Today is my last class with you. Awe, I'm sorry, too. You guys are a lot of fun. This has actually been the most interactive 7.

1 I've ever had. Usually there are a couple of people who perk up and say things, but you guys are great because all sorts of people are willing to contribute. So, I've had a wonderful time and it certainly seems like you guys have learned a lot.

What I'd like to do for my last lecture is pick up again a little bit like I did with genomics and try to give you a sense of where things are going. I always like doing this because I get to talk about things that are in none of the textbooks that, well, I mean, it's just stuff that many people working in the field don't necessarily know. And that's what's so much fun about teaching introductory biology is because it only takes a semester for you guys to get up to the point of at least being able to understand what's getting done on the cutting-edge.

Even if you might not yet be able to go off and practice it, you might need a little more experience for that, but you'd be surprised, it's not that much more.

Take maybe Project Lab and you'll be able to start doing it already.

It's really wonderful that it's possible to grasp what's going on.

And, in many ways, you guys may have an advantage in grasping what's going on because, as I've already hinted, biology's undergoing this remarkable transformation from being a purely laboratory-based science where each individual works on his or her own project to being an information-based science that involves an integration of vast amounts of data across the whole world and trying to learn things from this tremendous dataset.

And, in that sense, I think the new students coming into the field have a distinct advantage over those who have been in it.

And certainly the students who know mathematical and physical and chemical and other sorts of things, and aren't scared to write computer code when they need to write computer code have a really great advantage. So, anyway, all that by way of introduction. I want to talk about two subjects today of great interest to me. One is DNA variation and one is RNA variation.

The variation of DNA sequence between individuals within a population, and in particular our population, and the other is RNA variation, the variation in RNA expression between different cell types, different tissues. And the work I'm going to talk about today is work that I, and my colleagues, have all been involved in. And it's stuff I know and love.

So, feel free to ask questions about it. I may know the answers, but what's reasonably fun about these lectures is

if I don't know the answers it's probably the case that the answers aren't known.

So, that's good fun because it's stuff I really do know well, and I love. So, anyway, here's some DNA sequence. It's pretty boring.

This is a chunk of sequence from, let's say, the human genome.

How much does this differ between any two individuals?

If I were to sequence any two chromosomes, any two copies of the chromosome from an individual in this class or two individuals on this planet, how much would they differ? The answer is that much.

That's the average amount of difference between any two people on this planet. Not a lot. If you counted up, it is on average one nucleotide difference out of 1, 00 nucleotides on average, somewhat less than one part in 1, 00 or better than 99.9% identity between any two individuals. Now, that is a very small amount, not just in absolute terms, 99.9% identity is a lot, but in comparative terms with other species. If I take two chimpanzees in Africa, on average they will differ by about twice as much as any two random humans. And if I take two orangutans in Southeast Asia, they will on average differ by about eight times as much as any two humans on this planet.

You guys think the orangutans all look the same.

They think you all look the same, and they're right. So, why is this?

Why are humans amongst mammalian species relatively limited in the amount of variation?

Well, it's a direct result of our population history.

It turns out that the amount of variation that can be sustained in a population depends on two things. At equilibrium, if population has constant size N for a very long time and a certain mutation rate, Mu, you can just write a piece of arithmetic that says, well, mutations are always arising due to new mutations in the population and mutations are being lost by genetic drift, just by random sampling from generation to generation.

And those two processes, the creation of new mutations and the loss of mutations just due to random sampling in each generation, sets up an equilibrium, and the equilibrium defines an equation there, Pi equals one over one plus four and Mu reciprocal which equation you have no need to memorize whatsoever and possibly even no need to write down. The important point is the concept, that if you know the number of organisms in the population and you know the mutation rate, those set up the bounds of mutation and drift, and you can write down how polymorphic, how heterozygous random individuals should be at equilibrium.

That is if the population has been at size N for a very long time.

Well, the expected amount of heterozygosity for the human population -- Sorry. For a population of size 10, 00 would be about one nucleotide in 1300. We have exactly the amount of heterozygosity you would expect for a population of about 10, 00 individuals. Yeah, but wait, we're not a population of 10,000 individuals. Why do we have the heterozygosity you would expect from a population of 10, 00 individuals? We're six billion.

It's a reflection of our history.

Because remember I said that was the statement about what the population heterozygosity should be at equilibrium?

We haven't been six billion people except very recently.

The human population has undergone an exponential expansion.

It used to be a relatively small size, and then it very recently underwent this huge exponential expansion. If you actually write down the equations, the amount of variation in our population was determined by that constant size for a very long time.

And then a rapid exponential expansion that's basically taken place in a mere 3, 00 generations, it's much too rapid to have any affect on the real variation in our population.

What do I mean by that? What's the mutation rate per nucleotide in the human genome? It's on the order of two times ten to the minus eighth per generation. In a mere 3,000 generations, a tiny mutation rate like two times ten to the minus eighth is not going to be able to build up much more variation.

So you might as well ignore the last 100,000 years or so.

They're irrelevant to how much variation we have.

The variation we have was set by our ancestral population size.

Now, don't get me wrong. Eventually it will equilibrate.

A couple million years from now we will have a much higher variation in the human population as a function of our size, but the population variation we have today is set by the fact that humans derive from a founding population of about 10, 00 individuals or so.

And that means that the variation that you see in the human population is mostly ancestral variations, the variation

that we all walked around with in Africa. And, in fact, that makes a prediction. That would say that if most of the variation in the human population is from the ancestral African founding population then if I go to any two villages around this world, in Japan or in Sweden or in Nigeria, the variance that I see will largely be identical.

And that prediction has been well satisfied.

Because when you go and look and you collect variation in Japan or Sweden or Africa and you compare it, 90% of the variance are common across the entire world. Most variation is common ancestral variation around the world, and only a minority of the variance are new local mutations restricted to individual populations.

This is so contrary to what people think because there's a natural tendency to kind of xenophobia, to imagine that world populations are very different in their genetic background.

But, in point of fact, they're extremely similar.

So, anyway, there's a limited amount of variation.

That's why we have such little variation in the human population.

Now, that variation, humans have a low rate of genetic variation.

Most of the variance that are out there are due to common genetic variance, not rare variance. If I take your genome and I find a site of genetic variation at the point of heterozygosity in your genome, what's the probability that somebody else in this class also is heterozygous for that spot? It turns out that the odds are about 95% that someone else in this class will also share that variance.

So that the variance are not mostly rare, they're mostly common.

And it turns out that some of this common variation, that is most of this variation is likely to be important in the risk of human genetic diseases. So human geneticists have gotten very excited about the following paradigm.

If there's only a limited amount of genetic variation in the human population, actually, if you do the arithmetic, there are only about ten million sites of common variation in the human population, where common might be defined as more than about 1% in the population. There are only ten million sites.

Folks are saying, well, why not enumerate them all?

Let's just know them all, and then let's test each one for its risk of, say, confirming susceptibility of diabetes or heart disease or whatever? After all, ten million is not as big a number as it used to be. We now have the whole sequence of the human genome. Why not layer on the sequence of the human genome all common human

genetic polymorphism?

Now, that's a fairly outrageous idea but could be a very useful one.

Some of these variance are important, by the way.

We know that there are two nucleotides that vary in the gene apolipoprotein E on chromosome number 19. Apolipoprotein E is also an apolipoprotein like we talked about before with familiar hypercholesterolemia. But, in fact, it turns out that apolipoprotein E is expressed in the brain. And it turns out, amongst other tissues, that it comes in three variances, the spelling T-T, T-C and C-C at those two particular spots.

And if you happen to be homozygous for the E4 variant, homozygous for the E4 variant, you have about a 60% to 70% lifetime risk of Alzheimer's disease. In this class 13 of you are homozygous for E4 and have a high lifetime risk of Alzheimer's.

And it would be fairly trivial to go across the street to anybody's lab and test that. Now, I don't particular recommend it, and I haven't tested myself for this variant because there happens to be no particular therapy available today to delay the onset of Alzheimer's disease. And, therefore, I don't recommend finding out about that. But a number of pharmaceutical companies, knowing that this is a very important gene in the pathogenesis of Alzheimer's disease, are working on drugs to try to delay the pathogenesis using this information. And it may be the case that five or ten years from now people will begin to offer drugs that will delay the onset of Alzheimer's disease by delaying the interaction of apolipoprotein E with a target protein called towe, etc. So, this is an example of where a common variant in the population points us to the basis of a common disease and has important therapeutic implications.

There are some other ones, for example. 5% of you carry a particular variant in your factor 5 gene which is the clotting cascade.

It's called the leiden variant. Those 5% of you are going to account for 50% of the admissions to emergency rooms for deep venous clots, for example. The much higher risk of deep venous clots. And, in particular, there are significant issues if you have that variant and you are a woman with taking birth control pills. Some of you were at higher risk for diabetes, type 2 adult onset diabetes.

There's a particular variant in the population that increased your risk for type 2 diabetes by about 30%. 85% of you have the high-risk factor, so you might as well figure you do.

15% of you have a lower risk, et cetera. And one I'm particularly interested in here, it turns out that HIV virus gets into cells with a co-receptor encoded by a gene called CCR5.

Well, it turns out that if we go across the European population, 10% of all chromosomes of European ancestry have a deletion within the CCR5 gene. If 10% of all chromosomes have that deletion then 10% times 10%, 1% of all individuals are homozygous for that deletion. Those individuals are essentially immune to infection from HIV. They are not susceptible. It's not through immunity, it's through lack of a receptor.

Yes? You certainly can. It's not hard. It's a specific known variant.

You could test for it. Absolutely.

Now, of course, that only helps the 1% of people who have that variant. But what it did do was point to the pharmaceutical industry that the interaction between the virus and that variant is essential. And now companies are developing drugs to block the interaction with that particular protein.

And that tells you that it's an important protein. Yes?

Over the whole world?

I just specified European population for that one.

That one, interestingly, is not found at as high a frequency outside of Europe, and no one knows why, whether that might have been due to an ancient selective event or a genetic drift. By contrast, the apolipoprotein E variant, at that frequency of about 3% of people being homozygous and being at risk for Alzheimer's, is about the same frequency everywhere in the world. So, there's a little bit of population variation in frequency. Now, the HIV variant is found elsewhere but at considerably lower frequencies there.

And that's an interesting question as to what causes that variation.

So the notion would be, I've given you a couple of interesting examples, but, look, there's only ten million variants. Just write them all down.

Make one big Excel spreadsheet with ten million variants along the top and all the diseases along the rows, and let's just fill in the matrix and then we'll really, you know, this is the way people think in a post-genomic era. Now, could you do something like that? You would have to enumerate all of the single nucleotide polymorphisms, or SNPs we call them, single nucleotide polymorphisms. Now, to give you an idea of the magnitude of this problem, as recently as 1998, the number of SNPs that were known in the human genome was a couple hundred.

But then a project has taken off. In 1998 an initial SNP map of the human genome was built here at MIT that had about 4, 00 of these variants. Then within the next year or so an international consortium was organized here and

elsewhere to begin to collect more of these genetic variants.

The goal was going to be to find 300, 00 of them within a period of two years. In fact, that goal was blown away and within three years two million of the SNPs in the human population were found.

And as of today, if you go on the Web, you'll find the database with about 7.8 million of the roughly ten million SNPs in the human population already known. Now, that isn't all ten million.

And it takes a while to collect the last ones, you know, collecting the last ones are hard, but we're already the hump of knowing the majority of common variation in the human population.

Not just a sequence of the genome, but a database that already contains more than half of all common variation in the population.

So, we could start building that Excel spreadsheet.

Now, it turns out that it's even a little bit better than that because if we look at many chromosomes in the population, here are chromosomes in the population, it turns out that the common variance on each of those chromosomes tend to be correlated with each other. If I know your genotype at one variant, like over at this locus, I know your genotype at the next locus with reasonably high probability. There's a lot of local correlation. So, instead of looking like a scattered picture like that, it's more like this.

If I know that you're red, red, red you're probably red, red, red over here. In other words, these variations occur in blocks that we called haplotypes. Here's real data.

Across 111 kilobases of DNA there's a bunch of variants, but it turns out that the variants come in two basic flavors.

98% of all chromosomes are either this, this, this, this, this or this, this, this, this, this.

Then there tends to be sites of recombination that are actually hotspots of recombination where most of the recombination of the population is concentrated. And you get a couple of possibilities here. So, the human genome can kind of be broken up into these haplotypes. Blocks that might be 20, 30, 40, sometimes 100 kilobases long in which within the block you tend to have a small number of haplotypes, or flavors as you might think of them, that define most of the chromosomes in the population.

So, in fact, I don't actually need to know all the variants.

If they're so well correlated within a block, if I knew this block structure I would be able to pick a small number of

SNPs that would serve as a proxy for that entire block of inheritance in the population. So, what you might want to do is determine that entire haplotype block structure of hwo they're related to each other, and pick out tag snips.

And it turns out that in theory, a mere 300,000 or so of them would suffice to proxy for most of the genome. So, you might want to declare an international project, and international haplotype map project to create a haplotype map of the human genome.

And indeed, such a project was declared about a year and a half ago through some instigation of scientists and a number of places, including here. And this is $100 million project involving six different countries. And, it is already more than halfway done with the task, and it's very likely that by the middle of next year, we will have a pretty good haplotype map, not just knowing all the variation, but knowing the correlation between that variation, being able to break up the genome into these blocks. By the next time I teach 701, I should be able to show a haplotype map of the whole human genome already. That will allow you to start undertaking systematic studies of inheritance for different diseases across populations.

And in fact, people are already doing things like that.

Here's an example of a study done here at MIT like this, where to study inflammatory bowel disease, there was evidence that there might be a particular region of the genome that contained it, and haplotypes were determined across this, and blah, blah, blah, blah, blah, blah, blah. And this red haplotype here turns out to confer high risk, about a two and a half or higher risk of inflammatory bowel disease.

And it sits over some genes involved in immune responses, certain cytokine genes and all that. And, things like this have been done for type 2 diabetes, schizophrenia, cardiovascular disease, just right now at the moment, a dozen or two examples.

But I think we're set for an explosion in this kind of work.

In addition, you can use this information to do things beyond medical genetics. You can use it for history and anthropology as well. It turns out rather interestingly, that since the human population originated in Africa and spread out from Africa all the way around the world arriving at different places in different times, you can trace those migrations by virtue of rare genetic variants that arose along the way, and let you, like a trail of break crumbs, see the migrations.

So, for example, there are certain rare genetic variants that we can see in a South American Indian tribe, and we can actually see that they came along this route because we can see that residual of that.

In fact, we can do things with this like take a look at Native American individuals and determine that they cluster

into three distinct genetic groups that represent three distinct migrations over the land bridge.

And, you can assign them to these different migrations.

You can do this on the basis of mitochondrial genotype, etc. You can also, for example, determine when people talk about the out of Africa migration, there's now increasing evidence that there really were two, one that went this way over the land, and one that went this way following along the coast into southeast Asia.

And, it looks like we're now beginning to get enough evidence of these two separate migrations by virtue of the genetic breadcrumbs that they have left along the way.

So, it's really a very fascinating thing of how much you can reconstruct from looking at genetic variation, both the common variation that allows us to recognize medical risk, and the rare genetic variation that provides much more individual trails of things.

None of this is perfect yet. There's lots to learn. But I think anthropologists are finding that the existing human population has a tremendous amount of its own history embedded in pattern of genetic variation across the world. You can do other things.

I won't spend much time on this. Well, I'll take a moment on this, right? There's some very interesting work of a post-doctoral fellow here at MIT named Pardese Sebetti who has been trying to ask, can we see in the genetic variation in the population, signatures, patterns of ancient selection, or even recent selection in the human population? Now, hang onto your seats, because this will get just slightly tricky.

But, hang on. It's only a couple of slides. Here was her idea.

You see, when a mutation arises in the population, it usually dies out, right? Any new mutation just typically dies out. But, sometimes by chance it drifts up to a high frequency. Random events happen. But it usually takes a long time to do that. If some random mutation happens, and it happens to drift up to high frequency with no selection on it, then on average it takes a long time to do so.

If you want, I could write a stochastic differential equation that would say that, but just take your gut feeling that if something has no selection on it and it's a rare event that'll drift up, when it drifts up it's kind of a slow process. It was a slow process.

Then over the course of time that it took to drift to high frequency, a lot of genetic recombination would have had to have occurred many generations. And the correlation between the genotype at that spot and genotypes at other loci would break down.

And there would only be short-range correlation. So, in other words, the amount of correlation between knowing the genotype here and the genotype here, maybe allele A here and a C here.

That is an indication of time. It's a clock almost. It's like radioactive decay, right, that genetic recombination scrambles up the correlations. And, if something's old, the correlations go over short distances. But suppose that something happened.

Some mutation happened that was very advantageous.

Then, it would have risen to high frequency quickly because it was under selection. If it did so quickly, then the long-range correlations would not have had time to break down, and we'd have a smoking gun. A smoking gun would be that there would be a long-range correlation around that locus, much longer than you would expect across the genome.

Things even out of this distance would show correlation with that, indicating that this was a recent event.

So, we just measure across the genome, and look for this telltale sign of common variance that have very long range correlation that indicate that they're very recent. So, a plot of the allele frequency, common variance, sorry, if something has a common high frequency and long-range correlation, you wouldn't expect that by chance.

So, something that was common in its frequency and had long-range correlation would be a signature of positive selection. So anyway, Pardise had this idea, and she tried it out with some interesting mutations, some mutations that confer resistance to malaria, one well-known mutation causing resistance to malaria called G6 PD and another one that she herself had proposed as a mutation causing resistance to malaria, variants in the CD4 ligand gene.

And to make a long story short, both the known and her newly predicted variant showed this telltale property of having a high frequency and very long range correlation.

Well that's very interesting because she was able to show that each of these mutations probably were the result of positive selection.

But what you could do in principle is test every variant in the human genome this way: take any variant, look at its frequency, and compare it to the long range correlation around it, and test every single variant in the human population to see which ones might be the result of long range correlation. Now, when she proposed this, this was about a year and a half ago or two years ago, this was a pretty nutty idea because you would need all the variants in the human population, and you would need all this correlation information. But in fact, as I say, that information's almost upon us, and I believed that this experiment, this analysis to look for all strong positive

selection in the human genome will in fact be done in the course of the next 12 months.

So, I'm hoping by next year I can actually report on a genome-wide search for all the signatures of positive selection.

Now, this doesn't detect all positive selection.

It will detect sufficiently strong positive selection going back pretty much only over the 10, 00 years. When you do the arithmetic, that's how much power you have. Of course, 10,000 years has been a pretty interesting time for the human population, right? The time of civilization and population density, and infectious diseases, and all that, and I think we'll have an interesting window into the change in diet. All of that should come out of something like this. So, there's a lot of really cool information in DNA variation to be had. All right, that's one half. The other half of what I would like to talk about is totally different. It's not about inherited DNA variation. It's about somatic differences between tissues in RNA variation. So, let's shift gears.

RNA variation: let me start by giving you an example here.

These are cells from two different patients with acute leukemia.

Can you spot the difference between these? Yep? More like bunches of grapes and all that. Yeah, it turns out that's just a reflection of the field of view you have if you move over to look like that. But I mean, that's good.

It's just that it turns out that that isn't actually a distinction when you look at more fields. Anything else? Yep? White blood cells like different. They look broken. There's more of them in this field of view. But you look at 100 fields of view and it turns out that's not either. Well, the reason you're having trouble spotting any difference is that highly trained pathologists can't find any difference either. I generally agree there's no difference between these two if you look at enough fields of view.

But you can convince yourself if you look that you see things there.

But these actually are two very different kinds of leukemia.

And, these patients have to be treated very differently.

But, pathologists cannot determine which leukemia it is just by looking at the microscope, it turns out. This is the work of this man, Sydney Farber, namesake of the Dana Farber Cancer Institute here in Boston, who in the 1950s began noticing that patients with leukemias, some of them seemed different in the way they responded to a certain treatment, and he said, look, I think there's some underlying classification of these leukemias, but I can't get any reliable way to tell it in the microscope.

And he put many years into working this out, first by noticing certain difference in enzymes in the cells, and then people noticed certain things in cell surface markers, and some chromosomal rearrangements.

And nowadays, there are a bunch of test that can be done by a pathologist when a patient comes in with acute leukemia to determine whether they have AML or ALL. But it turns out that you can't do it by looking. You have to do some kind of immunohystochemical test of some sort in order to do that.

So this is a triumph of diagnosis. After 40 years of work, we can now correctly classify patients as AML or ALL. And they get the appropriate treatment. And if they don't get the right treatment, they have a much higher chance of dying.

And if they do get the right treatment, they have a much higher chance of living. So, this is great.

There's only one problem with the story. It took 40 years, 40 years to sort this out. That's a long time. Couldn't we do better? Surely these cells know what they are.

Surely we could just ask them if they are. Well, here's the idea.

Suppose we could ask each cell, please tell us every gene that you have turned on, and the level to which you have that gene expressed. In other words, let us summarize each cell, each tumor by a description of its complete pattern of gene expression to 22,000 genes on the human genome.

Let's write down the level of expression, X1 up to X22, 00 for each of the 22,000 genes of the genome. So, ever tumor becomes a point in 22, 00 dimensional space, right?

Now clearly, if we had every tumor described as a point in 22, 00 dimensional space, we ought to be able to sort out which tumors are similar to each other, right? Well, it turns out you can do that now. These are gene chips, one of several technologies by which on a piece of glass are put little spots, each of which contains a piece of DNA, a unique DNA sequence. Actually, many copies of that DNA sequence are there. Each of these is a 25 base long DNA sequence, and I can design this so whatever DNA sequence you want is in each spot. The way that's done is with the same photolithographic techniques that are used to make microprocessors.

People have worked out a chemistry where through a mask, you shine a light, photodeprotect certain pixels; the pixels that are photodeprotected you can chemically attach an A, then re-protect the surface. Use a light. Chemically photodeprotect certain spots. Wash on a C. And in this fashion, since you can randomly address the spots by light, and then chemically add bases to whatever spots are deprotected, you can simultaneously construct hundreds of thousands of spots each containing its own unique specified oligonucleotide sequence.

And you can get them in little plastic chips.

And then if you want, all you do is you take a tumor.

You grind it up. You prepare RNA. You fluorescently label the RNA with some appropriate fluorescent dye. You squirt it into the chip.

You wash it back and forth. You rock it back and forth, wash it out, and stick it in a laser scanner. And it'll see how much fluorescence is stuck to each spot. And bingo: you get a readout of the level of gene expression. I guess each spot, you should design it so that this spot has an oligonucleotide complementary to gene number one. And the next one, an oligonucleotide matching by Crick-Watson base pairing complementary to gene number two and gene number three.

So, if I knew all the genes in the genome, I could make a detector spot for each gene in the genome. And of course we know essentially all the genes in the genome. So you can make those detector spots and you can buy them. So, you can now get a readout of all the, I mean, this is like so cool because when I started teaching 701, which wasn't that long ago because I ain't (sic) that old still, the way people did an analysis of gene expression is they used primitive technologies where they would analyze one gene at a time, certain things called northern blots and things like that, right? And, you know, you'd put in a lot of work and you get the expression level of a gene, whereas now you can get the expression of all the genes simultaneously, and it's pretty mind boggling that you can do that. How do you analyze data like that?

So, we still use northern blots. It's true. So, every tumor becomes a vector, and we get a vector corresponding to each tumor. So, this line here is the first tumor, the second tumor, the third tumor, the fourth tumor.

The columns here correspond to genes. There are 22, 00 columns in this matrix, and I've shown a certain subset of the columns because these genes here have the interesting property that they tend to be high red in the ALL tumors, and they tend to be low blue in the AML tumors, whereas these genes here have the opposite property. They tend to be low blue in the ALL tumors and high red in the AML tumors. These genes do a pretty good job of telling apart these tumors.

So, here's a new tumor. Patient came in. We analyzed the RNA, squirted it on the chip. Can somebody classify that? Louder?

AML. Next? Next? Congratulations, you're pathologists. Very good. That's right, you can do that.

It works. And in fact, in the study that was done that was published about this, the computer was able to get it right 100% of the time. Not bad. So now you say, wait, wait, wait, but you're cheating.

You're giving it a whole bunch of knowns. Once I have a whole bunch of knowns it's not so hard to classify a new tumor.

What Sydney Farber did was he discovered in the first place that there existed two subtypes. Surely that's harder than classifying when you're given a bunch of knowns. And that's true. So, suppose instead, I didn't tell you in advance which were AML's and which were ALL's, and I just gave you vectors corresponding to a large number of tumors, do you think you would be able to sort out that they actually fell into two clusters?

Could you by computer tell that there's one class and the other class? Turns out that you can. Now, I've made it a little easier by not listing most of the 22,000 columns here.

But think about it. Every tumor is a point in 22, 00 dimensional space. If some of the tumors are similar, what can you say about those points in 22,000 dimensional space?

They're going to be clumped together. They're near each other.

So, just plot every tumor as a point in 22,000 dimensional space, and your question is, do the points tend to lie in two clumps up in 22, 00 dimensional space? And there's simple arithmetic you can learn using linear algebra to get some separating hyperplane and ask, do tumors lie on one side or the other? And, it turns out the procedures like that will quickly tell you that these tumors clump into two very clear clumps. They're not randomly distributed. And so, if you get these tumors, and you do gene expression on them and put the data into a computer, the amount of time it takes the computer to discover that there were actually two types of acute leukemia is about three seconds marked down from 40 years. That's good. So, you can reproduce the discovery of AML and ALL in three seconds. Now you know what the pathologists say about this. They say, oh, give me a break.

It's shooting fish in a barrel. We know there was a distinction.

Big deal that the computer can find the distinction.

We knew that there was distinction there. I know the computer didn't know it and all that. Tell us something we don't know.

That's a fair question. So it turns out that you can ask some more questions. You can say, suppose I take now just the ALL's. Are they a homogeneous class, or did they fall into two classes? It turns out that extending this work, folks here were able to show that we can further split that ALL class. There was a hint that you might be able to do so because there's some ALL patients who have disruptions of a gene called MLL.

And this tends to be a little more common in infants, and tends to be associated with a poor prognosis.

But it was really very unclear whether this was simply one of a zillion factoids about some leukemia patients, whether this was a fundamental distinction. So, what happened was folks took a lot of ALL patients, got their expression profiles, and lo and behold it turned out that ALL itself broke into two very different clusters. This is an artist's rendition of a 22,000 dimensional space. We can't afford a 22,000 dimensional projector here, so we're just using two dimensions.

But, the two forms of ALL were quite distinct from each other, and so actually ALL itself should be split up into two classes, ALL plus and minus, or ALL one and two, or MLL and ALL.

And it turns out that these forms are quite different.

They have different outcomes and should be treated differently.

It also turns out that a particularly good distinction between these two subtypes of ALL is found by looking at this particular gene called the flit-3 kinase. The flit-3 kinase gene, whatever that is, was of great interest because people know that they can make inhibitors against certain kinases. And so, it turned out that an inhibitor against flit-3 kinases, against this flit-3 kinase gene product.

If you treat cells with that inhibitor, cells of this type die, and cells of this type are not affected. So in fact, there's a potential drug use of flit-3 kinases in the MLL class of these leukemias, and folks are trying some clinical trials now. So, not only did the analysis of the gene expression point to two important sub-types of leukemias, but the analysis of the gene expression even suggested potential targets for therapy. So, I'll give you a bunch more examples. I have a bunch more examples like that there.

They are examples of taking lymphomas and showing that they can be split into two different categories, examples of taking breast cancers into several categories, colon cancers.

Basically what's going on right now is an attempt to reclassify cancers based not on what they look like in the microscope, and based not on what organ in the body they affect, but based on, molecularly, what their description is, because the molecular description, as Bob talked to you about with CML and with Gleveck, turns out to be a tremendously powerful way of classifying cancers because you're able to see what is the molecular defect and can make a molecular targeted therapy.

So, these sorts of tools are quite cool, and I've got to say, in the last year we've begun using these expression tools not just to classify cancers, but to classify drugs.

We've begun an interesting and somewhat crazy project to take all the FDA approved drugs, put them onto cell

types, and see what they do, that is, get a signature, a fingerprint, a gene expression description of the action of a drug.

And then we hope, here's the nutty idea, that we can look up in the computer which drugs do which things and might be useful for which diseases, because we'd put the diseases and the drugs on an equal footing. All of them would be described in terms of their gene expression patterns. So, I'll tell you one interesting example, OK? This is an interesting enough example. I don't even have slides for it yet.

It turns out that these patients with ALL that I've been talking about, some of the patients with ALL will respond to the drug dexamethasone. Some won't. If you take patients who respond to dexamethasone, and patients who are resistant to dexamethasone, and you get their gene expression patterns, you can ask are there some genes that explain the difference?

And you can get a certain gene signature, a list of, say, a dozen or so genes that do a pretty good job of classifying who's sensitive and who's resistant. Then you can go to this database I was telling you about of the action of many drugs and say, do we see any drugs whose effect would be to produce a signature of sensitivity? If we found a drug X, which when we put it on cells turned on those genes that correlate with being sensitive to dexamethasone, you could hallucinate the following really happy possibility that when you added that drug together with dexamethasone, you might be able to treat resistant patients because that drug could make them sensitive to dexamethasone, and that you could find that drug just by looking it up in a computer database. So, we tried it and we hit a drug.

There was a certain drug that came up on the screen, yes? That's very much in the idea too. We found a drug that produced the signature sensitivity, and tested it in vitro. In vitro, if you take cells that are resistant and you add dexamethasone, nothing happens because they're resistant. If you add drug X, nothing happens. But if you add both drug X plus dexamethasone, the cells drop dead. It's now going into clinical trials in human patients. It turns out drug X is already a well FDA approved drug, so it can be tested in human patients right away, so it's going to be tested.

So, the gene expression pattern was able to tell us to use a drug which actually had nothing to do with cancer uses in a cancer setting because it might do something helpful.

Now, what's the point of all this? We can turn up the lights because I think I'm going to stop the slides there. The point of all of this, which is what I've made again, and I will make again, because you are the generation that's going to really live this, is that biology is becoming information. Now, don't get me wrong. It's not stopping being biochemistry. It's going to be biochemistry. It's not stopping being molecular biology. It's not stopping any of the things it was before. 45:57 But it is also becoming information, that for the first time we're entering a world where

we can collect vast amounts of information: all the genetic variants in a patient, all of the gene expression pattern in a cell, or all of the gene expression pattern induced by a drug, and that whatever question you're asking will be informed by being able to access that whole database. In no way does it decrease the role of the individual smart scientist working on his or her problem.

To the contrary, the goal is to empower the individual smart scientist so that you have all of that information at your fingertips. There are databases scattered around the web that have sequences from different species, variations from the human population, all of these drug database, etc., etc., etc., etc. It's a time of tremendous ferment, a little bit of chaos. You talk to people in the field, they say, we're getting deluged by data. We're getting crushed by the amount of data. I don't' know what to do with all the data. There's only one solution for a field in that condition, and that is young scientists because the young scientists who come into the field are the ones who take for granted, of course we're going to have all these data.

We love having all these data. This is just great, couldn't be happier to have all these data. We're not put off by it in the least. That's what's going on. That's what's so important about your generation, and that's why I think it's really important that even though it's 701 and we're supposed to be teaching you the basics, it's important that you see this stuff because this is the change that's going on, and we're counting on this very much to drive a revolution in health, a revolution in biomedical research, and we're counting on you guys very much to drive that revolution. It has been a pleasure to teach you this term. I hope many of you will stay in touch, and some of you will go into biology, and even those of you who don't will know lots about it and enjoy it. Thank you very much. [APPLAUSE]