ATUL BUTTE: So last week I talked about-- what did I talk about? I talked about an introduction to molecular biology for about 10 or 12 slides. And then we were going pretty quickly I covered a lot of this material on gene measurement techniques. And so we've already had one question, or one request, to talk about something that I actually flipped pretty quickly through, this concept of SAGE and how is SAGE used to measure microarrays. And actually it's a measure of RNA, and it's quite important that we actually talk about SAGE, I think.

But I'm going to start the discussion with SAGE talking about microarrays again for a second. So I we talked about microarrays, how these are made grids of DNA. Each spot in the grid is looking for one particular RNA. And in fact, more than one spot might be actually hybridized into one RNA. It doesn't have to be one to one It could be many to one.

And I just want to talk about the differences of the two technologies for microarrays and then talk about how that actually leads itself to the SAGE discussion. So for the past coming up on 10 years now there have been two technologies to make microarrays. There's the oligonucleotide approach and the cDNA approach. The analogy I like to use is this.

If you think of a microarray as, say, Manhattan with all the skyscrapers. If you had to rebuild Manhattan, there's two ways you could do it. You could take one skyscraper at a time and put them into place, or you could just build all the first floors, all the second floors, all the third floors, all the fourth floors. And that's essentially the differences between these two technologies.

So the oligonucleotide arrays are really made by Affymetrix. They're made-- each of the skyscrapers-- the equivalent, each of the strands of DNA, are only 25 nucleotides long, but they're built one base pair at a time. Because they're only 25 nucleotides long and genes are much longer than that, that's why they have multiple probes going against a particular gene.

The cDNA, on the other hand, are essentially strands of DNA that will hybridize to the RNA, but the strands of DNA that are just put into place using primarily using something like a robotic spotter. So it might pick up four at a time, or 16 at a time, or just one at a time, but that's how these two are made.

And beyond the technology difference here there's a major difference in what exactly these microarrays are measuring. The cDNA arrays are measuring relative expression amounts. So you have to put two samples on there, one of which the RNA might be colored red, another of which the RNA might be colored green. And you'll look at-- what people will do is after they've hybridized it, they've taken a TIFF image of this, they'll look at the relative ratio of how much red is there compared to how much green. So in that you always have a relative expression level.

In theory the oligonucleotide approach is different. Now this is relative, because each of the strands of cDNA has its own properties. Each of the strands might have a different set of genes and As and Ts and Cs, different melting temperatures. And that's why you can't really use that as an absolute measurement. And intensity might be lower on one of these arrays, not because there's less of a gene present, just only because the actual cDNA had a little bit different biochemical properties and wasn't hybridizing efficiently.

The oligonucleotide arrays on the other hand, are, in theory, designed so that each of the probes has similar characteristics in terms of melting temperature, and some of these other characteristics. So in theory you can compare one spot to another spot. And because of that, these are made-- these are hybridized using a single sample. So you put one sample on one array here, and if you want to do a relative difference you use two arrays.

Now again in theory this is an absolute expression amount. If people want to get to the most absolute measurement, like how many strands of RNA are present, then there's another technology called SAGE, and that's how I just wanted to introduce that discussion here.

SAGE stands for serial analysis of gene expression. I think the original paper on this was in *Science.* It was a one or two-page paper back in 1995. by-- it was Victor Velculescu at Hopkins. And so this actually was developed around the same time as microarrays but actually gained a lot of popularity before microarrays became so commonly available.

But essentially the way SAGE works is that it gives you an absolute expression level for a particular gene, but it measures gene expression by sequencing. And automated sequencers were just coming around at the time, and people realized that you can use this technique to use them to measure RNA expression levels.

So the way this works is this, is that in theory every RNA-- or let's say every sequence, every gene-- has what they call, is a SAGE tag here. And all the SAGE tag is is, let's say, 10 nucleotides that hopefully uniquely identify that gene. But the catch is it has to be after a CATG, and we'll see why in a moment.

So you look at a gene, you look for a CATG, and the most downstream 10 nucleotides are going to be the tag for that gene. Hopefully that's enough to make it unique, but that's not always going to be the case.

So the way this works is this. If you start with the strands of RNA that you're going to measure, the first thing you want to do is to make a strand of DNA. RNA is very fragile. You want to convert it to DNA as quickly as you can. Otherwise it's going to start to degrade.

So the techniques for taking a strand of RNA and making a strand of DNA are commonly available. It's a reverse transcriptase. You can buy this off the shelf, or you can buy it from a store. And essentially it's going to start by going at one end. You might give it a primer to get started, and then it's just going to go all the way until it falls off the end. So now you've made a cDNA copy of this.

And what you do now is, you can do that for the other strand as well. So now you've gotten rid of the RNA, and you have two, you have a double strand of DNA representing the RNA that you had in your original sample. Now it gets a little tricky, because now we're essentially going to use a lot of molecular tools that bacteria have to actually do this kind of fancy cutting and pasting so that we can sequence this and figure out what the genes are and what the gene expression level is.

There's a particular restriction enzyme. So a restriction enzyme is actually a component. It's a protein certainly, but it's an enzyme that's present in bacteria that goes around cutting DNA at particular sequences. There's a lot of restriction enzymes out there. If you get the current catalog or the current poster, there might be on the order of, let's say, about 300 of these restriction enzymes.

And there are many ways to define them, but one way people define restriction enzymes is based on the sequence it cuts. And so there's some that cut DNA, giving absolutely flat ends, and there's some that actually cut leaving sticky ends.

And so here, this particular restriction enzyme called NLA3 takes any strand of double-stranded DNA and cuts it at the CATG, but it leaves the CATG as an overhang. It doesn't cut it flat. So these guys saw this, and they said, wow, we could do some interesting things, a little bit of cutting and pasting, and actually take advantage of this.

So to be clear, here's the sequence. Here's the CATG and the unique tag. And basically we're going to-- even though the strand might be longer than this, it's going to get cut like this. Now to use some more fancy parts to this, remember how I told you most messenger RNA has a strand of A at the end, because as it's being copied off the DNA it starts to stutter in some ways and just leaves a poly(A). And that you can bind to it, or you can actually find it by using a poly(T).

Using their technique what you can do is instead of just starting with T, start it with Ts and something sticky on the end. In other words, it's labeled in such a way that we can then fish these out after we apply a magnet to it, for example. Or some people use streptavidin biotinylation as a common way to actually pull, fish these things out when you run through a column, for example, with particular beads, however you do it.

So now we have a whole bunch of the double-stranded DNA, except we have this little overhang here. We've got all the As and Ts on that end. We fished them out of this big cellular pool, this cellular mess. Now what we're going to do-- it starts to get a little tricky. But now we have this tag here. Keep your eye on the red. That's all I can say here in this technique.

Because what you're going to do is, in the end you're going to see how we're going to be able to use a sequencer to tell these RNAs how much of each RNA is present. So we've got one RNA going this way, and we actually got this light blue for another RNA going this way. And you're going to see how they're going to end up being next to each other.

So what we have to do then is add an adapter that has another sticky end here, and add it to this side here. And we're going to use another restriction enzyme that's going to cut at the CATG, but it's going to cut after the tag. There's one particular restriction enzyme called BsmF1 that looks for CATG but doesn't cut at the CATG. It cuts 14 base pairs downstream from there. That's why our tags have to be 10 nucleotides long.

A lot of fancy biochemistry here to get this to work. So you got a tag here, and you got tag here, and you got a little adapter here that we've added on this side. And then finally, what we're going to do, we've cut it here, and we've cut it here.

What we can do is add something called a ligase that basically takes all blunt ends and just tries to make pairs together. It tries to bring any small strands of DNA with blunt ends and tries to attach them together. So we have a blunt end that was here, and a blunt end that was here. Here's the first tag, and here's the second tag. And we're going to have a whole bunch of these.

And what we can do is just run it through a sequencer, basically. We're going to do some more fancy footwork to get these to be as compact as possible. And in the end we have one stretch of DNA with just one tag after another, with some CATGs in the middle as spacers so we can make sure we're staying synchronized.

Now what you do is run this through an automated sequencer, and then you can just get-- you can read off the sequence just like this, one then the other, then the one then the other, and then search databases to see what gene do we think this is.

Now the reason why this is an absolute gene expression measurement, for sure, is that if this gene is over-represented, it's two times more than this one, you're going to see two times more tags than this one, in theory. There are efficiencies to each one of these steps that we talked about. And this is not as simple as a microarray.

Literally you can take the sample if you're a physician at Brigham Women's Hospital, or any of these hospitals, and essentially take it to a core and get an microarray done. SAGE is a lot more involved, but in theory this is a great way to pick up low expression-- genes with very low expression levels.

So that's why it was worth at least talking about SAGE. Because if you're looking at absolute expression levels, this is where it's going. For now this is a back burner kind of technology. People still use it. There are people who are religiously devoted to it, but most people end up using microarrays today. Yeah?

AUDIENCE: Can you go back and explain just a little bit how the tag got there that way?

ATUL BUTTE: The tag just happens to be a sequence. So what we're calling a tag here is nothing we put there.

AUDIENCE: OK.

ATUL BUTTE: It's something that is next to a CATG that's on this end of the gene. If there's nothing there like that, then we're not going to find it, so that's another catch. If the first one is all the way up here, we might not even get it. So they're saying the average, if you do the math, there's at least one of these within 256 base pairs, because it's like 4 to the fourth. So there should be one there, but there might be degeneracy, or it might not be there.

So that's the whole technique. It's 10. It's CATG only because of the peculiarities of the enzymes being used to actually cut these things apart and then just look at the tags. Any questions about that? So that's basically-- that was the last thing I want to talk about in this section of gene measurement technologies, or gene measurement techniques.

So now Zack was suggesting that I talk about some of the stuff that we've been doing with diabetes, only because this is supposed to be genomic medicine. We can give you a real world example of how we've used microarrays to help us with diabetes.

So I'm a pediatric endocrinologist, and I've spent most of my time over at the Joslin Diabetes Center studying type 2 diabetes. It currently affects about 15 million people in the United States, and in fact, the CDC just estimated a couple of months ago that a child born in the year 2000 now has a 1 in 3 risk of getting type 2 diabetes.

So it's 33% likelihood now of a child born just two years ago to get type 2 diabetes. Why? Because of the obesity problem. So we all know kids have less exercise today in school, because they have to pass these tests. And now schools have to give more courses so that they pass the MCASes and things like that.

There's more TV. There's more channels. There's internet. There's all these things that cause kids to basically sit in one place. They don't get much exercise. I think there was something on CNN last week about how even toddlers one or two years old are just not getting as much exercise as they used to. So this isn't even school age. This is starting at basically age one or two.

So the question we had here was this. We know how to define-- the diagnosis of diabetes is actually sort of arbitrary. In fact, the definition of diabetes keeps changing, it turns out. But for example, at one point the definition of diabetes was having a fasting blood sugar-- so morning blood sugar-- level of 127, I think it was, milligrams per deciliter qualified you for having diabetes.

But what we know is that it's not just an on and off type of thing. All of a sudden you don't have diabetes, and now all of a sudden you do. Because diabetes is a very heritable trait. It's controversial, but some people say, for example, it's a very genetic trait. Let's put it that way.

Some studies that show that in twin-twin studies if you have one twin that has type 2 diabetes, you have a 98% likelihood that the other twin is going to get it, if they don't already have it. So it makes you think that it's incredibly genetic, but the problem is the incidence is so high. This is a common disease. It's not very easy at all to find a particular mutation that causes it. We haven't been that lucky so far.

We have found-- the field has found a number of mutations that do lead to diabetes, but they might explain the sum total of maybe 0.01% of the people who have diabetes. Those are very rare forms. The common form, which seems like it's incredibly genetic, it's still a big mystery.

The point of what we were going to use genomic medicine for here was trying to define the patterns of how people might go from non-diabetes to diabetes. There's a middle group in there. In fact, the ADA, the Diabetes Association, now calls this pre-diabetes, people who don't have a normal handling of sugar but don't strictly meet the criteria for diabetes today.

The question we had here, using microarrays, can we define patterns of gene expression in human subjects with diabetes, specifically in their muscle samples, to identify those that are at high risk not just define what genes are different between diabetes and non-diabetes but also what genes were already different in those that were at high risk of getting diabetes. Specifically these are offspring of type 2 diabetics.

So in this particular study that we did was in Mexican-American subjects. We have diabetes and family history of diabetes. So we basically have three groups, family history negative, diabetes-- so obviously these are control-- and then we have family history positives. So we got three groups here.

And you can look at the subjects that we used for this particular study. We had one that we use for the arrays, and we had another that we actually did validation work on, some of the hypotheses that we came up with. And so most of these characteristics are not different, but they are different in a number of important ways.

So fasting glucose is 99, 92, but 200 in the diabetics obviously, because they have diabetes. The two-hour glucose was higher in the diabetics, but the hemoglobin A1C was already high in the family history positives. What is the hemoglobin A1C?

Well that's one of our measurements for how high the blood sugar levels are over a span of 120 days. So in fact most physicians now who take care of patients with diabetes, we measure this hemoglobin A1C at least four times a year so that we can get a gauge as to what the blood sugars are like over the past 120 days, because that's the lifespan of a erythrocyte or a red cell. And as the glucose levels are higher in the blood, they actually glycosylate the hemoglobin in the red cell, so it's a byproduct of this.

But it's already high in the family history positive. They're not exactly like the family history negatives here. Fasting insulin already high. So the problem with type 2 diabetes-- just to make sure everyone's on the same page here-- it's not a problem with making the insulin from the pancreas. The problem is that insulin is there, it just doesn't act like it's supposed to in the target tissues.

Now to use my analogy from last week, I just had this great food from the Chinese truck again. My glucose levels are going up, my insulin levels are going up. And so insulin is now telling a number of my tissues to do a number of things. It's telling my liver to stop making sugar. I just ate a whole huge meal. I don't need the liver to start making more sugar, because I just ate a whole bunch.

But more importantly, it's telling that my muscle and my fat cells to start taking in the sugar. And what happens in type 2 diabetes is that that signal from the insulin to what the tissues are supposed to do is muted for some reason. We just don't know exactly why. But that's the problem. The insulin is there.

In fact, now these guys have a higher insulin level, because already their bodies are realizing the same amount of insulin isn't doing what it's supposed to. So the pancreas sees that the sugar levels are not going down after the meals, for example, and it's saying, I need to ramp up the amount of insulin I'm making. But the primary problem is felt to be in the target tissues, insulin resistance.

So the groups are already different, and that's going to actually make this analysis complicated. So we have five diabetics there, four without diabetes but with a family history of diabetes, and six controls here. And when you do this, you can compare each group to each other group.

So we can take-- we took muscle samples from these individuals. If they were on medicines for the diabetics, they were actually taken off for a number of days. They had no strenuous exercise beforehand. And it's all informed consent, and you get the muscle samples. You extract out the RNA. We put them on these microarrays. The microarrays when we did these experiments measured about 7,000 genes. Today when you do this for the same price you get 44,000 genes.

And so we can use a number of different analytic techniques that you're going to be learning about over the rest of the semester-- exactly how we make the distinction of what genes might be different, what genes are not changing. But if you compare family history negative to diabetes, we find 187 genes that are different.

Even if you compare family history positive to diabetes, there's 166 genes that are different. So fewer, and there's some overlap. So the 55 genes between-- there are 55 genes that were in this list and in this list. Most of them coded for all sorts of interesting proteins. For example, mitochondrial proteins that are involved in actually energy metabolism, or ATP synthesis, oxidative phosphorylation, which we'll talk about in a moment.

So this is one common way of how we actually like to look at these genes. It's called a heat map, or even a dendrogram, depending on the different programs that you use. But the green indicates-- so each of the rows indicates a gene. Each of the columns typically represents a sample. And the green indicates a gene that's higher than the mean, and the red indicates a gene that's lower than the mean.

And what we're showing here are a subset of those 55 genes that actually have something to do with energy metabolism. And primarily most of the genes that have something to do with energy metabolism are higher in the family history negative, they're lower in the diabetics, and they're already low in the family history positive, so the ones that are at risk for getting diabetes.

So now we've got a piece of this puzzle. We see that transcription of some of these mitochondrial genes, metabolic related genes, is down. That might be causing decreased oxidative phosphorylation. So what happens is, fat is building up in the muscle and the fat cells.

And if you have less of these genes and presumably less of the proteins that we're using them as a proxy for, we're getting decreased lipid oxidation. As fat or lipid builds up in muscle, it actually can cause the insulin to not work as well. So maybe that's how we're getting to insulin resistance because of an increase in the lipid in the muscle.

But there's a huge half of this that's still unknown today, and that's the part I want to focus on for a second. People show this slide all the time in all sorts of different contexts. Huge formula here, huge formula here, the miracle occurs in the middle. And still the biggest problem with this type of study is that this miracle still has to occur to figure out why do we have this list of genes.

What is it about this list of genes that's special? How are we getting to this set of genes? What's the biology behind it? So although many of my colleagues will argue with me about this, you cannot ignore the biology here. It's getting harder and harder to just say, here's my list of genes, go run with this. You need to know what is it about this list, what's causing this list of genes to be so special.

And unfortunately right now to informaticists it looks like a miracle occurs, that someone has the insight to look at one particular factor, and you actually get a causal chain of events, and then you can get your list of genes. Here what we realized is that most of these genes are downstream of a transcription factor called NRF-1 one and another one called PGC-1alpha.

It'd be great if we had some database of these things. Of course, we don't. It'd be great if NRF-1 was even on the chip. It's not. It's on today's chip, but it wasn't on these chips from three generations ago. So no amount of informatics and network building and Bayes networks or any of that would have got us to this particular hypothesis, because the gene wasn't even being measured at the time.

But all of these genes that I'm showing here-- again, a subset of the previous subset-- are all downstream of NRF-1. So that led us to actually think about the hypothesis, maybe NRF-1 is down, and that's why all these other genes are down. Because it's much easier to come up with a causal explanation requiring just one thing to be wrong than 50 things to be wrong. That's Occam's razor. Simpler explanations.

So when you look at NRF levels-- again this measured a different population using a different technology, RT-PCR. You have the family history negatives here, positives here, diabetics again. And the gene actually has a statistically significant drop in the diabetics.

But what's puzzling is that it's not dropping in the middle group. But we saw those genes were down in the middle group, so this isn't a perfect explanation at all. Another piece of biological knowledge led us to realize that NRF is actually itself is downstream, or actually it's co-activated by PGC-1alpha. So when you actually measure that, you also get the clue that PGC-1alpha is down in both of these two groups, as well as another form, PGC-1beta, which is down in these two groups.

The biggest problem, though-- so now the hypothesis is that drops in PGC-1alpha and NRF are actually what's causing the difference between diabetics and non-diabetics, and it's already showing a difference in the pre-diabetics. The biggest problem, though, with this study is that what's the cause, and what's the effect here?

I'm already showing you that the glucose levels and insulin levels are already high in that middle group. How do I know this isn't just an effect of having high sugar levels? How do I know this is causing the high sugar levels? And that's the problem with this type of approach. We can easily get samples now. Samples aren't the hard part. The arrays are not the hard part. The analysis is not the hard part. Informatics is not the hard part.

There's plenty of people around now, plenty of software you can download. The hard part's the interpretation in the end. Even though we have all of these genome scale tools, we still have nothing that makes the interpretation that much faster or that much easier.

So now it's still routine. If I go to the Joslin Diabetes Center, I look at their core facility. I was just there a couple of days ago. They've literally built a pyramid of all the arrays they've done so far, on the order of about 1,500 almost 2,000 microarrays already done at the Joslin Diabetes Center.

There's maybe been about two papers published so far with all of this array data, this one which had about 30 samples, and another one that had about 10 samples. And still there are thousands of arrays that have already been collected. The data is already sitting there. Analysis has already been done.

But people cannot make the leap to the next step. Why is it this list of genes? What's so interesting about that list of genes? I think that's going to be the hard part now in the next few years. This is not that novel to take samples and put them on arrays. That's not what the medicine part of this is going to be about. It's going to be trying to get down to simpler explanations for what we're seeing here.

So this is a final hypothesis here, perhaps that PGC-1 is down, and that, along with NRF, is actually causing this list of genes to be different. Any questions about that so far? Any thoughts?

AUDIENCE: Do you get any-- we look at that AMP kinase pathway activation in terms of toxins?

ATUL BUTTE: A great question. So the question is, have we looked at the AMP kinase pathway? So we know, for example, when people exercise, their sensitivity to insulin improves, even if they have diabetes. People who exercise, the same molecule of insulin has more of an effect. And that's mediated through a different pathway, through the AMP kinase pathway.

Did we look at AMP kinase levels here? No. Did we look at the expression levels that might have been on the chip but they weren't different? Did we look at activity levels or protein levels? No. But you can bet in that towering pile of 1,000 arrays-- Lori Goodyear and others have done these arrays already-- but they're still stuck the same way.

So what is it about these lists of genes that's going to get us to a hypothesis here? It's very unwieldy. It's very hard to deal with these lists of hundreds of genes. It's not very validating anymore. It's not very fun to work with these genes. The hard part is to get down to the causal mechanism for those genes, and that's where we have the least information right now, like what's upstream of what, for example.

AUDIENCE: Can you go back to the chicken and egg conundrum [INAUDIBLE] I was just wondering how do you sustain a direct decrease in ox pause, lipid oxidation. So the idea is that you're not using your glucose efficiently. when you blood sugar is high.

ATUL BUTTE: Exactly.

AUDIENCE: You're not metabolizing [INAUDIBLE]

ATUL BUTTE: Primarily you're not generating ATP as efficiently as you could in a muscle.

AUDIENCE: Yeah, well if you're generating ATP inefficiently, and that's the equivalent of having off proxy coupling or something. [INTERPOSING VOICES]

ATUL BUTTE: Exactly.

AUDIENCE: This is you're actually not generating the ATP and instead--

ATUL BUTTE: Yeah, the pipe is--

AUDIENCE: --substrate.

ATUL BUTTE: Exactly. The pipe is not being filled, exactly. The pipe is not being filled as efficiently as it can be.

AUDIENCE: Right.

ATUL BUTTE: Absolutely.

AUDIENCE: If we use that and we go back maybe to sort of backwards reasoning, wouldn't that basically suggest that the initial perturbation is in NRF and CTAR?

ATUL BUTTE: So there's other ways we can-- there's another piece of information that we can bring to try to figure out the ordering of them. So there are sequenced polymorphisms that are known in these genes, specifically in PGC-1, that go along to having diabetes.

So we think, well, if there's people who have an abnormal sequence and they get diabetes, maybe that's what's causing it.

AUDIENCE: Right, right.

ATUL BUTTE: That's what puts this upstream of actually having the high sugar levels. But in this sample, without that a priori knowledge, you can't tell that. That's the issue.

AUDIENCE: Doesn't it suggest that, though, or was that a giant leap? In other words, if the glucose level was high, if the initial perturbation was the elevated blood glucose--

ATUL BUTTE: OK, yep, if it were that.

AUDIENCE: --that doesn't explain why you have upregulation of-- or is it downregulation of NRF and PGC, whatever it is. It up or down?

ATUL BUTTE: The NRF is down. All I'm saying is with just this experiment, I could put this square up here and say that is leading to this, leading to this, leading to this, leading to this. I can't make that leap that this is causal with this.

[INTERPOSING VOICES]

AUDIENCE: More suggested that the square's down there, or am I not thinking about this correctly?

ATUL BUTTE: With the sequenced polymorphisms and things, it makes me think that it's down here. But just the experiment by itself without the a priori knowledge, without knowing what the literature tells you, how would you know that, because these patients already have some degree of resistance. That's the issue here. You need to know what-- the biology. You cannot ignore the biology.

AUDIENCE: Right, but the question is just the-- [INAUDIBLE] Because if you look at the control group, the body mass index seem to be much lower.

ATUL BUTTE: Absolutely.

AUDIENCE: [INAUDIBLE]

ATUL BUTTE: Exactly right. So you picked up on that as well? The body mass index is different, too, because it's hard to find. I mean, that's life in America now. Everyone is obese. And so even teenagers that are obese might be in this middle group now. They will have some impaired glucose tolerance.

AUDIENCE: Of this case in the control group, control group was normal.

ATUL BUTTE: The control group was-- I mean, their body mass index is still 30. That qualifies as obese today. And the family history group is actually slightly less. But they're all obese, basically, especially the controls.

AUDIENCE: But you can see the distribution on the graphs on the next page.

ATUL BUTTE: Absolutely, absolutely. This one here.

AUDIENCE: Yep.

ATUL BUTTE: Yep, exactly. This is one gene versus the other, and you can see the body mass index here, exactly. And you got one all the way over here, exactly. That's destroying the mean. So these are not otherwise equal groups. That's the problem.

Now you have to-- so why am I even showing all this, if I have all these problems with this analysis and this data? The whole point of this kind of course is not just to teach you how to do these types of experiments. To me it would be great if you knew how to interpret these things that come out in the *New England Journal* every day.

Because every month now we see a microarray paper, say, in the *New England Journal,* or *JAMA,* or other equivalent publications. And I think we have to teach people how to read these things with a critical eye. Because it looks so fancy to come out with 44,000 genes, and these are the ones that are actually diagnostic.

But there's actually problems with the experimental design. And if the experiment was not designed in exactly the way to answer the question that was being asked, no amount of fancy arrays is going to be able to help you. You can't salvage that. That's what I'm trying to get across, is how to be critical here.

I think the same thing-- when you see a bunch of genes that distinguishes ALL from AML. Let's think about all the other things in their context that could be making a difference here. When were the samples acquired? Where were they acquired? Are patients with AML preferentially sent to one center, and the ones with ALL sent to another center?

My favorite example I love to give is this. Let's say you're looking at a solid tumor, and you're trying to distinguish the genes that are different between the metastatic form of this cancer and the non-metastatic form. If you've spent any time in the wards, you know how surgeons work.

If you have a patient, and if a surgeon is looking at their caseload for the next day, and they have a patient with metastatic and patient with non-metastatic, the patient metastatic, it's already metastasized. The surgery is going to be a hell of a lot longer, let's say, because they have to go, check this, check that, clean this out, clean that out.

The surgeon might say, I'm going to put that patient as the first case of the day. And the simpler case, I'm going to put on as a follow-on or an add-on case. Keep them waiting in the waiting room or in the recovery room ready to go, and then we're going to do them after this first case is done.

So now imagine a study where you look at the difference between metastatic and non-metastatic forms of this particular cancer, and you've got these samples from these surgeons who preferentially put the hard cases in the morning and the easy cases in the afternoon.

I'm an endocrinologist. I know that are hormones that are different between morning and evening, like insulin, growth hormone, cortisol. Who knows how many genes are downstream of those hormones? How do you know that the signal that you're seeing isn't because of that bias or that confounding there?

Now this is not revolutionary stuff. If we looked at clinical tests, the clinical trials in the past, we knew to look for these things. But all of a sudden now that we have these fancy red/green-ograms, and we have these microarrays, people have forgotten about all of this traditional-- these traditional confounders.

We're measuring across the entire measurable genome. The signal that you get just from endocrinology, from circadian cycles, might be stronger than the signal that you actually think you're measuring. Now you go back and look at the last three microarray papers that are in the *New England Journal of Medicine* and just look to see if you can figure out what time of day these samples were acquired. No way.

AUDIENCE: How sensitive is something like that? Well, two-part question. One, is it a matter of is there an 8:00 AM, an 8:30, and that sort of thing. And is that even known?

ATUL BUTTE: Right. So it might not be known, but then what I would say is we should randomize across that variable. You could turn the lights on there. We should randomize across that variable. So for example, someone should be looking at the clock and say, well, these are all acquired-- there's some that are here, some that are here, some that are here, and just convince yourself that it's not just a gross bias. I'm just making a fake story here that's believable.

AUDIENCE: Couldn't you just do normal controls and do a raise on them in the morning and the afternoon?

ATUL BUTTE: Sure. You can try to compensate for some other way. But it's going to be relative to their hormone levels, too, probably.

AUDIENCE: I guess the example that you gave and just going out there, but I guess it's different than in a laboratory setting, when [INTERPOSING VOICES] cells and eventually they die.

ATUL BUTTE: Exactly. Exactly. We've taken this technology that we've used in micro-- used in cells, in cell culture from cell lines and rapidly moved it to patients that exist in a context. Even the middle ground of just using them in lab mice, look at the same mice, same conditions, same everything. There's genes that are different.

One of the mice was an alpha male in that little cage, and one wasn't. There's differences between things that we think are otherwise the same. But now we figure it to this extreme, because there's a course, you're in a course called Genomic Medicine, and that's what we're doing today. These are the problems that we have with that type of approach.

I kid you not. Go back to the last few microarray papers in the *New England Journal,* you will not see what time of day, even if you go to the supplemental material. Nobody might not-- I think people don't even write that stuff down. And it's a problem. I've only seen one paper so far that's addressed this issue.

AUDIENCE: So what would your suggestion be, I mean, obviously to us?

ATUL BUTTE: Control for that.

AUDIENCE: It's to control for it, but in the interim, while there's labs that maybe really aren't thinking about this.

ATUL BUTTE: Yep. You can do what [INTERPOSING VOICES] Sunil is saying.

AUDIENCE: Stop what you're doing.

ATUL BUTTE: No, but you can try to model that. So take some other tissue that's related and see just across a sampling of time points what genes are different, and maybe you can then subtract out that effect. But more importantly what I would say is, if you're going to make a list of genes that you think are diagnostic, we're going to have to-- as physicians, we're going to have to look at that with a very skeptical eye.

Because you cannot control for everything here. But we're measuring across so many genes and have so few samples, that no matter what, we're going to be overfitting the data here. That's what I'm worried about. So we'll be overfitting what we have. No amount of control samples is going to help us fix that problem, I think.

How are we doing for time? OK. Let's talk-- let's go a little bit more into the diabetes. And so what I was planning to do is talk about one more diabetes project that we've been doing. Then I'm going to talk about some web-based resources for doing some interesting hypothesis generation, and we'll just end it at that point.

So all of you are-- so not all of you are actually are physicians or medical students, but nonetheless I'll present this case of a patient that I saw. This was a female presented with Acanthosis nigricans, random glucose of 162. So Acanthosis nigricans, do you know what that is?

There's a dark patch that you can get in the back of your neck, and your elbows, and your underarms. And what that means is that your insulin is too high. Because insulin is telling the melanocytes through some unknown mechanism, it's telling the melanocytes to become darker.

So you just look at people on the street now, and look to see if they have a dark patch here. It's not dirt. You can't scrub it off. That's a sign of Acanthosis nigricans, so someone's having a problem with insulin. It's too high.

Did more studies, the LDL is sky high. LDL is high, total cholesterol sky high, triglycerides are very high, HDL is low. So it's a metabolic syndrome, or otherwise called Syndrome X. Fasting glucose 133, that meets criteria for diabetes. Insulin is 27, which means this is high, but this is very high, bringing that glucose trying to compensate for that.

Hemoglobin A1C is 7.3. The normal range is 4 to 6. And again to drive the whole point home, this is a 12-year-old girl with adult onset diabetes. So this is her body mass index. This is a 97th percentile curve, and this is where she is, well over the curve here. And now she has type 2 diabetes. She meets criteria for type 2 diabetes. She was not even entering seventh grade when I started doing summer vacation last year. 1 in 3 kids is going to be like this. That's a problem.

So one way to study this problem is to study obesity, and one way to think about obesity related to diabetes. So one way to study obesity is to think about-- specifically what I'm interested in studying is how adipogenesis, or the process of making fat cells, might be related to how insulin works.

And the best established way to study that is to look at the insulin receptor. So now we're zooming into one of these target tissues-- remember, liver, muscle, and, fat-- and the insulin receptor is a protein that sticks in the membrane of the cell. Like any other protein, it's got a gene coding for it in the DNA, gets transcribed, gets translated. The protein goes out to the surface. There are known mutations in the insulin receptor that can cause you to have problems with glucose handling and diabetes.

In fact, there's a syndrome called leprechaunism, where these kids look like leprechauns actually. They die pretty quickly after birth, but they have mutations in the insulin receptor. It turns out you can actually be born without an insulin receptor. There's been a null person. There's been one patient, I think, that was found to have no insulin receptors. And obviously the insulin level was sky high in this baby at birth, because it's not acting anywhere.

But what happens is-- what's supposed to happen is insulin binds to the insulin receptor, and then these phosphate groups are added to itself. And that changes the shape of this molecule in such a way that it can actually start to interact with other molecules. And in fact, the insulin receptor also adds phosphate groups to other things.

In fact, these are insulin receptor substrate 1, insulin receptor substrate 2, 3, and 4, and there are some others here. So in collaboration with Ron Kahn, President of the Joslin Diabetes Center, he's created some mice where they've basically knocked out IRS1, 2, 3, and 4 in separate mice. So these otherwise normal mice missing this one particular gene.

And if you take fat cells from these mice-- in fact, if you take pre-fat cells or pre-adipocytes, you can make them go into fat cells using a standard cocktail of hormones. This has been done for the past two decades. You take pre-fat cells and make them into fat cells with a particular cocktail.

And what happens is if you take normal pre-fat cells and make them into fat cells and stain for the fat or the lipid, you get a big red circle here. This is a dish of the cells. If you knock out IRS 1 and try to do this, you cannot get the fat cells to form. So again, pre-fat cells, the same cocktail, you don't get fat cells in here. There's no stain. There's barely a stain for any red here.

And it turns out if you knock out IRS3, 2, and 4 in the middle here, you get a gradation of the phenotype, interestingly. So it's not all or none. If you knock out the other IRS molecules, you get a gradation of the phenotype, curiously.

So what we did is in collaboration with Ron Kahn, we actually made a list of genes that behaved in this pattern. We you don't have numbers here, but we're just trying to find genes that go along with that pattern. So let me show you an example of one of these things.

Here's a gene. Here's wild type 1. Here's wild type. Here's IRS1 knockout, so no fat cells, easy to make fat cells. And we're taking the genes-- we're taking RNA before we try to make them into fat cells. The hypothesis here is, what genes are different before we try to make them into fat cells that might be impacting the process of making fat cells?

Now we'll take a second here, and you're probably asking me, wait a minute. We were just talking about humans. We were just talking about this kid that was obese, and now we're talking about mice here. I'm going to bring this back to the humans in a minute, because I want to intersect this with other data sets and show you where I'm going.

But suffice it to say we have about 80 genes that follow this kind of pattern. They're high in the normals. They're low in the IRS1 knockouts before adipogenesis, and they're in this kind of monotonic pattern here, or vice versa, going down or up. Came up with a list of 88 genes that fit this kind of pattern.

Now put those 88 genes on hold for a second. Let's talk about a second disease. I happened to be involved with studying this particular syndrome called progeria. Progeria is advanced aging syndrome in children. In fact, it makes-- it's very rare, so about 1 in 8,000,000 births, and it makes children look like they're getting older. We've all heard of these kids and this kind of story.

And specifically they lose their hair. They have alopecia. They're certainly short. They're shorter than age-matched kids. That the skin changes that make them look old. They don't go through puberty. They have poor weight gain. And usually they die at age 15 because of cardiovascular problems, atherosclerosis, get strokes, they get heart attacks.

Now even though they look older, they don't get everything associated with being old. They don't get Alzheimer's. They don't get cancer. It's only the atherosclerosis and these other changes here that they get.

Now as you might know, the gene for this was just recently discovered. Very amazing that they found this gene. I got involved with this particular syndrome only because of my friend Leslie Gordon, who was a medical school classmate of mine and who has a son with this. And she's got an MD-PhD. Instead of just studying this disease herself, she built a foundation called the Progeria Research Foundation. Got Francis Collins interested, the guy who's basically doing the Genome Project, the head of NHGRI.

And within 18 months of forming this foundation, they found the gene for this. Incredibly difficult problem. There are people who on the first day would not believe this is still a genetic syndrome. One in 8,000,000 births is way less than the mutation frequency. We have a whole bunch of diseases that at 1 in 40,000, 1 in 100,000. 1 in 8,000,000 seems like it's so rare. How could a mutation be that rare? So it's very questionable.

And secondly, they had a very hard time finding any families where two of these kids were in the same family. So it was dumb luck that they were able to find this gene, because in one of the individuals-- at any point there's only about 50 of these kids on the planet, really.

And one of these kids, in one of the samples, someone with a sharp eye saw a piece of chromosome one was actually inverted. The bands, those banding patterns, was actually different in that one individual. And that got Francis Collins to think maybe it's something on that particular arm of chromosome 1, and they sequenced all the genes there, and they found the particular mutation.

It's a gene called lamin A. And it's not just lamin A that causes it. Lamin A has actually been involved in a lot of other diseases, which we'll talk about in a second. It's just this one mutation in this one gene that causes it.

Now amazingly enough, in the same issue of *Nature,* less than a year ago, the mouse model for this was published. How is that? Because this group at National Cancer Institute realized-- lamin A is known to be involved in certain forms of muscular dystrophy. They said, well, they wanted to create a mouse model of a particular form of muscular dystrophy. They accidentally created a mouse model for progeria.

So the same issue-- you got the gene that causes it. You got diagnostic tests. And you got a mouse model to study it, and just in a tour de force issue there. Imagine how long it takes for these things to happen for a disease. Start to finish, 18 months here for something that affects maybe 50 people. So if you put your mind to it, you can solve these diseases. It's amazing what the tools and the people can do today.

So in collaboration with the Progeria Foundation what we've done is this. It turns out over the past 20 years when people thought they had kids with progeria, that they were trying to treat kids with progeria, they had no idea what to do. So one thing they did is they took fibroblasts and they saved them in national repositories.

So there's a cell repository called the Coriell cell repository where you can order fibroblast cell lines from patients with exceedingly rare disorders, including progeria. So we have gotten three cell lines from patients that were stored in the Coriell collection over the past 10 years, and we ordered them three times to make sure that we're controlling at least for passage number and things like that.

And we compared them to age-matched normal fibroblasts from the same Coriell repository. Looked at what genes are different between progeria and non-progeria. And these are more up-to-date arrays. These are U-133 arrays on two-- the whole genome on two chips-- that gave us about 33,000 genes, and used a particular analytic technique, and we got about a list of about 366 genes that were different between fibroblasts in progeria and fibroblasts from age-matched controls.

And so if you look at where these genes are and what categories they're involved with, you see interesting things like development, signal transduction, cell adhesion. The only thing I want to point out is this. These tables are made using a catalog called Gene Ontology. You're going to learn about that more in the future. And every week or every other week, the list of known properties of genes and proteins gets updated. But still the majority of genes unclassified, 161.

No matter how interesting or how comprehensive we think Gene Ontology is, the vast majority of genes just have no-- we have no idea where they are, what they do, what roles they play. These are where the proteins that they're coding for actually take place. Nucleus, membrane, unclassified 131 again. No idea where these proteins even are in the cell. That's life today. We have a great taxonomy here of all of these terms and structure vocabulary, but we have no-- there's very little data for most genes in this still.

And there's a bunch of genes that were upregulated that seemed to already be known to play a role in atherosclerosis. So this paper, again, talking about genome medicine, this paper is going to end with this list, only because a number of these are targets, and pharmaceutical companies actually have drugs against a number of these already.

For kids with progeria it's basically a death sentence today. We have no medicines. We treat them with nothing today. Maybe some people treat one or two kids with growth hormone to see if growth hormone does something, but that's basically like a voodoo medicine kind of thing. We have no proof of that doing anything. Now we can at least get to some information to possibly even think about clinical trials here.

But let's be frank here. These were fibroblasts that were frozen for years in the cell repository. And now we're making a claim as to what their blood vessels look like. And not just the blood vessels, but the intima of their blood vessels. This is not the same cell. The fibroblasts might have been taken from their cheek, but we're just making a claim here because we don't have access to the real tissue.

So one of the reasons why I'm mentioning progeria here is remember that in picture you just saw, one of the ways that these kids get diagnosed is because by the time they get to age two, they start to lose their hair. But even more interestingly is they lose all their fat cells.

It turns out in progeria you lose all your subcutaneous fat. And by death at autopsy it's very tough to find any fat in the body. It makes you think something is up with the actual process of making fat cells in these kids. In one study these kids have been known to have some insulin resistance, but it's just one study almost 10 years ago.

In the mouse model, the last paragraph of that *Nature* paper says that they saw some process where muscle was re-differentiating into fat. They said they think there's some developmental problem going on, but it's not very clear. Other mutations in lamin A cause lipodystrophy, which is a syndrome where you lose all your fat cells. So it makes us think that there is legitimately something wrong with fat cell development in kids with progeria.

So now I told you about-- where are we here? Hold on. I told you about 80 or 90 genes that were in the mouse models going along with fat, so about 360 genes that are different between progeria kids and age-matched controls, and they have some problem with fat.

I have a third data set here, which was published by Gary Ruvkun's group over at Mass General Hospital. This came out about last year. And now this is in a totally different species. This is in the worm. To give you an idea of what we can do today, what this group did is they knocked out every single gene in the worm genome.

So a worm has about 16,000, 17,000 genes, and they basically serially knocked out each and every one of these genes to look at what happens to the worm. Now in a worm it's actually pretty easy to do. You can actually get bacteria to make a particular sequence called an RNAi that interferes with the worm's ability to make that gene. So you don't even have to change the DNA. You could just effectively shut down the amount of that gene that's being made in the worm.

And basically it's very easy. You don't even have to inject these worms. You essentially set up 16,700 buckets of bacteria, and just put the worms in it, and they eat the bacteria, and basically they knock out the gene themselves. That's literally how it works. All you have to do is get the worm to eat the bacteria that makes the RNAi. It's very easy. It knocks out the gene in that worm.

So what these guys did is-- they had back-to-back*Nature* papers on this process. The first one was on a whole bunch of different phenotypes, but the second paper was looking at that, amazingly enough. And here there's a picture of a worm, and they're using the same stain for lipid that I showed you two or three slides ago, what we were using on the mouse.

And basically they're saying, for example, if you knock out daf-2, you get an increase in the fat. If you knock out daf-2 and daf-16, you get a subsequent decrease, daf-2 and daf-3, slightly more. And they can quantitate the amount of fat that results after you've knocked out any of these genes.

And essentially they've made a list. It's available on the internet, amazingly. If you knock out any of these 112 genes, it increases the amount of fat in the worm. If you knock out any of these 305 genes, it decreases the amount of fat in the worm.

So now we have three big data sets, one in the mouse, one in the human, one in the worm, two microarrays, one RNAi, spanning all of these. But they all have something to do with each other, because they're all thinking about fat and fat storage.

Fat storage in the worm is very different than a human, but still a number of the genes are actually intersecting. And really what we're talking about is intersecting these three data sets. We got the worm, we got the mouse, we get the progeria, and we're looking at the intersection of all these three.

So this is what I've been working on over the past six months, actually. And this is what we're calling integrative biology or integrative genomics. Because these large data sets are not-- they don't just exist, and a large majority they're actually publicly available.

You could go to the National Center for Biotechnology Information, the NCBI GEO, and pull down any number of these kinds of data sets. And you could do it today. It's for your final project. You can go download any of 600 experiments where the microarrays have already been collected. The only hard part now is to have the interesting question.

You can't just pair something with something else. You want to have to actually think about an interesting question, because the resources in all likelihood are already there to actually try to answer your question. Whether you're interested in circadian rhythms, or fat development, or anything, that's the hard part now, is knowing enough biology to actually ask an interesting question in this way.

So it turns out if you look at the genes in intersection, just to give you an idea, these WNT genes-- WNT is actually a hormone that's made. WNT genes are actually well known to be involved with adipogenesis, and they are actually living in the intersection of these lists.

So the worm egl2, which is called WNT7B, increases the amount of fat if it's knocked out. The mouse WNT6 and 10A increase with dysadipogenesis, depending on the IRS knockouts. And WNT5A and WNT7B are down in progeria. So that's an example of a gene that lives in the intersection of these three data sets.

I'm just amazed that we got any of these things to intersect, because they're so different. Three different species. We're not even looking at the fat cells here. With the progeria kids we're looking at a fibroblast, which is a proxy for the pre-adipocyte, which is a proxy for the adipocyte. But still you can get some signal this way.

This picture is just the same. Even my own daughter knows the difference between a human, a mouse, and a worm. So there's caveats to all of this. So we're doing a lot of this validation ourselves now in the bioinformatics group over at Children's. But the point here is integrative biology. To me this is where I think a lot of fun is going to happen for me in the next few years.

We have a lot of these data sets, and now I'm trying to come up with interesting questions so I can put these together to try to get a massive handle on these processes that other people are just using one of these days that might not be able to get a grasp on.

We've got gene expression. We've got proteomic data. We've talked about that. We've had genome scans. Joel Hirschhorn is going to be talking about how you actually try to find genes that are associated with a particular phenotype using SNPs. We have the clinical measurements. We have all of these data points that are collected in all of these hospitals. I'll talk about each of these separately.

ENU mutagenesis is a new technique where-- it's actually not so new-- but you could take mice and give them this particular chemical, ethyl-nitrosourea, that causes mutations in their gametes. So then all of their offspring basically are mutants. And you do this for a couple of generations, and you find the interesting phenotype that-- wow, all of a sudden, this one has diabetes. You go back and see, what did we mutate?

We basically mutate it to see as much as you can. You characterize the output and then go back to see what-- exactly where did we put the mutation in? That's ENU mutagenesis. RNAi we start to talk about with the worm, where you basically can just knock out any particular gene very quickly.

People are starting to do this with humans and specifically with mouse models, but there's already a number of pharmaceutical companies within a five-mile radius of here trying to come up with drugs like that. And of course, most importantly, the prior biological knowledge. This is what's going to drive the interesting questions here. Any questions about that so far?

All right. I'm going to talk about something just to bring this home so we can talk about-- at least you can see how easy it is to actually do this. Let me skip a couple of slides here.

So yeah, we can actually talk about some of this. So let's say you're interested in doing some of this for a final project, that you want to actually put some of these data sets together. There are quite a number of sources where you can get microarray data and other genome size data today.

And if you see some things in the description that you're actually biologically interested in, you might be able to run with that for a final project. I just want to point out a few of these data sources and then talk about one way how you can put this together even just using a web browser.

So the cardiogenomics program is a source where I think there's over 150 microarrays now. The headquarters of this is over at Beth Israel Deaconess across the street. They have mouse models of cardiac development, so basically they have embryos, and they have sections of their heart. And basically all the genes are turned on and off at different parts during development.

They have all sorts of different components of cardiac growth. So for example, people know that if one pathway or another pathway is involved in a heart failure, they have mouse models where those components are already knocked out. And they have a bunch of measures made in time series.

So they take a normal mice, and they put them in this tank to swim, and they make them practice swimming a certain number of hours for weeks then look at their hearts. And this is what swimming does your heart, for example. They have those lists of genes available today. These slides are going to be online, so you don't have to copy down every single URL. If you just even write down the heading, then which one to look for.

The Whitehead Institute, obviously the mother lode of these arrays. At least 12, maybe 14 publications now. Many different types of cancer, that's their primary focus there. A lot of clinical measurements with some. A lot of these are clinical samples, but they might have one or two columns of clinical measurements.

Was this a smoker? Was this a non-smoker? That kind of thing. How long did this patient relapse, not relapse? But you've got leukemias, you've got solid tumors here, brain cancers. Things like that are in this data set.

DC Children's, or Children's National Medical Center, 500 arrays for many interesting human diseases. Muscular dystrophy, dermatomyositis, so that's rare. A bunch of rheumatologic things. Heart failure as well as mouse, rat, and dog models of spinal cord injury, if you want to see what are the similarities and differences there. Pulmonary disease, including asthma and heart failure. I mentioned heart failure twice there. So that's DC Children's and Johns Hopkins.

The Human Gene Expression Index I think is Brigham and Women's. 121 microarrays just from normal tissues, 19 normal human tissues. It's also an interesting list to think about. Stanford's microarrays is probably the largest besides the GEO, which we'll talk about a second. Stanford has probably close to 4,000 arrays now, measured across 11 species, covering more than 80 publications. Many different fields, most of which are yeast, but a whole bunch of human ones as well.

The National NCBI GEO, it's basically right next to PubMed now. It's the same website. Over 8,000 arrays from over 100 different types of microarrays. Never mind how many experiments. Their database can handle over 100 different companies, 100 different special microarray products that can integrate across all of these.

TREX is from TiGER, which is The Institute for Genome Research over in Gaithersburg in Maryland, outside of NIH. 500, 600 arrays for mouse and rat models of sleep, infection, hypertension, pulmonary disease. So like for example, if you're interested in asthma and you want to see differences between human asthma and the mouse model, the data is there today. I will guarantee you no one else has done that intersection.

If you find some interesting genes that you think might code for proteins, that might be in the blood and you can use that as a measurement, you can get some interesting person, one of any collaborator here in the Longwood area to actually measure this in the blood, that's it. You have a finding right there. Because the data is there. Hardly anyone knows how to intersect these things at all.

And so what I'm going to talk about here for the last few minutes is this concept of discovery portal. So what do I mean by that? Let me skip this part.

We start with in-vivo, experiments done in humans and mice. Then we had in vitro and cell culture. Now we have NHTTP. So here's an idea. Here's how you can go from an actual thought to a pretty well established hypothesis without even leaving your web browser, because all of these tools are available today.

Now for the next maybe 15 slides, I'm going to flip through them pretty quickly. I'm going to cover a number of different databases where you could get all sorts of inspirational ideas from. So I'm going to start with an interest, move to a rat model. We're going to cover QTLs while we're there and physiological findings. Go to a rat genome map. Go to multiple species expression data, and end with the hypothesis here.

This is live. You could do this yourself today. We're going to start at a website called PhysGen. Some of us in the room know about this. This is at Medical College of Wisconsin. What these guys do is they have a huge pipeline of things they do to mice and rats to measure things, measure all sorts of different quantities. How much does this rat pee versus that rat? How much does this mouse breathe compared to this mouse? Et cetera et cetera, and they have tables of all of this data publicly available.

So if you go to PhysGen, if you click on Data, and if you click on Animal Model, some genotype information for all strains. Now as I scroll down a little I see a whole bunch of the rats. Now I might remember, because I went to the Diabetes Association conference last year, I remember hearing that's the Brown Norway has some problem with diabetes.

I might not have even gone to the seminars on the Brown, but everyone kept talking about Brown Norway, Brown Norway. I heard a Brown Norway has a problem with disease. Let's go click on the Brown Norway strain report. Scroll down a little. This is the BN, or Brown Norway. And I can see, yes it does, it has diabetes on here, diabetes QTL.

So what is a QTL? So we've got the rat, and it's got its own genome, like any other creature. And what a QTL is a quantitative trait loci. So a particular portion of a particular chromosome appears to be statistically significantly associated with a particular phenotypic measurement. That's a quantitative trait loci.

So in other words, this particular region which we're going to look at in a second on chromosome 2 is statistically associated with the insulin level, the fasting insulin level of a rat. So notice if you have one version of this, you might have a low fasting insulin level. If you have another version, you have a high fasting level.

But something in this region of a chromosome is associated. We're not at the gene level yet, but we're just in a region of a chromosome as a starting point and and ending point of this chromosome, and some piece in the middle is associated.

There's all sorts of other measurements here. You've got X-ray hypersensitivity. What does that mean? Well something on this region of chromosome 1 means you get more sensitive to X-rays. You can think about how esoteric these measurements are. If you're interested in radiation effects, there's your measurement right there. Arthritis severity, blood pressure, body weight, blood pressure, blood pressure.

They're all sitting there, waiting to be tapped into. No one has enough time to look at every single one of these, but we're mandated by NIH to just put them out on the internet for people like you to just study. If you're interested in blood pressure, there's three QTLs right there that people by chance, people have probably never even looked at.

Let's click on this particular QTL to see what it looks like. Here's the name of the QTL, Nidd/gk2. It's in the Brown Norway. And here's the marker for this particular peak. So a marker means this is a unique sequence in the chromosome that is going to tell me whether I have the one version of this or the other. Let me click on that marker, D2Wox23.

And here is the PCR test I would need to do in the rack to tell me, do I have the one letter or the other letter? Or in other words, do I have the high insulin version or the low insulin version of this particular peak?

So it even gives you the test. You could just go order those oligonucleotides, find a rat that's walking down Longwood Avenue at night, and just test to see, is this a high insulin or low insulin? You can do it today.

Now out of dumb luck it just so happens that they've already done the work, and they've told us in this region of the chromosome we know there's a gene there. It doesn't have to be, but there is a gene there, S100A4. That's a gene that exists in this peak of significance.

Now what we're going to do is go take that piece of information, that name of that gene, and let's look more at that QTL region, that region of the chromosome. We're going to go to National Center for Biotechnology Information, click on Genomic Biology. I'm going to click on the rat.

And I'm going to just type in S100A4, the name of that gene I told us was in that region. Where is it in the rat chromosome? Here it is on chromosome 2. It's S100A4. If I click on that, if I zoom out a little, here's S100A4, and here are all the neighboring genes.

That's interesting. There's a bunch of other interesting genes. First, S100A4, there's A3, A6, A8, A9. This gene is one in many of a gene family. So how new genes are often formed is through duplication events like this. But to me more interestingly I see things like natriuretic peptide receptor.

Natriuretic peptide is made in a number of different sources, including the heart. It has to do with blood pressure and blood volume regulation. That's an interesting hormone. Let's see some others. Interleukin 6 there, might be some interesting things there. But let's just keep running with this S100A4 here.

Now what I'm going to do is, I'm going to look at where is this gene expressed. I got it in a statistical peak. I've seen what its neighbors are in the genome. Now I want to know, why is it that this particular gene is associated with diabetes or insulin level?

Well let's see. Where is that gene expressed in a whole bunch of different samples? This is one particular search engine where you can just type in the name of the gene and see where is this gene expressed in about 1,200 different samples. This is specifically cardiovascular.

I type in S100A4, it says, which do you mean? The human, the mouse, or the rat? Well we started with the rat, but let's just see where the human ones are expressed. Click on that. It's measured in 10 microarrays by Johns Hopkins. That's not that many. Let's go back. Let's look at the mouse one.

Mouse is measured by 142 by this group, 100 by this group, but 84 by this group. Let's look at the 84 first. And we can see that top samples. So this is the expression level. This is the percentile of that gene on that chip, on each chip. Muscle, muscle, muscle, muscle, muscle, muscle, muscle, muscle. And at least for Johns Hopkins, this gene is most highly expressed in their muscle samples, specifically regenerating muscle.

We can look to see exactly what sample there is, but let's go on. Let's hit the Back button, and let's look at cardiogenomics, so that's these guys at the Beth Israel. And again, it's highly expressed in banded smooth muscle. Now it says band here. That says band 48 hours, band 48 hours, 24 hours. What exactly did they do? What exactly are these samples?

You click on Overview, it takes you to their website. It tells you what was the experiment they did. They actually did an experiment where they took these mice, they opened them up, and they pull a little rubber band on their aortas. In half of them. The other half, they just sewed them up again. They didn't do anything.

And so they made a list of genes that go up in the heart as a response to this pressure overload. The genes are sitting there. I can guarantee you again, not a single paper has been published on this list. Because we're so busy creating more data, we have not had time to publish anything on this data.

So they were trying to study pressure overload induced cardiac hypertrophy. So now here's the hypothesis. I started with the Brown Norway rat, and the Brown Norway rat has signs of diabetes because of this gene, S100A4, because S100A4 was under that significant peak. But I think it's because S100A4 is expressed in muscle under a variety of different conditions, and I know muscle is an insulin target tissue.

So now I can get some clues to what experiment I can do next. Maybe I should go look up muscle cells for S100A4. If you actually do PubMed searches, you can see S100A4 is associated-- at least it's being studied in diabetes today. Now I just picked the first peak I saw on that website.

You can go back and find any number of new genes there and come up with hypotheses, and we didn't even leave the web browser. You didn't even have to download a file to do any of this. So you could do this today. You start just surfing from one of these sites to another from something you might have heard down the hall to an actual hypothesis.

We have about 10 minutes. I'm just going to call it quits now to see if you have any questions or thoughts.

AUDIENCE: So at what point with these various websites that have the data, what place does that data get put on the website? Are they after publication versus [INAUDIBLE] sort of like--

ATUL BUTTE: Good question.

AUDIENCE: [INAUDIBLE] structures [INTERPOSING VOICES]

ATUL BUTTE: Great question. So NCBI GEO, you could submit data to them, and then they'll keep it hidden until you say it's been published, and then they could turn it on. These other personal websites like the Whitehead one, Stanford, they typically appear after publication.

For these other websites in this PGA, the cardiovascular data I keep showing you, that data appears within 60 days of its creation, regardless of publication. NIH is more and more mandating that to be the model now. Because if they're paying for all of this genomic work, they don't want just one lab to really be able to run with it, because there's too much data for any one lab to deal with.

So for some of their grants they actually insist that this data be made available well before publication. So right now on cardiovascular and lung disease, and a little bit of sleep as well, you can get a lot of data, 1,500 arrays.

AUDIENCE: [INAUDIBLE]

ATUL BUTTE: Yep, these PGA- these Programs and Genomic Applications. PGA was funded by the Heart, Lung, and Blood Institute. That in particular has a lot of data that's already available that hasn't been published on.