ISAAC SAMUEL KOHANE: So the overview of today's class-- I'm going to try to convince you that the future is now, that all this talk about the geniculate revolution is not pie in the sky stuff for venture capitalists for the next 10 years, but it's stuff that is of immediate practicality today. I'll spend a little bit of time in distinguishing between genomic and genetic more for your taxonomical edification and for any substantive reason. We have to touch upon heredity and what that is all about.

And then we have to talk about the fundamentals of the resequencing of the diagnostic process that's being put into effect by the availability today of all this genomic data and then I'm going to finish up about how all this is resulting in accelerating consumer activation. And for those of you who are practitioners of medicine, and there's a subset of you who are, this should be alarming-- not that it's bad that consumers get involved but that they're ahead of us in their knowledge and application of this knowledge. So the future is now.

So just because some of us, in fact, claim not to have taken any biology, I just want to review some basics. This will be old hat to some of you, but since you may not have that old hat, I'm going to review it. This is the old dogma, and this dogma is flawed.

And it's not quite true. Hey, welcome. Just introduce yourself.

AUDIENCE: [INAUDIBLE]

ISAAC SAMUEL KOHANE: And you are a student at?

AUDIENCE: Oh, I'm sorry. I'm a student at MIT. HST student.

ISAAC SAMUEL KOHANE: HST student, chemistry background. All right. Todd, have ever had a basic genetics course in your life? I ask the simple questions. Yes?

AUDIENCE: Yeah.

ISAAC SAMUEL KOHANE: All right.

AUDIENCE: Undergrad.

ISAAC SAMUEL KOHANE: Undergrad. All right. So I'm just reviewing the basic dogma, which is DNA gives rise to transcribed RNA, which gives rise to translated protein. So let me ask a question of the class, and I start saying you because I can't remember your names yet. Is the DNA in each cell in our body the same?

AUDIENCE: No.

ISAAC SAMUEL KOHANE: Why is it different?

AUDIENCE: Which as the cells differentiate making it every time.

ISAAC SAMUEL KOHANE: Good because you just stated the misunderstanding that most people have. In fact, with exception of your sperm or eggs, all the genome of the entire body is exactly the same. The full genome is available. In fact, the reason you can clone an animal from any somatic cell in the body in theory is because with some caveats which I don't want to get into right now, it's, in fact, the same darn genome. The exceptions are gametes, which have half of the genome, and red blood cells, which have no nucleus.

Now why is that important? And there's some other things like methylation and the tips of chromosomes which I don't want to give you details of. But why is that important?

Because in fact, what makes a liver cell different from a white blood cell is not the DNA but what RNA is being transcribed and also what protein is being translated from the RNA. And the reason I bring this up is because as we measure DNA using sequencing or genotyping or we're measuring RNAs using transcriptional profiling or we measure translation using proteomics, we have to start thinking now, even those of us who are not MDs.

As clinicians, where can we actually obtain this biological material? And if you're doing a DNA study, anything is good. You can get white blood cells. You can get a blood sample.

But if you wanted to study, for instance, brains and understand what makes someone susceptible to brain tumors, you could not get an RNA sample without actually getting a hunk of someone's brain. And so that immediately tells you some limitations of these technologies. So if you want to do RNA or protein profiling to look at someone's likelihood to get a disease, for instance, you have to get the right tissue, and that tells you the kind of limitations that we face when we're doing clinical studies. I mean, certainly, we can hack out brains out of mice or drosophila, but we can't readily do that with humans.

And yet that's frustrating because on this end, we're closer to function than we are to this end. This is the master code, but this is what leads to the actual functioning the interacting of the proteins. And so people are very excited about proteomics, for instance, because it's very close to function.

But when you're looking about what are we going to actually be measuring the proteins on, we're fairly limited to things like serum and urine. We can't take pieces of tissue out of human beings. And we can talk on another day about how to get around these obstacles.

So what's the magnitude of the task? From one perspective, it seems relatively piddling. I mean, there's, after all, 46 chromosomes, and these chromosomes are these little microscopic things, bushy things of hyper coiled DNA. And it seems fairly discrete. And I know that you know it's 3 billion bases, but I want to give you an appreciation for how much 3 billion bases are.

So let's say we have a necklace of beads. Each of those beads represents one base pair. Can I take it for granted that y'all know what the base pairs are? ACTG, right. Good.

So if you took a bunch of these necklaces and gave a necklace to everyone in Shea Stadium, it will still take 1,000 stadia to get enough necklaces to get all the bases that we have in the human genome. So that's a lot of beads, and it gives you some sense of the coding challenge and the deciphering challenge. But you'll have technology nerds say the following.

Some say that Madonna's music is actually more information than Madonna the person because-- and I actually heard this from-- someone in Microsoft say this in a big [INAUDIBLE] talk because they said, look. A DVD with her music video is 4.7 gigabytes whereas her genome, 3 billion base pairs at two bits per base, since there's four bases, and 8 bits per byte, is only 3/4 of a gigabyte. So what are we making the big deal about?

And in fact, they went on to say correctly that every human is on average only less than 0.1% different than any other human. So we could actually store you all as just a difference file of everybody else. So you can it turns out if you did that, you can compress you down to the size that could fit nicely on an obsolete diskette.

But that's a the question that I want to raise. Is Madonna her DNA sequence? And the answer, I hope you realize, is no because what makes Madonna particularly Madonna is not only the way she developed in her womb even without environmental influences as programmed but all the developmental influences she had through her life-- the good knocks, the bad knocks, all those experiences, environmental exposures that bring her state, her organismal state, into the current megastar state it is. And there are many other states that could have been depending on what those environmental influences were. And if you really want to capture a state, it's not only-- it's therefore the evolution of her physical state, not just her genome, but all the alternative splicing.

As you'll hear about next lecture, genes can be spliced in multiple ways to create RNA. So that's on average about three per gene. The proteins which are derived from RNA can have a multitude of different modifications-- adding of glucose, glycosylation, adding of phosphorous, phosphorylation. And in fact, this is underestimated I'm giving you.

That's on order of 100 to 1,000 modifications. So that's greatly increasing the complexity. And then in a multi-cellular organism, each of those gene products can be in any particular compartment in the cell and then can float around to any cell in the body on the order of trillions.

So Madonna now is several orders of magnitude larger than her music, as we are all. And this is just as a caveat to note that although it's really impressive that we can sequence an entire genome, all we're showing is the master code that can result in one of many potential fates, and to avoid the reductionism that often happens, you have to remind yourself what were the environmental influences. And we'll come back to that again.

Now getting on, again, a more optimistic technological note, it is nonetheless true that we can now measure today with commodity technology large parts of your genome for commodity prices. And what do I mean by that? Let's look at RNA microarrays, and I know some of about these and some of you don't. But I just want to remind you of them.

If you take a tissue-- let's say, liver. Or tissue-- by the way, let me just interrupt myself to say that all these PowerPoint presentations-- thank you. All these PowerPoint presentations and the video are going to be available through the MyCourses website. If you don't have an account yet through MyCourses, you can definitely get one.

And so all the PowerPoint and all the video will be available. So don't feel compelled to take everything down, and just be more sure that you're asking questions in the flesh. So take liver or, let's say, liver under influence of, perhaps, insulin.

You take it. You grind it up. You label the RNA that was extracted from that ground up liver with a fluorescent compound. That means you attach a fluorescent compound to that labeled RNA.

And then you purchase. And I used to say purchase or make, but that's becoming less and less true. It truly has become a commodity, and companies such as Affymetrix or Agilent will sell you these chips, and they can do it probably with higher quality, unfortunately, than we can in the homebrew fashion even though if you go to the Stanford website, microarray website, they'll tell you how to build your own microarray manufacturing facility out of RadioShack parts for less than $10,000. Nonetheless, it's easier and more standardized at this time to unfortunately buy it from this Microsoft-like monopoly called [? Affymetrix, ?] for instance.

In any case, so you buy this chip, and pre-positioned on each spot on this chip-- and I will have a whole lecture devoted to microarrays, so don't worry about the details too much. Pre-positioned on spot on this microarray is a sequence of DNA corresponding to the gene that you're looking for. So this might be the sequence of DNA corresponding to growth hormone, a sequence of DNA corresponding to insulin, a sequence of DNA corresponding to the insulin receptor.

And what's going to happen is through the hybridization reaction that you all learned in basic genetics, the RNA that you extract from the tissue will hybridize with the matching sequence of DNA that is present on this chip. And suddenly, the chips are no bigger than this if not smaller. And then you scan it using a simple HP scanner of no higher quality than one that you used to scan in your family photos. You scan the image of a chip, and it fluoresces proportional to the amount of fluorescent RNA at any given spot.

And so what you have is a readout now at every spot of how much of that RNA was present in that sample. And the current density of Affymetrix microwaves is now on a single chip. They have spotted-- and this is using, actually, a photolithographic process. They have spotted every single gene in the human genome-- in fact, so many so that they are actually putting extra ones, such as different alternative splices. So more than 30,000 genes on one chip, and you can buy it through Harvard pricing for about $250.

So you can measure the expression of 30,000 genes for $250. Now that sounds academically as an interesting thing, but how does that translate into genomic medicine? So let me give you my poster child for a pulse cycle study which has launched 1,000 chips-- chips, but chips as well. It's launched several million chips.

This was a study that was done out of Stanford. And they did the following thing. They had patients with large B-cell lymphoma.

Large B-cell lymphoma is a type of cancer that is on a sort of medium bad disease-- not as bad as lung disease, not as benign as thyroid disease, thyroid cancer. And this is a disease that was previously monolithic. In other words, the patient came to you, and you look on the microscope and said, ah, you have large B-cell lymphoma, and you had a rough estimate of how long they were going to live, and you'd give them all the same treatment because you could not distinguish between subclasses of these patients.

So what they did through techniques that you will learn about in this class is they performed clustering analysis on these patients when they took RNA extracted out of the lymphoma of several dozen patients. And when they did a clustering experiment, they were able to see that there were two groups of patients based solely on their gene expression profile that there was one group of patients that have one gene expression profile and another group of patients that had a different gene expression profile.

Then and only then, they asked themselves after the fact what was different about these patients. And very rewardingly, they found that corresponding to 1 gene expression profile-- in other words, one set of genes being switched on or off-- was this group of patients in red where here on the x-axis is years of survival.

Here on the y-axis is probability of survival. And so what you see here is one group of patients who die very fast, most of them, by two years and one group of patients who have better than 50% survival in the 12 year outcome. And that was remarkable because it told us that previously, where we had not had a sub-diagnosis, we have now a new diagnosis-- high risk and low risk large B-cell lymphoma. Two, we have a new prognosis. We can tell patients, unfortunately, this profile, you're at much higher risk than you would have told them previously.

Three, you have a new clinical opportunity, a new therapeutic opportunity, because without discovering any new drugs, you're just going to hit these ones much harder with the chemotherapeutic agents that you already have because you know they essentially have a much more dangerous kind of cancer. And then fourth, you have a new research opportunity. Why are these patients dying off?

Are they pumping out-- do they have a transporter that's pumping out chemotherapeutic agents out of the cell? It's unclear, but we already know that they're using this as research to-- for instance, it turns out this is a B-cell marker. This group is enriched for B-cell markers.

B-cells are a type of immune cell, and they're actually using a therapy directed against that B-cell to actually treat these patients. So that's fairly remarkable. From a trivial clinical research experiment using this commodity chip, we have a new diagnosis, a new prognosis, a new therapeutic opportunity, and a new research opportunity.

And this has been reproduced multiple times for lung cancer, for breast cancer. You can now actually purchase a test in the Netherlands based on the needle biopsy of the breast. We'll actually stratify the woman on her prognosis based on the needle biopsy.

And it's not only for cancers. Now it's been done for a variety of inflammatory conditions-- inflammatory bowel disease, rheumatoid arthritis, a whole bunch of conditions where you're actually getting a much broader insight into the physiology by looking at 30,000 genes than we were by just a patient asking a few questions and measuring just one or two variables in their blood. So this brings us to this cartoon that says you're not ill yet, Mr. [? Blondel, ?] but you've got potential.

And of course, this is the motivation behind a lot of genomic medicine that we're going to be able to by having these highly forecasting markers, these predictive markers, we'll be able to move away from acute and expensive and not very effective interventions to much more cost effective, cheaper, more effective prevention techniques. And it's not theoretical. Right now one of the largest groups-- by the way, how does chemotherapy actually get devised? So how do these [INAUDIBLE] get devised?

So there are these large cooperative oncology groups, literally dozens of hospitals and medical centers that cooperate on devising how much of cisplatin, how much Adriamycin, how much azathioprine to give to various patients with the various chemotherapy protocols? And they treat these thousands of patients, see how they do, and then issue a new protocol based on outcome. Very highly funded on the order of hundreds of millions of dollars a year.

And today, because I'm part now of the cancer and leukemia group B, one of these cooperative groups is now-- this entire group is now coming up with new protocols that are all based on genome-wide understanding of expression. In other words, they're going to put you into a different treatment category based on your gene expression profile. And this should worry you because as you have learned in this class, there's a lot about this analysis, this cluster analysis-- I gave you the shining great story about it.

But even that initial analysis I showed you of large B-cell lymphoma is flawed. And yet people are moving ahead very fast to turn these into clinical use diagnostics. And because they don't understand fundamentally the limitations of their clustering techniques, for instance, and the limitations of the measurement platforms like the microarray analysis, they're actually going to do patients a disservice by taking patients who don't need the therapy and put them into a high risk class and vise versa.

So I think it's hugely important. One of the things that I want you to learn in this class is how to read skeptically the literature because in fact, it will help you become better leaders in this area of high dimensional medicine. What do I mean by high dimensional? Well, you're looking at thousands and thousands of variables.

Well, here, we're just looking at expression. But as we'll get to later in the course, it's, of course, looking at all the different SNPs on your genome-- the different variations, I should say-- and your protein state. So this is the current state of the art for those of you not doctors. It's the stage of the cancer, where is it in your body, the grade, how ugly do the cells look under the microscope, and what kind of cell type are they is typically what drives diagnosis today.

But even today, it's now starting to be driven in real protocols being deployed for real patients by expression profiling. And that begs the question that I think is now upon us, which is, are we now on the threshold of a new taxonomy of human diseases? Human disease has been classified according to manifestation, and it's actually not so long ago that we used to think of fever as a disease rather than a symptom.

And if you look very closely at most diseases like the inflammatory diseases like arthritis, they're still, for the most part, based on phenomenology and not on some deep mechanistic etiology or cause. And as Thomas Lewis pointed out in 1944, the diagnosis of most human disease provides only insecure and temporary conceptions because in fact, what we're describing is always the tip of the iceberg, the things that the patients complain about. And there's only so many things you can complain about.

This hurts, this hurts, this hurts, and I can or cannot do this. And there's a limited set of all the ways it can manifest. But it's actually reflecting a much larger multitude of possible things that can go wrong in this very complex mechanism that is the human being. And in fact, of the main common diseases only infectious diseases have a truly mechanism-based nomenclature.

You see how people name infectious diseases. It goes right down to the organism causing it. The rest of it is pretty loosey goosey as it relates to the mechanism. And here we have part of the challenge in genomic medicine is if we're measuring comprehensively the entirety of the physiology of the patient state, we can actually come with a much more objective and universal nomenclature, which is a nontrivial thing for medical students who are being-- for those of you-- is there any medical students in here?

Yes, who are being assaulted by the-- I was told once in medical school that I had to learn on the order of 20,000 new terms, and you can see they're already outdated. There's also a reason why today we should be thinking about changing the use of everyday medications. Shown here is a protein that codes for a potassium channel it's a protein that allows potassium in or out in a selective fashion in and out of cells.

And it turns out if you have a misspelling or a polymorphism of one of your nucleotides here, then it will result in a new amino acid in one spot, and that will change the electrical properties of this potassium channel such that individuals are much more prone to cardiac arrhythmias, bad rhythms of their heart, if they are given a sulfa drug. Now sulfa drugs, for those that you don't know, are a broad range of antibiotics, and in pediatrics, for instance, we give almost all kids sulfa drugs.

Bactrim has part of it a sulfa drug. And so what we're saying here is there is a subset of children who have this mutation. And if you give them sulfa drugs today, you're going to put them at risk of death, which seems like a bad decision for just treatment of earache.

Now if you're still awake, you should be asking yourself the following. Well, Zack, you're telling me about this, but why should I really be worried about this? It's obviously a rare thing. It's probably less than 1 in 10,000. And why should I do this expensive re-engineering of the medical process in order to be able to avoid harming these children?

So I'm going to ask you a question. How much do you think in a high throughput laboratory like the Channing Laboratory across the street, how much do you think it costs to see if your genome has this polymorphism, and therefore, we should be careful about treating you with sulfa drugs? How much do you think it would cost us to process your blood and obtain which spelling you had for that one polymorphism? And I've asked this of the entire class, and I've asked this endless times for very senior people and very junior people.

And unless you really know the answer, then don't answer. But let's go for it. "Price is Right" rules work.

AUDIENCE: Well, you assume that you have a blood sample already there? Just the cost of running the test?

ISAAC SAMUEL KOHANE: Yes, I'm not asking you to pay for the blood draw. Right. How much?

AUDIENCE: 30 bucks [INAUDIBLE].

ISAAC SAMUEL KOHANE: Less than a buck.

AUDIENCE: [INAUDIBLE]

ISAAC SAMUEL KOHANE: 130 bucks.

AUDIENCE: [INAUDIBLE]

ISAAC SAMUEL KOHANE: 5 bucks?

AUDIENCE: [INAUDIBLE]

ISAAC SAMUEL KOHANE: How much?

AUDIENCE: $1

ISAAC SAMUEL KOHANE: $1.

AUDIENCE: 50?

ISAAC SAMUEL KOHANE: What? 50?

AUDIENCE: I think I [INAUDIBLE].

ISAAC SAMUEL KOHANE: 100?

AUDIENCE: No, less than $1.

ISAAC SAMUEL KOHANE: Less than $1. So let's give you reality. Today, you'd be charged-- charge, which is not the question I asked you. You'd be charged approximately $500 to $1,000.

The actual cost if you're running the test in time of your technician and materials, so the full cost is, if you're extremely inefficient, $0.50. If you're a very efficient, $0.10. So that's important to realize because we are now today in an emergency room routinely setting off the SMA 20, which is 20 different things, like sodium potassium, chloride, carbon dioxide, BUN, creatinine, and a few other tests. And you're charged on the order of-- or your insurance, hopefully is charged on order of 100 to 200 bucks to do that. The marginal cost of this is no higher than of doing those electrolytes.

So what this tells us is that today, it's possible to run, for uncommon diseases, 1,000 different snip tests at a cost effective way. So even though your polymorphism is rare, if it has clinical significance, we can find 1,000 different rare polymorphisms that in aggregate may end up being a lot of people. And I am not the only one to have recognized this, and there are now a bunch of companies who understand this.

So the companies who previously were charging $1,000 per tests are under some cost pressure, and to give you a sort of reality-based feeling for this, probably what you'll see over the next 5 to 10 years as this technology, as-- understand, this technology is actually a commodity level-- is that price is going to be crushed down from the 1,000 level to a 10. Of course, they're going to resist every step of the way and try charge every last dollar along the way, but it's sort of an internet kind of effect because the true costs are trivial, and we're charging the government these bulk rate prices all the time for our human association studies.

So this is a hugely important understanding to have. And why is this so important to understand? Because if you look at the number of publications of different polymorphisms that occurred with time, this is 1980. This is 2005.

And what you see here is that it's been growing steadily and not so slowly every year so that we're now close to 8,000 publications per year about reporting on different polymorphisms. Now as you'll hear on two or three lectures from now, well over 50% of these are wrong reports, and one of the things you have to learn in this class is what constitutes a good versus a bad report, a protective factor based on a polymorphism. Nonetheless, it's telling you that there's thousands of polymorphisms out there.

Even if 50% of the reports are wrong, thousands of polymorphisms are associated with clinical disease that are being reported on literature per year. So that's a huge number. And as we'll get to shortly, genomics is not about single gene diseases. But nonetheless, the number of diseases that are being found to have a monogenic basis is growing even faster exponentially from 1981 till the Millennium. And so what I told you until now the $0.05 figure was, I wanted to make sure you understood that was true today.

The following is not in the so distant future. It's actually doable now, but it's at a research basis only. This is a big chip. It's a chip made out of many Affymetrix chips sold by a company called [INAUDIBLE], which is a wholly owned subsidiary of Affymetrix. And I'll tell you more about this chip two or three lectures from now.

But suffice it to say that instead of interrogating for a different gene in every spot, it's interrogating every single possible misspelling of the entire human genome or-- and I think this is-- yeah, the entire human genome on this collection of chips so that in one fell swoop-- this is actually just for chromosome 21-- you can actually fully genotype the entire chromosome of an individual within two hours. And that's available today. It's commodity technology. It's just too expensive to be able to give you the same sense of reality as I did for single SNP genotyping.

And I mean, this just shows that they're able to comprehensively show the frequencies of distribution of different SNPs across these individuals in one fell swoop. And that reminds me within the next day, I'm going to start uploading to the My Courses website-- I'll get to you in a second-- to the My Courses website PDFs of all the class reading. So today, I'm going to backfill it with things that would be relevant to today's lecture. Nonetheless, for future lectures, you will have the PDFs for the lecture of that day prior to that day. So-- you're going to ask a question.

AUDIENCE: Yeah, [INAUDIBLE] do it [INAUDIBLE]. So these are just identifying a sequence of genes, or is this coming out of the genome project, and it's based upon [INAUDIBLE]?

ISAAC SAMUEL KOHANE: So what it's based is it's a confluence of several industrial-- well, what is it? Each little sequence on this microarray corresponds-- name on the blackboard. Each sequence corresponds to a stretch of DNA, and the center of that little stretch of DNA that stretch of all of the nucleotide, 20 bases. In the middle of that is the polymorphism.

And so let's say we want to capture all four possible polymorphisms. For that stretch of DNA, you'll have four nucleotides corresponding to the ACE or C or T or G that's different on that central base. So on here are all the different polymorphisms that are known for chromosome 21.

So how do they get that? Is a combination of the verified SNPs from the Human Genome Project, the verified SNPs from Celera which they had to pay cold cash for-- but they're a company, they have no problem with that. And they also licensed SNPs from ABI, from [INAUDIBLE] Biosystems.

And finally, they did their own resequencing so they would know what are all the SNPs, the common SNPs, for this region. So you're able to, in one fell swoop, directly assess through hybridization of DNA now, not RNA, which are the genotypes of the ones that are known. Now if it was not known, then it wouldn't show up here.

AUDIENCE: So just as a follow up question, seeing as how these are SNPs that are known and taken from the sources and we know that those sources are basically a few sample subjects.

ISAAC SAMUEL KOHANE: Yes.

AUDIENCE: --are we going to talk at all in the course [INAUDIBLE].

ISAAC SAMUEL KOHANE: Class.

AUDIENCE: --the class about the implications of a very small pool of individuals an compare with other people's [INAUDIBLE]?

ISAAC SAMUEL KOHANE: We are definitely-- so what you're referring to is, for instance, the Human Genome Project and the Celera project were based on well under 20 individuals each. And so I have two columns. One, a column which I'm not too comfortable with but nonetheless true is that if it's a common SNP, it'll show up typically in one of those 20 more or less. The danger is the opposite, is that a lot of private mutations-- in other words, you may share a SNP with and your immediate family, wherever you came from, that's not present in anybody.

But if you're one of the 20, that now has this major representation in this database, and it's probably the other direction that we have to worry about. Nonetheless the answer to it is very simply this-- is massive resequencing of a lot more individuals. And as this technology becomes commoditized, that problem is going to go away. So the HapMap map project, for instance, is a project, which you'll hear about during this class.

And it's a project to actually define the structure of SNPs across larger populations and across ethnically diverse populations because as I hinted to you in that term private mutations, you can imagine or you will be taught that there are different subpopulations that have these SNPs that are only present in a subpopulation either within a subpopulation called your family or a subpopulation called people that came from that continent that you originally came from. So the short answer is it's a problem, and it's a problem more on one side and the other. In other words, since-- let me rephrase that.

There's a certain frequency of common SNPs, about one per kilo base. But if you look at rare SNPs, they have a much, much higher frequency. In other words, I have a bunch of SNPs that are just unique to me and a few thousand people that have a shared heritage with me. And those probably don't have, by the way, for the most part, any clinical significance. But if I were one of the 20, it'd be over-represented in that sample if it's not a common snip.

Ah. So this makes the following simple point. Why-- it's answering the following question. Why is computation such a central part to genomic medicine?

And simply this-- these two 1999 style stock curves, stock market curves, correspond to the following two processes. This shallower curve is the number of publications in "Medline," the online repository of articles, and goes all the way back to a century, Index Medicus. And it's [INAUDIBLE], say, exponential curve. But because of the industrialization of gene sequencing, the number of sequences have risen at a far higher exponential curve.

And what you can think of this, as a proxy for the knowledge gap. If you take this to be genomic data, the genome sequences being genomic data and this being as a proxy for our knowledge, this gap is. Growing and it's growing even faster than it's portrayed here because it's not just DNA sequence. It's genetic maps physical, maps, polymorphism, structure information, gene transcription patterns, protein translation activity. And all of these end up in databases. And if you are like me in the middle of a Medical Center,

You'll find that there's very, very few people who are trained as clinicians and who have the requisite computational skills.

So I have a PhD in computer science, but that's not the typical medical path a lot of you in this room actually have the nice joint skill set, but you're in the minority. And consequently, because all these data types end up in a database, trying to translate this into medicine is fundamentally devolved to the responsibility of individuals who can both understand the biological problems and translate through a variety of computational techniques all this data into more knowledge. So the hubris, the conceit of [? bioinformaticians ?] is that somehow we can use computational techniques to raise this to that level.

We did the [INAUDIBLE] already. So now let's talk about what is genomic versus genetic because you hear these terms a lot as if they're interchangeable. And they kind of will be, but they're not quite yet.

And it depends on your perspective. So let me just tell this joke. I hope you think it's a joke.

An engineer, a physicist, a mathematician, a computer scientist, and a statistician are on a train heading north and had just crossed the border into Scotland. They look out the window and they see a black sheep for the first time. The engineer exclaims, look. Scottish sheep are black.

The physicist yells, no, no, some Scottish sheep are black. The mathematician looks irritated and says, there is at least one field containing at least one sheep of which at least one side is black. The computer scientist says, oh no, a special case.

And in fact, it's how you view genomics and genetics actually corresponds to where you find yourself on this spectrum. So that will probably inform you why there is some global confusion. And finally, the statistician, I forgot, says it is not statistically significant.

So here's one view of genomic versus genetic. So genetic medicine, the kind of medicine that's been around for us for a long time, has to do with low frequencies of high penetrance genes by which I mean genes that if you have that particular mutation in that gene is going to give you a high likelihood of having that disease. And there's thousands of these relatively uncommon diseases.

The most common of these, hemochromatosis, which is iron deposition in the liver, has a frequency in the population about 1 to 300. And these diseases previously had been mostly assessed indirectly or focused on single genes by which indirectly linkage-- for instance, linkage within a family. In contrast, genomic medicine is about not in a rare disease but just common diseases.

It's diabetes, cancer, heart disease, inflammatory diseases. And therefore, almost by definition, the geniculate risk for common diseases will often due the disease producing areas with relatively high frequencies-- that is, greater than 1%. And all these genes in this perspective may be disease-causing.

And they're not for uncommon diseases. They're for common disorders, and they're due to the interactions of multiple genes and environmental factors. And because of the industrialization in genomic medicine, we do direct experimental access to the entire genome in the fashion that I just illustrated.

Someone else, [INAUDIBLE] and [INAUDIBLE], have given a related similar distinction between genetic and genomic. In genetics, you're looking at basically a structure of genes, the sequence. In genomic medicine, you're looking at the function of these genes.

In genetic disease, you're looking at genomics. You're looking at DNA. In genomic medicine, we're looking at more than just the gene. We're looking at the RNA and the proteins.

This is what I referred to already. Map-based gene discovery-- in other words, probabilistic linkage between loci versus direct sequence based gene discovery. I mean, now, diseases are being discovered computationally. I don't know if I'll get around to it, but there's a great example now of how people are discovering cholesterol risk factor genes purely based on our shared attributes in our genome between man and mouse. Without having ever poured a beaker, people are actually being able to computationally just identify new risk factors in that fashion.

But perhaps I will get to that. It just occurred to me that I don't have lecture anymore on comparative genomics, which is how do we actually take advantage of the genomes of other species to inform us about human medicine. Maybe I should retroactively fit that in. And again, in genetics, it's monogenic disorders. In genomics, it's multifactorial disorders.

And there's a specific DNA diagnosis in genetic disease. You have this disease. As opposed to genomics, it's more you're at risk for the following factors because it's not, in fact, a high penetrance disease. And looking at gene action and genetic disease, and you're looking more at genetic regulation in genetic medicine.

And etiology, specific mutation versus-- because a mutation just tells you there's an error which is associated with a disease whereas in genomic medicine, you're saying, how did this error come to give rise to this disease? And in genetics, it's one species. In genomics, it's several species.

Now I think you could in good faith actually argue that there's a lot of overlap in these two, but that's broadly how the world seems to be divided. And I believe that will be blurred soon. But I think the important take-home message is that in genomic medicine, we really look comprehensively at all data types. And basically, you view the patient, if you want to look at them in a reductionist fashion, as the following kind of matrix. We're looking at environmental data, single nucleotide polymorphisms, proteomic measurements, gene expression measurements, and clinical measurements such as your history, your physical exam, laboratory studies, and imaging studies all across time.

But although there is missing data and although these are incomplete, we can now view the patient as this big time series that you can apply a lot of machine learning techniques to cluster patients together, create new diagnosis and predictive modeling techniques to figure out where in this space of possibilities of these attributes is the patient likely to be in the future. And it's this comprehensiveness that I think characterizes genomic medicine. As we will get to later in the course, I will identify how unfortunately, this is becoming a commodity.

Only genomic measurements are cheap, as I just illustrated you. But this part, because we have to deal with these dumb doctors, it's very expensive because it's labor intensive. Trying to assess the phenotype of the patient is a labor intensive process just as medical care is, just as education is.

And so why do we have to be comprehensive? Other than just liking to be comprehensive, what is our motivation? And the answer is very simple. It comes from an understanding that the phenotype is not just directly just probabilistically linked to some gene.

So in the old genetic medicine, you would be looking at some gene, some marker, some polymorphism on some gene, and say, ah, this gene has a G in position 291, whenever this person has a risk for heart disease. And what you'd really be saying is when this gene has this spelling, it's associated probabilistically with some other gene that you don't know that is directly causing a change in the phenotype. So in genomic medicine, we want to measure this gene but also every other gene that might be contributing to this phenotype because as we well know, even if you stuck in the same room the three individuals with the same polymorphism of that gene, they would actually have different diseases.

If you look at Huntington's disease, Huntington's disease is this awful disease where you lose control of your limbs at age 40 and become demented because of an expanding number of [? CG ?] repeats in the Huntington's gene. The longer that number repeats, the larger the number of repeats, the more acute is the onset, the earlier is the onset. But if you take on average, if you take a set of individuals with the same length of repeats and study them, they have a wide distribution for that given length of repeat and the onset of disease.

And why is that? Well, it's because they have a different genetic background, and these other genes, these other genes involving the neurotransmitters, structural genes having to do with the structure of synapses, and so on, are also influencing it. And therefore, this particular gene's action on the phenotype is conditional on all these other genes. And furthermore, let's not forget about the common environment and the individual environment and cultural factors which can further influence this.

And I'm going to shortly point out to you that this is all too real. So in genetics, just to remind you, genetic medicine, basically the study of what was the probability of a particular phenotype happening given a particular genotype, and genomic medicine would try to assess all the genes directly and capture these other factors and see how they interact. And this creates a fundamental problem, and the question you might ask yourself is, why in genomic medicine are biostatisticians not front and center as part of the process?

After all, in terms of the quantitative analysis of clinical outcomes, for instance, they have been quite omnipresent. And the answer is simply this. In a traditional clinical study like the Nurses' Health Study or the Framingham Heart Study, you would have on the order of tens of thousands of patients and at most hundreds of variables. And there's a lot of tried and true autoregressive models-- for instance, T-tests, parametric and non-parametric tests-- which work extremely well for this kind of data set.

Unfortunately, in genomics, we have quite the opposite. We have 10,000 or hundreds of thousands if you proteomics variables, and we only have, on a good day, hundreds of patients, especially if you're looking at something in the brain. As I explained to you at the beginning, you're only going to get a few hundred brains, let's say, with brain cancer.

And so let me put this in sort of grade school terms. If you have one variable, how many questions do you need to solve it? one.

If you have two variables, how many equations do you need to solve it? Two. And if you have 10,000, how many do you need to solve it? A lot.

If you have fewer equations than that, then it's undetermined. In other words, there's many possible solutions to the relationships of the variables than you have available. And this is so darned undetermined that all these traditional statistical techniques really did not work well. And therefore, frankly, the statistical community until recently were just not interested in this area.

Whereas the computer science community, probably because they're just not really smart enough to understand how difficult a problem this was, just started applying a bunch of techniques that they had applied for other high dimensionality data sets such as vision, face recognition, for instance, and hearing recognition, which also have a high number of features, applied these very same techniques to these data sets. And that just gives you a cultural reason why computer scientists are the bioinformaticians right now the center of genetic medicine rather than statisticians, although they are now coming into play. So let me give you a feeling for the new pharmacology. This was a study that I did with a Atul Butte now four years ago.

The National Cancer Institute has 60 cancer cell lines called the NCI 60, which they test for efficacy a variety of different drugs. And they have, in their banks, thousands of pharmaceuticals. And they have thousands of pieces of dirt, leaf that has been collected around the world that they test-- literally some leaf in the rainforest, a piece of dust in China-- that they test to see how effective that it is in inhibiting these 60 different tumors. And so what we did back then in collaboration with one of my former interns, Todd Golub, we extracted the RNA from these 66 cancer cell lines and measured how much of the RNA was being expressed in these 60 cancer cell lines.

And we obtained from the National Cancer Institute, for a subset of the 50,000 compounds they had tested, 5,000-- let me see how I'm doing for a time-- 5,000 drugs that had been tested against the 60 cancer cell lines and how much they had inhibited the growth of the cells. I'll get to the details of this experiment in-- the common cell details in another lecture. But suffice it to say, this is the overall picture.

You have 60 common cell lines, 6,000 genes, 5,000 anticancer agents. So you have 11,000 variables with only 60 cell lines. That sounds like a miserable thing to do.

And nonetheless, what we were able to do is the following. In this complicated looking diagram, we picked a correlation coefficient between all-- if we computed all the possible correlations between all the 11,000 variables, that gave us on the order of 68 million different correlations. And if you pick the correlation coefficient threshold of 0.8, we found 202 networks that were joining 834 variables-- in other words, only 7% out of the 11,000-- and only 1,200 links out of the 68 million. 1,200 correlation coefficients out of the 60 million were above the threshold.

Only one link was above the threshold between a gene and anticancer agent. And let me just ask you a quick question. Why is it that we found a lot of correlations between drug effect on the cancer cell lines and another drug effect on cancer cell lines? So why did we find a lot of high drug-drug correlations, a lot of high correlations between the inhibitory effect of one drug and the inhibitory effect of another drug? Why did we find a lot of correlations, high correlations?

AUDIENCE: [INAUDIBLE]

ISAAC SAMUEL KOHANE: What?

AUDIENCE: Were they chemically similar?

ISAAC SAMUEL KOHANE: They were chemically similar because in way a lot of drugs are created are by copycat drug discovery, doing small permutations of an existing drug. So of course, they have similar reactions. Why were there a lot of gene-gene correlations?

AUDIENCE: [INAUDIBLE]

ISAAC SAMUEL KOHANE: What?

AUDIENCE: [INAUDIBLE]

ISAAC SAMUEL KOHANE: Not quite. It's more fundamental insight than that.

AUDIENCE: [INAUDIBLE]

ISAAC SAMUEL KOHANE: Right, because basically, in order for the darn cell to live, genes have to have coordinated action. So not all genes, but the genes that make the ribosome, that makes the transcription complex that run oxidation, all have to be synthesized into RNA at approximately same. Time. Otherwise, you're not going to have coordinated action. There's a lot of exceptions to that. But by and large, it's true.

And in fact, the whole area of expression analysis clustering, as I'll get to in another lecture, would not work if that intuition was not true. But God or evolution did not evolve genes and chemotherapeutic compounds together. So there's no particular prior reason why the inhibitory pattern of a drug and the gene expression profile should be linked. And so we were darn lucky to find even one, but we did.

And it turned out that links were this gene and this drug. Before I explain to you what they are, let me summarize what it means. It means that across all these 60 cancer cell lines-- and these were different very different cancers cell lines. They were liver cancer, breast cancer, skin cancer, blood cancers, 60 different cancers.

The more of this gene was expression in the cancer, the more sensitive the cancer cell line was to this chemotherapeutic agent. And it turns out that this gene is L-plastin, also known as LCP1 or Lymphocyte Cytosolic Protein. The drug is this unpronounceable thing.

I mean, a lot of these are just drugs that they obtained essentially wholesale from pharmaceutical companies. But it turns out it's a [INAUDIBLE] carboxylic acid derivative. Those of you who know anything about diabetes know that other [INAUDIBLE] carboxylic acid derivatives are known to inhibit tumor cell growth. And by the way, these are the same class of drugs as the oral hypoglycemic agents, the pills that type two diabetics will take.

And our role for this gene in [? tumorigenicity, ?] the generation of tumors, have previously been postulated. And in fact, subsequent to our publication of this paper in [INAUDIBLE], in fact, a group at the National Cancer Institute was able to show the dose response that we had, in fact, calculated. Yes.

AUDIENCE: So if I'm understanding the description of the [INAUDIBLE] experiment and the results--

ISAAC SAMUEL KOHANE: Yes.

AUDIENCE: --is it safe to say as a generalization that whenever you find a link between a particular [INAUDIBLE] and a gene that [INAUDIBLE] across the [INAUDIBLE] cell lines these are genes that are typically going to have a high conservation rate across the [INAUDIBLE]?

ISAAC SAMUEL KOHANE: No.

AUDIENCE: [INAUDIBLE]

ISAAC SAMUEL KOHANE: In fact, if it had been true that the gene had a high concentration rate across all 60 cancer cell lines--

AUDIENCE: [INAUDIBLE]. In other words, like, this gene data [INAUDIBLE]--

ISAAC SAMUEL KOHANE: Yes.

AUDIENCE: --that this drug is acting on the different cell lines, is acting on this one particular gene. But there probably are differences in [INAUDIBLE].

ISAAC SAMUEL KOHANE: Let's just have our conversation [INAUDIBLE] for the rest of the class. We don't know how it's acting. All we're observing right now is a correlation.

AUDIENCE: OK, right.

ISAAC SAMUEL KOHANE: Right? But I still think that you have a question in there which is still true regardless of what I just said. So the more of this gene expressed, the cancer cell line is going to be provably more sensitive to the chemotherapeutic agents. So you had a question still.

AUDIENCE: Right. Well, it was that it didn't have to do with the concentration of the gene but rather just that the-- just for argument's sake, let's say it's actually acting on [INAUDIBLE].

ISAAC SAMUEL KOHANE: OK, fine. Yeah. Right.

AUDIENCE: [INAUDIBLE]. So let's say it's acting on the gene product and across the cell lines, then it would be a safe assumption that the gene product would be very similar. In other words, there would be this [INAUDIBLE]. Well, there's not-- whereas with other gene products, where there is no correlation between a therapeutic agent--

ISAAC SAMUEL KOHANE: Yeah.

AUDIENCE: --you have a high variability of different types within those cells.

ISAAC SAMUEL KOHANE: I may be misunderstanding your question, and I'll give you back exactly what I believe my misunderstanding is. And maybe it's yours, and we'll clarify it right now. For a given gene, what we're measuring is a gene product, which is the amount of RNA being expressed. And basically, what we know is that the gene product is the same for every cancer cell line.

It's just they don't have a new gene. They have the same complement of genes. But just from the point of view point of view of this experiment, the only thing that's different between the 60 cancer cell lines in a very reductionist point of view, just looking at the gene expression, is how much of each of those genes they're making.

AUDIENCE: So there aren't any [INAUDIBLE].

ISAAC SAMUEL KOHANE: No, you're addressing a whole other layer of complexity that may be very, very true. But we just didn't measure.

AUDIENCE: All right, because I thought that one of the take home messages was that for sort of fundamental gene products within a pathway.

ISAAC SAMUEL KOHANE: Yeah.

AUDIENCE: Those are what you're going to find links to therapeutic agents more than, say, something, you know--

ISAAC SAMUEL KOHANE: Well, we got out of this. So the answer is not in the way that you conceive it because if you're doing-- and by the way, it's a good point because it reminds us again of the distinction I drew at the beginning where that's a very simple diagram about DNA, RNA, and protein. There would be a study where you'd be looking at the genotype and saying-- doing the very same study and looking at all the possible genotypes and saying, which genotypes are the most predictive of this kind of expression?

But here, we're not worrying about individual differences in the genes we're just looking at the amount of the gene being expressed. Nonetheless, if there was, for instance, a direct effect-- in other words, if the more this gene is expressed, the more sensitive the cancer is, short of actually going and understanding the biology, which is the right answer, you still ask yourself a question. So if I make this gene more highly expressed with a vector to make it more expressed, will this cell line become more resistant, more sensitive to chemotherapy? That becomes a therapy in and of itself.

But thanks for asking question. So just now-- and this would give you an idea why in*The Economist* just appeared within last two weeks the following cartoon that was part of an article called "Malignant [? Mass," ?] where they're arguing that actually, coming up with these comprehensive models of which this is a very simple example of susceptibility of cancers to treatment and risk to cancers is actually become so mathematical that we can actually have a whole new class of researchers who are going to take advantage of these models to actually find new therapeutic agents. So let's go to-- we've talked about how much of these data measurement techniques are available today in a large scale.

We've discussed [INAUDIBLE] genetic. Let's now talk about hereditary-- heredity, rather. So the way a population geneticist thinks about heredity in the so-called broad sense with a capital H is the following.

Heredity includes all genetic influences on the phenotype whether due to additive, dominant, or interactive effects. So when you hear in a technical paper heredity, what they mean is the ratio of the variance due to the genotype over the total variance of the phenotype where the variance of the genotype is additive variance, dominant variance, and interactive variance between the genes. That's the technical definition of heritability.

So let's think about that in a concrete fashion. Don't some people just eat and not get fat? Isn't that true?

Isn't it in their genes? In other words, these poor people who just, no matter how little they eat, they're just going to balloon up? It's just tough. And it's really their problem that they have this genetic background that makes them at risk for this obesity. So we've got a real big problem.

So the way to think about it is an experiment that I only dared to do four years ago at MIT. I asked all the Asian students in the classroom to raise their hand. They raised their hand and I told them to leave.

They're upset with me, and they walk out of the room. Then of the remaining students in the class, which, since this was MIT, was about 30% left, I say, take your pulse. And we took a pulse, and I found a nerd in the room.

And I said, fire up Excel, and everybody called out their pulses. And I had the nerd take the average and the standard deviation. Welcomed back the Asian students and went through the same process with them. Everybody takes their heart rate. We tally it, mean, standard deviation.

And then we show them the board. And in fact, there's a more than two standard deviation separation between the Asians and the non-Asians. And I explain to them, as you well know, the basal heart rate of non-Asians is quite a bit lower than that of Asians. And they sort of look puzzled.

Did I know that? Really? Really?

Maybe. I don't know. And then I say, well, is that really right?

And then suddenly it dawns on what's happened here. We're not talking about genes. What are we talking about? What?

AUDIENCE: Work.

ISAAC SAMUEL KOHANE: Work-- we're talking about environment because basically, I pissed them off by doing a potentially racist thing. So the catecholamines were pouring through their veins, and I made them get up. And so I did two major environmental stimuli which changed their physiology.

But if you didn't know that, that hidden variable, you could legitimately come to a conclusion that there's some racial stratification around this phenotype and therefore is with reduction to genes. And you could ask the same thing about the overweight people. You can say, look, they're eating just as much as I am, and somehow, they're getting fat and I'm not.

So obesity is just a massive epidemic, and that's the point. It's an epidemic. This is the prevalence of obesity, note, just from in the last 40 years. Exponential rise.

Now I defy you to find any possible genetic evolutionary evolutionary model that explains how our genome changed so fast in 30 years. And note, by the way, this was the same time as the amount of calories from fat has been steadily decreasing. In fact, some argue that's the cause of it, by the way, because we've substituted that those fat calories with carbohydrate calories.

AUDIENCE: But it's US only.

ISAAC SAMUEL KOHANE: It's US only. By the way-- but there's a lag. Even third world countries are now beginning-- the moment they become caloric sufficient, they actually are showing this.

And so obviously, it's not genetic, and it's actually conditioned on some significant change in our environment most likely having to do with our combination of our exercise patterns and our dietary composition. And yet if you did not understand that, you'd just be hammering on the genome. And it's that kind of reductionism, that we have to avoid.

But if you-- in terms of genetic medicine, not genomic medicine, where you really think comprehensively, you could easily run into this kind of error. And so how do we define environment, diet, daily habits, environmental insults, medical care? And so the genotype does not capture the individual patient states.

It's what I told you before, but I think it's a graphical example of how that's true. So we need to capture and quantify the environmental influences. We need to capture the effect of the genotype and the environmental effects on the phenotype.

And these two comprise-- history, physical, laboratory studies, imaging, which, in fact, you'll all recognize those of you getting your medical training is the basic medical history. And so if we really want to do the comprehensive kind of studies that we need to be able to dissect environmental and genetic interactions, we have to have these data items. And that's why this is part of genomic medicine-- because those of you who are going to be doctors or clinical researchers or researchers studying clinical phenomenon, somehow you're going to have to be able to get this accurately and at large scale. And that's kind of boring, frankly, but nontrivial.

And so this is just repeating what I've said, but I just really want to emphasize it. There's more to the state description than the genome. Given the necessity to capture both environment genetic state and the interaction, it's only then that we'll be able to elucidate the variation of environment due to the genome and through the interaction between the environment and the genome.

For example, you're only going to figure out the risk effects of smoking on lung cancer if you can quantify well the environment, which is both the smoking and the exposure to other harmful chemicals in the environment. And it's required for effective new therapies. It's required for deeper understanding of mechanism, and it requires capturing the aforementioned interaction, and the less we capture, the more undetermined, in the sense that I gave to you, the system is.

Well, we're running out of time. One thing I want to make true for this class is that I always end on time. So we'll just talk about the last bullet, and at another time, I'll talk to you about accelerating consumer activation.

Here's a standard pediatric question. You have here the standard growth curves in height and in weight. And you have a patient who's falling off their height curve and sort of falling off the weight curve.

And in order to diagnose them, you have heights and weights and family history. You can take an X-ray of their wrist. You can measure breast development or size of the testicles.

And the disorders show characteristic patterns on a growth chart. But if you just have the misfortune of sending them to my clinic, we'll also do some other tests. We'll look at thyroid function.

We'll look at a protein called IGF1 that's made in response to growth hormone. We look for inflammation, looking at your blood count, and so on. And if we still have evidence that you're really not going well, then we bring you into the hospital where we expose you to insulin and to glucagon.

And that makes a kid hypoglycemic. So they pass out and feel crummy, and then they get nauseous from the glucagon. So they puke, and then they pay $2,000 for the pleasure of having been with us in the morning.

And the interpretation remains controversial nonetheless because there's a significant false positive rate, and there's a significant false negative rate. And why is that? Because we're really not capturing the underlying process.

And for instance, obese patients sometimes don't quite secrete enough growth hormone. If you're before puberty, sometimes if you don't have enough sex hormones at all, you might not, in fact, secrete any growth hormone even though you're ultimately going to secrete growth hormone. And we're totally going to miss the following thing.

Here at the tip of the X and Y chromosome is a gene called SHOX, Short Stature Homeobox. And those of you who are doctors or medical students, you know what Turner syndrome is? Do you know what Turner syndrome is?

AUDIENCE: [INAUDIBLE]

ISAAC SAMUEL KOHANE: Close. You have the right category. X no Y, and they look like girls.

And they're short. What [? neck? ?] Excellent.

They're short. They only have an X and no Y. And after class, I'll give you the full Turner syndrome so I can end on time.

And they're short. Why are they short? Because they only have one dose of SHOX.

Now it turns out-- because they don't have two X chromosomes. It turns out the most common chromosome is, of course, of short stature in males, and it happened in 2.5% of idiopathic-- that means without a known cause, short children have SHOX mutations, causing them to be short. And these mutants have perfectly normal growth hormone.

But if you treat them with growth hormone, they'll grow to a normal size. And this is known in the literature and just not tested. I actually tested this because of this darn class. I made sure I was fully familiar.

But like I said, in my own clinic, now probably less than 25% of the doctors order this. And yet this is a fully treatable cause of short stature. And unlike the-- and it's not as expensive as the big insulin glucagon test I gave you.

And it's much more certain outcome. If you have this mutation, you're definitely going to be short, and you're definitely going to respond. So this is just a bit of a teaser to show you how today well within my reach, I can order the test today, and I do so.

It's just not done. So this brings us to the other problem, which is education. If there can be thousands of mutations that are clinically significant, how in the heck are we going to teach them to our medical students who are already overloaded? And the answer, I think, and unless you can come up with a better solution, it's one or two-fold.

Either create a whole bunch of experts whose only job is just to look at the patient as a doc and say, do they have X, Y, or Z, or we have to bring information technology to bear right in the process so that as you enter characteristics of patients, all the genetic possibilities start getting updated. And that's a problem because this is a decision tree for the treatment of prostate cancer based on an old proteomic test, namely prostate specific antigen. And basically, what do you do to someone who has basically a big bump in their prostate?

And the problem is, if you look at the literature-- I did this last year. That's 7,325 articles, 7,000 articles-- sorry about the color or lack of it. 7,000 articles are saying, what is the right ordering of decisions based on this one test? So depending on different PSA levels.

Now that's just one proteomic test for one gene. What in the heck are we going to do with these decision trees when we have to measure-- or we will measure, as we can today, thousands of gene expression levels or protein levels makes up for a much more complex tree, and we'll never have enough patients, by the way, to sample all the different possibilities. But that's a fundamental challenge of genomic medicine.

So this is what I was alluded to, and just in the interest of time, I'm going to short circuit it, although this slide will be on the web shortly. If you only had 10 diseases, 10 hypotheses, and you had five binary tests-- so there's not like a prostate-specific antigen. It's not different levels in a continuous fashion and just say it's normal or abnormal.

So five binary tests-- the analysis requires knowing 63,000 probabilities to be able to calculate the probability of all the possible outcomes under all the different tests. And that's just very hard to imagine how we're going to do that comprehensively. And in fact, the number of diseases is much larger than this.

The number of tests that we can do is much larger than that. So that's a fundamental methodological challenge for new medicine. And in the same sense, I'm going to just-- now this will be my last slides. Evidence-based medicine is a big movement in medicine which says, let's look at the literature and try to find out what the best practice.

And they've done this for a number of different disease areas, and there's literally thousands of doctors and researchers who are filling out these databases of evidence-based medicine, the largest being the Cochrane collaborative. But if you look at the genetic section, it's just a few highly penetrant genetic diseases. There's very little to guide-- in fact, there's nothing to guide the practitioner for anything of these genomic diseases for these very common diseases, which is where I stop today's class because we ran out of time.

And for another time, I'll tell you about the problems about patients who are able to select tests, order tests over the web today. You can actually send your own samples to be sequenced and a dubious interpretation or undubious interpretation is given to you. And then you show up in a doctor's office, and he doesn't know or she doesn't know what to do with it. But that'll be for another time. All right, any administrative issues, feel free to--