Good morning. Good morning.

I don't know about you, but I can't take too many more nights like this.

I confess, I haven't gotten a thing done for so many nights in a row now, but what a game! How many of you saw the game? Excellent.

Very good, very good. You have your priorities straight in the world. Very good. Well, if it's possible to get your minds off Curt Schilling last night, and off more importantly tonight.

Perhaps we can spend a bit of time this morning in the meanwhile with whatever spare neurons you have talking about recombinant DNA for a bit, OK? What we talked about last time was different ways to clone your gene based on its properties. We started off with cloning by complementation, right, the idea that if you took a library of clones, you would be able to put it into bacteria and select a bacterium whose phenotype had been restored by virtue of having the plasmid. You would complement the defect.

You'd find the clone you wanted because it complemented the defect.

That's great if you can put it into an organism that has a defect.

You can do it with bacteria. You can do that with yeast.

It's harder to do with large organisms because you can't inject enough of them with different clones to be able to make that practical unless you're working in cell culture or some very small, fast growing organism. We talked about being able to use a protein sequence, reverse translating that protein sequence in the computer from amino acid sequence to nucleotide sequence, and using the nucleotide sequence to design a probe to hybridize back to the genome. That works fine if you have a protein sequence.

But the last topic we talked about that I wanted to just touch on again this morning was suppose you were trying to clone the gene that causes a certain human disease, and you have no idea what the protein was. Then, you can't use its amino acid sequence because you don't have the protein. What can you possibly do when all you know is that you have a gene which causes a genetic defect that causes a disease? And I said you could clone it using the ideas of genetic mapping, position, the things that Sturtevant developed. And, I touched on it briefly, and I want to just touch on it a bit more because some people had some questions about it. And I've set up a very simple example to show you. Suppose that, to make it easy, we're working in a fruit fly first. We're working drosophila, and suppose that the true picture of the underlying chromosome is like this.

There's a locus that could either have a mutant allele M or the wild type allele plus. There's a bunch of other loci

along the chromosome. And, let's suppose we know all of where they are and all that. And, they have two alternative alleles. At this locus the alleles are orange or pink. At this locus I'll call the alleles orange or pink.

Now, these are different loci. These are different alleles. I've just called them orange and pink in both cases so I don't have a rainbow of colors up here to confuse us. But all I mean is there's two possible alleles here, two alleles here, two alleles here.

This is the diseased gene we're interested in, and these are passive markers. These are other markers along the chromosome. If we were to set up a cross between heterozygotes, a heterozygote here, and a heterozygote here, and it were the case that on the chromosome bearing the mutant allele, it happened that at these three markers we had orange alleles. I don't know what they are, but whatever these orange alleles are, they might be a visible phenotype, forked or yellow or bristled. They could be a DNA sequence difference. They could be whatever you want, but let's suppose the M chromosome has a set of alleles that are different in each location than the plus chromosome. Then, when we look at the offspring that come out of this cross, let's only, for the sake of simplicity, look at those offspring who are homozygous mutants.

Well, in general, if there's been no crossover here, then the M chromosome will have orange, orange, orange, orange, orange, orange. If there's been a crossover, however, it could go orange, orange, pink on one of those chromosomes. Or if there's been crossovers like this, it could go orange, orange, pink on one chromosome, and orange, pink, pink on the other chromosome. It could even, in the extreme, have had crossovers very close to the gene maybe here, and even maybe here. And you've got orange, pink, pink, and pink, pink, pink.

But if we look at the many segregates, you know from genetic mapping that the closer the locus is to the disease gene, the more strongly correlated the inheritance will be, the tighter the linkage will be. This is nothing more than linkage mapping. But now, suppose we were doing linkage mapping, but for the sake of argument the whole genome had already been sequenced. Suppose the genome had been sequenced in a cross, and the whole genome of the fruit fly had been sequenced which it has been sequenced.

And, we looked at a cross and we looked at the mutants.

And what we did was we tried different positions along the genome.

And at each position, we had some genetic marker.

And that genetic marker might be as simple as the fact that at that position, maybe there is an A in the DNA sequence on one of the chromosomes, and maybe I don't know a G in the other sequence.

And over here, this marker might be, there's a T in some particular position, and there's a C in some particular

position.

If we could assay that, if we could tell, we could look whether this spelling variation is closely correlated with the mutant.

And this spelling variation is closely correlated with the inheritance of the mutant allele. And we could just try up and down the genome, different sites of spelling difference as if they were genetic markers in our cross because they are genetic markers in our cross, and see which one is most tightly correlated.

The minute we get any genetic sequence difference, that shows co-inheritance linkage in this cross, we know that this spot in the genome must be nearby our mutation.

So, we'll try one closer, and we'll try one on the other side.

And, what you do is you test sites of genetic variation, first to find one that shows any co-inheritance.

And once you've got that, you try ones closer, and closer, and closer. Last time I talked about the process of, if you had one of those markers you could use it to isolate the next clone and the next clone and the next clone. But you know what I realized? That's so old fashioned. We might as well deal with the fact we have a sequence of the genome. No more would you ever isolate the next clone and the next clone and the next clone.

You just look it up in the computer. So, even if you have the whole sequence of the genome, we have to figure out what part of it was co-inherited along with this disease, and that's the way you do it, OK? Genetic mapping, ust as Sturtevant invented it, can be applied if you have a whole sequence of the genome, and enough sites of variation. And, I've drawn it for a fruit fly cross, but this could equally well be cystic fibrosis.

The only difference is if we're doing this in human families and it's cystic fibrosis we don't have as many offspring.

So, we have to pool data from many families. And, we can't arrange it so that every family has exactly the same orange alleles up here and pink alleles down there, but computers can deal with that. They can still figure out the correlation across many families, and you find the spot in the genome where for many, many, many families the kids who all got the disease show correlated inheritance with this marker. And that eventually pins you down to a region of the genome. It pins you down to those genetic markers that show the absolute tightest correlation, tight correlation, and that's where you look.

And in that fashion, people went being able to map the location of Huntington's Disease in 1984 to, by now, mapping the locations of more than 1, 00 different human genetic diseases where people didn't know the protein in advance. They did it entirely based on this positional mapping. So, Sturtevant's idea, which I like so much, has

played itself out so beautifully now in the area of modern molecular medicine. OK. So, onward.

I want to talk about a few other variations on the theme rather quickly, and then I think I want to talk about how you analyze your clones. First, variations on cloning, I should just at least mention it. We talked about cloning in an autonomously replicating plasmid in a bacteria.

So, you go to a bacteria. They have some autonomously replicated pieces of DNA. There are circles. You can clone in them, and you can typically, these things are on the order of, I don't know, 1,000 to 2,000 to 5, 00 bases can be readily cloned in these plasmids. You can do more, but that's a typical kind of number is the insert size, typically. But we in the lab go up to much higher numbers like 10,000 sometimes. You can also, if you wanted to study yeast, it turns out yeast happily have plasmids as well, and you can do a similar sort of thing for yeast.

It turns out that instead of using plasmids, you can use bacterial viruses. These bacterial viruses have all different shapes as we've talked about, circular or linear, and they can typically hold, oh, 15, 00-40,000. Some of these viruses are quite big.

The bacteriophage lambda tends to carry a lot of stuff.

And, it can replicate. So, you could do the same thing to that. You can even use viruses that infect mammalian cells and there are all sorts of viruses now that people clone in again, linear or circular. I don't know, for mammalian cells, you often, the viruses like 1,000-5, 00. You can even make artificial whole chromosomes now. You can do this in yeast.

Artificial chromosomes are called YACs. They have all the little machinery, little telomeres on them, little centromeres. They have a selectable marker, and then you can clone into it your piece of DNA. And these can take up to a million bases of DNA.

So, if you wanted, there are bacterial artificial chromosomes.

They're called BACs if they're in bacteria. And recently, people have developed artificial chromosome systems for mammalian cells, and specifically human cells. And they're called unfortunately MACs and HACs and things like that. Basically, any molecule that can replicate in any system, some smart molecular biologist will come along and say, how do I use that for my purpose, to stick my DNA in it, and get it to replicate in this organism?

And so, if something's not on this list, it will be soon, OK? Now, here's another thing. This is cloning chunks of DNA.

Just to have the piece of DNA in a library, but suppose we want to do more than just have the DNA sitting there in

the bacterium, suppose what I'd really like to do is take a bacterium, E coli, and put it to work for us. Maybe what I'd like to do is take a plasmid and insert in that plasmid the gene for human insulin.

So, I'm going to take the DNA locus corresponding to human insulin, clone it into my plasmid. Maybe I'll have isolated it from my library because, let's see, insulin's protein sequence is known so I could reverse translate it to a nucleotide sequence. So, I could probe a library.

So, I could find the clone that has insulin. Now what I'd like to do is persuade this bacteria not just to carry the DNA but to make insulin for me. Would that be useful? Yeah, how did people used to get insulin? Cadavers, dead bodies; it would be much easier to get them from a fermenter, right, to get insulin from a fermenter, if you could just ask E coli to make it.

So, if we put it into E coli, will it make insulin for us?

Here's the human locus, DNA for insulin. Will it make insulin? Let's see, how do you make a protein?

You've got to start by making RNA, right? You've got to transcribe the gene. Will E coli transcribe this gene?

Well, why? It's got a promoter, right? It's got the insulin promoter. There we go. The insulin promoter is here.

So, E coli will come along to the insulin promoter and start making RNA? No, it turns out that promoters in humans and promoters in bacteria are sufficiently different. They don't work across species.

They won't recognize the human promoter. Too bad. Any ideas?

Yep? Stick a bacterial promoter there. Good, you're acting like a good molecular biology designer here. Let's put a bacterial promoter here.

It will recognize its own promoter. That's great. Then, let's put the DNA for the human insulin gene here. And now, maybe we'll put the Lac operon, and when it has lactose it'll start making RNA from the human insulin gene. And it'll start translating it.

And, we get insulin. Any problems? Well, will it make any, for starters? What's another aspect of mammalian genes that's different from bacterial genes?

Processing, what kind of processing with the RNA? And the splicing, ooh, the insulin gene has introns that have to be spliced out.

So, this is going to make some RNA, insulin RNA, and it needs to be processed like this. Will bacteria carry on our splicing for us? They don't do splicing. Yep? Well, that's a very interesting question because we haven't. But, what

do you propose? You see, I've just taken a piece of human DNA from the human genome, which encodes the introns and the exons. But, you seem to have a solution to our problem, and what would that be? So, instead of making a library of genomic DNA, what you're suggesting is a radical idea.

Let's instead take human RNA. Here's some human RNA, lots of human RNA, a big collection of human RNA. What was at the end of the human RNA: a poly(A) tail. And what I understand you to be suggesting is if we take human mRNAs, a whole collection of them, you want me to turn these mRNAs back into DNA and clone them instead of using the chromosomal DNA. How do I turn an RNA back to DNA?

Is that possible? What do you use: reverse transcriptase.

We have to give it a primer. So remember, five prime to three prime, we'd like to put a primer going over here.

Any ideas for a good primer? Poly(T), isn't that convenient?

One of the reasons that mammalian messages have poly(A) tails is so that we are able to reverse transcribe them using poly(T) primers. No, that's actually not true. So, we use reverse transcriptase. And what we can do is we'll copy this RNA into a strand of DNA. There we go.

Then what we'll do, next step, is we'll take the DNA, and we'll copy back into a second strand of DNA.

And now, we have double-stranded DNA whose sequence matches the already-processed mRNAs. Sorry? So, the sequences would match the mRNAs. So what you could do is instead of taking human DNA from the nucleus, you could take RNAs, turn them back into DNA by reverse transcriptase, and make a library now that consists of zillions of inserts, each of which has what's called a cDNA, a copied DNA, copied back from the RNA. The great advantage of this is that the human cell has already done the splicing, and so there are no introns left. Now, when you stick it in a bacterium, the bacterium is able to express this. It's able, if you give it its own bacterial promoter, to make an RNA.

And if you don't ask the bacteria to have to splice, if you just give it a pre-spliced piece of DNA that doesn't need splicing, it can translate that DNA. Now, notice we used all of our tricks. You had to know about reverse transcriptase, poly(A) tails, structures of genes, introns, exons, yes, question?

It doesn't. You do this in the test tube. You purify human mRNA in the test tube. You take that mRNA in a test tube, add reverse transcriptase, add poly(T), make this reaction of RNA to DNA in the test tube go back.

Where does it come from? Viruses that copy themselves back for a living, right? So, again, every single thing we're using comes from some living organism that does this kind of stuff. And, when I teach you about the facts of

how viruses replicate or what the structure of mRNAs look like or whatever, it's because every bit of knowledge we get about the way biology works turns into an incredibly powerful tool as it's turning out for us to actually be able to further study biology. So, great. So, where does reverse transcriptase come from now? Originally they come from viruses that turn themselves back from RNA to DNA. Now, how do you get reverse transcriptase? Catalog, right, very good. All right, so this is called, finally, a cDNA library. And, if you had made a cDNA library, you would be able to screen the cDNA library to find the gene for insulin.

Is this useful? This happens to be, for example, one of the consequences of this was the biotechnology industry. OK, so if you have any doubts about the usefulness of understanding these abstract things about E coli and bacteria and stuff like that, one of the consequences was Genentech, Biogen, and Amgen, and if you just simply walk around Kendall Square, within a mile of this place you will see laid out before you the consequences of this ability, OK? It's transforming Cambridge. Yes?

And the world. Yeah. Indeed.

It might be that producing large amounts of insulin was bad for the bacteria because there would be so much protein it would clump and kill the bacteria. It might be that insulin, for various reasons, might not fold appropriately in the bacterial environment.

And, this is why the biotechnology industry has lots of smart people working in it because you're totally, 100% right. You might decide that instead of cloning it in bacteria it's better to clone it in some insect cell in culture which, in fact, people like to work with, or some other cell, or a mammalian cell. And so, I simplify by saying put it in coli, but in fact that might test six different cell lines, six different host possibilities.

They might have to take the insulin out and refold it in vitro and things like that. You're totally right.

This is actually something that requires work to do it right, just like building an airplane requires work.

I could tell you Bernoulli's principles, but then Boeing does more than just writes down Bernoulli's principles.

OK, so onward. Now, I'd like to turn next to analyzing your clone.

Analyzing the clone, so suppose we have, maybe it's by positional cloning, maybe it's by cDNA cloning, but one way or the other we've got us a clone that we're very interested in.

Maybe it has the insulin gene. Maybe it has the Huntington's disease gene. Whatever it is, we're going to want to study it.

And at the moment, I haven't told you how I would even read its DNA sequence or analyze its DNA. So, the first

step is, of course, I have to purify the plasmid. And, it turns out that that can be done.

There are simple biochemical techniques, as I mentioned in a previous lecture, that allow you to grow up a lot of the bacteria, crack them open, and the plasmid being a little circle, and being a little more tightly super-coiled and wound up has somewhat different physical properties. And you can use those to purify the plasmid. So, plasmid preps are not hard to do. You can get a fairly pure collection of the plasmid.

Now, suppose I've done this for, oh, I don't know, let's take my first example, orange mutants. Suppose I tried to rescue bacteria that were orange minus, and suppose I found that 50 different plasmids rescued my orange mutant because I transformed a lot of plasmids in, I plated it, and 50 colonies grew up.

Are they all the same thing or are they different?

Is there any quickie way to take a look at these 50 plasmids and see if they're identical or fairly close, or obviously different? Well, I'd like to take some way to take the DNA from the plasmid and analyze it kind of easily. I might want to see, like, how big is the insert? Right, that'd be one way, if they had different sized inserts so they couldn't be the same thing.

So, maybe what I could do is how do I clone this? I used EcoRI sites I recall. So, I have EcoRI sites here. Suppose I were to take this DNA, and I were to now cut the DNA from the plasmid with EcoRI.

Then, what I would get is two separate molecules.

I would get the vector and the insert. How could I see how big they were? Gels, gel electrophoresis is the way to do that. So, I take a gel. A gel is a slab of gelatin, Jell-O, OK, and normally it's laid flat, but I'm going to do it vertically here. I load into the top of it here a little bit of my DNA, this whole mixture. I take the plasmid. I cut it. I put it in here. DNA's positive charge or negative charge? Negative. So, where should I put the positive pull? On the bottom, well done.

That's often not done, and to the detriment of the experiment.

If you put the positive pull here, it goes the wrong way, and everybody has to do that at least once. So, what'll happen is the DNA fragments move through, and the smaller fragments move faster than the big fragments, right? If something's little, it'll move fast. If something's big, it moves slowly: little, big.

Smaller moves faster because it wiggles through the little pores in the gel better. So, suppose I were to do this for a bunch of plasmids, and what I saw was this.

First order, what do you guess? Sorry? Top road's probably the plasmid vector. This is probably the vector, and

what do I know about the inserts? At least two inserts, at least two distinct inserts. Now, if I wanted to be sure that was the vector, maybe what I could do is take another row, and run a known amount of the vector, take the vector alone and I could check that the vector alone runs over here. And maybe I might take some other known molecules. These would be called molecular weight standards. So, if I run some knowns in one of the lanes of the gel, I can even measure and say, ah-ha, the insert is somewhere between the size of this one and the size of that one. And so, I get a little ruler that I can put on the gel. So, in fact, that's the first thing you would do is you digest your clone that way.

Now, does the fact that these guys have exactly the same, apparently, size on the gel mean that they're the exact same piece of DNA? No, because you can't even actually tell it's exactly the same.

There's a limit to how precisely you can measure it.

So, what else could you do? You could try another restriction enzyme. It turns out that since there are so many restriction enzymes in the catalog, if I take a piece of DNA, maybe that Eco fragment, I could try cutting it with HinDIII.

And when I cut it with HinDIII, I'm going to get three distinct lengths. I could try cutting it with, oh, I don't know, pick another enzyme, BamHI. When I cut it with BamHI, I'll get some other lengths. And, how to get these lengths by adding these, by running them out on a gel and looking at their sizes.

What if I added both HinDIII and BamHI to my test tube?

I'd cut at both sites. So, I'd cut here, here, here, here, here. So, this is cut with HinDIII, here cut with BamHI, here cut with both and I could measure these lengths. So, suppose I gave you this as a computer problem, I have a string and it's an unknown string, and I cut it at two places and I get these lengths, X1, X2, X3. And then I take that same string and I cut it at other positions, Y1, Y2, and Y3 are the lengths that result. And then suppose I now cut it at both of the sites, and I measure it, and I get Z1, Z2, Z3, Z4, Z5. If I gave you all those numbers, could you figure out where the sites must be? Probably. It turns out to be a reasonably doable computer problem, although it can get a little hard in places. And you could try a third enzyme and a fourth enzyme, and it's a cute exercise to write yourself a little piece of code that will figure out where the sites are based on the lengths. The reason it occasionally gets funny what if Z3 and Z4 are exactly the same length and they run on top of each other in the gel, and there are special cases.

But you can kind of reconstruct where those restriction sites must be just by writing a good piece of code that'll put these pieces together. This is called restriction mapping, and it's great fun. Everybody likes to do this once.

But, it's only a limited amount of information, right, because you get where the sites are, and I guess if I gave you ten clones and they all had exactly the same restriction maps, the exact same positions of these restriction sites, you'd feel pretty confident they were the same clone.

But you still wouldn't really know much about the clone other than it had two HinDIII sites and two BamHI sites, and here's where they were.

What do you really want to know about this clone? It's DNA sequence, right? Let's not settle for anything less than the exact nucleotide sequence of the clone. So, that's really the last key topic is sequencing DNA.

How are you going to sequence DNA? Well, suppose I give you some double strand of DNA, five prime to three prime, five prime, three prime, double stranded DNA.

Let me heat it up. What happens when I heat up DNA?

It melts the hydrogen bonds, the non-covalent hydrogen bonds here break, and I got my two strands separated. Now, what I'd like to do is I want to start reading out this DNA sequence.

So, I'm going to make me a primer. Now, golly, here's a primer.

You're going to ask me, how did I even know what primer to use if I don't know the DNA sequence? How can I make a primer?

Hold that question. Make sure I remember to come back and answer that, OK? But for the moment, grant me that I have a primer here.

What I'd like to do is add DNA polymerase. So, let's add some DNA polymerase. And, I'd like to add nucleotide triphosphates, dNTPs, the dATP, dCTP, the dGTP, dTTP, and if I add DNA polymerase and I add my nucleotides, what does Arthur Kornberg tell us will happen? It'll start polymerizing, right? And, it'll stop there. So, the polymerase knows the bases, right? It knows what base to put in because polymerase is very smart.

So, the bases get put in correctly. The only problem is, how do we get polymerase to tell us what it just did?

Here's a cute trick. This is, by the way, a cute trick that won the Nobel Prize. So, suppose my primer is like this: five prime, T, A, A, T, T, C, T, and the template strand here, A, T, T, A, A, G, A, now let's keep going, A, T, G, C, C, A, A, T, G, G, A, T, T, A, five prime. So, there's my primer.

There's my template. I'm going to start adding. Well, let's add our polymerase. Let's add our dNTPs, polymerase, dATP, dCTP, dTTP, dGTP, and then I want to add a special extra good old ingredient into this.

The special extra ingredient I want to add is a defective T, a defective dTTP. What do I mean by defective? I mean chemically modified in such a way that it can't be extended, that you can't extend past it. So now, let's follow my reaction.

I'm going to start with, I'm just going to write them down here, T, A, A, T, T, C, T. What's the next base I'm going to put in?

T, OK? Is that a defective T or a good T? I don't know.

It could be. Maybe it's a defective T, which I'll put a little star there, OK? If so, what happens to my polymerase?

It stops. It can't go any further. It can't go any further because the T's defective. But what if it wasn't a defective T?

What if it was a good T? Then what goes on? The polymerase will put in, keep going guys. A, C, G, G, and what does it put in now? T, right? Now, is that a defective T? Maybe. We don't know.

If it is a defective T, it stops there. Otherwise, polymerase goes here, and the next space is what?

T, and is that a defective T? Maybe.

And, if it's not a defective T, then polymerase goes on, puts in an A, puts in a G, a C, C, and then a T.

And maybe that's defective. All right, which of these possibilities is what polymerase does when I throw it in?

Well, all of them. There's a lot of molecules there.

Some of the molecules, by chance, happen to install a defective T, and they grind to a halt here. Sometimes, a good T's put in and the molecules stop here. Sometimes they stop here, and if I start with a big collection of primers in a lot of my template DNA, I'm going to get this whole collection of different molecules of different lengths. What lengths do I get?

The lengths correspond precisely to the positions of the Ts.

I get a series of molecules whose lengths perfectly match the positions of Ts. Well, first off, how do I measure their lengths? Run a gel, bingo, run a gel.

So, if I could run a gel that could separate nucleotides based on length that two next to each other, another one up there, I'd see a small molecule, length one, two, three, four, five, three, six, eight, I'd see one of length eight. I'd see one of length, what's the next one, 13, eight, nine, ten, 13, 14, so eight, nine, ten, 11, 12, 13, 14, 15, what's that, 13, 14, 15, 16, 17, 18. OK, those would be the positions at which I would see this T. So, I'd need to have a

special kind of gel that's so accurate that it can separate single nucleotides, right, that the lengths, but that can be done.

There's acrylamide, the polymer that will do that.

That'll tell me the exact lengths of the T's. What else do I do?

Well, let's obviously do it from the other bases.

Let's try defective A, defective C, defective G.

Let's see, if I got it right, which we'll try, it ought to end up looking something like that. And if not, you get the picture, that this ought to match up as to which columns have which lengths.

OK, I think I got it right. That tells me the lengths of the molecules. So, I could read off at sequence.

The sequence of that molecule ought to be, starting over there, the sequence of what I've added in, ought to be something like T, A, C, G, G, T, T, A, C, C, T, yep, it worked.

It's exactly right. Bingo. I can now read the sequence.

Fred Sanger, a brilliant scientist, thought up this method of just exploiting E coli's own polymerase or other organism's own polymerases.

So, copying and all the chemistry that had to be done was thinking up a defective nucleotide that could not be extended.

It could obviously be inserted. It can't be extended. So, one question is, what's a defective nucleotide? Well, you will recall that our nucleotide in the sugar phosphate chain is sitting like this. Let's see, hanging off the one prime carbon is the base. This is the one prime carbon, the two prime carbon, the three prime carbon, the four prime carbon, the five prime carbon. What do we know in DNA at the two prime carbon?

Normally in ribose there would be a hydroxyl here, right? But in deoxyribose, there's just a hydrogen.

So, if this is deoxyribose, so a dNTP really means a two prime deoxyribose, where do I now attach my next base in the sugar phosphate train? Three prime ends, and what do I attach it to: the OH.

What do you think would happen if there's no OH there?

You're stuck. All you've got to do is take off that hydroxyl.

No hydroxyl group. If you made nucleotides that don't have that hydroxyl group, they can't be extended.

So, instead of these being just deoxy at the two prime position, they are dideoxy, deoxy at two positions.

They are two prime, three prime, dideoxynucleotides.

That's it. Now, if you needed to get two prime three prime dideoxynucleotides, they're in the catalogue of course, right, because Fred Sanger had to make them himself and all that, but you can just buy them now. And so, you can do the sequence.

A few other little details here, though, guys.

How do we see the DNA and the gel? One possibility would be staining it. There are some dies like ethidium bromide, and for doing your restriction mapping, using a dye that sticks to DNA like ethidium bromide does is pretty good. And then you put it under fluorescent light and you look. For sequencing, the amount of DNA is so little that it's hard to see with a dye by the naked eye, which is what you do with restriction map. So, sorry?

So, the first thing people did was radioactive. What they did was they took a primer, made it radioactive, and you did this whole sequencing reaction with radioactive primer.

Then, when you run the gel, you take your gel and you expose it for some number of hours, eight hours maybe, a piece of x-ray film, develop the x-ray film, and you'll see that picture. So, one solution that you could do to visualize is using radioactive nucleotides. So, we got the defective nucleotide.

We now need to visualize our DNA. Let's visualize the sequence.

One possibility: radioactive. The second possibility, someone already mentioned it, a fluorescent dye.

Now, here, a fluorescent dye could be put on, and you can't read it with your eye, but lasers are very good at reading.

So, you might run a whole gel here and have lasers scan it.

But, you can actually do better than that. Suppose I put my fluorescent dye on my dideoxynucleotides.

Suppose I put it on my dideoxynucleotides, and suppose I even had enough chemistry at my disposal that I could put a different color of fluorescent dye on each of my nucleotides. Then, whenever the dideoxy is put in to terminate the chain, it carries with it its own color.

Wouldn't that be cool? And, that's what's done. Not just can you buy dideoxynucleotides now, but you can buy the

four different dideoxynucleotides each with its own dye attached to it.

So, there are di-dideoxies I guess, sorry, but it's different di's, right? They're dye-dideoxies. So, you could do that.

And then what you get would be that in this column you get this color.

And in this column, you'd get this color. And in this column you'd get this color, etc. I'm not worrying about where they are here. And they'd all be different colors and it would be very pretty. You know what? Why do we need to run separate lanes anymore? If we got a laser, we can tell the laser scan it to tell it different. Stick it in one way. In fact, what's done is stick it in a capillary tube, throw in all four at the same time now, and as these fragments come by, each has its own color. And all we need is a laser scanner capable of sitting right here. Here's my laser scanner. And the laser scanner, positive here, negative here, as the DNA flows by through this polymer, the laser scanner reads off which colors just went by.

And it goes A color, C color, T color, G color. That's it. So, there are actually machines now that have 96 different capillaries.

These are called capillary tubes. And you can have 96 of them with laser scanning across, and in each column now, it turns out that you can read almost 1,000 letters, 1,000 bases per column per capillary times about 100 capillaries.

Or in other words, you can read out about 10^5 bases of information.

You can read out 10^5 bases of information in about two hours.

Of course, you can do that ten times a day. So, you can actually read out 10^6 or about a million bases of information per machine. And here at MIT, we have 100 of these machines. So, we actually can read out a little shy of 100 million letters of DNA sequence per day, which I mean is a lot.

We read about 40 billion letters per year here at MIT, and this is how we do it. How much does a machine cost?

List, or do you want a deal? They list for $300, 00, but if you buy in bulk, I can do better. [LAUGHTER] We buy it in bulk, by the way. So now, how are we going to get our primer there? That was the only little bit we were missing is where did our primer come from? The last little detail: here's my vector, remember, and I want to sequence this insert.

How am I going to get a primer in the insert? I don't know what its sequence is. How do I even start this? Sorry? Well, but that won't tell me what the sequence is that I have to, I mean, I was looking to try to get a primer that matches the insert.

And I don't know what the insert is. So, how am I going to get a primer?

Oh, I know the vector. The vector is well known.

It sequence is in the catalog. Let me instead just use a primer that happens to sit in the vector, and I'll match to a known sequence to start with, and then I'll sequence into my unknown territory. So, this is how you get the initial primer was you arrange that your initial primer is sitting in known vector sequence. All right, so you can now sequence DNA. I've got to say, I've taught this course for a little more than a decade, and being able to say, now we can routinely sequence about a million letters per machine, and 100 million letters per day, and things like this was not routinely the case. When we started teaching this course, I was describing what we di