Genes and Genomes: Impact on Medicine and Society

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 100,000 that I teach my, that I taught my students at MIT for a decade, more like 30,000, and 30,000 increasingly looks like an overestimate, I think we're clearly into the 20,000s, 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, 300,000 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 paradoxus, mikatae and bayanus, related species that we think differ by up to about twenty, 30 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. Ah ha, 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, "Ah ha, 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 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, paradoxus, 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-1α, 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 10 million variations, common variations in the human population, and more than 5 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.