

What we've talked about in recombinant DNA so far is how to get a piece of DNA from somewhere and make a whole lot of copies of it. So, instead of working with DNA extracted from my cells, which there's 3 billion different base pairs of sequence, we can find a little stretch of DNA. But simply being able to clone it, it was a long way from him being able to get a lot of a fragment of DNA to be able to figure out what that sequence is. That sequence of that gene is actually from the [zero derma? gene, the gene that's broken and the zero derma pigmentosa variant patience, missing one of these translesion DNA polymerases that can copy over a thymine, thymine, [permadine? ], [dymer?], this induced by UV light. So the reason they have that problem with their skin after sunlight is because they are missing a polymerase that can't copy over accurately over this very common lesion caused by DNA damage. But how do you get from having a piece of DNA and having the sequence? So, the first thing that people learn to do, and you still do this all the time in any molecular biology lab. We're sort of switching now in engineering as you'll see. You're going to see in the next things that I'm going to say, proteins that were, I talked to you about because of their biological roles. DNA polymerase is, ligases, we learned what they do. And now you're going to see them used in manipulative ways. Restriction enzymes: they have a biological purpose, too. They weren't put on Earth for me to cut up into fragments and clone. They were there to give the bacteria is some kind of primitive immune system. But the first thing that you often have to do, and you have, let's say, a plasmid into which we've inserted a fragment. And let's say it was the kind of cloning we described the other day, where I had cut to the vector with an ECO-R1 site, and the other DNA with an ECO-R1 site. So these, the junction between the inserted fragment and the cut vector has re-created two ECO-R1 sites now. And if I cut with ECO-R1, I'll just undo what I did in the cloning, and we should get the vector DNA and the insert DNA back again. So, I can go from this vector to give us an orientation. I'm going to imagine that it has one more restriction site. This one is called [sal 1? . It just recognizes a different sequence. So, if I take the plasmid DNA and cut with ECO-R1, I just reversing the cloning. So I should get is the vector DNA and the insert that I generated in the first place. But I have to detect them somehow. Unfortunately, they don't just look in the test tube like that. So, what people do is they use a very simple principle. It's called gel electrophoresis. And the idea is you just make a gel of something. In this particular case, it's just made of agarose, which is agarose. These are polysaccharide products, often derived from seaweed or something like that. They have the property that if you warm them up, they're liquid, and that if you let them cool down, they're a gel. You've run into Jell-O which has the property. That's actually made of a protein rather than a carbohydrate. But it's that kind of principle. So it's very easy to pour something and that let it solidify. Now you've got a slab. And it's just a network of things that interact. And the principle of the thing is that, so you have to get the molecules to move. Well that's pretty easy with nucleic acids because they're charged. They have all those phosphates. They've got a lot of negative charge. So if you apply an electric field, they'll move. And the principle of the thing is that if you're big, it's harder to wiggle through this network than if you are small. Or you can think of a big, fat person trying to go through a forest with a lot of trees, and a little skinny one. And if we let them have a race, eventually the skinny one will emerge from the forest first. And so, if we had a set of markers down on the side where this is big, and this is small, and we take this piece of DNA, we're going to get two fragments. The bigger one would be the vector and the smaller one would be an insert. So from that, we could say, oh, if I didn't know what I started with I could say what must be at that plasmid? Is the vector? And I can run at all by itself and see it's exactly the same size. And I got an insert of this particular size. And now, can I learn anything more about that just using restriction enzymes? And let's say I now take the same. Actually, maybe I'll do it over here. So let's cut, this time, with ECO-R1 plus another restriction enzyme. They all have these weird names: [bam H1?], and let's see what happens. Well suppose I do that and I get something like this. Well it looks like the vector wasn't cut at all. That still seems to be the same, but it looks as though the insert got cut into two pieces. Since it was linear, it must have one site in it. And so, this molecule that I cloned could look, be one of two kinds of ways. It could be like this. Let's say this is the insert. Here's the ECO-R1. I'll use this sal 1 to orient us. So, the bam site could either be close over here. Or it could be over on the other side. Does that make sense? The logic is pretty simple. How can I tell which of those is correct? Just doing the kind of stuff I'm doing. Beautiful, beautiful. So if we cut with the sal 1 plus the bam H1. In one case, one would get a fragment like that. In the other case, get a fragment like that. That should feel uneasily familiar to you. It should feel just like what we are doing what we did that phage cross, and we had some genes that were lined up. And we were trying to figure, was the orientation this way? Or was the orientation that? That was exactly the same principle. And so this is usually, in the lab you'd call this restriction mapping, or making a restriction map. And it enabled people to manipulate fragments of DNA and make inferences about their orientation and other features before we can actually even sequence DNA. And that's just part of routine sort of stuff you do a lab. The equipment is disarmingly simple. It looks something like that, usually you're putting some colored dye so you can see that the things are moving down the gel. And the way you visualize the DNA is you add a molecule. The name of it doesn't particularly matter. It's called ethidium bromide. But its property is it doesn't fluoresce when it's just in solution. But it's a flat molecule, and it can interpolate in between the base pairs in DNA. They have all those stacked base pairs going down a helix. This molecule's flat, edit likes to slip inside. And now it's a much more hydrophobic environment. It's hidden from the water, becomes florescent. And so, DNA that's soaked up this dye then will fluoresce when you put a UV light on it. So if I take the gel out of there after I've run it, and soak it in this dye and then shine a little handheld UV light on it, it would look something like that if I photographed it. And so, you would end up with those patterns that look exactly like that. Oops, I guess I took the other one out. But you can, of course, depending on how complicated it is, you could have a lot of different fragments. OK, so the next big thing that had to happen in order for us to really move to where we are in today's molecular biology was somehow, DNA had to be sequenced. And as I say, when I was an undergrad, or even when I was just about to start, when I was a postdoc anyway, just again it seemed like how would you ever do it? Because every nucleotide was joint by a phosphodiester bond. The only difference was the

base that was there. It seemed very, very difficult. It was hard to imagine you would ever be able to sort out the sequence of a billion base pairs. Of course, you could clone. Now you've got maybe a fragment of DNA that's a couple hundred base pairs long, and at least the problem becomes smaller. Maybe you could work it out. Now, there were a couple of different ways of doing it. One was by Wally Gilbert, who's up at Harvard who got half the Nobel Prize for doing this. The other principle, the other one that's proved to be most generally useful is Fred Sanger from England. And he had Wally shared the Nobel Prize for discovering sequencing. And the principal was disarmingly simple. I think it's one of these great ideas do you look back at afterward and think, I could have thought of that. You guys already know everything you need to invent how to sequence DNA. I've told you all the stuff already. But nobody's come down to tell me that you've got it. And I didn't think of it. So here is the principal. What we've talked about, if we take a DNA polymerase plus the four deoxynucleotide triphosphate's, remember we talked about deoxyribonucleotide, the adenosine triphosphate, and so on. There's four different ones. And we take a primer. And there's a three prime hydroxyl right there. And so this is the other strand is going the opposite direction. If we add that, I think you all know what's going to happen. We're going to get an extension to the other end. And what happens every time we add a nucleotide is that three prime hydroxyl attacks the phosphate of the triphosphate. We lose two of the phosphates. This is called pyrophosphate, and we've created a new five to three prime linkage. That gives us a new three prime hydroxyl, and we repeat the process, right? That's what we talked about. So, what would happen, let's spike in a little, let me do it. It's a little deoxy TTP. So this is dideoxy. But what would we mean by that? Well, if this, remember where the deoxy came from? The ribose has at the two prime position has a hydrogen instead of a hydroxyl, and at the three prime position it has a hydroxyl. If we made a dideoxy, what we do is we'd make that. What could that nucleotide do? Well, as long as the polymerase thought it was useful it would use this end, it would have its triphosphate up here. So, somebody else's three prime OH could come down and form a bond to here and we'd lose this. So it could get incorporated. That chain is finished. It can't be elongated anymore. So, let's think what would happen if we had, let me stretch this out a little bit here, and let's imagine we had a few A's in the sequence. So, we are just going to spike it a bit. So, most of the things will not see a dideoxy. So, this primer will put, we'll try elongating this. So when we get to this point, this point many of them will put it an ordinary A, but a few will put it a dideoxy. And those will finish. At that point, they can't go any farther. The rest of them keep going, [the various?] nucleotides. When we get to the next A, most of them will put them a good T, but the ones that put in a dideoxy will stop, and they will generate a fragment that looks like that. You get the idea. Out of this reaction, we are going to get a set of fragments. And each one terminates where there was an A up there. Now, in this newer emulation of this thing, we have a T. And the trick is to put a dye that you can attach to this nucleotide, so it has a particular color. So, suppose we had something that was yellow. Then this particular set of fragments would be yellow. And maybe you can begin to see what would happen now. If we did the same game three more times, each time using a different deoxy, next time maybe we'll use dideoxy A. And we'll put a different colored dye on it. Then every time, in this case we come to a T in the template, it would stop, and we'd get a little fragment that's stopped because it incorporated a dideoxy A, and those would be, let's say, green. So, by the end of this, we would have all possible fragments if we mixed them all together and the last nucleotide on each fragment would say who it was by its color. So, if you were to, then, take this whole mixture of DNA fragments and you run them down a gel, in this case it's a difference of polyacrylamide gel because you have smaller trying to get things to go by smaller fragments. You could sort of see what would happen. The big ones would be at the top. The small ones would be at the bottom. And you'd see each band would have a different color depending on the dideoxy that terminated its chain. So, if you had a little scanner that just goes along, it can read this, and it will print out something. And these are always slightly idealized. This is a real one. But this is the sort of stuff you get back. If you send a piece of DNA over to a sequencing center, they'd send this back as a file or something. And you'd sit there. And it's very good these days. The technology wasn't as good, but they can almost always now get the sequence. Occasionally, you'll get something like a run of G's that gets a little hard, but what they'll do is they'll sort of what they call sequencing [bow strands?]. You can see this way, but really only looking at the information of one strand. So, if we took the other strand. So if we took the other strand, and we did the same thing, but we should get the complementary piece of information. So, what this DNA sequencing allows you to do, then, is determine the exact sequence of nucleotides in some kind of piece. And much of the art from the rest of it then comes, how do you assemble all of those things together? In the case of a bacteria or something, it wasn't so bad because its DNA was small enough. You could cut it into a bunch of sort of big fragments, and then take each one of those, and then the sorting problem was relatively simple. In the case of something like humans, it was really complicated because there were so much more DNA. And the other thing is higher organisms such as yourselves have a lot of repeated DNA. It's just the same sequence, and sometimes there's quite a bit of it, a bunch of repeats. And so, if you see that at the end of your thing, you don't really quite know where you are in the genome. So a lot of other tricks had to be brought into play, including knowledge of the human genetic map. And so you could get yourself anchored at various places because you do on this particular piece of DNA, because it was associated with some gene, had to be here on the chromosome. And therefore, things at least decide beside it were there on the chromosome. And there were a whole lot of tricks to putting it together. But the very basic principle of how we sequence DNA has at its heart the same process that I was talking to you about as when we were doing DNA replication, except in this case it's just used in a very clever way. And that was an amazing idea. It got a Nobel Prize, and you've been sitting here for the last month with all the knowledge to do it. You keep emphasizing that you've got to have that three prime hydroxyl. But some of the great ideas often when you look back you could see it was the hurdle was kind of small. And they didn't even have to do this with dyes at the beginning. In fact, that was a later innovation. The key thing was just the dideoxies stopping in each place. I was lucky enough to live through some of this, the development of this technology. OK, so I've got one more really big thing to tell you, which again was extraordinarily clever, but extraordinarily simple once you heard about it. And it was one more technological advance. It wasn't a big insight into biology in and of itself. but it was a technology

that opened up just incredible experimental possibilities. And it's something known as the polymerase chain reaction. And this allows, in principle, someone like me to go and to grab a single cell from you, take it to DNA, and get a copy of any gene I want from your genome. And I can look and see whether you have any mutations in that genome, or whether there are different polymorphic alleles in the population, in which one you've got from your mom, or which one you got from your dad. So, you take from a single DNA molecule, I can make as much as I want. And this is just like DNA sequencing. You guys already know everything you need to know to invent this technique as well. It has very much that same property. It's another one of these very brilliant insights that you just had to put things in the right place. So let me explain the principle. So, suppose that I would like to know there's a gene that I know there's a family history of something, and I would like to know, but I happen to get the allele that carries that? Or did I get the one that didn't? So, in principle what I would like to do is to get a hold of the piece of DNA for that gene from my own cells. But all I've started with is my entire DNA. Well, I could clone it. I could make a recombinant library. I could do everything else. But there's this other simple way. And one way this involves, what it involves taking, is since I know the sequence of the genome now, I know that almost everything is going to be the same. There will be little differences between individuals. I'll make a little primer that corresponds to the sequence that one end of the gene, and another primer that corresponds to the DNA at the other end of the gene, or whatever fragment I want to use. And that's all I have to do in terms of getting anything made. Now the rest, we are just going to play games with DNA, with DNA polymerase, and nucleoside triphosphates, just all the stuff I dragged you through talking about DNA replication. So here's the idea. So here's my DNA, let's say, or part of it. If I could actually see the sequence, I would know, let's say, the gene I'm interested in is in here. So, what I would do is make a little primer. It just has to be enough to confer specificity for something with humans. If I make something probably 30 nucleotides long, that's enough. It'll only bind one place in the DNA. And I make one, let's say, for the opposite strand over here. So remember, this is five prime, three prime, five prime to three prime. So the principles will heat to 95°C, and will denature the DNA. And we'll add an excess of the two primers. And let's say we'll cool to 55°C, or something. And we'll cool it down enough so that we can get the primers on. But we are not going to go all the way and let all the strands find their way back. And we'll add a DNA polymerase plus four deoxy nucleoside triphosphates. Well, what will happen? Well, here's one of the strands. And we'll prime it here, let's say. So, it will copy down here and go as far as it can go. And the other one starts here. And it's going to go down all that way. Let's just repeat the whole process now, OK? What'll happen? Now when we pull them apart, we ought to have four strands. We'll have the original ones here, and when I repeat the process, the same thing's going to happen again. This one will go here, and it will copy out. This one will go here. It will copy out. But what about this guy? So, this one becomes this one here. So the primer that it does will copy it, and it can't go any further. I just generated a piece that's exactly what I wanted. And the same deal here: as long as I don't get lost, which what did I do? So, we've got this guy here. So, it starts there. So this one becomes this one, and we'll prime it here. It'll go along and it will stop. So there's the complementary strand to the one here. And I think this is sort of like doing a math problem. You can't just look at it and say, we'll maybe you will get it. But there's nothing like sitting down with a pencil and paper, and take yourself through several cycles. What you will believe is how quickly you get to get being nothing but, almost nothing, but the sequence that you are trying to amplify. And so, this again has an astonishing effect. This is why you hear about DNA testing all the time in forensics, because you can take a tiny bit of DNA from saliva, or semen, or blood, or whatever they might find on a crime scene, and then they can amplify little pieces and they compare. And there's a trick they use in forensics, and that is that there are sequences within the human genome where the little variable repeats like GT, GT, GT, GT, GT, GT, and I might have 14 of them in one of my chromosomes. The one I got from my mom might have 40. You might have 24 and something else, and so on. If you were to do PCR around a little region that was known to be variable, if you had 14 repeats you'd get a shorter fragment. And if you had 40 repeats, you get a longer fragment. So, I'll come back to that in a sec. So, if you were to, for example, take something with a long [repeat and a?] short peak into this kind of thing. We get two fragments, say, one from the paternal. And if you do this with several such sites around the genome, pretty soon you run into situations where the odds of a particular combination of a long one at the site, a short one at the site, and so on, becomes statistically improbable that it's anyone other than yourself. So on a crime scene, if they did this, they, for example, might have three individuals that they were thinking was possible. And they'd generate patterns like this, say, using three different loci like this, and then have the forensic sample. And it was pretty evident who didn't do it, and who at least remains a suspect. This probably would improve it. The very last thing, just to close us off, is when people develop this PCR technique, you had to sit there with your pipette because every time you raised it to 90°C to denature the DNA you killed your enzyme. So, you cool it down to 55, escort it in a new DNA polymerase. And then someone finally said, another brilliant idea, what if I had a thermoresistant polymerase? Where would I find those? Well, Penny was talking to you about those events where it's really, really hot, and those black smokers and everything, so maybe you got a bacterium from there. It would have a temperature resistant polymerase. So here you are from the New England [Biocatalog? , [vent exominus?] DNA polymerase, deep vent DNA polymerase. People went to grab those bacteria from there, grabbed the DNA polymerase gene. And now, the DNA polymerase just sits there. It just laughs and you bring it up to 90°C. And when you cool it back down and give it a substrate again, it will do its thing. And so, this whole thing can be done automatic and you don't have to sit there and pipette something in at the end of every run, another little cute sort of engineering trick that combined ecology together with biology. OK, see you on Friday.