TODD GOLUB: So I come at this actually originally from a pediatric oncology perspective. So I'm going to start by giving examples of a couple of patients that I saw in the Jimmy Fund clinic at the Dana-Farber that were typical.

So the first patient was a nine-year-old girl who presented to her pediatrician with-- turn these off? With a fever and bruises. She got a blood test and then a bone marrow test that revealed that her bone marrow had been replaced by acute leukemia cells-- acute lymphoblastic leukemia cells, or ALL. And so she was enrolled on what was a standard chemotherapy protocol, which was nine different drugs in rotation and combination. She entered remission in three weeks and is still alive and well.

And then, a few months later, there is a second patient-- this kid happened to be a boy, about the same age-- same presentation, same diagnosis-- acute lymphoblastic leukemia-- was enrolled on the same treatment protocol. So got the same drugs. Same hospital settings. So it was as close to a controlled experiment as one could do in a human being because, of course, response to any kind of therapeutic intervention is not only a measure of the treatment itself but how it's delivered. so that was controlled, But this patient didn't respond, unfortunately, and died about six months later.

So the overarching question, that I'm sure you've been addressing in this course, also, is generally how to understand this clinical variability and to discern whether there's a molecular underpinning to that variability. So we're particularly interested in this patient who responded particularly well and so did what is pretty straightforward, which is to do a standard chromosome analysis, karyotyping experiment using some molecular techniques known as fluorescence in situ hybridization.

Details don't really matter, except to say that, while it wasn't apparent at the routine morphological analysis of the chromosomes, if you look molecularly, it was clear that there is a translocation between chromosomes 12 and 21 that fused two genes-- actually two transcription factors-- one called TEL, one called AML-- to make this tell AML1 translocation.

And it turned out that, while it hadn't been previously discovered, actually, the majority of known genetic abnormalities in childhood leukemia-- the lion's share of those actually have this translocation. And of those patients in retrospective and now perspective analyzes, about 90% of those patients survive. It turned out that patient number two had a different translocation-- a 922 translocation-- that fuses two genes BCR and ABL-- that you've probably talked about. And it's known that these patients, give them the same therapeutic regimen, only have about 10% survival. And so these are molecular tests done at the time of diagnosis.

So what I think this really sets up-- which is now perhaps generally accepted but was really emerging at this time-- was the notion that cancer is a genetic disease, it seems like, and that the outcome clinical outcome is predictable based on molecular determinants at the time of diagnosis as long as you know what to look for. Yeah.

AUDIENCE: Was this before Gleevec?

TODD GOLUB: This is before Gleevec. Turns out that, even with Gleevec, which targets BCR-ABL, in these patients, it's not particularly effective.

AUDIENCE: What is the one [INAUDIBLE]?

TODD GOLUB: Uncharacterized. [AUDIO OUT] balance translocations that identify a single oncogene at a breakpoint. So we started thinking about alternative ways to try to think about molecular classification of cancers more generally. And this is just the obvious experiment-- two biological states. They could be clinical states or biological states-- that you collect some kind of genomic information on, like microarray data, and then have this pesky problem of actually trying to figure out how to interpret the patterns that emerge. And I'll say quite a bit about that part because it's tough.

Have you talked about the non-analytical but these laboratory aspects of microarrays at all? Let me just spend one slide talking about this. So all of the microarrays are based on the basic principle of somehow labeling mRNA or its derivatives from a cell and hybridizing to probes on some of a solid support, whether it's a microscope slide a silica wafer or something. They're cDNA arrays, and they're oligonucleotide arrays. And you can either make them yourself or you can buy them commercially.

cDNAs generally take the form of two-color hybridizations, where you simultaneously hybridize a test sample and a reference or a control, whereas the oligonucleotide arrays generally use a single color. This is completely and entirely a historical artifact of how these things were developed. There's nothing intrinsic about cDNA arrays that makes them require two-color hybridization. The only reason for using two-color hybridization is, if the quality of arrays is so bad that you need an internal control for every spot. Otherwise you can't interpret the data. And such was the case for the earliest cDNA arrays, which were the first ones to be made.

And I would venture to say that two arrays, in general, are going to become obsolete very quickly, in favor of oligonucleotide arrays-- probably commercial ones. The genome is finite. Once you have the genome represented and identified, probably not much advantage to making your own. And most of the arguments that say, oh, we can make them much cheaper making them ourselves are usually Enron-type accounting practices where you don't-- well, we didn't count the cost of all the people involved in making these things. We didn't include the fact that we spent three years trying to figure out how to do these and they didn't exactly work. When you really get right down to it, usually buying them is cheaper, and they're higher quality.

We should also mention that there are non-microarray based expression profiling methods, like serial analysis of gene expression or something called MPSS, that are transcript-counting-based methods, where you use essentially, DNA, sequencing to enumerate the precise number of copies of a given transcript in a given cell. And the proponents of this method say-- and they're correct-- this is the only way to really know how many copies of a particular transcript that are in a particular cell.

I would say that's absolutely true, but it's actually not that interesting because, for most biological questions, it actually doesn't matter how many absolute copies of transcript there are. It's not that useful information. What's useful is to have some kind of comparative experiment that tells you there are more with statistical significance-- more in this sample than that sample. And rarely do you want to know that there are 682 copies, but it is true.

But I would say that the low throughput and cost of these counting based methods greatly outweighs the benefit that you get from this absolute counting-based method. Yes.

AUDIENCE: If you have more copies of [INAUDIBLE] this is what we're talking about with [INAUDIBLE], with the same gene. They are copies of the gene?

TODD GOLUB: Well, copies of the transcript, not of the gene. So you could say-- we sequenced-- we identified a million total transcripts in a cell of which 684 came-- were transcripts from-- encoded by gene X. So that can be useful quantitative information, but not so useful to-- these experiments are still many thousands of dollars per sample.

AUDIENCE: When it comes down to a pharmaceutical company trying to figure out the mechanism of the disease and actually how to go in and develop target drug [INAUDIBLE] that might interfere-- I'm assuming that this technique would be useful. For example, if you did one of the above two [INAUDIBLE] and you don't know the quantitative number of the transcripts you have present, you're just going to, qualitatively, more or less compared to disease state or not or [INAUDIBLE] people [INAUDIBLE].

TODD GOLUB: Yeah.

AUDIENCE: If you know a number, doesn't it give you ability to make speculative speculation about how those transits are being processed and whether that's an important process and the technology becomes easier [INAUDIBLE]?

TODD GOLUB: I would say it doesn't, actually, because if you tell me-- if God tells me that there are 684 copies of a given transcript, I don't know what that means. I do know what it means that-- if there are five times as much of a given transcript in a disease state compared to a normal state, I can at least say that having one fifth the number of transcripts isn't sufficient to give you the disease state, whereas, if you just give me an absolute number without a comparative thing, I don't know what that means, And I'm not sure that saying, there's 600 in this state and 3,000 in that state-- to me, that's not really more useful information than knowing relative abundance.

AUDIENCE: So it's really a question of what the resolution is, in terms of common [INAUDIBLE]. When you get an exact number from the bottom [INAUDIBLE] up to the top get a sufficiently resolved difference between samples that you can say--

TODD GOLUB: That's right. That's right.

AUDIENCE: 104 through 500.

TODD GOLUB: That's right. And I would say, for most biological and clinical applications, knowing that the ratio is 1 to 5 gives you about 95% of the information that's useful, compared to saying, it's 500 versus 100. You might be able to think of some special experiments where you really want to know the number. But usually, it doesn't help you very much, in my opinion.

There are all sorts of sources of variability that we won't discuss. Except, there's only one that matters. And that's the biological and clinical variability that goes into these experiments. There are a lot of people that spend a lot of time hand-wringing over these very technical things, and none of them make any difference, really, as far as I can tell because they're overwhelmed by biological variability. So making microarrays that are somewhat more accurate or precise in their measurement won't actually make that much of a dent in the problem because that variability is exceeded so tremendously by biological variability. That's all I'll say about that.

So let me give some examples of applying this. And I know you've discussed these general methods before, but that's OK. So here's an experiment where we are interested in the differentiable clinical outcome of children with a brain tumor called medulloblastoma. Have you discussed this particular example?

AUDIENCE: [INAUDIBLE] but not this one.

TODD GOLUB: OK. So we had 60 pre-treatment biopsies of patients with this brain tumor. Tumors were biopsied. And we knew the long-term clinical outcome of these patients-- whether they survived from their disease or they died despite therapy. They had the tumors removed and were treated with chemotherapy and radiation.

And we knew that some of the patients survived and some of them did not. And based on that, we said, well, perhaps there are two classes-- subclasses of the disease. So we clustered the data. This was across, I think, 6,800 genes on a microarray. So each dot represents a different patient sample. We said, well, if there are two classes, let's cluster into two groups because we have a clinical suspicion there are two classes.

And of course, if you ask the algorithm to cluster into two groups, in this case, it's something called a self-organizing map. But it doesn't matter. You could take the two major branches on a dendrogram from a hierarchical clustering dendrogram. Whatever. You get the same thing. You get two classes. And the patients are shown there-- about equally distributed.

And now, if we fill in the labels of the patients-- that is, whether their survivors-- turned out to be survivors or not survivors-- you get this picture, where-- you don't need a statistician to tell you that there's no correlation between this class structure and survival.

So what do you take? What do you think of this experiment?

AUDIENCE: What do you guys think of this?

TODD GOLUB: What can you conclude?

AUDIENCE: What do you do with this? How are you going to rescue this? Are you saying [INAUDIBLE] was [INAUDIBLE] signature that [INAUDIBLE] survival? What do you say?

AUDIENCE: There's a difference, but it may not be related to [INAUDIBLE].

AUDIENCE: But maybe there was [INAUDIBLE] would work somewhere else. [INAUDIBLE]

TODD GOLUB: Right. I think you're both hitting on the relevant point here, which is that the unsupervised learning clustering algorithm found some structure-- dominant structure that made these two classes. They just don't happen to have anything to do with the question that we're interested in, which was one of survival.

And so that gets back to this basic notion of two general approaches to data analysis, which you've probably discussed but I think often gets confused. So unsupervised learning-- which is not exactly synonymous with clustering, but it's a reasonable first approximation, or supervised learning classification.

So here, you're interested in finding a dominant structure defined only by the intrinsic gene expression patterns in a given data set, irrespective of anything you happen to know about the samples, such as their clinical outcome. Here, you're saying, whatever. Maybe there's some other interesting biological structure. But I'm not interested in that right now. I want to know whether there's a gene expression pattern that's correlated with the thing I care about in this particular example outcome.

So we take the same data set-- the same matrix of data samples by gene expression values and now apply supervised learning approaches. This happens to be a K nearest neighbor classifier. Again, it doesn't make any difference what you use here. And this happens to be an eight-gene model.

You classify the samples using a leave-one-out cross-validation approach so that you don't-- you attempt not to overfit the data-- the model to the initial training set. And then you ask, well, , of the two classes that are predicted how do those patients actually fare? And here is a survival plot in terms of months-- months of survival for those patients who are predicted to be alive versus those that are predicted to be dead. So what do you think about this?

AUDIENCE: This? So we'll pick those genes?

TODD GOLUB: Is this significant? So if you look in a basic biostatistics textbook about how to calculate statistical significance of a Kaplan-Meier survival curve such as this is, they'll tell you to do the log rank test. And if you did that, you would get a p-value that, if you've looked at a lot of these things, would match your intuition for this degree of separation. So it would look something like that.

Is that reasonable? Well, that looks quite significant. But you should ask, well, how come this model is eight genes, for example? How did you choose that number?

Well, it's quite easy. We chose that because I worked the best. It worked better than six, worked better than 10, or 50. So we had to pay some penalty for overfitting a model, potentially, to this particular data set. So the ways that you could then really test the statistical significance of this model would be to apply it to another data set. That's the gold standard.

But short of that, a reasonable thing to do to better approximate the significance is to take into account the fact that there are a number of parameters of this model that were optimized to fit this data set. And you can do this by doing a permutation test, where you don't scramble the gene expression values themselves, but you randomize the class labels in terms of, are the patients alive or dead, and you go through the same procedure of attempting to build an optimal classifier and including choosing the number of-- optimal number of genes to ask, if you really try hard with these machine learning methods, how often can you make a classifier that works as well as this one does?

And when we did that 1,000 times, 9 times out of 1,000, we could do this well or better. So we estimated the significance of this model here, which is still decent. But you can see we took a hit of a couple of orders of magnitude on this p-value. So if you had a nominal p-value of 0.05 or something, that result would entirely vanish when you appropriately attempt to correct for such multiple hypothesis testing. And this is independent of what particular classifier you used. So I would say, much of the literature-- and everyone's figuring on how to do this as we go along. But much of the literature and worries about failure to reproduce an initial model are due to the problem of overestimating the significance of an initial model because of these overfitting types of problems.

So let me just make a couple of general comments about supervised learning. And some of them may seem obvious, but turn out to be actually problematic. And this is one example. So the first step-- establish the class labels of what you're trying to classify.

So in one of our first experiments where we were trying to classify the two different basic types of acute leukemia-- acute lymphoblastic leukemia or acute myeloid leukemia-- the way you build a classifier is to choose examples of one class, examples of another class, and then find gene expression patterns that are correlated with that.

Well, who's to say that we have these right, right? What should you use as the gold standard for these things? It's not always obvious, particularly for the very things that we want to build better molecular classifiers for because the current clinical diagnostics are so poor. It doesn't really feel like a good gold standard to go back, rely on the clinical labels as the gold standard. And so this is, in general, a.

Real problem for survival studies, if you force the question into a simple two-class problem-- survivor or non-survivor-- well everyone's a non survivor at some point. So at what point do you declare that a patient is a survivor of their tumor, for example? That requires some judgment call in terms of which bin to put the samples into. And I would say that much of at least our effort and time has gone into trying to figure out how to get this right. And there are some approaches that one could take to not have to be so rigid on how you assign these labels, but it's a challenge.

So the second general step in making the classifier is selecting the features that you're going to feed into a model-- features in this case being genes. The details don't matter-- their whole long list of ways that you can rank genes. This is a simple one that is based on the mean expression level in the two classes and their standard deviation. It's a relatively unsophisticated way to select genes because it assumes that there's a uniform behavior of these marker genes in the two classes, which, in many cases, is not at all the case. There are many other methods as well.

And then I'm sure you talked about these at length. So I think maybe I'll skip this, that you then take these things and classify. So for unsupervised learning, there are all these methods. And I think they basically don't matter-- that, either for clustering or supervised machine learning methods, if you get a result that is obtainable with only one magical-- one person's really special algorithm, I would worry deeply that there is a problem with-- there's an information leak somehow or something's not right because, at least in my experience, when there's really biologically or clinically meaningful structure to be found, you can find it with a number of different approaches. And in fact, that's a reasonable sanity check to make sure that you can recover structure, whether you're using machine learning or unsupervised clustering algorithms-- that you can recover it with multiple different methods.

There are some examples to that that can be interesting. But on the whole, I'm confident saying it doesn't really matter, and it's really the input to these data sets that matters the most-- that is, are you really sampling the diversity of, for example, the disease process that you're studying with the samples that are in your initial data set?

What's more challenging is, how do you evaluate the output of these clusters in terms of their biology, in terms of really knowing how robust the structure is, given any algorithm? And then how do you actually know when you-- having seen a given structure once, how do you actually apply it to another data set to know whether you see it there as well? It's not obvious.

Any questions about that general stuff? I wasn't going to say any more because I know you've covered it. Let me give a couple other examples of applying these principles to some data sets. Did you talk about this one?

AUDIENCE: No.

TODD GOLUB: OK, so here-- again, focusing on childhood leukemia-- most kids with childhood ALL respond to chemotherapy. I told you about this subgroup of BCR-ABL patients that does not. Another group that does not respond well are infants less than a year of age, who generally don't respond. It turns out that most of those patients have translocations into a gene called MLL.

But it's clinically of interest because these patients don't respond to conventional chemotherapy, and this just shows you that, using standard clinical criteria, they're hard to distinguish. So what if we take conventional ALL samples, these infant MLL rearranged leukemias, and some AML-- the myeloid leukemias and we apply an unsupervised learning approach? This happens to be principal component analysis, but could be your favorite clustering algorithm. What do you see here?

AUDIENCE: What do you see?

AUDIENCE: Three different classes.

TODD GOLUB: Three different classes. Why do you say that?

AUDIENCE: Jose, speak up.

AUDIENCE: I mean, you're maximizing the separation so you see some separation-- some MLL [INAUDIBLE].

TODD GOLUB: OK, so right here. So, yes, you see three classes, but only if you have the colors filled in. So if you imagine this is just a group of leukemias, you might get the sense that there was something going on over here. But if you imagine these are all black, it's not so obvious, maybe.

AUDIENCE: [INAUDIBLE] analyzed or have been optimized [INAUDIBLE]?

TODD GOLUB: No, this is completely unsupervised. So that's the first point, is that these things often look clearer when you actually impose knowledge on them. Even though the structure here is done in an unsupervised way, you get the impression that it's really clean result if you superimpose knowledge afterwards. That's the first thing.

But let's say, yes, they are three classes. And I think you can appreciate that. One question was, well, maybe these infants with the MLL-- rearranged genomes in green-- maybe they don't respond to therapy because they're babies and you know that this is a metabolic post-metabolism problem. And their leukemias are the same as the conventional ALLs shown in dark blue. This would argue that it's actually not the case, that they're fundamentally different leukemia. Is this helpful?

So it's helpful, maybe, from a taxonomy perspective, but does it tell you what to do for these patients? So what if you wanted to gain some biological insight into what was different about these MLL green infant leukemias? What might you do? You've got these data. You see that those patients define a different class. What could you do?

AUDIENCE: Any of you have any idea? What would you do with this?

AUDIENCE: Inspect the component? Look for the things that have higher weight? Try to define biological function related to the [INAUDIBLE]?

AUDIENCE: So the genes whose weights explain the most is separation.

TODD GOLUB: That's right. You could do that. As it turns out, in this case, there are a lot of genes that actually have relatively equal weight. So you still have a large list.

And the three principal components don't perfectly separate the classes. So you could go back and say, well now I believe that these MLL leukemias are a distinct entity. That would be reasonable. That's a reasonable thing to do.

But the other way that you could do is to say, well, now this tells me that I believe that these MLL leukemias are a distinct entity. Now let's use supervised learning types of methods to identify the genes that are most correlated with the class of interest-- for example, high in the MLL class versus the others. That would be a straightforward thing to do. So you could rank the genes according to that distinction.

So we did that and did what I think-- yeah.

AUDIENCE: [INAUDIBLE] In the first case, [INAUDIBLE] compare between the different classifications-- the difference-- the different genes-- [INAUDIBLE] expression or whatever? And the second case, you're just thinking comparing the one classification? Is that correct?

AUDIENCE: No.

TODD GOLUB: No, I think it's more that, if you didn't have these colors to look at and you said, ah, there's some structure here. I don't know what it is. What's the biological basis of this structure? Looking at the weights of the genes that are driving this distinction would be a reasonable thing to do.

In this case, we had a specific question. Are these leukemias unique, or are they admixed with the others? Having determined that they are unique, it's a little bit cleaner to say, all right, let's use supervised methods to find the genes that distinguish one class from the other. Of course, if you had perfect separation, it would reduce to the same experiment. But because it's imperfect, there are some advantages to using class labels here.

I should mention also that-- you see this blue guy here sitting in a sea of green? So this is a patient that, based on gene expression, one would predict to be MLL rearranged. But the clinical record for this patient study was not. But when we went back and actually looked at this, it turns out that there was a missed translocation into the MLL gene that you could recover by FISH.

So this is not a public health menace-- diagnosing these leukemias properly. But there are examples of missed diagnoses that I think can be-- I think looking at these multi-parameter gene expression readouts can serve as a unifier, an integrator of lots of upstream genetic activity. And so I think the power to detect those upstream events is going to be higher when you look at some downstream pattern, such as an RNA pattern, as opposed to developing specific tests for each of the individual genetic abnormalities that could cause the same phenotype because, in the end, all you care about is knowing whether the molecular program has been activated.

So you rank the genes according to this distinction and just start at the top of the list-- here is a gene that was top of the list of 12,600 or whatever that we're on on the list. And any time a tyrosine kinase rears its head in a cancer classification-- cancer biology experiment, you pay attention to it, particularly given the Gleevec story.

So what do you think about this? I tell you, oh, look at this. The RNA level of a kinase-- so a receptor tyrosine kinase called FLT3 is characteristically high in the MLLs, compared to the others. What do you think of that, in terms of therapeutic-- potential therapeutic significance?

AUDIENCE: What would you do with that?

AUDIENCE: [INAUDIBLE]. Or maybe this already popped out, but can't you check the levels in the other two classifications?

AUDIENCE: Because these [INAUDIBLE] slower.

TODD GOLUB: Right. So we define this list by virtue of the fact that it's high in the MLLs compared to the other two combined.

AUDIENCE: So how do we get the therapeutic K out of this?

AUDIENCE: For the patients patient are responsive [INAUDIBLE]

Are they classified as far as the response, the FLT3 inhibitor?

TODD GOLUB: To a FLT3 inhibitor.

AUDIENCE: Yeah.

TODD GOLUB: So you want to treat patients with a FLT3 inhibitor? Well, that's not an FDA-approved drug, so you can't do that.

AUDIENCE: OK, so what else do we have?

AUDIENCE: [INAUDIBLE]

TODD GOLUB: So there is. The hypothesis would be that MLL leukemia cells are dependent on FLT3 kinase activity for survival. If that's not the case, then you don't care. Unless that's the case, the overexpression of this thing is totally irrelevant from a therapeutic perspective.

So you could do that genetically-- for example, using RNA interference to knock down the expression, or you could do it pharmacologically, if there was a compound in development-- not yet a drug-- but a compound in development that inhibits kinase activity. And so that's what this experiment is.

AUDIENCE: So [INAUDIBLE]

Doing RNA interference-- is that something that you can just do in a person?

TODD GOLUB: You can't do it in vivo in a person, but you can do it in human-derived cell lines.

AUDIENCE: So clinically, that would not be [INAUDIBLE].

AUDIENCE: You could test the hypothesis that you want to go down that path.

AUDIENCE: I see.

TODD GOLUB: That's right. Now, you can make the argument that, well, doing these things in cell lines in mice-- that's not real disease. And so I don't care what any of this stuff shows. But still, if your hypothesis is that a given gene-- the overexpression of a given gene is important and you do the experiments to ablate the expression of that gene and nothing happens, that should deflate your enthusiasm a little bit.

So here's the experiment, though. Here, now, taking patient-derived human cells that have been engineered to express firefly luciferase genes so that they glow-- and you can monitor in vivo tumor burden. So here, mice that, on week one, you inject in the tail vein infant leukemia-derived tumor cells. And you see, over 4 weeks time, the amount of luciferase activity increases as the cells grow and the mice start to die around week 4.

And here is a cohort of mice also injected but treated with a drug once a day by mouth that functions as a FLT3 kinase inhibitor. And you can see that the development of the leukemia is significantly abrogated, which, at least to our first approximation, validates the hypothesis that FLT3 overexpression is not just a diagnostic marker of this class, but it's actually a potential therapeutic target.

And so, based on this and some other preclinical data, the clinical trial that you wanted to do with a FLT3 inhibitor is being planned to treat patients. Yeah.

AUDIENCE: [INAUDIBLE]

TODD GOLUB: So the cells are infected with a retrovirus that contains the cDNA for the firefly luciferase gene so that, if you inject these mice with the compound luciferin, they will emit the same enzyme that fireflies do, and they will glow. So usually, this is done in vitro in test tubes. But here, you introduce it into the cells and the animal so that you can monitor.

What you used to have to do would be to inject a whole bunch of mice, kill some of them here, kill some of them here, and then examine the bone marrows to evaluate the progression of the disease. What's nice here is that you can follow a cohort of mice non-invasively.

AUDIENCE: Let me ask a dumb question, because I've never actually done this. When you actually look at these mice, can you tell that they're fluorescent?

TODD GOLUB: No. No.

AUDIENCE: They don't actually look-- no.

TODD GOLUB: No, you need to use a special device that can measure, I think, in the near-infrared range. There are green fluorescent protein mice that actually do glow. And you can tell that they're green.

AUDIENCE: The mice that we use-- are they immune [INAUDIBLE], which means you don't have a massive immune response?

TODD GOLUB: So you have to do this in immunodeficient mice so that they don't reject the human tumors.

AUDIENCE: Does that factor at all in your determination of the degree of proposed-- the spread of the tumor cells and whatnot because they're just there? The immune system can direct attacks against-- so when you're considering these experiments and saying, OK, I see this spread across the entire mouse and this level of [INAUDIBLE] how do you factor that in?

TODD GOLUB: You don't. You factor that in by saying that there are many things that are occurring in these models that don't recapitulate what happens in the mouse. Most people don't get cancer by having intravenous injection of a tumor into them. Most patients have an immune system.

So I think it's just one of the limitations that-- it would not be worth the time and expense to have a drug development project around every little inkling that comes out of a microarray experiment. So you need to do something, even though it's deficient in many ways-- and you've hit on some of them-- to say, is this interesting or is it not? I think-- not yet at the point where one can do this entirely computationally and have any kind of confidence.

That being said, these so-called xenograft models, where you put a human tumor into a mouse model, are not particularly predictive of efficacy of a drug in the human clinical trial. But in the absence of anything better, it's still what most people do first.

AUDIENCE: [INAUDIBLE] more robust, in other words, if there's no effect on the xenografts, then you're really a loser if you go to the human [INAUDIBLE]?

TODD GOLUB: No. If anything, it's the opposite, that, if you show some activity in the xenografts, you often see activity. But failure to see activity-- failure to see activity in xenografts is not particularly-- particularly for molecularly targeted therapies, where it may be that you can show in the mouse that you've really shut down the pathway. Let's say you've inhibited FLT3 completely.

Drug companies are starting to use graphs in that way to say, all the mouse is is a test tube so that I can ask-- have I inhibited FLT3 enzymatic activity? Yes or no. If I have and I believe that FLT3 is a good target, I don't care whether the tumor's actually shrunk or not. I'm going to bring it forward to clinical trial. But you need something to convince you that the target is reasonable, yeah.

AUDIENCE: An interesting [INAUDIBLE]. So the small spots in the [INAUDIBLE] three mouses are-- the purple ones are the xenographs, right? The purple spots are the xenographs.

AUDIENCE: Are the tumors.

TODD GOLUB: Yeah.

AUDIENCE: But that means the spots are all very fixed in specific areas of the mouse.

TODD GOLUB: Yeah, I think you see them there because the cells-- they're injected intravenously in the tail vein, but they home to the bone marrow and you're seeing large bone marrow cavities, which is why you see them over the flank there. I think I'm going to skip this.

OK, did you talk about this?

AUDIENCE: No.

TODD GOLUB: Good. So when you do these experiments, the data usually present you with two-- you either have one or two problems after you do all this appropriate correction for multiple hypothesis testing that I told you about. Despite having done that, you still have a list of genes that's too impossibly long to bring biological understanding to, or you've corrected away everything and you have the impression that there's actually nothing that is differentially expressed in your two cases.

And so let me show some of the more recent approaches to dealing with this because it's really substantially changing our thinking about how to do these kinds of experiments. So this is not a cancer example but a diabetes experiment where there were patients who were either-- adult patients who either had type 2 diabetes or they were normal, as defined by having a normal glucose tolerance test. And they underwent voluntary skeletal muscle biopsies under a euglycemic clamp.

18 of these patients, 17 of these patients-- it's a simple two-class problem. Get the expression data to identify those genes that are differentially expressed in these two classes. Do the appropriate permutation testing to make sure that you correct for multiple hypothesis testing. And here's the result. Nothing meets significance.

So out of the 20,000 genes on the array, even the top-ranked gene doesn't meet statistical significance. Possible that this is the case. But the question is, are there other ways that you might go about recovering a biological story here? And so the way that Vamsi Mootha and graduate student, Arvind Subramanian took to this was to define groups of genes or gene sets whose activity as a collection of genes could be interrogated in these data sets.

And we could have a rich discussion about, how does one define such gene sets? You could do it based on the literature. So ask Zach what genes are important in some pathway that he knows something about. That could be a list.

Or you could say, we don't trust that about-- that'll bring Zach's bias. We're not interested in that or anyone else's intuition. Let's just experimentally derive lists of genes by one way or the other-- perturb cells, get the gene expression change, and that makes a gene set. And you can collect as many of these as you could stand. For these experiments, we made 150 of these gene sets.

AUDIENCE: Before we go on, [INAUDIBLE] twice on the slide.

TODD GOLUB: Oh, yeah.

AUDIENCE: So he's enriched one.

TODD GOLUB: He's enriched. Yeah, he should be. So then how do you do this?

So first thing you do is rank all the genes on the array-- 1 through 20,000-- according to how well they're correlated with the distinction. I already told you this top one-- even that one doesn't meet significance as a single gene. And then you interrogate each of these gene sets and ask, are they enriched?

And so here would be an example of a hypothetical gene set, so each gene in the gene set of a dozen genes or whatever, that is not enriched towards the top of this rank-ordered set list, whereas here is a hypothetical gene set. It's not perfect. But it's non-randomly distributed on this rank order list. It's enriched towards the top.

AUDIENCE: Do you see this as a similar operation to the following-- there's a bunch of proteins that they would look at [INAUDIBLE] giving microarray result at gene ontology, and they'd say, what classes of gene ontology are overrepresented given-- in this set of genes?

TODD GOLUB: Yes. So you can define these gene sets based on a gene ontology annotation. That's an example. The important part is to make sure that you appropriately correct for testing all the gene sets. So now, instead of 22,000 genes, we have 150 gene sets. But that you should think of 150 hypotheses. So you should do the same permutation type of testing and say, if I randomize-- in this case-- the diabetes versus normal distinction, is my favorite gene ontology class still enriched? And that's what some of the current approaches to that kind of annotation don't do.

And so you can codify this in something called a Kolmogorov-Smirnov statistic. It doesn't matter. You can come up with an enrichment score for these things. And if you do that in this example you essentially get 1 gene set which meets--

AUDIENCE: [INAUDIBLE]

TODD GOLUB: Which gets quite high statistical significance for a set of genes. So how do you reconcile this? How do you reconcile this thing and this thing? How could that be?

AUDIENCE: That's my point. I want to understand what you just were trying to explain. So the first thing-- you're saying that you didn't pick up any difference in the expression--

AUDIENCE: On a gene-by-gene basis.

TODD GOLUB: That's right.

AUDIENCE: But then, when you group a couple of them together, all of a sudden, there is a difference.

TODD GOLUB: Right. So how could that be?

AUDIENCE: How could that be?

[INTERPOSING VOICES]

AUDIENCE: --getting more information out of that. Maybe a weak signal [INAUDIBLE] sample [INAUDIBLE] the coherence.

AUDIENCE: [INAUDIBLE] a couple things things being combined [INAUDIBLE]. Make it one up and one down. [INAUDIBLE].

TODD GOLUB: So the microarrays themselves-- the precision of these arrays is not so fantastic. And so you can imagine if there's a subtle signal. On a gene by gene basis, it's difficult to detect it. But if you consider the coordinate regulation of a group of genes all in the same direction, as a group, this might be quite striking. And this is shown right here, which is really quite amazing when you think about it.

So here, look at the mean expression level of all the diabetic patients versus all the normal patients. All the genes on the array are shown in gray. And so you would expect there are no outliers. There's nothing really way off the diagonal. If they were, those would show up as single genes that were differentially expressed. Of course, you could have one massive outlier that could screw you up with looking at the means. But still, you get my point.

And here are these oxidative phosphorylation-- the gene set that was defined by those genes that are involved in oxidative phosphorylation. And you can see that, with only a few exceptions, they're all lined up just below the diagonal. Their change in gene expression is only about 20% compared to normal, but it's all in the same direction. So 20% change in this number of genes is quite significant.

AUDIENCE: Are you getting that? Let me try to-- because it's an incredibly important point. The chance-- if you look at any given dot of thing, what does it means to be 1-- near the diagonal, one side or the other. I'm not going to make [INAUDIBLE] story. [INAUDIBLE] about. It's going to be a quick one-- the fact that it's on one side or the other.

And then, just by dumb luck, they're all on one side of the diagonal if they're put diagonal. That's going to be incredibly unlikely. And so each individual gene is one side of the diagonal. The fact that all genes that we have pre-assigned beforehand of the other type-- in this case, [INAUDIBLE] phosphorylation-- that will end up on one side of diagonal-- that's hugely unlikely, the fact that you can just, by some luck, put them all on one side of the diagonal.

AUDIENCE: So it's less tied to the enrichment [INAUDIBLE] the probability that you define [INAUDIBLE] that, given that you've said these are genes that should be related to some [INAUDIBLE].

AUDIENCE: Yes.

AUDIENCE: That makes sense.

TODD GOLUB: That's right.

AUDIENCE: [INAUDIBLE] together, the classes can be formed.

TODD GOLUB: That's right. Because if you look-- if you take this point in isolation, there's no way that that's going to be significant because it's right in the middle-- in this thing. So this is an eye-opening experiment, and it's causing us to go back and reanalyze, using this methodology called gene set enrichment analysis, some old data sets.

Let me give you a couple of other examples of unpublished examples that in a slightly different-- use it in a slightly different way. So I told you about our medulloblastoma outcome prediction experiments before. And around the same time, there was a paper published that looked at the same question, essentially. Non-metastatic versus metastatic medulloblastoma-- different patients, different arrays, different groups, whatever.

They made a classifier that was centered around the PDGF receptor alpha gene-- was a predictor and also a number of the downstream players of PDGF receptor alpha. And so we asked, are any o-- when we look at our classifier of outcome, which I showed you is pretty decent, where's the PDGF receptor alpha pathway on there? And neither PDGF receptor alpha or the genes in that pathway were among the top predictors-- top 50 genes in our data set, one which I think would lead 1 to believe that one or both of those data sets is wrong or the models derived from them are wrong.

But if you take this PDGF receptor alpha-related genes as a gene set and ask, is it enriched in our data set using this methodology shown schematically here, it's enriched. This list is 12,000 genes long or so. So you can see they're not all stacked up, like 1 through 50, but they're non-randomly distributed, which we take to mean that, actually, the two data sets are consistent. If you had data sets of infinite size, then you'd start to see convergence of the markers being overlapping at the very top of the list. But with these smaller data sets and the clinical variability--

AUDIENCE: This puts some formalism around what [INAUDIBLE] was waving his hands and yelling about. For five years, people said, well, this microarray [INAUDIBLE]. You can't tell the difference between them.

TODD GOLUB: Yeah.

AUDIENCE: This is the pattern. It's the overall pattern. This is a much more formal way-- a pattern that's released to pick [INAUDIBLE].

TODD GOLUB: So here's another example of that thing. It's also unpublished. So we looked at lung cancer-- human adenocarcinoma in the lung and identify some-- we just drew the line at 50 because it's a nice number-- predictors of outcome in the Boston lung cancer patients. University of Michigan did the same experiment, published around the same time. Overlap in gene of these two-- listed 50 genes-- zero-- concerning.

But if you look in the space of gene set space and you ask, what gene sets are enriched in one data set? What gene sets-- which you can think of loosely as pathways. They're not really pathways, but it's reasonable think of them for this purpose. There's really quite significant overlap in gene set space.

So I think what this is saying is that Botstein is right, that there is more biologic coherence in these data sets. It's just we haven't been smart enough to really know how to see it.

AUDIENCE: How did you choose your gene set? [INAUDIBLE] What do you use-- [INAUDIBLE] function or is it a pathway? Or how do you actually [INAUDIBLE]?

TODD GOLUB: We now have about 450 or so such gene sets-- some of which are good, some of which aren't particularly useful. They include some go annotation. I don't think those are particularly useful because the granularity isn't fine enough.

I think, in the end, the most useful types of gene sets are going to be those that are experimentally derived. But this is a mixture of those, and we're not yet at the point where we even started try to understand-- of these 35 enriched sets of genes, what are they and what's the biological story?

AUDIENCE: You threw at it on the order of 60 gene sets?

TODD GOLUB: No.

AUDIENCE: No?

TODD GOLUB: No. We threw at it 400-and-something gene sets and asked, how many of those are enriched in the Boston data set? And the answer is 35 plus 18. And 35 plus 12 were enriched here. And so the majority of the sets enriched in one were also enriched in the other.

AUDIENCE: That's helpful.

AUDIENCE: Are those cancer-specific or just biological?

TODD GOLUB: No, they're not cancer specific.

AUDIENCE: And those gene sets are manually annotated by their group, or is the [INAUDIBLE]?

TODD GOLUB: There a combination, as I said. Some are these [INAUDIBLE] pathways that are so-so annotation. Some are entirely computationally derived. That is, they're the nearest neighbor genes of a given index gene in a data set. They are various things. And what the definitive collection of gene sets would actually look like isn't obvious to me.

AUDIENCE: [INAUDIBLE] learn more and more about mechanism.

TODD GOLUB: That's right. On the one hand, I think it will be useful to just not fret about it too much and worry about exactly how to define these things. Just get them in there. The nice thing about this GSCA methodology that I didn't really go through in detail is that it's forgiving-- how you calculate these enrichment scores is forgiving of the definition of the gene sets because you're looking for enriched-- non-random enrichment of the gene. Set so the fact that a third or a 1/2 of the gene set may actually be inappropriately there doesn't make any difference because there's still enough that significantly enriched to detect it.

AUDIENCE: You also [INAUDIBLE]?

TODD GOLUB: You can. We didn't happen to do it here, but you can. Again, like any of these other things, there are going to be a number of different metrics that you could apply to measure significant enrichment. The most important thing is just to make sure that you correct for the possibility of whatever metric you use-- that you're detecting something beyond what you'd expect by chance.

AUDIENCE: So I have two questions about [INAUDIBLE]. One would be-- so I'm assuming that you can also detect-- you have a [INAUDIBLE] of a particular gene, which you haven't mentioned so far. So is that [INAUDIBLE] an example of an actual [INAUDIBLE]. Is that something you would find by going back to your healthy samples and comparing-- looking for enrichment relative to your disease?

AUDIENCE: A positive score?

TODD GOLUB: So actually, the way you calculate this is the metric doesn't specifically look for enrichment towards the top. It looks for a non-random distribution. You would find depletion.

What you could also find-- which I think is-- this score will capture and is not desirable-- would be something that's concentrated in the middle, which is very uninteresting. So there are some false positives in there.

AUDIENCE: That may not actually be so interesting-- for example, in T-cell activation, you get normal [INAUDIBLE]. You get upregulation of certain proteins and you get downregulation of others. And so what I haven't heard yet is how you account for, perhaps, enrichment of part of your genes [INAUDIBLE] increasing another with that--

TODD GOLUB: So there is another version of this that tries to dissect the gene sets into those components that move coherently in one direction versus the other, because you're absolutely right.

AUDIENCE: [INAUDIBLE]

TODD GOLUB: Yeah. If, for example, you take a-- use GO annotation or something like that, or some pathway, if half the genes in the pathway go up and half the genes go down, that could look like no enrichment at all, whereas, if you separate those somehow, you could see it. How are we doing for time?

AUDIENCE: Well, you've got 20 minutes-- 18 minutes.

TODD GOLUB: OK. So let me push this to-- not classification but some newer directions that we're thinking about-- how can you use these signatures for useful things, particularly to think about something that's closer to drug discovery. So this is the way the usual discovery pipeline would look like. You have some disease process or biological process you care about. You do some microarray experiments.

And then a miracle's is supposed to occur whereby you develop sufficient molecular understanding of what the data are telling you, that you can identify the smoking gun target. And then you partner with a drug company and say, screen for a small molecule that inhibits this critical therapeutic target. The problem is that this part is really tough.

And so what we've been thinking about is, well, could you bypass the understanding part, at least initially, whereby you screen for small molecules based on their ability simply to perturb a signature of interest. And then, once you have those in hand, you could use them to further dissect the biology or, if you're lucky, think about them like drugs.

And so the proof-of-concept experiment is shown here, where-- here are two biological states, for example-- a leukemia cell, which is undifferentiated, and a normal blood cell, which is fully mature and is differentiated along the myeloid pathway-- a peripheral blood neutrophil. It's not known what the critical targets are of this pathway. So it's hard to do a small molecule screen to induce this process, which would be nice, if you could simply induce your leukemia cells to turn into normal cells.

So the question is, could we define a signature of this state, a signature of this state, and then screen for compounds that trigger the signature? So the details don't matter here, but the concept is, define signatures of the two states of interest. So we call this thing GEHTS, for gene expression based high throughput screening. Define the signatures-- now standards-- what we've been talking about-- our microarrays, where the experiment would be-- treat cells with various different chemical compounds and ask whether any of those compounds trigger the signature of interest. And to make this feasible, we simplify these complex signatures into a handful of genes that you can measure by multiplexed PCR.

AUDIENCE: I read the paper. Was just purely a cost issue as opposed to going directly to the microarrays?

TODD GOLUB: It's a cost and throughput issue. Yeah, so if you wanted to--

AUDIENCE: We're good. [INAUDIBLE]

TODD GOLUB: If you wanted to screen tens of thousands of compounds, not really feasible if it costs you $500 a pop in the throughput issue. So yeah, it's a practical matter here. This part doesn't matter. Suffice it to say there's a method for how to measure a simplified signature in high throughput.

So we screened a couple thousand compounds and asked, do any of them trigger this little mini gene signature? And some of them did. Details don't matter.

But then the question should be, well, maybe these things just trigger these five genes that are-- I'm sorry, these compounds trigger the five genes but actually don't do anything. So one way that you could sort out whether they actually do anything biologically is to now step back and look across the whole genome again, take cells, treat them with these candidate compounds, and ask, did you actually recapitulate the overall molecular program, not just of these five genes but of the whole thing-- of the whole molecular program?

So if you turn back to genome-wide arrays, you can see that a number of these compounds recapitulated the molecular program of differentiation. Does that make sense? So you use the simplified high throughput assay just as a readout of whether you've triggered the signature or not. And then, with those candidates in hand, you go back and interrogate them.

AUDIENCE: Does everybody follow that?

TODD GOLUB: So the genes in this little signature itself--

AUDIENCE: We don't if they're actually doing anything to the rest of the cell.

TODD GOLUB: I'm pretty confident they don't, actually. But it's irrelevant. So you define these signatures not based on their being important or somebody thinks, oh, that's a good target, and that's important for leukemia or differentiation or whatever. It simply--

AUDIENCE: Represents the class.

TODD GOLUB: It represents the class. All you care about. And that you can measure it well. So sometimes you find a good candidate marker that, for some reason, it doesn't behave nicely in this assay so you chuck it out and replace it with something else.

So here's just an example to say that then, as you might expect, when you treat leukemia cells with these candidate compounds, discovered solely based on their gene expression changes, they do the thing-- the cells do the things that maturing leukemia cells should do, like-- they become phagocytic. They start engulfing--

AUDIENCE: [INAUDIBLE] much more differentiated like in their behavior.

TODD GOLUB: Yeah, so it's a promising idea. This is one unpublished example that says, well, it is known that blood cell differentiation is largely governed at the transcriptional level. So maybe that's why you can define these transcriptional signatures of the differentiation process and screen for things.

But here's an example of defining a gene expression signature. Again, the signature itself is devoid of any real biological meaning, other than it reads out, in this case, activation of a signal transduction pathway. So the usual way of thinking about this is, well, if you want to look at signal transduction-- proteins talk to each other, RNA has no place in that. RNA profiling has no place in that.

But here the idea is, if we stimulate a signaling pathway-- in this case, by stimulating cells with platelet-derived growth factor and then capture the transcriptional response at the RNA level, could we use an RNA signature as a readout for PDGF receptor activation and a screen for inhibitors of the signal transduction pathway using RNA as a readout? And here, this is work of a MIT graduate student-- chemistry graduate student, who pulled out a compound called aurintricarboxylic acid as an example, which turns out to be a previously unknown inhibitor of the PDGF receptor itself. But it was discovered by looking down here at a signature.

So I think this is going to be useful for various-- being able to screen for things that you can't otherwise screen for. So this is a chemical structure that is not currently being explored in people who know much more than I do about kinase inhibitors because no one thought to look at it. And it was discovered simply by using as a signature, as a readout.

So the last five minutes, let me push the signature idea perhaps further than I should.

AUDIENCE: But it's the end of the class.

TODD GOLUB: But it's the end of the class, and I'm excited about the idea. But it has less data surrounding it. And that's the idea of using these signatures-- so you can see, at least in-- my thinking has shifted over the past year or so, much away from finding, oh, what's the needle-- using these microarray types of experiments to find a needle in a haystack-- what's the gene that's responsible for something I care about? To thinking about the power of these signatures as readouts for various things.

So the idea here is to use the signatures-- RNA signatures as a vehicle for establishing connectivity between components of the genome and each other, that you manipulate through perturbation. It's a little bit different than relevance networks, I think, but it's conceptually similar.

AUDIENCE: [INAUDIBLE]

TODD GOLUB: Establishing connections between drugs and drugs and drugs and genes. So the idea would be that, if you can define a gene expression signature of, let's say-- comprehensively of all drugs-- there are only about 2,000 FDA-approved drugs. It's actually amazing that that's just not publicly available information, what happens to cells when you treat them with drugs we give to patients.

If you had that and, for example, you had a signature that was a result of ablating each gene in the genome sequentially-- only 26,000 genes in the genome, and there are now reagents coming online with RNA interference where you can actually do that experiment-- then you'd have a systematic matrix of perturbations whereby you use these gene expression signatures as the universal bio assay to connect genes with genes, genes with drugs, and drugs with drugs. Does that make sense? Do you have a question?

AUDIENCE: [INAUDIBLE]

How do we prevent from [INAUDIBLE] to the [INAUDIBLE] cancer cells [INAUDIBLE] cancer cells. But because of [INAUDIBLE]-- because of [INAUDIBLE].

TODD GOLUB: Well, we're arguing that-- don't bother doing this in lower organisms because you can do the experiment in human cells. So we're going to do it in human cells. And I would say-- the idea, again, for establishing these connections is not to be able to--

AUDIENCE: Be perfect.

TODD GOLUB: I think people who think that you can create a wiring diagram and reverse engineer a cell are out of their minds based on these data. It's just not feasible. So the question is, can you find-- can you find, for example, enrichment in the same GSCA kind of thinking? Can you find enrichment of one signature in another, thereby establishing connectivity, in which case yes, you'll get some of it wrong because the context isn't right or the species mapping isn't right. But will you be able to see enough connections to establish connectivity?

AUDIENCE: [INAUDIBLE]

TODD GOLUB: Yep.

AUDIENCE: Sure.

TODD GOLUB: Yep, certainly some. We're finding, actually-- I'll give you a couple of examples that-- while that is certain to be-- that context-dependence is certain to be the case--

AUDIENCE: There's lots that's shared.

TODD GOLUB: There's lots that shared. And so, when we first started talking about this project, there were a lot of objections to the idea-- oh, still are, that, oh, what if you don't choose the right cell line to do this in? Ideally, you would do it in like 100 different cell types. But then it gets then it gets to be a serious experiment. Even in one cell line, it's a huge experiment.

AUDIENCE: I think I told this group that I've seen the same thing in [INAUDIBLE] relationships. Many, many cell types that people wouldn't think they would be there. They're just there.

TODD GOLUB: Yeah, exactly. So let me give a couple examples and then we'll end. And it also-- again, comes back to this idea of enrichment, looking for enrichment using a Kolmogorov-Smirnov enrichment test. So here is an experiment that was published not by us, but by a group at Abbott Laboratories, where they are interested in this class of drugs called histone deacetylase inhibitors.

They took-- what is it? Five of these things, treated cells, took a common set of genes that were regulated, defined a signature of 22 genes. It's a gene set. It's the HDAC inhibitor gene set.

We took breast cancer cells-- not the cell type they used-- and treated them with a bunch of different drugs, including one, valproic acid, which is actually used to treat seizures, as it turns out. It was later discovered valproic acid actually has histone deacetylase inhibitory activity. It's weak, but it's there.

And we ask, can we see enrichment of this signature in any of those compounds? And the answer is, yes. So Trichostatin A was actually-- where is it? Here's the top one-- was one of the drugs that they used to define the signature. So we recovered it-- not surprisingly. But it's a little bit cheating because it wasn't a new example.

But here, you see sodium valproate, third and fourth ranked on the list, which was not used to define the signature but simply based on the signature connectivity. Had we not known it, we could have rediscovered that-- we could have discovered that sodium valproate was an extract inhibitor because of this connectivity. You can also see here-- Trichostatin A-- so the signature is defined-- I can't even remember-- in one cell type. And we see it triggering the signature-- the HDAC inhibitor signature in breast cancer cells and in leukemia cells. So it's robust [INAUDIBLE] context.

Interestingly, we also ran our little connectivity-- mini connectivity map across the oxidative phosphorylation signature that we define.

AUDIENCE: Does the [INAUDIBLE] protease cause hypoglycemia?

TODD GOLUB: It does. We didn't know that. We thought we discovered something new.

AUDIENCE: It's the wrong specialty.

TODD GOLUB: In the wrong specialty. Yeah, it was reported about 20 years ago that valproate causes hyperglycemia and modulates oxidative phosphorylation-- separable activities, separable from the anti seizure activity, separable from the HDAC inhibitor activity, and they trigger separate gene sets.

So one last example-- a signature defined of a drug called rapamycin, which sits in this pathway simplified here of PI3 kinase AKT, and a protein called mTOR. And it turns out that, if you define the signature of rapamycin treatment here, its defining T-cells published by David Sabatini, and apply that signature to our connectivity matrix, you see that rapamycin itself is recovered in two different cell types. But also, this thing-- LY294002-- we didn't know what that was initially. But then you look it up. It's actually an inhibitor of PI3 kinase, which is upstream of mTOR. So it puts, in the same pathway, two drugs that act together and trigger the same signature.

| again, I don't think it's going to be possible in the near term to actually reconstruct signal transduction pathways in their entirety. But to be able to put either genes or drugs together in a pathway not previously known to be in the same pathway, I think, is going to be possible with this approach. So we're committed to figuring out how to launch what would be a large scale public domain connectivity map project, where we would do these perturbations and put the data in the public domain so that people could use the data to find their own connections.

AUDIENCE: [INAUDIBLE]

There is a database that's recording some molecules [INAUDIBLE]. What will be the major difference between this project and that one?

TODD GOLUB: I'm not sure I know what--

AUDIENCE: I think she's talking about the [INAUDIBLE]. An expression, right?

AUDIENCE: Yeah, very expression when conditions on the molecules--

TODD GOLUB: On response to the molecules. This is actual response to the molecules. So that data set is the resting, untreated expression of the cell lines, which you can then correlate with how they respond. This is the acute changes of response, because the nice thing about that experiment is that you only have to measure each cell line once and then you can correlate it with all those things. Here, you actually have to do a microarray for each different--

AUDIENCE: Condition time point.

TODD GOLUB: Yep. OK, so the last slide. Whoops, let's skip that.

There's some future challenges. Just to remember that these prognostic signatures that everyone's developing are really a function of therapy. They're likely only to be useful if you continue to use the same therapy. So in this medulloblastoma example, we can predict things pretty well as long as we're trying to predict response to therapy that was given 10 years ago. But the therapy has evolved, and it's not certain that our classifier will still hold.

This is going to be challenging because the clinical trials are generally small and underpowered to do this kind of thing. Still not clear how, once these signatures are actually vetted and validated, what form will they take when they actually go to the clinic? Will they be a microarray? Maybe.

I've given you some ideas about how one might do signature-based chemical discovery. But turning a chemical into a drug is a big deal and not easy. And so for that reason, pharmaceutical companies aren't dropping their current approach to drug discovery in favor of this. And in general, how to integrate these kind of signatures into the drug development process is still something to think about.

But I still think this notion of using these tools-- whatever personalized medicine means-- but to gain more insight into the particulars of a given individual's disease to better match them with an existing or new therapeutic is likely to be here to stay even though there aren't a lot of examples of it happening yet. That's all I have.