we are at a very interesting junction in genetics, but first we should state what the project is all about, what is the project of genetics? And it's quite simple. Biological systems are unique in the world of natural complexity; they are the only naturally complex systems that contain an internal description. I once went to a meeting where we were addressed by a Buddhist—he was called an archbishop, but I'm sure they have other names—and he was asked What is the Buddhist definition of life?" And he said in true Buddhist flexibility, he said, "Well, some Buddhists think everything is alive; mountains are alive, rivers are alive." So I stopped him and said, "Mountains are not alive." So he asked a good question, he said, "How do you know?" I said, "You can't clone a mountain." A mountain contains no internal description, as all living systems do. And I think our task has been always, is, how do we map, if you like, genotype onto phenotype? Because if you think about this lineal sequence of bases within all of us, and if you think about the complexity of structure and function that we all have, that we human beings have a nervous system, a very complex thing, and what are, we might ask, the transformations of going from this linear sequence into the structure of the organism?

The Sequencing Revolution

Now, of course, there was a second revolution in biology, in molecular biology, which took place about 1975, and that was the development of techniques of cloning DNA, and in particular development of techniques of sequencing DNA. Because up to that point we had only been able to study the DNA sequence through the rather opaque lens of genetics, that is, the only way you could tell changes in the sequence was by mutation. We could map mutation, with fine structured genetics with a kind of failed attempt to sequence DNA by its effects on the phenotype. But now for the first time we could get direct access to the genetic material, and we could read the base sequence there. This alters the problem into one which goes from an inverse problem to a forward problem, and science is always interested, as I shall point out, in solving forward problems.

And of course it raises the question of can we compute organisms? That's the forward problem, can we compute organisms from DNA sequences? If we solve that we solve all the questions in biology. And this is the issue that I want to take up here, because it is the issue of what is the theoretical basis that we will need to construct in the future in order to put that problem at our disposal and aim to work towards a solution of it.

Now, of course, just to reflect back again, the ability to sequence DNA, to get the direct access to the sequence, generated the Human Genome Sequencing Project in 1985 when, I should remind you, that the biggest genome sequence to date was 45,000–46,000 bases, the genome of lambda bacteriophage. And this rather heady group decided they would take a jump from 45,000 to 3 billion. Of course that will need technology, that will need money. And it was put together. I for one thought the technology was not yet able to deal with that size, but of course they hadn't realized the importance of the techniques of DNA sequencing which have come home. So I must point out to you that DNA sequencing is a totally unique technique, technology, nothing else is like it. You can take DNA from any source, plants, animals, bacteria, your mother-in-law, and you can put it through a machine which extracts the essential information, which is the lineal sequencing bases. That means you can essentially make a factory of this, and this means that if you tackle big projects you just have big—more machines. So it's something I called "3M science." It stands for money, machine and management. And the concept of factory production, which I have to say was first put forward by Wally Gilbert actually, and he saw very clearly that effectively you could expand this technology and obtain the sequences of anything, simply because it's unique. Nothing else in biology is like that; proteins are not like that, cells are not like that. We cannot do everything by this method because the information, relevant information, that we have to extract is not a lineal sequence of bases but is at a different level of complexity. So I disbelieved strongly in all -*omic* science beyond genomics simply because of the uniqueness of this.

Now, make no mistake, I don't disbelieve in parallel observations, that's a different thing; it's a subject we used to call spectroscopy. So I think we should do spectroscopy which would make measurements as effective as possible, but I don't think, as you will see in a moment, that we can extract anything significant from it. Now that's, of course, a hard thing and probably a daring thing to say at this stage, but I would like to develop the logic of this. There's now a great movement that we will pin everything we know, hang everything we know, onto the genome. It is in fact true that ultimately the fundamental description lies at the level of the sequence of bases. But it is not true that it is implemented in that way. So we have to distinguish between something that in fact Schrödinger has failed to distinguish between, that is, the nature of the information that exists in the DNA and it's the execution, or the implementation, of that information. In fact Schrödinger has stated in his book *What Is Life?* that the DNA contains, as he says, a program for development and the means of executing it. It doesn't—contains a program for development and it contains a description of the means of executing it, but not the means themselves. And that was, of course, earlier enunciated very beautiful in the '40s by John von Neumann, who showed what was essentially required in terms of logic for self-reproducing automata of which you could see biological systems as one example. In fact, I strongly believe that that is probably the first, most important theoretical concept to have been produced in which we can have a level of abstraction that could have actually led one to consider DNA simply from a logical basis, but never did, because von Neumann's work was not known to biologists.

Starting the Cell-Map Project

Very well. So what is the best way to look at this information in terms of how we are going to put it all together and how we're going to actually use it? Now I believe very strongly that the fundamental unit, the correct level of abstraction, is the cell and not the genome. In other words, I've been quoted as saying "forget the genome," you know, we don't want to forget it, we'd like to thank all those people for their sterling work and give them all a gold watch and send them home, or better still send them back to the factory to sequence more genomes. But what we've got to do now is to get away from that and look at how we're going to give the true biological picture of it.

And I want to say this because this is quite important, and a student many years ago came up to me after a lecture and said, "Dr. Brenner, what is going to be the breakthrough in the nervous system?" And I said, "You're about fifty years late, it's already happened. It's called the neuron hypothesis." And of course if you look at what people thought about brains before the neuron hypothesis, it was very clear they would absolutely nowhere. But it was the idea that there were these units called neurons, that they were connected in various ways that underpin the framework of all explanation in the future. And of course the neuron is just another cell. And the cell theory, which was put forward more than a century ago, 150 years or so ago, that is the one that just said everything's just made out of cells, you start off with one cell, have cell division, cells grow, produce a lot of different cells, and so that's clearly the way to get back to looking at organisms.

And I think if I can explain. So of course all projects should get a name, because then they get—I call this the "Cell Map Project." It exists only as a plan, but I think it is one I hope that everybody will participate in the future, because that is the way, I think, the only way to organize our activity. And I want to just sketch out roughly how we—I would see this developing in the future.

The Cell as a City

Well, let me begin with an analogy, because I think that's very useful to think about. Let's look at a city, New York. Now, all of the people who live in New York know how New York works. All of the strangers who come here have to learn how it works, otherwise they just don't understand it at all. And cities in general work at least to, let us say, a martian observer as follows. There are a whole lot of units, they're called houses. And every morning they dissociate subunits. The subunits then go in all different directions where they aggregate again into other complex assemblages, which have names on them like banks, hospitals, stores, etcetera, etcetera. And if you study the function of the city, you can see that the function must be defined in terms of these assemblages. So for example, you'll see that banks, money goes in, lots of it, very little comes out. Schools, all of these functions. Right. So I want you to envisage all of that complexity, all of those, if you like, all of those interactions. And they're not to be considered as a matrix of interactions of all the people in New York with each other, but simply that there are—that matrix indeed is quite sparse, but there are, if you like, strong interactions and weaker interactions. The strong interactions between all the employees of the bank and another set of strong interactions of all the employees in the hospital.

Now, I would like you to think of the genome sequence of the white pages of the telephone directory. Now, what we hope is that they've got it all accurate, there are no mistakes. When you call the dean, you won't find yourself in the House of Pleasure, for example. So we will have that. This is a list and of course we want that, and of course the one thing you can do with sequencing is obey what I call the CAP Principle: complete, accurate, and permanent. You never have to do it again, it's a permanent resource in biology once we have done it.

What people are thinking about doing now is to compile the yellow pages; this is the annotated genome. Now I mean it's a great thing to know there are seven plumbers on one block, in fact, just the existence of plumbers might make you think there must be something for them to plumb, so maybe there are pipes hidden underneath those streets, which plumbers will get into. But, you see, you will still have a fragmentary description, just glimpses of pieces of it. So you must reflect the organization of it, and you must put that organization in right from the start.

Well, let me give you one quick example of what I propose, and that is the following. Everybody will hold their hands up in horror and say, "Look in a cell! Maybe we have twenty thousand genes expressing. How are we going to cope with twenty thousand different polypeptide chains?" The one lesson you should learn is that if you can't cope, neither can a cell. And cells have learnt to view the world as composed of something like income tax. In other words, it's criminal to evade, but there's a legal means of avoiding paying income tax. And I think what we have to say is when we find things that prima facie would just seem impossible, you can be sure that evolution has found the way of avoiding them. All right. And this is exactly what nature has done.

The Cell as a System of Gadgets

So the first thing to realize is a cell is not a bag of protein molecules all buzzing around and interacting with each other, it's not like that at all. When you go into it you find that no gene product hardly ever acts on itself; it usually acts within a molecular assemblage. And some of these assemblages are quite complex. The assemblage that actually splices out introns contains the products of 65 genes. And so when you go through this, you'll realize immediately that there is a strong reduction of the complexity, perhaps by an order of magnitude, so that if I call these things "devices" or "gadgets" rather than the term "molecular machine" because I think these have much more important functions to reveal than you might have expected from a mechanical device, if you call these "devices," we have condensed the problem immediately into two thousand such gadgets that exist in the cell.

And the next thing you observe is that the cell is not homogeneous, it has a topography. What goes on in the nucleus is different from what goes on in the cytoplasm, and what goes on at a membrane is different from what goes on in the mitochondria. And just for argument's sake, let us just think of this as again ten topographical regions in the cell. So that the problem is reduced then by thinking about it through this kind of modularization. And now we're only thinking of two hundred gadgets or so per topographical region, and that means that we can start to look separately at how these gadgets communicate with each other, and the problem then becomes totally digestible, in my opinion. And if you look at the way it's done, it's done stepwise. And let me say that it's all written in the genome like that, because it's written in the genome of a protein that it should have a certain sequence of amino acids in a certain configuration so it can bond with another protein which has similar properties, and that these then can bond with yet others. So that is what I call strong interactions.

And so you come to a model of—the cell model being a theory that effectively there is a—we would look at it as a graph of gadgets connecting with signals passing between these, and you'll see into that logical structure we can easily assemble not only true signaling, but even metabolic pathways can be thought of as signals, chemical signals, passing between gadgets and connecting between them.

The Next 50 Years: Networks of Cells

Now, of course one question we'd like to know, and this has been purely a sketch, is how many different kinds of cells are there in the body? Well, you can go to a textbook of histology—I used to teach histology—and maybe there are two hundred different cell types, smooth muscle, striated muscle, etcetera, etcetera. You can write a list. Now when we come to the nervous system we just don't know. It is quite feasible that in the retina that 26 kinds of amacrine cells have been characterized. So there may be, for all we know, in terms of what I call noncontingent states, that is, states that do not depend on environment; there may be as many as a thousand such states in the brain. That remains to be found out, and effectively I think that is the first task that we have to do, is to discover how many of the noncontingent states there are.

Now of course, there are contingent differences as well, differences due to stimulation, differences due to learning. But I don't consider a neuron that has learnt something as a different cell type from its naïve neighbor, it is just something—because the capacity to learn, the capacity to learn is in effect a contingen—a noncontingent state, which allows the thing then to, so to speak, fill in the form with what its connections with environment [are].

So if I were to return to Earth, as I'm threatening to do for one day on April 26, 2053, which is the hundredth anniversary of the publication of the Watson-Crick model, in order just to have a look around and see what's happened—promise to go back from wherever I came from—just want to have a look, but I hope that by 2053 we would not be groaning as we are today saying, "What are we going to do? What's all this information?" Because I think that if the last year, the last fifty years, has seen us elucidate the nature of biological information at the chemical level, if you like, it's just been the chemistry, which means also the physics of biological information—I think what we are going to do in the next fifty years is to elucidate the biochemistry, if I can call it that, of biological organization. And I think we will come to see that we will be preoccupied with networks, which are graphs, that we will see, I hope, organisms as a network of cells, with cells connecting with each other, just as we will see the cell as a network of molecules. And that once more, I hope, that we will use the structure into which we will feed the genome, and not try to hang everything on it.

Rethinking the Gene

So I just want to finish up with one last remark. When after Herculean efforts it had been discovered that we only have thirty thousand genes, amazing to some, insulting to most, that they had eight times the number as *E. coli*, people were very puzzled as to why we have so few genes. When it was discovered effectively as could be well predicted that all vertebrates will have the same number of genes, because evolution can only take place at this level, at the level of reorganizing what you do with the same genes rather than having explicitly human genes—in fact, the way we think about a gene is totally wrong today. We think about a gene for some function, and I like to tell the story—which I happen to have invented—about the difference between chimpanzees and humans. And everybody has felt that when they get the final sequences humans will have one extra gene, the one for language, what we call the Chomsky gene. But of course what you don't realize is that chimpanzees have learnt that talking gets you into trouble—just look what's happening to us—and so they evolved a language-suppressor gene. So they'll have the extra one, and it'll be called the Chimpsky gene, of course. But I give this as a lesson, you see, because now language must surely be a very complex thing; it involves vocalization, you can find a huge list of things, and you don't go from not speaking to speaking with one gene. I mean otherwise we would see the mouses, the perfect model for dyslexia. Mice cannot read, and therefore we should study dyslexia with your ordinary mouse.

But one thing that's interesting is complex systems can be broken in many different ways and [in] the way of breaking them, each of these appears as a unique gene. So what we are really interested in are the genes for normality, if you like, and we're interested in finding out how those work. And only when we find out how those work will we understand then how you do this.

Solving the Forward Problem

Now, I said I would follow up on something which I think has only come home to me in the last few months. There are people who will tell you the following: you should study the system. In fact, I believe a whole subject has been created, called systems biology, which we used to call physiology, actually, but if they want a new name for it, fine. But systems biology has a program, or at least certain proponents of it have a program, which we should ask whether it's feasible at all to do it this way. Now the idea being that when you study a lot of things at the same time, there are some things called emergent phenomena, things will happen there. Because everybody will tell you that a system is greater than the sum of the parts, and when you put a lot of things together new things emerge.

Now that statement itself is nonsense; it's nonsense because they haven't quite the correct definition of a system. A system, the whole, is the sum of the parts and their interaction, okay? Because it is true that the whole is greater than the sum of the parts studied in isolation, that "studied in isolation" is what has been left out. The whole can not be more than the sum of the parts and their interactions, or else we're getting off into a nonscientific view of how things work. There can't be mysterious essences flowing around.

So, if you make a set of multiple observations, could you deduce what's going on in the system? Could you model the system as it's expounded and deduce what's going on? These are classic inverse problems—look at the result, try to work out the causes, look at the effects, try to work out the causes.

Now, such systems are generally also called ill-posed systems, because they don't give rise to unique solutions, they're ambiguous, they're not continuous; this whole thing was specified as to what a well-posed explanation will be. And what I want you to do is to think of the following: if someone plays a drum and you record the sound and you make an analysis of the sound, could you work out the structure of the drum, the physical structure of the drum? Now that is a classic inverse problem; it is also classically ill-posed because information has been lost due to interference and all sorts of other things. And so therefore only if you make certain assumptions, which is called regularization, can you effectively use the sound to tell you about the drum.

But there's another way, and that's the way I think science should go. Get hold of the drum and find out what it looks like; what's it built out of, how big is it, and then you can solve the forward problem which is easy, you can play the drum yourself. And that is what I think is going to be the essential methodological difference. So if we were to call systems biology of going from the end result, from the phenotype, to try to deduce what is inside the organism, I think that it is doomed to failure. On the other hand, if we find out what is in the organism and solve the forward problem, that is the only path to success. So if want names, we'll call the latter part computational biology because we're going to compute what there is.

And as a last word I would say the following: although people have said "more is better," I would like to point out that the least is best. That is, find the least you have to find out and you can predict the rest. That's what science is all about, and I hope that we will encourage people very strongly not to become factory hands in some vast institute, but to really get out there, find those drums, you can play them yourself.

Thank you very much.

Isidore S. Edelman: We're now scheduled for what is generally called a coffee break, but you're not obligated to drink coffee. The—it is my intention to restart the symposium at 10:30 sharp. Thank you.

Svante Pääbo, Ph.D., Max-Planck-Institute for Evolutionary Anthropology, Leipzig, Germany

The Genome as a Mosaic of Human Evolution

Introduction by Isidore S. Edelman

Tempest Fugit. I'm going to ask the people in front to sit down. All these eminent people are keeping us from our program. Okay.

The next talk will be given by Dr. Svante Pääbo, entitled "The Mosaic in Human Evolution"—I'm sorry, "The Genome as a Mosaic in Human Evolution."

Dr. Svante Pääbo was born and educated in Stockholm, Sweden. He completed the curriculum in medicine and the Ph.D. degree at the University of Uppsala in the 1980s. After three years of postdoctoral research at UC Berkeley, he was appointed to a full professorship at the University of Munich in Germany. In 1997, at still a young age from the standpoint of a career in biology, he was appointed director of the Max Planck Institute for Evolutionary Anthropology in Leipzig.

His major contributions include detailed analysis of the genetic basis of human diversity, a subject of enormous interest today, and the crucial role of gene expression in speciation. We know that the genetic composition of all primates is highly similar, but speciation goes behind the straightforward genetic composition. His recent work on genes involved in the evolution of human speech and in olfaction are at the very forefront of human evolutionary anthropology.

Dr. Pääbo's exceptional contributions to biological science have been recognized by numerous awards and honors, too many to list. I will mention honorary doctorates, the Max Delbruck Medal, the Ernst Schering Prize, an election to the Royal Academy of Sciences in Stockholm. Dr. Pääbo.

Studying Related Genomes

Well, thank you very much, ladies and gentlemen. So, what I'd like to start out by saying then, is that as we have learnt from so many celebrations this year, this year 2003 is indeed an historical year, not only because this great university was founded 250 years ago, but also because fifty years ago the structure of DNA was discovered by Watson and Crick, and in this very year, almost on the day fifty years later, the almost complete structure of one human genome was made available. The challenge that many of us in biology now stand in front of is trying to make sense of this enormous amount of information presented to us. And it's also clear, I think, to most of us that one major way, if not the major way, of doing that is to perform comparisons between this genome and the genomes of other organisms. And the complete genomes that are available to us today are genomes from quite distantly related organisms, from the worm, from the fly, from the mouse, from the rat, from the dog, and so on.

So what we then do when we compare the human DNA sequence, a little segment of our genome, and compare it, for example, to a fish, we'll find that most parts we see no similarities at all, and then we find regions of conservation. And those are, of course, very important, that's what we look for, evolutionary conservation, because these features might point to features that are crucial importance for being an animal, for being a vertebrate, for being a mammal, and so on.

However, as we now get sort of the capability to sequence many vertebrate genomes, we will progressively encounter situations where we can also study closely related genomes. And we are about to enter that era for vertebrates, as now the chimpanzee genome will be available in a few months. And what we will then have to look for is not conservation, because almost everything will be conserved, but instead we will have to look for differences. So in the case of humans then what we will compare ourselves to is our closest evolutionary relative, and that is the two chimpanzee species, common chimps and bonobos that separated in the order of maybe five million years ago from our lineage. And that then translates to quite little DNA sequence. So for parts of the genome that we can actually line and compare, the average amount of difference is somewhere around a bit over 1 percent. So that means then that in this case when we compare ourselves to a chimpanzee, almost everything is identical, and here and there, on an average of one base in a hundred will differ. So there's very little difference.

However, when we now multiply this with three billion bases in our genome, we will end up with a lot of differences, in the order for forty million differences, and so on. This is still an underestimate because it doesn't sort of include rearrangements and duplications, and so on.

Making Sense of Our Differences

So the challenge we will stand in front of is trying to make sense of all these differences, define the differences that may be of importance among all these differences. And why will this be important? Well, it will be important, I think, because we're interested in finding the genetic basis for human-specific traits. And many of these traits we're interested in are sort of normal human traits, the genes for normality in humans, if you like. So this may be for sort of complex cognitive abilities, an example being language, aspects of our aging process, and so on. There's also a number of sort of medical conditions that are specific to humans; some of them are specific to humans simply because they affect some of these human-specific normal traits, and others are more trivial, such as susceptibility to certain infectious diseases. And it's clear that most of these traits we're interested in will be very complex, very many genes will influence those traits. For some of them, though, there may be fewer sort of critical genes, and in some cases there may be even single critical genes.

And it's also clear already now that this will be a tall order to find these genomic regions, and it will not be enough simply to compare the genomes of the typical graduate student here with the genome of a chimpanzee, because wherever we sort of find differences, the first problem we will stand in front of will be that we don't know what the ancestor out there looked like, so we don't know if a difference we observe here in the genomes is a change that happened on a line to the chimpanzee and is then of particular interest to the chimpanzees, or if it's happened on a line to humans and is of particular interest to us. So the first thing we will actually need is a third primate genome—and that will be available within two or three years, and it will not be the orangutan depicted here but the rhesus macaque—because then we will be able to say whenever the rhesus is identical to the chimp it's very likely that we deal with a change from the human lineage; when the human is identical to the rhesus, it's likely to be change here.

What we will also need to study, sort of functions of the genomes, the transcriptome and the proteome, and I'll show you our, some of our sort of humble start at doing that in a few minutes. But I'd like to submit to you that what we also need to do, much more than we've done so far, is to study what I'd call the phenome of humans and the great apes in a comparative way, because for many of these features we are interested in we actually have no clear clue how they differ in a detailed way from the chimps. It's easy to say, for example, that humans have language and chimpanzees do not, but we do know that we can teach a chimpanzee something like two hundred words and symbols, and the chimpanzee will also use them in an adequate fashion. So that is sort of much of the idea behind the institute that we started around five years ago in Leipzig where we tried to, many of the researchers tried to, study in a comparative fashion humans and the great apes from various different aspects. There is a department of primatology, for example, that has research stations in Africa where they study chimpanzees and gorillas to try to define what is common to the cultures of chimpanzees in different regions and what differs between different chimpanzee cultures. There's a department of comparative psychology that does similar experiments in children of humans and of the great apes, so they have a primate facility in the zoo where they have social groups of all the great apes, the bonobos and chimps, gorillas, orangutans, and try to define the actual differences, especially during the first year of life, how we explore the world. There's the department of linguistics that try to find common features among human languages. There's a department of paleontology and a department of genetics that we will focus more on now.

So there are of course two general approaches we can take from a genomic standpoint when we would like to approach sort of human-specific traits. One is to look across the whole genome to look for interesting patterns, and one approach to that is to study how genes are turned on and off, the transcriptomics—I'll mention a little bit about that. And this then produces candidate genes that you can also recognize from other studies, for example, disease mutations. So the other way into this are candidate approaches and there are two examples of that I'd like to mention.

Transcriptomics of the Brain

So if we start then with the transcriptomics, this is particularly the work of Philipp Khaitovich in our lab, and it actually starts by studying then transcriptome of the brain and is expanding at the moment to many other tissues in the body. But if we start by looking at within-species expression differences in the brains of humans and chimpanzees, we look at eight different regions, five regions of the cerebral cortex, very different regions, say from the prefrontal cortex to the primary visual cortex, or to the limbic system, and so on, and two other regions, one the cerebellum and one the caudate nucleus. And we do comparisons primarily within an individual to minimize differences due to different environmental influences. So if we look on the overview over what you will find in a multi-dimensional scaling plot you will find that you—from the pattern of gene expression over, in this case, twelve thousand genes, you can easily distinguish a caudate nucleus from the cerebral cortex from the cerebellum in three humans.

However, what is of interest is that these different parts of the cerebral cortex—what was surprising to us is that we cannot distinguish the cerebral cortex according to the different parts of it that we looked at, rather the cluster according to individuals, that these are the cerebral cortex samples from one individual, from another individual, and a third individual. So this means then that the transcriptional pattern in my prefrontal cortex is more similar to what I have in my primary visual cortex than in the prefrontal cortex of someone else in this room. We don't know what this depends on; it may be simple environmental influences, having to do with how much coffee I had for breakfast or so, or it could be more interesting things in how the brain is formed during development. We see exactly the same pattern in the chimpanzees, with different—these big regions—and don't differentiate cerebral cortex parts here. We can then ask what are the genes that do differ between different parts of the brain, and we can use sort of functional annotation of genes to ask that question. So the Gene Ontology Consortium has three sort of hierarchies: cellular component, biological process, and molecular function of gene functions ascribed to genes, and it would test statistically. Each of these actually differ significantly if we look at different regions of the brain in comparing the genes that have changed their expression and those compared to all detected genes.

So if we focus in on each of these, for example, cellular component, and these are the things that have significantly changed as groups, we find that that involves membrane proteins and plasma membrane proteins. If we look at the biological process, that involves single transduction as a big group, and the other big group is neurogenesis. If we finally look at molecular function what we find there is single transduction again, and for the other big group are kinases.

So if look on sort of the genes that differ between different parts of our cerebral cortex or different more coarse parts of our brain, they are—the groups, the big groups that differ—are plasma membrane proteins involved in single transduction. If we then go and compare between species, looking now between humans and chimpanzees, we have three humans here and three chimps—nine comparisons for each region—and we require then that there is a significant difference in all nine comparisons to score this as a difference.

The first thing we find is that you actually find surprisingly many genes that differ. Fifteen percent of genes differ in their activity in at least one of the six regions of the brain that we compared in this way. We can ask what genes differ between the species now, and now only one of these hierarchies here, biological process, shows significant amount of change. And we focus in on the groups that change; it is in a sense quite disappointing. There is no big single group that changed, and it's very hard to make sort of functional sense of these reference groups together.

However, there are things here when we look more carefully at the data that do look interesting. For example, if we say over, in this case, forty thousand genes assayed, how many genes do we find that are unique in their expression pattern to the caudate nucleus, eight hundred genes. To the cerebellum it's something like three thousand genes, and to these two different parts of the cerebral cortex it's a lot less because they are so similar to each other, 25e here and 29 there. But within parenthesis here I have written the number of genes that is unique to, say, cingulate cortex all in humans or all in chimpanzees, or to the Broca's area here and here. And what you will see is that although there are very much fewer genes unique to the region in the cerebral cortex and elsewhere, proportionally the ones that are unique to one species or the other is much, much more. So it may indeed be that one way in which the cerebral cortex is unique in humans and in chimpanzees is that we have more species-specific differences between different regions of the cortex. So obviously those genes now that show a species-specific difference between different parts of the cortex in humans, not in chimpanzees, are candidate genes that we want to focus on and study much more carefully. And this work is only beginning, and it then feeds into these candidate approaches. And the candidate approaches I'd like to sort of illustrate to you by two other classes of candidates that come from other types of studies.

Olfactory Genes and Pseudogenes

And the first involves many, many genes involved in olfaction. And this is work by a graduate student, Yoav Gilad, in the lab, and it's a collaboration with Doron Lancet in Rehovot, Israel. And as you know, I think, at this university particularly well the olfactory receptor genes is the largest single group of genes in our genome; around a thousand genes deal with olfaction, and it may be surprising maybe that such a large—the largest group of genes deal with the sense that we in our everyday life don't regard as extremely important. However—and we most often think of our nose, the organ that you use for this, as not very prominent in our bodies. However, if we would unfold the olfactory epithelium and have an olfactory disc instead, it would be about this size in our faces. And a thousand genes is also not that much if we consider the fact that we can sense in the environment millions of different odorants and discriminate them from each other even at very low concentrations. And that's achieved by the fact that most odorants will bind to several different receptors, different cells in our olfactory epithelium, they will signal to the brain that will put together a picture that is sort of called a smell.

What drew our attention and the attention of others also to the olfactory receptors is that if you look in the mouse genome sequence how many of the thousand or a little over a thousand genes that the mouse has for smell are pseudogenes, have mutations in them that inactivate the gene, it's a rather small percentage, something like 15 percent of the genes in the mouse are nonfunctional. If you look in the human genome sequence, it looks very different. There, over half of these thousand genes are nonfunctional, even taking a very strict criterion we expect to see a stop codon in the protein coding reading frame. So from this olfactory disc, then, a large part of it seems to be nonfunctional in humans, whereas a quite small part nonfunctional in a mouse.

So what Yoav said was let's go and look at our closest relatives, the great apes, on olfactory repertoire. And what he did was to by a random-number generator select fifty olfactory receptor genes, make primers for them and first sequencing them in the human and the mouse to see that they're indeed representative of what we know is the situation, and indeed the human has 56 percent pseudogenes, and the mouse has 16. So the question is now for the great apes are they like humans in having a lot of pseudogenes or like the mouse in having little? And the answer is somewhere in-between. They have around 30 percent, so significantly more than the mouse, but significantly less than the human. So there are here two jumps, if you like, from the mouse to the great apes and the rhesus macaque, and from the great apes to humans. I'd like to say a few words about this increase in pseudogenes and then a few words about that.

Human Olfactory Pseudogenes

So, if we now just look in the phylogeny of primates here on the number of mutations that are called pseudogenes, and not are genes, it looks like this. So we see a rather steady rate according to the length of the lineages, but on the line to human there's something like three or four times more such mutations that knock out genes than on the chimp line. So the question then becomes is this a process that is still ongoing or is it something that happened in some defined time in our past? And how would we see if it's ongoing? Well, we would see that we find segregating pseudogenes, that we would find in the population today people that carry an allele that's functional and other individuals that carry nonfunctional alleles or stop codons. So Yoav selected 32 genes that are annotated in the human genome as pseudogenes, having a stop codon in them, a single stop codon and sequenced them in 14 humans from all around the world. For 13 of these 32 genes it was what you expected: all individuals carried pseudogenes. For 7 of them it was more disturbing, because all these 28 chromosomes in these 14 individuals carried intact genes. So we don't know what's going on; it may be an error in the human genome sequence, which does occur, or it may be that the human genome sequence happened to catch a very rare variant.

More interesting were 12 of these, a third or so. We found indeed that some individuals carried functional variants and some nonfunctional variants. So indeed it seems to be the case that we are losing the sense of smell as we speak, so to say. So what is then going on is that now and again a gene is knocked out. That of course means that an odorant that would bind only to this receptor will not be sensed at all by us. But that's very rare, as we said. More common is that an odorant will bind this receptor as well as other receptors, and then losing one of these receptors probably means that we have less discrimination, that we are less able to distinguish closely related chemical odorants, and perhaps also that we need a higher concentration in environment to sense it, although we don't know that.

This also means that we differ a lot from each other in our olfactory receptor repertoire, and this is now data from Doron Lancet's lab, it doesn't come from our lab at all, but these are 26 segregating pseudogenes that they analyzed, in this case in a hundred individuals. Wherever an individual has two functions it reads as dark green; when it's heterozygous, it's light green; and when it's homozygous for pseudogenes, it's red. So you will see that olfactory receptor genes were hardly—where it's very common to not carry any functional variants where others where almost everyone does it, and you also find individuals that have a third of loci homozygous or pseudogenes, others that have very much fewer. There are good smellers and bad smellers around.

Interactions Between Sensory Genes

So a segment then of this is pseudogenes, we don't know quite the proportion of it yet, and this was then what we at the moment know about this jump here. If we then looked at the difference between mouse and the great apes here in this data, what you want to do is to go and look in very many other primates, so not only the great apes but Old-World monkeys and the New-World monkeys that are further away from us. And in each individual species, sequence a hundred olfactory receptor genes in these eighteen species. And if we just plot the fraction of pseudogenes now, in humans it's around 50 or 60, in apes it's around 30, and in a mouse it's 15. If we go to the other Old-World monkeys, they look exactly like the great apes. If we now go to the New-World monkeys, they look like the mouse and the lemur.

But those of you who are observant will have seen that there seem to be one funny guy here among the New-World monkeys that look like the great apes and the Old-World monkeys. And who's that? Turns out to be the howler monkey, it's this individual that lives in South America, and this is surprising because in this phylogeny then it means that twice independent in the history of primates, have we lost around 15 percent of our olfactory receptor pseudogenes, once in the common ancestor of apes and Old World monkeys here, and once on the lineage to the howler monkey. So this then begs the question, is there something else that Old-World monkeys, apes and us have in common with the howler monkey? And there is; otherwise, I would probably not bring it up. And it's full trichromatic color vision or routine trichromatic color vision. So seeing colors with three pigments in the eye is common to all these creatures, and it's also present in the howler monkey, whereas other New-World monkeys have either just two pigments, even just one pigment, or has trichromatic color vision as a polymorphism in females.

So there seems to be this correlation here between the loss of olfactory receptor genes and the gain of full trichromatic color vision in all members of the species. It's of course very hard to prove that this is really sort of related to each other, but I think it points to a very interesting area, namely sort of interactions between the part of our genomes that deal with the senses. So there may be this connection between olfaction and vision. Olfaction and taste is another obvious one, of course, and we're looking at all the taste receptors now in the great apes and humans and other primates. Hearing is another one. And I think that sort of when we now analyze these genomes, the part of it that deals with the senses will directly interact with environment and will be particularly interesting.

The Evolution of a Speech Gene

So for the last few minutes I'd like to talk about one other candidate gene, and that involves language and speech, and it's a collaboration with Tony Monaco's group in Cambridge—in Oxford. And they studied this very famous family in the UK, the KE-family, that goes over four generations, and where a severe language and speech problem segregates it, so severe that naïve listeners will generally not be able to understand affected individuals, but family members, for example, will learn to. And you will see here that it segregates in a family pretty much as an autosomal-dominant condition. And this allowed Tony to find the gene—it's called FOXP2, it's a member of the forkhead-domain transcription-factor genes, and they had actually two different families with disruptions in these genes that had very similar clinical features.

So what we did, together with Tony, was to study the evolution of this gene. And it encodes a protein of 715 amino acids, and we started very naïvely by saying let's look in the mouse, a mouse doesn't speak, so it shouldn't have this gene or have it in a very different form. And we sequenced it in the mouse—it wasn't available in the mouse database at that time—and it turned out to have only three differences to the human. It's one of the most conserved genes we have between humans and mice. So it obviously had very many important functions that we share with the mouse.

It looked a little bit more interesting, though, when we went to the other great apes because it turned out if we look in the chimpanzee, for example, it has only one difference to the mouse. So if we look in a tree sense from, say, the chimpanzee or the common ancestor of human and chimps to the mouse, there's only one amino acid came from this gene over, say, 140 million years ago or so, whereas over the five million years to humans there are two amino acid changes here.

And if you look where they are located within the gene, they are close to each other in one exon here. Now this looks compelling, but it's so few observations, of course, it could still be just by chance that these two events happen to happen in the human lineage. However, what Wolfi argued was that if this was indeed positively selected recent in human history, we should be able to see it in the pattern of variations surrounding these two mutations here. They went in sequence, forty chromosomes for 14-kb adjacent to this mutation. We have to become a little sort of theoretical here and say that what would be expect to see in the frequency spectrum if this was positively selected.

Well, if some time in the past we had a number of chromosomes in the population with variation in them, the white balls here, and then a positively selected mutation happened, perhaps one or both of these amino acid changes that allows carriers of this chromosome to reproduce better than the others. What will happen is that this region of the chromosome will sweep through the population, become fixed, become present, in all individuals. And then slowly mutations will start accumulating around it again. However, it's not only that simple, because we also have recombination going on, and the effect of recombination will be to sort of this sweep, lift in all variations adjacent to this selected mutation here. So if we now with recombination compare what we expect to see after a selective event with a region where no selection had happened, there was no selection that happened. We expect to see an excess of rare variants present, say, only one of the forty chromosomes compared to here, and also an excess of frequent variants present, say, in all 39 chromosomes except one compared to here.

Genetic Adaptation for Language?

So if now look at our data for FOXP2 and what we would expect under a model of no positive selection, if we just plot the number of sequence variants present in one chromosome, in two chromosomes, in three chromosomes, up to 39, we'd expect a distribution like this. And the actual data we find looks like this, so we see this excess of rare variants in the sample and also, very tellingly, an excess of very common variants.

Now you may ask is this a common feature? If we look on every other gene on the genome maybe they have this pattern. That is not the case. If we, for example, look on data from [inaudible] where they looked at this excess of rare variants in over three hundred genes, FOXP2 was the second-most extreme case that we know of. So clearly FOXP2 is an example of a gene that was positively selected recently in human evolution. Does that then mean that this is the Chomsky mutation, this is the gene that if we put it in a chimp the chimp would start to speak to us? I certainly don't think so. First of all, the only thing we know is that if we knock out a copy of this gene, we have a language and speech problem. We also know that it was positively selected recently in human history. We don't know that language was necessarily the thing that caused the positive selection. It's the most reasonable hypothesis at the moment.

And even if we make that leap of faith and say that it was language that caused that, then of course if we look on the phenotype in the family, it seems that the main problem that affects them is muscle coordination, the millisecond coordination I need between my vocal chords, my tongue, and my lips to produce articulate speech. So if we make that second leap of faith and say that this is what was selected in FOXP2 that, of course, is something that would only be selected in a society where already vocal communication was quite important. So where we already have some kind of protolanguage, and FOXP2 would then be kind of an adaptation to language or an improvement of language. And the members of several Chomsky-like mutations we know already—one other name of this larynx moving down during childhood—that has sort of adapted us to language.

An interesting question, though, is of course when this would have happened in human history; there are interesting time points that we would like to sort of check. For example, when human forms left Africa two million years ago, was that already then driven, if we believe this, by language as facilitated by FOXP2? Or if we see modern humans separated from other work, you know, separated around half a million years ago from the *Neandertals*. Do we share this with the *Neandertals* or not? It's very hard to sort of date these fixed sweeps; you have to make a lot of assumptions that I will not go through for you, but under these assumptions, under this model, you can then count an approximate 95 percent confidence interval of a time point when this sweep became fixed, or later. So it is within—under these assumptions—within the last six thousand generations, so 120 thousand years or so.

Our assumptions are somewhat shaky; we may have to push this back maybe equally much as this date, but then we would still arrive somewhere around 200, maybe maximum250, thousand years. So we can certainly exclude the common ancestor with the *erectus* here, and probably we can also exclude the common ancestor with the *Neandertals* and say that this selective sweep happened on the human lineage. But it's still, of course, only a hypothesis that this had to do with language.

Future Studies on Speech Genes

So I would like to end with a few words about the problems we now stand in front of with this new class of genes, where FOXP2 is one of the first, which are sort of probably involved or may be involved in human-specific traits. And that is how do we come forward to actually prove that this hypothesis would be right? But what we would like to do is to sort of take the FOX-chimp, FOXP2, and put it in the transgenic human; we'd like to take the human FOXP2 and put in the transgenic chimp and see if they can speak to each other. But there are problems with getting that through the ethics committees.

So what can we do with these type of genes to come forward? Well, one thing we could do is sort of screen a lot of humans for backmutations, so to say, to their ancestral state in FOXP2. So far we've looked on over five thousand individuals and haven't found any such mutations. We can also try to develop tissue culture models; for example, FOXP2 with transcription factors. You can express the two forms in a cell and see what genes have turned on and off. And you can try to move to an animal model, that is, putting the chimp and human versions in mice, or even better maybe putting in the human-specific changes in the mouse FOXP2 gene. So you would then go and move down now, put in the two critical amino acid changes that we believe might be critical into the mouse FOXP2.

That mouse had just been born, but then you stand in front of the next problem in this, you're studying a human-specific trait in an animal model. So we will obviously try to speak to the mouse, but in case that doesn't work you then stand in front of trying to design assays for the phenotype, where there may be things here that you can guess from the family that may be able to test in the mice, such as perception of tone length, for example. However, I think this points to problem in this field for the next few years, namely that it will take a lot of sort of biological insight, creativity, and in fact intuition perhaps even, to design sort of models for human-specific traits in animals. So although I then end on sort of stressing some of the problems we stand in front of, I do hope that I have convinced you that the next few years will be extremely interesting when we will now in a systematic fashion be able to study not only the changes in our genome that have happened since we diverged from the chimpanzee and that are unique to the human lineage, but identify those changes that have been positively selected and therefore have been of—per definition, of importance in forming the unique characters of our species.

I thank you for your attention.

Michael Levine, Ph.D., University of California, Berkeley, CA
Whole-Genome Analysis of *Drosophila* Gastrulation

Introduction by Isidore S. Edelman

Isidore S. Edelman: Dr. Fischbach opened this morning's symposium with a description of the role of the fruit-fly model in the broad field of genetics. And now we bring this morning's session to a close with the return to contemporary fruit-fly research.

Dr. Michael Levine will speak on whole-genome analysis in *Drosophila* gastrulation. Dr. Levine was born and educated in the United States and obtained a Ph.D. degree in biology from Yale University with distinguished scientist Alan Garen. His postdoctoral research with Walter Gehring at the University of Basel in Switzerland and with Gerry Rubin at UC Berkeley put him on a fast track in the field of molecular developmental biology, and he proved up to the task. His first faculty position was in the Department of Biological Sciences at Columbia University, followed by progressively more senior appointments at the UC San Diego and UC Berkeley. He is currently professor and codirector of the Center for Integrative Genomics at UC Berkeley.

Dr. Levine has made major contributions to our understanding of how specific genes contribute to embryonic development in the model fruit-fly system. He was awarded the prestigious Monsanto Prize in Molecular Biology by the United States National Academy of Sciences, and was elected to the academy in 1998. Dr. Levine.

The Levine Connection to Columbia

Michael Levine: Now I don't know if this gizmo works okay, I've got a little hip microphone here so I can wander around. Can you hear me in the back, is it okay? Yeah. Happy birthday, Columbia University. I'm going to show you that it's possible to link the origins of this university with my last name, Levine, and it goes like this. When you think about Columbia, of course, you have to think about Kings College, and when you think about Kings College you think about its most illustrious student, Alexander Hamilton, the great American. Now it's true that Hamilton did not graduate because of the Revolutionary War—pretty good excuse. Now at the time that Hamilton's father married his mother, Rachel, Rachel's name was Rachel Fawcett Levine, L-E-V-I-N-E, all right? That's pretty—that's a shocker on Shock Street. But

[Eric Kandel calls out something from the audience.]

Okay, but see now Eric Kandel's getting nervous. Alexander Hamilton, however, did not have a Jewish mitochondria, okay, what happened is that Rachel, a Huguenot, at a young age married a Jewish merchant named John Michael Levine, and then remarried James Hamilton. So, you know, he's clean but still, you got the link. Levine—okay.

Now for a long time—that's, I don't know—that's my favorite trivia question, you know, what was Rachel's last name? All right.

Evolution and Gene Regulation

So for a long time my lab has studied gene regulation in the early fruit-fly embryo, the early *Drosophila* embryo, and most recently we've really focused these studies on the process of gastrulation, and before I launch into that I'd like to spend a few minutes giving a general introduction to the importance of gene regulation in development and evolution.

Now this first slide shows the seven different animal genomes that had been sequenced and assembled at the end of 2002. And as you'll probably hear later from Eric Lander, this slide is already hopelessly out of date; the number of genomes is skyrocketing, this number's likely to double every year or two for the foreseeable future. But this is what we had at the end of 2002. And you've already heard from Sydney the shocking revelation that came from these genome assemblies, namely that the human genome contains only thirty thousand genes, probably fewer than thirty thousand genes. Although I would argue, since much of that human genome assembly is based on Craig Venter's genome itself, there may be some key genes missing. You know, it could be ten thousand genes that give rise to decency, courtesy, you know, the—okay, never mind. I'm teasing, I'm teasing—for crying out loud. You're New Yorkers, you don't know how to take a joke?

All right, so humans have thirty thousand genes and they have about the same number in all other vertebrates, including the primitive pufferfish. And what's worse is that it's possible to align most of the human genomes one-to-one with another gene, corresponding gene, in another vertebrate. So this suggests that the evolutionary diversification of different vertebrate groups did not depend on the invention of new genes. And this kind of logic—again as Sydney alluded to—can be extended into the invertebrates such as sea squirts, fruit flies, and nematode worms. A typical invertebrate genome contains 15 thousand genes, half the number in vertebrates, but the main difference in gene number between invertebrates and vertebrates is not the invention of new genes but rather the duplication of old ones. So, for example, the fly genome contains three FGF genes; a typical vertebrate genome contains more like twenty FGF genes, which arose from duplication. So it's an aggressive and oversimplified view, but one can make the argument that animal diversification during evolution really involves deploying the same basic set of genes in new ways, in different ways. So evolution, at least in part, could be viewed as a problem in gene regulation.

Animal Regulatory DNAs

The next slide summarizes different regulatory DNAs that are seen in animal genomes. So the diagram on top shows two hypothetical genes, one to the left and one to the right. Now let's say that the gene on the right is regulated by this enhancer and silencer. Enhancers are the most prominent regulatory DNA which in general turn genes out, let's say particular cell types of particular tissues. So if this a mammalian genome, this enhancer may very well turn this gene on in the liver. Silencers actively repress gene expression in the wrong tissues, so this silencer, for example, might keep this gene off in the kidneys to ensure that it's only on where it's needed in the liver. The third major regulatory DNA, this insulator, is responsible for ensuring that the regulatory DNAs that regulate the gene on the right do not inappropriately regulate the gene on the left.

Now each of these three major regulatory DNAs is about three hundred base pairs to one thousand nucleotides in length, and contains clustered binding sites for multiple regulatory factors, and I'm going to come back to that point in just a moment. These three regulatory DNAs are already seen in invertebrate genomes such as *Drosophila*, but it's possible that vertebrates have done something creative, little innovation, in the DNA, for example, down here you see some very complex regulatory DNAs called the LCR (locus control region) and the GCR (global control region). These coordinate the expression of linked genes within complex loci. So, for example, the LCR leads to progressively older expression of these linked genes in developing mouse and human fetuses, and so far something like an LCR or a GCR has been identified, for example, in the *Drosophila* genome. So this could be a unique vertebrate innovation.

But the enhancer is the most prominently used regulatory DNA for determining where and when genes are on and off. And it may very well be that the enhancer is the single most critical determinant of organismal complexity. So for example, the nematode worm, *C. elegans*, contains twenty thousand protein-coding genes, significantly more than the 14 thousand genes seen in the fruit fly, *Drosophila*, yet *Drosophila* . . . This is not, you know, organism chauvinism, but it's a fact that flies, despite having fewer genes, exhibit a broader range of cell types, morphologies, and behavior than the worm. Now it's hard to get a hard number on this, but it looks like the typical worm gene may be regulated by one or two enhancers, while the average fly gene is regulated more like by three or four enhancers. So in other words, the nematode worm may be built up from twenty to thirty thousand different total patterns of gene expression whereas the fruit fly is built up from forty to fifty thousand different patterns of gene expression, significantly more than the worm, even though it has fewer genes. So the enhancer may be the measure of organismal complexity, and I'll come back to that.

Genes Regulated by Multiple Enhancers

This is an example of a gene that's regulated by multiple enhancers. The gene is called "*even-skipped*," or "*eve.*" It's a segmentation gene in *Drosophila*, which is expressed in a series of seven stripes in the early embryo. And this gene is regulated by five separate enhancers, two located upstream of the gene and three located downstream of the gene. And each enhancer generates a subset of the total pattern, just one or two stripes; together the five enhancers generate the complete composite pattern, seven stripes. Each enhancer is five hundred base pairs in length and contains cluster binding sites for multiple regulatory proteins. So for example, the best characterized of these enhancers, the *eve* stripe 2 enhancer, which is 480 base pairs in length, contains 12 binding sites for 4 different regulatory proteins, 2 activators and 2 repressors.

Each individual site does not have that much information—each individual site is maybe six to eight nucleotides. These are degenerative sequences. But together, looking at the clustering of multiple sites, gives you more information, and in fact that clustering can be used to identify new enhancers using the computer. And one example is shown on the next slide. So—that's rather dark, but I think you can see it—this is a gene locus called "*ftz*," another segmentation gene. Like *even-skipped*, it's expressed in a series of seven stripes in the embryo. So *ftz* works along with *eve* to subdivide the embryo into a repeating series of body segments.

Vince Calhoun, a graduate student in the lab, simply used the computer to look for clustering of binding sites for regulatory proteins that we know are present in the early embryo and are important for segmentation. And he found there such clusters. These two correspond to known enhancers that were identified in classical studies done nearly twenty years ago in Walter Gehring's lab. The third cluster, however, corresponds to a new enhancer that was missed in the earlier studies, and what Vince did here was to simply take a genomic DNA fragment that encompasses these clusters of binding sites, attached it to a *lacZ* reporter gene, and put it into a transgenic fly embryo. And you see that the cluster of sites generates two of the seven *ftz* stripes, stripes number one and number five. So it is possible to use the computer to identify enhancers, new enhancers, but this is really a daunting problem when you think about the human genome which is likely to contain at least 100 thousand different enhancers scattered over one billion nucleotides of DNA, and with our current technology most—the vast majority of these enhancers are invisible, you can not identify them by simply reading off DNA sequence.

Regulatory Gradients In Development

So a major goal of these kinds of bioinformatic studies is to try to decode the regulatory DNA, establish a direct connection between primary DNA sequence and predicted patterns of gene activity. Will we ever be able to read off the human genome and infer that a given gene, for example, will get turned on in the prostate of a 50-year-old man based on the flanking regulatory sequences? And I'm just going to give you a progress report in trying to decode the regulatory DNA in the simpler genome, in *Drosophila*.

Most of our studies along these lines—I can't see this, can you guys see that? No. Can I have lights down at least on the stage, and I would just as soon be dark and invisible up here. Is it possible to get the lights down? I'm at the limits of my technological manipulations here with this pointer and moving back to the podium. What you will soon see is really very beautiful. What will be revealed to you is a blastula-staged embryo stained with an antibody against a protein called Dorsal. This is a regulatory protein; it binds to specific DNA sequences; and it specifically regulates the transcription of particular target genes.

Now we may be able to get by with—I've got more pictures of this, so maybe it'll come up later. You can see it now? Eric Kandel can see it, but I can't, so what does that mean?

So this is a side view of a blastula-stage *Drosophila* embryo, stained with this antibody against the Dorsal protein. It was—it's kind of an inverse staining method, so at the bottom you see black holes—this is where the Dorsal protein has entered nuclei—and on top you see bright spheres—these are nuclei that lack the Dorsal protein. So the Dorsal protein is distributed in a broad nuclear gradient from the future belly to the back, from the ventral surface to the Dorsal surface. This Dorsal nuclear gradient controls the differentiation of several embryonic tissues by regulating a number of target genes in a concentration-dependent fashion. And this is a generally pervasive process using gradients to control different cell identities. The Dorsal gradient is formed by the differential activation of a cell surface receptor called Toll. So Toll is distributed throughout the surface of the early fly embryo, and it is selectively activated mainly in bottom regions, in ventral regions, by a localized ligand called Spätzle, an extra-cellular signaling molecule. And this summarizes studies from a variety of labs, including those of Nüsslein-Volhard, Roth, Stein and Anderson.

But the point here is that the selective activation of the Toll receptor by the extra-cellular Spätzle gradient leads to a corresponding regulatory gradient of Dorsal within the embryo. And this is seen over and over again. Extra-cellular gradients of signaling molecules are transduced into regulatory gradients of one or more transcription factors within an organ, a tissue, and in this case within the entire embryo.

Dorsal Gradient Regulates Gene Expression

I'm going to talk about how the Dorsal gradient regulates gene expression and controls complex processes like gastrulation, but of course we're only dealing with the fruit fly, so to try to engage your interest a little bit. I will point out that a very similar kind of process is used for far more glorious processes, such as the patterning of the vertebrate neural tubes. So now we're looking at the crown and summit of the organ structures in the animal kingdom.

And this summarizes studies from primarily Tom Jessell's lab here at Columbia, and what Tom has shown is that a specialized structure at the bottom of the developing vertebrate neural tube, the floorplate, is a localized source for an extra-cellular signaling molecule called Sonic hedgehog, and Sonic hedgehog is secreted from these cells and forms a concentration gradient. Cell identity, the formation of different neuronal cell types within the neural tube, depends on the amount of Sonic hedgehog they receive, which in turn depends on their proximity to the floorplate. So cells close, located near the floorplate, receive the highest levels of Sonic hedgehog, and become one particular neuronal cell type, whereas cells located farthest from the floorplate get the lowest amount of Sonic hedgehog and become a different neuronal cell type. And so this process, at least in principle, is very similar to what I'm going to describe for how the Dorsal gradient specifies different cell types in the gastrulating *Drosophila* embryo.

As in the case that I showed you with the localized activation, the Toll receptor, forming a Dorsal nuclear gradient, here again this extra-cellular Sonic-hedgehog gradient almost certainly leads to a regulatory gradient of one or more regulatory proteins, including a transcription factor called Gli.

What you would see here to the left—and now you can use your imagination and perhaps even produce a more beautiful gradient in your mind's eye than what we actually generate with antibodies—is you'll see, you would see, a gradient of the Dorsal protein, high magnification, ah, it is beautiful if you could see peak levels of the protein here in ventral regions, lower levels in lateral and more Dorsal regions. And what this Dorsal gradient does is to regulate a variety of target genes in a concentration-dependent fashion. I think you could at least see in outline the expression patterns of some of these target genes.

The *snail* gene, for example, is activated by the highest levels of the Dorsal gradient, so *snail* expression is restricted to the bottom, the ventral regions of the embryo, where it's important for the specification of the embryonic mesoderm. Low levels of the Dorsal gradient, believe you me, right here in the middle part of this panel, activate a different target gene called *sog*, which is expressed in a very broad lateral stripe that helps define the future neurogenic ectoderm, and *sog* is kept off in these bottom regions where there's high levels of Dorsal by the Snail protein which works as a repressor. And then finally the Dorsal gradient not only turns on genes like *snail* and *sog*, but it also represses genes such as *dpp*, which in principle could be activated throughout the embryo, but *dpp* is kept off in ventral and lateral regions by both high and low levels of the Dorsal gradient. So this differential regulation of Dorsal target genes leads to the formation of three basic embryonic tissues: the mesoderm, which will form the muscles and internal organs; the neurogenic ectoderm, which will form the ventral nerve cord, kind of the poor man's spinal cord of the fly; and finally the Dorsal ectoderm.

Gradients Controlling Gastrulation

Now a postdoc in the lab, Angela Stathopoulos conducted a genome-wide microarray screen to try to identify every single Dorsal-target gene, to get a comprehensive view of what does it take for this morphogen gradient to control embryonic patterning in general, and gastrulation in particular. And she identified a total of fifty genes which show these types of localized patterns of expression across the dorsal-ventral axis of early embryos, and we estimate that at least thirty of these fifty genes are directly regulated by the Dorsal transcription factor. That means at least thirty genes have enhancers with binding sites for the Dorsal protein which mediate the regulation by this gradient that you can't see. About half of these Dorsal target genes control the process of gastrulation, which is summarized here.

So the circles represent cross-sections through early embryos, and here you see the Dorsal nuclear gradient for the first time in diagrammatic form. Low levels of the gradient lead to the localized expression of two key signaling molecules, *dpp*, or TGF-β, on the top of the embryo, in yellow here, and a couple of newly identified FGF genes in the future neurogenic ectoderm. High levels of the Dorsal gradient turn on the FGF receptors, and these invaginate the cells that express the FGF receptors, invaginate into the blastocele of this gastrulating embryo, and now receive the FGF signal from the neurogenic ectoderm, and this causes the activation of the receptor and changes in cell shape, which in turn result in the spreading of the internal mesoderm. So the mesoderm forms a monolayer on the inner wall of the neurogenic ectoderm, and the mesoderm that reaches the Dorsal-most regions now comes into contact with the Dorsal ectoderm, which expresses this TGF-β signaling molecule. TGF-β induces the internal mesoderm to form heart lineages.

So it is possible to see gastrulation as a discrete series of threshold readouts of the Dorsal gradient. For example, the exact limits of this inductive interaction between the Dorsal ectoderm and the internal mesoderm is really set by readouts of *dpp* expression and *sog* expression, two direct targets of the Dorsal gradient. *Sog* encodes an inhibitory protein, which blocks *dpp* signaling and restricts *dpp* signaling to the dorsal ectoderm. So I think it is possible, in fact, to read out the genome, understand the genome, in terms of revealing a complex cellular-based process such as gastrulation. So Dorsal gradient thresholds control gastrulation.

And now I want to talk about some of the bioinformatics methods we've used to try to understand how the Dorsal gradient generates different patterns of gene expression, how does it regulate gene expression in a concentration-dependent fashion? And I don't know; I'm afraid to press the button and look at the next slide.

That ain't so great. Oh well. Once again we're going to have to use our imaginations, but you can—you can—okay.

Characterizing Dorsal Enhancers

Now, this is supposed to show you a series of embryos with an abnormal anterior-to-posterior Dorsal gradient. Never mind how we did that; we've engineered these embryos so instead of having the normal up-and-down gradient, they have a gradient that spans the head-tail axis, with peak levels at the head or anterior tip, lower levels in more posterior regions. And what you see here are the expression of genes that respond to different levels of this anterior-to-posterior Dorsal gradient, with *snail* being activated by the highest levels of the gradient, and then these other genes being activated by progressively lower levels of the gradient, until you wind up with the *sog* gene, which is activated by the very lowest levels of the gradient.

So the Dorsal gradient generates at least five different patterns of gene expression. But we believe that there are three major threshold responses to the Dorsal gradient: high levels turn on *snail*; intermediate levels turn on *sim* and *vnd*; and, low levels turn on *ind* and *sog*. The different patterns within each threshold depends on transcriptional repression. For example, the *sog* border here is set by *snail*, whereas *ind*, which responds to the same level of the Dorsal gradient, it is duly repressed by both the Snail protein and the VND protein to set back this forward border. So you get these different patterns due to transcriptional repression.

So how does the Dorsal gradient regulate gene expression in a concentration-dependent manner, how do high levels turn on genes like this, intermediate levels and low levels regulate these other genes? And to get at this process what Angela Stathopoulos and Michele Markstein did in the lab was to isolate a large number of Dorsal target enhancers, mainly using bioinformatics methods. So I told you that there's at least thirty such enhancers in the *Drosophila* genome, and by now we've isolated and characterized 17 of these 30 enhancers.

Some of the enhancers were identified quite simply in the following way: Angela took the microarray data, so this could be done in any process for which there's microarray information. Angela just took genes that she knew, exhibited localized patterns of expression across the dorsal-ventral axis, and simply asked where are there potential clusters of Dorsal binding sites? And so, for example, one such cluster is associated with the gene called *vnd*, which is activated by intermediate levels of the Dorsal gradient in these lateral stripes. This cluster is found in the ectopic region and when she attaches this genomic DNA fragment to a *lacZ* reporter gene, this fragment recapitulates the normal *vnd* pattern, response to intermediate levels of the Dorsal gradient.

Another cluster was found five-prime of this *mes3* gene, which encodes an insulin-like growth factor, and once again a genomic DNA fragment that spans these potential Dorsal binding sites recapitulates the normal *mes3* expression pattern, namely activation by the highest levels of the Dorsal gradient in the mesoderm.

Finally another—this is just another of many examples—Angela identified a potential cluster of Dorsal binding sites located far of five-prime of the predicted promoter for a gene called *neu4*, which encodes one of the newly-identified FGF signaling molecules. This cluster of sites is about 15 thousand nucleotides, five-prime of the promoter, but nonetheless a genomic DNA fragment that encompasses these fragments gives an authentic *neu4* pattern of expression in transgenic embryos, namely activation by low levels of the Dorsal gradient throughout the neurogenic ectoderm.

Surveying the Genome for Enhancers

Another way that we've isolated new Dorsal target enhancers is simply summarized here, and this is kind of a weird collaboration between my graduate student, Michele Markstein and, of all people, her parents, Peter and Vicky Markstein, who are senior computer programmers at Hewlett-Packard. And, you know, this is no joke, I mean it was a good collaboration even though they all hate each other and it's a fairly dysfunctional family, but it led to the publication of a nice paper—I think a nice paper—by Markstein, Markstein, Markstein, and Levine. And I used to say well, no, it's not a deli and it's not a law firm, it's a scientific article. And then I consulted with a Talmudic scholar here in New York who said that you never have four names associated with a deli, at most it's two, like Sherm and Al, you know, something like that. And so if it's four names, four Jewish last names, it's got to be a law firm, and frankly Markstein, Markstein, Markstein, and Levine sounds like a kind of a low-grade law firm, kind of maybe ambulance chasers. Okay.

But here's what they did in the study. They wrote a computer program called Fly Enhancer, which allows you very quickly to survey the entire *Drosophila* genome for clusters of binding sites for regulatory proteins or any combination of regulatory proteins. And in this first experiment what they simply did was to ask for the regions in the *Drosophila* genome that have high-density clusters of actually optimal high-affinity Dorsal binding sites. And they found 16 such clusters, although one of these corresponds to a previously identified enhancer which was actually used for the analysis. The enhancer in question has four optimal high-affinity Dorsal binding sites; it was the sequences of these sites that were used to survey the genome. So 15 new clusters were found. This means a window of three hundred base pairs or four hundred base pairs containing at least three high-affinity Dorsal binding sites. Fifteen clusters and Michele systematically tested each one; she took genomic DNA fragments for each of the 15, attached these to reporter genes, and examined them in transgenic embryos. And it turns out that 5 of the 15, one-third, correspond to authentic enhancers. And without going into the gruesome details what you see on the right, here are the genomic DNA fragments, encompassing one or another of the Dorsal binding clusters attached to a *lacZ* reporter gene. Two of the enhancers mediate activation by the highest levels of the Dorsal gradient in the ventral mesoderm, one of the enhancers responds to intermediate levels of the gradient in ventral regions of the neurogenic ectoderm, and finally two of the enhancers give broad lateral stripes of expression throughout the neurogenic ectoderm in response to the lowest amounts of the Dorsal gradient.

So the hit rate is 5 out of 15, 1 out of 3, which is pretty good, I mean that's useful for isolating enhancers, but if there is a regulatory code, as I discussed earlier, then we should be able to get all thirty or so of the estimated Dorsal target enhancers scattered throughout the *Drosophila* genome without any false positives and without any false negatives. Here we got 5 positive hits and 10 false positives, 10 negative clusters that do not work as enhancers that respond to the Dorsal gradient.

Improving Bioinformatics Analysis

So what is it going to take to improve the precision of this kind of bioinformatics analysis? We should surely get a higher precision if there's anything like a regulatory code that links primary DNA sequence to predicted pattern of gene activity. And so what Michele and a postdoc in the lab, Albert Erives, have done is to develop computer programs to look for shared features among coordinately regulated enhancers, and this very much improves the precision of whole genome searches for new enhancers.

And this just shows one example of their analysis. So here you see three different enhancers that range in size from three hundred base pairs to five hundred base pairs, which are activated by intermediate levels of the Dorsal gradient in lateral stripes within ventral regions of the neurogenic ectoderm. These three enhancers do not share any straightforward DNA-sequence homology. You would never get them through any kind of cross-hybridization, either experimentally or in the computer. They are associated with unlinked and unrelated genes. And so what—but they nonetheless give similar patterns of expression.

So what Michele and Albert did was to ask do these three coordinately regulated enhancers share common features, common sequence motifs and even perhaps a common organization? And the answer is yes. But initially the problem when you do this, asking for shared sequence features is if you take any two or three DNA fragments of several hundred base pairs in length, the computer gives you an impossible list of potential shared motifs. Bear in mind that a typical transcription factor, as I mentioned earlier, binds to a sequence of just 6 or 8 nucleotides and generally you can get degeneracy in 2 or 3 of those positions, so it's not a very strong signal, an individual binding site. And so what they did was actually kind of clever; they used the ten negative clusters that I talked about earlier—remember I said when you survey the *Drosophila* genome for simply Dorsal binding clusters then you get 15 hits, 5 are real enhancers, 10 are negative, we know they definitely—for whatever reason—do not work as enhancers. And so they used that as a filter to focus on the important shared motifs and get rid of the background noise.

And this led to the identification, in addition to the Dorsal binding sites, which of course we knew were present in all 3 of these enhancers, 4 different sequence motifs, and we mostly know what proteins are likely to bind to these motifs. But the main point I want to make here is that now when they repeat the Fly Enhancer screen and look for tight clusters of all five of these binding sites, Dorsal and these four newly-identified shared motifs, they find 12 hits in the entire *Drosophila* genome, and at least 6 are real enhancers that give this kind of pattern of expression. And it's possible that in fact 9 of the 12 are real hits. So in other words, they're getting a precision of at least 50 percent positive hits in a whole genome survey when they ask for clustering of multiple regulatory factors, and that's getting us closer to this problem of trying to decode regulatory DNA and link primary sequence to gene expression patterns.

Constraints on Enhancers

Albert went on to obtain evidence that enhancers are not simply loose collections of binding sites for regulatory factors, but they may contain some constrained organizational features, and the more constraints, for example, on the spacing and orientation of these binding sites the better the prospects are for elucidating a regulatory code. And Albert found evidence that coordinately regulated enhancers share common organizational features by looking at the mosquito genome. So *Anopheles gambiae* is the carrier of malaria, and so it was sequenced and assembled, it's a fairly high quality product, but it's sort of an untamed and unknown genome with respect to regulatory DNA.

What Albert did was to take these five shared sequence motifs, found in the type-two Dorsal target enhancers that are regulated by intermediate levels of the Dorsal gradient, and scan the entire *Anopheles* genome, and he got some hits. And a couple of these hits are associated with genes that are orthologous to the known Dorsal target genes in *Drosophila*. One of them is shown here, the *Anopheles gambiae vnd* gene.

Now I should mention that *Anopheles* is a dipteran, like *Drosophila*, but they last shared a common ancestor 250 million years ago, and unlike the situation in vertebrates, these guys don't share any sequence homology outside the protein-coding region. Forget about non-coding homology; you could barely identify orthologous genes in a one-to-one relationship between flies and mosquitoes—outside the coding sequences there is nothing.

So Albert identified a cluster of these linked sites in the *Anopheles* genome which maps within the intron of the *Anopheles vnd* gene, similar position to where the enhancer maps in the *Drosophila* gene. He would've never found this enhancer by any kind of straightforward sequence homology, so this is a very nice situation. You have two orthologous enhancers that are completely devoid of any sequence homology or extraneous sequence identity. And so what you're left with is the bare-bones skeleton key—what organizational features do these two enhancers share in common? And in fact, in addition to containing binding sites for all five of these regulatory proteins, there is a two hundred base-pair domain within the enhancer which contains a very similar organization of the binding sites that's drawn out here. These binding sites are on the same side of the double helix and show similar spacing. And so it's possible that enhancers do contain organizational constraints; it looks like this basic organization was probably present in the ancestral dipteran enhancer for the *vnd* gene, and this organization has been maintained over 250 million years of evolution. And again the point here is that the more such organizational constraints the better for elucidating a regulatory code.

I should mention that the idea of enhancers containing constraints is not new. For example, Dimitrios Thanos, here at Columbia University has shown that the enhancer for the human Interferon-β gene has highly organized binding sites for different regulatory factors so stringently organized that this is called an enhancesome. But the enhancesome seems to be rare—it applies to a couple of genes. Typical enhancers do not show that kind of global organization, although I would argue that typical enhancers may contain some subtle organizational constraints, particularly close linkage of two different activator sites which must work synergistically in order to get gene expression.

And then the next slide of the series simply shows that the enhancers, the putative clusters—or the putative enhancers of the clusters of binding sites found in the *Anopheles* genome in fact worked when introduced into *Drosophila*. So for example, this is the DNA fragment I talked about in the first intron of the *Anopheles vnd* gene. When attached to a *lacZ* reporter gene, it gives a crude lateral stripe of gene expression in response to intermediate levels of the Dorsal gradient, just like the orthologous enhancer in *Drosophila*. So it's been possible to use the computer to use predictions linking sequence to gene activity, to actually go into this fairly uncharted genome and identify a new enhancer which gives a predicted pattern of expression.

Now I realize you can't see half my slides, I realize that most of you don't know what the hell I'm talking about—I'm not sure I do—and that I'm the only person standing between you and lunch. How much time do I have, chairman? He'll say, "Negative five minutes." I got ten minutes. Okay, great. Nine and a half minutes, and he means business. Okay.

Sea Squirts as a Simple Chordate System

So in these last minutes I'm going to talk about sea squirts. And I will simply try to connect what I've told you before to what I'm going to finish up with now, by saying that this idea of using clustering and organizational constraints within coordinately regulated enhancers is not a peculiarity of the *Drosophila* genome, but it's also seen in a chordate genome, admittedly a very primitive chordate, *Ciona intestinalis*, the sea squirt.

So this is an adult sea squirt—if you can see it. It's a simple tube with two siphons that basically suck in and spit out the sea water. And they live in shallow ocean water. Their extreme simplicity led Aristotle to classify them as mollusks, and they were stuck with that lowly designation until 130 years ago when the great Russian embryologist Kovalevsky noticed that despite the simplicity of the adult, the embryos and larvae look a lot like vertebrate tadpoles. So here you see a ten- to twelve-hour *Ciona* embryo, and you could see the basic organization of chordate features. It looks like a very primitive, stripped-down frog tadpole. We have a head and a tail; the head has a cerebral vesicle, the tail has a prominent notochord and a dorsal hollow neural tube. So there's no doubt that's a chordate, despite the primitive and hideous appearance of the adult.

Now this system has three really nice features for doing whole-genome analysis on gene regulation. First of all, it's constructed from very simple lineages, similar to what Sydney originally showed for *C. elegans*. This simple tadpole is composed of just a thousand cells, and these cells can all be traced right back to the fertilized egg. The notochord, which is composed of thousands of cells in vertebrates, is composed of only forty cells in *Ciona*. A second nice feature of the system is that it has a small and compact genome; at 150 megabases it's the same size as the *Drosophila* genome. In the *Ciona* genome, both the *intestinalis* and *savignyi*—related *savignyi* genomes—have been sequenced and assembled, they contain about 15 to 16thousand genes, similar in size and density to what's observed for *Drosophila*, but it is a chordate. And then the third feature of the system which is nice—and this is my favorite—is that it's very easy to identify tissue-specific enhancers because you can introduce transgenic DNA into developing *Ciona* embryos using electroporation methods. You don't to use tedious methods of injection and then wait for the animal to grow up and get propagation through the germ line, as is done in mice, and this takes many months. Here what you do is, for example, take a *cns* enhancer, a genomic DNA fragment from the *Ciona* genome that directs *cns* expression, attach it to a *lacZ* reporter gene, take this fusion gene, mix it with fertilized eggs, zap it in the electroporator, you know, go out and get a meal, ten, twelve hours later this is the pattern you get.

So this is the simplest system I know of for rapidly identifying tissue-specific enhancers. And I should tell you that this problem in trying to link primary sequence to patterns of gene activity, trying to identify the hundred thousand or so enhancers present in the human genome, one reason why we don't know these enhancers when we see them is because so few have been identified. When you get a new protein coding region you've got a 50/50 chance of being able to infer the function of that coding region because of all the vast amounts of structure function studies that have been done with different proteins over many years. In the case of regulatory DNA there is not so; probably only something like a hundred enhancers altogether have been identified in all animals combined, at least in terms of looking at them in transgenic developing embryos, a very small data set, so we don't know the rules for what it takes to get gene expression in the nervous system or in the notochord. Okay, so this system allows us to get a bunch of enhancers.

Tissue-Specific Enhancers in Sea Squirts

And the next slide—I really am almost done—I can see you fidgeting, I'm getting there—what Dave Keys and Naoe Harafuji did was very simple. They just took random two-kb pieces of the *Ciona* genome and asked how many of these give tissue-specific patterns of expression. And here are some examples. Here's a random piece of genomic DNA that mediates expression in the notochord and in the neural tube. Here's another one that mediates expression in the sensory palps at the front end of the tadpole. Here's one that mediates expression in the hindbrain region, and so forth. Altogether they found one enhancer roughly every ten or so kb. So there's something like ten thousand to twenty thousand different tissue-specific enhancers in this fairly primitive creature. So once again it's easy to imagine that humans have a hundred thousand enhancers or more.

And so now just to finish up, what Albert Erives has done is to classify these various enhancers that have been identified, which were generating very, very rapidly, put them into different groups, those that are coordinately regulated in a particular tissue, say the tail muscles, and ask do they share common features. And the answer is yes, they share common sequence motifs, binding sites for regulatory proteins, and more often than not these binding sites do show some constraints in their organization.

Here are a bunch of notochord-specific enhancers, and this analysis is just really picking up steam right now. Here are seven different notochord-specific enhancers, and many of these share a number of sequence features, and Albert's looking now to see if these regulatory binding sites are also organized in some constrained way, as seen for the muscle enhancers and for the enhancers I described in *Drosophila*.

So if I could finally have the lights on I will just read off my summary of what I tried to tell you today. I first talked about how the Dorsal gradient controls gastrulation in the early *Drosophila* embryo, then I discussed the use of the computer for trying to decode the regulatory DNA that's controlled by the Dorsal gradient, and made the point that you can find enhancers using the computer, new enhancers, that the precision goes up as you put more and more binding motifs into the algorithm, and that a number of the binding sites show some constrained organization. So enhancers may be organized, and I think this improves the prospects for elucidating regulatory codes. And then finally I ended with my belovéd sea squirt showing you that it's possible to very rapidly identify tissue-specific enhancers which will be useful as a starting point for decoding regulatory DNAs, underlining gene expression in chordate tissues and vertebrate tissues. With that I'll stop. Thank you.

Isidore S. Edelman: I discovered that the surest way to move an audience is to announce lunch, that does it. We will reassemble at 1:30 and Andrew Marks will moderate that—this afternoon's symposium.

Genes, Genomes and Medicine
October 16, 2003

Eric S. Lander, Ph.D., Whitehead Institute Center for Genome Research
Massachusetts Institute of Technology, Cambridge, MA
Beyond the Human Genome Project: Biology as Information

Introduction by Andrew R. Marks

Andrew Marks: Okay, welcome. Happy birthday, Columbia. My name is Andy Marks, I'm the chair of the Physiology Department here at Columbia, and I have the pleasure of introducing our speakers for this afternoon.

Our first speaker is Dr. Eric Lander. He is currently a member of the Whitehead Institute, and founder and director of the Whitehead Institute Center for Genome Research, and he is one of the world's leading genome investigators, having been a principle leader of the Human Genome Project.

He is also a professor of biology at Massachusetts Institute of Technology. He received his undergraduate degree at Princeton, and a Ph.D. in Mathematics from Oxford University, and has the distinction of having been not only an outstanding biologist, but a professor at Harvard Business School from 1981 to 1990.

He is going to talk to us today about the utility of taking comprehensive views of DNA, RNA, and protein across many tissues, individuals, and species, and how this is driving a revolution in biology and medicine. So we are delighted to have Dr. Lander here.

The Last Fifty Years

Eric S. Lander: That's working. Thank you very much, and the relevant part of the biography that was omitted is that I'm actually a kid from Brooklyn, grew up in Brooklyn, New York, and went to high school in Manhattan, at Stuyvesant. And in particular in my connection to Columbia in wishing you a very happy quarter of a millennium, I was one of the students, one of the many students in the city of New York, who participated in the Columbia Science Honors Program, which was really formative. I, for me in high school, being able to take the train uptown to Columbia, and sit in my first-ever university-level math course in Galois theory about two buildings down, a class in immunology up at the medical school, and a class in astronomy, was really formative. And so I think for many many kids who grew up in New York who don't necessarily go to Columbia as their undergraduate degree, they still very much benefit from Columbia. So, in saying happy birthday, let me also say a thank you to Columbia.

So, my title today, "Biology as Information." Biology can be thought of in lots of different ways, and over the course of, you know, hundreds of years, thousands of years, people primarily have thought about biology as organisms, the study of biology as the study of organisms. Indeed, for a very long time, what else could you think it to be? Well, somewhere in the end of the 1800s, an important shift began to occur, with the recognition that we could study biology not just in terms of organisms, but in terms of molecules. The birth of biochemistry, the understanding that you could purify molecules such as enzymes from the cell that would subserve vital functions led to the idea that one could break up biology into these pieces, and that by understanding the structure of the pieces, you could gain insight into function.

By the middle of the twentieth century, the most beautiful example of that biochemical understanding, the purification of that rather unlikely function, heredity, in a single molecule, DNA and perhaps the most beautiful of all structure-function relationships, the double helix, to the explanation of the transmission of information, occurred. And in those fifty years since the double helix, biology has blossomed out, at least this view of biology has blossomed out, into a view of biology as elaborate molecular machines, and, at least in some minds, an understanding of all of this, someday, in terms of these molecular machines.

But around mid-century with, with the DNA double helix also came an understanding of biology in a third way. They're all complementary of course, but biology as pure information. Just as the molecules are purified away from the organism, information can be purified away from the molecules at some level. That DNA sequences, in fact, can be read out, that other types of readouts of the cell can be had, and you can begin to think about biology in this complementary view as information.

The Next Fifty Years

By mid-century, by fifty years later, or so, we have things like complete sequences of the human genome, etcetera. Where does it go in the next fifty years, what is the picture that goes there for biology as information? I think it's this: I think it's biology as extraordinary library, vast and beautiful library of information. This is actually just downtown, it's the Rose Reading Room in the New York Public Library, I don't have a shot of Low Library or something, but this, this is pretty pretty, to me this library. And I imagine on its shelves here volumes corresponding to each organism, sequence of the organism, every species. Also, every individual within each species, every tissue within the individual. And not just genome sequence, but levels of expressions of RNAs and proteins and modifications. And all of that, in principle, can be had. We currently only have a few of the volumes up there, but we can gather all of that, and increasingly the question becomes how do we use that kind of information, how do we make sense out of that information?

And when you think about it, it's pretty powerful. It represents laboratory notebooks. It represents evolution's laboratory notebooks. For three and a half billion years, evolution's been running experiments. It gets up each morning, changes a few nucleotides, sees how things work. If it likes the results, it keeps the notes, if it doesn't like the results, it discards the notes, which we now regard as, as incorrect procedure, but you've got to cut evolution a little slack in this regard, because it started before that was codified. So we at least we have the lab notebooks for all the successful experiments that evolution has run. Well how much can we learn by reading someone else's lab notebooks? We didn't design the experiments, that's the problem. Usually, we like to design a specific manipulation ourselves, and then it's easy to interpret, at least we think so. Here, we have lab notebooks where we didn't design the experiments, evolution did, but what makes up for it is they're massive, they're voluminous. Evolution is much more patient than we are. Evolution is better funded than we are. It undertakes experiments on the scale of millions and billions of years, and it's very patient, it's willing to mutagenize every nucleotide in the genome. And so, we would be wise to incorporate this view into our own experimental view. Now of course, anything you learn you're eventually going to have to go test back at the bench, but how can you not look at that kind of experimental lab notebook data?

Well the way to do it is what we've got to work out. How do you, how do you read this stuff? Well it's got to be by comparison. You've got to be able to compare things, and, and it could be comparisons within a genome, it could be compared—with, within the human genome—it could be comparisons between species and their genomes, it could be comparisons between individuals, different variations, it could be comparisons between cell states in their expression readouts of RNAs and proteins. And in all of this, we need tools to take those multiple views and integrate them together to figure out how to, how to extract from this picture of data insights, not just, as Sydney sometimes calls, clutter.

So I'm going to touch on a number of different subjects, but I'll try to pick a few examples that I hope will illustrate the potential power of information. The theme throughout the talk will be, just how much can we squeeze out of this information, and I think the answer is quite a lot.

Comparing Mouse and Human Genomes

Well first, a foundation for this. The Human Genome Project, about which I'm going to say almost nothing, other than, it existed, it happened, it was a good thing. It was biology's first attempt to collect a large amount of information. It was about a 15-year project involving many many different people around the world, resulted in draft sequences of the human genome being published in February of 2001, and happily, in April of this year, a finished sequence of the human genome going up on the Web. Finished is a term of art; there's still about 1 percent of the human genome we can't clone and sequence by any available techniques. They will take loving, those regions of less than four hundred gaps will require tender loving care to sort out, but the rest of it is in finished form. And we can draw pictures here of chromosomes, and infer genes and as has already been referred to, there seem to be fewer genes than we thought. Not the hundred thousand that I teach my, that I taught my students at MIT for a decade, more like thirty thousand, and thirty thousand increasingly looks like an overestimate, I think we're clearly into the twenty thousands, which is good. We're doing our part to simplify biology in that sense of decreasing the gene count, but it's still going to be quite complicated.

So enough about the Human Genome Project. To really make sense out of this, you have to start running comparisons, you have to look at the experiments evolution has done and, and draw inferences from that comparison. Well the, the, the most logical one is the mouse and the human, the mouse being the leading experimental system for biomedicine, the human being us, and so what you'd like is the sequence of the mouse genome. And happily, we have the sequence of the mouse genome. Less than a year ago, a highly advanced sequence of the mouse genome was produced and published, with something like 96 percent of the mouse genome in it, again there are holes to fill in, but it's mostly there, and you can start looking at it.

So you can lay out the mouse genome, all along its twenty chromosomes, and you can start taking any region of the mouse genome, and start looking for where it matches up in the human genome. And if I take a region of say, oh I don't know, the human genome here, and I take a point and ask where's its best match any where in the mouse, anywhere in the mouse it's there, and this spot here, anywhere in the entire mouse it's there, and this there, and this there. And the fact that all of these points have their best match in the mouse genome in the same region in the same order is of course no accident, it's a reflection of the fact that this region, in fact, in both mouse and human descends from a common ancestor some 75, 80 million years ago, and while there have been various changes that have occurred, we can still pick out very clearly from its sequence that this is the same region. Now of course this perfect lockstep correspondence doesn't continue forever, there have been breaks and reunions, and so one can build a map that relates any part of the mouse genome to a corresponding part of the human genome. This bit here on mouse chromosome number 18 I guess corresponds to human at 18. This bit here on mouse 15 to human chromosome number 8, and you have an across lookup table with three hundred or so blocks between mouse and human. Now when we look up close at those blocks, and I'm going to attempt, because I know it's a half technical audience and half nontechnical audience, to see if we can hit both levels, when you look at a string of information, like human sequence and mouse sequence, obviously sequence is really ATCG. But conceptually a string of information here, what you'd like to know is, what hidden messages are there? So can you see the hidden message here? There it is. This is hidden. Computers are very good at running along sequences and picking out what is conserved between them, what parts match up. And so you could do that to the genome. You could take regions of the mouse and human genome, and begin to look for these *this is hidden*'s by which I mean stretches of DNA that are more heavily conserved, more highly conserved than the background rate of conservation due to random evolutionary drift.

And when you do that for one of my favorite genes, PPAR-γ, you find that there are a bunch of spots here that show a high degree of conservation that correspond to known axons, protein-coding regions of genes. But there's about an equal number of spots along here that also, about at least a hundred base pairs long, that show high degrees of conservation, and they are not axons, they do not encode protein. And do you know what they are? Me neither, that's the very interesting thing, is we don't know what they are. I think one of the most interesting surprises about the sequencing of the human genome was this something like very roughly half a million conserved elements in these genomes, and we can't even ex . . . It says fifty-fifty encodes proteins, it's probably less, 40 percent of these things encode proteins, the other 60 percent we don't know. They relate very much to what Mike Levine was talking about. I suspect that lots of them are these regulatory sequences. They may also encode RNA genes, they may also be structural elements in chromosomes, but what's really clear is this was a wake-up call to those of us who did genomes, that in fact the vast majority of what evolution cares about, we don't yet know how to read, and we didn't even know it was there, and we would have estimated that the coding regions constituted the majority of what evolution was lovingly conserving, whereas in fact it appears to be a minority of it. So this is then a very important challenge to biologists, is make sense out of these quarter of a million, 300 thousand other things in the genome. And Mike Levine has already referred very elegantly to studies in *Drosophila* and to *Ciona* about doing this and I too in fact will in fact refer to that.

Connecting Sequences to Functions

Now the way to do this is both experimental, as has already been described, and evolutionary. The comparison of mouse and human pointed us to regions of likely conservation, but it didn't pinpoint them very exactly. It's a very fuzzy sort of a mapping of what regions might matter. If we also had the chimpanzee and the dog and the kangaroo and the cow and the this and the that, we'd be able to refine that picture much more exactly. And happily, we do increasingly have those things. The chimpanzee genome is largely sequenced, it's up on the Web. The dog genome, there's a light coverage of the dog genome that was just published, and the NIH is engaged in the sequencing of the dog genome, and we should have a very dense coverage of the dog genome, we should have a dog by Christmas basically. The possum is coming, the cow is ambling along, and we will have a series of genomes, then, to be able to do this evolutionary conservation and try to refine very exactly what these elements are. So it's going to take a little while to have them, so in the meanwhile we want to practice. How are we going to attach function to all of these conserved elements?

Well, the best place to practice is on a really good model organism. As I say, Mike has already talked about even about *Drosophila* and *Ciona* model organisms. I'll talk about an even simpler model organism, where we have tremendous power to try to extract this evolutionary information: the yeast *Saccharomyces cerevisiae*. Well the yeast *Saccharomyces cerevisiae* was sequenced long time ago now, in 1996 it was published, and it's the backbone for eukaryotic biology. And when it was published, it was reported it had 6,200 genes, and all sorts of other things. Well, we wanted to see how much more you could learn with more yeasts. So we sequenced *Saccharomyces paradoxu*s, *mikatae* and *bayanus*, related species that we think differ by up to about twenty, thirty million years or so. We sequenced them, we assembled their genomes, we lined them up, and just like I showed you with mouse and human, they line up pretty nicely. Across very large stretches you see essentially the same genes in essentially the same order. So, the first thing we did was just said, "Well what about the genes?" We knew, that when the genome of *Saccharomyces cerevisiae* was first published, they declared about 6,200 genes, but they knew that the computer programs . . . well you had to make a choice one way or the other. They declared anything that was an open-reading frame capable of encoding a hundred amino acids to be a gene. But they knew that by chance such things could arise and would arise in the yeast genome, so they either had to over-call or under-call, so what do they do?

Well now with the multiple species we can line it up and ask about whether all of these reading frames appear to be conserved, and to make a long story short (I won't go into the methodology), what you can do is look for lots of frame-shifting insertions / deletions. And it's an absolutely whistle-clean distinction. Real genes, you almost never see them, these other genes, the distributions essentially don't overlap, and you're able to go back and re-annotate the yeast genome. And it turns out about 528 of the alleged genes in the yeast genome are not genes, they were just spurious open-reading frames, and you can get rid of them. And we have good evidence that in fact these are not genes. You find about 43 novel genes that were below the limit of detection before. In 34 cases you merge genes that people thought were distinct, and it turns out the stop codon that separated them was a sequencing error or a mutation. And you can pick that up, and in lots of cases you can find that the first start codon and the stop codon, in fact, that had been annotated are not correct, and evolution can just tell you these things pretty effectively. There's lots to say to defend all these statements, but basically I think the yeast genome now comes down to about 5,695 genes, plus or minus a dozen or two. There are about 20, 25 where we really can't say for sure from this analysis, but there's no doubt that evolution's lab notebooks contain enough information to be able to sort these questions out, but a few more species would even nail those.

Comparisons to Characterize Non-Coding Regions

But we can go further. We can take these conservation bits, the sequences that are well-preserved, that are not corresponding to coding regions, not corresponding to the protein-coding genes, and we can ask about them. Well, between two known genes where we know a lot, we find that known binding sites for transcription factors, like the GAL-4 transcription factor, are well conserved. The GAL-4 transcription factor binds to these sites, it's been worked out experimentally, and sure enough, those sites are well conserved. And when we look at what else is well conserved, it corresponds in this little region to known factors, so it's a pretty good match.

Well, could we use that to find all of the regulatory apparatus of yeast? Well, who knows about *all*, at least a lot. The idea is this: the GAL-4 binding site, the binding site for this protein, is actually pretty crummy. As Mike's already alluded to, these things often bind at six-base pair, seven-base pair sort of things. GAL-4 binds at CGG, 11 bases of space, CCG. Well that's such a small signature that you're going to find it many times at random across the *Saccharomyces cerevisiae* genome. But what if I—and sometimes it's a gene, sometimes it's an intergenic region—but what if I look at all four species together, and I look for those occurrences of the sequence which are in all four species, many fewer. And the ones that occur in all four species are greatly enriched for occurring in intergenic regions. They're four times more likely to be in intergenic regions that genic regions, which is more surprising than you might think, because genic regions are better conserved on average. A random pattern is three times more likely to be conserved in a genic region than an intergenic region, but GAL-4 shows the opposite pattern. So GAL-4's 12 times more likely than a random control pattern to be conserved in an intergenic region. Aha, that's a signature that you could apply to another pattern and with a computer to every pattern. So we applied it to every pattern, you take all the patterns of three bases, a bunch of spaces, three bases, most of them, sure enough, show better conservation in coding regions than in these intergenic regions, but some show much better conservation in intergenic regions and our friend GAL-4 is amongst them but by no means unique in that.

Well, to make a long story short, you can take all of those patterns that show enriched conservation by number of different tests, extend them, merge them, collapse them, I won't inflict on you the work that gets done, and they come down to about 72 regulatory motifs that very clearly stand out from the genome. That list of 72 regulatory motifs that are found automatically by this analysis include the vast majority of all previously characterized yeast regulatory motifs. Very few failed to be on this list. So if you didn't know them already, you would have discovered them. In addition, you'd discover a bunch of other new ones that weren't previously known, and we, you know, we didn't know what they are. Now could we go further than just saying "Aha, these are 72 regulatory motifs that stand out by evolution." Could we figure out what they do, could we attach function, meaning to those things? Well, maybe.

Here's a way to do it. Take our friend the GAL-4 motif again. Let's ask, "What genes does it occur in front of?" Well, if we just look at one species, *Saccharomyces cerevisiae*, since the GAL-4 motif is kind of degenerate and it's all over the place, it occurs all over the place, it's hard to know. But suppose we restrict our attention to only those genes where it occurs in front of it in all four species. Then we get an extraordinarily clear overlap with carbohydrate metabolism. Unambiguous, ten to the minus twenty-eighth, is a significance level I am comfortable with. So, you can say, even if you didn't know in advance, simply by knowing categories of genes you would be able to attach this GAL-4 motif to carbohydrate metabolism. Obviously, you'd need to know something about carbohydrate metabolism to have that as one of your categories, but you didn't have to do an experiment relating to that. You could look it up against a the library of experimental information. You could do this with already annotated biological categories, you could do this with gene sets that were determined by chromatin immunoprecipitation experiments, you could do this with clusters of genes that are identified by mass spectrometry as hanging out together, you could do this by clusters of genes that are organized by coexpressions patterns. And in all of these cases you could correlate the set of genes in which you see conserved appearances of these factors with a set of genes that have some set of biological properties, and you can find some extremely strong matches and for all of the known genes, I'm sorry, known regulatory factors, regulatory sequences, the meanings you attach make a great deal of sense, you would be able to determine what these things had done if you didn't know it already, and for many of the new ones you're able to attach them to meaningful categories.

So in fact, there's a lot of information there, if only we can squeeze it out. Now why not do this to the human? I'm going to skip over this next slide here, which just says, for the aficionados, you can also work out combinatorial control between these things by virtue of their co-occurrence, etcetera. So how relevant is this to the human? Well it turns out that this evolutionary tree we've been working with, *Saccharomyces cerevisiae*, *paradoxu*s, *mikatae*, and *bayanus*, matches up in terms of its evolutionary distance to human, lemur, dog, and mouse. Those are roughly the same distances. So it's a very relevant tree. So why can't we do this right away with human, lemur, dog and mouse? The reason is only that there's more noise in the human genome because it's bigger. The tree has the right distances, but because in the yeast genome, genes represent most of the genome, in the human they're a small amount, we need to have more species on this tree in order to be able to squeeze signal from noise. But if we had a dozen or a dozen and half species on this tree, we can extract the same amount of information as we're extracting from these yeasts. And that's no longer an unreasonable prospect. It may, in fact, be that the best way to get at human regulatory things is go with a set of sequence coverage from a large number of mammals filled into such a tree, and we ought to be able to extract this from yeast. So that's why there's so much interest in sequencing more and more species, because it should refine our experiments down to crisp regulatory elements, allow us to begin to correlate those elements with functional categories, etcetera, and then of course very much couple up to the experimental systems in *Drosophila* and in *Ciona*, and that's why it's exciting that a bunch of *Drosophila* are getting sequenced and various *Ciona* are getting sequenced, because of course we have to do this in all of these different experimental systems.

Applying Sequence Information to Diseases

All right, so that's one feat, is the information contained in the genome book, per se. Now I want to turn to the second feat, which is extracting information about cells, because Sydney says we, you know, we should be paying attention to cells, and cell states, and things like that, and I agree. So I will now turn for the second part of the talk, to the same theme, extracting information, but applied now to, to cell states and particularly to disease problems.

So I'll start with a simple problem, just to illustrate. A relatively rare Mendelian disorder called cytochrome oxidase deficiency, this is a relatively rare disorder, it is a disorder that is at particularly high frequency in a human population, the Saguenay-Lac-Saint-Jean region in Quebec. They're French-Canadian isolate here. And in that region of Canada, it occurs at a frequency comparable to cystic fibrosis in the European population. I won't go into the details of the disorder other than to say it's a severe disease and the children who are homozygous for this die by the age of 12.

To make a long story short, using tools of human genetics we, which is Tom Hudson and colleagues, mapped the gene to a particular region of human chromosome II, and the problem was there aren't that many families, so you couldn't map it to a very narrow little region, you were left at the end of all this mapping with many megabases of DNA, and it was clear it lived there, but what gene was it? Well we weren't prepared to resequence the entire region base by base in order to discover which the gene, where the mutation was. And so it sat on the shelf for about a year, year and a half, until a postdoc came along, a postdoc named Vamsi Mootha. And Vamsi said, "If I claim that, in effect, by going to this vast library of biological information, we could find the gene with no more lab work." Now we could afford that, that was a reasonable proposition, so we decided to have it a . . . to give it a whirl. And so what Vamsi did was he said, "Look, I want to take all of the information already known about DNA, RNA, and proteins, and lets intersect it, and I claim there's going to be one or only a few genes that could possibly match the characteristics that a gene has to have for our disease." So DNA, well, in the meanwhile the sequence of the genome had become available, there were gene predictions across there, so you had the DNA, and you said here's likely genes cross with RNA.

Well, what Vamsi's idea was, let's take all the genes we know and figure out which ones are likely to pay a role in the mitochondria, or those with the mitochondria. Well, of course how would we know that? Bootstrap. What he did was he went down the hall to somebody who studied cancer, and he said, "Can I borrow all your gene-expression experiments on cancer? I want to cluster all the genes from all the gene experiments on cancer—on all the gene expression experiments on cancer, and I'm going to look at those clusters that happen to have a lot of already known mitochondrial genes, and I'm going to guess that the other genes that are in those clusters are also likely to be mitochondrial but that haven't been previously identified as such." So he gave every gene in the genome a kind of mitochondrial score for how likely it was to be a . . . how much of a fellow traveler it was in terms of its expression pattern with known mitochondrial genes.

He then did a third thing, which was he himself—he happens to be an expert on mitochondria—had been doing proteomic experiments purifying mitochondria, purifying proteins from mitochondria, digesting them, flying in mass specs, and collecting databases of proteins that appeared in mitochondria. He took the three data sources, DNA, RNA, and protein, and he intersected them. Exactly one candidate gene. In fact, better than that. Any two of those sources intersect in exactly one candidate gene, and the third source confirmed it. At that point, I confess, he broke down and did an experiment: he ordered PCR primers, resequenced the gene, and sure enough, it has a major mutation in Saguenay-Lac-Saint-Jean and it has a secondary mutation. It's the right gene, and it even has a very weak homology now that we see it, to something in yeast that has to do with mitochondria, etcetera, etcetera, and it's quite clear it's the right gene, but, the point was, the information was sitting there in the database, if only we figured out how to extract that.

Now he's been doing this, by the way, for other diseases now, too, where we don't necessarily know the organ or the tissue, so we don't have this trick of matching up with mitochondria. So instead of finding things that have the same expression pattern as mitochondria, he's been matching it up to—let's see if I can get this right—genes that have similar expression patterns to genes that appear in papers about this disease. And it turns out that that works too, that hook, hooking it up to the bibliographic database turns out to be remarkably powerful and I think for several Mendelian disorders we've been able to pin it down in that fashion.

Screening Individual Diabetes Genes

Anyway, there's a lot more data out there than we know how to use. That will be the same theme that I want to hit with respect to the second and last broad topic, which is type II diabetes. So, we have a group of people in the lab and at the Center who are very interested in type II diabetes, and, have again, been trying to take an integrative approach to this. And so type II diabetes as you know is a serious disorder, most it effects something, 7, 8 percent lifetime risk for Americans. It's increasing in frequency together with obesity, and it's serious because of the secondary consequences of type II diabetes, a contributor of coronary disease, stroke, etcetera. And it has significant genetic components, I should mention as well. For this purpose, if you're not a diabetes expert, I just want to remind you that an important feature of type II diabetes—insulin is produced by the pancreas, it's target tissues, muscle and fat, and that in type II diabetes, the adult diabetes, insulin-mediated uptake of glucose, the uptake of glucose in response to the insulin, is diminished, there is decreased insulin sensitivity, or insulin resistance. And as a result blood glucose goes up. The diabetes aficionados know that there's also problems with insulin production, but I'll focus primarily on insulin resistance in some of what I'm going to say, and so I just want to remind you of that.

So what causes this insulin resistance? What's the basis of diabetes in these, in the target tissues of muscle and fat? Well, the problem is not a lack of pathways to explain it. The problem is too many pathways to explain it. And cases have been made in the literature for the defects being in the insulin-signaling pathway, the PI-3 kinase signaling pathway, the TNF pathway, etcetera, etcetera. There's no shortage of potential explanations, there's just not clarity as to which of these explanations are really relevant or not.

So we tried an experiment. We tried a couple of experiments that we've been going at, and these are, these are related to each other. I'll start with a human physiology experiment. This is work that David Altshuler and Leif Groop in Malmö have undertaken, together with several of us at the center. And here's how the first experiment went. We got diabetics, we got normals in Leif's lab in Sweden. We put the individuals on a euglycemic hyperinsulinemic clamp, that is, we kept them at normal blood sugar, high insulin, for two hours, then removed muscle tissue, did a muscle biopsy, made RNA, and looked at what genes were responding to this hyperinsulinemic state. You've got 18 diabetes, 17 normals, you put them on gene-expression chips, you get all the samples, and you can imagine you can do some mathematical analysis that lets you find genes that are differently expressed amongst the diabetes than the normals, and in every gene you can see how different the expression level is, and you could sort the genes, the ones that are most different and least different.

The traditional approach is to now write a paper singing, you know, the praise of the top genes on the list, saying, "Aren't these interesting?" The problem with that is that they might be there just by chance; something has to be on the top of the list, how do you know it means anything? Well here because we have 18 diabetics and 17 normals we can do a scrambling experiment; we can randomly scramble the categories between the patients, redo the whole thing and say, "Now what's at the top of the list and how extreme is that?" And that way you can rule out those differences that aren't any more extreme than you would expect by chance. So when you do that and you apply a rigorous statistical criterion, the genes that are significantly different really now mean something, and the answer is we found none.

So this was depressing because this was not a cheap experiment, and it was an experiment we had high hopes for, and so it—it sort of sat around for a while. Now this is—I've got to note—if we had not had 17 diabetics—18 diabetics and 17 normals and done this thing, we would've been able to write a paper. It is only because we had more patients and were able to do this scrambling experiment that we knew we didn't know things. So this is a case where having more knowledge means you know less, although what you know is more right. In any case.

Screening Sets of Genes

So what do you do, how do you get information out of this? So an interesting idea was proposed, again by Vamsi Mootha and by Aravind Subramanian. The idea was this: suppose I take a—basically they said, Aravind said, "Don't look at one gene at a time, look at sets of genes. Supposing I give you a set of genes that I think are physiologically important, and I look at where did those genes turn out on my ranked list of gene differences? If this gene set has nothing to do with the physiological process, it'll kind of be randomly distributed. But if it matters, it'll be near the top of the list."

Now this could be a lot more sensitive than looking gene-by-gene. So we do some test. Let's start with one that we know to fail, the urea cycle. That can't have anything to do with diabetes—you test genes involved in the urea cycle and just as you expect, scattered around the list at random means *bupkes*. Now we go for the gold, insulin signaling, scattered around the list doesn't do a thing. So we had to get serious. So we began to curate sets of genes from the web, some publicly curated sets and sets that Vamsi curated. We also took anonymous gene sets, just based on clusters of expression about which we knew nothing because we thought maybe we don't know everything, we put it in there. We put each of these through the test. One gene set gave a screaming signal, right there, that is, oxidative phosphorylation. Oxidative phosphorylation is an unambiguously strong signal. Any one of those genes, as I'll come to in a second, is not a strong signal, but as a set very strong.

Oxidative phosphorylation, as you know, is involved in ATP production, etcetera. I won't have time to explain if you don't know, but oxidative phosphorylation makes a lot of sense. You should know exercise, for example, increases oxidative phosphorylation; caffeine, by the way, and caffeine is known to help decrease your risk of diabetes. Caffeine is good for you with regard to diabetes;

it's bad for other things but it's good for diabetes. Have a cup of coffee, that's good.

Now what was striking about the OXPHOS genes is that the OXPHOS genes on average were only diminished 20 percent. Those of you who do gene-expression experiments on chips know that 20 percent means *bupkes*, all right, in New York, I can use the technical term. Twenty percent means *bupkes*; you're not going to be impressed by 20 percent because you know that the noise on these gene chips is much bigger than 20-percent variation of expression level. But to see a pathway of 106 genes where the vast majority of them are down by about 20 percent is highly significant, and when you think about it in terms of metabolic flux through this pathway, it is indeed potentially very significant.

As an aside I'll mention that you might think that this produces measurable physiological consequences, and we measured total aerobic capacity of the patients, VO2 max. It turns out that this gene-set predictor is a highly statistically significant predictor of total body aerobic capacity, and a better predictor than your diabetic state. So in fact, you can see by that 20-percent reduction in the pathway measurable traits in patients there, about VO2 max.

Now, we're going to do—we'll get a little technical near the end of the talk—it wasn't the only gene set that turned out to be important. The other gene set that gave us a strong signal, not quite as strong as OXPHOS, was an anonymous set of genes that had been identified by clustering genes from other experiments, a set of coregulated genes, which we didn't know what it meant, but we threw it in. It, too, gave a signal. It turned out that almost all of the signal came from the intersection of those two sets, a subset of the OXPHOS genes that were also coregulated with each other. Well what are they?

So here it took a little bit of biological insight and Vamsi guessed that this subset of the OXPHOS genes that were coregulated, found by this completely computational method, might be the targets of a common transcriptional coactivator, and he guessed it might be PGC-1α. So PGC-1α is a coactivator implicated in mitochondrial biogenesis, and so he made a transgene with PGC-1αα, put it back into mouse fat cells, and in fact looked at the expression pattern and bingo, it is very clearly the targets of PGC-1α. And you could read it out from that.

So that's the first story there, where there was a lot more information in those expression patterns than we'd realized. And I must say what the take-home lesson to me about this was that for five or six years we've been doing expression experiments. I bet we've missed most of what those experiments are trying to tell us. And we have to go back now and reinterpret all those experiments, and we're beginning to do that for cancer and for other things, and it's clear there's a lot more there than we had been seeing, so we're collecting gene sets. If you have good gene sets let's us know, we're going to try to get on the Web a public set of gene sets. And of course this is just one of many, many ways to kind of filter data to get stuff out.

Cell-Culture Models for Diabetes

So let me turn to the second topic—which I'll do briefly—which instead of human physiology is cell-based models related to diabetes. So this is the work of a graduate student Nick Houstis. And very briefly here you can make cells insulin-resistant in culture. So Nick decided to take 3T3 adipocytes, treat them in various different ways to make them resistant to the effects of insulin. It's not that hard to make cells insulin-resistant. Curing them when resistant is a big deal, but making them insulin-resistant there are some treatments. Tumor necrosis factor alpha does, dexamethasone does. So he treated cells and what he did was he got their expression patterns, treated with these two agents that cause insulin resistance, and his goal was to find out what were common effects of both of these independent things that could produce insulin resistance, figuring that each of them would produce many of their own specific effects, but things that were in common might be more interesting.

So he has the same kind of setup, cells treated with dex and TNF, he's looking for things that are in common between those, but differing with those, same sort of thing, make some mathematical measures, sort the genes, what do you see? And here an interesting set of genes emerges, genes involved in reactive oxygen species generation. So reactive oxygen species, as some of you will know and others of you may not, are things like superoxide, hydrogen peroxide, and hydroxyl radical. They are produced by mitochondria. Again, they're [inaudible] now are related sort of to OXPHOS here, they're produced by high flux through the electron transport chain, and thought to be very damaging cellular agents, although whether that's the way they're acting is another question.

So our interpretation of the experiment says, "Hmm, you're seeing genes related to reactive oxygen species. Maybe that's telling you something about the physiology that's going on here. So you have to look."

So first off, do you actually see higher levels of reactive oxygen species. It's one thing to see it on a gene chip that there are genes, but can you measure reactive oxygen species? And the other is, is that just a passive marker, or is it potentially causally involved in the insulin-resistant state? So what he did first was he took cells, treated them with TNF and dexamethasone and confirmed indeed, as he had done before, that they had become resistant to the effects of insulin. This is glucose uptake. Then he developed some ways to measure reactive oxygen species, three different assays measuring reactive oxygen species, and he said, "When I make my cells insulin-resistant, is it really true that I see higher levels of reactive oxygen species?" And the answer was yes, he sees higher levels of reactive oxygen species.

Then he asks—so that says yes, it's not just an artifact there now is it causal? In theory if you could treat the cells with an antioxidant you ought to be able, if you believe this, to reverse the effects of insulin resistance. So he worked out a couple of ways to do that. There are some small molecules that have antioxidant effects, and there are some transgenes that—catalase for example—that should have antioxidant effects. And he therefore treated cells with small molecules and also made transgenic versions of the cells and asked, "Do you in fact see some restoration of insulin sensitivity?" And the answer was yes, these treatments on average restore about 50 percent of the insulin sensitivity of the cells, saying that it's not merely a passive bystander but it is somehow causally involved. I don't want to overstate that because it's not a total restoration, but it's not that easy to find agents that will reverse insulin sensitivity, so it's quite a meaningful observation there, and I'm sure that the pathway is not a simple one. It may be that high levels of ROS trigger all sorts of cellular responses which as a bystander consequence reduce sensitivity to insulin.

Anyway, I don't mean to make too much about that other than to say, ah ha, another time the inherent information in the library is pointing us to very reasonable pathways there, if only we can figure out how to extract that information.

Finally I'll mention just—oh actually I should mention for the aficionados—there are two rare human genetic diseases, Lou Gehrig's Disease, ALS, and Friedreich's ataxia, both of which affect proteins that are involved in reactive oxygen species. A little-remarked-upon observation in the clinical literature is that patients with both ALS and Friedreich's ataxia show insulin resistance. And so that's obviously a circumstantial point, but it's not an unrelated point here to saying that this may indeed be an interesting clinical connection.

Screening Genomes for Variations

Finally the last point, human genetics—I'll just say very briefly—the other kinds of data you can bring to bear on this are matching up not expression differences between cells but inherent DNA-sequence differences between individuals. To make a long story short, you can look at all the different variations up and down the human genome to look for genes that have sequence variations that correlate with diabetes. David Altshuler a couple of years ago did a study like that with many different candidate genes, and I summarize it in just one slide. When he puts all this together, one gene comes through with a screaming signal, and this has now been confirmed in many labs around the world, PPAR-γ in fact, affects your risk of type B diabetes by about 20, 25 percent, and so sifting through genetic data tells you an interesting target. And I raise it because PPAR-γ is a partner of the PGC-1α coactivator that comes out in that experiment. How exactly all these things put together is a subject of much debate; I'm not going to pretend that an exact pathway can be put together other than lots of circumstantial evidence connecting these, some evidence at least here and here of causality, not just pure correlation.

But mostly I present it to you not for a study on diabetes, but to say that there's lots more information in the clutter, if we can figure out how to extract it. And I think that's the exciting thing, is that biology as information, as a large collection of information, it's obviously very complex. But the tools to be able to read that information in a sensible and sensitive way, I think the next generation of students are going to be developing lots of them. That's what is really exciting about this next era. And it pushes us to a slightly different paradigm. The long-standing paradigm has been you formulate your own question, you collect your own data, the experiments of individual scientists. The complementary paradigm—obviously that will still go on—but the complementary paradigm is you formulate your question but go consult common data produced by many scientists across the world, representing in fact the experiments of nature, and that that is at least as important a complementary point of view.

But if that's the point, if it is the consultation of common data that will help us there, we need lots of it. We need common data about genome sequence, much more broad sequence across evolutionary trees, from deep sequence from related species, so we can pick out these regulatory factors of the sort that Mike and I have been talking about; we need information about the DNA variation in the human population so that we can correlate it with risk of disease, and in that connection I can only say there's only about ten million variations, common variations in the human population, and more than five million of them are already in the database. So this is not a crazy thing to imagine doing.

RNA profiles, protein profiles, across tissues and cells, across disease and healthy states, and under chemical and genetic perturbations. We need to collect all this, and I must say if the point is to use this common data, it must be freely available to everyone, because it is only the free availability of the whole set of data that makes it powerful, and that's why it was such a big issue to the people involved in the Human Genome Project that all these kinds of data get out there for everyone to use.

I shall stop there and simply say to Columbia happy quarter of a millennium. The next quarter of a millennium I suspect will be even more exciting than the past, and it's a pleasure to come down to New York and celebrate it with you. Thanks very much.

Andrew R. Marks: Well, that was great, although I'd like to know how a kid who grew up in New York could become a Red Sox fan, but topic for another day.

[Eric Lander responds from the audience.]

That's even worse.

Michael Brown, M.D. and Joseph L. Goldstein, M.D., University of Texas Southwestern Medical Center, Dallas, TX
Fatricide: When Genes and Diets Collide

Introduction by Andrew R. Marks

Andrew R. Marks: Our next two speakers are Joe Goldstein and Michael Brown. Joe is the Paul J. Thomas Professor of Medicine and Genetics and chair of the Department of Molecular Genetics at UT Southwestern and the regental professor at University of Texas. Mike Brown is the Paul J. Thomas Professor of Molecular Genetics and director of the Johnnson Center for Molecular Genetics at UT Southwestern. Together they were awarded the Nobel Prize in Medicine or Physiology in 1985. Their work has touched the lives, I would venture to say, of everybody in this room, because it seems that everybody today is either on the cholesterol-lowering drug statin or knows somebody who is.

Indeed the work of Brown and Goldstein, as they're affectionately referred to, stands as one of the most powerful models for the application of basic studies, in this case the discovery of the LDL receptor that provides the molecular link between cholesterol and heart disease. These basic studies forming the basis of developing novel therapy, in this case cholesterol-lowering agents, have treated millions of patients.

Today they will tell us about recent work identifying genes that control metabolic pathways that have become maladaptive in our modern lifestyle, despite millions of years of evolution. Welcome to Columbia.

Cholesterol and Coronary Atherosclerosis

Joseph L. Goldstein: Thank you for the nice introduction, and also Tom Jessell, thank you for organizing such a great program, and Mike and I are delighted to be here to help Columbia celebrate their 250th anniversary. And so speaking of celebration, all year we've heard about the most celebrated molecule in biology, DNA. And now Mike Brown and I would like to tell you about another celebrated molecule, cholesterol. And although it lacks the panache of DNA, cholesterol is important to medicine because it is the root cause of coronary atherosclerosis , which is the most frequent cause of death in the U.S. and other industrialized countries. It's surprising that coronary heart disease was recognized as a major clinical problem only in the early years of the twentieth century. But by 1950 the disease had reached epidemic proportions, and today coronary heart disease causes more than one-third of all deaths in the Western world, and that's the bad news.

But the good news is that in the twentieth century, scientists have discovered that the disease is caused by cholesterol, and most specifically by lipoprotein carriers that transport cholesterol in the blood. And by the end of the twentieth century, scientists had developed powerful drugs that lower these toxic lipoprotein particles and prevent heart attacks.

So in part one of our joint lecture I will tell you how the link between cholesterol and coronary atherosclerosis was established, and then I'll tell you about the disease familial hypercholesterolemia and how it led to the discovery of the LDL receptor which is a molecule that controls the blood-cholesterol levels in humans. And in part two Mike will tell you about some of recent research on the transcriptional regulation of cholesterol on metabolism and its link to garden variety forms of high blood cholesterol.

Now here's a cross-section of a coronary artery of a 50-year-old man who died suddenly of a heart attack. And this atherosclerotic plaque began to form forty years ago when this man was a teenager, and it began when the cholesterol-carrying lipoproteins infiltrated the bloodstream and entered the artery wall where they underwent oxidation. And the oxidized lipoproteins initiated an inflammatory reaction that over forty years led to this damage that you see and scarring and eventually the buildup of a plaque that narrowed the channel of the blood vessel. And this insidious process of plaque formation was augmented by aggravating factors, such as smoking, high blood pressure, and a high-fat diet. And when a plaque becomes unstable it eventually ruptures, leading to the formation of a blood clot, a thrombosis, that suddenly blocks the blood flow of an artery, and then the heart muscle supplied by the artery dies from lack of oxygen and it produces what's called an acute myocardial infarction, or a heart attack.

Now the lipoprotein that initiates this a atherosclerotic plaque is called low-density lipoprotein, or LDL. And LDL is made in the liver, and it's secreted into the blood, and in humans LDL is the major cholesterol-carrying lipoprotein in the blood plasma. And now look. The cholesterol that LDL carries, as shown right here, is a hydrocarbon molecule composed of four rings and a side chain, and it's totally insoluble in water. And this insolubility allows cholesterol to perform its most vital function in the body, which is to act as a component of the lipid-rich plasma membrane that forms a water-resistant barrier around all cells in the body. But the insolubility of cholesterol creates a transport problem; in order to reach its sites of metabolism outside the liver, the cholesterol must be solubilized so that it can be transported in the bloodstream, and this is the function of LDL.

These LDL particles are spherical particles that are around 22 nanometers in diameter, about the size of a small virus. And the core of each LDL particle consists of 1,500 molecules of cholesterol ester, which is a cholesterol to which a long-chain fatty acid is attached. And this hydrophobic core is surrounded by a polar coat composed primarily of phospholipids and a protein called apoprotein

B. And this polar coat solubilizes the particle in water and allows it to be transported in the plasma. And the link between LDL cholesterol and coronary atherosclerosis is based on four lines of evidences—epidemiological, genetic, experimental and therapeutic. And by itself any one of these lines of evidence might not be completely convincing, and in fact for many years there was a cholesterol controversy. But when one considers all four lines of evidence in concert, the argument becomes irrefutable.

The History of Cholesterol Research

The first link between cholesterol and atherosclerosis was established in 1913 when a Russian pathologist, Nicolai Anitschkow, fed pure cholesterol to rabbits and produced atherosclerotic plaques. And this was the first experimental production of atherosclerosis, and Anitschkow's classic experiment has now been repeated many thousands of times in virtually every species, from pigeons to humans.

At the time of these experiments, pathologists believed that a thrombotic occlusion of an atherosclerotic plaque in a coronary vessel was always a fatal event, and the clinical syndrome of nonclinical myocardial infarction was not recognized until 1918 when the American clinician James Herrick made the first use of the electrocardiograph, the EKG, to diagnose heart attacks in patients who presented with crushing chest pain. And Herrick provided the first demonstration that thrombosis of a coronary artery was not always a fatal event, and that coronary heart disease was responsible for the syndrome that had been previously diagnosed as indigestion or apoplexy.

In 1938 the connection between cholesterol and heart attacks to humans was firmly established on genetic grounds when the clinician Carl Müller in Norway described families in which high blood cholesterol was present from birth, and in which early heart attacks occurred in these hypocholesterolemic relatives. The disease came to be known as familiar hypocholesterolemia, and I'll have more to say about that in a moment. The mounting clinical interest in cholesterol led to an intense effort to understand the pathway by which cholesterol was synthesized in the body, and this pathway was worked out in the 1950s by four biochemists, Konrad Bloch, Feodor Lynen, Cornforth and Popjak. And actually Bloch's work began here at Columbia P & S when he was a postdoctoral fellow with Rudolf Schoenheimer in the Department of Biochemistry.

This slide shows you the cholesterol biosynthetic pathway as worked out by the four biochemists. And at the time it was a tour de force in biochemistry; no pathway of this complexity and magnitude had ever been worked out, and the challenge was to figure out how the two carbon acetyl-CoA could be converted to the 27 carbon cholesterol which has these 4 rings and a side chain. And to make a long story short, it involved a series of 25 different enzymatic steps, and a key enzyme that you'll hear more about in the talk from Mike and me is, which formed malonic acid. One of the really brilliant insights in working on this pathway was the realization that squalene, a 30-carbon straight hydrocarbon, could be folded in such a way that with the proper enzymatic conversions it would then lead to the classic sterol side chain.

So remember, as I mentioned before, cholesterol is totally insoluble in water, and in order to be transported in the blood it must first be incorporated into LDL particles. And LDL was first identified as a risk factor for coronary disease in 1958, when John Gofman, a medically trained biophysicist at the University of California in Berkeley, used the newly developed ultracentrifuge to separate plasma lipoproteins by floatation. Gofman found that heart attacks correlated with elevated levels of plasma LDL, and Gofman was also the first scientist to discover that heart attacks were less frequent when the blood contained elevated levels of another lipoprotein, HDL. So Gofman's discovery of LDL as a risk factor for heart attacks was followed over the years by an avalanche of confirmatory epidemiological studies, such as the Seven-Country Study of Ansel Keys, and the Framingham study that was supported by the NIH.

Now according to Richard Peto, an eminent epidemiologist at Oxford, the total body of epidemiological data over the last 45 years implicating LDL as the cause of atherosclerosis is more convincing than the total body of data implicating smoking as a cause of lung cancer. It was at this point in the century of cholesterol and coronaries that Mike Brown and I entered the picture. In 1973 we discovered that the level of LDL in blood is controlled by a cell surface receptor that we named the LDL receptor. And we also found that Müller's familiar hypercholesterolemia was caused by mutations in the LDL receptor gene. And this was the first molecular link between cholesterol and atherosclerosis.

In 1976 a Japanese scientist Akira Endo discovered a fungal metabolite that could block cholesterol synthesis by inhibiting the enzyme HMG-CoA reductase. And this was the first statin drug . And Mike and I collaborated with Endo to show that the inhibition of cholesterol synthesis led to an up regulation of LDL receptors, which explained how these drugs could selectively lower LDL, the bad cholesterol, without lowering HDL, the good cholesterol. We encouraged the Merck company to develop these drugs, and in 1986 the first statin was approved for human use. This year the statins will be consumed by more than thirty million people worldwide.

In 1994 a landmark epidemiological study called the 4S Study was completed, and conducted by scientists in four Scandinavian countries. The 4S Study was the first of six large prospective trials to show that statins by lowering LDL could not only prevent myocardial infarctions but they could actually prolong life. And this therapeutic triumph provided the final link in the cholesterol-coronary chain, complementing the experimental link, the genetic link, and the epidemiological link.

Familial Hypercholesterolemia

So now let me tell you how Mike Brown and I got interested in cholesterol, and a little bit about our work on LDL receptors, which has contributed to the genetic link in this story.

So our interest in cholesterol was stimulated by taking care of patients like the little girl shown here. This little girl has a homozygous form of familial hypercholesterolemia which I'll refer to homozygous FH. Her plasma LDL level was elevated tenfold above normal since the time of birth, and her total plasma cholesterol level was 1,000 milligrams per deciliter. Some of the excess LDL particles deposited in her skin, as you can see here, forming these bumps which are called xanthomas. Whenever this little girl bruises her skin, the capillary blood vessels break, allowing the excess LDL to escape from the bloodstream and enter the skin tissue, where the LDL particles undergo oxidation. And the oxidized LDL deposits in macrophages and initiates an inflammatory process that produces these xanthomas.

And this same process that you see here on the slide, the formation of these xanthomas in skin, occurs in her coronary arteries, and this little girl suffered multiple myocardial infarctions before age seven. In FH homozygotes LDL is really the only risk factor for atherosclerosis. This little girl does not smoke, she has no hypertension and she doesn't have a type A personality.

Well, Mike Brown and I first saw two children with this disease when we were research associates at the NIH 35 years ago. We were awestruck by their striking clinical picture, and we decided to work together to try to figure out how a genetic defect in a single gene could produce these high levels of LDL which ultimately led to the severe atherosclerosis.

Although homozygous FHs is a rare disease, the heterozygous form of FH is the most common genetic disease in humans throughout the world in all populations that have been studied. And this next slide summarizes the clinical features of the heterozygous and homozygous form of FH; the heterozygous occur in one in five hundred individuals in every population that's ever been studied. These individuals have a two-and-a-half-fold elevation on average in their plasma LDL from the time of birth, and because of the sustained increase in their LDL, they begin to have myocardial infarctions at around 35 to 45 years of age, if they are not treated. Now, because of treatment, this number fortunately is being delayed to later years in life.

The rare homozygotes who have to inherit a gene from both of their heterozygous parents to have this disease in a severe form occurs at a frequency of one in a million; this is like the little girl I just showed you the picture of. They have a plasma LDL level that's greater than six-fold from the time of birth and because of the magnitude and very high increase in LDL from the time of birth, these individuals typically have heart attacks between the age of 5 and 15 years, and we've actually seen one FL homozygote in Toronto who had a first heart attack at 6 months of age.

Now, 5 percent of all individuals that have a heart attack under age 60 have the heterozygous form of familial hypercholesterolemia which, as I mentioned, occurs in one in 500 people, so this is an important public-health problem as a single-gene disease.

Identifying the LDL Receptor

Now our approach to working out the molecular defect in FH was to study the skin fibroblast in tissue culture, comparing LDL metabolism in cells from normal subjects from those with patients with the homozygous FH, and we began our studies in 1972 in UT Southwestern in Dallas. And to make a ten-year story short, we found that fibroblasts require cholesterol for growth in tissue culture, and that cells could obtain this cholesterol from two sources: they could either synthesize cholesterol from the cholesterol synthetic pathway that I showed you earlier; or they could take up cholesterol by the uptake of LDL that was derived from the fetal-calf serum in the tissue-culture medium. In order to take up this LDL, the cells produced a specific cell-surface receptor that we call the LDL receptor, and cells from patients with homozygous FH turned out to lack LDL receptors, and since these mutant cells could not take up LDL from the serum, they obtained all their cholesterol from growth from the endogenous synthetic pathway.

The availability of mutant FH cell lines that lacked LDL receptors allowed us to work out the biochemical steps in this LDL receptor pathway, which is a prototype for the general process of receptor-mediated endocytosis. LDL receptors are concentrated in specialized regions of the plasma membrane called coated pits. These coated pits pinch off to form coded vesicles, the coded vesicles shed their coats and ultimately fuse with other vesicles to form an endosome, which has an acidic pH compared to the coated vesicle, and once the receptor-bound LDL is in the endosome the acidic pH causes the LDL to dissociate from the receptor, and the receptor is recycled to a recycling vesicle, going back to the cell surface where it then concentrates again in coated pits to pick up another particle of LDL.

Each LDL particle, as I mentioned, contained 1,500 molecules of cholesterol ester, and each receptor makes about 1,200 trips in and out of the cell during its 30-hour life span. Each trip is about 15 minutes or so, and so one has an enormously efficient mechanism for transporting a molecule like cholesterol into a cell; it's probably one of the most efficient transport mechanisms ever designed in nature.

Back into the endosome where the LDL dissociates from the receptor in the acid environment, and that LDL particle, this part of the endosome then fuses with other vesicles to enter a lysosome where the hydrolytic enzymes of the lysosome degrade the LDL particle, releasing the cholesterol which is used for various structural purposes and regulatory purposes that Mike will talk about later in his talk.

LDL Receptor Structure

So by 1982 we were able to purify the LDL receptor and then with the help of all the exciting technology that had come around with recombinant DNA, we were able to clone the cDNA in gene for the receptor in 1984, and the deduced sequence of the receptor was quite revealing. It showed a molecule of 839 amino acids that had five distinct domains. The first domain consisted of multiple repeats that served as a ligand binding domain the apoprotein B of LDL. And then the second domain, this A-B-C, is a region of the receptor that senses the acidic pH in the endosome and somehow leads to a conformational change in the receptor in such a way that the LDL is released from the receptor, so that the receptor can recycle back to the cell surface. And then there is a region that has sugar molecules, and then there's a region that anchors in a receptor in the membrane, and then there's the targeting signal in the cytoplasmic domain that tells the receptor to go to coated pits.

Now this is a very complicated structure for structural biologists to solve because it turns out to be there are cysteine residues in this receptor, there are thirty disulfide bonds, and so it took many years for Hans Deisenhofer in Dallas to solve this structure, and just in the last year he and his postdoctoral fellow Gabby Rudenko, after 15 years of work, have been able to solve the structure of the extra cellular domain in the LDL receptor, and it revealed a very interesting point to illustrate this recycling mechanism that I mentioned to you.

So here we are looking at—this is the structure by Rudenko and Deisenhofer—this is these repeats, these seven repeats, that bind the apoB of LDL, so a neutral pH LDL would presumably be bound up here, and this part of the molecule would not be where it is right now. But in acidic pH what happens is that this A-B-C part of the molecule which exists as a β-propeller in the three-dimensional structure now binds to several of these repeats, these repeats 3, 4 and 5, in such a way that it ejects the LDL particle that's bound up here, and this now allows the receptor to recycle to the cell surface to pick up a new particle of LDL, and then releases the LDL to now go into lysosome. At least that's the hypothesis. The structure's only been done in one confirmation and now has to be done in various different confirmations, in different pHs, to prove whether this is in fact right or wrong, but it's consistent with the biological information that Mike and I and our students and postdocs have derived from mutagenesis studies. But in any event, it's gratifying to have a sequence to begin to think about after all of these years.

Now, as I've already mentioned to you, the FH had mutations in their LDL receptor gene that disrupt receptor structure and function. This just happens to be a map of the LDL receptor gene on the short arm of chromosome 19. There are 18 exons, and this just shows the position and the types of the first 100 mutations that we and our colleagues in Dallas worked out from 1984 to 1990. And then after the first 100 we decided to move onto other things. And now other people have gotten involved in a big way and there are more than 1,200 different mutations in this very relatively short LDL receptor gene that have been identified in patients with FH. With few exceptions, virtually every unrelated family that has this disease, familial hypercholesterolemia, has a different mutation in the LDL receptor that affects the structure and ultimately the function of the receptor.

Now I'd like to turn to the question of how does these LDL receptor mutations produce a disease, how does the deficiency of receptors raise the plasma LDL levels in patients with FH? And so our studies showed that in normal humans, the majority of LDL receptors in the body are expressed in the liver where they act to remove LDL from plasma and by receptor-mediated endocytosis. The FH heterozygotes, as I mentioned, have half the normal number of functional LDL receptor genes, and they thus have one normal LDL receptor allele and one mutant LDL receptor allele, and they remove LDL from plasma at half the normal rate. And as a result they have twice the level of LDL in plasma compared to a normal individual. And then the FH homozygotes, the ones shown here, have absolutely no receptors, have a very huge level of LDL in their plasma that's only removed at a very low rate by nonspecific, non-receptor-mediated process.

So a physiological approach to therapy in the FH heterozygote would be to increase the activity of the receptors encoded by the normal allele. If one could get them to work twice as hard, then one would be able to create a state that would look more like the normal, and then these increased number receptors would remove the LDL from plasma at an enhanced rate. So this would be a pharmacological form of gene therapy. And this brings me to the next topic, designing a rational therapy for FH heterozygotes.

Treating for Familial Hypercholesterolemia

And now the mechanistic insight into how drugs could be used to lower LDL levels came from knowing that cells can obtain cholesterol from two pathways, both of which are under the same feedback control. So when cholesterol synthesis in the liver is decreased, the cell responds by increasing the synthesis of LDL receptors, which in turn leads to an increased removal of LDL from the plasma. And this is precisely how the statin drugs appear to work. Now the statins are competitive inhibitors of the HMG-CoA reductase enzyme, and when you ingest the statin drug, it goes directly to the liver where the drug inhibits HMG-CoA reductase, which in turn reduced cholesterol synthesis and raises LDL receptors in the liver because of the existence of this feedback system which Mike will actually tell you about in more detail.

But the overall effect of administering such a statin drug is to decrease the plasma LDL level, and when the plasma LDL now enters the liver and its cholesterol is now available in the liver, one has a restoration of the normal cholesterol content of the liver. So one has a drug that actually works by lowering the plasma LDL without essentially affecting the level of cholesterol in the liver; it's almost a perfect system. You could not ask for a better drug or a better system for a drug to operate on.

Now, there are 6 statins that are now approved for human use and are taken by 30 million people worldwide, and the 3 most commonly prescribed statins are Lipitor, Zocor, and Pravachol, and I'm sure that these drugs are very, very familiar to many of the older professors in the audience, as Andy pointed out earlier.

This slide shows you how different individuals respond to the statins. The FH homozygotes who have no receptors show a very poor response, as would be predicted from the model that I showed you. The FH heterozygotes who have one normal LDL receptor allele and one mutant LDL receptor allele when maximally treated with statins have a very good response, a 40-percent drop in their cholesterol level. And normal subjects actually who have two genetically normal LDL receptor alleles show an even greater response, when maximally treated, and that's because their receptors are metabolically suppressed for reasons that Mike will tell you about in a moment.

So these statins clearly lower plasma LDL levels in FH heterozygotes and in so-called normal hypercholesterolemic individuals. Do they decrease heart disease? And this is a summary, actually a meta-analysis of the first 5 clinical trials that were done between 1994 and 1998 on 30,000 subjects followed for years, half receiving statins, half receiving placebos. And the plasma LDL in these clinical trials dropped by 28 percent and the heart attacks over the next 5 years dropped by 31 percent. These—you didn't see the larger response I showed you on the last slide because these patients were not really maximally treated; they were taking many different drugs and they were given one standardized dose.

The most recent study done by Collins and Peto at Oxford was just published last year. The Heart Protection Study involved a single study of 20,000 subjects followed for years. Plasma LDL was reduced by 33 percent, heart attacks and strokes were decreased by 33 percent over this 5-year period.

Now these individuals, these sixty thousand, fifty thousand people were only followed for five years with a 30-percent reduction in coronary events. And the question is what happens if they had been taking this drug for 10 years, or 20 years, or 30 years, would one ultimately prevent heart attacks had they started

10 or 20 years earlier? And that's a question that will eventually be answered over many years, but many people believe that one will see a dramatic effect if these drugs are started at a much earlier age.

Now the vast majority of these fifty thousand people in this study are not heterozygous FH, but they're people with genetically-normal LDL receptor genes, the people who are so-called polygenic hypercholesterolemia, a garden-variety hypercholesterolemia. And that brings me to the question why do people with normal LDL receptor genes have LDL levels high enough to cause heart attacks? And I just can't answer that question; I have to ask Mike to come up and answer that question for you.

"Normal" and "High" Cholesterol

Michael S. Brown: Okay, if you could just leave that on, it's the same talk, go back to where it was before.

Well, let me just start by saying that I am not the descendent of Alexander Hamilton's wife's first husband. Nor did I attend any sessions at Columbia when I was in high school. But I have a very high regard for Columbia and it's for that reason that, I think, Joe and I were very honored by the invitation to—we're going through somebody else's talk, we just have to go back to that, that's it now—what we want to do is go down to this thing, the slide show, just cancel it. Now how do we go back onto—okay. All right, so, all right.

Joe has told you about LDL receptors, and how their normal job seems to be in part to protect—not only to deliver cholesterol to cells but also remove LDL from plasma, and thereby to protect us against high levels of LDL which lead to heart attacks. And the question then is if 499 out of 500 of us have normal LDL receptor genes, why do one-third die of heart attacks? And the answer comes down to the definition of normal.

Here you see the plasma cholesterol level in industrialized countries like the United States, Western Europe, and the rest of North America, and what you can see is for a 40-year-old man, the mean cholesterol level is about 210 milligrams per deciliter. And we know that the incidence of heart attacks rises as the plasma cholesterol level rises, but it's already high in people with normal, at least statistically normal, plasma cholesterol levels. And that is very hard for some people to accept. Why should we have heart attacks with normal LDL cholesterol levels; LDL can't be the cause of this because we have normal levels.

Well, this is only normal when you look at these industrialized countries. If you were able to measure the cholesterol level of everybody in the world, this is the curve you would get. Now what you would find is that the median value for most of the world is a plasma cholesterol level of about 150. Now there's a simple reason for that; there are a lot of Chinese people in this world, but the fact is that for the majority of human beings who live outside of industrialized countries that's what the cholesterol level is, that's what the cholesterol level is in primates, that's what the cholesterol level is when we're first born before we've done anything to change the cholesterol level. And so one can look at our whole society and see that we're all above the ninetieth percentile. So it's not surprising that we're seeing heart attacks in this whole group.

Now what's the reason why we're faced with these high cholesterol levels in North America and Europe? The answer is not a mystery. This is from *Time* magazine. I don't know if any of you can see the—to show you how old this is, this is Walter Mondale and Gary Hart up there. So we've known about the diet and cholesterol for a long time, and in fact there's a huge amount of data that says the reason we see these high cholesterol levels in Western countries is because of the diets that we eat that are rich in cholesterol and saturated animal fats.

And the question is why should these diets raise the plasma LDL cholesterol level, why shouldn't the LDL receptor just keep removing the cholesterol from the blood? Well the answer is contained in something that Joe referred to, and that's the regulation, the feedback regulation of the LDL receptor. So if one considers the liver, so then the liver of somebody living in China has lots of LDL receptors and LDL levels are kept low; if somebody has familial hypercholesterolemia in the heterozygous state there's a 50-percent reduction and plasma LDL builds up to twofold above normal, but in people living in—who eat high-fat diets, the cholesterol is absorbed in the intestine, and it's delivered directly to the liver. So all of the cholesterol that you eat does not hang around in the bloodstream very long; it's cleared in the liver almost in one pass through that organ. And the cholesterol is taken into the liver by a totally separate receptor called the chylomicron-remnant receptor, and that receptor basically concentrates—all of the cholesterol that you eat gets concentrated in your liver cells. And in fact, it's contained within the liver membranes, because all of the cholesterol in cells is contained within membranes in the cell, since cholesterol is insoluble.

And then what happens at that point is something that Joe illustrated, and that is the feedback mechanism that controls the production of LDL receptors comes into play. And a signal is sent into the nucleus of the cell, and the gene encoding the LDL receptor is partially suppressed. It's not suppressed by 50 percent, or we would all look like FH heterozygotes, but it's suppressed by about, oh, say, 10 or 20 percent, which is probably enough to raise the plasma cholesterol level from its normal 150 up to the 210 to 300 values that we customarily see in this country. And so it's the down-regulation of LDL receptors by dietary cholesterol and saturated fatty acids that leads to the elevation of LDL in plasma.

How LDL Receptor is Regulated

And the question we'd set out to answer several years ago is how is this information transmitted to the nucleus? After all, cholesterol is insoluble, it's only in the membranes of cells, and how does the nucleus of the cell know that the membrane of the cell is saturated with cholesterol? And so several years ago Xiaodong Wang and Mike Briggs in our laboratory purified transcription factor from nuclear extracts that controls the production of LDL receptors, and we call it the sterol regulatory-element binding protein, or SREBP. We pronounce it S-R-E-B-P, some people pronounce it "srebp," but it's an unpronounceable name basically, like an Al Capp cartoon character.

But anyway, what we purified from the nucleus of cells was a classic transcription factor; it has a DNA binding and dimerization domain called the bHLH -Zip domain, and this part of the LDL receptor binds to an enhancer in the five-prime flanking region—this part of SREBP binds to an enhancer in the five-prime flanking region of the LDL receptor gene and also other genes involved in cholesterol metabolism, and turns on their transcription. So this part of the—which we purified from the nucleus is a classic factor that stimulates gene transcription.

But the SREBP differed from all of the other transcription factors that were known at that time, because when we cloned the cDNA we found that this transcription factor was actually just the first third of a very large protein of 1,200 amino acids, and following this transcription factor domain there's a membrane-attachment domain that has two transmembrane helices that is inserted into the membranes of the endoplasmic reticulum of the cell. So SREBP is not a nuclear transcription factor, it's a membrane-bound protein. And then at its carboxy-terminus, it has another six hundred amino acids that act in a regulatory fashion.

Now in order to see whether this SREBP could ever get into the nucleus, we made an antibody against this amino-terminal Zip domain and used it to stain cultured human cells. And what we found was that if we grew cells in tissue culture in the presence of sterols, so they were basically overloaded with sterols, they don't want to transcribe the LDL receptor or the cholesterol biosynthetic genes, and as a result the SREBP is not in the nucleus, it's outside of the nucleus, and it turns out that this staining is in the endoplasmic reticulum. But if we simply deprive the cells of sterols so they become hungry for sterols, then there is a proteolytic reaction that occurs that sends this Zip domain directly into the nucleus of the cells. So the transcription of the LDL receptor gene is controlled by the controlled proteolysis of this membrane-bound transcription factor, SREBP. And over the last seven or eight years we've been able to work out the mechanism by which this occurs. And let me summarize it here.

When SREBP is made, it forms a complex with another protein in the endoplasmic reticulum, and that protein is called SCAP (SREBP Cleavage Activating Protein). And that protein is the key to the whole regulatory system.

This SCAP-SREBP complex, after it is formed, if cells have been deprived of sterols, this complex now is incorporated into vesicles that leave the endoplasmic reticulum and move to the Golgi complex of the cell. And within the Golgi complex the SREBP encounters two proteases, first a serine protease called site-1 protease that clips the SREBP here in the lumen of the Golgi, and it separates the two halves. But this bHLH domain is still bound to the membrane because it has one transmembrane helix. And at that point it's cleaved by a second protease called site-2 protease, which is a zinc metalloprotease, a very unusual membrane-embedded zinc metalloprotease, that actually clips the SREBP within the transmembrane helix and releases the HLH domain so that it can go the nucleus and activate transcription.

When sterols accumulate within this membrane of the endoplasmic reticulum, the sterols bind to SCAP and cause a conformational change in SCAP. And when that happens, the SCAP-SREBP complex no longer leaves the endoplasmic reticulum, it no longer moves to the Golgi. The SREBP that's in the nucleus is rapidly degraded, and no new SREBP is made because the SREBP is trapped in the Golgi, and as a result of that the gene transcription declines. So cholesterol turns off SREBP by blocking the movement, by building up in the membrane of the ER, endoplasmic reticulum, and blocking the movement of SCAP to the Golgi.

Now, how is it that this cholesterol buildup in the ER causes the SCAP-SREBP complex to be retained? Well, this is the most recent part of the story. We've discovered a protein in the ER membranes called Insig, and Insig acts as the anchor for the SCAP-SREBP complex. So when the sterol content of this membrane is low, the SCAP-SREBP complex gets incorporated into these budding vesicles and goes to the Golgi. But when the cholesterol content of this membrane is elevated, SCAP undergoes this conformational change which we can detect biochemically, and that causes it to bind to Insig, and Insig holds it back in the ER.

Studying the SREBP Pathway in Mice

So we've been able to identify the transport protein SCAP, the holdback protein Insig, and the two proteases that process SREBP; all of that has been done through the use of somatic cell genetics in tissue-culture cells. But the question is then does this whole system apply to the liver? And is this system, this SREBP feedback system, really the reason why the LDL receptor gene is suppressed in people who eat high-fat diets? Well, to answer that question in experimental animals we've returned to the SREBP pathway, and we want to study the role of this SREBP SCAP complex in the liver of mice. And to do that we took advantage of a mutational form of SCAP that we had discovered in tissue-culture cells, in tissue-culture cells that were mutated and became resistant to the feedback actions of cholesterol. And the mutation that occurred, that made cells resistant to the feedback regulation by cholesterol was a point mutation within the SCAP gene that substituted an asparagine for aspartic acid at position 443.

When SCAP has this one-point mutation, then it still binds SREBP and it still carried SREBP to the Golgi, but it's no longer turned off by cholesterol. This mutation prevents SCAP from binding to Insig. So SCAP can't bind to Insig, and therefore the SREBP continued to be processed. And in tissue-culture cells that have this mutation, they continue to take up LDL and can't turn off the receptor, and therefore they become with cholesterol. And to test whether this system really works in the liver, we made a transgene encoding—well, first of all before I get to that, let me just show you the biochemical phenotype in tissue-culture cells.

So now this is an immunoblot using an antibody against the amino-terminal domain of SREBP. And it's performed on nuclear extracts from CHO cells in tissue culture. If we take wild-type cells and grow them in the absence of sterols, the SREBP has been processed and we find the—the nuclear form in the nucleus. But if we just add a small amount of sterols, the SCAP no longer moves to the Golgi, and SREBP is no longer processed. But if the SCAP in cells that contain this mutant form of SCAP, they're markedly resistant to feedback regulation by cholesterol.

So we made a transgene encoding of that mutant SCAP, and together with Bob Hammer in Dallas we made a transgenic mouse that expresses, simply expresses, this transgene. Of course this transgene creates a dominant defect because when it binds SREBP, it carries it to the Golgi, so that's a basically—and it can't be turned off by cholesterol—so that's basically a gain of function. And so what we see here is that we get the same phenotype in the livers of these animals that we do in tissue-culture cells; that is, when wild-type mice are placed on a chow diet, we find SREBP in the nucleus; when the animal is fed a diet containing cholesterol, the SREBP is no longer in the nucleus, the LDL receptor gene, and all the enzymes of cholesterol biosynthesis are repressed. But if the animal has this single copy of this mutant SCAP gene, then the amount of SREBP in the nucleus is elevated, and there's a marked resistance to feedback suppression.

And what that does is when the animal is on a—not even fed cholesterol, just because of the overproduction of cholesterol that occurs with this SREBP in the nucleus, the livers of these animals are quite enlarged and they're white here because they're absolutely filled with fat, not only cholesterol but also triglycerides and fatty acids. We call this *foie gras de mouse*.

So this one experiment says that the liver has SREBPs, that the SREBPs are responsible for feedback regulation, and that if you defeat the feedback regulation by putting in a non-suppressible form of SCAP, then you get this overload of the liver with cholesterol and the suppression of LDL receptors. So we believe that this SREBP mechanism is what controls LDL receptor activity. And of course the converse to this thing happens when animals are treated with one of the statin drugs that blocks HMG-CoA reductase. Here you see immunofluorescence of a normal hamster liver, and the SREBP is not in the nucleus, it's outside of the nucleus, because this animal is making cholesterol and keeping its own SREBP partially suppressed. But now if you block cholesterol synthesis by giving this statin drug, you see all these nuclei of the liver lighting up. So we believe that this is the mechanism by which the statins lower cholesterol levels in humans.

Searching for Other Polymorphisms

Well now to summarize my talk and Joe's talk, let me just say that—let's go back to this bell-shaped curve, and this is the bell-shaped curve when looked at through the eyes of a LDL receptorologist. So here's the bell-shaped curve of cholesterol again in Western countries, and the people down here at the low end are people who manage to have very active LDL receptors that keep their plasma LDL level low. Now not all of these people are eating a low-cholesterol diet, some of these people down here are eating quite a high-cholesterol diet. But they obviously have other mechanisms, something's different about them, that prevents their LDL receptors from being suppressed even though they're eating a high-cholesterol diet. On the other hand, we have the homozygous FH patients who are not even on the curve; they have two mutant LDL receptor genes so they can't make any LDL receptors, and they're way off the curve. The FH heterozygotes are stuck up at the top here because they only have one normal gene and one mutant gene. And we believe that all the people in here have been moved out of the normal range because they've eaten this high-cholesterol, high-saturated-fat diet and thereby metabolically suppressed their receptors through the SREBP feedback system.

Now, so as I say, this—the problem from a medical standpoint—point of view—the interesting thing is the splay in this bell-shaped curve. I mean even in our society this, you know, we see people with cholesterol levels of 160, like me, despite the fact that I don't eat a very healthy diet. And on the other hand, there are some people who just look at an egg and their cholesterol goes sky high. So there are clearly a lot of genetic influences that influence our response to this environmental challenge, and some of those genetic influences are very strong, and I think the answer to the whole problem is shown on the next slide.

As far as we could tell, at least what they told us publicly was, that Bill Clinton had, you know, very low cholesterol levels and was the picture of health, despite his pizza diet. So if we can understand Bill Clinton's genes then maybe we'll understand what it is. And maybe by some of the approaches that Professor Lander described to us, we may be able to understand all of the other modulatory genes. One can imagine, for example, that there might be polymorphisms in genes involving the absorption of cholesterol, or in the conversion of cholesterol to bile acids, or in other metabolic pathways that cholesterol can have. There also could be polymorphisms in these regulatory genes; we're looking for polymorphisms in SCAP and in Insig and in the SREBPs themselves that could underlie some of this variability. Our feeling is that whatever it is, it's going to be complex. We know there are other single-gene mutations that can elevate LDL, but most of the garden-variety LDL elevations are going to be due to this high-fat diet, playing on the background of multiple variable genes that make us all different one from the other. And that's the work that we're continuing to do in our laboratory. Thank you very much.

Cornelia Bargmann, Ph.D., University of California, San Francisco, CA

Genes, Behavior, and the Sense of Smell

Introduction by Andrew R. Marks

Andrew R. Marks: So it's my pleasure to introduce the next speaker, and first I want to apologize for the lack of coffee out there. I announced a coffee break and people were asking where the coffee was.

Dr. Cori Bargmann received her B.S. at the University of Georgia and a Ph.D. from MIT. As a postdoc in Bob Horvitz's lab at MIT, she established the worm *C. elegans* as a biological system for behavioral analysis of chemosensation. She is currently a professor of anatomy at UCSF and investigator in the Howard Hughes Medical Institute. She identified the first olfactory receptor for a specific odor, and used the *C. elegans* system to achieve seminal discoveries matching the sense of smell to behavioral responses. Today she will tell us about her remarkable work exploring the links between genes and behavior. Cori, welcome to Columbia.

Genes and the Brain

Cornelia Bargmann: Thank you for that introduction, and am I audible? No. I'm on? Okay, thank you. And thank you very much for inviting me for this celebration of Columbia's 250th anniversary.

The human brain is a wonder of nature. It's the seat of the mind, the seat of our perceptions, emotions, thoughts, memories, and feelings. The human brain is also a biological organ, and like other biological organs, it's built by genes, which are the tools of biology. The object of the brain is to generate behavior, and at some level that means that genes affect the behavior of animals and of humans.

Now to say that genes affect behavior is not to say that genes control behavior; no serious person believes that in any way, but no serious person denies either the fact that genes have a role in generating behavior. William James, the father of psychology, in the nineteenth century had already written that one clear implication of Darwin's theory of natural selection is that the behaviors of animals would also be selected for in ways that were appropriate for their survival. But going from that to understanding how that occurs is a different question.

Now it's been appreciated for a long time, throughout the history of human genetics, that genes can affect the brain. In fact, one of the first genetic diseases identified in humans was a neurological disease, phenylketonuria, identified in 1934. And this disease is associated with behavioral issues, late social skills, mental retardation, and various movement disorders, and it's related in a straightforward way to an inborn error of metabolism, a defect in handling the amino acid phenylalanine, one of the eight essential amino acids in the human diet.

Now phenylketonuria, this disease, remains a paradigm for human genetics in a whole set of ways and helps to illustrate—this is the pointer?—great—and helps to illustrate principles of thinking about how genes affect behavior in ways that remain to this day. First of all, the gene does not act directly on the behavior but rather through a set of intermediate steps. The gene is interpreted in the context of a cell. In the particular context of this disease the mutation phenylalanine hydroxylase leads to smaller and fewer neurons. This in turn leads to altered function of the brain and brain circuits, and these altered circuits lead up to a person who has altered behavior. And so understanding genetic influences involves understanding all of these steps along the way in the generation of the behavior.

Now the success of genetics as a tool to find this gene also led to a success in terms of the ultimate health and well-being of the patients affected with this disease. As a disease of amino-acid metabolism, PKU can be controlled by a low-protein diet, and a low-protein diet remains the treatment of choice for this disease to this day. And the success of genetic approaches has also been quite prominent in studying all kinds of neurological diseases, neurological diseases being distinguished from psychological diseases or psychiatric diseases by the fact that neurological diseases result in something you can see. And the major neurological diseases all consist, at least in part, of diseases that have genetic origins or genetic components.

Genes and Human Behavior

So to what extent do genes affect behavior, however, as opposed to the structures of the brains, and the viabilities of nerve cells? And what is the evidence that genes are involved in behavior, specifically human behavior? Probably the strongest evidence for a genetic component to human behavior results from studies of identical twins who are essentially genetically identical at all loci, and a comparison of the traits of these twins with other siblings or first-degree relatives, who share only half of their genetic material. Now twins are raised in identical environments, but much of the power of this approach has been the ability to extend it to twins that are reared in different environments, identical twins reared apart, exemplified here in the eyes of the entertainment industry, to illustrate that twins reared apart might share only some and not all of their characteristics. So in this case both twins grow up to be highly paid movie stars, but only one of them is the governor-elect of the state of California.

And in fact for many different kinds of behavioral and psychiatric traits, it turns out that there's a substantial but by no means overwhelming genetic component, that about 40 percent of the sorts of quantifiable behavioral assays, personality indices, psychiatric diseases, can be ascribed to genetic causes. And this has been very notably the work of many groups, but particularly Tom Bouchard and his colleagues at the University of Minnesota.

So this problem is interesting not only as sort of an intellectual problem, but in terms of an enormous amount of human suffering, human disease, disability, lost work, economic indices. Depending on exactly how you count, seven of the top ten reasons that people take time off from work because of disability are directly related to psychiatric disease, including depression, drug and alcohol abuse, obsessive-compulsive disorder, and schizophrenia. So understanding something about the genetic bases of these diseases is important, and in fact the genetic risk for these diseases is quite substantial. If you look at the genetic risk associated with either an identical twin or a sibling for these common psychiatric diseases, it's comparable or higher to the genetic risk associated with type 2 diabetes, which is clearly recognized to have at least a partial genetic origin, as Eric Lander just told you. So these genetic diseases, however, are still waiting for their PPAR-γ, the specific molecular information that would lead you to be able to design some sort of an intervention to help the individual.

So to know that these diseases have a heritable component is at some level to know nothing. It doesn't give you any information that you can use to understand more about the underpinnings of the disease, and it doesn't give you any information that you can use to intervene in a useful way. And so what we'd like to understand about genetic variation and behavior is how many genes are involved, what kinds of molecules and pathways are involved in generating behaviors, what kinds of changes distinguish one kind of individual from another, and how do those changes affect behavior. And somehow this information is imbedded within the thirty thousand genes of the human genome. But we don't really know where to look.

This is a pie chart from one of the papers about the human genomes dividing genes up by category of function, and I draw your attention here to the fact that the largest piece of this pie, about 41 percent of all genes, fall in a category called "molecular function unknown." So despite what we do know about human genes, there's a great deal that we do not know. And the tools that have been valuable for reaching and understanding here are really tools that were conceptualized for the first time for animals by Thomas Hunt Morgan and his colleagues working at Columbia University, which was the idea that simple animals would represent tools that could be used to understand general laws, initially of chromosomes and heredity, later of developmental biology, and ultimately of behavior. And Morgan himself was interested in the question of behavior, although he did not address it in his own career.

And we now have a quantitative sense of what Morgan appreciated qualitatively, which is that most human genes are shared with other organisms, that the number of genes that are specific for humans alone is going to be extremely small, and even the genes that are specific for vertebrates represent only twenty-some percent of all the genes in the human genome, that the overwhelming majority of genes are shared by all animals, and many even shared by unicellular organisms, and that these much simpler and genetically tractable organisms would represent a place to create the power to understand these problems.

Circadian-Rhythm Genes

Now Morgan ultimately moved to Cal Tech and the genetic analysis of behavior really travels in pretty much of a straight line to Cal Tech through the work of Seymour Benzer and his colleagues, who used Morgan's experimental organism, the fruit fly *Drosophila melanogaster* to make the first inroads into understanding the genetic basis of behavior. And the behaviors that they chose to study are behaviors that have circadian rhythms, patterns of behaviors that fruit flies exhibit over long periods of time. So flies are diurnal, they buzz around and do their feeding and mating and carrying on during the day, and during nights they have low activity levels. These chits here represent some sort of activity, flying around of a fly. Many animals have circadian rhythms. The interesting and remarkable thing about circadian rhythm is that they are not directly controlled by the environment, but are in fact internally generated by the animal's nervous system. So, for example, if you take an animal and shift it to constant darkness or constant light in the absence of any information about the day, the animal will continue on a twenty-four-hour cycle involving high levels of activity punctuated by low levels of activity.

And Konopka and Benzer in 1971 reported the first genetic analysis of this complex, long-range, innate behavior by identifying mutations in which this periodicity of the animals when shifted to an environment without cues was disrupted, and in particular they identified mutants that had short-day rhythms, as short as 18 or 19 hours, long-day mutants that might be as long as 28 hours instead of 24, and mutants that were completely lacking in circadian rhythms. And remarkably all three of these classes of mutations could be isolated in a single gene, implicating this gene as having a central function in the generation of these regular behavioral rhythms. And the name of that gene is *per*, and molecular studies and further genetic studies initiated, starting from that particular starting point, has led to a molecular understanding of a network that generates behavioral rhythms in the fly.

And that network consists of a gene regulatory network that consists of genes, including *per* and several other related genes, which regulate each other's synthesis positively and negatively in such a way that the levels of these gene products rise and fall with a regular twenty-four-hour oscillation. And the names of these genes are *per*, *tim*, *timeless*, *clk*, and *bmal*. The result of their positive and negative regulations is that these gene products oscillate on different cycles, and therefore different ones dominate at different times of the day. And just to sort of simplify what that entails, a gene called *clk* is active through most of the day, and the genes *per* and *timeless* are active through most of the night. And so the alternating activities of these two gene sets account for the differences in activities and many behaviors that insects exhibit.

Now the power of this approach was evident to Joe Takahashi and his colleagues at Northwestern University in the late 1980s and early 1990s. They wondered if you could take the same kind of single gene-forward genetic-mutagenesis approach to comparable behaviors in more complex animals, such as rodents. And indeed Joe in a genetic screen identified a mouse mutant which in the same kinds of activity logs when shifted to constant darkness has ever-increasing day lengths, in other words, had a long circadian rhythm, and which could in addition, when homozygous, be shifted to a completely arrhythmic pattern. So mice, like flies, have endogenous circadian rhythms. They're the opposite of flies in that they run around at night and sleep during the day, but these kinds of activity monitors look quite similar between vertebrates and invertebrates, and the mutations look remarkably similar in terms of the kinds of effects that they have.

The mutations are also remarkably similar in terms of the molecules that they affect. And so the central circadian oscillator of mouse circadian rhythm is essentially identical, with a few modifications, a few duplications and a few bells and whistles, to the central circadian oscillator that accounts for fly circadian rhythms. And so it's the same network of oscillating genes, in fact the *clk* gene, one of these genes, that was first identified in the mouse, later in the fly. The others were first identified in the fly, later in the mouse. One is able to move seamlessly between these experimental organisms, really validating Thomas Hunt Morgan's idea that you would be able to study genetic processes in simple animals to understand more complex animals. And the same ideas of these oscillating gene products that are dominant at different times of the day explains the behavior of the mouse.

Genes and Human Sleep Disorders

The final example and the completion of this idea of genetic conservation and the role of genes and behavior is indicated by a rare group of human patients who suffer from something called "advanced sleep-phase syndrome," and these patients can be shown in normal conditions to have very unusual behaviors compared to most individuals in our culture. They wake up at about three o'clock in the morning when left to themselves and have an enormous amount of difficulty staying awake past six o'clock at night. And they never go to any good parties, they never get into any of the cool clubs, they're absolutely wretched and they complain to their doctors about the effect on your social life of essentially not having the free evenings, which are when in our culture you have most of your social interactions. And sort of in order to characterize these further, a set of ASPS patients volunteered to be placed in conditions where they had no external information about day and light cycles, so no television, no clocks, no radios, and their activities were monitored. And when these patients were monitored over a period of days the shift in their activity cycles, toward the left side, indicates that they had shortened circadian rhythms, that their clocks actually run fast, on the order of twenty hours instead of twenty-four hours. So they're constantly pushing themselves to try to stay on a schedule which in fact wants to shift earlier and earlier every day. And this behavior pattern is virtually indistinguishable from rodents or flies that have short circadian rhythms, where you have shifting patterns that move to earlier and earlier periods of time.

Louis Ptacek and Ying-Hui Fu at the University of Utah identified a mutation associated with this advanced sleep phase syndrome in these patients and discovered that this mutation is precisely a mutation in the human ortholog of the fly *per* gene; that is, the identical gene that affects circadian rhythm in the fly is responsible for this behavioral difficulty in humans. And in fact it's found in an exact residue which is known to regulate the cycling and the turnover, and therefore the rate of oscillations of these gene products. So the conclusion of this is that it's possible to understand a fairly complex long-range behavior in terms of molecules that are conserved through evolution that organize whole groups of behaviors, in this case transcriptional molecules.

And I have to say one of the reasons that this was such a powerful model was that sleep is a fabulous behavior, so one of the difficulties for understanding behavior is the difficulty of scoring it. You know when organisms are asleep, you know when they're awake. They don't have to tell you, they don't have to write it down on a form. And as a result sleep genetics is something that of the behavioral systems we probably know more about than other kinds of behavioral disorders.

So a second example of that is the syndrome of narcolepsy / cataplexy. This is a very unusual disorder, it's quite rare, it's observed in human patients, and it represents a fragmentation of the separation between sleeping and waking behavior, such that behaviors appropriate for sleep creep into everyday waking life. These patients are sleepy—they fall asleep early, they enter REM sleep prematurely, and can even enter REM sleep while they're awake and therefore have hallucinations. They can have a loss of muscle control under excitement and a paralytic response, and that's the cataplexy response. Again, paralysis is appropriate for sleep but not for waking states. And this is a human disease which has been mapped to susceptibility locus, for this has been mapped by linkage mapping, but at some level this was a failure of genetics. And the reason it was a failure of genetics is that the linkage mapping identified a particular genotype at the HLA, the major histocompatibility locus, as the source of narcolepsy / cataplexy risk. And so on the one hand that tells you something about what gene is responsible, but on the other hand it tells you nothing, because what it tells you is that there's something, probably some sort of immune disorder, probably some sort of autoimmune destruction of something that leads to this sleep disorder and this curious behavioral fragmentation. But it doesn't tell you how to go about finding out what exactly has been lost, and how exactly these different manifestations are coupled to each other.

The Mechanism Behind Narcolepsy / Cataplexy

So the same group that worked on the human genetic linkage, Emmanuel Mignot's lab at Stanford had a backup plan in case that didn't work out, and their backup plan involved a genetic analysis of the dog. This dog is Prancer, is a Doberman Pinscher, and Prancer is lying on the ground in a state of cataplexy, because Prancer, along with a few but not most Doberman Pinschers, has a genetic syndrome indistinguishable from narcolepsy / cataplexy syndromes in humans. So this has been observed several times in several different dog breeds. It's not something that every animal in the breed will have, just a few individuals have it, but since these are animals that you can breed with one another you can do genetic mapping, and you can try and find out what genes are involved in this disorder, and ask if they're related in some way to the genes involved in the human disorder.

And so the Mignot lab working in dogs, and actually in a parallel study at the Yanagisawa lab working in mice, identified a particular signaling mechanism as essential in the generation of this narcolepsy / cataplexy syndrome. And that mutation was in a G-protein-coupled receptor called the hypocretin-2 receptor, also known as the orexin receptor, whereas the mouse disease was in hypocretin orexin itself, the ligand, the small peptide ligand for this receptor. So these are molecules that are expressed in nerve cells. It was known that they were expressed in the brain, it was not known what they did; in fact, they were somewhat misnamed with the idea that they might have something to do with feeding, whereas the genetic results suggest that they have something to do with the sleep-waking regulation.

Now neuropeptides represent a very interesting kind of molecule to think about, genes that might be binding together groups of different responses in the form of coherent behaviors. So you can contrast neuropeptides with classical neurotransmitters in the brain. Neurons signal to one another using both classical transmitters and neuropeptides. Classical neurotransmitters generate rapid, precise, local transfers of information involved in rapid, precise processes like perception and thought. Neuropeptides can be released from the same neurons, but they work in a very different way. They act over much longer time scales of seconds to hours, rather than milliseconds. They can act over a distance, many different groups of neurons or over circuits. There are many different neuropeptides and many different neuropeptide receptors, dozens or hundreds, and they're all expressed in very specific patterns and specific groups of neurons. And they're slow action in their ability to act over a distance. It's quite interesting in terms of thinking about how you would bind together a set of different behaviors in regulating, for example, the transition between sleep and waking.

Further insight into this came from asking where orexin was expressed in the brain. It turns out that in the human brain there are about only 2 thousand neurons out of the billion or so neurons of the human brain that express orexin, this peptide. They're localized in the hypothalamus, which is involved in regulating many instinctive behaviors, hunger, thirst, sexual behaviors. And this particular set of neurons, once they were characterized, could be looked at carefully, and they were found to project to many different regions of the brain that were already known to be involved in sleep and arousal. So there many parts of the brain that were known to be regulating sleep and waking, but what wasn't appreciated until the genetic study of narcolepsy was that these parts of the brain were essentially bound together by a master control region in the hypothalamus that expressed a particular neuropeptide that communicated with all of them, and gave them the potential to coordinate with one another.

So this gives an insight into how the brain works that was not understood prior to the genetic insight. It also enables an insight for going back and understanding the autoimmune basis of the human disease, because the human patients who suffer from narcolepsy have a specific loss of the hypocretin-containing neurons in the hypothalamus, associated with the gliosis. Compared to normal individuals it suggests that they're undergoing some kind of an immune destruction. So the use of genetics in multiple organisms was actually powerful for understanding something that at one genetic level in the human was already understood, but which required the other organisms to go back and understand it at the level of the brain and the circuit, these intermediate levels of going from a gene to a behavior.

Species-Specific Social Behaviors

What other kinds of behaviors are interesting to study, what other kinds of behaviors are interesting to think of from a genetic basis? Social behaviors represent behaviors that are innate to species and shared by all individuals within a species often. All species of animals have to be able to recognize other individuals of the same species, for example, so that they mate with the right individuals, but how elaborate their social structures are in addition to this varies enormously between different species. Social behavior has different advantages and disadvantages, and it's evolved and been lost many times independently during evolution. Because of things like game theory, there's times that it's advantageous to cooperate, there's times that it's advantageous to go it alone, and different species will exploit these different strategies to different extents. So on one extreme you have the elaborate social constructs of humans or of social insects like ants or bees. At other extremes you have animals that really only get together to mate, but in all cases these responses are innate, they involve recognition, and they're relatively fast evolving. So one example of the fast evolution is that species that are quite closely related can exhibit different social strategies. Among the big cats, tigers and lions are closely enough related genetically that they can mate and have offspring, but tigers have a solitary life pattern—in Brooklyn—and lions spend their entire lives in a social group.

So what kind of genetic changes might occur in the mouse species or individuals to be different from one another? You can actually break this problem down further to think about differences between individuals within a species. And one slightly artificial but intuitive example of differences in social behavior is represented by the behavior of different breeds of dogs. So dogs have been bred by humans for particular traits to work with humans, and many of these traits are quite specifically and explicitly behavioral traits. So, for example, golden retrievers were bred by English hunt clubs to be able to work during hunting with any number of the hunt clubs, and so these dogs are relatively easy to get along with, easy to work with, don't have strong preferences for particular individuals, they basically have an extroverted sort of character. By contrast, German shepherds and other shepherd dogs have been bred specifically to work with one individual, and in particular to work with that individual to protect the sheep from thieves or from other predators, and so they tend to encounter new individuals with some degree of skepticism and mistrust, and instead form very strong bonds with a single individual, so strong actually that German shepherds are no longer used as seeing eye dogs because they form such a strong bond with their initial trainer that they have difficulty reforming a bond with another individual.

So these are traits that you can expect to be reliable among golden retrievers and among German shepherds, but we have no idea what the genetic basis would be that would lead to their differences in social strategies. And so we'd like to understand that and take this idea of Morgan's of working with a simple animal to see if we can use genetic tools to figure out what kinds of molecules or pathways could be involved in this sort of behavior pattern. And the simple organism that we work with is the soil nematode *C. elegans*, a millimeter long, which feeds on soil bacteria. As Mike Levine pointed out to you it doesn't have particularly fascinating developmental patterns, but I would argue that it has fabulous behaviors and—at least as good as flies buzzing around. A little sibling rivalry here.

Social Behaviors in *C. elegans* Strains

And different isolates of *C. elegans* have been isolated for many different places in the world and grown in the laboratory. And this is the standard laboratory isolate that Sydney Brenner introduced to the world. It's a British strain called N2, and if you place it on a lawn of bacteria in the laboratory it observes a characteristic feeding pattern in which the animals pay no attention to each other whatsoever. And this we call a solitary feeding pattern.

Now by contrast, other isolates of *C. elegans* from different places, such as this strain, RC301, isolated by Randy Cassada in Germany, exhibited a different feeding pattern in which animals gather into groups of dozens or hundreds of animals. So there are bacteria throughout this field, but the *C. elegans* will not feed on the bacteria until they have first gathered into a feeding group, and only then do they exploit the bacterial source. And these traits are reliably different from one another, and they've both been observed many times. So about a third of all strains are solitary, and these include strains isolated from England and from various places in the U.S. About two-thirds of all strains exhibit the social feeding pattern, and these include strains from continental Europe, Australia, Hawaii, as well as the continental U.S. And these two different kinds of nematodes appear to be able to coexist with one another. Both a solitary and a social worm were isolated from the same compost pile in Pasadena, California, implying that these two different strategies are both under some sort of positive selection which provide advantages to the animal under different circumstances.

So the power of this experimental organism is the simplicity of its lifestyle and the rapidity of its generation time. These animals grow up in about three days, and you can use genetics to understand what the basis is of the genetic difference between these true breeding variations.

And what Mario de Bono found when he was studying this is that a single gene, the neuropepide receptor *npr-1*, can explain the difference between social and solitary strains. So this gene, like the hypocretin-2 receptor implicated in narcolepsy, is a neuropeptide receptor, a molecule expressed on neurons that responds to peptide ligands. *C. elegans* has about a hundred neuropeptides and about a hundred neuropeptide receptors, again a complex network of these molecules that are expressed on neurons in specific patterns. And what Mario found was that this molecule was reliably different between the social and the solitary strains, that the solitary strains invariably, in five out of five cases, has a valine, a particular residue, implicated in the strength and specificity of G-protein coupling, whereas the social strains invariably, in twelve out of twelve cases, had a phenylalanine residue at that same location. And this was true regardless of the original site of isolation of the animal, there was a perfect correlation of these amino acid residues.

And Mario has gone to show that these residues result in functional changes in the receptors that are detectable in their G-protein-mediated responses to their neuropeptide ligands. So different ligands, the solitary isoform represents a high level of activity of this receptor that enables it to respond strongly to its ligands and to respond to a broader variety of ligands. The phenylalanine or social version represents a low level of activity of this receptor in which a higher level of ligand is required to stimulate this activity. So the activity of this gene is to stimulate solitary behavior when expressed at high levels, or at high levels of activity.

So what's the argument that this gene is not only correlated with but central to these behavior patterns? In fact, that argument is based on genetic manipulations that you can do in either the solitary or the social backgrounds that indicate it functionally in these different outcomes. So the particular 215 valine allele is necessary for solitary behavior. If the npr-1 gene is inactivated and it is the only gene inactivated in a solitary strain on animal, these animals are fully converted to social feeding behavior, which is actually a complex constellation of behaviors that include social feeding and several other responses to the environment.

Conversely, this allele is also sufficient for solitary behavior in an otherwise social background. So if we take different social isolates from Europe or Australia or the United States and introduce into them a transgene which contains only this gene in the allele that corresponds to the solitary version, these strains are immediately and completely converted to the solitary feeding pattern. So based on the arguments of genetics, this represents a necessary and sufficient argument that indicates that this gene accounts for the difference between these two strains.

Brain Circuits in *C. elegans* Social Behavior

So what I've just told you is that when we grow these animals in the standard lab environment, the level of activity of this gene, whether it's low or high, is sufficient to determine the behavior of the animal. So a low level of the activity of the gene corresponds to social feeding, a high level to solitary feeding. But having said this, at some level again we've said nothing. We've said that there's a gene and we've said that there's a behavioral output, but we've missed the stages of translation in between, the understanding of the neurons and the understanding of the circuits that will enable us to see what it is that these behaviors truly represent.

And so David Tobin and Jesse Gray, two graduate students in the lab, have tried to understand that, in particular using this gene and other genes as tools to go into *C. elegans* and dissect the circuits in the same way that the orexin gene was done, to dissect the narcolepsy and sleep and waking circuits in the human brain. And what they discovered was a set of different sensory neurons and sensory inputs were actually essential for generating these different kinds of behavior. And in fact, those sensory inputs could be modified, at this point we can modify them at will, in ways that lead us to believe that we understand something about the genetic nature of this behavior.

So if we move animals from a standard lab environment to a high-stress environment—and there are many different ways of doing this that involve food deprivation or various kinds of environmental stressers—the animals exhibit social feeding. And in fact in a high-stress environment, animals, regardless of their phenotype at this neuropeptide locus, will exhibit a social feeding pattern.

Conversely, if the lower the level of stress in the animals' environment—and again there are various kinds of sensory manipulations that we use based on our understanding of this sensory circuit, and the identification of the olfactory neurons and other sensory neurons involved in it—we can make animals solitary feeders, regardless of the genotype at the npr-1 locus. So this experiment, or this series of experiments, tells us some things rather generally about the way that genes interact with the environment to generate behavior, that first of all the genes do not determine that ability to generate the behavior, because animals at either genotype are capable of exhibiting solitary feeding behavior or social feeding behavior, but rather what the genes are doing are regulating the probability that the behavior will be generated in a very specific set of environmental conditions. And second, they give us some sort of an idea, which we have been able to explore further using these exact environmental conditions, about what the nature of the behavior is that we're looking at. And that is in fact that the social behavior that we're studying is a behavior that is induced by stress. So social behavior, based on evolutionary theory, can carry a variety of advantages to an animal: it can carry advantages for feeding, it can carry advantages for defense against predators or toxins, it can carry advantages for mating and reproduction. But social behavior does come at a cost: you have to share your food with other individuals, and you're more susceptible, for example, to infectious diseases which occur at high density. And so the decision whether to engage in a social behavior will be regulated by other kinds of stimuli that tell the animal whether it's appropriate. And *C. elegans* at high stress levels, the animals engage exclusively in social feeding; in low stress levels, they're more exploratory and engage exclusively in solitary feeding. And what we believe is that the genetic polymorphism in fact represents a polymorphism in the response to moderate stress levels, because it turns out that our laboratory conditions are actually a little bit unnaturally and moderately stressful to the animal, and that in that particular environment the animals that have the social genotype are cautious and assume that the moderate stress is stressful enough that they respond as though it was a high-stress environment, whereas the animals that have the higher activity or solitary version of the gene are bolder and they exhibit the behavior appropriate for a low-stress environment.

And this kind of a behavioral axis between cautious and bold is one of the behavioral axes that you see varies between individuals in almost all species. There's almost always a range of behavior that can include these two extremes, because they'll be valuable depending in a very online way in the particular environmental conditions. And so you see, for example, in subspecies of fish in different lakes if there are predators within the lake, very rapidly you start to select for cautious fish, and if there are no predators in the lake, very rapidly you start to select for bold fish. But you can imagine that the introduction of a predator would immediately switch these things back. So this is a kind of behavioral axis that would toggle back and forth easily in evolution.

So one of the reasons that we find this result engaging is that we think that perhaps there are ways of thinking about social behaviors and genes involved in the behaviors that might extend to multiple systems. Work of Tom Insel and his colleagues has studied a different level of social behavior, and that is the evolutionary components of social behavior, in particular the differences of social behaviors in mammals. Now all mammals are social because mammals have to raise their young and therefore interact with other individuals, at least partly, in their life, however there are many different social strategies in mammals. About three percent of all mammalian species are monogamous, whereas the other species tend to be polygamous or what's called promiscuous, mating freely and not maintaining long attachments. And sometimes very closely related species will show these differences. So voles, which are closely related to rodents, very similar individuals can exhibit monogamous, strongly pair-bonded forms, or polygamous, promiscuous individuals that only get together to mate, and in fact the mothers don't even stay around the pups very long. And Insel and his colleagues have over the years implicated a set of neuropeptides strongly in the variation and the creation of these different behaviors.

So in this case it's a different neuropeptide receptor, in this case the vasopressin V1 neuropeptide receptor. And the activity of this receptor is essential for the monogamous behavior of male monogamous voles, blocking its activity blocks their appropriate behavior, and moreover the expression of this receptor in the brain is very different between the monogamous and the polygamous species of vole.

So this expression pattern, this molecule that belongs to the same set of molecules, and the antagonist experiments provide evidence that these molecules are actually engaged in social behaviors. A transgene taken from a monogamous vole and introduced into a mouse which has more of a polygamous lifestyle appears to stimulate affiliative behavior in mice, suggesting that, as in the case of the neuropeptide receptor system that we studied in worms, this neuropeptide receptor may actually be able to modify behaviors and bind together groups of behaviors in a way that correspond to the affiliative behaviors of rodents.

Research on Human Psychiatric Diseases

I'd like to close by talking a little bit about genes and complexity, and returning to human disease and the importance of genetic systems in understanding how diseases arise. And I would like to start with some of the things that we do know about human psychiatric disease.

Probably the strongest influence on thinking over the past twenty years or so about the fact that psychiatric has a strong biological component has come from the discovery of effective drugs that in many cases, somewhere between 20 and 50 percent, can actually treat major affective disorder, particularly depressive disorder and anxiety disorder. And these are the so-called SSRIs or antidepressants, known by their names of Zoloft, Prozac, Paxil. And the role of these molecules is to increase the effective levels of the neurotransmitter serotonin. Serotonin is a small molecule that sort of acts somewhere between a classical neurotransmitter and fast neurotransmitter and a neuropeptide neurotransmitter in terms of the fact that it can act on slow-ish time scales and it can act across distances. The involvement of the SSRIs in affective disorders like depression and anxiety has really provided evidence and very strong thinking in people's minds that in some way serotonin is involved in these affective disorders.

Now the exact molecular target of these drugs is known, and it's worth saying that these targets, that this whole link was discovered completely by accident based on side effects of medications used for cardiovascular patients, and that if we understood more about these diseases it would be quite possible that we would understand more about ways of designing effective interventions for patients.

But although this molecule is known, this molecule has been sequenced in many, many patients, the link between the serotonin reuptake transmitter, the molecule affected by Prozac and this family of molecules, and the genetics of human affective disorder was extremely weak and discouraging. And this just gives a sense of how difficult it is to do human genetics, because this link is now starting to be made but it required an enterprise of immense proportions, the so-called Dunedin Multidisciplinary Health and Development Study. And what this represents is a study done in New Zealand initiated by Silva, carried on by Poulton and colleagues which has been going on for thirty years, where a thousand children born in the same hospital in New Zealand in 1972 have been tracked, starting from the age of 3 to most recently the age of 26, and they've been monitored very closely for all kinds of health outcomes, for environmental influences, for behaviors, and also for genotypes. And using this very large cohort and doing the kind of effort that it takes to understand human genetics and human behavior it's actually been possible to demonstrate not only that the serotonin reuptake transmitter, the target of Prozac, is involved in depression and depressive disorders, but that it's involved in a very specific way that involves an interaction with an environment.

There are two major genotypes at this locus, and this represents truly natural variation; about half of all individuals have one genotype or the other in the Caucasian populations. And if you look at the risk of a depressive event, depending on the genotype, both genotypes actually have plenty of depression. These individuals between the ages of 21 and 26, about 17 percent had one depressive episode, so this represents about two hundred events in these thousands patients. So the gene itself is not instructive. Both kinds of people got depressed, and that fits the lack of correlation observed in many studies.

But what's remarkable is what's seen when looking at the S-genotype when actually tracking these individuals for what was happening in their life over this period of time. And it turns out that individuals that have an S-genotype when confronted with multiple traumatic and stressful events, such as poor health, the loss of a job, the death of a family member, a divorce—if they had three or more events of this sort, their risk of affective disorder shot up in a way that those of L-genotype did not. And in fact if you work through these numbers, it appears that about 25 percent of depressive disorders within this cohort can be explained specifically by the interaction between the S-genotype and a series of environmental insults. And the depth and intensity of this study is wonderful because it really lets you exclude many trivial explanations for this result. For example, you can ask what's cause and what's effect by looking at depression between the ages of 18 and 21, and then depression at later ages, and figuring out when the events happened and when the depression came, so you can figure that out.

So I think is a very powerful and very helpful insight into the real basis of human disease, again, the importance of neurotransmitter system in behavioral disorders. But at another level we know almost nothing, and there's a lot that we still don't understand about serotonin means in disease, like where, when, and how it acts. And so for this I think the importance of the animal models and of genetic systems for understanding behavior and psychiatry is as great now, or even greater, than it was. And I want to close by just mentioning the work of René Hen and his colleagues, to end at Columbia University as I began, in this study.

René's work on serotonin and particularly on anxiety-related disorders in the mouse has helped to localize, for example, the brain regions and circuits that appear to be affected by serotonin, including the hippocampus in the cortex, in generating affective disorder like syndromes in the mouse. And the timing of these events, which is fascinating—at least one receptor appears to act very soon after the mouse is born—for its adult behavior, its adult level of anxiety. It reflects the activity of this gene very early. So this represents a place where the classical ideas of psychiatry, of formative early childhood experiences, and the ideas of genetics may come together. And there's even hints from René and others about the mechanisms of some of these pathways, that they may act, for example, by stimulating neurogenesis of neurons.

And when I come back for Columbia's 300th anniversary I hope that there will be answers to many of these kinds of questions as well that will emerge from these studies. Of course there are many genes still left to go, and having one insight into depression leaves much of human psychiatric disease unknown. There's more of the brain to know about, there will be much to learn about how genes and the environment and the brain and the environment interact. It's increasingly clear that experience alters gene expression, just as the fat in your diet changes the expression of LDL receptors in your liver, the experiences in your life change the expression of significant genes in the brain and its biological properties thereafter. And ultimately we can hope to understand more than disease, but rather how genes make us what we are, and how we in turn transcend them to become what we can become.

Thank you.

Roy M. Anderson, Ph.D., Imperial College of Science, Technology and Medicine, London, UK

The Evolution of Pathogen Genomes

Introduction by Andrew R. Marks

Andrew R. Marks: So, now for the last speaker in this session. Roy Anderson is professor of infectious disease epidemiology and head of the Department of Infectious Disease Epidemiology at Imperial College, Faculty of Medicine, University of London. His seminal studies have examined the transmission, evolution, and control of infectious disease agents in human communities, animal livestock, and natural ecosystems. His work has been the basis for designing childhood immunization programs, strategies for controlling the spread of AIDS, and control of Mad Cow Disease in Great Britain. Today he will discuss how genetics in pathogens and their hosts, which are us, determine the mortality of the major infectious diseases, including malaria, AIDS, TB, and the emerging disease SARS. Welcome to Columbia.

Origins of Population Genetics

Roy M. Anderson: Well thank you very much. It is a real pleasure to be here, and a very warm happy birthday on your 250th anniversary. Coming from London University, I very much understand the problems of being a university in a very, very major city with all its disadvantages in one sense, but the many, many advantages of the sort of liveliness and the cultural environment that New York presents.

I was very surprised to walk into your university this morning and see many flags with the royal crown on, and clearly a little bit of thought, thinking about history, explains that. But I was also drawn to think that perhaps this had something to do with "Genes and Genomes" because clearly royalty throughout Europe are quite an interesting case for study in terms of inbreeding coefficients.

My association with Columbia is very, very recent. I've had the privilege to serve on the advisory board for your new Earth Institute, under the directorship of Jeff Sachs.

And what I'd like to talk to you today about actually, I was asked to address the problems of population genetics and genetic epidemiology, I'm going to very briefly refer to that, but I want to move more to the genomes of pathogens and how one studied genetic diversity, and in particular I'm going to talk about two problems which are new and evolving problems. One is the description of genetic diversity and understanding its linkage with biological characteristics, and the second one is a more technical area, it's the beginnings of understanding how we can simulate genome evolution for simple organisms within computers.

If we think about the origins of population genetics, really these were laid down—the techniques and much of the statistical methodology and indeed mathematical methodology—was laid down in the first half of the twentieth century, particularly by R. A. Fisher at Cambridge, but also very importantly by C. C. Lee here in the United States. The term "genetic epidemiology" first emerged somewhere between '54 and '70 —there is some debate about who was the person who first coined it—but the second half of the twentieth century really saw the very, very rapid growth in human genetics, and much of the early growth was related to observational studies, family studies, and the development of DNA markers.

We haven't—it's not the sort of place to talk about the statistical methodologies, but there are some very, very important developments in this area. We've heard repeatedly today about you can let the computer do this, that and the other. Well that's not strictly true. Some people have to develop rather clever algorithms containing a set of biological assumptions to instruct the computer to do x, y, and z, and there's quite a lot of theoretical developments in that area when you are dealing with huge linear sequences.

Ah, this is an interesting concept—I hope this is not persistent. This is the translation from a PC to a Mac. There is a picture there which was a copy of the front of the *Journal of Science*, and it was in 2001 and it was essentially a cartoon about different animals talking to each other at a cocktail party, saying, "Well, have you had your genome sequenced?" What I want to mention today is that we've clearly been focused on the human genome, but it's very, very important to remember that the genomes of most of the important human pathogens have now been sequenced. Sydney Brenner mentioned phage. If you think about very shortly after that, a series of very small viruses were sequenced, one of the first was foot-and-mouth virus, and if you think about the SARS epidemic recently, something like two weeks after the identification of the etiological agent, the four-genome sequence is out for this new virus. The only organisms where the genomes are not available at present—but they will be in the coming two years—are the larger pathogens such as the protozoa, the *falciparum* malaria, and the helminths, the worms, which have very, very big genomes. But I'm sure that these will tumble out in the coming years. And just as for humans, these genomes, we have complete sequence, dense marker maps of the genome, and increasingly sophisticated methods of analysis.

Approaches to Genetic Epidemiology

Now what is genetic epidemiology? I mean in some senses it's very obvious; it's the study of the role of inherited factors in disease etiology, and for infectious diseases, for example, this involves two genomes: that of the host laid upon that of the pathogen, and very complex and subtle interactions happen there that we understand very, very little about at present. In the early years this was largely concerned with family-based studies, or monogenic disorders, but if you think about recent years you think about increasingly a focus on complex multifactorial diseases; arthritis is one, and there are many others. I'm not quite as optimistic as others that we will be able to make rapid progress in this area. If one has six or seven or more genes involved, then genome scanning and linkage studies are going to be exceedingly difficult. In the United Kingdom recently the Medical Research Council approved the funding of a project called BioBank, which is obtaining DNA from half a million people, adults, and then following in detail their clinical and medical histories over time. And the debate really centers on, is that sample size big enough if you've got a very complex disease with many genes involved? And that's far from clear, very far from clear.

If you're interested in infectious diseases, you look with fascination and some wonderment at the increasing number of noninfectious-disease problems that are beginning to be linked with an infectious event or a persistent infection. We have the extraordinary example of gastric ulcers and *Helicobacter*, but more recently some really interesting other examples, one of which is atopic, in particular asthma and the occurrence of a past exposure, which leaving serological markers of Hepatitis A virus. Much remains to be unraveled here, but one suspicion is that infectious agents—particularly something like asthma and arthritis, which involves the immune system—infectious agents may have a very, very important role to play.

Now the two strategies to these linkage problems, there is the sort of brute-force approach of genome-wide screens, using markers, etcetera. An alternative strategy to the genome screening is that of candidate gene studies based on a knowledge of a gene product involved in a particular disease. And it's important to remember in history that many more genes have been discovered by identifying the protease involved in the disease and working backwards to determine the genes that encode. So this in Sydney Brenner's term was sort of going backwards, but it is important to remember that that is where the majority of our information has come from, and it requires a biological understanding of pathogenesis of some detail to lead you to that particular gene target. And clearly in both of these areas the new technologies offer extraordinary scope for rapid progress, as we've seen so clearly today.

If we think about genetic epidemiology and its coming out of population genetics, familial aggregation of disease, the backbone of this approach, but I should stress that increasingly huge population-based studies—where you have differences in the instances of disease, controlling for other environmental factors—are playing an increasingly important role. And in fact, Cori in the last talk talked about this example in New Zealand. Iceland, many of you will be familiar with, a population of a quarter of a million, and a company DeCode, has been involved in studying and extracting genetic material from consenting children and adults, for a nationwide screening where you can link genetic background, hopefully, with the occurrence of certain disease types. It's helped enormously in Iceland by the existence of a complete genealogy from the founders of that society. However, I was a postdoc in Oxford at a time when there was a series of studies called the Otmoor studies in villages around Oxford on human genetics, in part influenced by Walter Bodner. And what was so remarkable is that genealogies are often false; in other words, something like 10–20 percent of births do not come from the parents in that family. So genealogies may or may not be true, but it is a very powerful tool within Iceland.

Now many population geneticists—Richard Peto is a good example, and our own group also—have been looking with longing eyes towards China where you have a population of over one billion, and at the moment a very adherent population to directives, as we saw so clearly in the SARS epidemic, and clearly many very, very important epidemiological genetic studies are going to be carried out in China and are already underway over the coming decades.

Technology is a very important aspect in population genetics. Just to pick on one technology—whole genome microarrays. Increasingly in the pathogen area, we have these designed for very specific problems. One illustration might be drug resistance in tuberculosis, mycobacterium tuberculosis, which facilitates in the research laboratory extremely rapid screening and understanding of the genotype or phenotype of a particular organism. It is interesting to note, however, that these technologies have not made rapid entry into the public-health arena, so if you look across CDC in the United States and equivalent bodies in Europe, you will not find the rapid take-up of these technologies, as there should be. They offer extraordinary scope, even given problems of accuracy.

Rapidly Evolving Infectious Diseases

Now I'm now going to turn for the majority of this talk to the two problems I want to introduce to you, and that's genetic diversity to start with, and I'm going to focus in particular on bacteria here which present many, many interesting problems.

First in passing, I make the observation that the list of disease, infectious disease, and a human genetic influence—in other words, a gene or a product for a gene—is growing exceedingly rapidly. But this is hardly surprising for those of you who've ever done experiments with pathogens with mice in the laboratory; everybody knows that you use highly inbred strains as choice, and ideally you'd like cloned mice, because as soon as you move a pathogen from one inbred strain to another, you have totally different pathogenesis. Rolf Zinkernagel's system is a very good example of that. So we know that genetics matter enormously in determining the typical pattern of infection. So this list is huge and the origins started way back, probably with the blood-group disorders related to malarial infection.

If we look at infectious disease today, there's been a resurgence in Western medical schools in students interested in this discipline; it is worth reminding ourselves that if we look at the world in total that the leading cause of premature mortality remains infection, sadly. In our own developed and industrialized societies, of course, that is not true today. But in the majority of the world's population these are the agents of the most important source of mortality.

If you think about our world today, there are three factors that tell us that evolution is going to be speeded up in terms of the emergence of new infectious agents. The first is a very obvious thing. Our internationally mixing world, our increasingly global mixing world, means that agents pass quickly from high-density populations to new ones. A rough order of how our mixing has changed, we probably had a thousandfold increase in effective epidemiological contact rates between different countries between 1980 into 2003. Now that's an epidemiological observation, but remember that's also an evolutionary-biology observation, because transmission from one host to the other gives the organism an opportunity for evolution. So concomitant with that transmission increase, there is the huge increase in the evolutionary opportunities. If we look within countries, again we've had extraordinary changes in the pattern of mixing within human societies.

We've also had other changes which are relevant. There is obviously increased population size, there are improved technologies for detecting new pathogens, but there are these factors, pathogens like big, highly dense populations, and if we look at the growth in what the United Nations called "megacities" —that's cities over ten million people—we see that there is going to be considerable growth, and there has been considerable growth, particularly in Asia. It's also of interest to note that the fastest growth in air traffic between different cities has also occurred in Asia over the last ten years. So these Asian—the origins of new influenza strains or the coronavirus is not entirely by accident, there are epidemiological reasons why that might happen.

The SARS Coronavirus

So remember we've got all these genomes, and we have the capability for studying genetic diversity in some detail. And remember also in certain fields there is repetitive whole-genome sequencing; HIV is one, but there are a number of others, where in other words, sequential whole-genome sequencing—influenza is another good example—enables you to discover or to investigate genetic diversity not just in single genes but across a whole range of them.

The coronavirus, this new one that we suffered this year, is an interesting example—this is another interesting case when PowerPoint transferred from one system to another—underneath that red bar is the new coronavirus, and this was Malik Peiris' first pass based on the whole sequence in looking at the relatedness of this new virus to other viruses. And you can see the red bar falls between the avian, bovine, and murine. This was actually wrong in hindsight, there's another beast called the civet cat in which the virus, coronavirus, exists in, that is very, very closely related to this new human form. But this question is still very open, where this virus came from, and a lot more viruses need to be isolated from other mammalian species, particularly in these animal markets such as exist in Qangdong Province. So evolution is constantly happening, usually with combination events where two viruses enter the same cell, and there's a jumbling of the genomes.

This epidemic of this new pathogen had the most extraordinary effects in terms of loss to the economy, particularly within Asian countries, WHO took the brave step of issuing travel directives which limited greatly, not as a sort of order but in terms of voluntary limitation of people traveling to certain parts of the world. The sort of postmortem of this event tells us that we were extremely lucky. First, if these major introductions had occurred in North America and Great Britain or Europe, I'm not sure we would have controlled this particular virus, for the simple reason that the draconian measures that were put in place by China, by Singapore, by Hong Kong, I don't think we could've put in place in Western societies where politicians are concerned with reelection.

There's another very important biological reason why we were lucky, and this is a more genetic one. This virus—ah, interesting, we're getting variants—that picture there was a picture of the SARS coronavirus [inaudible] micrograph—never mind, the important information is not in the picture. This virus had trouble replicating in the human body, and a very, very unusual situation pertained where peak viral load, in other words peak infectiousness, occurred significantly after the onset of clinical symptoms, such as acute temperature. In other words, if you were to implement isolation and quarantine, it would work with this virus. If you look at viremia in clinical epidemiology of influenza A and the onset of clinical symptoms, peak infectiousness coincides with the onset of clinical symptoms. So in other words you've got a chunk of infectiousness before the patient is aware they're sick. With this beast, very fortunately, because of its poor replicative ability in a human host, we were lucky; it was not infectious, so isolation and quarantine worked exceedingly well.

Now this continued evolution occurred in China, much of it not understood, because not enough sampling of coronaviruses has taken place, and we await with some trepidation what might happen this autumn to see whether this virus has indeed increased its evolutionary transmissibility.

The other aspect of this virus—this is one of the most pathogenic I've ever seen, the case-fatality rate in the over-60s was over 60 percent. Case-fatality rate in the under-24s was essentially zero. So that's a mortality rate which is very much linked to age. And remember influenza, which we regard as a highly pathogenic organism, has a much, much lower case-fatality rate than this beast. So we were right to be duly concerned internationally.

Genetic Diversity in Pathogens

Now how do we understand and sort of interpret and study genetic diversity in these pathogens? There are a variety of approaches, and this perhaps relates a little bit to Eric Lander's comments about the information that is available. Multilocus sequence typing is one approach used for bacteria, where in an international effort via Web-based information entry, you sequence certain gene fragments on some housekeeping genes, not the ones expressing surface antigens, and you enter these sequences into a common international database with information on where that isolate came from, and if possible clinical information on the pathogenesis induced in that patient. So what you're trying to do is to build up a huge databank on gene sequence and the clinical pattern of disease created by that isolate.

Now this database approach has now been instigated for a wide range of very important human bacteria, and some of them are listed here, ranging from sexually transmitted agents through to agents that cause pneumonia, etcetera. Now the problems with these databases, and some of you would probably pick this out instantly, is that bacteria are dreadful organisms to study at the population-genetic level because they have very, as it were, promiscuous genomes. They're constantly recombining; they constantly suffer the problem of mobile genetic elements, where huge chunks of genetic information is moving around. Your gut is a very good illustration; we've only typed probably 10 percent of the bacterial species in the human gut; there are many, many tens of thousands of species present. There's enormous movement of genetic material between different species. Recombination was always regarded as rare in viruses. It's very interesting, now that we've got whole-genome sequencing for viruses what is coming out time and time again is recombination is very, very frequent in viruses as well. HIV is a very good example; recombination frequency is exceedingly high.

Now the trouble is when you have these international databases is that trees are extremely difficult to interpret, and this is one called *Streptococcus pneumoniae*, the pneumonia-causing bacteria, or one of them, and if you look on your left-hand side, you'll see the phylogenetic tree of relatedness between all these strains, and they're exceedingly difficult to interpret.

Now the question is, can we get a reliable intraspecies phylogeny for a bacterial species, given the problems of recombination and mobile genetic elements, or can we only hope to look at the most recent clonal groups and to try and associate those, the genetic diversity there, with occurrence or nonoccurrence of serious disease?

And this is the work of Brian Spratt in my own institute and cohorts of postdocs with Brian, and they have very much been addressing this question recently, and have come—for most important bacterial pathogens of humans—come to the conclusion that these phylogenetic trees—beyond the end point on the right where you look at clonal complexes—are almost meaningless. Now the reason they've come to that conclusion, as I'm going to show you the techniques later on, is they're beginning to simulate the evolution of genomes in the computer with recombination, and then reconstructing the phylogenetic trees, when in the computer you know exactly what the evolutionary events were. And these evolutionary phylogenetic trees of relatedness are for most bacteria where recombination is frequent almost meaningless past the most recent events. And that's the depressing conclusion.

A New Approach to Genetic Diversity

Now what can you do with the most recent events? Can we have new paradigms for interpreting and looking at genetic diversity? And this is one approach, and I only mentioned one. There are a number of others evolving at the moment; it has a rather trendy title called eBURST, and Ed File has developed this in London. It identifies all the clonal complexes within the multilocus-sequence database. You can define what a clonal complex is; it might be a group of isolates that have at least six to seven alleles in common with any other isolate in the complex, and then you need a simple biological model of how this complex is evolving, and then you need a way of robustly estimating relatedness, and then reflecting the tree.

So you could divide the database into non-overlapping groups, use a defined similarity, and then attempt by various statistical methods to predict the center. And here's some illustrations. Remember, with most bacteria you've got two chromosomes—sorry, you've got double strands, you've got double-locus variants and single-locus variants, and here you can see an attempt from [inaudible] to link in distance terms the relatedness between different isolates.

If you move from a simple example to a more complex one—I'm sorry this hasn't translated either. Let me try another one. Ah, here's one that has. This is *Streptococcus pneumoniae* from a very large international database where you're searching groups of isolates that are clonally expanding, where within that group of clonal expansion you've got limited number of genetic changes, you've got clinical data on the seriousness of disease in particular patients. And then by this method—and I'm not showing the linkage of this with the clinical data at the moment, that remains to be done—you can begin to understand more detail about the recent evolution of these organisms.

So it identifies clonal complexes; it provides measures of statistical support for what was the founder of that complex; it's a very conservative approach; and it doesn't attempt to reconstruct what is very difficult to reconstruct, the recombination events in the past.

Evolution of Influenza A

Now lastly, to end I want to turn to how one can use modern computational techniques with large computers to simulate the actual evolution of the genome. And I'm going to choose as an example influenza A, for very good reasons—it's a simple virus—and I'm going to look at how one might interpret its evolution.

We have a lot of information epidemiologically on influenza A and B, and its close variants. We have extraordinary information in some systems. The United States doesn't have these. It always surprises me that the CDC in computational terms is a bit backward—they need a heavy dose of quantitative methodologies—it's very surprising given the advances that have been made in the molecular-genetics areas. So here for France, for example, showing my lack of bias towards our European neighbors, is this wonderful spatial evolution of influenza A epidemics, which could be tracked day-by-day, week-by-week, by sentinel survey settings.

Now what we're interested in is evolution. We're interested in three problems. One is evolution—what are the determinants of the pattern of it. The second one is how likely you are to evolve drug resistance—we have two new drugs for influenza A which will be used in the next pandemic of a seriously pathogenic strain, and are used in old people's homes for the vulnerable and in hospital-care settings. And then thirdly the design of effective vaccination programs. I'm only going to talk about one of these problems today, and that's the first one. What are the ecological and immunological determinants of influenza A evolution?

I should remind you that if you take something like measles prior to wide-scale immunization, we always thought that was a highly transmittable infection, average age of infection about 5 in North America and Britain. But if you look at the annual incidence per head of population prior to immunization, it was only about 1.5 percent. By contrast, if you take influenza A, this attack rate is about five times bigger, at least. So influenza A is one of the most transmissible agents that we know in detail.

Interpreting Influenza A Ecology

Now the first study of significance in interpreting ecology was done some years ago in the very early stages of the development of immunological tools, and it was a very clever study by Stuart-Harris. Essentially what he did is he isolated virus from patients and took serum from patients. You type the virus and put it into its various strains, and then you looked how the serum in patients who had different viruses cross-reacted. And here you can see a very important biological problem, you can see that there's not a great deal of cross-reactivity outside of the specific strain. And that began to lead to the notion that once you've had a strain you're probably immune for life and the fact you get influenza A is because the strain has changed.

Now today, of course, with genetic information we know a great deal more. And if I just focus for the purposes of this description on the surface glycoprotein, haemagglutinin; it contains about 329 codons, and these are the ones I want to simulate the evolution in. And it's very interesting that 35 percent of replacements occur at a very small fraction of the genome. In other words, that's clear evidence of very heavy selection. We can describe these replacements and selection in very formal mathematical terms, and we can use large computers to simulate that evolution patient-by-patient in a large population.

Here's what happens in reality. If you think about influenza phylogenetic trees, they're like Canadian redwood, they're a sort of straight trunk with little branches coming off. And that's in huge contrast to most other viruses; the most extreme would be HIV. HIV is like a flat-topped acacia with a trunk, and then evolution has gone *boomph*, like that. So this one is the easiest problem to address in the first instance.

Ah, another picture gone, sorry. In that box up there was the acacia tree of HIV-1, so it was like a flat-topped acacia, and these are the influenza A trees, and remember there's two processes of importance—drift, slow evolutionary change—and then we have antigenic shift, which is the recombination event for an avian strain that causes a dramatic change in the virus, or significant change in the virus.

Now what surprises me at first sight here is given the high transmissibility and high mutation rate, why are these trees so linear? So the scientific question relates to this, this is looking within influenza A. We've got A subtype H3, subtype H1, and then through to B. You can see A subtype H3 is very linear; A subtype H1 is a little bit more diverse; and then B there's been a major bifurcation in the past. And we've got two distinct serotypes at present. These bifurcation events, of course, are clearly very important in vaccine development. So we've got a slender trunk for H3, H1 is more similar, and influenza B reflects this branching.

If we look at any one year, we find the interesting phenomenon that you have many strains present at one time, which is a slight problem for immunization effectively, but you also, if you do a time-series analysis of this data, you find that there are strong correlations here. One being very prevalent dictates that another one will be less prevalent, and that may fluctuate from year to year. And we know something about the rate of evolution for the different major types.

Computer-Simulated Genome Evolution

Now to move to what shapes the topology of these phylogenetic trees of genetic diversity. Now the only way I can really think of doing this is that you can't do real-time experiments with humans. You could, of course, with an animal model, but again you'd like large populations of animals like some of the very early studies done in the UK by a variety of people on populations of mice who, in the basement of the London School of Hygiene and Tropical Medicine, introduced ectromelia virus and let it spread for months on end in this free-running population. But in those days they didn't have the genetic tools to study the evolution of the virus.

So why not use modern computational power to actually explore the evolution of the genome? So mathematical models can play a very important role here, and these have to be structured with some degree of sophistication; they've got to include age and spatial structure—and I'll come to spatial structure as well because that's very important—and then the evolution of the genome.

So we have a very trivial individual framework, susceptible person infected and infectious, recovered, immune to all strains, transient nonspecific immunity—and I'll come to that in a second—and then immune for life to a specific strain so you can circle back and be reinfected with another strain which is significantly distinct.

Now when you start to simulate these evolutionary trees within an artificial genome, what you do is you have a patient who's infected, you're simulating the mutational changes across a set of codons, and then you're replicating that in millions of people, and then you're also simulating transmission within that group. You cannot do population genetics here without including transmission. Much of population genetics assumes arbitrary frequency-dependent functions, totally arbitrary. You must have the real frequency-dependent function which is the nature of population density and the transmission of the virus.

And here's four examples: one, this is run for fifty years of evolution. B has no short-lived immunity post-recovery from influenza, C has no short-lived immunity, and has a very low transmission rate, D has over-restricted diversity achieved with very specific parameter values, and then E has—reducing the intensity of cross-immunity across the different strains. So slowly you can build a picture of how each biological factor influences the shape of this tree.

And here's one very trivial example, cross-immunity. Using the information from Stuart-Harris, C is an inverse measure of cross-immunity, if it's 1 it's perfect, if it's 0 there's no cross-immunity, and you can see how changing the parameter C changes the shape of these evolutionary trees. And these calculations were done by Neil Ferguson and the group, who's a theoretical physicist by background who when interviewed for the post when asked why he wanted to come to biology from a very good theoretical physics, he said, "Biology is much more interesting because there are so many more nonlinear processes." It wasn't the furry and the featherers in the gene sequences, it was the fact that the mathematical nonlinearities excited him, and there are some very significant ones in these problems.

Now I work with trepidation to see whether this functions. The reason space is important—as illustrated in this slide—this is a study of the spatial spread with these genome evolutions going on. And each color is a different strain. And these are two loosely connected patches in the country, or between countries, and so evolution depends on transmission between the patches.

You might ask how on earth can theoreticians describe mixing of human societies. Well, very interesting, one facet of our modern society, namely mobile phones—and you're probably not aware of this—provides an extraordinarily digital record of what you do. This is not so true in North America, but it's very true in Europe with a high density of masts. Every time if your phone is switched on you move between a different mast, a digital imprint is left on the computer there that your phone is logged onto that mast. There are teraflops of data per year, so you need bloody big analysis setups, but you can then map for location X the probability distribution that an individual moves a distance Y. And that spatial probability distribution underpins these spatial transmission events.

And what you see is that frequency-dependent selection operates very much in a spatial domain. So a strain sweeps through a population, you've got heavy frequency-dependent selection with the evolution of a new variant. Then it moves onto another patch. The thing with influenza is it goes through the world in about two years because of its very, very high transmissibility.

Now one can map these epidemiological and genetic things in three dimensions, and here's an illustration. On the axis going into the graph you see substitutions, which is the genetic information, at a defined set of sites. On the horizontal axis is time, and on the vertical axis is the epidemiological variant, which is the frequency that that strain is in the population. So now we're beginning to get the computational tools to meld the epidemiology with the genetics, where you can study the evolution through time and you can study the epidemiological, the prevalence and instance of infection. These are very, very early beginnings with a very, very simple virus, influenza A, but I see no reason why one shouldn't extend this to much more sophisticated systems. We just need rather large computers.

Human Genetics and Pathogen Diversity

So in conclusion, evolution in pathogens, of course, occurs in a very, very fast time scale. And for many of the human ones, the important point is that evolution is speeding up, not slowing down, because of our mixing, because of our population size, and so forth. So they're very good models for understanding evolution in animal populations. We have increasingly sophisticated methods for study, but we very rarely bring all these together in one laboratory or team. And that's so apparent in our centers for disease control or public health authorities.

The need to link the human genetic background with the pathogen diversity I haven't touched today, but it's so obvious. They're two gloves that fit together—pathogenesis against one genetic background for one virus may be totally different with a different virus or against a different genetic background. We've hardly begun to think about how to address those problems. We can do it in the laboratory context, but not in real human populations as yet. But, given the fascination in human genetics and the accumulation of information like the BioBank project, given the ability to isolate in sequence pathogens, the real issue here is sampling, not the lack of technology. We are very, very bad at taking representative samples of pathogens at one point in time internationally.

And lastly I do mention—which I'm sure will come up tomorrow in much more detail—the very important ethical issues for epidemiology that must be resolved. If some current legislation and discussion in the European parliament were to go through, it would almost inhibit a huge chunk of epidemiological research. It's the question of informed consent and the holding of biological samples such as blood, serum, etcetera. So there are very, very important problems there indeed that scientists must play a role in discussing with the public.

Thank you very much for your attention.

Genes, Genomes and Society
October 17, 2003

Lee C. Bollinger, President, Columbia University, and Jonathan R. Cole, Ph.D., Columbia University
Welcoming Remarks and Introduction

Welcome by Lee C. Bollinger

President Lee C. Bollinger: I would just like to welcome everybody briefly to the second day of the symposium. I have heard comments after yesterday's session that makes me think that the impossible may have been reached here, which is that a symposium has actually been one of the most creative and important experiences intellectually for people who were fortunate enough to be here. I want to thank all of the people who have come as panelists and speakers from outside of Columbia to participate in this and to help us celebrate the 250th anniversary of Columbia.

The subject of genes is of course immensely important, and we all know that. We are, outside of science, woefully ignorant about what this really is and the implications, although through discussions, through media, we have inklings that this is perhaps one of the most important revolutions in human knowledge within several centuries, or right up there among the very top.

The other panel, the other symposium that we have going on yesterday and today, is on constitutions, constitutionalism, and for my mind, to my mind, these are again two of the most important areas of development in knowledge in human activity. The subject of constitutions of course in the United States has a sort of magisterial presence. It has really been not only a way of defining the government in a fundamental way, including civil liberties of course, but also a way of defining the basic values of American society. Today in this symposium we take up also a connection between the discoveries that have been made or that well be made, we hope will be made, and the implications those have for society.

The moderator is Jonathan Cole, whom I very pleased to introduce. Jonathan is enviably on a sabbatical in Paris and has given up a few days of that to come back and be with us here. No one understands the interconnections of knowledge, whether from constitutions to genes, or from science to humanism, technology to social sciences, better than Jonathan Cole, whose own work as a sociologist covers and thinks about the sociology of science. It is a great, great pleasure to have Jonathan back, and I give him to you. Thank you very much.

Introduction by Jonathan R. Cole

Jonathan R. Cole: Good morning, and thank you very much, Lee, for those kind words. It is good to be back at Morningside Heights, despite the pleasures, culinary and otherwise, of Paris. And it really is my pleasure to serve as the moderator of this morning's session, "Genes, Genomes and Society."

I should also like to take this opportunity to thank our guests at Columbia, particularly the extraordinary women and men of science who have participated in this exciting program. I want also to thank Gerry Fischbach, Tom Jessell, and Joanna Rubinstein for organizing the program and for the immense care that they've taken in putting together a balanced and extraordinarily interesting set of speakers.

Now over the past day we've heard some extraordinary examples of the way the biological revolutions involving genes and genomes are rapidly advancing, and are apt to change the way we think about medicine, disease, about creating and sustaining life. We have also been given a tantalizing introduction into the puzzles of science, and how scientists go about trying to solve them. We have seen the beauty of science and its unending quality, that is, the way in which the solving of puzzles create new puzzles to be addressed. If these scientists see further by standing on the shoulders of giants, they in turn are creating the shoulders on which others will sit and gain advantage in viewing the beautiful and ever-continuing landscape of scientific puzzles. Watching my old friend and colleague Michael Levine and others here yesterday, I thought of Newton's reflections with perhaps slightly obsessive humility on doing science. He said, "I do not know what I appear to the world, but to myself I seem to have been only like a boy playing on the seashore, and diverting myself and now and then, finding a smoother pebble or a prettier shell than ordinary whilst the great ocean of truth lay all undiscovered before me."

Now on an entirely different but equally important level, there are, of course, profound social implications of these revolutionary scientific changes that were discussed yesterday. All of the science created and developed by these extraordinary men and women is embedded in larger cultures and societies. Fully understanding these revolutions requires that we place them in the proper societal context. The values, attitudes, beliefs, and allocation of scarce resources affect the progress of these revolutions and in turn affects how we use the outcomes of this amazing work. It's therefore critically important that we consider the way societies have had an impact on this science, and the ways that science and technology can have an enormous impact on societies. And we should consider how these discoveries can affect the relationship between the richer and poorer societies of the world.

There are a set of important and relatively new questions that we must also attend to. A sharply reductionist view of science is yielding to a deeper understanding of the critically important interactions between biology and the larger social environment, if you will, to the critical interaction between the biological and the sociology and anthropology of social systems. This in itself opens up the possibilities of the emergence of critically important new areas of social-science study that are related to the knowledge being generated by the biological revolutions.

How do genomics and genes influence social orders and social systems? How do social systems influence biological outcomes? How do biological and social factors interact to affect the expression of genes or influence types of behavior? How do we reconsider old and problematic concepts, such as race? How should we, in the purely normative sense, use the fruits of our biological discoveries to improve the condition of the populations of developing nations in the world? Can biological discoveries, if put to proper use, attenuate the growing inequalities between the rich and poor nations of the world? How do we strike a proper balance between the needs of scientists to experiment and the needs of individuals to be protected from improper modes of experimentation? How do we deal with the ethical issues related to the science of cloning and the use of specific strategic materials, such as stem cells? Can we protect science from the dangerous intrusion of political ideology, which can destroy otherwise healthy scientific organizations? Can science police itself adequately? How do we close the gap between the rate of advance of science and the rate of advance in institutions that must use or misuse the fruits of the discoveries?

It is towards consideration of such questions that we turn to in this session "Genes, Genomes and Society." We have four extraordinarily distinguished speakers with us this morning, and biographical sketches of each of our speakers can be found in your programs, so I will only highlight a number of their many achievements.

Now in the first of this morning's sessions before we break for some coffee, we shall hear from Professor Anne McLaren and Professor Andrew Colin Renfrew.

Anne McLaren, Ph.D., Wellcome/CRC Institute, University of Cambridge, U.K.

Bioethics: Embryo Research, Genetic Diagnosis and Therapy, Stem Cells and Cloning

Introduction by Jonathan R. Cole

Jonathan R. Cole: Let me introduce Professor Anne McLaren. She's our first speaker. Professor McLaren was educated at Oxford University, where she did both her undergraduate and graduate degrees. She was director of the Medical

Research Council of Mammalian Development Unit in London for 18 years until 1992. Prior to that she worked for the Agricultural Research Council in C. H. Waddington's Institute of Animal Genetics in Edinburgh. Her research has ranged widely over developmental biology, reproductive biology, and genetics, including molecular genetics, using the laboratory mouse as a model. Since 1992 she has been working at the Wellcome Trust/CRC Institute of Cancer and Developmental Biology at the University of Cambridge. She was a member of the UK government's Warnock Committee on Human Fertilisation and Embryology. She's a member of the European Group on Ethics, and advises the European Commission on social and ethical implications of new technologies. She's published widely in all of these areas, and most recently coedited a book entitled, *Adoption of Opinion on Ethical Aspects of Human Stem Cell Research and Use*. She was elected a fellow of the Royal Society in 1975, and from 1991 through 1996 she served as foreign secretary and vice president of the Royal Society. In 2002 she was awarded the Japan Prize for Developmental Biology. It is a great honor for me to introduce to you Professor Anne McLaren, who will speak to you this morning on the subject "Bioethics, Embryo Research, Genetic Diagnosis and Therapy, Stem Cells and Cloning."

Louise Brown, First IVF Baby

Anne McLaren: Can you hear me at the back of the room? Good. I feel hugely privileged and honored to have been invited to participate in this remarkable, remarkable symposium. So many thanks to the organizers for inviting me. I'm also delighted to find that Columbia University, which I've known for quite a while on and off, is, as it were, descended from Kings College of New York, because I'm a fellow of Kings College in Cambridge, so that makes me feel even more—even more friendly towards Columbia.

So, 250 years of Columbia, 50 years of DNA, 25 years of Pope John Paul, and 25 years of IVF, *in vitro* fertilization. It was 25 years ago, in 1978, Louise Brown, the first IVF baby, was born in the UK.

Now, the developments that led up to the birth of Louise Brown occurred in the UK rather than the USA, and the birth of Louise Brown in Britain was greeted by the British public and the media as a miracle baby and not a Frankenstein baby. And those two things, had they been otherwise, I think would've made subsequent events very different. Much of the ethical issues in this area relate to the status of the human embryo, and this was really first brought to public attention by the birth of Louise Brown and by the popularization, as it were, of IVF. Because by now something like a million babies all over the world have been born by IVF. But the British government realized that there would be social and ethical problems, so they set up a committee to report on infertility, surrogacy, for instance, and human embryo research. It was chaired by Mary Warnock, the moral philosopher, a moral philosopher, I think a very sensical moral philosopher, and an excellent, excellent chair.

Now there was wide consultation and the Warnock Committee got piles and piles and piles of documents, letters, from individuals, from groups, church groups, women's group, trade-union groups, neighborhood groups, many, many groups, and quite a few of these did comment on the moral status of the human embryo. Of course some thought that embryo meant something with arms and legs and a head and a nervous system, but not all that many; most people realized that one was dealing with maybe a dozen or so cells.

There were some who thought that this aggregation of cells was really just an aggregation, and had little or no moral value, but that few wasn't and isn't very common. But the other extreme, of course, the official Roman Catholic view and the view of some other people as well, is that from the one-cell stage onwards, from fertilization onwards, the moral value in embryo is equal to that of a newborn baby or an adult human being. But the majority of people seemed to think that the moral value of the embryo increases as it develops, first into a fetus and then into a baby. And that was in fact the view that the Warnock Committee eventually came to, and they concluded that of course the human embryo is entitled to respect because it is human and it has potential, but unless that embryo is going to be put into a uterus, that respect can be weighed against the benefits to be derived from the proposed research. But the research should only be permitted, given certain requirements. Some were obvious: for instance, the informed consent of the donors of the embryo; a local ethics committee; the work would have to be scientifically valid; the need for human, not animal, embryos; enough animal research would have had to have been done first to make it essential now to check the results on human embryos. And they decided that it would have to meet acceptable objectives, purposes of clinical relevance, and also that no embryo which had been the subject of research should be transferred to the uterus, because of course that would imply research on the fetus and on the woman, because by definition with research you don't know what the outcome is going to be.

Also, if there was going to be legislation, then there had to be a time limit for *in vitro* development. One couldn't just airily say yes, you can do embryo research, there had to be a time limit, and after some discussion the Warnock Committee decided on 14 days, which is the last stage before twinning is possible, before the formation of a primitive streak, so it's the beginning of individual development, in contrast to fertilization, for example, which is the beginning of genetic identity. And this 14 days, which I'll say a little more about later on, has been adopted by many countries who do allow some degree of human-embryo research.

The UK Human Fertilisation and Embryology Bill

Well, there were several years of consultation, and finally the British government brought the Human Fertilisation and Embryology Bill before parliament. There was by that time—after considerable debate—there was wide support for the view that the archbishop of York put forward in the debate in parliament, namely that if IVF was to continue, research must also continue, because it would be totally unrealistic and indeed immoral, he said, to continue IVF without a proper backing in research because imperfect techniques without a backing in research are bad practice medically, and I believe wrong morally.

Well, the vote in parliament supported that view. In 1990 the act came before both houses of parliament in favor of embryo research, you see both the House of Lords and the House of Commons large majorities, and also both houses of parliament voted in favor of allowing embryos to be made for research under certain very strict conditions, namely that it was only allowed if the project was essential and couldn't be done otherwise, couldn't be done on any other form of embryo.

The other thing that the act did was to set up the Human Fertilisation and Embryology Authority, which was a statutory body, and it has to license, monitor, and inspect every year all clinics that carry out IVF or donor insemination, also to license and monitor all centers carrying out human embryo research, so every project has to be discussed. It may be accepted, it may be rejected, it may be modified, but it has to be licensed before it can be started. They regulate storage of gametes and embryos, keep a register of information about donors, treatment, and children born, and produce a code of practice.

There's quite a bit of debate going on at the moment in the UK about donor anonymity, because according to the act, both sperm and egg donors should be anonymous. But a lot of people feel now that that's wrong, that children have a right to actually know the identity of their genetic parents. I can see both sides of that argument, I haven't made up my mind, and parliament hasn't made up its mind either.

The other thing that the UK government has done quite recently is to set up a national stem-cell bank for both embryonic stem-cell lines and fetal and "adult" stem-cell lines; it's only just getting going, but I think it's an important development. The cell lines will be available to UK academics and overseas academics free of charge except for the cost of shipping and all of that. Industry will be charged. The ownership of the cell lines remains with whoever deposited them in the bank, and of course there'll be material trade agreements involved.

Stem Cells

So what are stem cells, what are stem cells? Well, stem cells are defined as cells that are capable of self-renewing; they could make more cells like themselves, or they can make cells which give rise to one or more differentiated cell types, so they're cells with a choice, they're cells with a choice. They can make one cell just like themselves, and another one which may develop into nervous tissue or muscle.

It became rapidly clear that stem cells had great promise for a number of reasons. First of all, of course, in experimentally basic research, they are enormously important because one can study in culture in the lab how cells develop into these different tissues. The pharmaceutical companies are very interested in stem-cell lines for drug development and toxicity tests. And in the future they hope to use a lot of stem-cell lines, human stem-cell lines, rather than animals, for instance, for toxicity testing. And again there's the possibility of cell and tissue therapy, because although sometimes whole organs need to be replaced, quite often it's enough just to replace some cells, some tissues, repair, as it were, and that can be so both of the—of the blood, bone marrow, nerve cells, the heart muscle, pancreatic eyelet cells, in diabetes—and indeed there are a whole number of diseases that I've listed here; some of them are neurological diseases, Parkinson's, Alzheimer's, stroke, Multiple Sclerosis, then we have the liver, the heart, diabetes, rheumatoid arthritis. And these tend to be, unfortunately, very common diseases. Diabetes, Parkinson's—I suspect we all know somebody who is suffering from one of these diseases. They're intractable, no real cures are known, and they cause a great deal of suffering. But they could be alleviated or even cured if there was the appropriate cell type in a large enough quantity and appropriate condition to use for therapy.

Now, stem cells can come from different sources, they can come after birth, which is usually called adult. There are stem cells indeed in the brain. It used to be thought that we were born with all the neurons we were ever going to have, but now it's known that there are stem cells in the brain that can repair the brain, and animal experiments have shown that they can be removed and used for cell therapy. Bone marrow is already used, of course, for transplantation, and has been for a decade or two. Cord blood from the placenta is a possibility; satellite cells are muscle stem cells. Then it's possible to get stem cells from fetuses, on termination of pregnancy, and also embryonic stem cells, and it's these where the ethical issues are the most pressing.

Now, adult stem cells. The problem is that they're present in the body in only very small numbers, and they don't proliferate all that well in culture. So there's a numbers problems with adult stem cells if you want a large mass of stem-cell tissue for use in a hospital. Embryonic stem cells come from the blastocyst-stage embryo, and I'll show you in a minute what that means, it's a stage where 100 and 150 cells in an embryo not yet implanted in the uterus. They proliferate indefinitely in culture, are chromosomally stable, unlike transformed cancer cells, so one can get a very large number of cells from embryonic stem cells. They're pluripotent, they can give rise to any tissue if appropriate treated, but when a stem-cell line gets old, then it tends to be less capable of giving rise to all the desired tissues. And important from the ethical point of view, embryonic stem cells can't on their own make an embryo. An embryonic stem cell is very different from an embryo, and it can't make an embryo. But the derivation of these cells, which are called ES cells for short, the derivation of these cells involves the destruction of the embryo from which they come, and of course that is what raises the ethical problems.

Embryos

If we look at the first couple of weeks of human development, here we have fertilization, and of course fertilization is when the new genetic constitution is first established. Genetic uniqueness comes much earlier. Eggs and sperm—every sperm is genetically unique, different from every other sperm, different from the man who produces them. But the new genetic constitution starts at fertilization. Then you get cleavage into two cells, four cells, eight cells—this is the hundredth cell stage from which the ES cells can be derived from these inner cells here. That embryo will then after—about a week after fertilization it will implant in the wall of the uterus. That's about seven days after fertilization. And then during the second week the implantation process continues in the uterus, until you get this large mass of tissue derived from the fertilized egg, of which less than 1 percent, about 0.1 percent from the very middle here, this layer called the epiblast is going to develop into the fetus and the baby. And that's what's called at 14 days the "primitive streak stage," it's the last stage at which twinning can occur. One can get instead of one primitive streak two or, in the case of the Dionne quints, five monozygotic twins. Sometimes one doesn't get a primitive streak at all, and sadly one then just gets a tumor instead of a baby. But this is the stage at which individual—in the sense of undividable development—begins, at the end of the second week, once implantation is completed.

Now, embryos are of different sorts, and the source of the embryo determines the ethical problems associated with it. First of all, and most importantly, there are the embryos that are derived by fertilization in the course of IVF treatment. Because almost always there are more embryos produced, the woman is treated with hormones; otherwise, she would only produce one egg at a time, and IVF would be extremely inefficient, so she produces more eggs and more embryos, and the spare or supernumerary embryos are usually frozen for the couple's own use. But the time comes when they may no longer be required for the parental project; either the woman has got beyond the age of reproduction, or the couple have got the family, one or two children, that they wanted, or sometimes where IVF is private, rather than on a national-health system, they can't afford to have another cycle. So there are spare embryos, and the couple has the choice then of either donating them to another couple or donating them for research, for example, for research on stem-cell derivation, or just letting them die, and that is the fate of in fact most spare embryos at least in the UK.

Then, as I explained earlier, there are embryos that are made for purposes of research, donated oocytes, unfertilized eggs, which are fertilized for a research project, for instance. It may be a research project on fertilization itself, because fertilization is a process that quite often goes wrong and produces abnormalities; we don't know why, we need research.

And finally there could be embryos derived by somatic-cell nuclear transfer, cloning. So far that has not been successful in the human, but in principle one could have embryos of that sort.

International Laws on Stem-Cell Research

Now different countries all over the world have very different views as to which, if any, of these types of embryo should be used for research or for stem-cell derivation. Within Europe there are very big cultural differences between different countries, and therefore the European Commission has a problem in deciding what should be done. Should they give money for research on human embryos, or what? The European Group on Ethics gave an opinion to the European Union which said that for those countries where it was legal to do research on embryos, there was no reason not to develop treatments for serious diseases, which of course means stem-cell derivation, and no reason to deny European Union funding. But that would be on spare embryos, and then it went on to say that while spare embryos are donated, fertilization of eggs specifically for stem-cell research is not ethically acceptable, and furthermore they said derivation of embryos by somatic-cell nuclear transfer, cloning, for stem-cell research would be premature at the present time.

So if we just look at what the situation is in Europe and elsewhere in the UK, Belgium, Sweden quite recently, and also China, all those types of embryo that I listed could in principle be used—it wouldn't be against the law. In a lot of countries, spare embryos can be used for research on stem-cell derivation, but not any other type of embryo. France and Spain are still—the law isn't quite through yet—but it looks as if that's what it's going to be. Australia has an additional condition, which is that it's only embryos that were frozen, fertilized and frozen before April 1 this year that can be used for stem-cell derivation. But that applies throughout Australia, both to government-funded and also to privately-funded research. Four countries in Europe, Austria, Germany, Ireland, and Norway, human-embryo research and stem-cell derivation are entirely prohibited. In Germany it's allowed to carry out research on imported stem-cell lines, and mainly those stem-cell lines are imported from Israel.

In the Czech Republic, Israel, Italy, and Portugal, there is no legislation yet. You might think that Italy would have strict legislation on these matters, but it has no legislation at all, anything goes. Czech Republic and Israel have both made stem-cell lines and, as I say, Israel is exported them, but there's no legislation, no regulations.

And then we come to the United States, but I think you know more about this than I do. Government funding is not available for human embryo research or for derivation of human embryonic stem cells, but scientists are allowed to do research with government funding on embryonic stem-cell lines that were made before August 9, 2001. On the other hand, if privately funded human embryo research and derivation of new stem-cell lines is not illegal. So I think this is the only country that actually has a distinction between government funding and private funding for the ethics of human embryo research.

Stem Cells for Gene Therapy

Now, another possible use of stem cells, which isn't often drawn attention to, is in fact for gene therapy, because gene therapy—namely the attempt to cure genetic diseases by replacing or substituting gene products—there are single genes which could be replaced like for cystic fibrosis, Duchenne muscular dystrophy, and SCID, severe combined immunological deficiency. This is what is sometimes called bubble babies, babies that have to be kept in bubbles because they have no immune system. And this is one of the more successful types of gene therapy, but gene therapy is proceeding very slowly at the moment. Then there is the possibility that genes for therapeutic proteins might be introduced into the body, and this is where the stem cells could be useful, because stem cells that were producing these therapeutic proteins could be used for treating genetic diseases.

At present the vectors that are used for introducing these genes are less than satisfactory; a number of viruses are used, but they can be risky. Liposomes, which are little fat droplets, DNA conjugates, they are not very efficient. So engineered stem cells are looking more promising, for instance, mesenchymal stem cells; they're cells that can give rise to all sorts of cartilage, bone, muscle, fat cells, and they could be engineered to express some of these useful protein products.

Now, any new therapy, and this applies to gene therapy as well as any other sort of new therapy, any new therapy is risky. And there are two ethical issues that are really in conflict with one another. One is the precautionary principle, be careful. Now if one was totally careful, one would never introduce any new therapy. But then what about the serious diseases for which new treatments are urgently needed? You need a risk / benefit analysis. And then there are other risks; for instance, informed consent can be tricky. It's very important that false hopes should not be raised in patients' minds. I think that's happened in the past with gene therapy, at least the media have promised greater successes than have been apparent, and perhaps particularly in this country the role of litigation if things go wrong. Because indeed things can go wrong, for stem-cell therapy, good manufacturing practice, good clinical practice, are essential. One can't use mouse fetal layers, animal serum, that would be risky. There's a possibility that the stem cells might make the wrong sort of tissue—that could be very dangerous—or that if there were undifferentiated stem cells there they might cause tumors. For gene therapy, most of you probably know the tragedy of Jesse Gelsinger, a young man who died. I think it was an adenovirus vector. There were questions raised about whether the regulations were being followed, whether there had been adequate reporting of previous cases, and the suggestion that perhaps competition between different groups might lead to premature trials. But undoubtedly that case made something of a setback to gene therapy.

And then the SCID trial—I mentioned the SCID, the babies in the bubbles—that was going well, the 11 children were treated and appeared to be totally cured, clinical benefit; they came out of the bubbles, were leading normal lives, but 2 of the 11 then developed leukemia. And it was discovered that that was because the inserted gene had inserted in the wrong place, and that was what caused the leukemia. So that can be controlled, and since there is satisfactory clinical outcome in the other cases, and there is no other way of treating these very sad children, those trials are continuing, certainly in the UK.

Now, all of the gene therapy that I've been talking about is so-called somatic gene therapy, that is, treating the individual patient, the body, as with any drug. But there is also the possibility that's been proposed of germ-line gene therapy, that one could treat a very early embryo, before the germ line that's going to give rise to sperm and eggs is developed. And in that case, the defective gene would be replaced not just for that generation but perhaps for future generations. But it would be risky, unpredictable, and it is not a good option because there is an alternative, which is not risky and which is frequently done, which is pre-implantation genetic diagnosis. Because there a couple who are at risk of producing a child with a genetic defect can have the embryos screened before they're implanted in the uterus to make sure that the embryo that is being put back doesn't carry the affected gene for which they're at risk.

This is what the implantation genetic diagnosis looks like. Usually at the eight-cell stage, one cell is removed from a number of embryos, and by either—usually by a polymerase chain reaction, which is a very rapid method of looking at the genes in the embryo, there may be two that carry the defect, three normals. The three normals will be replaced in the uterus or frozen for future use. It's also been suggested that germ-line gene therapy could be used for enhancement; in other words, to make genes better. Well that's not technically possible at present, and it's widely regarded as ethically unacceptable for reasons of equity, autonomy, and so forth.

Even pre-implantation genetic diagnosis has sometimes been termed ethically unacceptable, because it has a flavor of eugenics. But eugenics is a very confusing term. Historically it really implies coercion, and in the opinion of the geneticists who met in 1998, new genetic technology should be used to provide individuals with reliable information on which to base personal reproductive choices, that's the essence, not as a tool of public policy or coercion. Informed choice should be the basis for genetic counseling, and the term eugenics is so confusing, it really should no longer be used in the scientific literature.

Cloning

Well, finally what about cloning? Now, cloning has never been illegal in the UK, because the 1990 Act didn't prohibit making embryos for research, but equally it's never actually been done. Now, cloning involves first of all the removal from a donated unfertilized egg of the genetic material, and then replacement by a somatic cell, a body, say the nucleus of a skin cell. This has been done in animals, Dolly the cloned sheep was the first one, it's been done in cattle, goats, cats, not dogs, mice, not monkeys—it doesn't work in monkeys. And in animals the progeny born from this cloning technique produce many abnormalities, many deaths during pregnancy, deaths at the time of birth, and problems later in life. Some, just a few, are healthy, but so few that most people feel that cloning for babies in humans would be criminally irresponsible, much too high risk of fetal neonatal deaths, malformations.

Of course it might one day be made safe and effective; if it were, would we want it? There are many varied ethical objections. On the other hand, there are people who say that it could be valuable for couples who are irremediably infertile, the replacement of a dead child—of course a dead child can never be replaced—or do-it-yourself female reproduction, where the woman produces the donor eggs that are going to have the genetic material removed and her own nucleus put in, and then they can be put back into her own uterus so as to produce a baby which is her identical twin, or I don't know what you'd really describe it as, but that's science fiction, that is not science fact at present. And certainly in the UK human reproductive cloning is now forbidden by law.

However, if such an embryo is not placed in the uterus, it could be used to make stem-cell lines to derive embryonic stem cells. And it's been suggested that that could be a useful way of getting around transplantation protection. Here you have the patient, the somatic cells, the body cells, are taken, the nuclei are put into a donated egg which has had its own genetic material removed, the embryos are cultured, stem cells are derived, and then differentiated to make muscle or nerve or pancreatic tissue, whatever the patient needs for their particular disease. And of course the stem cells will not be rejected because they will be the same tissue type as the patient.

Now, ideally of course, one would turn the somatic cells, the body cells, directly into stem cells without going through the embryo, but that's not possible at the moment. And I have to say that I think this whole process is never going to be clinically realistic. It's far too labor-intensive and costly to think of doing it for an individual patient. I really don't think it's practicable. There are other ways that one could envisage of getting around graft rejection, as well as this somatic nuclear transfer. You could take adult stem cells from the patient, though, as I said, there are rather few of those. You could genetically manipulate the stem cells to eliminate the antigens. You could have a large stem-cell bank so that you could select stem-cell lines that were reasonably compatible with the patient. You could induce specific immunological tolerance in the patient, with a lot of research going on along those lines. And of course there are immunosuppressive drugs which we use today for kidney and heart transplants. They are getting better all the time.

But I think it's important that the technology of somatic-cell nuclear transfer should not be prohibited, because it could be extraordinarily valuable for research. There are a whole number of very rare genetic diseases which at present are so rare that we really know very little about them, and it's difficult to get enough material to investigate them. If you could make stem-cell lines from such patients, you could have an indefinite amount of material to study. And the same is true of common but complex diseases. If we had stem-cell lines for some of these diseases with multiple causes, it would make it easier to study them and again. Research on the actual nuclear reprogramming might help to allow the sort of somatic to stem-cell conversion that I was talking about earlier. And in fact only last month sixty academies of science from all over the world signed a statement calling for a ban on human reproductive cloning, but requesting that cloning for both research and therapeutic purposes should be excluded from such a ban. And that statement was sent to the United Nations General Assembly, because United Nations has a committee on cloning, which is at the moment considering a convention.

So finally—I'll skip the next slide—what are my conclusions? Well, I think that ethical objections are usually more appropriately addressed by regulation than by prohibition. I think stem0-cell therapy may become important in clinical practice sooner than gene therapy, but gene therapy could use stem cells as vectors. And my own view is that cloning is unlikely to play any part in clinical practice, but it's going to be a valuable research tool.

And I'd just to end by showing you two passages from a Canadian report that I thought were really rather sensible, namely that the public tends to demand prohibition of conduct that's universally opposed but expects issues of moral ambiguity to be regulated, and criminal law should be an instrument of last resort to be used only in response to conduct which is culpable, seriously harmful, and generally conceived of as deserving punishment.

Thank you.

Colin Renfrew, Ph.D., McDonald Institute of Archaeological Research, Cambridge, U.K.
From Genes to Civilization

Introduction by Jonathan R. Cole

Jonathan R. Cole: Thank you very much. We now turn to our next speaker, Professor Colin Renfrew. Colin Renfrew received his Ph.D. in 1965 from the University of Cambridge. From 1986 to 1997 he was master of Jesus College, Cambridge, and since 1981 he has been the Disney Professor of Archaeology at Cambridge, and since 1990 the director of the McDonald Institute of Archaeological Research. In 1973 Professor Renfrew published *Before Civilization: the Radiocarbon Revolution and Prehistoric Europe*, in which he challenged the assumption that prehistoric cultural innovation originated in the Near East and then spread to Europe. Professor Renfrew was elected as foreign associate of the United States National Academy of Sciences in 1996. In 1991 he was awarded a life peerage, and chose the title Lord Renfrew of Kaimsthorn. Today Lord Renfrew will speak to us on the subject "From Genes to Civilization," and it's a great pleasure to present him to you.

The Record of Human Population History

Andrew Colin Renfrew: Well, good morning, ladies and gentlemen. And first of all I'd like to say what a very great pleasure and privilege it is to be here on this occasion to celebrate your 250th anniversary, and I'm very grateful to the organizers for the organization and for having invited me. It seems to me that I'm probably the first person in this meeting to be concerned with not just the historical dimension but with history, including prehistory. Many of us are interested in asking the question what are we? in the sense of how have we become what we are. And that really is the focus of my talk today, and I want to ask to what extent molecular genetics, genes, understanding of the genome, has clarified for us that process. In some aspects it clearly has, it gives us insights into how we have become *Homo sapiens*, but I want to emphasize that the story involves more than that, and that there are some aspects which we have not yet addressed very successfully, and I think the question today that's worth asking is whether genes and molecular genetics, as currently deployed, have yet taught us about the most significant transformations in the human story. That is to say, as I shall argue, that as it happened in the past 10 or 15 thousand years, and to ask whether we're yet ready to undertake the appropriate analyses.

Well, when in New York, where better to start than with *The New Yorker*, so I give you first of all this slide, that in a rather subtle way, asks us what are the differences between today and cocktails at a private view in Madison Avenue perhaps 15 thousand years ago. And I want to move on to a cartoon by Chas

Addams from again, *The New Yorker*, which I think poses problems about the nature of mind, which have still not been very effectively addressed. So I think this is a conundrum. As you can see there are some creatures, perhaps ants, who are constructing what looks like a pyramid, while those undertaking the picnic, no doubt cholesterol-rich picnic, are making the comment, "Well at least they're not bothering us." But we're all aware that some ants, termite ants, construct nests, so what is so disquieting—because I hope you find this image disquieting—what is so disquieting about this image? And one element—it may not be easy to analyze completely—one element is that these ants are doing something which is clearly the product of planning, and planning—I'm not sure about purposive behavior—but planned purposive behavior is surely something that is unique to the human species. So as you'll see, this slide is rather relevant to my talk this morning.

Now, my suspicion is that I was invited to this symposium not to give you the bad news that genes and genomes have as yet made limited contribution to our understanding of human history, as I shall go on to say, but because I have also been involved, to a limited extent, in the new discipline of archaeogenetics, which may be defined as the study of the human past using the techniques of molecular genetics. And archaeogenetics has already been of great value in indicating, mainly through studies of mitochondrial DNA for female lineages, and non-recombining Y-chromosome studies for male lineages, that our species, *Homo sapiens sapiens*, originated in Africa over 100 thousand years ago, and came to populate the globe following disbursals out of Africa less than 80 thousand years ago. And despite some notable successes with the study of ancient DNA, some of them pioneered by Svante Pääbo, who spoke to us yesterday, despite those successes on DNA recovery from hominid or human remains, it is very fascinating that most of the data in the field of archaeogenetics have come from the comparison of molecular genetic sequences obtained from living individuals in different parts of the world, so that, in that sense, the record of human population history is imminent within us.

And this is a slide from one of the very first papers to make progress in this respect. This is a slide of mitochondrial DNA from the work of Cann, Stoneking, and Wilson, when they were comparing the mitochondrial DNA from living individuals in many different parts of the world, and they achieved the first divide, as it were, from African individuals versus the rest. And this led them to the conclusion, which has since been supported by other work, that the first divide in the history of the diversity, the dispersal of our species and the diversity which comes with it, is the result of the emergence of our species, *Homo sapiens sapiens* in Africa from our *Homo erectus* ancestors, and that subsequent dispersal process. And there we see the map produced at that time, which although it's been significantly improved on by later studies, shows the history of those dispersals of our species out of Africa sixty or seventy thousand years ago, by a slow process of migration or diffusion, and the subsequent peopling of the world by our species, *Homo sapiens sapiens*.

And more recently comparable work has been undertaken using non-recombining Y-chromosome studies, which give you the male lineages and so it gives you a separate insight into these processes, and gives you information about perhaps somewhat different aspects, but the conclusions are substantially the same.

The Emergence of Homo sapiens sapiens

Well, this brings me to offer you what I consider to be one of the paradoxes, if not exactly in human history, in our understanding of human history, the sapient behavior paradox. Because so far as we understand it from molecular genetic studies, the hardware, the actual physical composition of ourselves as members of the species *Homo sapiens sapiens*, and our brains, is not significantly different, so far as we've been able to determine, from our ancestors, our sapient ancestors, who left Africa 60 thousand years ago, or if one were talking about Europe, who arrived in Europe and peopled Europe something like 40 thousand years ago. And so that notional confrontation is a very interesting one, because if you were to meet your ancestor of 40 thousand years ago, that ancestor in terms of the genetic composition and in terms therefore of the brain at the time of brain would not be very significantly different from ourselves, or at any rate that's what seems to be the case. And we have some clear understanding in outline, through fossil studies and then through archaeogenetic studies, of the history of the evolution of our species.

And so what I want to do now is to give you something of a caricature of the way the human revolution is often portrayed, the human revolution being the emergence of our species, and then its embarking on the course of cultural development which ensues. And it's often suggested, or often implied, that it is the role of the scientist to show how the emergence of *Homo sapiens sapiens* came about, and then that's the job done, that effectively tells the story. So let me illustrate this account, and this account is one of the triumphs of archaeology, prehistoric archaeology, over the past fifty or so years.

So here we have a slide, which I find a very evocative, a slide of the footprints in volcanic ash of our ancestor *Australopithecus* at Laetoli in Africa, something like two million years ago. So here you have footsteps, not yet human footsteps, but hominid footsteps, of this particular creature, where you see the skull, as I say, *Australopithecus*. But by two million years ago we have our more immediate ancestor, *Homo habilis*, and the term, the genus *Homo* is ascribed to this hominid, who was capable of using stone tools, what we would see as rather simple stone tools, such as were found by Louis Leakey in Olduvai Gorge. The finds of these tools are so far restricted to Africa; it's not clear that *Homo habilis* ever migrated outside of Africa. But the story continues and something like a million years ago you have again our more proximate ancestor, *Homo erectus*, and *Homo erectus* did indeed diffuse out of Africa and is found in the Orient, in

China, in Southeast Asia, and indeed in Europe. And it is with *Homo erectus* that much more sophisticated industries are found. As archaeologists, we have to speak mainly about stone tools at this early period, because we find very few other direct indicators of human behavior than the stone tools which are so effectively preserved.

But one of the characteristic products of *Homo erectus* in many parts of the world, including Africa, was the production of what has often been called the hand ax. And these, as you can see, and as you probably know, really are very sophisticated tools, which were made by chipping, of course, a flint core, but using very controlled techniques, so that for you or me, unless you're experienced in the matter, trying to produce one or two of these today would be a painful and sometimes rather a bloody business. And if you try it, you'd better wear gloves, though I doubt if *Homo erectus* wore gloves in the production of these tools.

The Emergence of Language in Humans

Now, one of the fascinating themes which we've already been discussing at some length, is when in the evolutionary line down to ourselves was language first practiced? Well how would you know? It's not clear how you would know, that's why we don't know the answer to the question. It seems reasonably sure that *Homo sapiens* in general, our own species, was and is capable of sophisticated language involving a wide vocabulary, involving syntax, involving tenses for the past and for the present. And we know that mainly because all living *Homo sapiens* groups have this capacity, although we diverged historically, as we were looking at from the earlier map, some considerable time ago. So we can reasonably confident that our *Homo sapiens* ancestors of 40 thousand or 60 thousand years ago should have been capable of that, first of all because the descendents in all the diverse areas of the globe have much the same linguistic capacity; despite the diversity of language, there's about six thousand different languages spoken in the world today. And secondly, of course, because we have come to the conclusion on the basis of the genetic reconstructions which we were just looking at that the genetic composition of ancestral *Homo sapiens sapiens*, 40 thousand years ago or 60 thousand years ago, was not significantly different from that of ourselves. Although it should be added that we do not yet have ancient DNA successfully recovered for *Homo sapiens sapiens* of 40 thousand years ago. Through the work of Pääbo and his colleagues we do have that for Neanderthal man, *Homo sapiens neandertalensis*, but for technical reasons the risks of contamination, or problems of contamination, it's probably likely that we will not have that documentation by actual analysis of ancient DNA of humans 40 thousand years ago that they were not significantly different genetically from ourselves. But that still seems to be a reasonable assumption.

So when would language emerge? And how would you know when it emerged? Well, there used to be an argument that you couldn't possible learn how to make stone tools of that degree of sophistication without being told how to do so. But many archaeologists now feel that the power of mimicus, of learning through imitation, should not be underestimated, and perhaps therefore it would be possible to learn and to pass on the skill of making stone tools in that way. And we're talking about half a million years ago, these hand axes are up to half a million years old. Perhaps that could be done without the use of a developed language.

Indeed to my mind the best evidence that we have for the use of language, and all of this is inferential, is that around that time our *Homo erectus* ancestors must presumably have been constructing and using boats or rafts, because in Indonesia the island of Flores was—you find stone tools of *Homo erectus* there from contact something like 350 thousand years old, and the understanding is despite fluctuations in sea level, the island of Flores was an island, in distance some thirty or forty kilometers from the nearest mainland, as it were, throughout that period. And I find it difficult to conceive how an island like that could effectively be populated by a human or *Homo erectus* population without the use of a boat or a raft, and I don't really see, even if you're a very plausible *Homo erectus* how you would persuade your lady friend, if you were a *Homo erectus* man, how you would persuade your lady friend to enter the raft without some pretty plausible story. And I don't see how that would be effected other than linguistically, although that's a matter of inference and perhaps for debate.

Well, here now is a skull, Cro-Magnon, a skull of Cro-Magnon, which was one of the early fossil remains found in France a century ago. This particular skull is some 32 thousand years old, and is of our own species, *Homo sapiens sapiens*. And when you look at the tool kits of *Homo sapiens sapiens*, they're not so radically different to the casual observer, and since I'm not a specialist in the Paleolithic period I rather think of myself as the casual observer, I don't instantly say, "Good gracious me, that is an Aurignacian tool kit, how much it differs from the Mousterian tool kit which was used in the preceding period."

But there are those enthusiasts for the human revolution who rightly point out that together with the new tool kits, the composite tools which *Homo sapiens* was producing, we do have documentations of changes in behavior, and they emphasize above all, and not unreasonably so, the production of cave art. This is one of those early paintings. This is about thirty thousand years ago from the Grotte Chauvet in France, and it is still, I think, breathtaking that you can visit the painted caves of France and north Spain and you can see these astonishing animals at Lasko or Altamira or the Grotte Chauvet which were made by *Homo sapiens sapiens* something like thirty thousand years ago.

And not very younger than that are the stone figurines, sometimes called Venus figurines, just three or four inches long, and also bone figurines. This is a very famous example found more than a century ago, the Venus of Villendorf , and so these figurines are something of the order of 20, 25 thousand years old.

So the conventional view of human origins, which I'm not disputing but I'm suggesting it's insufficient, the conventional view is that there was this great transition and that with the transition from *Homo erectus* to *Homo sapiens*, there emerged our own species, with new behaviors reflected in the tool types, certainly with that advanced linguistic capacity which all *Homo sapiens sapiens* communities share, with the capacity to go out and people the world, as we've been seeing, and also with these extraordinary products of cave art and mobiliary art, the small figurines. The word *art* is perhaps rather a modern one, but these representations of the world in paint and in stone, if we don't like the word *art*.

Well, that's all right so far as it goes, but I'd like first of all to point out that the notion that cave art and figurine art, mobiliary art, are universal features of early *Homo sapiens sapiens* in the Upper Paleolithic period, that is to say, prior to 10,000 B.C., is an erroneous one. If you look at the distribution of cave art, that Franco-Cantabrian cave art, it's shown there on the map in that shaded area of north Spain and southern France, and the dots on the map show you the finds of the figurines we've been speaking of, which are mainly in Central and Western Europe with a few further east in Siberia. And although it's certainly the case that in the Upper Paleolithic period in Australia you do have some designs on rock walls, they are not at all in the style of the Franco-Cantabrian cave art, nor do you find carvings of the kind that we've just been looking at. So it's a mistake to think of those products as a general feature of our species in the period between forty thousand years ago and ten thousand years ago.

Homo sapiens sapiens Lifestyle

And now I want to go on to show you what I think were some of the most astonishing transformations in human existence, whereas what we've been looking at up to now, the lifestyle of *Homo sapiens sapiens* in the period between forty thousand years ago and ten thousand years ago, to my way of thinking wasn't overwhelmingly different from the lifestyle of their *Homo erectus* ancestors. They were, of course, hunter-gatherers, they were living in small mobile groups or bands, perhaps groups of twenty people, and they had all kinds of wonderfully ingenious strategies for extracting food from the world. They had had the use of fire for a considerable length of time. They were indeed in the Upper Paleolithic period using bows and arrows. Certainly by then they had learned to construct boats or rafts. So there are lots of quite sophisticated elements of behavior, but they were still leading a hunter-gatherer life.

And now I want to take you to the time period around ten thousand years ago in the Near East and in Anatolia, and this is a slide from the site of Çatal Hüyük, which is one of the earliest towns, one of the earliest settled communities in human experience, around 7,000 B.C. And there you have a large town, and I could show you many slides to document that, a settled life. It's a town of many acres or hectares in area, and it, too, produced paintings, this time on plaster.

There was burial, although burial was something—deliberate burial, sometimes with grave goods—was also a feature of our *Homo sapiens* ancestors, as much as twenty thousand years ago, and the creation of effigies of terra-cotta, which in some cases it may not be extravagant to claim as deities. This lady from Çatal Hüyük—the original head is not preserved, but she's seated on a chair or throne which is supported by two feline animals, two sort of tiger- or lion-looking animals, and this is a rather strange procedure, unless this is a woman with rather particular and special powers. And elsewhere in the Near East at this time or a little earlier, around 8,000 B.C., you find new practices, new practices of burial, you find skulls with plaster faces which are difficult for us to understand, but beautifully made plastered faces, such as you find at the site of Jericho, for instance, around 8,000 B.C. And above all the settlements, some of them, there are many permanent settlements, the settlements are large-scale. At Jericho there was what appears to be a defensive wall.

In other words, you find categories of behavior that are now completely different, represent very marked developments in comparison to those of our hunter-gatherer *sapiens* ancestors of the Upper Paleolithic period. And it's clear that a very significant transformation has occurred, and it's not just in the Near East that you find this transformation. Accompanying the transformation, I should emphasize, is the development of food production, is the development of domesticated plants and domesticated animals. You find similar developments a little later in India and in China, in Egypt of course also, and you find analogous developments, again some millennia later, in Central America, in Mexico, and also in southern America. And so far as we can tell, these developments in those different areas, the move towards permanent settlement in village communities and the exploitation of domesticated plants, these are developments which occur independently one from another.

And then we have a whole series of remarkable and very different trajectories of development which lead on in each of these areas to what can be described as state societies, or as urbanization. Here from Mesopotamia is the Lady of Warka from the Sumerian civilization, what is now Iraq, somewhere around 3,000 B.C., and it was that civilization, of course, which around that time produced the first literacy, as also did the contemporary early civilization of Egypt. And here is one of the first emblems of kingship in Egypt, the so-called Narmer palette, from around 3,000 B.C., which also has some of the earliest Egyptian hieroglyphic symbols, indicating the development there of writing.

And there we move onto further developments, Egyptian pharaohs, the pharaoh Tutankamen, and so on. And I've just chosen rather casually a few slides to emphasize the very differing trajectories which are being followed in different parts of the world as social complexity develops. So here is a great temple at Ushmal in Mexico, and with the developments of the Mayan civilization and the other contemporary civilizations, again come remarkable works of figuration. This is one of the wonderful stalea from the Maya site of Yashjilan [phonetic], and as you know comes a different kind of literacy. One of the great developments in archaeology of the past twenty years has been the effective decipherment of the Maya script.

But in Europe also, in northwestern Europe, around the same time as the Pyramids were developing and the first ziggurats of Mesopotamia, we had developments of their own monumentality. This is the Ring of Brodgar in Orkney, which is about the same time as the Pyramids. And certainly if you look at some of the extraordinary built stone tombs, this is the huge site, the megalithic site of New Grange in Ireland, built before 3,000 B.C., something like 3,500 B.C., and therefore before the pyramids of Egypt, you have sophisticated accomplishments, although not accompanied by literacy.

And there is one of these very early Mesopotamian tablets, Sumerian tablets, from around 3,000 B.C., one of the first sophisticated recording systems. And then I'd like to highlight in the Indus Valley civilization, we have systems of weights, and although it's very different for the archaeologist to reconstruct the thought processes of people in prehistoric times, you can certainly see how their thought processes were working. And if you as an archaeologist, or as a student today, weigh those cubes of stone, weigh in the modern sense, you find that in a modern sense they're multiples of unit of weight. And it's difficult to conceive how that could be so unless the people in question had a system very closely analogous to our own system of weighing, and were probably using these cubes in order to weigh one commodity against another.

And very appropriate to pay respect to the Acropolis at Athens, and to Greek achievements in the fifth century B.C., the statue of Apollo from the Temple of Zeus at Olympia.

The Sapient Behavior Paradox

Now the sapient paradox is this: if we are seeking to ascribe or to explain the human achievement which I've been summarizing in this rather superficial way by referring to these trajectories of development, by the genetic transformation which accompanied the emergence of our species in Africa something like 100 thousand years ago, or the appearance of our species around the globe, and I've chosen the date of its appearance in Europe as 40 thousand years ago as a convenient time point, in what sense is the emergence of the modern brain—and I've said I don't think we doubt that the modern brain emerged 40 thousand years ago and earlier—the sapient paradox is if that's the explanation what took so long? Why did it take another 30 thousand years before you get the settled communities such as we see at Çatal Hüyük or the earlier settled communities in Mexico or in China, and before these differing trajectories of development got underway?

And that is the central question which I do not believe that molecular genetics has begun to answer for us, and I doubt if molecular genetics will indeed give us the answer. Now I'd like to summarize a few conclusions in the form of questions, and then I'll just give you an inkling of how I think the discussion may need to continue, but the discussion will need to continue of course using our understanding of the capacities of the human brain, which of course are genetically determined, those capacities which were already present forty thousand years ago. But what kind of explanation is that if you've, as I say, got to wait thirty thousand years after that for the effects to be made manifest? It simply is not an adequate explanation. So I'd like to ask a few questions and give what would be the current answer.

Was the genetic transformation by which our species, *Homo sapiens sapiens*, emerged largely or entirely accomplished by eighty thousand years ago, that would be in Africa, or by forty thousand years ago for instance in Europe? And I assume the answer is yes, I've not heard that disputed.

Secondly, is the study of human demographic history after this time, after sixty thousand years ago, by the techniques of archaeogenetics, notably lineage studies by non-recombinant Y chromosome and mitochondrial DNA data, mainly the investigation of diversity among phenotypically neutral markers, while the operational genetic unity of the species is not effectively called into question? And I think in current understanding the answer to that question is also yes. We do understand human diversity better through mitochondrial DNA and Y-chromosome markers, and that allows us to say what were the demographic processes, where were the population movements, where were the striking increases in population, when did one population replace another population? And we have all these episodes being elucidated for us by archaeogenetics. But in terms of the human genome in a general sense, did that make very much difference? To which the answer is probably no.

Thirdly, was the practice of carving small three-dimensional representations of the human form during the Upper Paleolithic period largely restricted to a tract of Eurasia from Franco-Cantabria to Siberia, while the practice of wall painting of animals at that time in the Franco-Cantabrian style was restricted to the western part of that tract, with simpler and rather different representations in Australia? The answer I think is yes. And do we know why the distribution was so geographically restricted over that time period? I think the answer is no, not a clue. I've never heard anybody give the faintest indication of why that should be.

Is there any suggestion that this behavior had specific genetic determinants in the sense that the humans in the areas where these representations were produced were significantly different in genotype or in phenotype from their contemporaries elsewhere? I think the answer is no, not so far. It could be argued that the ancestors of the Basques had some genetic endowment that allowed us to do these things, but I've never heard that seriously argued.

After ten thousand years ago, we see different trajectories of rapid development in different regions of the world. Is there any suggestion that these had specific genetic determinants, that is, specific to that region, or more precisely were the local genetic features specific local polymorphisms which made possible or facilitated or enhanced the innovations in each developmental trajectory, and which were specific to some of the participants in that trajectory, and which were not present in the participants in other trajectories? Just a long way of asking the same question. And I think the answer is no.

Human Behavior in Settled Communities

So genetic variability, genetic homogeneity, are not giving us the handle we need upon these processes. Well now, that is my main point. I'll just take two minutes longer before I conclude to tell you where I think the answer may lie, and I'm sure we shall need more insights into human behavior and how human behavior in general is genetically governed, genetically facilitated, and genetically determined. But I suggest that the crucial transformation in human experience was the development of settled, permanent communities in different parts of the world. How the conditions came about for those communities to develop is a further question, but it comes about partly from demography, partly certainly from climate change, and partly from no doubt other factors. But I suggest that once human communities were able to live permanently together, completely new forms of human interaction, social interaction, were able to develop. Indeed I think of it in terms of property, power and piety, the three *P*s. And I suggest that it was not until you had settled life that it was possible to have heritable property. And I suggest that until you have heritable property and transferable property some forms of economics are not possible. There were always subsistence economics with any living species, but until you had exchange and property of the kind that I'm speaking of, further economic developments were not possible. And you can see how that would be. If your family now lives in a house, that is your house, and you're going to dispute the right of another group to have access to that house. If you are now sowing fields of wheat and barley, you expect to reap where you have sown, and you do not expect other people to step in and do so. If you are looking after herds of sheep and goats and cattle, or your family are doing so, you don't expect somebody else to go and slaughter and eat one of those sheep on a suitable occasion.

Now these, the notion of heritable property is a new relationship. I could—I won't go on to talk about power relationships. I will just point out to you that some of the earliest images that we have in any trajectory are related to things that appear or may appear to relate to superhuman powers. In other words, the development of religious concepts and ultimately the formulation of deities—we've still got a deity on the screen—may have been an important part of that story at very early development of settled communities.

Now I've no doubt that when we have learnt to analyze these procedures more effectively, and we're talking about the formation of mind, we have the brain, the hardware, determined by that human revolution of sixty, seventy, eighty thousand years ago, but the mind, the software, comes about in those conditions that were created in those earliest settled communities.

So what I'm saying to you is, I think this is a point that's not sufficiently appreciated, it's a point that merits much further investigation, and I've no doubt that that will lead us back to ask what are the properties in the human genome, the general human genome, not talking about diverse elements in different parts of the world, what are the properties of the human genome that make it feasible for us to conceptualize? Of course the existence of language is essential, you must have language before you can have a shared understanding of what is property that is hereditable and so on.

Well I can't go on to develop that theme, but I just put it to you as a question mark, because so far the molecular genetics takes us to that tantalizing spot, but hasn't really crossed the threshold yet, and I think there is a threshold there that we have to cross.

So I'll leave you with a final image: "A word of advice, Dirk. It's the Mesolithic. We've domesticated the dog, we're using stone tools, and no one's naked anymore." Thank you very much.

Jeffrey Sachs, Ph.D., Columbia University
Plant Genomes, Food, and the Developing World

Introduction by Jonathan R. Cole

Jonathan R. Cole: Welcome back. It's now my real pleasure to introduce two of my distinguished colleagues at Columbia. First we shall hear from Jeffrey Sachs. I could go through the many features of Jeffrey's meteoric rise to preeminence within his chosen discipline of economics. I won't do that, but let me just briefly summarize some of his attributes in the past. Educated at Harvard, Jeff was at the time the youngest person in Harvard history, I believe, to receive tenure, although I think Larry Summers disputes that somehow by a matter of a couple of weeks. At Harvard, Jeffrey was a member of the Society of Fellows and the director of the Center for International Development. Of course Jeffrey has become a counselor to many nations on problems of economic development; he has become a trusted colleague and special advisor the UN Secretary-General Kofi Annan. He has been elected to many honorary societies, including the American Academy of Arts and Sciences. He has published and continues to publish scores of scholarly articles, over two hundred thus far, and takes his responsibility as a public intellectual to write on matters of economic policy for a broad audience. But that would not capture what Jeffrey is trying to do and who he is.

Jeffrey came to Columbia from Harvard only recently to become director of the Earth Institute, and the Quetelet Professor of Sustainable Development and professor of Health Policy and Management. The Earth Institute, which Jeffrey directs, is one of the largest interdisciplinary or multidisciplinary projects at Columbia, and perhaps one of the most ambitious academic experiments in the United States. There is no one that I know who understands better the interdependency of knowledge generated in various scholarly fields, if we are to address the most complex and most important social problems that are crying out for solution.

What makes Jeffrey so unique is his deep appreciation for the depth of knowledge that comes from mastering at the very highest levels a single discipline, while he also understands that without combining the knowledge discovered by these disciplines, we are not apt to solve the world's most vexing problems. Consequently Jeff has learned the language of many disciplines and collaborates with leading scholars and scientists in a number of them.

Jeffrey Sachs is interested in economic development, and cares deeply about the social, economic, scientific, technological, and medical challenges faced by the poorer nations of the world as they seek better life chances and opportunities. He is also interested in sustainability of the planet and the interdependence of biological, geophysical, and social systems. Because of all these activities, Jeffrey Sachs has been one of the most influential economists in the world, not only because of the brilliance of his scholarly work but because of his active role as an advisor to leaders of many developing nations. In my 14 years as provost and dean of faculties at Columbia, we made a great number of extraordinary appointments of scholars and scientists of the very first rank. I know of none more significant than that of Jeffrey Sachs, who will speak to us today on the subject "Plant Genomes, Food and the Developing World." It's a pleasure to give you Jeffrey Sachs.

Science and Technology in World Society

Jeffrey D. Sachs: Ladies and gentlemen, thank you so much for including me in this wonderful workshop and celebration. And Jonathan, thanks especially for including me in this wonderful university, because you played a unique role in this, of course. This move that I've been so fortunate to experience in the last year and a half is certainly the most thrilling and exhilarating part of my recent life, without question, and Columbia is everything and much more than I could have imagined and hoped it to be. The dynamism of the thinking at the university and the readiness of the university to confront the challenges of the world are simply startling and as empowering as is conceivable, and I just could not be more pleased that a year and a half ago Jonathan gave me a call, almost out of the blue as I had started to commute to advise Kofi Annan and said, "How about coming by for a chat? And I think Jonathan would also confirm that he laid out his vision of Columbia and his vision that he had championed and pioneered of the Earth Institute. It was within that meeting itself I said, "This is a good idea. Give me a few weeks to check it out but I think I'm going to do it." And it was just what he promised it would be, an environment of unparalleled intellectual commitment and more collegiality I can say and more readiness to work across disciplines than I had ever seen in the academic world before. And so I just want to thank you for this tremendous chance.

We're talking about genes and genomes and society, and I want to talk about the society part, because first I am an economist and I understand comparative and absolute advantage. I certainly have no absolute advantage in this room perhaps in almost any aspect of this topic, but my comparative advantage certainly lies on the society end.

And I want to talk about world society. We spend a tremendous amount of our time talking about our own society, our own immediate environment, and what we sometimes think of as the world, the world of the United States or the world of the rich countries dominates our thoughts, except when we're worrying about what other places are doing to us, and we don't spend enough time in the world, unfortunately, understanding the nature of a world of 6.3 billion people living in extraordinarily diverse conditions, and some in extraordinarily adverse conditions. And we have not really come to grips with how science has a fundamental role in shaping a world of such diversity and perhaps ameliorating the conditions of such extreme want in many parts of the world.

We're living at a time of a great scientific revolution that we've been hearing about from all of the speakers and that scientific revolution is already making wondrous changes in our own lives in raising life expectancy, in improving the quality of life in many ways and, yes, of course, in throwing up new and unparalleled challenges of ethical dimensions that were unimaginable even a few years before. And that scientific revolution is diffusing, as scientific revolutions do, it is diffusing to other parts of the world. In fact genetics, in its more traditional variance, has, of course, played an enormous role all through the world in the twentieth century, and genetics discoveries, even in their most practical and particularly in their most practical forms—for instance, in how to improve plant varieties—did diffuse to much of the world from the United States, from the epicenter of science in the second half of the twentieth century. But, just as with the traditional genetics and now the new biotechnologies, that process of diffusion is haphazard, incomplete, and so critical in its character and nature that we need to understand much more deeply how it works and how it doesn't work; we need to understand that the scientific enterprise is also a social and an economic enterprise, as well as a purely knowledge-driven enterprise; and, therefore, one that is susceptible of improvement and guidance for the sake of the human needs.

I want to talk this morning, therefore, about how the new biotechnologies and the new genomic sciences could do vastly more to play a fundamental role in improving the quality of life in the world, but why they will not on the trajectory that we're currently on. We can predict, unfortunately with a high degree of certainty that, if we continue in the manner of our public policies with respect to the science enterprise at large, that the failures to help to spread the benefits of science more widely will mean year-in and year-out the continuing, unnecessary mass suffering of large numbers of people on the planet. And yet, on the other hand, if we simply think a bit harder about this, at relatively low cost to us as I'll stress, we could have a monumental effect on improving the human condition more widely, so powerful are the new tools of science that we've been hearing about in the last two days.

Fighting Infectious Diseases

While we meet today, about 16 thousand people will die in Africa of AIDS, TB, and malaria. It's ironic to begin with, of course, that even under the current scientific conditions all three diseases are to a significant extent preventable, and all to an important extent are treatable. There's actually no excuse, even aside from the scientific issues, for us to be as complacent as we are day-in, day-out at the mass deaths, which amount to six or seven million per year in Africa, and add another two or three million from those diseases in other parts of the world, despite the fact that they could already be addressed with existing technologies.

But the technologies that we're talking about also have brought us to the threshold of potentially huge advances as well in finding new drugs, new vaccines, and other approaches to the control, and perhaps even one day effectively the elimination, of these diseases. But as I'll stress, we are not on a direction to cross that threshold. Science can take us so far, social action, collective will, political choice is needed to take us across the threshold.

In addition to those 16 thousand or so Africans that will die of the three diseases, another 5 to 10 thousand Africans will die today of undernutrition. And of course some of the deaths from AIDS, TB, and malaria are multifactorial, where chronic undernutrition plays an extremely important role. But undernutrition will play an even more dramatic role in deaths from diarrheal disease that would be passed off as transient and far from fatal episodes in properly nourished populations. As with deaths from other infectious diseases, the million children or so that will die of measles this year in developing countries will die in significant part because of the immunosuppression that has accompanied chronic undernourishment of very poor communities. So we face a mass crisis that is almost unrecognized, or is taken for granted, of extraordinary suffering from preventable and treatable diseases, diseases that could even be addressed much more fundamentally by further scientific advance now at hand. And similarly in a world of ample aggregate food supply, not just hunger and the pangs of extreme poverty, but actually mass death when you take a hard and cold look at it, that come from the fact that approximately one billion people on the planet still lack access to reliable diets, certainly reliable even in the macronutrient sense of caloric protein intake, perhaps up to two and a half billion people when one adds the micronutrient deficiencies of iron, vitamin A, and other micronutrient deficiencies. And again areas where scientific knowledge and current technologies could massively ameliorate the conditions, and where the new promises of biotechnology could change fundamentally these risks.

It's often summarized, as you know, by the fact that at this point southern and eastern Africa probably has a life expectancy little more than half of the United States. We're pushing eighty years now with our continuing advances in longevity, increasingly based on a deeper understanding of our biology, nutrition, and the increasingly sophisticated interventions for extending life. And at the same time, Africa's life expectancy in southern and eastern Africa is falling to around 40 years now, on the back of an absolutely unconfronted AIDS pandemic, unconfronted not only by Africa itself but by us, because we are the ones that have the technologies and the resources that could help African countries to confront this, but aside from one good paragraph in one speech of one president of the United States, we have almost nothing to show for our role in this pandemic in more than twenty years in the poorest of the poor countries. So if you feel good about our 15-million-dollar program, understand that not a penny has been spent, and that the United States has to this day, 22 years after the start of the AIDS pandemic, put fewer than one hundred people on antiretroviral drugs invented by modern science in the United States, Britain, and the rest of Europe. We've done nothing except the president has made us feel good that we're doing so much. So I unfortunately am here to make you feel not so good.

Science and Technology and Economic Advance

Now the essence of economic development when you look at it, ladies and gentlemen, is really science. The essence of advances in economic well-being, which began in an unprecedented way two hundred years ago, because up to that time while there were ebbs and flows of economic well-being, there was no such thing as sustained economic growth. Those advances over the last two hundred years have come to societies that have been able to master science and technology. It's science that brung us to where we are. We shouldn't forget it as we imperil the scientific venture every day with the mythologies, with the attacks of irrationality, that are so profound and seemingly growing in our societies as well.

Science can't be measured so easily in its direct inputs to rising standards of living, but about fifty years ago the Nobel Laureate Robert Solow created an ingenious method to ask how much of economic growth could be understood on the basis of things that economists usually talk about, like saving and investment and capital accumulation. Well, economists didn't necessarily pick up on

Professor Solow's lesson to us in that Nobel Prize-winning paper of 1957. What he found was that only 13 percent of the rise of per capita income in the first half of the twentieth century could be explained by what we call capital deepening, a rising ratio of observed physical capital per person in the country, and 87 percent was the advance of technology that wasn't explained simply by saving and investment. So Professor Solow told us that what we know today is a cliché but what we don't necessarily reflect in our public policy or in our way of thinking about the world more generally, that 87 percent wasn't the thing we argue incessantly about of whether the tax cuts are going to give this or that incentive for rich people, who already probably have more than they know what to do with, but whether we could stimulate the advance of knowledge. We don't talk about that even though we live, as a cliché, in the knowledge economy, we don't talk adequately about it. Well, 87 percent seems to be from the advance of knowledge, and hundreds of studies since then have confirmed that it is science and technology which provides the fundamental force of economic advance.

But the great lesson for the world, a striking lesson, is that that advance has been so remarkably varied across the planet, and for complex reasons, so that what is the underlying impulse of development differentially reaches the human family in dramatic ways. I brought you a couple of numbers just to give you an idea of this.

In the year 2001, the most recent data, 166,000 patents for new inventions were taken out in the United States. This is a pretty good indicator of science and technology reaching a commercializable base, commercializable state of advance. Now, the U.S. Patent Office gives patents for American as well as foreign inventors. Of those 166,000 patents, 87,000 were taken out by U.S. resident inventors, and the balance, about 70,000 by foreign inventors. Of those foreign inventors, the 70,000, Germany had 11,000, Japan about 30,000, South Korea about 3,500, Israel 1,000, all of tropical sub-Saharan Africa with its 500,000,000 people had 10 registered patents for the year. And that was up from the previous year, which is a good sign. It was 2 the previous year.

Essentially the part of the world that is struggling for survival. And it's not only Africa. It's the Andes region, which is in explosion these days from Bolivia through Peru, Colombia, Ecuador, It's central Asia, which is in explosion from Afghanistan, Nepal, Kazakhstan, Turkmenistan, Tajikistan. You name it, it's not a pretty picture. These are places that are cut off from the world of economic advance, and fundamentally cut off from the world of science and technology which is the prime mover of long-term economic change.

Poverty and Geography

Now, it behooves us to understand these facts because without understanding them we make terrible mistakes. And the most important we make in our world to our own disadvantage, I'm afraid, is that we blame the poor for their problems without understanding either the roots of our success or the paths out of extreme crisis of the poorest of the poor. We believe that the problems are necessarily problems of bad governance or corruption or other mismanagement of the poorest countries, rather than fundamental shortcomings in existing technology, the lack of ability to mobilize science in appropriate ways, and the lack of efforts to get to the deeper problems which afflict the places which have been unable to make the breakthrough to economic development.

The evidence is clear, for example, that in regions with holoendemic malaria on the planet, with high transmission burdens of malaria, there has been almost no economic advance. And those regions more than anything else are ecologically determined. Malaria transmission is based on climate and mosquito vectors. High temperatures, adequate precipitation for breeding sites, and specific species of *Anopheles* vectors that like to bite humans rather than animals are the recipe for holoendemic transmission of malaria. It turns out that one can look through the whole historic record through the genetic record indeed of humanity to understand that far before there was economic growth, technology, capitalism, the United States, or anything else, already West Africa was the epicenter of global malaria. We know it from the hemoglobin-S anomalies, which tell us that the burden of disease of malaria has been a thousands-of-year phenomenon in West Africa in a uniquely burdensome way, well before any of the current factors which we use to blame Africa for its problems existed.

When you have populations where 10 or 20 percent of the children are carrying the sickle-hemoglobin anomaly, you know that the deaths due to malaria by simple genetic principles are extraordinarily high, as they are in the humid tropical forests of West Africa, up to 40 or 50 percent of all deaths attributable to malaria, if you have population density so high of these mutant traits, which are killers in their homozygous form. And so one can look at the record and understand that the problems of the poorest of the poor, to a very important extent, result from deep geographical forces of climate or soils or disease ecology or other forces which contributed to the isolation, impoverishment, high mortality rates, and other burdens of places to prevent them from benefiting from a more general advance in the global society, which began economically around two hundred years ago.

And once one starts to realize how geographically conditioned so many of the problems of the poor are, then the question of the nature of the science-and-technology enterprise becomes more clear. The reason is that so much of the science and technology that is produced in the world by the richest countries diffuses along ecological gradients and does not diffuse across ecological space where it's not relevant. We are doing a lot of work in our country at NIH and at Columbia University and in other leading scientific and academic centers on diseases that affect Americans, on diseases that affect the whole world incidentally, such as many of the diseases now conquerable by immunization which are worldwide transmission diseases, but we do almost no scientific research or technology development on diseases that are specific to ecologies distinct from our own.

So on malaria, for example, perhaps the greatest barrier to economic development in Africa in the long historical dynamic, the amount of worldwide research underway on malaria control is probably on the order of 50 to a 100 million dollars a year right now, out of an annual biomedical research budget on the order of 75 to 80 billion dollars a year. By simple calculations, one crude measure which I think underestimates the reality considerably, puts the burden of disease due to malaria at about 3 percent of the worldwide disease burden, which would suggest an annual spending on R and D, if you made the simple calculation of a comparable portion of total biomedical research spending, to be on the order of two and a half to three billion dollars a year. We're running at one twenty-fifth to one-thirtieth of that amount of research right now.

Worldwide Diffusion of Innovation

Fundamentally, in my view, what's happened in shaping of the world economy in the last two centuries is that for a whole range of reasons that we need the full university to properly understand, a scientific and technological revolution did get started in England, and it spread to other parts of Europe. The Industrial Revolution spread to the early United States. It was for many reasons both intrinsically and by dint of historical accident a temperate-zone phenomenon; the technologies improved as the markets grew, and as the markets grew the incentives and capacity to advance the technological revolution grew further. Science was driven by markets, and markets drove science, in very subtle, important and complex ways, both through pure private market forces, as well as through public economics of an increasingly share of public and collective action, whether formal federal and state budgets or philanthropies and foundations or donor beneficence, enabling the growth of the market to support the growth of science, and then an ongoing dynamic positive feedback process, which economists call endogenous growth.

The vast proportion of global scientific advance, including the genetic revolution and the genomics revolution, was centered in the rich countries. And as wide apart are the gaps of rich and poor, in wealth today a gap of, say, one hundred to one as the little illustration about patents suggests, the gap of scientific organization is orders of magnitude greater than that.

These technologies did diffuse for the benefit of the world in many cases, where they could from a geographical and ecological point of view. Many vaccines can work anywhere. Penicillin can be of service for the entire world. But many of the most critical conditions of poverty, and principally two areas where the biotechnology revolution is most essential, public health and agricultural productivity, are strongly, if not fundamentally, ecologically centered. So only a partial public-health revolution took hold throughout most of the tropics, because the suite of tropical diseases never was addressed with the extent nor the success, nor the investment is the point I'm making, of the suite of worldwide or temperate-zone-specific diseases.

Steering Crops to Poor Nations

And in agriculture the same phenomenon is plainly evident and a wonderful study by Robert Evenson of Yale University last year has documented this more clearly than ever before. The agricultural revolution of the twentieth century, aside from the chemical revolution of fixation of nitrogen through the Haber-Bosch process, the part due to improved varieties of crops to better germ plasm, was a revolution born in the temperate-zone world for temperate-zone crops, and only incidentally diffused to other parts of the world, except in rare episodes where there was a conscious attempt to steer that development.

Asia's green revolution started in Iowa in the 1920s. It started with the development of dwarf varietals where plants were bred through traditional genetic mechanisms to put more of the biomass in the edible part of the crop and to have shorter stalks so that the crop wouldn't lodge, wouldn't collapse, as it grew at a faster rate under the beneficent nutrient bed of higher fertilization and irrigation. And through that breeding these high-yield varieties of dwarf plants for wheat and hybrid corn were developed in the United States. In the 1950s the Rockefeller Foundation had the wonderful insight to appreciate that in a world of rapid population growth and poverty, these same techniques needed to be extended to poor countries, and the first project, of course, was asking Norman Borlaug in Mexico to adopt the wheat high-yielding varieties to the Mexican context. And Borlaug did that in the 1950s, and then took that to the International Rice Research Institute in the Philippines in the 1960s, and they developed the variety IR8, which is the short-stalk rice which saved Asia and which at the core was the trigger of China's and India's dramatic escape from poverty in the last generation. Without the preceding green revolution of East Asia, there would be no Chinese economic miracle, and without the green revolution, India would have continued to be subject to the famines on a repeated basis that were expected of India even just 25 years ago, which India has completely superceded now to the point where India is a food exporter and just mildly tripped over last year's monsoon failure as if it was a passing curiosity, where it would have been a human disaster thirty years ago.

Rockefeller Foundation, Norman Borlaug, took a ready stock of technologies to these new conditions but for temperate-zone crops. Rice, while obviously a subtropical and in many cases tropical crop, fortunately has a very large inventory of technology in the temperate-zone world, from Japan, United States, and was able to be transmitted, and wheat and maize even more directly so. But there is no backlog of cassava high-yielding varieties for Africa. There is no backlog of millet and sorghum high-yielding varieties for the arid tropics of Africa. And what Evenson and Golan and collaborators found on this really wonderful study that they completed last year was that one could trace the speed of diffusion of these traditional genetic advances across the ecological zones, depending on where the translational work was less or more, and where the gap was too great, which is almost always Africa for almost every aspect of these biologically, ecologically based problems. The work often of scientific translation from basic principles to working, on-the-ground technologies has hardly gotten started.

In my opinion, this basic mechanism of science being an increasing returns-to-scale proposition— where science promotes markets and markets promote science, and science being largely an imperfect diffusing mechanism, where science diffuses in neighborhoods or in shared ecological and geographical space, but not so effectively across very different ecological and especially disease-ecological and agroecological conditions—is the most important shaper of the divides of the world today. Because the developing world, to the extent that it catches up, catches up largely by diffusing technologies developed elsewhere, translating them for local applicability, and advancing on that basis and, if they're good and lucky and very self-conscious about it, passing through the phase of translational work to innovation in their own right, as China's clearly doing right now, and as Israel, Korea, Taiwan did a generation ago, but almost no other developing countries have been successful in that. But the countries farthest away in the world in geographical and ecological space, countries that are too remote to interest anybody, usually in the hinterlands of the great continents, whether it's Bolivia in extreme disarray and chaos this morning, landlocked in the Andes Mountains at 12,000 feet above sea level, or whether it's Afghanistan, continually in disarray for about the last five hundred years, since Vasco da Gama did them the disfavor of finding a better trade route between Europe and Asia, and it's been all downhill for Asia since then. Or the interior of Africa stretching from Mali through Chad, Niger, Sudan, in the Sahel or Central African Republic, places too remote, too ecologically distinct, too poor to be able to adopt the technologies, just find themselves falling farther and farther behind.

Mobilizing Science for World Benefit

What the Rockefeller Foundation accomplished in the last century was probably more for economic development than any other organization in the world, much more, I would say, than the World Bank or than other official donor agencies. They got the idea to take science and put it to operation for development. They kept the *Anopheles gambiae* out of Brazil in the 1930s, they developed the yellow fever vaccine, and they supported the green revolution, just to name a few things that rolled off the fingertips. But the basic model of that foundation, which the Gates Fountain is now pioneering in a way in the twenty-first century, is to recognize that if knowledge really is at the core of economic development, then mobilizing knowledge through our universities, through our scientific institutions, through NIH, through our private sector, which has to be made interested in some way, because it's not interested in it on normal market grounds, may be the most effective key for finding the long-term solutions. And in my view it is the right approach because the problems that I'm discussing are not problems that markets can or will solve. Markets view the problems of the poorest of the poor as no problems at all, thank you, because one should understand that perfectly efficient markets are designed to ignore the poor. They're only designed to respond to people with purchasing power. And if you're owning shares of companies devoting huge amounts of money to impoverished people, sell the shares, and give prizes instead to these CEOs, but markets are not designed to handle these problems by themselves.

So when we think about the genetic revolutions and the genomic revolution and society, and by society I mean world society, the point I want to leave you with is that this remarkable revolution, which will do so much for our benefit, only accidentally or incidentally will reach those in most need of what it offers, unless we absolutely consciously plan for it to be brought to bear to the needs of the poorest of the poor. We have to overcome market forces, not rely on them, if this is to be done. Extending intellectual-property rights, that may have merits in certain circumstances within rich countries, but that doesn't bring medicines to the poorest of the poor, nor does it generate research within those countries. The markets are too small, the scientists are here rather than there, because the other aspect of science is that scientists like to do their work in communities, such as this wonderful university, not in isolation, because science itself in its own production function is an increasing-returns-to-scale proposition, so it's better to have communities together, which is why the lone inventors and the lone scientists in Africa don't flourish, they leave. It's not good to be the only one and somehow expect to have the high productivity, which is true in other fields where you don't want the competition, but it's not true in science. You need your competitors as your colleagues in order to get your own work done.

So we have to think of mechanisms, both in the public-health world and in the agricultural-biotechnology world, to make it possible to translate these benefits for those who won't otherwise have them. I'd mention in health very briefly that the pharmaceutical industry and the NIH has essentially looked aside from these issues during the past generation. Even the NIH itself is perhaps devoting only 1 percent of its overall budget to tropical disease, to those diseases specifically within the tropical ecologies. And for the large patent-based pharmaceutical industry and the smaller biotech industry hoping to sell to the large pharmaceutical industry, there's almost no work underway in that whole range of issues that are heavily focalized within the poorest of the poor countries.

The Example of Agrobiotechnology

In agriculture the telling fact from my perspective is that the wisdom forty years ago of establishing a worldwide network of tropical-agriculture research centers, the so-called CGIAR, the Consultative Group on International Agriculture Research, which included the units CIMMYT in Mexico, IRRI in Los Baños, Philippines, that made the translation of the green revolution to Latin America and to Asia, that wonderful conception is so starved for cash that the total operating budget of the 16 CGIAR units for the whole world of the two to three billion people who depend on the tropical crops and agroforestry and livestock that they investigate, total operating budget is about 350 million dollars which is roughly 60 percent of Monsanto's R and D budget alone. So one biotech agriculture firm, a firm I rather like, as much as it's despised in many other parts of the world, I think they're doing excellent science and deserve the praise for it, not the opprobrium that they've received. It's still one company, which is a modest part of the overall agriculture research enterprise of the United States, is roughly 1.6 times the entire combined tropical-agriculture research budget of the public system in the world. What we're seeing, therefore, in agrobiotechnology, the same way as in medical technology, is markets driving science, science driving markets, and the goods not reaching the poorest of the poor.

The agrobiotechnology advances in recent years have been almost entirely temperate zone. They have involved two major discoveries and innovations commercially, implanting a bacterium gene, the Bt gene, which is a disease-resistance gene from the particular bacillus that holds this, it fights off the corn bore and other known attackers of major crops, that was successfully implanted into corn, cotton, and other crops by Monsanto, it turns out. And also the introduction of herbicide-resistant genes against something that is done and was done through normal genetic manipulation, but now is done in a much more targeted way through targeted identification and transfer of specific gene sites. So it's old breeding methods simply updated, nothing new it terms of its actual nature within the plants. Both of these have been introduced to tens of millions of hectares of planting in the United States, Argentina, China, and soon Brazil. Why those countries? Temperate-zone crops, temperate-zone application. We see the same process of diffusion taking place that where the backlog exists is where the rich world has made the backlog. But the amount of work that could go on to get transfered saline-resistant genes that mangrove crops have, that could be used for the saline-stressed environments of tropical irrigated rice plots, or genes to target specific bacterial rust threats for the crops of the tropics, that has hardly begun to this point.

And similarly the advances, which we know again are on the threshold of bio-fortification, for example, genomic transfers of genes that can produce the pro-vitamin A, that we metabolize into vitamin A, we know how to . . . we! —I always . . . how nice, the economists really know a lot . . . I'm told that the scientists know it and I've seen the labs where they do it, but it's not being done again, because the places that need that bio-fortification lack the funding to do it.

The Need for Collective Action

I'll just end by saying that this misshapen world in which at the same time we have the prospects of unparalleled and unrivaled scientific advance, and at the same time we have one billion people fighting for daily survival, and thousands every day losing that fight because of extreme poverty, is an understandable situation. I hope I've helped to clarify it analytically, but it is a shockingly inadequate situation for the world. It leaves a world of immense suffering and pain, and it leaves a world of immense instability in its wake. We find ourselves inevitably dragged into the problems of anywhere in the world. If you would ask an American three years ago "Where's the place in the world least likely to bother you?" you would've put your pointer in the middle of Eurasia, said "Well, at least that place is too poor and too far away and too remote to reach us," probably right at Kabul. And the world, however, doesn't operate that way anymore. The diseases come to haunt us, the AIDS pandemic which began as a zoonosis seventy years ago in the jungles of West Africa, according to the best genetic clocks we have, didn't respect Africa's continental edge, it has traveled through the world. And the instability of extreme poverty also finds its way, not only through terrorism but mass migration, refugee movements, disease, state failure, criminality, drug trafficking, and a hundred other ways, to affect our interests most directly.

We have to understand how much we have the opportunity to help shape our own future. One foundation, the Rockefeller Foundation, probably did as much as anything to feed billions of people on the planet, and that was a relatively small, extremely well-targeted investment. Bill Gates is doing his part right now, but I've had the enormous pleasure to be able to say to him on several occasions, "Bill, even you can't do it by yourself," which is a very empowering feeling, because he comes as close as anyone in the world literally to being able to do it, but he can't. His foundation spends about a billion dollars a year in total, roughly 800 million dollars a year for global health and related activities right now. It's monumentally important. What's shocking is that he single-handedly outcompetes the United States in this effort as a whole.

We have a lot to do, and the evidence is that for much smaller amounts than we're devoting to keep those soldiers in Iraq, we'd be able to change the tide and save millions of lives abroad every year, and truly make the world a safer one for the twenty-first century. Scientists are the ones that have brought us to this point, but it's going to be all of us that take us across the threshold.

Thanks very much.

Philip Kitcher, Ph.D., Columbia University
What Genes? Whose Genomes? Which Society?

Introduction by Jonathan R. Cole

Jonathan R. Cole: Our last speaker is Philip Kitcher, and Philip Kitcher is one of the world's most distinguished philosophers of science. Professor Kitcher was born in England and obtained his B.A. from Christ College, Cambridge, and his Ph.D. from Princeton University. He has published over a hundred scholarly papers, but more importantly his work on the philosophy of science, in particular his books—*Abusing Science*, *The Case Against Creationism*, *The Advancement of Science*, *The Lives to Come: The Genetic Revolution in Human Possibilities*, and *Science, Truth and Democracy*—are considered path-breaking discussions of the nature of science and its relationship to larger social values. It is worth noting that many of the leading philosophers of science over the past fifty years had backgrounds in mathematics and physics and focused their attention on those fields. Philip Kitcher has focused to a significant degree on the biological sciences, with great interpretive results. Philip currently is the John Dewey Professor of Philosophy at Columbia, and he has been widely recognized for his work, including election in 2002 to the American Academy of Arts and Sciences. Beyond all of this, Philip is a remarkable teacher and colleague, and Columbia has benefited greatly by his move here in 1999. It's a pleasure to present to you our final speaker in this program, Professor Philip Kitcher, who will talk to the questions "What genes? Whose genome? Which society?" Philip.

Two Key Revolutions in Science

Philip S. Kitcher: Thank you very much, Jonathan. It's an enormous pleasure to be here today, and it's an enormous pleasure to take part in this celebration, even though I'm afraid, like Jeffrey before me, I'm going to end up by being a little gloomier than seems appropriate to a party.

I want to begin, though, by looking back into the past, back into the pre-Colombian past, as it were, and to an episode in history that was really quite remarkable. And I want to invite you to imagine yourselves as living in Europe and being intellectually well-connected in 1643. Now, looking back on the intervening century, you might feel inclined to celebrate the work of a Polish monk, Nicholas Copernicus, who'd published—or rather had been handed on his deathbed—a copy of his groundbreaking work on the revolutions of the heavenly bodies. Now you might also reflect on the achievements of Signore Galileo who had invented the telescope, and who had made remarkable discoveries in astronomy, who had written a very controversial book in 1632 for which he'd been summoned to Rome and shown the instruments of torture. You might also know something about Herr Johann Kepler, whose important announcements about the movements of the planets were buried in abstruse numerological speculations in his *Astronomia Nova*. And you might know of some of the work being done in France and Holland by people like Monsieur Descartes. But there's a lot you wouldn't know.

You wouldn't have known a year before, the year in which Galileo died, in a provincial town in a very provincial country, a boy was born, on Christmas Day, appropriately enough, and this boy would go on to revolutionize mathematics and write down equations governing the motions of bodies that could be applied to a physical system, not just of bodies on Earth, but of the entire universe. You wouldn't also know that Galileo's dream of an institute of learned investigators would be realized in 1662 in the founding of the Royal Society. And you wouldn't know that the revival of atomism would come to dominate the fields of chemistry, the physics of heat, and so on and so forth. So even though you might think of yourself as in the middle of a time of immense intellectual change, you'd be missing all sorts of important things about where it was headed.

Now I want to compare that with the only revolution in the history of science so far that compares with it. These are I think the two most important transformative episodes in the history of the sciences. And I begin not in 1953. I was glad to hear yesterday Sydney Brenner slightly downplaying the importance of 1953, but with the 1940s, and in particular with the experiments that identified DNA as the genetic material. Those experiments, the great accomplishment of Watson and Crick, the identification of the roles of RNA, the discovery of the genetic code, recombinant DNA, and the very important and I think underrated work, even though she won the Nobel Prize for it, in integrated molecular genetics with development by Christiane Nüsslein-Volhard, the Genomes Project, the Human Genome Project, all of these have unfolded enormous promise, and yet for the 2000s we don't really know where it's going. We're in the middle of something, it's very big, it's very exhilarating, and it holds enormous promise.

The Human Genome Project

Now I want to start by reflecting on the Human Genome Project, which is the most visible symbol of where we are right now. And I think the Human Genome Project, not the Craig Venter's Genome Project, is the important word there to stress. I'll come to that later on.

Now, luckily people no longer characterize this project as the search for the grail, the thing that will make us understand who we truly are. We know from what we heard from Dr. Svante Pääbo yesterday and from Colin Renfrew this morning that there's an enormous amount of work to be done beyond simple sequence deciphering in trying to figure out what is special to our species, *Homo sapiens*. It's an immensely complicated task, and as we look into the future we don't know how far we'll be able to get with it, or exactly how it will unfold.

The other great advertisement for sequencing the human genome was the idea that we will be promised cures, treatments, for various forms of disease. And as in the case of the rhetoric about unfolding the essence of humanity, here too recent discussions have offered a much more sober assessment. It's plainly one thing to identify the allele or alleles that are involved in a particular disease, quite another to devise a reliable means for addressing that disease, to bring relief to those who suffer from it or to forestall it. Knowing the sequence of a normal allele and the mutants that give rise to people with disease and disability, you can start to take steps towards probing the molecular mechanisms that are involved, and with some luck you can devise useful strategies of intervention. And the story that was told yesterday about cholesterol, the unfolding story across several decades, really brings home the importance of how much ingenuity, how much clever work, is needed to fathom these molecular mechanisms.

Sometimes we will get there, but the time scale on which we will achieve the results that we hope for is quite uncertain. I think we should not despair; after all, there are success stories, like those that were told about cholesterol and heart disease yesterday. But it is also worth pondering on the other hand that the molecular basis of sickle-cell disease has been known since the 1940s without an appreciable amount of progress. I like to think of us as buying tickets in a very large number of lotteries, so many that we would be unlucky if we didn't win some prizes, but where we will win these prizes and how big they will be is at the moment rather uncertain.

Now a third common image of the Human Genome Project is that now we are on the verge of a new era in which gene therapy will be widely available and in which some people fear we will have the capacity for engineering life. Now I think Anne McLaren pointed out very clearly this morning that these ventures are often misunderstood and overblown. Current ventures in gene therapy attempt to solve the problem of dealing with people whose cells contain mutant alleles. And what one tries to do is to insert into some of these cells a copy of a normal allele that will generate a needed protein. Now that doesn't turn an abnormal cell into a normal cell, what it does is it replaces one sort of abnormality with another. In some cases the predicament is so severe, as with the kids who suffer from SCID, the bubble babies. In that case it seems worth injecting some DNA or sending some DNA into the cell because that is the only way we know of relieving an incredibly difficult problem. But we are a very long way away from the stage at which we can hope to offer the world genetically enhanced people. We're a very long way away at the moment from knowing how to restore people to full health by these means.

Now in all of these cases the upshot is the same. We are going to need a lot more knowledge. The Human Genome Project is only the beginning. And the good news is that, as our speakers have explained over the past day and a half, there are lots of promising ways of going forward from where we are, most of those ways involving returning to other organisms, organisms that are more tractable, organisms like *C. elegans*, or the fruit fly *Drosophila melanogaster*. These are wonderful organisms to investigate because in part brilliant experimentalists have already discovered so much about them. And it seems to me that the fanfare that attended the sequencing of the human genome would've been much more appropriate when the genomes of nematodes and fruit flies were sequenced.

Now if you look back not to the pre-Colombian past, but about a century, after the rediscovery of Mendel's laws, scientists who were interested in heredity and in the application of new understandings of heredity to human problems faced one of two alternatives. They could go full bore ahead trying to apply those things immediately to human medical problems, and indeed some people favored that, or they could take an indirect route. And of course the most famous person who took that indirect route was Columbia's very own T. H. Morgan, to whom Dr. Fischbach referred at the very beginning. He set up the Fly Room in Schermerhorn Hall—and let me put in a plea at this 250th anniversary, won't somebody please turn that room into a museum, it's one of the greatest rooms in the history of American science—and Morgan's indirect route eventually made possible the molecular discoveries of the 1940s and 1950s. I really enjoy telling my students in Philosophy of Biology here that, you know, a sophomore at Columbia, Alfred Sturtevant, discovered gene mapping. Why can't they do something equivalent?

Genetic Tests for Diseases

Now the moral of what I've said so far is I hope obvious. We're in the middle of something, the future benefits for which we hope are not simply the consequence of what has been so far but also knowledge we will obtain, knowledge that will come crucially from investigations on nonhuman organisms. But of course we do know a lot right now, we have knowledge that we can apply right now. And in the rest of my talk I want to consider that knowledge, that ability we have right now, and how we might use it.

Now one of the great advantages of having sequence knowledge and being able to correlate bits of sequence with various diseases is that doctors are now enabled to perform genetic tests. And here we can distinguish two basic types of tests, diagnostic and predictive. And there's been a great tendency, I think, in popular presentations about the Human Genome Project to underrate the value of having diagnostic tests. In many cases having a quick way to diagnose a disease that was typically previously difficult to diagnose is a great boon, and there are also great advantages if a new genetic-testing technique enables you to disambiguate forms of the disease so you can treat the patient more appropriately. But of course the great excitement has been on predictive testing, the idea being that one will be able to take some sample from a newborn infant, perhaps, and then predict the disease susceptibilities down the road.

Now the trouble with these forms of advance testing is that in all too many cases knowing in advance that a person is susceptible to a particular disease does not enable you to offer any particularly good advice to that person for what they should do other than advice that you could've offered them anyway. This is perhaps most stark in the case of Huntington's Disease. In advance of the discovery of a genetic test for the disease, when Huntington's at-risk people were asked, you know, "Do you want the test?" close to 80 percent of them said they would actually like to have the test. Once the test was available it's about 17 percent of them who want to take it. And that is because even knowing what we know today about the genetics of Huntington's Disease there's nothing very much that can be offered to them.

Now over this entire area, this area of genetic testing, hangs a cloud, and that's the cloud of potential genetic discrimination. Information about risks is in some cases potentially valuable to somebody, but that information is useful not only to the individual but to others as well. Employers might want to have it, insurers might want to have it. And anybody like me who has given public talks on genetic testing over the last few years has encountered lots and lots and lots of people in this country who come up afterwards and tell a story, either about themselves or their family or about somebody they know who has been denied insurance or has lost a job because of some genetic testing, some obvious genetic problem.

Now many of the problems that arise in this area could easily be solved. And in the early days of Nancy Wexler's pioneering work on the ethical, legal, and social implications of the Genome Project, it was assumed that they would be solved, and solved quickly, that we would have in this country a comprehensive system of health care which would enable people to get out from under the burden of potential genetic discrimination. It seems to be scandalous that in the society in which genetic testing is most likely to be prevalent nothing has been done to set any such scheme in place.

Social Implications of Prenatal Testing

Now the trouble with this problem is one that actually infects many of the social problems that arise with respect to the implications of our new genetic abilities. And that is that the problems here are not readily detachable from wider moral, social, and political problems. It was assumed in that grand gesture of giving funds for the investigation of the ethical, legal, and social implications of the Genome Project that one would somehow be able to detach these problems, to solve them, as it were, piecemeal without their being tied up with other larger political problems. That is not the case. And the insurance issue brings it home. It's because we don't have the will—the collective political will as a society, to do the obvious things to ensure people against diseases and disabilities—that the task of people investigating how to cope with our genetic knowledge has been made not difficult but impossible. There seems to me no satisfactory and adequate way of taking the era of genetic testing and putting into it adequate safeguards and protections from people who may suffer future genetic discrimination.

But this I don't think is the most serious problem facing us in North American society right now. Far more worrying is the problem posed by the possibility of testing prenatal, and that is a genuine problem. And it's a problem but also a cause for some rejoicing. The incidence of Tay-Sachs Disease worldwide has been reduced in the past thirty years to about one one-hundredth of what it was. And that is something for which we should all be profoundly thankful. And that's because of a kind of eugenics. And I'm not actually afraid to use this word because I think eugenics is now forced upon us. We are forced to make decisions about how to use the knowledge we have in order to consider the ways in which lives come into the world. We have no alternative. Even if we said we won't use this knowledge at all, that would be in itself a kind of eugenic decision, because we would be saying, "Look, we value babies born in ignorance more than we value the idea of administering tests to find out about certain kinds of genetic conditions." As Anne McLaren rightly pointed out, what is crucial here is that the form of eugenics we practice should be one in which there is no coercion, no social coercion. Unfortunately I think that is actually more difficult to achieve than one might think.

Now there are dreadful sufferings that are inflicted on human beings and their families by rare, single-locus mutations, okay? Tay-Sachs is one of these, there are many others. Fortunately they are very rare and it seems to me it would just folly on our part not to make sure that prenatal tests for these are widely available so that people who choose can terminate pregnancies when they find that one of these conditions is present.

On the other hand, I suspect that within a decade we will be able to test fetuses cheaply for thousands of conditions, including the possibility of testing for alleles which affect all sorts of things and are correlated with temperament, behavior, and mental capacities of all kinds. Now if you think about this in the context of present American society, it leads to some obviously very troubling questions. After all, we live in a city in which many people I know are profoundly worried about what schools their children will go to, and before that they're profoundly worried about what nursery schools their children will go to, and before that they're profoundly worried about what daycare centers their children will go to, and so it goes. Those same people are likely, I think, to be profoundly worried about their children's genetic material.

Now we know that people in other regions of the world want to use various kinds of prenatal tests in order to shape the kinds of offspring they have. In India and China, for example, there are places with remarkable skewed sex ratios because amniocentesis is used to determine in advance the sex of an unborn child. Now the same kinds of things will become increasingly available to people who are concerned enough and well off enough in our society. And some of these things will not be decisions that are made entirely freely, but they will be made on the basis of a kind of social coercion. The prospective parents will say to themselves, "Well, you know, I'm actually not prejudiced, as it might be, against short people or people who are attracted to members of their own sex, or whatever it is, but you know, we live in a prejudiced society and the best I can do for my child is well, possibly to go in for massive in vitro fertilization with genetic selection of which embryos to implant."

Genetic Determinism of Behavior

Now everything I've said precedes on an assumption, and the assumption is that one will indeed find the kinds of genetic connections between alleles, parts of the genome, and forms of human behavior. Now I think Dr. Cori Bargmann was absolutely right yesterday when she said, "Of course there are genetic factors that are influencing behavior." And it's also likely that we will discover genes or forms of genes that in particular environments express themselves in particular ways. But here we have to beware, we have to beware of the casual assumptions of genetic determinism.

About a decade ago the first Bush administration became quite excited by reports about genes of crime. Molecular genetics, it seemed, looked very useful to it in the war on crime, and there was a nascent project to pursue this line of inquiry. Nothing much came of it, but it was pursued in what in now seems to me just an amazingly deterministic and socially unacceptable way. The idea was to try to find genes for crime and then presumably to try to identify people who bore these genes and nip criminality in the bud.

Now this is both, it seems to me, dubiously ethical and methodologically misguided. Plainly there are correlations between crime and various gross environmental variables. Certain kinds of crime occur most frequently in degraded urban environments, places in which people don't have much chance of gaining the rewards Americans look for. It would seem to me to be extraordinarily foolish to suppose that there is some genotype which is both a cause of crime and that has as a pleiotropic side effect the tendency to make those who bear out seek out degraded urban environments. Now there are better ways to use our molecular techniques in the analysis of human behavior. And again Dr. Bargmann's presentation yesterday on a similar problem in nematodes pointed the way.

For years behavioral geneticists have struggled with the difficulty of finding monozygotic twins reared in suitable different environments. Well with molecular genetics there's the promise of doing better than that, the promise of understanding how when people with the same genotype at a particular locus or a particular set of loci grow up in different environments, their characteristics are expressed differently, and thereby using the genetic tools as ways of dissecting the environment. Now Dr. Bargmann showed yesterday how the feeding behavior of the nematode worm can be understood in terms of a response both to genotype and to environmental variation. Doing that in the nematode was a difficult task, doing it for an organism with as complicated a social environment as ours is going to be much, much harder. But, it seems to me, that that is the direction in which molecular behavioral genetics ought to go.

The Fair-Share Principle

So let me sum up the perspective I've been developing, and you'll be relieved to hear that I've now said most of what I want to say, but it's not the whole thing.

This is what I'll call the perspective from the affluent world. First of all, and let's not forget it, we're in the middle of a great genetic revolution; it's one of the two great transformations in the history of the sciences. It's not finished, there's a lot more to do.

The next point is that down the road we are going to get, I think, relief from some diseases, but what we have now is an ability to predict and to test. And I think we shouldn't underrate the diagnostic tests. Now the age of testing already brings social problems, and it brings them most especially for those who are ill served by current institutions in our society, and it's most obvious in our lamentable failure to assure health care and disability insurance for all people.

And the hard problem we face is to understand prenatal genetic testing wisely, and it's going to be difficult, I think, to craft a benign form of eugenics, and it's going to be difficult to pursue human behavioral genetics in a methodologically sophisticated and ethically reasonable way.

Now I could stop right there, but that I think would be profoundly wrong, because I think we need a more global perspective on this, and here I'm going to connect up with what Jeffrey Sachs said, and it's interesting. The two of us actually made up our presentations independently, and I'm afraid we're going to sound as though we're sounding the same themes. I apologize both to Jeffrey and to Roy Anderson for the crudeness of the statistics I'm showing. I'm sure they could offer more refined things. These are from basic World Health Organization reports over the last few years.

You look at the annual burden of death, it's roughly sixty million, and from disease roughly forty million. And if you look at the way this is distributed, here are some diseases with their burdens. Now as Jeffrey pointed out, these deaths, or deaths from particular diseases, are very much skewed to different zones around the world. And if we look at the case of tuberculosis, for example, in the United States it's virtually negligible, whereas if you take the other countries I have listed on this slide about a million people die each year from it.

Now if you actually look at the ways in which research dollars are funneled into biomedical research, the NIH offers almost half the total of seventy billion. I think this may be a little bit out of date. And Africa, Asia, including Russia, and South and Latin America altogether contribute less than three billion. It comes then as no surprise that some diseases, especially infectious diseases, are radically underfunded. Now I think Roy Anderson made it very clear yesterday how genomics points to all sorts of possibilities in forestalling or fighting tropical infectious diseases. One can't just say, "Look, there's no chance of using our current molecular tools for doing this." Combining what we know about genomes of pathogens, genomes of hosts, ecological conditions, there must—there must be ways of proceeding it seems to fight these diseases. And yet very little funding is given to them.

Let me define something that was implicit in Jeffrey's talk. There's something to which I feel we should subscribe, I mean this is very roughly, and I'll call it the Fair-Share Principle. And that is the idea that biomedical research allocations ought to accord with the fair shares of diseases. Now I'll define a disease's fair share as the total research dollars, that's seventy billion, times the number of deaths due to that disease, divided by the total number of deaths from disease. Now obviously the Fair-Share Principle is crude; one could offer a different measure of the suffering produced by disease, such as the number of years lost or something like that. There are alternative methods, they don't make a lot of difference to the argument. One could also adjust the Fair-Share Principle in the light of promise of research, and I think one would be hard pressed to claim that infectious diseases are less promising as targets of molecular research than the diseases on which we lavish so much attention.

Advancing Worldwide Social Justice

Now this is just to recapitulate some of Jeffrey's points. The figure I got for research on malaria was 85 million dollars, and it's fair share would be roughly twenty times that. For tuberculosis the factor is a hundred to one; that is, it gets a hundredth of the amount it should. For the assorted respiratory diseases that kill so many people, they're radically underfunded, and the same goes for diarrheal diseases.

Now I just want to make this point sharply and starkly. The people who suffer from these diseases are far away. The methods we have to secure ourselves from these diseases, the public-health measures that have been so successful in the affluent world are not exportable to them. We can work to try to make those public-health facilities exportable to them, and so we should. But in part they're not exportable because in fact they don't trust our intervention. And one way of proving our bona fides would be to invest in the diseases that concern them. Just because they are far away doesn't detract from the fact that they are still human beings. That's what I said the Human Genome Project was such an appropriate name and such an appropriate symbol. We have an obligation to them.

This may also be couched in the terms of prudence, as was said yesterday by Roy Anderson. We live in a world in which diseases can be transmitted very quickly from one region to another because of the ways in which people travel about, so just prudentially we might want to do something about some of these diseases. But I think the fundamental argument is a moral one. We have a moral obligation to do something about this, and no discussion of the future of molecular medicine and molecular possibilities can be complete, I think, without acknowledging that obligation and pledging ourselves to remedy the dismal conditions of current distribution of biomedical research funding, where the funds that the NIH spends on tropical medicine are almost a rounding error in its budget. And, I may say, the diseases that it studied under that rubric are very largely diseases that afflict people in places to which American forces are likely to be sent.

Now I want to end with a last turn of the screw, and it echoes another of Jeffrey's points. Because of this acute imbalance in the world, because of the perception of the affluent world, and in particular the United States, as callous and indifferent, as not using its fruits for human well-being broadly construed, we live in a world of great instability, and that instability is only going to get greater. The eminent British cosmologist Sir Martin Rees has made a bet, it's a bet he doesn't want to come true, and that is by 2020 a million people will die in some bioterrorist attack.

Now you've heard a lot very casual talk over the last couple of days about inserting genes into this organism and that organism. And that's because the technology has gotten so good and so easy. And it's easy enough that a group of people designed to go for some bioterrorism would, I think, before 2020, probably be able to manufacture a pathogen to meet their needs. And what are those needs? Well, in an age in which people who believe in "the true cause," whatever it may be, are prepared to blow themselves up for that cause, they can also be prepared to serve as disease vectors. And all you have to do is find the right combination of factors to produce a pathogen which makes you not conspicuously infectious, that is, no more infectious than anybody who's suffering from a common disease, and get yourself into a position where you can by ordinary means, coughing, spitting, handling food, shaking hands with a lot of people, whatever, infect a large number, and you have a ready-made device for bioterrorism. That is why Rees' bet I think is not a stupid bet.

Now it would be silly to say that, you know, we have to shut down biology. Can't do that, that's just obviously the wrong solution. What we have to do is think about the ways in which biological resources might be used to guard against these possibilities. But even that by itself is not enough. We all have to do something, we have to pledge the social conditions into which these wonderful biological tools are being thrust. And what needs to be done, it seems to me, is much greater attention to social justice on a global scale. How can we expect other countries, countries whose populations are suffering from terrible burdens of disease to control, to help us police groups or individuals who are out to wreak their way through hideous acts of bioterrorism?

Okay, now I don't want to end completely on a morbid and ghoulish note. Let me go back to the beginning. We are living through one of the great transformations in the history of science, I think as great as what is commonly called the scientific revolution, which is why I began with that. But there's a difference. If one held a comparable meeting in 1643, predicting the future would've been a matter of sort of party-time gossip or idle curiosity. But for us today at the beginning of the twenty-first century, science matters. We have to think about where it's going, acknowledging its unpredictability, and we have to try to develop new ways, in terms of social structures and in terms of scientific research, to make the world a safer, juster, and better place.

In the early 1660s, just after the birth of the Royal Society when Boyle invented the air pump and started to conduct experiments on air pressure, King Charles II is reported to have laughed that the learned gentlemen were weighing the air. We may be exhilarated by what's going on in laboratories today, we may be frightened at the prospects, we may be determined to overcome them. One thing we shouldn't be doing is laughing.

Thank you.

David Hirsch, Ph.D., Columbia University
Concluding Remarks

Jonathan R. Cole: I want to thank you very much. I want to once again thank all of our speakers, and we have run beyond what we expected to in terms of time, but there is one more duty that I want to perform which is to turn the stage over to my good friend and colleague David Hirsh who will have some final concluding remarks. I must say that as I do that I marvel at the ability of our speakers—these scientists, philosophers, economists—to describe, analyze, and explain phenomena. Their increasing understanding of nature of behavior is absolutely extraordinary to me, but I wonder whether any of them can give me as we leave a highly plausible explanation for "The Curse of the Bambino." Thank you.

David Hirsh: Now that is a genetics problem. I just wanted to say a few words in these closing minutes, I know we're running behind schedule. Is this not . . . we all right now?

First, I do want to make a couple of formal announcements. One is on behalf of Columbia University, I want to thank once again people who have been responsible for this. I know you've heard the names a number of times, but at this moment I think it's appropriate we say it one more time. To Joanna Rubenstein, who really put in an enormous amount of effort and good sense in putting this wonderful program together. Tom Jessell for his usual clarity of thought, constructed it in this way that has made it such an exciting program. And of course to Gerry Fischbach, the vice president of Health Sciences. And of course to Lee Bollinger, who has been at the pinnacle of this intellectual endeavor. And it's on that note that I just want to make a couple of closing remarks.

Here you've heard the history just a moment ago of the great scientific revolution and now the genomics revolution. What do we look forward to the future, and in particular what do we look forward to at Columbia? We're in this privileged position to listen to all of these speakers and perhaps do something about it, because at this moment we are in a phase of development, of growth, of wanting

to expand the natural sciences at Columbia. The natural sciences as they exist in the 9 departments at the Morningside campus, and the 9 basic-science departments in Health Sciences, in addition the 18 clinical departments at the Health Sciences, and most importantly to listen carefully to what we just heard in order to find solutions by combining these sciences in new interdisciplinary ways, both at the basic-science level and then between the clinical and the basic sciences, to render some of these new advances in a more accelerated form and a more applied form.

So what I have heard going forward . . . I would like to think that what we have a responsibility to do is to expand upon our great strength in neuroscience, to understand better all of these phenomena at two kinds of levels. One is—and I'm happy to say it said to me that molecular biology still exists as a powerful discipline—one is to understand the individual elements, the genes, their readout, and in particular now to start to examine the noncoding regions of the genome, as Mike Levine pointed out to us, and to start to uncover the regulatory elements, whether there are a set of codes there that can help us going forward, or to reach into biology with the cell as the basic element, as Sydney Brenner expounded, and to start to use new physical methods, perhaps from the physical sciences, to understand more about those units that are incidentally of the nanoscience dimension—the ribosome, the spliceosome, the topology of the inside of the cell—and to make that the defining feature.

But then we heard from a number of people here—Eric Lander, Roy Anderson— that really what we're dealing with too in this new era are populations, and the analysis of populations, polymorphisms. And we also heard a note of caution from Roy Anderson, be careful in these multigenic diseases, they're going to be difficult and right now probably impossible to decipher through population genetics, difficult.

But almost everything we heard relies on both of these, understanding the element. You just heard reference once again to Mike Brown and Joe Goldstein's seminal work on the LDL receptor, where it was basic molecular biology applied to a disease phenomenon, but what it is now becoming is a population problem. What are the modifiers that even once you know the mechanism of all of the components of the LDL receptor and regulation of its synthesis, where are the modifiers within the population, that lead to different responses to your fat diet, and where do we look for them? That will be a population diagnostics effort going forward, and I believe we are poised to develop both of those approaches at Columbia.

Cori Bargmann also showed us the combination of the two, and as we go forward in this area of trying to link genes to behavior once again, for example, in the serotonin reuptake receptor, there we have a definition of the molecule, there we can identify polymorphism within a population, as in the New Zealand study. But the overriding future of that will also rely on population and the interaction between the individual and the environment, which is going to have to be done not only at a unitary gene level but as a population phenomenon.

So there we have this combination, and that is what we're going to have to prepare ourselves for going forward. And I think we're going to have to do it particularly in neuroscience, genes and neurons, understanding the structure and function of the brain, perception, taking in this environment, processing it, and the link, as I've just referred to, between behavior and the environment. Genome regulation will undoubtedly continue to be studied here. But now I believe also Columbia must garner its strength not only on top of its strength in neuroscience, but the relationship between neuroscience as we understand it, behavior, and behavior in its broadest form, but also now we have to stretch into infectious disease in a more powerful way.

The last two talks brought us to this issue of our responsibility as a great university and the global view of science. And I think Columbia is well prepared and has a history of participating in that way. But I do believe also, in addition to participating in the ways that both Philip Kitcher and Jeff Sachs rallied us to address, we have to be careful, we have to be careful that our job first and foremost is creative scholarship, that we are careful to apply our best efforts not to build yet another iron lung, to use that polio example, but in fact to develop the polio vaccine, to address each of these kinds of problems facing the world today.

It's been a pleasure to be here throughout this wonderful symposium celebrating this remarkable time in history, this remarkable time in science and this remarkable moment, the 250th birthday of Columbia. I would harken back to one reminder that in 1754, when this university was founded, it was faced with some of the very kinds of epidemiological and intellectual and moral questions that we face today. In 1754 Boston was hit with an epidemic of smallpox; in 1775 the whole east coast was hit with another epidemic of smallpox. More people died in that smallpox epidemic than in the ensuing Revolutionary War. At that same time Ben Franklin and his intellectual contributions to science and politics were here is also the moment in history when Linnaeus brought us our classification system in biology. The problem was that one of the great infatuations at that time was to define race and the meaning of race and to justify some preconceived notions that are rather distasteful. It was founded at a moment of infectious disease, of philosophy of science, and of science, but at this moment we have heard too in this past day and a half how moving forward we have an opportunity to springboard from these great discoveries that you've heard so much about, and I want to thank each of the speakers for their contributions to this moment, and I want to thank each of you for being here.

I also would like to announce that this event has been archived on the 250th Web site where alumni and others can also now review it or view it for their very first time.

Thank you again. It's been a marvelous day and a half. I hope we, as a great university, fulfill some of the promises of the technology, of the intellectual activities that were described in this past day and a half. Thank you.