ATUL J. BUTTE: So I can and have in the past talked for about 6 hours on this subject. Today, we're just going to talk about the first of these, microbiology for the [? bioinformaticist. ?] And if we have time, then we can talk about gene measurement techniques, not just microarrays, but all sorts of different technologies depending on how fast or slow we go. So let's just get started then.

So I have about 10 slides of basic biology, OK? If I see you start to yawn, then we'll just go faster. There's a lot that we can cover. So how many of you remember this kind of thing?

Do you remember this? Do you remember this? Do you remember this? OK, so we can go pretty fast. So obviously, the key problem in biology or in all of life science is that organisms need to produce proteins for a variety of different things that they need to do over their lifetime. So the goal is proteins, but the organisms have to make the same proteins again and again in a very set pattern type of way.

Some of these proteins are important. They're enzymes that catalyze reactions. These reactions would otherwise take 10,000 years to complete. Because you have the enzyme [INAUDIBLE], it can take 10,000 microseconds instead.

Structural support-- otherwise, a bag of water would just collapse. These are things that are going to hold the cell wall together, to give it some shape. And certainly, we have hormones to signal from one part of an organism to another part of the organism or for one organism to another organism.

So the key problem in life science over the past 50 years is how to encode the instructions for making these specific proteins because they have so many disparate shapes, disparate lengths, characteristics. How does the organism know or how do the cells know how to make these proteins? And the first step is obviously nucleotides. That's the most basic element of these blueprints that goes into making the proteins.

So have adenine, cytosine, guanine, and thymine, H, C, G, T. And as we all know, 50 years ago we've learned that these were actually arranged in a chain, actually two anti-parallel chains where the As base pair with a Ts and the Cs go along with the Gs. And these are not just arbitrary chains, but these are chains with polarity. So you have a start and an end to one chain. It's lined up with the start and the end to the other chain. And this is the double helix. This naturally forms a double helix if you melt this to the right temperature.

If you heat the DNA, it starts to denature. But if you cool it again, it'll start to come back together again. It'll re-nature. So it's a natural formation depending on the temperature of the bonds here, the temperature needed to break these bonds. And so you get an idea of what this double helix looks like here.

So now, let's take a step back. We have the DNA. We know that's going to be the end point, but we haven't had to even sequence the As, Gs, and Ts, and Cs to make genomic or genetic diagnoses. So even for the past more than 50 years, we've been able to make genetic diagnoses even before sequencing was invented.

How? Well, obviously, you can actually look at the chromosome structure. So you can get a blood sample from any human, isolate the white blood cells, which still has DNA, and stain them in a particular way, take a photograph, a photomicrograph of that, cut out the chromosomes, and line them up. And so people have been doing this for decades, more than half a century.

And basically, these chromosomes have been numbered based on the original estimates of the size of these chromosomes. So chromosome 1 is thought to be the biggest. Chromosome 22 is thought to be the smallest, but it turns out that 21 is actually smaller than 22. And there's a couple other missed pairings like this now that we have the exact length. But this is how we've been able to make genetic diagnoses.

Each chromosome, so each chromosome is a single double strand of DNA from end to end. And obviously, it's wound and coiled and rewound and recoiled in such a degree that this looks nothing like a double helix only because we're at a big picture here. We're at least five orders of magnitude up in terms of magnification from actually visualizing a double helix.

Now, like I said, we've been able to make diagnoses for more than 50 years using-- come on in-- in genomics. Anyone want to guess what this diagnosis is? Let's keep this interactive here.

AUDIENCE: [INAUDIBLE]

ATUL J. BUTTE: This is Down syndrome or trisomy 21. So you can see, obviously, it's highlight here with the little arrow. But there's three copies of chromosome 21 here. And that happens to be a viable phenotype.

You can born with this. And you can actually live for quite some time, though you have marked phenotypic problems or phenotypic issues. You can live with that. This is trisomy 21. There's two other chromosomes that you can have an entire trisomy of and still survive one way or another. Now, we have 22 pairs of chromosomes here.

AUDIENCE: Interruption.

ATUL J. BUTTE: Yes.

AUDIENCE: [INAUDIBLE] question. Explain the following fact. So this [INAUDIBLE]. How come with exception of testicular cancer Down syndrome people don't get cancer?

ATUL J. BUTTE: OK.

AUDIENCE: Look at the genes, figure it out. [INAUDIBLE] who don't get cancer.

ATUL J. BUTTE: So like I was saying, we have 22 pairs of chromosomes that are ideally paired, although we can get trisomies rarely. We also have a pair of sex chromosomes. Either you have two Xs or you have an X and a Y. And those are the normal conditions. There are abnormal conditions as well, but these are the sex chromosomes here.

Most of the Y pairs with a piece of the X, but the Y has some unique material. And the X certainly has a lot of unique material that doesn't pair at all. There's some very important things on the Y, for example, a piece of the androgen receptor and things like that. The coding for that is on the Y chromosome.

So this is what an actual photograph looks like. And this is the idealized version of this ideogram. The bands essentially are actually quite specific with a particular stain. When a particular stain is applied, when you apply different stains, you get different banding patterns. And it turns out these bands actually have a lot to do with the base pair composition of those regions. For example, GC-rich, or regions that have a lot of Gs and Cs, will turn out to be one color versus the As and the Ts.

So now, we've seen the big picture from the chromosome. We started with the DNA and the nucleotides. And we've already covered the double helices. So just as an introduction, I got only 10 slides of basic biology refresher, and then we'll get to some interesting stuff.

So how do we get from the DNA to the genes, right? Because even though we have 3.5 billion base pairs, only 1.5% of that codes for the proteins. We said the proteins were the final goal. And out of the 3.5 billion base pairs, only 1 and 1/2% actually are in a coding region per se. So let's talk about what we call a coding region exactly.

This line here, this red line here, represents the double helix. That represents the As, Cs, Ts, and Gs. And the structure of a gene is such that there's a piece here, which is going to serve as our blueprint for the protein called an open reading frame. It's got a start initiation. It's got an end or a termination. And this piece of that is going to be spliced in and spliced out as part of this blueprint. And we'll talk about the splicing in a second.

The piece of the double helix that's upstream or before this, towards the 5 prime end, it's called upstream. And the piece of the double helix that's after this termination codon, it's called downstream when it's on 3 prime end. Now, here's the most crucial point.

We all love diagramming these things like this, but there is no punctuation in the genome. There's no highlighter. There's no caps lowercase. There's no italics bold. All you have are the letters, right?

So one of the hardest problems is trying to figure out this structure from the letters because there's no flashing arrow that says, here is where the gene begins. All you have are the As, Ts, Gs, and Cs. And it's only informatics that actually deciphers this, OK?

So we have algorithms that people use commonly now that have been trained on manual experiments and the results of where this gene began and that gene began. They learn the patterns. They code it into algorithms, and the algorithms run on the entire genome to try to make predictions. But in the end, they're still only predictions of where these genes are, and they're still commonly wrong.

So each gene, a gene, encodes instructions to make a single protein. The DNA is called, before, is upstream. And a lot of the regulatory elements for this gene are actually in the upstream component. So let's say more than they are in the downstream component.

Regarding the splicing, imagine if an architect made a series of blueprints, but the architect said, well, if you don't want the Jacuzzi in your house, you can take this page out. If you only want a one car garage instead of a two car garage, you could take this page out. That's what the splicing is. You start with a whole set of what could be in the plans and pieces could be taken out or pieces could be left in. That's what the exons and introns can serve as.

The introns get spliced out. They can be thrown out. And exons remain to actually be part of the blueprint, which we'll talk about in the next few slides. Different tissues may splice in and out different components here. It's not always the case. Even though I'm drawing it here as exon, intron, exon, in a different part of the same organism that intron might be left in.

That's still a mystery as to exactly why that happens in some tissues, not in other tissues. And even the mechanism, the proteins that are involved in cutting and splicing these things, are reasonably well-defined. It's not clear why is it sometimes specific to one tissue versus another.

These are called alternative spliced products. If I have the same open reading frame from start to end-- that's how we define an open reading frame-- if I have the exon intron exon in one blueprint, but just the two exons in the other, those are alternative splice products. And because they're two different blueprints, you're going get two different houses or two different proteins at the end.

So we talked about the code at the start, a code at the end. And again, the biological system can figure out the promoter regions, where is the boundary between the intron and exon, where the start, where is the end using just the sequence syntax. There's no highlighter. There's no bold or uppercase, lowercase.

So we have really 3.5 billion base pairs, but only 35,000 genes. And the reason why-- this was an early estimate of 3%. It's actually only 1.5% of the genome, so only around 60 or 70 or 80 megabases or million bases out of the 3.5 billion base pairs. So what makes up the rest?

If I just showed you the structure of the gene-- here is the gene that I just showed you, let's say. What makes up the space between genes? Well, we're not going to call it junk here, but the 50% of the regions between genes are repeated elements. So in other words, there are at least four different types of repeated elements. And those four are repeated many times in the genome. That makes up 50% of the genome itself.

Now, these repeated elements-- so the repeated sequences are interesting. Because early on, when we had to actually-- when people were using DNA in a forensic type of way, the number of repeats can be different between individuals. So what do I mean by that?

So here is one type of repeated element. This is a LINE, or Long Interspersed Nuclear Element. It doesn't matter what it is, but here you see 1, 2, 3, 4, 5, maybe 6, 7, 8 copies of this line between this gene and this gene. So you might have 998 repeats between these two parts, but I might have 997 of them.

And if you do that between a whole bunch of different areas of the genome, you can actually measure it. You can actually find differences. You can do simple things, like figure out is a child a descendent of one or both parents. You can certainly tell, to some certain degree of confidence, whether a sample left at a crime scene is from this individual versus the random population. So just counting these repeats is actually pretty useful. Even though today we have much more accurate ways of figuring out whether a sample came from one person versus another, this is actually what was first used almost 20 years ago in the forensic view of DNA. So these are the repeated elements here. And the number of repeats can be different between individuals. Now, the repeat--

AUDIENCE: Another [INAUDIBLE] idea. Look at the number and placement of these repeats. And as you learn about [INAUDIBLE], you can measure comprehensively especially these genomes. We have thousands of [INAUDIBLE] in the public. And we actually determine what is it about the spacing or number of these repeats that does or does not influence the expression. Do they have a purpose given that they're [? matured ?] quite widely throughout the species?

ATUL J. BUTTE: So the repeats themselves are also interesting for one other thing. They repeat, so it makes you think that these are copies of an original. And in fact, when we look at the sequence of these repeated elements, we can tell what the original looked like and how deviant is this sequence from the original.

So how can we tell that? So what a repeat exactly is a repeat codes for machinery to make a copy of itself. That's all a repeat is. So let me be really clear. The repeated elements making up 50% of the genome codes for the machinery that goes back and makes a copy of itself.

That's why it's going to persist in the genome, right? If all you need is one or a couple of these things, eventually they're going to come back and make copies of themselves as the genome goes on through evolution. Now, it turns out all of these repeated elements are broken today in humans. None of them work today in humans. That's not true in other organisms. Repeats are still alive and well in the mouse, but not in humans. In fact, they're all dead in humans. Why?

AUDIENCE: You mean they don't work in--

ATUL J. BUTTE: They can't work. None of them are working. They cannot go back and make copies anymore. None of them code for that anymore.

AUDIENCE: OK. Generally, evolutionary what we see is the first of these classes are repeats that populate a certain branch in a tree.

ATUL J. BUTTE: Absolutely, I have a whole thing on this if we get to it, OK?

AUDIENCE: Oh, all right.

ATUL J. BUTTE: If we get to it. So just think about that, OK? So these different types of nuclear-- these different repeated sequences are all different ways to code for that machinery. That's all it is. We're the smartest organisms on this planet or so we like to think, so we must have the largest genome, right? Absolutely wrong. So we have 3.5 million bases.

So it turns out with 3.5 million bases and only four letters A, G, T, C, you can fit your entire genome on a CD-ROM. You don't even need a DVD for this. You can easily fit-- with no additional compression, you can fit your entire genome in 750 megabytes, which is what a typical CD-ROM can hold. So just think about that for a second.

So look at these other organisms and what their genome sizes are, the E. coli. So E. coli is a common bacteria that lives in your gut. It can be friendly. It can also be pathogenic-- has 4 million basis. Yeast, which is used to make bread-- 12 million bases.

But even the pea, the garden pea-- you got peas in your salad maybe for lunch-- has more of a genome than we do, 4.8 billion. Maize and wheat-- wheat has 17 billion base pairs. And we have 3 billion base pairs.

So the size of the genome has nothing to do with the intelligence of the organism at all it turns out. So where are these genomes different? A lot of the difference is in the space between the genes. So here are four organisms. Here's human, the same 50,000 base pairs I just showed you on the previous slide. And here are four genes within 50,000 base pairs and a whole bunch of repeats in the middle.

Here's yeast. And you can see how many genes there are and very few spaces. So it's much more compact. Here's maize. And wheat would be the same thing. And 50,000 base pairs, there's only one gene there.

And here's E. coli. In 50,000 bases, not only is it so compact, but there's actually genes on both sides of the genome, both strands. And they can even overlap it turns out. Then we talk about that more specifically here. This is a genome for plasmodium falciparum, which is the organism that causes--

AUDIENCE: Malaria?

ATUL J. BUTTE: --malaria, still the number one infectious disease killer in the world. This is chromosome 2. And you see the arrows here. The arrows represent which strand of the DNA the gene is on. They don't all have to be on the same strand, making the problem even harder. And they can overlap. There's many instances of a gene where another gene exactly overlaps. It makes a hard to measure one versus the other it turns out.

So this is the central dogma. We'll talk about central dogmas in a second, but this is basically the meat of how this happens. It serves as a blueprint. As we all know, the DNA-- when the cell decides to make this protein, somehow the DNA unravels in such a way to expose just the smallest piece of the 1.5% of the genome that's going to code for the protein it wants to make.

And so the DNA is kept locked up and nice and safe in the nucleus, the center of the cells. And the proteins have to be made in exterior of the cells. So a temporary copy is made of the DNA that goes out. And it's going to be worked on to make the protein. And the temporary copy is called RNA or, specifically, messenger RNA.

And so when you unravel, a messenger RNA is actually made in an opposite sequence, or this is actually built in the opposite complementarity of the DNA. So if there's an A there, there's a T here, et cetera. That gets exported out actively out of the nucleus and goes into the cytoplasm where it's made into a protein.

And to me, this is the most interesting part of the whole thing. Somehow nature realized that you can't just make proteins with the four letters. You have As, Ts, Cs, and Gs. And proteins are made up of 21 amino acids-- not 20, 21 amino acids.

The 21st is actually one that was just discovered about 10 years or so ago it turns out. And the coding for that one breaks all the rules. And we could talk about that if people are interested.

But there's 20 amino acids, let's say. And I have four letters. Obviously, if I made one amino acid for one letter, I couldn't do it. If I had two letters, I got four things in this one and four things in this one. That would code for 16 amino acids, but I have 20 to choose from. So I need three letters to code for at least 20 amino acids. So that's what's known as the genetic code. I'm blanking on the term.

So each position can be one of the four nucleotides. And nature evolved into using three nucleotides to code for a single amino acid. So that messenger RNA comes in here. And one of the most amazing pieces of nanotechnology, this ribosome, looks at the strand as it's coming in. And it's able to link that strand with other amino acids that are held in place by these transfer RNAs that recognize these triplet sequences.

So if there's a particular A, A, T, the tRNA that recognizes A, A, T, or rather the opposite of that, which is holding the right amino acid, comes into place. And the ribosome attaches it to build a growing protein chain. So in comes the RNA, out comes the protein. The micro machinery for this is astounding when you think about it.

Remember, this is happening many, many times per second in all of your cells. You don't have to think about it. It hardly ever fails. If it failed, we just couldn't do anything.

So here is this genetic code. So there's 64 different positions here. And so you see something like UUU. So Us are used in the RNA instead of Ts. It's just a slight difference there. UUU is code for phenylalanine. UCC codes for serine. And that's the code, essentially.

Now, obviously, these four different codes code for serine. It doesn't actually matter what the third base pair is. It's the first two are what are specific. So there's degeneracies in this code because we don't have to choose from 64. We only have 20 or 21 amino acids to choose from.

So this is actually the central dogma here. And these are all the different concepts we've talked about so far. We talked about nucleotides, which are held in a double helix. One single double helix makes up a chromosome. A chromosome holds the genes or DNA. And the entire set of all these genes is called the genome.

The genes code for messenger RNA. The transfer RNAs bring the amino acids, and the ribosome operates on that to make the proteins. We'll talk about signal sequence in a second. The amino acids are joined together to make proteins. And this sequence of arrows here is called a central dogma. Even though the arrows work bidirectionally now, this is the original central dogma.

Now, another fascinating part is this. Sometimes the cell needs to make something that would be quite detrimental to the cell. So for example, there are cells in your pancreas that need to make things to digest the food we eat. I just went to the truck. I had a little dumpling. Now, I need to break down that protein.

And it turns out I can break down protein with something called trypsin. Trypsin will help me break down proteins as well as many other things, but those things are proteins themselves. So here's the quandary here. My pancreatic cells, how are they going to make this trypsin without digesting itself? How do I make proteins that are going to kill me if I need to make them?

Well, so what happens is the first few base pairs of the sequence, in fact, the first few amino acids that get spit out by the ribosome, might code for what's called a signal peptide. And the ribosome might see this and say, hold on. The thing I'm about to make could kill me. I need to stop here.

And what happens is somehow the ribosome makes its way to a safer place to make the proteins called the Endoplasmic Reticulum, or the ER. So this is the ribosome. Here's the RNA coming in. Here's a growing protein chain.

And all of a sudden this red piece here gets recognized by the accessory part of the ribosome, the signal recognition peptide. It says, hold on. Don't keep going here. Because if I even have just one molecule of this, I'm going to start degrading the proteins within the cell.

So the whole thing goes over. It starts to spit the protein chain into this hole into the endoplasmic reticulum where that can actually then be built safely. This is in a separate compartment that can be exported. It's a pretty sophisticated piece of nanotechnology, this thing. It knows, because it's given the right codes, to not digest itself because it might need to make something quite toxic to the cell.

There are signal recognition peptides like this likely for a lot of different compartments in the cell. Don't build this into the ER. Build it somewhere else. The complete catalog of those is still unknown, but would be quite valuable to know.

So how did the cell decide to make this gene? This area of research is called transcriptional regulation. So why does this cell decide to make that messenger RNA? No one knows for sure in any one case how we actually get the DNA to unfold in just the right way for these things to actually be built, but we love to theorize. And we love to draw pictures like this.

And you'll even see pictures like this in*The New England Journal of Medicine* now. This is a commonly accepted way to draw a gene. So here's the start of a gene. The arrow here means that this is where the gene, a copy, is going to be taken off. And this is the upstream region. And these boxes represent things that can bind to that upstream region to turn on or turn off this whole process.

So the GR here represents the Glucocorticoid Receptor. So when this cell sees steroids, the steroids are going to bind to glucocorticoid receptor, and the glucocorticoid receptor is going to bind here. And this picture here shows that with the hormone, with steroids, this gets turned off. They put a little X here. And here, this is saying without, so with no steroids present, these two are actually bound in the natural state. But when you have that plus a glucocorticoid receptor, the thing gets turned on.

So you can imagine things, in fact other proteins-- the glucocorticoid receptor is another protein-- come back to the DNA and bind to the regions upstream of the genes to turn on and turn off other genes. So the whole thing is one amazing repetitive cascade. Genes code for proteins, some of which do their job, some of which come back to the DNA and turn on other genes or turn off other genes. And this cycle goes on endlessly. And we just know the smallest fraction of these cycles.

Remember, we're drawing nice pretty boxes like this here. There is no boldface, no highlighter, nothing upstream that tells us this is where it's going to bind. All we have are the letters. And a lot of slow motion biology goes into figuring out where these binding sites are. And we love to think that we have algorithms to do this. Some work and some don't. Huge area of research here if we could figure out this immense network of transcription.

What can start the process even though these are going on endlessly? Whole bunches, whole different actions, on cells can actually trigger specific responses. So for example, hormone action on receptors-- I just had lunch. My pancreas is now making insulin. That insulin is a small protein. It's getting exported out of my pancreatic cells into my bloodstream. That insulin is going to now go to the muscle and fat because my stomach is taking in all this sugar from what I just ate. I need to store that sugar somewhere.

So the insulin is going from the pancreas into the blood and triggering a response as I speak in my liver fat and muscle cells. So there might be a receptor up here, another protein. The insulin comes and binds and triggers a whole bunch of things to turn on and turn off different genes being expressed.

Shock or stress to the cell-- I just walked over here from Children's. It is frickin' cold out there. A whole bunch of things just started to be transcribed in my genome, new source of or lack of nutrients. If you have yeast and you put in one type of sugar, all of a sudden it starts to make things to deal with that type of sugar. If you change it to a different source of sugar, those genes go down. Another set of genes go up. So that could start a whole bunch of genes being transcribed.

Internal derangements of a cell, this is actually two different ways to make a transcriptional program. If I fly from here to San Francisco, I'm going to be hit with cosmic rays. And more than likely, one of them is going to hit one of my cells. And all of a sudden, if it hits something in the wrong place, I might start to make a gene that I shouldn't be making.

But even better, if my cell detects that it's making something it shouldn't be making, it might just decide to kill itself, save me the problem of having cancer. So those are two different ways internal derangements can actually trigger these things. Sometimes cells can decide to kill themselves for the benefit of the organism. And that's another common way to start these programs.

And of course, the list of things that can trigger a transcriptional program is endless. It's infinite, right? I can go like this, and I've just triggered some genes in my hand, for example. Anything could start these things going.

Any questions so far? I have about three or four slides left on biology. Is this stuff you all already knew? Are you learning anything at all so far? Perhaps, OK.

So this is a little bit more esoteric now, this idea of temporal programs. So one of the more important places where genes are getting turned on and turned off is during the development of an organism. So in our 10 months of gestation when we go from fertilized egg to actually leaving the womb, whole bunches of genes are getting turned on and turned off in a pretty well-defined program.

And the picture I'm showing here is the difference between development and-- I guess-- so the terms they use here are segmentation versus homeosis. Take a look at this picture, for example. These are two houses built in San Francisco in 1857. And they both started with the same blueprint. And they both started out-- they're essentially the same house right next to each other.

But after more than 100 years, the two houses look very different from each other. So this is given as an example on this cell paper as basically saying, the genes that are involved in the development of a particular tissue might not be the same ones that refine the tissue after it's been built. So the same tissue can have two different programs on two different time scales in terms of gene expression.

Now, let's talk about this process of making the messenger RNA because this is important. Most common, our most impressive way to measure genes now happens to do a lot with messenger RNA. mRNA can be transcribed at several hundred nucleotides per minute. That's not that fast, actually.

So if you have a gene that's 10,000 base pairs long, I'm only doing 100 a minute, maybe a couple 100 minute. It could take me hours to transcribe one gene. So in thinking about this process, this actually puts it into a real time scale that we can know and love here in terms of minutes.

So the gene dystrophin, dystrophin is involved in muscle. It turns out dystrophin has defects that can lead to muscular dystrophy. Dystrophin is involved with how muscles are formed and contract.

One dystrophin gene can take 20 hours to transcribe. So how are we able to make enough of this if it takes 20 hours to make, any ideas? If it takes 20 hours for me to make one transcript of dystrophin and I need to make a bunch, how am I going to do that?

AUDIENCE: [INAUDIBLE].

ATUL J. BUTTE: You have to have a whole bunch of ribosomes, but what?

AUDIENCE: Multiple copies of the gene.

ATUL J. BUTTE: OK, you can have multiple copies of the gene in the genome, but there's one other way that you haven't thought of. Look, I'm telling you it takes 20 hours at the start of the pipe to the end of the pipe, right? Now, I'm not telling you how fast I'm shoving things into the pipe. I can have one of the things making the [INAUDIBLE] right behind the next, behind the next, behind the next. So I could just stream off thousands of copies at a time, right?

I don't have to wait for one to be done to start the next. I can have them all go in one base pair off of each other, a couple of base pairs off from each other. That's how I can make more or fewer. But I still need all the ribosomes to make them into proteins. You're absolutely right.

So that's an important concept. Even though they are long, it doesn't tell you really how many we can make given any unit time. If the cell really wants to make a bunch of them, it could probably figure out a way to do it by just lining up the polymerases to make copies of the gene.

Now, it turns out a lot of messenger RNA happens to end with poly(A). So what do I mean by this? So as that little polymerase is working on the DNA streaming off a copy of the mRNA, when it's done it seems to stutter. So it adds a whole bunch of A, A, A, A, As at the end.

Lucky for us-- I'm sure there's some biological reason for this, but it happens to be very fortunate for us. Because if we have something that sticks to this wall that's just made of T, T, T, T, T, it's going to bind to the A, A, A, A, A. And all of a sudden, we can now have a big filter for all the messenger RNAs.

All we had to do is stream the messenger RNAs past my little thing stuck to the wall with the poly(T)s, and it's going to bind to the poly(A)s. That's how we can fish out all the messenger RNAs from the general pool of RNAs that might be out there because not all RNAs are actually coding for proteins. And in general, this is how we detect RNAs. You can take the sequence of RNAs and build the reverse complement of it.

So everywhere there's a G, put a C in my detection. And if there's an A, put a T. If there's a C, put a G. So we can use the reverse complement to help us tell whether this sequence is there or not depending on how we designed the probe. And we're going to talk about that.

So this is, I think, the last slide. So why are we trying to do this whole Genome Project? Why was it important to actually complete the Genome Project?

[? Eric ?] [? Weiner ?] basically uses this analogy that knowing all the genes is equivalent to knowing the periodic table of the elements for biology. So around 1850s, 1860s, 1870s, Mendeleev came up with the way of lining up all the elements in such a way to be able to predict characteristics of those elements. Now, back then, most of this was empty spaces because they hadn't seen some of these rare elements before. But they knew, by golly, there has to be an element right here that's inert, for example. Or there's got to be an element here that's magnetic.

They can make these predictions even though they didn't see what was actually there yet. They hadn't discovered it. The same way, the periodic table for biology for genomics isn't going to be a table. It might be a tree because that's how these genes are actually formed, through duplication events.

And so the idea is maybe we can predict that there must be something in the genome that's binding to this. So there must be something in the genome that's responsible for that. Now that we have the entire catalog, we can start to fish those things out. And what this whole course is is how this is relevant to medicine.

So most of the figures I've shown you here are from this book*Genomes.* In fact, there's a second edition now. It might be one of the required reading or something for the course.

AUDIENCE: [INAUDIBLE]

ATUL J. BUTTE: But it's freely available at NCBI. Definitely know this. So I would-- and in fact, all the figures are there, too, if you're interested in adding these things to reports and stuff. I highly recommend this more than anything else only because it's very readable, lots of great figures to help explain these processes, and it's very cutting edge, starts with proteomics, mass spec, microarrays, the whole works, much better than other books, relevant books in this field. And the Department of Medicine at this website has a primer on this. It's starting to get old now.

AUDIENCE: [INAUDIBLE]. You guys received another email OK?

ATUL J. BUTTE: This is starting to get a little old, but it's still quite relevant. Some of that stuff will never go out of date. OK, so I got another, let's say, 45 minutes or so?

AUDIENCE: Yup.

ATUL J. BUTTE: OK. Let's talk about gene measurement techniques. I will come up with something different for Tuesday. So what I'm going to talk about now-- so any questions so far? Yes.

AUDIENCE: [INAUDIBLE] time scale for the transcription?

ATUL J. BUTTE: Sure.

AUDIENCE: Does anybody know like said you can line up multiple transcription [INAUDIBLE] to combat these slower [INAUDIBLE], right?

ATUL J. BUTTE: Yup.

AUDIENCE: And I guess assuming that transcription complex binds to the start--

ATUL J. BUTTE: Exactly.

AUDIENCE: --and then faster than it moves along the strain, what's the lower limit? I guess the [INAUDIBLE] what's the fastest these things move?

ATUL J. BUTTE: So what's the fastest that gene can be transcribed?

AUDIENCE: Yeah. I mean because you've got the first transcription complex just move out of the way.

ATUL J. BUTTE: Absolutely. Yeah, exactly.

AUDIENCE: [INAUDIBLE].

ATUL J. BUTTE: You could probably do the math and figure this out, OK?

AUDIENCE: I think that's a great project, actually.

ATUL J. BUTTE: Right. So--

AUDIENCE: Helps with a-- no, seriously.

ATUL J. BUTTE: Absolutely. You can model this, right?

AUDIENCE: [INAUDIBLE] a sound basis understand everything we know about the initiation of the [? transcriptual ?] site and moving on and the letting go, all the things that have to happen. And there are quite detailed documentations of the [? characteristics ?] and processes. What is the upper bound for transcription and translation?

ATUL J. BUTTE: Well, basically, the model, the whole thing, is a pipe or a channel. And I already told you the length of this thing, how long it takes to make dystrophin. You just have to model how fast I can put the next one in. You might be able to model it based on just looking at the size of the polymerase and maybe how many base pairs it's covering up.

Maybe it's got to move out of the way before the next one can come in. You can model it just like that, for example. It would be quite interesting to see how that matches up with the actual published numbers for it.

My bet is that's probably still not the rate limiting step in this process. That's another interesting question. What's the slow part here? I'm not actually sure. It probably depends on the genes, actually.

There are some genes that are only-- there are some important proteins that are only 10 amino acids long. Insulin, for example, is one with 50 amino acids. So some of those can be spit out pretty quick. And then you got some that are huge like dystrophin and titin, some of these others. Any other questions so far?

AUDIENCE: [INAUDIBLE] in terms of [INAUDIBLE], what in terms of bind [INAUDIBLE]?

ATUL J. BUTTE: So again, remember, there's no nice line there. It's got to be in the sequence. So there are specific sequences that are known to be intron-exon boundaries, but clearly they're not 100%. Because in other cells, that same sequence might result in another transcript. So there are intron-exon boundaries about the start of the intron and the end of the intron.

And it turns out one end of it complements the other side. So these things are actually taken out in a loop type of form. So the thing that cuts this thing out lines up the mRNA in such a way that's able to put the start and end together and just basically snip the whole thing out like that. I can refer you to pictures in the genome exactly on this. But again, it's not 100%.

AUDIENCE: [INAUDIBLE] like GA, for instance, is one of the dinucleotides that sometimes initiates the splicing complex. But often it doesn't, meaning it's a very common dinucleotide. So [INAUDIBLE] the syntax that determines the splicing is not [INAUDIBLE]. So we have some heuristics which work some of the time. We have not reverse engineered [INAUDIBLE] the alternative splicing code.

ATUL J. BUTTE: You think about this. This is a massive decryption effort in the end. All we have are the As, Ts, Cs, and Gs. One level of coding is the triplet codon for the amino acids, but another more interesting one is the grammar here. Where is the start of the gene? Where is the end of the gene? When do I splice it? And when do I splice that? What are all the transcriptional regulatory elements? That's the Holy Grail here. That is the Holy Grail.

AUDIENCE: [INAUDIBLE]?

ATUL J. BUTTE: We know those.

AUDIENCE: But it doesn't always work either.

ATUL J. BUTTE: Absolutely. OK. So I'll give you-- I'll tell you exactly why. The code for the start, there's only one code for start, AUG, but there's three codes to stop. It turns out one of the codes to stop, if there's another code 100 base pairs later, that makes the 21st amino acid. So in other words, you can't even just look at the triplets. It's in the context of other things, too. It makes things messy again. It's so nice and neat with just triplet codons. But that stop plus something else 100 base pairs later gives, actually, the 21st amino acid, selenocysteine. That's life, messy.

AUDIENCE: So people discovered the selenocysteine.

ATUL J. BUTTE: People for the life of them couldn't figure out why. How are you making selenocysteine proteins? There's only four in humans, one of which is involved in the thyroid, which is why a whole bunch of endocrinologists look at this. It metabolizes thyroid hormone into one form to another, thyroxine to triiodothyronine.

That enzyme has a selenocysteine. And people couldn't figure out why it was coding for selenocysteine because it looked like a stop codon their. Well, you could start to figure these things out once you have the whole genome now. You can look at these exceptions.

AUDIENCE: [INAUDIBLE] the function of [INAUDIBLE] mice and humans. What [INAUDIBLE]?

ATUL J. BUTTE: OK, let's be very specific. The repeats code for a protein that comes back, looks for the sequence, and actually integrates the sequence somewhere else in the genome. It repeats code for protein like any other gene that finds its sequence and puts it in somewhere else in the genome. But for that to happen the code for the protein has to work. If it's damaged beyond repair, the thing can't work. And sometimes it makes mistakes.

So the thing's coming back. I made a protein. It's coming back. It's trying to make a copy of itself. Sometimes it grabs the wrong thing to make a copy of. So that's why we have copies of lots of things in the genome.

AUDIENCE: [INAUDIBLE].

ATUL J. BUTTE: Good question. It's able to recognize its own sequence, but it's not at all perfect. But what we can say about the difference between human and mouse is that the mouse has those sequences for that protein, a reverse transcriptase. It has codes for it that seem to be fully functional. And in a human, there are none that are still functional. That's what we're saying.

It's very curious why that is, this whole area of genomic archaeology that you can do with the genome now. In fact, what Zach was saying earlier is that we have so many millions of repeats that we can look to see how deviant they are from a working copy. And so we can age each of the repeats.

This one's 40 million years old. This one's 50 million years old. Because if it's got four places where it's different and this one only has two, the one with four is probably older than the one with two, right?

So like that, we can age all of these repeats. And it turns out a whole bunch of them were made at a certain point in time and then they dropped off. Again, no one knows why. I'll show you the graph in the second here.

AUDIENCE: So we can assume that the genome is growing all the time. So the [INAUDIBLE] much bigger than that.

ATUL J. BUTTE: Or smaller, too. There's ways to get rid of these things, too. If you do too much damage, the thing can't reproduce. And that's it. It's the end of the line. The most mind blowing thing that the Genome Project has done for me personally is that, now, I realize there's only one life form on this planet. It's DNA. Everything else is a side effect.

That DNA is doing everything it can to actually keep its code going. That's all there's is. There's a great book called--

AUDIENCE: *The Self Machine.*

ATUL J. BUTTE: *The Self Machine,* of course, but there's also a more recent one called-- it's on genomes. It's a paperback*The Autobiography of a Genome* I think it's called. It's like 23 chapters. And basically, they just pick one gene from each chapter and make a story.

AUDIENCE: Oh, yes.

ATUL J. BUTTE: It's a great book. I recommend reading that. It's a lay kind of reading.

AUDIENCE: Yeah, [INAUDIBLE].

ATUL J. BUTTE: It was on a bestseller list for a while.

AUDIENCE: That's right.

ATUL J. BUTTE: *Autobiography of a Genome* I think it's called, something like that.

AUDIENCE: [INAUDIBLE].

ATUL J. BUTTE: Some chapters were harder to write than others, I think, for this guy.

[INTERPOSING VOICES]

AUDIENCE: It's just a question about [INAUDIBLE] age, you know, and aging [INAUDIBLE]. Given that you can pretty much get rid of or add to or-- it seems like dating genes according to [INAUDIBLE] number of mutations against mutation frequency. And it seems kind of invalid now that we know that there are [INAUDIBLE].

AUDIENCE: Are you saying that because of these kinds of influences just looking at, let's say, the refresh rate of synonymous substitutions versus non-synonymous substitutions, it's not a good clock anymore?

AUDIENCE: I guess. I guess I'm doing-- what I'm driving at is it seems like one of the main-- one of the underlying points that you've been making is that there's a lot of change and a lot of mistakes that can be introduced from a variety of sources, none of which have to be all that magnificent or catastrophic. It's every day types of things. And yet, on the other hand, when we're talking about, well, this gene's probably over because it has four mutations, not two--

ATUL J. BUTTE: Right. It's all based on-- that's a great point. We can age these things because we think we know models. So given a fixed rate of nucleotide substitution where people have tried to calculate numbers for that, we can make a premise. But if that's not true, then it's not going to hold.

AUDIENCE: So first, if you have a transpose that on the [? box ?] right in the middle of the gene, then that would be invalid when you look at the rate of the mutation processing [INAUDIBLE]. It got a new puzzle. [INAUDIBLE] like a new to maybe kind of compare. The model [? is wrong. ?] But if you have a stretch, I mean--

ATUL J. BUTTE: Right.

AUDIENCE: --a stretch, it's more or less the same except for some [INAUDIBLE], then a lot of these-- not catastrophic, but more violent changes has to happen.

ATUL J. BUTTE: It's all a matter of time scales. These things are making copies of themselves over millions of years. It's not a day to day kind of thing at all. The cell in my lung can make a copy of itself and move it somewhere in the genome. On the grand scheme of things, who cares? It's not in my sperm cells. It's not going to go on.

So it's got to happen in the right cell at the right time. Across millions of years these things are happening when you think about it. It's got to be in the right cell, too.

AUDIENCE: When you look at the [INAUDIBLE] between genes, is there anything to say that the rate of the mutations would be about the same all regions of [INAUDIBLE]?

ATUL J. BUTTE: Yeah.

AUDIENCE: Actually, one of [INAUDIBLE] study. [INAUDIBLE].

ATUL J. BUTTE: Yeah, I think the rates are definitely different between different regions only because some of these processes have different rates. Remember the banding pattern I showed you in the AT-rich? That corresponds to what regions are rich in AT and rich in GC. Some processes occur at different rates between them because there are tighter bonds there.

AUDIENCE: [INAUDIBLE].

ATUL J. BUTTE: It might, but we don't have anywhere to model it.

AUDIENCE: That we don't know.

ATUL J. BUTTE: We don't have any way to model it.

AUDIENCE: [INAUDIBLE] when the genome was first presented in 2000 with [INAUDIBLE], people [INAUDIBLE] stopped seeing the genes in junk DNA. But if you look at the plate of sequences from [INAUDIBLE] of DNA from [INAUDIBLE] phylogenetic tree, the percentages are quite similar for the intergenic region. That's what the [INAUDIBLE] region tells us. [INAUDIBLE] speaking, this stuff is important whether it is to keep selfishly some [INAUDIBLE] to survive or because it's actually necessary to the organism's function.

ATUL J. BUTTE: Let me just show this one slide because you're asking so many questions about repeats. And this sort of summarizes a lot. So 3 million repeat copies started as working elements. That's how they can model this. There's 3 million repeats in the genome. So measure how far are they away from a working copy of them. Because through time, these mutations just happen randomly presumably at a fixed rate, but not necessarily so.

Most of the repeats predate the mammalian radiations, so before mammals were even formed. Most of these repeats are actually older than that in the genomes. But there's no evidence for transposon activity in the last 50 million years since we've diverged from monkeys. None of these things have been alive. But they're still alive and well in the mouse and the rat. So not-- go ahead.

AUDIENCE: Is that it? I mean, what's sort of the breakdown of how many species are still-- I mean are monkeys still doing this or they're [INAUDIBLE]?

ATUL J. BUTTE: I don't think we know yet because the monkey genomes are just being finished now. So we can better answer this question every month. By the end of this course, you'll be able to answer it better than today because these genomes are coming online. Literally, this is just only two years old information.

AUDIENCE: [INAUDIBLE].

ATUL J. BUTTE: What's that?

AUDIENCE: Severely [INAUDIBLE].

ATUL J. BUTTE: You're never going to be done sequencing every organism on the planet, but I think we will have a better estimate now than when this paper came out if you choose to do this. Here's the graph that I like to think of. So the different bars here, the different colors represent the different repeat elements. The lines, the signs, they're just different codes.

And here are less than 1% substitution from a working copy. And here's 34% substitution from working copy. And you can read that as millions of years basically, going back hundreds of millions of years. And you can see at 7% substitution there's a peak. And then it drops and then a slight peak, and then it drops. So these aren't fixed at all.

It makes you wonder what happened on the planet at this point and in evolution. Why were all of these things alive, for example, then all of a sudden they died off down to zero? So it's amazing when you think about that only because they model these as working copies. And how do they diverge from those working copies? It's fascinating. Genomic archaeology, this was not a field three years ago. The relevance in medicine still has to be determined, though, which is why we're going to move on.

OK, so I've got about 20 minutes left. Let's see how much of gene measurement techniques I can cover here.

AUDIENCE: [INAUDIBLE].

ATUL J. BUTTE: I'm going till 2:00, right? OK. Let's say 30 minutes. So the gene measurement techniques I'm going to talk about today are-- well, one's how do we measure DNA? How do we measure RNA? And how do we measure proteins?

If you can get this, you're going to know 95% of what you need to in this area. Let's start with DNA. So DNA sequencing was invented back in the mid-'70s. And when we measure DNA, we're sequencing it. So we want to know the sequence of As, Gs, Ts, and Cs. And so how did we get all 3 and 1/2 billion base pairs-- basically, just slightly more automated ways than what I'm going to show you here.

So what you can do is, if you have a DNA sample-- remember, it's As, Gs, Ts, and Cs-- we have tools to make copies of them. In fact, we can borrow the enzymes that the cell uses. And in a test tube, I can take my As, Gs, Ts, and Cs and make copies of them.

The process of making copies is a technique called Polymerase Chain Reaction, or PCR. All PCR does is make copies of things and amplify them. If you have one, you make it into two. Then those two become four. Four become eight, et cetera. And it grows to be pretty big, something that you can actually visualize.

So now, basically the way this technique works is this. As I'm making a copy of one strand of DNA, let's stop the reaction. So as I'm making a copy-- let's say I have 100 base pairs. Let's stop the reaction. And I stopped it after I have a whole bunch of copies going.

So some of the copies are going to be 50 when I stopped it. Some are going to be 99. Some are going to be 98. Some are going to be 97. So I'm going to have a whole variety of lengths of copies of that gene of that sequence. It might not even be a gene.

So then let's start it up again, but the last base pair I'm going to add to that copy I'm going to make a fluorescent. And we'll make it fluorescent in such a way that I'm going to make four colors, one color for A, one for T, one for C, one for G. So I have four different fluorescent colors that I could incorporate into the nucleotide, the last nucleotide I'm going to put on this particular sequence.

So now, I have a whole variety of lengths, and the last base pair might be a fluorescent color. Then what I can do is I can put that whole mix of length of DNAs and put them in a gel. I want to actually activate an electric current. The DNA starts to fall because DNA gets pulled based on its charge.

But as you would guess, the smaller sequences get pulled faster because it's easier to move than the larger ones. Even by one base pair you can separate them this way. It's called gel electrophoresis. That's why a one base pair difference we can see on a gel depending on the charge and all of this kind of thing.

So then what you do is you put a laser down at the bottom of the gel. So it used to be we'd run the gel out. And if it went off past the end, you'd swear because you just ruined your sequence. You lost it. It's in the puddle at the bottom of the gel.

But now, we want it to happen this way because we set up a laser at the end that excites whatever fluorescent is coming down there. And depending on the color, we can just read it off. Oh, here's a red peak. Here's a blue peak. Here's a green peak.

So we have a whole bunch of these peaks, and then software looks at this. In fact, the most commonly used software for this is something called phrap and phred, which is an open source, freely available program. People don't even use the software that came with the thing.

They wrote their own, and it's open source. It looks at the peaks, and then makes the call. I see this peak here. I think it's a G. I see this peak here. I think it's a C. There's going to be immediate problems with this. First, despite how fancy this looks, we still can't do more than 300 base pairs with this.

AUDIENCE: [INAUDIBLE] some flaws related to the size of the [INAUDIBLE].

ATUL J. BUTTE: They all, in theory, should be just one base pair apart, but there's all sorts of reasons. So why are they some jumbled here and some they're spaced out here? As you let the thing go-- it's not a linear thing, the spacing. So as you let it go, it gets wider and wider and wider in general.

But sometimes the copying process itself can have errors. Let me be clear. If the sequence I'm starting with has a whole bunch of Gs in a row, sometimes it gets confused. And it just happens to make one the wrong length. So if I have a lot of repeats, in the sequence I'm trying to sequence, I'm going to have problems. And I just told you 50% of the genome is repeats. It's not easy.

AUDIENCE: Different sequences might actually have different proficiencies in these reactions.

ATUL J. BUTTE: Yeah.

AUDIENCE: And so that might cause some slight shifts.

ATUL J. BUTTE: But basically, we can't do more than 300 to 400 base pairs than this, beyond this. And this is state of the art. This is the state of the art. So to go from this to 3 and 1/2 billion base pairs, you do it 300 at a time. That's it. There's no more magic than this.

There's also another problem. What if I have two peaks in the same spot? Remember? I have two chromosomes. On one chromosome, it might be one letter. On the other, it might be another letter. Happens all the time, 1 in 1,000 it turns out, which we'll talk about. That's a polymorphism. It could be this or that. So that's how you can read it by looking at the peaks as well.

That's basically it. You terminate the chain with the fluorescent nucleotide. And you just line them up, and you just read them off at the bottom. And phred basically makes the calls here. The software looks at the peaks and makes it the best guess here.

So all of these peaks, not just the letters-- I told you the letters would fit on a CD-ROM. If you wanted to keep the tracings, you're going to need a hell of a lot of storage, on the order of hundreds of terabytes for just one human. But they're all stored, and they're all saved for the Human Genome Project.

Because people go back and say, well, this peak wasn't so high. I don't think this is the right letter. But for the Human Genome Project, not only do we have all 3 and 1/2 billion base pairs, we also have an assessment of how good of a call it was for each letter based on these peaks.

So how do we get to the entire Human Genome Project? You automate the process. So this is a picture of the Genome Center, the Whitehead Institute. And basically, it's a whole bunch of robots just going off setting these things going. Most of the genome was sequence in just the 12 months prior to the finish of the genome.

Even though the process had been going on for 10 years, it takes time to develop these machines, develop the technology, and more importantly to develop the strategy, which we'll talk about in a second. And then the machines were built. And basically, they could do the whole genome in 12 months at the end. You could see the number of genes that were in draft form grew exponentially until the end even though it started back in the mid-'90s.

The Whitehead can run 100,000 sequencing reactions every 12 hours. Multiply that by 300 base pairs for each, and you can get on the order of 3 million base pairs. Let's say if I do 300, that's going to be three-- 30 million base pairs, let's say, every 12 hours now.

Robots pick all the colonies because these things are grown up in bacteria. OK, here we go, 60 million nucleotides per day is the current estimate of what they can sequence. And we couldn't do it even five years ago.

So then you have all of these little, little, little pieces of 300 base pairs. And somehow, you need to put this entire jigsaw puzzle back together. So there's essentially two ways you can do this. You can start with some strategy and say, well, I'm just going to take this piece of this chromosome and sequence that. Then let's move to the next piece and sequence that and take that and take that and just have some structure in mind as to the order you're going to do it and keep tract of the pieces. And then how they're going to go back.

The opposite approach is called shotgun sequencing. And that's basically saying, take the entire genome, split it up into these 300 base pair fragments, and split it up again. Take the whole genome and split up again into 300 base pair fragments using some other way.

So the place where I'm going to cut it for this method are going to be different than the ways I cut it for this method. Sequence the whole mess and just see what overlaps. Basically, this piece of this puzzle seems to go on this piece of that kind of puzzle. And you basically put the whole pieces together like that.

Now, that assumes a lot of computational power. You got to keep all these things in memory. You got to line them up and see which ones work best. So the best approach in the end was a hybrid.

You don't just put all the pieces together, but you also keep some idea of the scaffolding, where you got the pieces from, instead of just doing it randomly like this. So like most things in life, the hybrid worked the best. So here are all the pieces cut one way. Here are all the pieces cut another way.

And this is basically saying, well, A1 to A2 has no overlap, so we're stuck there. We don't know how they join. But A2 and A3 might have some overlap. And there's B here, and then you basically put them together based on the pieces. And hopefully, you have some overlaps between them.

This A goes with this A, but we also see B and B here. Here's B1 fits right here. And B3 fits like there, et cetera, et cetera. That's how these things are joined.

AUDIENCE: [INAUDIBLE] you look for overlap [INAUDIBLE]?

ATUL J. BUTTE: Obviously, you're going to need a whole bunch. You're going to need a whole bunch. And so there's a lot of holes.

AUDIENCE: So there's only 300 base pairs [? information. ?]

ATUL J. BUTTE: And again--

AUDIENCE: [INAUDIBLE].

ATUL J. BUTTE: Again, I told you 50% of the genome is repeats. So statistically speaking, you're going to get a whole bunch that look like they fit together. That's why this doesn't work that easily right out of the box because you have three small pieces that look so much like each other. So that's why hybrid approach works best if you remember that this piece came from chromosome 1 on this band on this thing, actually you then can put it somewhere physically in the genome.

So for a human they had 10 4-processor machines, so 4 gigabytes of RAM each, and a 16-processor with 64 gigabytes. That took 10,000 CPU hours. That's a huge amount of computational power just to put all the pieces together for the human. The overlaps take memory, 600 gigabytes of RAM.

And in the end, hybrid of putting it on a physical map is the way to go. Let's see if I have that there. So now, we're at a state where you can actually go to websites, like this one at the University of California Santa Cruz, which is probably one of the more popular ones, and you can just start to browse the genome.

So if you just pick-- you can even hit a button to take you to a random spot on the genome. Here's an arbitrary region of chromosome 19. It shows you where the bands are, where the GC-rich regions are, what you would see on a karyotype, where the predicted genes are, where actually people have seen different pieces of gene in messenger RNA.

So you can start to tell where the introns and exons are, where the polymorphisms are, all in one screen. And there's actually two different commonly used browsers for this. And I think the next screen is just a blow up of the UCSC browser. This is like a running thing of the GC content here. So that gets us to this.

So now, I already started to hint at this. Let's say, off the automated sequencer, you see peaks like this. Or here, I see a little bit of a hump here and a little bit of a hump here or here, here exactly, two humps that are exactly the same size. That makes me think-- I mean, the program puts a little n here because I don't know which it is, an A or a T. But what that really could mean is that that could be a polymorphism, that that human actually has both letters, one on one chromosome, one on the other chromosome.

And so if you have a chunk of DNA that you're interested in finding polymorphisms, you can get that gene or get that chunk of DNA, sequence in 100 humans, and just see what the different letters are in that sequence. That's one way to do it. But they found a whole bunch of these polymorphisms just by doing the Genome Project.

The Genome Project was sequencing 30 or so humans. And there are going to be differences in 30 people. So a whole bunch of them were found that way as well.

So you can either do this while sequencing the genome. You find arbitrary polymorphisms. Or if I'm interested in this gene because I think it's involved in asthma, I want to see if there are polymorphisms, I can go after that gene and sequence it the same way. Both ways are going to look like this on an automated sequencer, two peaks there.

You can also use microwaves, which we'll skip this one. But you can use an array to help you find these SNPs as well. So I think you're going to have electron SNPs. If someone--

AUDIENCE: Joel Hirschhorn, who is an expert for Whitehead, told you about SNPs and problems of interpreting SNP [INAUDIBLE].

ATUL J. BUTTE: So I think this is my last slide on DNA measurements. And I'm just going to end on clinical uses since this is genomic medicine. So again, SNPs are Single Nucleotide Polymorphisms. At a particular arbitrary spot in that genome, you have an A. I have a T, let's say. And it's usually just one thing or another. Rarely, there's places where there's three combinations. Usually, it's one or the other.

That happens, on average, rough estimate-- about 1 in 1,000 base pairs can be a polymorphism. So if you do the math, that's maybe 3 million, 10 million different spots in the genome where you could have one letter or another. And figuring out the association of those letters to differences in susceptibility to disease is another Holy Grail now.

If you have diabetes, I don't have diabetes. You have these letters. I have these letters. Maybe those letters have something to do with it is the basic idea.

Now, here's one particular figure from an article in *The New England Journal of Medicine* just to give you an idea. There's one particular gene called HLA. And this is specifically HLA-B. It's a particular HLA gene. And you can have one particular polymorphism on both your chromosomes. You can have a mix of one or the other, or you can have the most common form of this. So there's a common form of the letter. This is a rare form, or we could have a mix of the two.

And this one particular gene, if you're infected with HIV, that one spelling determines your conversion to AIDS. It can actually have a huge impact. So if you have this rare form of the polymorphism on both your chromosomes, homozygous, you can go from having HIV to AIDS very quickly within 10 years. If you don't have it, if both your chromosomes have the more common form, you can go on and on. And some people never get AIDS as far out as 18 years. And if you have the mix, you're somewhere in between here.

So associations of SNPs with diseases are exceedingly common to find now. In fact, every other week in *The New England Journal of Medicine,* this is what you're reading. It's going to be some association with this spelling versus that disease.

AUDIENCE: What's the most frequent form in this case?

ATUL J. BUTTE: The most frequent is going to be no HLA, in the black one. Yup. So let me ask you. How much do you think it costs to measure a SNP?

So if I have one arbitrary spot in all of our genomes that I want to measure, whether there's A, T, C, or G in any arbitrary genome, how much do you think it costs? Not how much the lab charges, how much do you think it costs to measure one base pair?

AUDIENCE: [INAUDIBLE].

ATUL J. BUTTE: How much does it cost?

AUDIENCE: I already [INAUDIBLE].

ATUL J. BUTTE: Guesses?

AUDIENCE: [INAUDIBLE].

ATUL J. BUTTE: You already guess? Yeah. So we come up with the answer of $0.05 to $0.10. It's probably going to $0.01 in the next few years, pretty cheap.

AUDIENCE: Stole your thunder, sorry.

ATUL J. BUTTE: Stole my thunder, all right. Oh, yeah. Last thing on SNPs, you didn't take this, did you?

AUDIENCE: I did, too.

ATUL J. BUTTE: Damn, I'll stop giving you my slides. How much time we have? OK. Let's talk about gene expression. Let's talk about microarrays, fourth slide, and then we'll talk about protein.

So we recovered DNA. Now, we're working our way out of the nucleus. Now, let's talk about the messenger RNA.

So like we said, if the cell decides to make a particular protein, it's going to start making the blueprints for that protein, the messenger RNA. It turns out it's very easy to measure these now. The most common way to measure them is using a microarray. All a microarray is is a 2 centimeter by 2 centimeter, about the size of your thumbnail, it's a man-made grid of DNA where each spot in the grid is looking for one particular RNA.

And so the density is about 500 by 500. So that's on order of a quarter million different spots on these arrays. And suffice it to say, you can measure the entire genome of transcripts with one array today.

And the way this works is have to get the tissue that you're interested in. You can't just do this from blood unless you're interested in blood. If you're interested in the brain, you've got to get the brain tissue to figure out what genes are turned on there. It might expose it to certain influence. Get the RNA, so isolate it. Remember, most of them are sticky with the little A. We just have to get the Ts and pull them out.

Make them fluorescent by making copies of them. Remember, I showed you that technique. You can make a copy of something using the PCR and just have it put fluorescent ones in there, and it'll become fluorescent. Let it sit on this man-made grid overnight. Come in the morning. Wash it off. Light it up with the laser.

The laser lights up the fluorescence, and you basically take a picture. And you get a TIFF image. If you scan in pictures at home, you know what a TIFF image is. But these are huge. These are on order of 10 megabytes, these files.

And then the biology takes a pause. The informatics comes in and starts to quantitate each of these. It's quantitative to 4 to 5 orders of magnitude, and it's not perfect. There's a lot of noise. Most people like this technique. Again, these are also commodity items today. You can get one of these for the entire genome for about $300, $350 bucks today on [? Longwood. ?]

The quantitative, there's at least two different ways to do this. There's an absolute measurement technique and a relative technique. So what does that mean? How much of this gene is present is one technique. How much is this gene is present relative to this other tissue is another technique.

And the relative one uses two fluorescent colors. The absolute one uses one fluorescent color. It depends on what's the nature of the spots are, but we're not going to go into those details unless you really want to.

The important point is that the genes are chosen arbitrarily. Statisticians hate this aspect of it. These are not drawn from some master normal distribution here. This year's chips are different than last year's chips because we know more of the genes.

Next year's chips are going to be different than this year's chips. So every year, these things are changing. There's not ever going to be one fixed genome here. And you need the functional tissue.

There are some diseases that are easy to study with microarrays, like cancer. If you have a solid tumor, the patient goes to the operating room. The tumor gets taken out. You have a hunk of tissue. You can extract the RNA. You can find a lot of genes this way. And so that's why most of the microarray papers involve cancer.

A slightly harder disease to study might be something like diabetes. Diabetes affects the liver, muscle, fat, pancreas, maybe brain simultaneously. You might want to get all of those tissues. And let me tell you, we don't normally do biopsies of these things when we diagnose diabetes. Well, we could use animal models, et cetera.

Slightly harder disease to study would be something like schizophrenia. I know the functional tissue, but just which part of the brain am I going to actually get and put on this array? I can't take an entire human brain to do this. Is it this part or this part? So I even know the tissue, but I don't know where I'm going to put the array, what I'm going to put on the array.

And then one harder disease to study would be something like hypertension. What's the functional tissue in high blood pressure? Is it the heart? Is it the blood vessel? Is it the smooth muscle of the blood vessel? What exactly am I going to put on the array to learn what genes are up and down? So this is not the panacea for all diseases. If you want to know what genes are functionally involved, you have to be able to think about the functional tissue here. Yes.

AUDIENCE: How long after death can you do this [INAUDIBLE]?

ATUL J. BUTTE: The easy way to answer that is that people have been able to get RNA from paraffin-embedded tissues. And they think that they're successful, but most people flash freeze these things. If you're in the OR with the tumor getting taken out, most people just put it into [? monosodium ?] right then and there.

AUDIENCE: The slightly longer answer is RNA does degrade as opposed to DNA, which can stick around. And that's why they can do *Jurassic Park-like* stunts.

ATUL J. BUTTE: But there's a reason why RNA degrades, because we degrade it. A lot of viruses are just strands of RNA. So that's on our interest to degrade it. So we have RNA, so we exude RNAses.

It's much harder to work-- I mean, high school students work with DNA. If you've ever done this in your high school where they go get blood from an abattoir-- and they just basically get the blood. And you can look at the strands of DNA. You can't do that with RNA because we all have RNAses. We're going to degrade it immediately. So you have to wear gloves. You have to use special techniques, et cetera.

AUDIENCE: But sometimes studies [INAUDIBLE] you have hours after death. And definitely, a muscle tissue four hours after death has a different [INAUDIBLE] factor [INAUDIBLE], for instance, [INAUDIBLE] in many other [INAUDIBLE]. And so if you really want to do it reasonably, then basically as close to get the moment of death is where you want to be.

ATUL J. BUTTE: This is what these arrays actually look like. This is the whole human genome array from a company called Affymetrix that makes these things. And they actually make a 96-well format of this. So here are 96 microarrays in just one plate. So you can imagine how much data you get out of this with 10 megabytes of each for each of these images, huge amount of data being collected this way. And we're not going to talk about the differences here. And we'll not talk about that.

So you can use these microarrays to tell which genes are up. So for example, if I have a whole bunch of patients and half of them had one type of leukemia and half had another type of leukemia, leukemia means a whole bunch of white blood cells go up in the blood. So you get easy access to that tissue. And you can find genes that are up in one and down and the other or down in one and up in the other, et cetera, et cetera. In the end, you can get lists of genes very easily.

Now, we all commonly do this. And this core facility is basically in every hospital. If you're a clinician at the Brigham and Women's Hospital, you can take your sample down to the core facility, and they'll give you back the text file-- this gene, this amount, this gene, this amount. It's just a matter of asking the right questions now.

But the list of genes isn't enough. Sometimes you want to go back and validate that a gene is actually where it's supposed to be or what you thought was actually happening. And there's two ways to do this, in situ hybridization and real-time PCR. So these are usually inserted after you've done a microarray analysis.

In situ hybridization is very easy. I take the same RNA that I'm looking for, make the exact reverse complement of it, and make it fluorescent again. And I'll make enough of it that I can then stain a sample, a tissue sample, with that and then light it up to make a picture. So wherever that transcript was, this is going to stick to it after I've washed it off. And it's going to light up.

And I can take glorious pictures like this. So I can use different colors to light up one RNA versus another. And that might help me tell that it's in the right tissue that I thought it was.

Real-time PCR, it's even-- it's actually a little bit harder to conceptualize, but basically real-time PCR uses that polymerase chain reaction technique to actually detect one sequence as compared to a control sequence. So it gives me a more quantitative measurement. This gives me a nice picture that I can put on the cover of *Nature,* but RT PCR actually gives me a number that actually can then be used to validate the microarray findings.

It was three times higher. It was four times higher. It's basically using a PCR technique to see how fast do I get a detectable amount using the repetitive the exponential growth aspects of PCR.

So now, let's get into proteins because we're coming close to the end of the hour here. So again, the Holy Grail is get to proteins. We all essentially use microarrays as a proxy for proteins, but we don't have an array. The same kind of array that we have for microarrays, we just don't have it for proteins.

Why? Because proteins are much harder to measure. The RNA, basically, is As, Ts, Cs, Gs. And it's essentially a strand. And sometimes it buckles and has twists, but it's pretty easy to measure using these microarrays.

Proteins can be positively charged, negatively charged. They can love water. They can hate water. They can be big. They can be small. If you look at it this way, it looks like one thing. If you look at it this way, it looks like another thing.

So there's no easy consistent way to measure proteins like there is for RNA. Just to give you an idea, here is a commonly used technique called 2D Polyacrylamide Gel Electrophoresis, or 2D-PAGE. Take one sample, a gemish of proteins. All the proteins are radioactive, and I'm going to spread them out this way based on the size.

Remember, if I put a charge, the smaller things move down faster. So here's size. And here's pH. Here's the pH that they like to hang out in, two different things, two different aspects of proteins. And in the one gemish, I have all of these spots. And if I do this again, they're going to be slightly different. If I do a third time, it'll be slightly different.

This is a picture. This is the most irreducible thing you can imagine. The gel is set just right. It might just come out a different way. So we wish that we could look at this and say, oh, it's got this size and this pH. It's got to be this protein-- can't do that at all.

So then what we'll do is we'll cut out one of these spots and use the next technique to figure out what protein might be in there. But literally, people look at these pictures and they say, well, here's a picture from cancer. Here's not cancer.

And oh, I see this spot here, and I don't see it here. Maybe it's a protein of interest. And there's a whole informatics trying to figure out and automate and image process these things, but they're terrible. This is not reproducible.

Look how these spots are streaky, right? I mean, it's just a mess. Each spot has dozens, hundreds of proteins in there. Here's one paper I picked. These guys looked at all of these spots and sequenced all of them just to identify all these different proteins. So they've numbered each one. So back in '99, this is the-- [INAUDIBLE] actually in this building. I'm laughing at this, but, I mean, it's a huge amount of work to do this.

So from that one spot, we might want to figure out what proteins are there. And this is a technique that most people use. It's this technique called mass spectrometry. And so the way this works is you get this little spot. Still, there's a whole bunch of peptides in there. Even though we separated them a lot in these two axes, there's a bunch of proteins here.

And if you actually shoot it into this detector-- so basically, what mass spectrometry does is is it takes a spot and hits it with a laser. And the things come flying off of there. And the way that they fly off of there depends a lot on the size of the protein and the charge of the protein.

So basically, in the end through all of this, you basically have one axis here, mass divided by charge. And I have peaks, basically, that are coming off of the detector. Oh, there's a whole bunch of stuff at this. Now, there's nothing. There's a whole bunch. There's nothing, like that, mass versus charge.

So the way this works is what you can do is, these proteins here, you can digest them. Remember how these are long proteins. That same enzyme that I have in my pancreas that's still working on my lunch, I can use it to my advantage.

I could take a spot and digest these peptides into successfully-- to have sequentially shorter protein. So here's one protein that's 40 amino acids long. Here's one with 39. Here's 138. Here's 137. Here's 136.

And basically, I do this in such a way so that all of these peptides get broken up into a series of smaller peptides. Why am I doing all this? Why am I doing all this? Because this is the trick.

If I want a protein that's 30 amino acids long and one the same protein, but it's only 29 amino acids long, the difference in size is the last amino acid. Then to go from 29 to 28, the difference in size there is that amino acid. So if I've basically successfully cut this one amino acid at a time and have a whole range of these, then I've got the difference in sizes to help me tell what that last amino acid was. And that's how we use it. We just look at the difference in peaks, the difference between the peaks is what we measure here.

AUDIENCE: [INAUDIBLE]?

ATUL J. BUTTE: No, not at all. It's a mess, right?

AUDIENCE: [INAUDIBLE]

ATUL J. BUTTE: It's a ratio. So there's a bunch of peaks here, but that same peptide might be somewhere else, too, if it happened to get an extra charge. Absolutely. It's not perfect at all. It's not perfect at all, a whole new area here that's dying for new algorithms.

You do it your best shot. You look at the peaks here. There's naive ways to do it, but they're not perfect at all. And of course, there's some proteins that love to have variable amounts of charge more than others. So some things are easier to detect than others.

AUDIENCE: But I mean, the technique itself, there are known methods of how amino acids [INAUDIBLE] species themselves that they take [INAUDIBLE].

ATUL J. BUTTE: We know the size of each amino acid, each of the 20, 21 amino acids. So we can tell if there's a difference of peak here. But remember, there are multiple proteins.

AUDIENCE: [INAUDIBLE] is not infallible, but there's characteristic [INAUDIBLE].

ATUL J. BUTTE: Absolutely, that's why we're able to use it at all.

AUDIENCE: And there's libraries of that. There's a lot of--

ATUL J. BUTTE: Exactly. In fact, we can predict. We can look at every protein in the database and predict what it would look like when applied with trypsin to cut it. What would the peaks look like?

And you can actually make predictions. And you basically take your pattern and compare it to the computer pattern. This might be predicted. This is the actual. And the computer, the algorithm here basically says, this is the protein.

Now, there's another important point here I want to bring up. This is not quantitative. This helps you identify a protein in a sample, but it's not quantitative. It doesn't tell you how much there is.

There's all sorts of newer, fancier techniques to compare this to some other sample to try to get some kind of quantitative measurement. But the way this is commonly used, it's just identifying. It's not quantitative, [INAUDIBLE].

AUDIENCE: Just a point, in *The Science Times,* in *The New York Times* this week, there was an article I saw about a test, a diagnostic for cervical cancer [INAUDIBLE].

ATUL J. BUTTE: Absolutely.

AUDIENCE: They don't know how. It's just a tumor.

ATUL J. BUTTE: I think I have that picture.

AUDIENCE: OK.

ATUL J. BUTTE: Yeah. Here, here's the one. So here's ovarian cancer-- same thing, right? So these guys use a particular chip to do this, but the chip is the smallest piece of it. It's basically still a mass spec that's on the end.

Here is unaffected, unaffected, cancer one, cancer two. Here are all the different bands. And they say, to hell with identifying them. Here's the pattern. Oh, wow. I see a band here, a band here, but no band here. This must be diagnostic.

I'll easily answer back. I mean, we're in danger of overfitting the data here. You have hundreds of thousands of peaks here and only four samples. How hard is it to find them that will answer your question?

I tend to believe these things when they have something to do with the biology. If we go ahead and identify and there's some causal mechanism, that's great. But others are going to be very happy with this, especially since we have no other test for ovarian cancer today.

That's fine-- or cervical cancer. For cervical cancer, we do Pap smears. For ovarian cancer, we have nothing. If there is something in the blood, it's better than nothing is what they would answer back.

So here's basically the peak differences here. So they say, here's this peak. Here's this peak. We think that's an arginine or-- that's basically how this works. You use the differences in peaks. These things are called peak lists.

Quantitative, we're not going to go there, but, basically, you can take one sample and another sample and grow the one sample in heavy water. Instead of H2O, we'll use H3O. And that extra H, it's not extra-- well, actually--

AUDIENCE: [INAUDIBLE]

ATUL J. BUTTE: It's--

AUDIENCE: Deuterium.

ATUL J. BUTTE: It's deuterium, right. So it's not just deuterium, but you can also use other different isotopes. But basically, you have the same atom. Like, carbon can have 12 protons, or you can have 12 plus a neutron. You can have all sorts of other things in the nucleus of the atom that could change how it looks on mass spec because you're changing the mass just slightly. And then you can tell which sample it came from and then try to quantitate that way.

But again, the charge could change. So one protein's not in one spot. It's in a whole bunch of slots. You need to sum them all up, and they're overlapping.

AUDIENCE: [INAUDIBLE] trypsin [INAUDIBLE] peptide one [INAUDIBLE] then--

ATUL J. BUTTE: OK, let me be clear. The trypsin cuts into manageable fragments. And there's another fragmenter in front of the machine that's actually cutting it up into these one amino acids. Go ahead.

AUDIENCE: You have [INAUDIBLE] different ways [INAUDIBLE].

ATUL J. BUTTE: Yes.

AUDIENCE: [INAUDIBLE] and you have maybe [INAUDIBLE] number of [? charges-- ?]

ATUL J. BUTTE: Absolutely. That's right.

AUDIENCE: [INAUDIBLE] pretty much [INAUDIBLE] on what exactly [INAUDIBLE].

ATUL J. BUTTE: Well, the thing is the only hard part is that you don't know where the protein is. You see a whole bunch of spots there. And you see that it's here, but it could be on either side of the spectrum you're looking at. And if you let the machine run long enough, you get a huge amount of data off of that machine, an unmanageable amount of peaks.

And we can look at and isolate one peak. Here, let me-- I think from this peak, we go to this peak. From that peak, we go to this peak. We've zoomed in so far to get the differences, but it could be elsewhere depending on the charge.

It's doable. People do get great results off of this, but by no means is the informatics done in this field yet. And as more people use this technique, it's screaming for new algorithms.

AUDIENCE: A quick statement-- is it obvious to everybody that the concentration of protein is not the same as concentration of RNA? Everybody should know that. In fact--

ATUL J. BUTTE: Right. Like I was saying, we're using it as a proxy. But let's be clear. If the cell has a whole bunch of this protein there and it's happy with that amount and it's not going anywhere, it doesn't need to make more. That's the simplest example of something where there might be zero RNA, but a whole lot of protein.

If that protein's degrading really fast, then the cell might need to make more. But there's a massive disconnect here. Some things are going to be correlating, a lot of things are not.

AUDIENCE: [INAUDIBLE].

ATUL J. BUTTE: All these things about proteins which we don't know-- depends on the protein.

AUDIENCE: I just have a question. I'm just not sure. When you do small molecules [INAUDIBLE] compare the [INAUDIBLE] you eliminate the charges, is there any equivalent way that we could prepare--

ATUL J. BUTTE: To eliminate the charges?

AUDIENCE: Yeah.

ATUL J. BUTTE: You asking a detail I don't know.

AUDIENCE: No, because that's the reason why [INAUDIBLE] small molecule [INAUDIBLE] proxy [INAUDIBLE].

ATUL J. BUTTE: Absolutely. But the thing is is that it's picking up charge, right?

AUDIENCE: Yeah, of course. It is a charge, ultimately. But the reason why [INAUDIBLE] use that as a proxy [INAUDIBLE].

ATUL J. BUTTE: Absolutely.

AUDIENCE: [INAUDIBLE] is not [INAUDIBLE].

ATUL J. BUTTE: Yeah. The biggest problem, though, is that these are longer than small molecules. I mean, these are much larger masses. That's the problem. And the proteins themselves, if you just chopped off the thing that's holding the charge or that loves-- if you just chopped off the polar part of it, it's going to have different properties. That next peak could be wildly different. It totally depends on that.

AUDIENCE: But for my microarray lecture is, why is it, if they, in fact, as [INAUDIBLE] explain it, protein function, protein synthesis is not in a 1 to 1 relationship with RNA synthesis [INAUDIBLE] so successful in defining disease classes and the [INAUDIBLE] just based on RNA expression? Why is it-- why should we [INAUDIBLE]? Or could it?

ATUL J. BUTTE: I think this is my-- this or the next one is the last slide here. Now, despite it being hard to measure these things, there are definitely companies that are trying to make protein chips here. One way to detect proteins is to have an antibody against a protein. That's how people have been doing things like Western blots for decades. And you can radiolabel them and stuff.

So you can make a plate here where each well has an antibody against one particular protein. Or you could make-- for example, people make a cytokine array. So this array just basically-- each spot looks for a different cytokine.

You can buy these things, but you can't do this comprehensively like can with microarrays. That's the point. If you look across all possible proteins plus alternative splice products in proteins, you just can't do it today. You can do functional assays as well. I mean, it's antibody array.

That's it. So we talked about sequencing, polymorphisms. We skipped SAGE. Most people are using microarrays now. We didn't even cover wafers. Let's cover wafers in a second-- 2D-PAGE, mass spec protein arrays. Let me at least cover the wafer thing because it's worth knowing that.

It's all slides. So the wafers blow my mind. And that's why I like to end with this. So those companies that make these microarrays, they don't just make them one at a time. They make them the same way computer chips are made. They're made 40 at a time in a wafer, and then they're just cut out.

So there's a company called Perlegen, which is owned by that other company Affymetrix, that basically said, why are we making 40 of the same array? Let's just make a wafer one big array. And with that, you have so many spots. You have 60 million spots in one wafer. It's about 5 inches.

I'm drawing it like this, but it's really this small. You can see the size here. This is a wafer. A 5 inch square has 60 million probes. So even two years ago, they showed how you can use these wafers to re-sequence an entire chromosome. And that was one paper.

You think about how many years it took to sequence of first chromosome, the Human Genome Project. They can just do that now over a weekend because they basically have each spot is a moving window of 25 base pairs. So what do I mean by that?

They took the smallest chromosome. I think it was, like, 22, 23. They took the smallest chromosome, and they took the first 25 nucleotides. And they put it in one spot. And then the middle one they said, well, it could be an A, T, C, or G. So that's four spots. Now, let's move to the next one, to the next one.

So they have a moving window of 25 nucleotides going from one end to the other end of the chromosome across a series of wafers. And once they have that done, they can just take anyone's blood and just basically get the entire pattern now. And they don't have to say, well, I think there's a SNP here or a SNP there. They just know all the SNPs now because it's just a matter of how many humans you put on the array. And it's just blood.

The big issue-- each scan, the TIFF image, takes 10 terabytes. We were just learning about gigabytes. Terabyte is 1,000 of those, and this is 10 of them for just one of these wafers. So it's well-known that life science data is growing much faster than this famous Moore's law. Moore's law is from Gordon Moore from Intel that said that microprocessor power doubles every 18 months. Life science data is growing way faster than that.

This is just one example of this. They've publicly stated that they can sequence all SNPs in a human in 10 days, not just the ones we know about. They just know all of them. And this is just one of many companies. There's many in the 95 Beltway here that are doing the same thing as competition.

So remember how I was joking around that you could fit your entire genome on a CD-ROM? It is absolutely conceivable that we can have this within the next one or two years if we want it. The technology is there to do this. They certainly have the hard drive space there at Perlegen to do it. All right, I think that's it. And we should end since we're a little bit over.