**ISAAC SAMUEL KOHANE:** Also, I forgot to mention at this point, the output of microarray studies is foreign to basic biology researchers. They're used to looking at three or four or five or 20 numbers and performing some easy analysis in an Excel spreadsheet. But the point-- or doing a BLAST of one gene at a time.

But the data from these microarrays does not easily load into standard tools. The analysis is non-standard. Excel does not cut it. Excel can barely load some of these data sets. And so all this is actually representing an interesting turn. So this is the other point of view, which is now that the standard biologist tools are no longer fitting nicely onto the desktop productivity suite that they previously had. And this is, in some way, a full circle.

It used to be, of course, that computers were run by the high priests in the computer center. And not only that were they-- did they know about how to boot-up the computer and to make the computer do whatever it did, but they were also pretty steeped in the math and in the computational sciences, so that they could actually interpret the results of a study. With the advent of the personal computer and incredibly useful tools, like Excel, there was a lot that was now democratized and brought back to the desk of the individual biology researcher.

But then when we scaled back up to large data sets that are not trivially analyzable with current desktop tools, then the biologist's education is-- then the biologist's computational education is lacking. The biologists desktop tools are lacking. And again, they are not in the loop.

And let me tell you this-- and I'll get to more interesting details in a second-- but something that you have to understand. This is creating a lot of resentment among standard biologists around this whole genomic revolution. Because not only-- I mean, what you'll hear mostly is voice of skepticism about the methodologies.

But what's really going on is that they're justifiably feeling quite threatened for a big part of what they've been doing previously. Because they're just going to be disintermediated by the commoditization and the quantification and computational expertise that is required.

So what are the characteristics of a microarray? How do you recognize a microarray when you are being marketed one? The first is that it should have a small form factor. So one part of it's -- capable of measuring a significant fraction of some "-ome," the genome, the transcriptome, or the proteome.

So I think small form factor is an obvious one. But it also needs to measure a large part of the total output of the system. Otherwise, it's not a microarray in the sense that we commonly understand it.

There has to be minimal labor in data acquisition, so that it's an industrial process and not a labor-intensive process. You need to have an automated data path to a digital electronic format. And it has to be sustainable, high throughput processing.

And that begs the question then, is this a microarray? Well, this is a very nice piece of technology from Sciomics, which has, on these pillars, these antibodies that allows you to measure, in a very consistent and systematic fashion, 200 protein measurements or typically 30 proteins measured in sextuplets, in sextuple for redundancy.

So it's kind of a microarray. It's small. It's an array. But as typically used, it's measuring between 30 and 200 proteins, and so, therefore, I would argue it's not a microarray. And basically, they're trying to sell their venture capitalists that this going to be as revolutionary as the standard Affymetrix RNA expression microarrays that I told you about.

But it doesn't scale. It doesn't give you the comprehensiveness. And, therefore, you know in advance that this is only a somewhat more efficient version of the ELISA test, the antibody-based test. It's not leveraged.

Similarly, when you're looking at two-dimensional gels, there you are capturing a large percent of the proteome. But it's neither small nor is there an automatic data path to analysis. Because you basically have to pick out these different blots and apply some intelligence about which ones you pick, and then feed them into the mass spec, and then do the deconvolution. It's actually not a high throughput data path.

And when you speak, as I have, to those who are really in the know about this area, and they don't have an immediate sales pitch to the venture capitalists on the horizon, they'll acknowledge to you that this is not ready for prime time.

Which explains why I had so much cognitive dissonance, in the late 1990s, when everybody is saying, forget this genomics, proteomics is it. And I'm saying, that's great. I'd love to go to proteomics. But I don't see these cheap, high throughput, comprehensive, electronically data acquired sources available today. That's the goal. I believe we'll reach it. But I think we're actually a full five years, at least, away from a high throughput version of this.

And what do we actually use these microarrays to look for? We rarely use it to find a single gene responsible for a process. But we frequently use it to find a pathway. We frequently use it to find a set of genes working in a coordinated fashion. The assumption being that there's structure at the biological scale.

And what do I mean by that? That genes behave in coordinated fashion. And so the questions that immediately are asked is why should this assumption hold and what evidence do we have that it does?

So why does clustering work in functional genomics? Well, what do we see in the night sky when we look up? We see a fairly, apparently uniform star field. But when we look with the right instruments, we see clusters of stars, from which we have been able to apply our classical and then Einsteinian mechanics to figure out what are the forces that hold these clusters together and make them behave the way we see them behaving.

Similarly, when the Genome Project essentially came to a conclusion in this year, what we had was the gene universe. We had the list of genes that form part of this universe. And here, for instance, is a bunch of such genes. However, this really doesn't tell us anything about how these genes interrelate.

And although there's a bunch of hoopla made about the Human Genome Project, it doesn't start to become really interesting until we can do something like this-- a figure that I borrowed from the Kyoto Encyclopedia of Genes and Genomes, KEGG, which is a pathway representation. Understanding that this is how the stars are interrelated in a galaxy, how these genes interrelate in pathways.

So first of all, why should-- why do I see this structure in the transcriptome? It's because, in fact, in order to get things done, like, in this case, make sure this cell kills itself, there has to be coordinated expression of genes. As we explained to you previously, there are many reasons why gene expression and protein expression may not be correlated.

But if there was no coordination at all of gene expression, two things would be true. One, the genes would be behaving randomly together-- in other words, would not be working together to make a process, which seems unlikely. And then clustering would never work. Because there would be no coordinated action of genes.

And the reason this coordinated action is-- in order to get a job done, these, out of principles of parsimony-- that is, you don't want to have a gene product unless you need it-- these things are only generated when, in fact, you need them for this process-- in this case, cell death.

And that allows people to do these clustering analysis. And there'll be another lecture on clustering. And I just want to give you a taste of what I mean by clustering.

So here's the world's smallest microarray. You have three experiments. And on this little, small microarray you're only measuring a gene three times. And across these three experiments, you therefore have three measurements of each gene.

And the most standard type of clustering is the so-called phylogenetic type tree or the dendrogram, where you calculate all the correlation coefficients between every column, in this case, every set of genes, and the correlation coefficients are shown, as here. You then figure out which genes are the closest together, based on their correlation metric, which, in this case, is 0.88. And you bring them together. And then you can summarize, in a variety of ways, that pair, let's say, by their median.

And from there, you add on the next closest gene. So very specifically, gene one and gene two get put together, because they're the closest. And then the next closest to the joint branch there is gene three. And recursively, this gets built up until you have one of these classical, large dendrograms, where, again, each row is a gene. The columns are different experiments or samples, in this case, a time course. But they were different, distinct samples.

And what you see here is that every row that's close to one another is close to one another because they have a higher correlation, between these two ones that are close to each other, than between two rows that are far apart. Now, there's something.

I'm eliding some complexity. For instance, you can toggle around a whole branch, so that you might actually have some blocks, some rows, which might not be together in some rearrangements of this tree. Nonetheless, by and large, within a subbranch, you'll find, typically, that genes, with a similar expression pattern, are put together because they have the highest correlation coefficient.

And the diagrammatic version of it is as follows. Higher expression level is red. Lower expression level is green. And you see that the red and the green bunches together, because, in fact, each of those genes, in each row, are clustered together by their expression pattern. That's all there is to it.

And then when you look what these bunches actually represent and what people do is, essentially, they draw lines. And it's unfortunate, it's not much more fancy than that, at least in the publications that occurred from 2000-- 1999 to 2003. We draw lines next to these groups. You can actually find, for some of them, some functions, such as cell cycle or coagulation or apoptosis. You'll find different groups of genes clustered in that fashion that correspond to those different processes.

Note, however, that for whatever reason, lines were not drawn against these other blocks of genes. And I'll argue, in another lecture, why that's a mistake. And furthermore, I'll argue in another lecture why this act of drawing lines, like these blocks, although it's ground in the basic truth that these genes more or less work together, where you draw the boundary of that line is an act of creative invagination and does not stand up well to statistical analysis.

In other words, based on the knowledge of a particular biologist, you might decide to change the boundaries of that box. And there's well-grounded information theoretic methods which allow you to draw those functionally related groups more soundly. And this is just a close-up of that.

So the point here is that, if you're not looking for a global pattern, as I just showed you, if you're not trying to see how things are working across a process, like coagulation or apoptosis, and if you're not looking for guilt by association, by which I mean, let's say you know the function of all these genes, and all these genes but not these genes, here, in the middle. Guilt by association says, essentially, that if you know the function of these genes and know the function of these genes, because these are coregulated in the same fashion, you can impute something about that process.

And that heuristic course, as it may be, has been actually exploited very successfully by many investigators. So if you're not using guilt by association and if you're not looking for global patterns, then microarray technology may be inappropriate and misleading. And it may be inappropriate and misleading, because microarrays themselves are much more noisy than people understand.

And therefore, if you're just looking-- specifically, if you trying to answer the question, do these three genes get expressed differentially, then you're probably using the wrong technology, for the wrong costs, and with much greater danger of being misled. And why is that a problem?

Well, let me give you a decision theoretic view of microarray analysis as we should all, I think, think about it. So the most basic micro experiment asks the following question, across a set of experiments is a gene up or downregulated? Or more specifically, of the thousands of genes that we're measuring, which are up or downregulated? And for some, you'll use some statistic. And you find it, it is indeed upregulated. And for others, it is not upregulated.

If a gene is said to be upregulated, is the gene relevant to the process being studied, let's say cancer? Yes. In that case, it's a true positive. It's perhaps a real target for cancer treatment. So therefore, big reward, because either you're a drug company or you're a researcher publishing a finding. If the gene is not relevant to the process being studied, it's a false positive. Therefore, you've wasted your money.

If the gene is not significantly up or downregulated, in other words, it doesn't look interesting by this experiment, if the gene is relevant to the process being studied, then you've missed it. It's a false negative. So then, again, you've lost an opportunity. However, if the gene is not significantly up or downregulated, and it's not relevant to the process of being studied, then it's a true negative. So that's great. You're not wasting your time.

And here's the problem. So you have 10,000 genes, let's say, or 20,000 genes that you're measuring in a microarray experiment. And the biologist and the big pharma can't bear the thought that they're going to miss the blockbuster false negative. And consequently, what they do is they change the threshold for declaring something to be a up or downregulated, so that more and more genes are considered to be up or downregulated.

But the problem is that, not only does this increase, potentially, the number of true positivees you find, but it also hugely increases your number of false positives. And the problem again-- this gets back to the biologists don't understand the difference between a northern blot on a gene they understand and doing a microarray on 10,000 genes that they know nothing about. Because they can be given literally thousands of false positives.

And the typical interaction you'll have with the basic biologists, when you do this kind of analysis for them, they'll say, Oh, Zak, I see that you said, statistically significantly, we can only place bets about 10 genes. But I see down the list, that you said not to be significantly differentially expressed, a gene that I know from my own research is, in fact differentially expressed. But it just doesn't have a P value that reaches a level of significance. So therefore, couldn't we actually just push the whole threshold down to at least the same threshold as the gene that I already know something about?

And the answer is, unfortunately, if you do that, then it'll still happen, you also-- there's nothing to say that these other genes, that are now included in the list, are anything at all like this gene that you happen to know a lot about. So you're using prior knowledge. You're not using the microarray. You're using prior knowledge to say something about it. And therefore, you get a lot of false positives.

And the problem is each one of the false positive is one post-doc going off to do a validation study. And it turns out, with tens of thousands of genes, you can launch a lot of post-docs. So the pharma industry, with it's big, deep pockets, actually got exhausted by this and, actually, quite turned off by a lot of these technologies. Because a lot of the genes that they thought they were going to look after, these top genes, which look kind of interesting-- you can always make a story and create a story about why this gene may be involved in your favorite disease process-- turned out not to be the case.

So by not understanding this decision theoretic view, a lot of money was wasted and a lot of disappointment with this technology, which is one of the reasons I wrote it, in the first sentence of the first chapter of my book, *Microarrays for an Integrative Genomics,* the functional genomic meltdown is imminent.

And I first wrote that in 2000. And everything that I have-- everything that's happened since shows it's true, that, although there's more interesting science being developed, the fact is a lot of the companies that were built around this, and some of the bioinformatics companies that are built around this, are just going down the tubes.

So let's just remind ourselves about-- by the way, I'm very optimistic about the whole field. But I'm optimistic about the whole field when we're careful in our use of the technology, both in understanding the actual engineering part of the technology and understanding the quantitative analysis.

So I know that we've gone through this, with you, several times, but I want to revisit. Actually, is everybody comfortable in the way the basic microarrays work? Do I have to revisit it? If you want me? I do not have to revisit it? Thank you. All right.

So now we can just remind ourselves about the difference between the two platforms. Affymetrix, as you know, does not represent, on each spot of the microarray, the entire gene. It just represents a few oligonucleotides that they have chosen to be representative of part of the gene-- more on the 3' end of the gene.

And what they will calculate, as we will get to, is some measure of the hybridization of these perfectly matched probes to this gene. Versus a probe, which is exactly the same to this one but has a central base which is a mismatch to the designated area in that gene. So, there, the measure of the gene is essentially the difference between the predict matches and the mismatches in aggregate across the entire probe set.

Whereas, in the Pat Brown inspired do-it-yourself spotted industry, what you're doing is you're spotting these probes right onto the glass slide. And then you're going to have a competitive hybridization between a reference set of cDNAs and the test set of cDNAs. And that has some implications which we'll get to shortly.

So we'll make a point, because I'm going to ask you a question later, that the spotted arrays have to end up on the slide through some physical process, such as a printing head. Whereas in the full lithographic techniques built by Affymetrix, what you do is, after you create the first layer, which is bonded to the glass or silicon substrate, you're going to chemically add, let's say, a nucleotide G masking off the rest of the array where you don't want to add that G nucleotide. And then let's say you want to add an A in the next step, you mask off everywhere that you don't want to have the A added, and so on. Clear so far?

Now the problem, of course, is that to do this for a 25 miR requires-- in other words, on the nucleotide that's 25 long-- is on the order of 100 masking procedures. Anybody have an idea how many masking procedures you do for a Pentium 4? On the order of 20 to 30. So they're having to do four times as much photolithography as for a Pentium 4. So that's going to limit their costs, at some level.

And furthermore, let me just say it right here, it turns out that 25 miRs are not the ideal length for hybridization. The longer-- if you get to a longer length, let's say a 60 miR or 70 miR, you get much more sensitivity and much more specificity. But they can't economically do it because of the limits of photolithography.

But the advantage of photolithography is that you can get extremely compact and high resolution. And so you can put a lot of genes on one chip, like the entire transcriptome. So this is just a summary of what I just said here.

Also, we like to think of these in some of computational abstraction. But it's important also to realize that these are real molecules that, during the hybridization reaction, are tethered to the substrate. And therefore, their secondary and tertiary structure may in fact influence how they're able to hybridize with the molecules from the sample. And it turns out that certain types of DNA probes just don't work as well as you'd expect, computationally, because of these effects.

And if you look at Affymetrix probe sets, there are some probes that never work. And despite the fact that they're well-designed from the point of view of finding a matching hybridization sequence. But it's because of these kind of effects. And note, also, that they are not all of the same length. The efficiency of the masking reaction is not 100% by any means. And what you will find in each one of those probes is, in fact, millions of DNA molecules, which have a distribution, which hopefully is close to the desired 25 miR, but, in fact, often is considerably lower. And unfortunately, there's no way for the end user to be able to ascertain just how good the job had done.

So given that fact, we can already see that, even though we can achieve a higher density with these Affymetrix microarrays, there is still a lot of variation that we can expect from different probes within a probe set. Which is why, for example, you might not see, in a single probe set, that the perfect matches have all the same intensity. You can see, for instance, if there's 11 perfect matched probes, five of them will be really lit up and six will be quite dark. And there's many other reasons why that might be the case.

And I just want to make the following point. As I said, when you're looking, for instance, in a sporulation experiment in yeast, what we're looking at is, let's say, a test sample versus a reference sample. And we are looking for their competitive hybridization to the target.

Now, a couple of problems immediately arise, such as, unfortunately, it turns out, for reasons that have yet to be made clear to me-- and in fact, if anybody in this room has any insight, I'd be very curious. Which one gets the C3 versus C5 dye actually makes a difference. So if you switch around the dyes, you'll get different ratios.

So something sterically about the attachment of the different dye causes a difference in the competition reaction. And so people who are practitioners of this and who are fastidious about it, will, in fact, perform both labels, both the Cy3, Cy5, and Cy5, Cy3 label to actually better understand the limits of that technology. It's not clear to me why that's happening.

The other point to note is, if you want to compare across different microarray experiments, you probably want to have the same reference sample. And that sounds like a trivial thing to say, because you'd like to see, for instance, let's say, this condition. The test condition is a bunch of different cancers. And you want to compare it against a reference RNA, which might be at Stanford.

What they do is they pool together all their different lymph node tissue types. Problem, of course, is, if you deplete your supply, you're going to get a different set of ratios with the new supply. And because you're going to have different abundance of RNAs in the reference sample, you're going to get different ratios.

And so you see drift in the ratios across experiments, over time, with this kind of setup. And it's highly problematic. At Mass General, they've developed a synthetic soup of spiked-in RNA controls. I think 256 different RNA controls. And that will give you a standard reference pool. But unfortunately, it's not quite the same as having the full eukaryotic reference pool. So the ratios you might get from that artificial pool might be quite a bit different from what you get in other experiments. So just want to make it clear to you that is not a done deal.

So how do you use, actually, Affymetrix to determine the intensity of gene expression? I just said that, for the spotted arrays, all you're doing is actually reporting a ratio, and a ratio that typically has a normal distribution. What you're doing in Affymetrix is a multi-step process.

First of all, you're going to take the average of the lowest 2% of the cell intensities in the sector. This is the background of each sector. And then you subtract the background from the average intensities of all cells in this sector. And you calculate a background noise, which is essentially a measure of variance of the pixels and the scaling factor and normalization factors, which are settings you comply to Affymetrix. And this is actually somewhat out of date. But I'll give you updates shortly.

So what is reported, for instance, most often as the intensity of gene expression is the average difference. Which essentially is the difference between the perfect matches and the mismatches across the entire probe set. So a probe pair, first of all, is said to be positive if the difference between the perfect match and the mismatch is greater than the amount of noise that is computed based on the background. And a probe pair is negative if the mismatch is greater than the perfect match by the same amount. So therefore, not all probe pairs will be scored as positive or negative. And then, what you have, you calculate a trimmed mean across the positive or negative probes.

That turned out, by the way, to be a highly fluky, non-robust measure of intensity. And that was ensconced in an algorithm called MAS, MicroArray analysis Suite 4.0. So when you read papers and you see Mass 4.0, that's the method that they're using to interpret the data. And the problem is, it was both oversensitive to outliers, because they're using the term mean, and it was also sensitive to the central moment.

So instead, in MAS 5.0, they are using the Tukey biweight statistic, which is a much more robust statistic of outliers. And I will upload a paper today to the My Course website, which will allow you to read the analysis of how they approached this. And this was actually done by Affymetrix. And it's now part of their standard interpretation software, MAS 5.0.

Another competing microarray or high throughput gene expression technology is a serial analysis of gene expression. Did Atul tell you about this at all? He did? All right. So suffice it to say that it looks digital, in that what you're doing is literally counting the number of tags with a particular sequence. But it turns out this has noise, as well, for a number of reasons. One, our sequencing technology is not 100%. It has error. And also, sometimes the tags are not perfectly chosen, either.

So it's actually, in the right-- excuse me. In the right hands, with the right set of genes, it does a nice job. But in actual practice, I've seen it to be fairly noisy and certainly not cost competitive with the standard microarray technology.

Quite interesting is this very nice technology out of Illumina. So they're using this for genotyping, but they're thinking about using it for doing expression microarrays. And it's as follows, you actually have a fiber optic bundle. And the tip of the fiber optic bundle is sculpted so that it fits in nicely. It has a groove that fits in.

Yeah, it has a groove, which fits a bump, nicely, on these optically coated beads, so that we can do the following. We can attach to each of the different optically coated beads a different piece of DNA. So we could literally have millions of different optically coated beads, each with their different DNAs. And then they will settle into the groove on the end of the sculpted fiber optic cable.

And then you can perform a hybridization reaction with a sample. And then, by sending down a laser, down that fiber optic cable, you can actually determine which of these beads has a hybridized piece of DNA versus which does not. And this allows you to do, in a very high throughput fashion, millions of different hybridization reactions.

It's currently being used for doing millions of different genotypes. So it's recognizing a whole bunch of different genotypes with different pieces of DNA corresponding to different genotypes. But, also, it's been it can be used to do expression microarrays. And there are some groups, around the country, that are now beginning to use it.

And for those of you who are in the enterprise program, looking at how Illumina was funded and how they got together with scientists to avoid being sued by Affymetrix, it's actually a great case history, which I don't have time for. But the point is that Affymetrix has a very broad patent around the measurement of expression in two dimensions, in a two dimensional form factor, essentially, with any way that you place these DNA spots down. And this gets away from that patent in a very creative way and also in an economical way.

**AUDIENCE:** Two questions.

**ISAAC SAMUEL KOHANE:** Sure.

**AUDIENCE:** Is the main benefit here that you don't have the tags or the data input down the line?

**ISAAC SAMUEL KOHANE:** There's a couple of factors. First of all, you're not limited in having this form factor, this big, where you're going to do this photolithograph. Even so, with the photolithographic process, because of the registration of scanning, the registration of the full lithography, there's limits to how many different probes you can have on one chip.

Whereas here, because each bead can have its own unique ID, so you know which one it is, and can have its own piece of DNA, it's going to self-assemble. And this is a very microscopic view of this. You can have many more different, because it's essentially self-assembly.

You swirl around the fiber-optic cable in the soup of beads. And they basically all self-assemble onto this tip of that cable. You get many, many more different types of beads and, therefore, of probes in one sample. And the process of labeling the beads and attaching the probe is actually much cheaper than the Affymetrix technique. So it's higher density, more economical, and it gets away from the patent.

Another technology is inkjet technology. The same people who brought you high resolution family photos said, if we're able to put down picoliters in precisely the right spot so that your kid looks the way your kid looks, perhaps we can use that same technology to spot a piece of DNA onto a flat surface. And in fact, they use--

So a spin-off of Hewlett-Packard, before it split off into Agilent, so a spin-off of Hewlett-Packard printers, actually developed the system where they actually have four colors, CATG, and they spit the four colors onto the glass slide. And they do successive in-situ synthesis of oligomers, 50 to 60 long.

And because your pictures of your baby have to look good and have much more high tolerances than anybody cares about in Pat Brown's group, let's say, these are very good looking spots, with very nice, consistent shapes. And it's also very cheap.

So this is the kind of shapes that you get with mechanical pens, which is the normal way you spot. And these are much more consistent and precise spots that you get with the inkjet.

But there's many other uses for microarrays. If you recall at the beginning of the lecture, I didn't describe a microarray as being about expression. I said it had to be able to have a small form factor, it had to be able to interrogate for a large amount of the -ome that you're measuring. And it had to have a direct data path to an electronic format, and so on.

So can we use these same kind of microarrays to study, for instance, gene expression control? So zinc finger proteins are proteins that actually bind to DNA. As you see here, so this finger is around the DNA molecule. And there's different parts of this that do the recognition and binding to a consensus site on the DNA.

And so for understanding that, understanding that it was a consensus sequence with a limited repertoire of possible values, you could actually create a oligonucleotide that had all the possible permutations for this recognition site. And then you then ligate that oligonucleotide to a surface, like the microarray.

And this is the actual results that you obtain when you take different transcription factor mutants. And you see that they bind at different spots, on the microarray, based on their affinity with a different consensus sequence. And since you know which spot corresponds to which consensus sequence, you can quickly develop a notion of what are the consensus sequences, what is shared and what is not shared among these mutants, and, therefore, what is important for the binding activity. As shown by this cartoon that represents the conservation of picker binding consensus sequence.

So a lot of this work actually was pioneered by one of our faculty members in HST, Martha Bulyk. And it's now been scaled up so you can actually buy a lot of microarrays that allow you to do much larger-scale assessments of thousands of different binding sites.

These are the protein microarrays that I discussed with you previously. And you can have a variety of different baits. A small molecule present on a microarray, an antibody present on the microarray, or a different kind of protein baits. Or you could even use a phage as a bait. The problem with all of these is that, as I said, we don't have yet, particularly for the proteins, high throughput methods of doing this.

So for instance, selecting and laying down the antibodies for the entire proteome is just not within our vision. We just don't know how to get there. Currently, in a reproducible fashion, we can do it for maybe hundreds of different proteins.

And these are the kinds of results that you get. You do, in fact, get these good looking graphs, where you can show that you can reliably identify the differentiable expression of different proteins. But again, these are from very small numbers of genes. Whoops.

Others have tried the notion of a universal microarray, where what you're doing is, rather than having to re-engineer the microarray all the time, you're going to create a generic microarray. And then you're going to create some custom technology to link that generic microarray to the system that you're interested in.

So this universal microarray has these zip codes, which basically are standardized sequences that recognize a bit of DNA with a complementary sequence, which you then add chemically to the sequence that you want to interrogate for. And such universal arrays don't require re-engineering of the zip code part of the probe, but only the part of the probe that you want to interrogate for. And it works actually reasonably well, as is shown in these experiments.

I want to point out that a lot of these technologies became popular when Affymetrix microarrays were more poorly performing and much more expensive. Four years ago, an Affymetrix microarray at Harvard cost, with their academic discount, about $2,000. Now it costs about $200 for eight times as many genes. So the price factor has gone way down. These other technologies are just going to go away. It's sort of a Microsoft effect, the technology is good enough and cheap enough that there's very little incentive to invest a lot of your own time to make these other technologies better. Plus, as I said, there's patent control by Affymetrix, which makes it harder to create a business out there.

**AUDIENCE:** Isn't it for things that were being published three, four years ago, maybe, were [INAUDIBLE].

**ISAAC SAMUEL KOHANE:** The short answer is, yes. And then I'm going to show you, quite literally, just how bad that intergenerational difference is. You have heard a lot about tissue microarrays. What are tissue microarrays? They're essentially these collections of hopefully homogeneous samples.

So for instance, you could take a tumor and could slice it and dice it into these small salami slices that you lay down on the microarray. And then you can stain it for a variety of things. And then you can look under the microscope. And if you're clever, get an image recognition program, to detect a particular color for a particular stain, to see how much of a process is present, based on the staining characteristics of that tissue or the morphological change in that tissue.

It's not yet a high throughput technique. And if sort of fails the microarray definition that I discussed previously. But I think that as we get better in the image analysis part of it, it does have a shot of becoming much more of a commodity microarray.

**AUDIENCE:** What have--

**ISAAC SAMUEL KOHANE:** Go ahead.

**AUDIENCE:** How could they ever do something like this, though, given the fact that in the chemical environments, themselves, [INAUDIBLE] incredible. Not only are you dealing with molecules, you're dealing with a very complex, poorly understood system.

**ISAAC SAMUEL KOHANE:** Right.

**AUDIENCE:** Well--

**ISAAC SAMUEL KOHANE:** Well, the answer is the following. The question that you just raised is almost always true in genomics. It's a more general question. With complex systems, how can we do this? And the answer is, low hanging fruit. In other words, there are some strong effects.

In other words, if there is a lymphocytic infiltrate into a tissue, you'll have a lot of immunoglobulins present as opposed to not. So if you stain for immunoglobulins, you'll see it light up. If you are looking for genes that-- if you're going to do an in situ hybridization for an apoptotic gene, those samples that have more apoptosis in them will light up more.

Now, the way you handle them is, without a doubt, going to influence it. Which part of a tumor of this piece of salami was originally sliced out will also have an effect. But if--

**AUDIENCE:** That was less what I was getting at. Let's take like a fibroblast.

**ISAAC SAMUEL KOHANE:** Yeah.

**AUDIENCE:** The degree to which you-- I mean, obviously, that isn't standardized in how you make these microarrays.

**ISAAC SAMUEL KOHANE:** Right, go ahead.

**AUDIENCE:** The degree to which you have confluent layer will dictate the morphology of the cell, just like under visual inspection.

**ISAAC SAMUEL KOHANE:** I see. I see where you're coming from. These are not cells. This is not cell culture. This is a hunk of tissue that was taken out of a patient, fixed, either through freezing or through some other process, and laid down onto that slide. It's not a tissue culture experiment.

**AUDIENCE:** OK.

**ISAAC SAMUEL KOHANE:** That's a very good question. And that speaks to the next generation of microarrays, which I'm not willing to speak about, which is when you have a confluent layer of living cells, as opposed to this fixed tissue that you're going to stain, that's another type of microarray. You can then, actually, for instance, target drugs, different drugs at different points in this microarray and see how these cells react. But god knows what that means in a set of fibroblasts spread as a monolayer. I think it's a big open subject. It's a very--

**AUDIENCE:** I think I'd laugh if somebody they tried to make any conclusion about something like that by entering individual cells into a layer.

**ISAAC SAMUEL KOHANE:** Well, I'm glad you're laughing, because--

**AUDIENCE:** The structure--

**ISAAC SAMUEL KOHANE:** I'm glad you're laughing, because I've been crying about the way these much more boring microarrays have been used in the past. And I'm going to start to give you the bad news, shortly. But I just want to tell you that people are publishing, in first class journals, the kind of thing I'm just describing now. And some of the luminaries in our field.

And so what you have to understand, guys, there's a lot of legacity in the field of genomics. And part of what this course aims to do is for you to learn the limitations. When Joel, for instance, talked to you about SNPs, you'll learn that half of all SNP studies are just wrong.

And what I'm about to tell you is at least half of all microarray studies are wrong. But big companies are being formed around this model they're targeting with drugs-- kind of stuff, Todd. So this is an opportunity for you to make a big name for yourself in the future.

In fact, so here comes the debunking. So Pliny the Elder said the following. "Indeed, what is there that does not appear fabulous when it comes to our knowledge for the first time? How many things, too, are looked upon as quite impossible until they have been actually affected." So this is a very optimistic statement about the future of science, technology, engineering. Build it, they will come.

And this is Pliny the Elder, who wrote this big tome called *Natural History,* which was very well regarded. He's a Roman patriarch. And note when he died-- 79 AD. And that there is significance to that date of death. Because what he was was, in fact, not a true experimentalist scientist, using the scientific method as we currently understand it. He was more of a description-based person, describing what he sees in incredible detail, annotating meticulously, taxonomizing it. But not really putting forward any hypotheses about how things work.

And because he did not understand the mechanism of things, when he went to visit his last scientific investigation, which was look at a local volcano, which happened to be called Vesuvius. When he went to the boat to visit Vesuvius, he misunderstood its basic nature. So, in fact, he died, as his son was trying to drag him from the island, probably from asphyxia from the vapors from this volcano.

And consequently, I ask the question, are we in danger of an imminent genomic Plinian eruption. Now Plinian eruption is actually a term of the trade used by geologists to describe a massive explosion that results in a lot of hot air and ash going up into the stratosphere.

And I think we can argue that we're in similar danger because of over-promising, the slowness to acknowledge the limitations of our measurement techniques, the challenge of linking genomic data to biological and clinical significance, and a lack of formal hypothesis testing, and the lack of sufficient multidisciplinary expertise. For all these reasons, the fact that Todd might roll his eyes at this tissue culture experiment doesn't mean that probably all VCs will think it's great. Because they won't have that multidisciplinary expertise. And so are functional genomicists over-promising? Absolutely, yes.

So let's go through it. Let's go back to our world's smallest microarray, three genes measured before and after an intervention. And in fact, the content of *Nature Genetics* science journal articles about microarrays, in the first four years of functional genomics, of lots of publications in functional genomics, which I think of as 1999 to 2003, so plenty of papers which did the following sophisticated calculation.

You take each gene before and after intervention. Take the ratio. And you report the ratios. And you sort them from high to low. So that, in fact, gene 2, which is 2.1/0.3, that gives you a ratio of 7, therefore, it's the most upregulated gene. It's lists like that that you'd get.

So for instance-- and I feel bad for this poor guy, because I always cite his article, because I just don't want to change my slides. But this is a good point. So this group, out of Yale, did the following interesting experiment, which you should think about as you have your next meal.

Rats, who are starved, live twice as long as unstarved rats. And so what they did is they starved these rats and compared them to non-starved rats and looked at the expression profile of the muscle tissue. And what they saw, using Affymetrix technology, was pretty interesting. There were a bunch of genes that are involved in mopping up free radicals, these bad things that screw up the structure of your proteins and DNA.

A bunch of genes that are involved in mopping up free radicals and combating the oxidative stress were downregulated in these starved mice. I want to point out that they are, for instance, down by 1.5 fold, down by 1.6 fold. Remember those numbers.

Well, I won't be cute about it. The fact is those fold changes, especially when they had a limited number of the replicates, like three replicates, are just not sustainable in any kind of analysis, even today, unless you do a lot more replicates and use much better platforms. So here, for example, is a study that we did using a Incyte chip. Incyte used to make a spotted microarray. They're one of the technologies that went away with the genomic implosion or meltdown.

And what you see, here, I've taken the microarray and I've strung out, into a line, into a vector, the 8,000 genes. And I'm showing you the ratio of these genes across two conditions in this case cardiac muscle of a mouse before and after cocaine.

And what you see is that most of the genes have a ratio of 1. And some are a little bit above and some a little bit below. What to make of it? Where do we draw the line? What is upregulated? Is this 1.4 fold upregulated and above, is that the significance level?

Well, I did the following thing. I applied a Fourier transform to this data, where the dimensional cross, across which I was doing the Fourier transform, was that of the linear position on the microarray. And as you know, the Fourier transform identifies elements which have periodicity.

And what I found was the following periodicity-- a periodicity of 4, a periodicity of 9, and some other periodicities, but a huge periodicity at 4. Anybody have any idea why there was this periodicity of 4 in the data? Remember, regardless of what gene it was, across this chip, there's a periodicity of four.

**AUDIENCE:** Chip.

**ISAAC SAMUEL KOHANE:** Chip what?

**AUDIENCE:** Chip [INAUDIBLE]. The printing?

**ISAAC SAMUEL KOHANE:** Print chip, what about it?

**AUDIENCE:** Maybe the form of the-- or in the tip, especially with the--

**ISAAC SAMUEL KOHANE:** You're getting close.

**AUDIENCE:** [INAUDIBLE] maybe, like, could there be any element--

**ISAAC SAMUEL KOHANE:** You got it. This is what the inside chip looks like. And the reason it looks like that is because they used four pins. Each pin took care of a different quadrant. And so if there's even slightly different physical chemical properties of each pin, in the engineering of the pin, you would get a slight systematic change in the amount of probe laid down. Consequently, it turned out that you got on this chip much more than 1.5 fold changes in the readout just based on which quadrant you were in. And that was just based on the physicality.

Now Affymetrix, in the past has not been guilt-free in this matter. For instance, depending on which way the hybridization solution was washed on, you can see different-- you can see intensity effects going this way or that way. And if there was a thermal gradient in the hybridization chamber, you could also see different intensities that were chip geometry dependent and independent of the gene. Yes?

**AUDIENCE:** What did you think to look for [INAUDIBLE]?

**ISAAC SAMUEL KOHANE:** Because, of course, I knew those four pins. I said, I wonder if there's anything different about the pins. By the way, the second-- and now, by the way, I almost do that routinely. Any series I do, as a sanity check, I just run a Fourier transform to see if there's regularity. The other periodicity was due to how often they replaced the printhead. But of course, the manufacturer would never tell you this.

Let's make things a little bit more interesting. So this is an experiment that Atul did with Morris White at the Joslin, using a now obsolete set of chips.

They had four patients with glucose intolerance. So it's not quite diabetes but it's the inability to bring down your blood sugar rapidly enough after you give him a challenge of glucose. And using a collection of three chips, you measured 35,000 genes. So we did it on patient one, patient two, patient three and patient four. And we repeated the same experiment with the same extracts of RNA from the patient's muscles. These were from the patient's muscles.

So let me ask you the following question. For gene five on this microarray, what should the ratio be between gene five, here, and gene five, here?

**AUDIENCE:** 1.

**ISAAC SAMUEL KOHANE:** Did someone say 1? Thank you. The answer is 1. What should the ratio be between gene five, here, and gene 5, here? Who knows? It shouldn't be 1. The answer is, who knows, because it depends on the individual, right? They are different individuals. Good. So now that we understand that, let's repeat. The intrapatient variation should be 1. The interpatient variation, we don't know.

But let's say this ratio, for the sake of argument, between gene five in patient one and patient two is a ratio of 5, OK? What should this ratio be of gene five from patient one?

**AUDIENCE:** Close to 5.

**ISAAC SAMUEL KOHANE:** Close to 5 as well. So what should the ratio of the ratios be?

**AUDIENCE:** 1.

**ISAAC SAMUEL KOHANE:** 1, thank you. So to repeat, the intrapatient variation ratio should be 1. The ratio of the ratios of the interpatient variation should be 1. Yes?

**AUDIENCE:** So the measurement's actually were pretty close in time and all that?

**ISAAC SAMUEL KOHANE:** Well, it's actually the same RNA sample that we took. And then we hybridized it all at the same time. But all good questions to ask. In fact, never reported typically in the journal article. In fact, so now I have to give you an anecdote. You'll learn about clustering in the lecture later by Steve Greenberg.

But Todd Golub did one of the first studies of the differences that you can find in two physiological conditions or pathologcal conditions based on expression. And he looked at Acute Myelogenous Leukemia versus Acute Lymphoblastic Leukemia, AML versus ALL. And sure enough, using a supervised learning algorithm that you'll learn about, he could do it without a problem.

He also had a test set and a training set. When I put all that data together, I found out I could actually distinguish another group, not AML versus ALL, but a test set versus the training set. And I said to myself, what the heck is this? And I went to talk back to Todd. Because, after all, Todd's one my former interns, so I have no problem calling him up.

I said, Todd, what gives? He says, they were hybridized on different days. And that wasn't published in the literature. But just doing it a different day-- same technician but different days, had slightly different hybridization reaction. And that could be picked up just by looking at the data.

Anyway, to remember where we were here. The intrapatient ratio-- 1. The ratio of the interpatient ratios-- 1. What in fact happens? Well, this is actually kind of crummy by 2004 standards, but it was not bad when we did it. Here is one chip, the other chip, one chip, the other chip, one chip, the other chip, for the four patients.

And you see, here, what looks like a 1-to-1 line, but it's kind of muddy. And the correlation coefficients were 0.76 to 0.84. Currently, anybody who does more poorly than a correlation coefficient of 0.97, I question how good their hybridization laboratory is. Nonetheless, that's what we have. But you still can bleed it to a good 1-to-1 ratio.

However, what was the ratio of the ratios? Now, here's the ratio of the ratios in one case versus the other case. And you don't have to be a statistician to realize that there's no signal here. These are blobs. Now what does a blob mean?

It means that, with one set of chips, this gene was 10 times higher, because of the log scale, in patient one than patient three. With the other set of chips, this gene was 10 times lower in patient one than patient three. So this is particularly heinous. We have not only the wrong magnitude, we have the wrong direction of regulation. And by inspection, you know this is happening a lot.

So right away, you should be asking yourself a question, how can this be? I mean, after all Zak's told me about how these microarrays are wonderful. How can this be that this is so bad? And after all, in the end, there is some signal here. And why is it so bad? Any idea why it is so bad? Anybody have any ideas why these results are so bad? We were not particularly worse than anybody else at that time.

**AUDIENCE:** [INAUDIBLE] different days.

**ISAAC SAMUEL KOHANE:** But we weren't doing that on different days, so some other kind of variation. So let me ask a lead-- go ahead.

**AUDIENCE:** All you have, [INAUDIBLE] if you're doing ratios to ratios.

**ISAAC SAMUEL KOHANE:** Right.

**AUDIENCE:** It compounds it.

**ISAAC SAMUEL KOHANE:** It does-- definitely compounds it. But which kind of measurements are the most sensitive to that kind of compounding? What kind of measurements? What, in the denominator, makes a ratio change the most? Small numbers, right?

So, when you have a gene expression, which goes from 0.6 to 0.3, that's a two-fold difference. When you go from 600 to 300, that's a two-fold difference. You add a little bit noise, this one flips on the left, and that one's pretty stable on the right. Well guess what? 2/3 of transcriptome is expressed at very, very low levels. You have one or two copies of RNA. One third or less of transcriptome, we have hundreds of copies of RNA, millions of copies of RNA of that gene. And those are easily picked up by these microarrays. But they're very noisy when you get down to these very low levels because of exactly that problem.

And so if you don't have a method that looks at the variance and does not take into account expression intensity, your hosed. And that's why just putting a fold cutoff that says, I'm going to look at everything that's two-fold or greater, you know that you're going to find a bunch of things at two-fold or greater at lower expression levels, but they're just wrong.

And thank goodness, in 2004, you can't get a journal article now published unless you do a variance analysis. But that was not true in 1999 to 2002, at the very least.

Even more worrying, so, as bioinformaticians, we were spoiled by the following fact. Is it really 1:45? and I have until when?

**AUDIENCE:** 2:00.

**ISAAC SAMUEL KOHANE:** All right. OK, time flies when you're having fun. We were spoiled in bioinformatics by having this international resource, called GenBank, where we put all our gene sequences into an international resource, and allowed the researchers to compare different things across different systems. And we thought we could do the same thing for microarrays.

So for the 60 cancer cell lines that the National Cancer Institute collects to test the thousands of drugs that they obtain, from pharmaceutical industries, from the rainforest, from China, and so on, to test for chemotherapeutic efficacy, they have these 60 cancer cell lines that they've been keeping around for years.

And so Todd's group, at the Whitehead, did the state of the art hybridization with Affymetrix. And Pat Brown's group, the best practitioners of spotted arrays, did the same analysis with spotted arrays. And shown here is something that you, again, don't have to be a statistician to understand the correlation of the genes that were in common across both platforms.

And the correlation was terrible. There was almost no correlation. That meant that you could not compare the results from one platform to the other. And when I first published this result a couple of years ago-- it was actually a result of a course like this. One of the students did it as a final project. I told him, why don't you compare these two platforms?

And I was convinced that he was wrong. When I looked at him, he wasn't. But I thought it might be some fluke. Since then, there's been about four or five papers doing this again. And it's getting better, but it's still pretty bad. Correlation now is up to about 0.6 across the different platforms. That's a real problem, obviously.

Here's one of the many reasons, subsequently, that we've discovered for this problem. Shown in orange is the ref seqs, the reference sequence that the National Center for Biotechnology Information maintains, a curated resource of what is the definitive mRNA subsequence of that gene.

Shown in Black are the Affymetrix probes, which part of the genome they interrogate. Well, watch it. They seem to be falling off the edge. So if you have, for instance, a cDNA that interrogates this part and an Affymetrix that is off the ref seq. It may have, in fact, a very poor reproducibility.

And we've subsequently done a study of that. We've actually looked at the position. By the way, Affymetrix previously had considered the exact sequences of these probes as proprietary. So it's only last year that they've revealed what those oligonucleotide sequences are, so that we can actually position them on the gene to know where, in fact, they're interrogating the gene. And consequently, when we eliminated those Affymetrix probes which fell off the gene, we had much better correlation across platforms than previously. There's other reasons, but that's, I think, a major reason.

**AUDIENCE:** Maybe try to find some correlations [INAUDIBLE]?

**ISAAC SAMUEL KOHANE:** No. So here is a study, that was done by one of our fellows, of Affymetrix going from HuGeneFL, which is one generation, so to U95A. So this is looking at human RNA for the same genes. We took the same RNA, from the same muscle, and hybridized it on the same day.

Looking at correlation coefficients, we got 0.7 and 0.59. Why do you think that is?

**AUDIENCE:** [INAUDIBLE] Again, you might be looking at different--

**ISAAC SAMUEL KOHANE:** The answer as they picked different subsets. As they learned more about the human genome, you could be more and more accurate about which oligonucleotides uniquely represent that gene, as opposed to matching another gene. And so what we found is that the larger number of probe pairs shared between the generations of Affymetrix microarrays, the better the correlation between those microarrays.

But to answer your question, that means that if it was done on a previous generation of microarrays, even within the Affymetrix family, reproducibility is not good, at least for those genes that don't meet those criteria, which is a huge waste. Because, remember what I told you, we can do DNA sequencing on blood. But the expression analysis has to be done on the tissue that you care about. So if someone used up some precious brain specimens to do that, they're gone. And it'll take them several years to recreate them.

When I actually first presented this at NIH, I actually heard a groan ripple through the whole audience. Because they had actually just blithely gone on from one generation to the other and assumed that they would just be able to analyze everything together. And that was millions of dollars down the tube. Please?

**AUDIENCE:** So you assume that all the old stuff is just junk?

**ISAAC SAMUEL KOHANE:** No. But what you do is try to figure out which stuff is reproducible, and you're careful about it. And it's actually quite doable. And we do it. But do we naively-- and sure, we can't do it wholesale, the way they thought they were going to do it. Yes?

**AUDIENCE:** The good thing is you can do [INAUDIBLE] process. You can just repeat the whole experiment.

**AUDIENCE:** If you have the tissue.

**ISAAC SAMUEL KOHANE:** If you have a tissue. Plus it's not-- even at $200 a shot to put on a chip, plus labor costs may be $500 a shot, so it's 500 times 100 patients, so it's looking like real money. And but I think the real problem there is limited tissue resources. It turns out to be a real issue.

For those of you have ever tried to do one of these experiments, getting the right tissue, with the right annotation, out of the medical system is very hard. In fact, that's why I really like the MEMP program, because it makes the engineers go and deal with doctors. So they understand what kind of social pressures are efficacious getting the right kind of biomaterials out of the medical system. It sounds like a trivial issue, but it's not.

So let me show you the following graph. Shown here, on the x-axis, is the amount of spiked-in probe. So this is a probe of known quantity. And here's the readout from Affymetrix on intensity. Shown in red are the perfect match probes. And you see, as you add more and more probe, it increases. Just as you'd hope, the signal increases. But what you see here is that it saturates at the high levels, right, as you'd expect it to.

Now, pretty interesting is the mismatched probes. These are probes that were designed not to hybridize with the target sequence. But they too, in fact, rise also. They lag, but they rise, as well, with the amount of spiked-in control. And in fact, they don't quite saturate.

What does that mean? It means that the average difference actually starts dropping down. Because even though the perfect match is saturating, the thing that you're subtracting out, the mismatch keeps on rising. So at the higher levels, you actually get a dropping signal. That's lousy. And that's with a clear solution background.

Here, with a eukaryotic background for the spiking, it gets even noisier. You get less sensitivity and even more marked effect. All of these things are true. And yet, you could actually do very, very good science with microarrays and actually discover lots of biology, but you just have to be aware of its limitations.

Let's get back to the dangers of dimensionality. Remember how I said, in the very first lecture, I gave my somewhat lame metaphor about, if every base was a bead on a necklace and the necklace was worn by everybody in Shea Stadium, it's take on the order of 1,000 Shea stadiums to have as many beads as we have bases in a single human genome. I was trying to impress you with, even though we'd like to think that gigabytes are trackable, it's still a pretty big amount.

So let me revisit that question. Given 1,000 stadia full of people, with necklaces, with beads of 10 colors, and let's say that 1 in 1,000 necklace beads are different every baseball season. And you notice that the third seat, on the fifth row of all games, has a yellow bead in the middle of the necklace, every year, in the season opener, in the largest New York stadium, in the last 26 wins, in the last 102 years that the New York Yankees have won the World Series.

In fact, let me restate that. Every time in the last 102 years that you, the very old sports fan, have seen that this equally old person is wearing a yellow bead in the same spot in the necklace, each one of those 102 World Series, the Yankees have won. I'm just asserting that as a fact.

How good a bet is it that the bead will be also yellow on that position next time the Yankees are in the World Series? Do any of you want to put money, their own money on that bet? Anybody give me $1 for that bet with a $1 million in exchange, with $1 billion in exchange? The answer is no, of course.

Because it's all too easy, after the fact, to look through all the games and through all the millions of combinations of necklaces, to find one necklace that was highly, highly correlated with a particular outcome, whether it be the Yankees winning the World Series or the color of the Coke bottle tops that day.

The point is, if you have enough opportunities to test something, you're always going to find it. Which brings us to the following point, if you have that much in the way of genes, and only hundreds of patients, it's going to be all too easy to find a correlation between the values of those set of genes and the outcome that you care about, let's say mortality.

**AUDIENCE:** Can I ask a question?

**ISAAC SAMUEL KOHANE:** Please do.

**AUDIENCE:** I mean--

**ISAAC SAMUEL KOHANE:** Go for it.

**AUDIENCE:** I actually need to think about it.

**ISAAC SAMUEL KOHANE:** Go for it. Go for it. In fact, that would be a good project, just to calculate that probability.

**AUDIENCE:** Well, I get what you're saying. And I would accept it wholeheartedly, without any reservations, if you were talking about just-- maybe I was misinterpreting--

**ISAAC SAMUEL KOHANE:** Yeah.

**AUDIENCE:** --just about any necklace in the audience or whatever. But as you begin to-- I mean, the example that you gave was starting to get more and more and more and more specific. And if I remember, correctly, the more specifications that you put on a probability, the harder it is to--

**ISAAC SAMUEL KOHANE:** Right.

**AUDIENCE:** --feel that. And so then you begin to get into correlations that actually are just-- are probabilistically significant right?

**ISAAC SAMUEL KOHANE:** Except that-- you're absolutely right. Everything you said is true. But you missed the following fact, which is I had the opportunity now to look at all the people in all these thousand Stadia. And all I had to do was find one bead on one necklace that would predict the game. I had millions of opportunities to find that.

But you see, that's very good, Todd. Because you just went through the same error that all the functional genomicists make. Because after the fact, I can always find, with the thousands of genes, some gene that was up or down by dumb luck across the phenotype that I care about, bad cancer versus good cancer. And that's an important intuition to have. And I'm glad you asked the question, because that's exactly it.

If indeed, I had been able to-- if I had a hypothesis I kept updating. And every time, it got stronger and stronger, then, boy, was I a genius by finding that yellow bead. But if after the fact, after having done all the experiments with 102 World Series, I look and search all the attendants to find which bead that correlates, you know I'm going to find it. It's certain that I'm going to find such a bead. But what's the likelihood that that bead is going to be useful for the next World Series? 0 or something.

**AUDIENCE:** Well, you can hypothesize. The correct approach for scientists-- and just keeping to your analogy-- would be to hypothesize that. You wouldn't say, it is. But you'd say, if this is truly a phenomenon that actually exists, then we should see it. And then if you saw it, we would be able to make a stronger statement about it.

**ISAAC SAMUEL KOHANE:** Right. That is correct. But that's not what happens.

**AUDIENCE:** That's not what people do.

**ISAAC SAMUEL KOHANE:** So remember, this is my poster child for you, that I gave you, of large b-cell lymphoma. And I like this paper a lot. It's groundbreaking. And it says the list of genes that distinguish between low and high clinical risk. Do you remember this paper? All right. This appeared, I think, in *Nature.*

2002, same disease, different microarray platform-- it's out of Todd's group. And they, too, find out high and low risk cohort that are predicted by a set of genes. I have a process point, which is, how the hell do you use, essentially, the same technique on the same disease and get into another first class journal? This was, I think, *Nature Medicine* or was it *New England Journal?* I can't remember which.

**AUDIENCE:** *New England Journal.*

**ISAAC SAMUEL KOHANE:** New England Journal, I think. *New England Journal.* Yeah. No, no, no, that was *Nature Medicine.* Then yet a third paper-- and no, I fully don't understand how this can happen-- same disease, same question, also using microarrays, and they, too, find a high and low risk group.

What was the overlap in the set of genes that predicted outcome in these? On the order of 20% to 30% depending on how you sliced it.

**AUDIENCE:** Chance.

**ISAAC SAMUEL KOHANE:** Which is getting pretty close to chance. Now, I think there's some signal there. There clearly is some signal there, but I don't know how much of this yellow bead phenomenon we have, which is basically the multiple hypothesis testing problem revisited in a very, very vicious way.

And the problem is-- this is more than actually a problem. I'm just irked, of course, that the same thing got published three times in first class journals. But the real problem is that people are being stratified today in oncology protocols based on expression profiles. So you know that, if they're not taking this multiple hypothesis testing really to heart, patients are being stratified the wrong way based on a subset of genes.

And remember, I told you that you could get together as groups of two for a final project? I would strongly urge at least one group to think about looking at something like this. Perhaps this very study, because no one to my knowledge yet has actually published this. The emperor is not, at least, wearing enough clothes.

What was overlap? And what was the characteristic of the genes that overlapped across those? And could you, in fact, come up with a robust-- more robust set of predictors based on those three studies. But that's highly problematic. And I think most of the fault is in this overfitting problem.

So why are these things inconsistent? Well, also, they were different populations. Not all humans are the same. And this brings me to another point, which is, why, all of a sudden, is it OK to do 100 patients and do a clinical study, whereas for everything else that we've ever did, with much fewer variables, we had to get thousands of patients? And the answer, of course, is because we have limited resources and so on. But I think it's somewhat delusional to think that we could get as good predictors, with only hundreds of patients, as we could get with thousands.

Now it's true that we can characterize two groups. Because we have all these thousands of genes to measure. And we can measure broad patterns. In fact, Botstein, who was one of the leaders in this area, never pretends that any single gene is a reliable measurement. He says, just give me an overall impression of the biology, of the pattern that's going on. But when you start then taking a set of specific genes and then hanging someone's prognosis on that, then you're perhaps getting close to a line where you have to be a lot more methodologically strict and understand what are the power and significance of your studies.

And again, some of them are using different measurement platforms. There's overfitting and some use of indirect measures. So that's exactly where I wanted to end up today.

I want to give you a heads up that, soon, you going to be getting your first problem set. It's going to be a very simple one, which is it's going to be a treasure hunt through all the national databases, to allow you to just make sure you touch all the databases, and so that you can be a modern biological researcher and find out what you need to find out about the different biological databases.

The other problem set's going to be one around clustering and classification. Just make sure you can do it correctly. But I'd like to start thinking about your final projects. And it's now mid-February. Again, I really would like to have settled on your final project no later by than mid-March. So, therefore, if you're having any doubts that you're honing in on a project, please talk to me earlier than later.

And there's nothing wrong. In fact, it's just the opposite. It's great if you can think of a problem that you would actually like to do further research on, there's nothing more motivating than that to address.