Course Name: I Think Biology

Professor Name: Dr. Sravanti Uppaluri

Department Name: Biology

Institute Name: Azim Premji University

Week:5

Lecture:25

W5L25_Genetics - II

Hello. In this second lecture, as part of the genetics unit for the iThink Biology NPTEL course, we will be picking up from where we left off, discussing going from basically Mendel to modern day genetics, right. How we can use our knowledge of genetics and inheritance for applications such as gene therapy, or even things like discovering our ancestry and personalized medicine. So we begin to do this by trying to understand the link between genotype and phenotype. We've not used these words yet so far. So genotype really refers to what is carried on our genes in the DNA, right? What is the code on the DNA that leads to the phenotype, so the physical characteristic often that we can observe. So the link between these two is not always clear, right? How do we know which genes are responsible for particular functions? For example, somebody may have sickle cell anemia, right? And sickle cell anemia we know as a genetic disorder, so we know the inheritance pattern.

And yet how do we discover how sickle cell anemia is encoded for in particular genes, right? So that would be an example of trying to identify a gene for a particular function. And so there are actually two approaches, right. The first is where we look at a phenotype and try to identify what the genotype is. And that is called forward genetics. That's called forward genetics. On the other hand, if we look at a genotype, right, we can then try to identify the phenotype that corresponds to a particular genotype. And that is called, as you may be able to guess, reverse genetics. Yeah.

So how forward genetics really works is where you identify a particular phenotype that you think is really interesting, and then try to go backwards. Sorry, try to go forward and say, okay, this must be the genotype that it corresponds to. On the other hand, the reverse process is where you choose a particular gene, right. So you already know the location on the chromosome in the DNA code that you want to target.

And then you sort of, you know, target that gene, perhaps introduce a mutation or overexpress

that gene or knock it down using RNAi, and then try to identify the phenotype that emerges, right. And once you see that there's a change in phenotype, you can then link what the genotype is to the phenotype. So what we're going to do is actually go through some examples to try to identify whether this, in this particular example, the individuals who are responsible for this work, have used a forward or reverse genetics approach.

So what we're going to do for the next few slides is to look at examples of work where people have tried to make this exact link between phenotype and genotype. And our exercise will be to really identify whether these examples correspond to a forward genetics approach or a reverse genetics approach. And by the way, all of the examples that we are going to discuss involve work that has led to a Nobel Prize.

So this particular example is work that has been done by Christian Nusslein-Volhard and Eric Wiecschaus. And they worked with the embryos of fruit flies, which was Drosophila melanogaster. And what they did was that they used known mutants. So they used flies that looked different from wild-type flies, right. Wild-type embryos. And you can see that on the bottom left.

So this is what the normal embryo looks like, right. That's what you see here. You see that it has these even segments. And then they identified different sort of phenotypes. They named them croupal hunchbacking terps, depending on the historical reason behind where these discoveries were made or what phenotype that came out of them. But basically what they did was once they identified these strange, let's call them strange abnormal phenotypes, they then went on to try to identify the mutations, the gene mutations that gave rise to these particular phenotypes, right.

And so if we refer back to our slide, this is actually an example of forward genetics because they have used phenotypes to go back and identify genotypes. Another example is work by Robert Horvitz, where he looked at the C. elegans, roundworm. That's what is imaged here, right. The C. elegans roundworm is really interesting because it has a very fixed lineage. What this means is that you can watch from the very first cell, right. So the zygote, the fertilized egg, you can watch as the cells divide and every single worm will follow the same pattern of cell divisions. And you can therefore follow what is called the cell lineage, right.

And what Robert Horvitz did was that he identified cells that actually die during this process of development. So as cells divide during the process of development, some cells also die. And they die through this process called programmed cell death or apoptosis, right. And he actually was able to identify all of the cells that died.

And there are exactly 131 cells that died during this process, right. So in the cell lineage, you have 10,090 cells. And by the time you get the adult worm, you have 959 cells, right. And what he was able to do was then look at fly, sorry, look at worms that did not follow this exact pattern, right. Where, let's say, the death genes were not activated for whatever reason. He didn't know

what those death genes were, but he was able to identify organisms that did not follow this 1090 cells to 959 cells, right. And so because he was able to identify them, he was then able to say which gene is it that is responsible for cell death, right. So again, this is actually an example of forward genetics because he identified mutants that did not follow the same cell lineage and then traced it back to the specific genes that were responsible for this phenotype.

It turns out that the genes that he isolated that were responsible for this programmed cell death also have homologs in humans, and they have been implicated in cancer. So you can imagine why this is the case, right. Cells that are normally supposed to die, if they don't die but instead continue to proliferate, this can be a precursor to what we know is the characteristic of cancer, which is uncontrolled cell proliferation.

Okay, another example is that of site-directed mutagenesis. This was work done by Michael Smith, and he basically developed this technique in order to be able to specifically identify a location, let's say, in a sequence of DNA and mutate it with exactly what you want it to be mutated with. So let me clarify that. Suppose you have template DNA that is shown here. This template DNA, one strand, let me read it out, is CTCGCCTTC, right. That's what is verified in the lower strand in yellow, green, and blue.

And through the process of PCR, you can engineer primers to bind to the template, right. So for the primer to bind to the template, you need to create strands of DNA, short strands of DNA, that's what primers are, that are complementary to the template, right. So in this case, since you have two strands in the template, let's take the lower strand from the template, and that's what you see on the left-hand side, and you create a primer that is complementary, except that there is one mutation. And what a mutation is is basically a change in what is supposed to be there of the base, right. So we know that in the template, we had a CAGCGGAAG, but in the primer, we've created a primer such that the C that was originally there, that's what I have here, has been replaced with a G. And so what happens, even though these are now no longer complementary, right, between the primer and template, because these are, the overall primer is fairly complementary to the template, you can use it as a primer to extend this strand, right?

So the polymerase enzyme, which you would have heard about in the biotechnology lectures, can extend the strand, and now you get a complementary sequence, except for that one mutation that you have introduced because of the primer. So now that you've extended it, you've created basically the complementary sequence to the template except for that one mutation, and let's zoom out of this image. And once you've created these complementary strands, you basically have created a complete plasmid, which is a circular element of DNA that has a mutation that you introduced, right. And so it's different from the initial template that you started off with.

Of course, I'm simplifying the process a little bit, but basically the idea is that now you can choose a gene, directedly introduce a mutation, right, and then see what happens. So what do I

mean by what happens? You can see what the outcome on the phenotype is. So you can imagine that if this gene is a protein coding gene, and you've heard about protein coding genes when we talked about the central dogma, if a protein coding gene is mutated, so there is something that is different, you can then look at the effect on the protein's function, right. So in this case, what you're doing is linking the genotype to the phenotype. So this is an example of reverse genetics.

One more final example, which is of RNA interference, you would have heard of RNA interference already in one of the previous lectures, but I'll just go over it quickly so that we can discuss whether it is a forward or reverse genetics approach.

So the idea of RNA interference is really based on the central dogma and the idea that double-stranded RNA can interfere with the process of the central dogma. So in the central dogma, you go from DNA, which is transcribed to mRNA, which is then translated to protein, right. So this mRNA, right, when you introduce a double-stranded RNA molecule into the cell, into the cytoplasm, this will, this can bind to the mRNA molecule and cause the mRNA molecule to be destroyed, right. So effectively what you're doing is you're shutting down the function of the gene because you're no longer allowing the gene to be transcribed or, sorry, you're no longer allowing the gene to produce its protein product, right. It can be transcribed, but it won't be translated to protein.

And I should mention that this RNAi technique has, can have different levels of penetration. So in other words, depending on how much of the mRNA is destroyed, you may be able to either knock down that protein fully from the cell or produce it in partial amounts, right. Because the mRNA is not, it's not just one mRNA molecule that you produce, but several. And so if you knock down this mRNA expression, not fully, but to some extent, you can actually try to understand what the effect of having, say, slightly less of the protein is, right. So you can take it, you can have a much more nuanced understanding of the function of a protein.

And this is a reverse engineering, sorry, a reverse genetics approach because again, you are going from knocking down the expression of the gene to try to look at what the phenotype is. So you're blocking the geotypic expression to see what the effect of the phenotype are, right. So the effect of looking at this knockdown versus the null mutant also allows you to have a more nuanced understanding of the function of a gene. Okay. So now we have a better idea of what forward and reverse genetics are, but I want to move on to an effort that started early in the 1990s. And that is the Human Genome Project.

The Human Genome Project was the first project of its sort that was an international consortium, open data sharing, and it really sort of changed the way that science was done, right. People shared data across the globe. They shared the work, they shared the results, and it was a very open project. And this was an amazing feat, not just for the idea that you could try to sequence the entire human genome, but also the way science was done in general, right. That you

could have these international collaborations, that you could share data openly and you didn't have to take credit for everything that you did and sort of do things in hiding. But this is really the idea that publicly funded work should be publicly shared, right. So this is also a very nice example of how science can be done. And you should keep in mind that at the time that the Human Genome Project was started, we really didn't have the sequencing technologies that we have now.

So the project was meant to generate the first reference genome. In other words, to sequence the entire human genome. And unlike what you may think, it was not based on a single human genome, right. But rather on 20 volunteers, okay, that were. First, so you can see the ad that had asked for 20 volunteers in Buffalo, New York in the United States. And eventually 12 donors were used, whose DNA was sequenced, the genome was sequenced.

Most of the sequence was actually derived from one donor, 70% of it. But they basically pieced together the genome from many individuals. This is just a roadmap to show you that the project began in 1990 and ended in 2003. So you can see that it was a long journey. And you don't need to be able to see all of this, but it's just the idea that it was a really long effort.

But an impressive one because in the process, right, many things happened. The outcome wasn't just that we got a reference genome, right. That we got the first human genome sequence. But it was also in the process that we developed a new way of doing science. We developed many, many sequencing technologies that did not exist before. And also developed a very deep understanding, or at least began to develop a deep understanding of how the human genome is organized.

Now since the Human Genome Project was completed in 2003, the cost of sequencing, right, the cost of sequencing one human genome now. So it took us 13 years to get a human genome. But now we can do this in a matter of hours. And the cost has dropped dramatically, right. So the Human Genome Project was a $3 billion project. And now the cost of sequencing of one genome is on the order of $1,000, right. So this is already a huge leap in terms of how much we can do, right.

It took us 13 years for one genome. And now we can do this in a matter of a day. So you can imagine how much this Human Genome Project has changed the landscape of science, but specifically human genetics. So how does sequencing technology work? This is just a very broad overview. Nowadays there are actually many different ways of doing this, but this is a guideline to tell you in general how sequencing works, right.

So we know that DNA is double-stranded. First we start off by having a DNA sequence, right, a double-stranded DNA sequence that is then separated into single strands, right. These single strands are chopped up into smaller pieces, okay. And these smaller pieces are mounted onto a

surface.

Generally, you will have a surface that is like a bead, and this is mounted onto a bead, right. And this sequence serves as a template, right, to guide the synthesis of a complementary strand. You'll notice that this is also somewhat similar to PCR where you have a complementary strand that is synthesized, okay. Now, the really key and important facet to this is that every time you have an extension, right, so let's take a blow-up of this. You have a guide strand, right, the template strand.

Every time you have a base added that is complementary to the strand, what happens is that the addition of this base releases some light, right. And this light, the color of this light, depends on the base that is added. So you can imagine, for example, if you have an A added, you might say that a red light is released. So this is a biochemical reaction that has been engineered. If you have a T added, you might have a green light released, and so on, right? And so a camera detects the light that is released, right, and the order in which this light is released reveals the sequence of the original strand. That's more or less the basis of sequencing technology.

There are variations on this as well. Okay, so let's look at the sequencing in a little bit more detail. This comes back, again, to, we start off with the genomic DNA. That's what you see at the very beginning, right. And there are sort of two really relatively important methods that you should know about when thinking about sequencing. The first is Sanger sequencing. This is named after somebody named Sanger, who also won a Nobel Prize for his work.

And the idea is that you start off with genomic DNA, and you amplify the region that you're interested in sequencing, right. And you do this by using PCR, which is a polymerase chain reaction, which you will also learn about in the biotechnology chapters. So what amplification really means is that you create many copies, right. That's what you see here.

You've created many copies of a specific region in the genome that you're interested in. And then you use chain termination PCR, which means that you take the sequence, right, and you take a single strand. And what you do is you keep adding bases, right. So you put in a mixture of bases, and you put in, let's say, in one reaction, you put in a C that is fluorescently labeled. And what happens is that when this fluorescently labeled C in your reaction tube joins, it doesn't allow any more additions of bases, right.

So basically, the template strand, the length of the template strand is determined by whichever base is added on to that section, added on to rather the elongating strand, right. And so the length of the strand, so for example, here, is determined by the base that is added that has terminated the PCR reaction. And in another case, you might see that an A has been added. And this A again is fluorescently labeled.

So you can then tell apart the sequences based on their length and the fluorescent label. So you can see that this strand will have a blue label, but will also be longer than this strand, which will have a red label, but will be shorter, right. And the length of these labels is indicative of where of the location of that particular base, right. So this is translated, in other words, by saying that PCR fragments, that's what you see here, right, are separated based on size.

And the DNA sequence can be read by the order in which the fluorescent signals are detected, right. And you can see that here, this is what is called, this is what you get out of a sequencing reaction. So this would have been the shortest strand, right. And this would have had a fluorescent signal of yellow, which is the T, right. The next length, the next strand would have been, let's say two bases long, and this is an exaggeration, a simplification, but basically you see this, it would have been two bases long, and it would have been blue in color, right. The fluorescent signal would have been blue.And so you can read what, this is called a chromatogram, and you can read off the sequence by looking at the peaks in color at every position.

And this is what is called Sanger sequencing. Sanger sequencing is very useful for, at least this version of Sanger sequencing is useful for relatively short sequences. There's also something called second generation sequencing, also called next generation sequencing. And the principle remains the same, except that here you take genomic DNA, you share it into many small pieces again, right. And then you add adapters to these, okay. These small fragments. So now you have the entire, you have a whole genome, or at least you have a much longer sequence. And these adapters that you've added to the short strands allow you to place these on a surface, right. And once you place them on a surface, they are amplified, okay. Amplification again is done through some kind of a PCR reaction.

And the same sort of process is followed where fluorescently labeled nucleotides are incorporated. But this time, as they're being incorporated, a light is released, and the release of the light is indicative of which base has been added, and in turn, the sequence of the bases, right. So you are able to sequence your, say, large strands of DNA using this method. But of course, don't forget that you've broken up the genomic DNA into many pieces, right. And so you will actually get the sequence of many pieces. What you have to do then is try to reassemble these pieces into a sort of cohesive structure. And you do this by these, so this, in this image, right, it's representative of the sequences that you've gotten from this process, right. The sequences that you've read.

And what you do is you look at overlapping regions and say, ha, these two regions overlap. Therefore, they must be part of a consecutive sequence within the whole genome. So this is called de novo sequencing, or de novo assembly. This is done using the overlaps. Or the other thing you can do is use a reference genome.

So for example, if you are sequencing a bacterial species, you might already have a genome, and you want to try to help, try to assemble what you are sequencing. So you can use the bacterial genome as a reference, and then try to align these smaller segments that you have to the reference genome. And this can help you then map the whole thing. Okay, so we talked about a $1,000 genome. But really, the $100 genome is the dream, right. So this is actually, you can see that this article is from 2022.

And this is, this dream has not been realized yet. But there really is a lot of new technology that is emerging, where it wouldn't be surprising if a $100 genome is just around the corner. The only problem with making genome sequencing cheaper and cheaper is that sort of the economics, right. It is cheaper as you if you sequence a lot of genomes at one time, right, but in the clinical setting.

So suppose you're a cancer patient, and you need your genome sequence to try to identify whether you have specific mutations that predispose you to cancer. The only way to do that is to have your individual genome sequence, right. And that becomes very expensive. Whereas if you have say 100 genomes to be sequenced at the same time, so sort of a bulk order, this is much easier to do, right, or rather much cheaper to do.

So the high throughput is really the problem with clinical settings. Okay, I just want to end with one more very, very interesting application to the sequencing. And that is the single cell RNA sequencing, are also called RNA-Seq. So this is actually an atlas of, let's say, single cells whose RNA expression profiles have been sequenced, right. And this basically what you might call an atlas indicates overlapping features, right, or overlapping RNAs across different cell types. And you can see that, let's say, if you look at the zoomed in area, cardiac muscle cells and skeleton muscle cells are much closer together on this atlas, because unsurprisingly, they express similar RNAs, right, because they have similar functions, they're both muscle cell types.

And so the idea is that now you can, you know, you're no longer sequencing just a genome, you can also sequence the RNA molecules that are present in a specific cell. And you can do this at a single cell level, so that you can actually gather a kind of atlas to see, let's say, in a single multicellular organism within each tissue, which, again, contains all the cells contain the same genome.

But you can also look at which genes are being transcribed in which cell and at what time. So you can imagine, for example, that this would be very interesting to look at in developmental biology, right. So how does a mouse develop over time? What is the expression patterns of different tissues as they develop, right. So this is really a very useful application, not just to understand disease progression, but also for developmental biology.

Okay, so that brings us to summarize what we've learned in today's lecture, we started off by linking genotype to phenotype.

We talked about different approaches that we use, right, forward and reverse genetics. We then looked at specific examples of forward and reverse genetics, right. Everything from the example of looking at cell lineages in worms, to looking at, you know, mutants in fruit flies, and RNA interference and site directed mutagenesis, all of which have been worked at, led to Nobel prizes.

We then said, okay, what is the, in modern day genetics, what is one of the most useful technologies or methods that's used to understand the link between genotype and phenotype, but also to understand beyond this, right, to look at, say, clinical applications or developmental biology applications, we looked at sequencing as a technology. We tried to understand the process of sequencing. And we also briefly discussed RNA sequencing, right, and how RNA sequencing can actually help us not just now move beyond genomes, right, but we can also look at expression patterns.

So how the genome is being transcribed into RNA, and presumably RNA is really much more representative of, or much closer to what the cells are actually doing, right, because though every cell has the same genotype in a single organism, they certainly have different expression patterns. So in the next lecture, we will continue our discussion on human genetics and varied applications and sort of interesting applications of our knowledge of genetics in human biology.