Next Generation Sequencing Technologies: Data Analysis and Applications

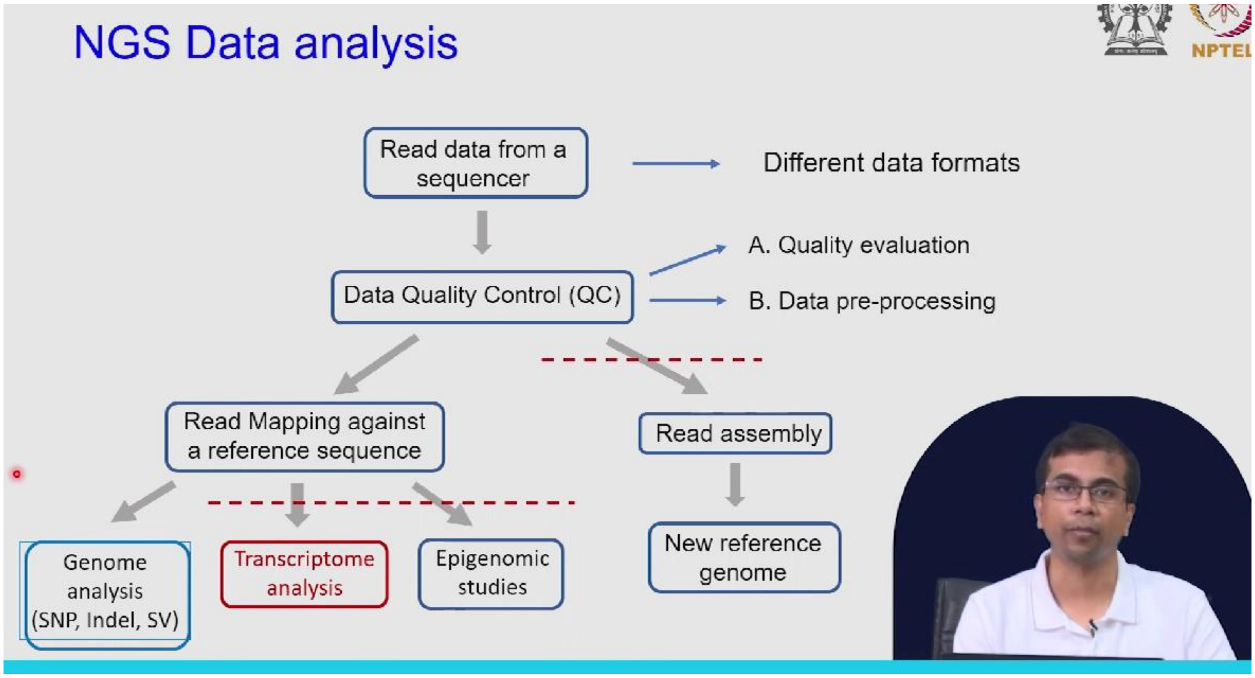
Introduction to RNA sequencing

Dr. Riddhiman Dhar, **Department of Biotechnology**

Indian Institute of Technology, Kharagpur

Good day, everyone. Welcome to the course on next-generation sequencing technologies, data analysis, and applications. In the last few classes, we have talked about the identification of variants and different types of variants. So, the time has come now to talk about transcriptome analysis, and we will talk about RNA sequencing. In this class, we will introduce what RNA sequencing is, the different aspects of RNA sequencing, and its advantages over the earlier methods. So, let us start.

So, we will talk about these RNA sequencing steps briefly, and we will also compare them with microarrays and talk about the advantages of this NGS-based transcriptomic analysis. So, these are the keywords we will come across: isoform, short read sequencing, and long read sequencing. So, we will start with that flow chart for NGS data analysis. As you see, we have the read data from the sequencer, which comes in different data formats; we have data quality control; and then we have read mapping. So, we have talked about all these steps in detail, and we have completed the genome analysis in the last few classes.

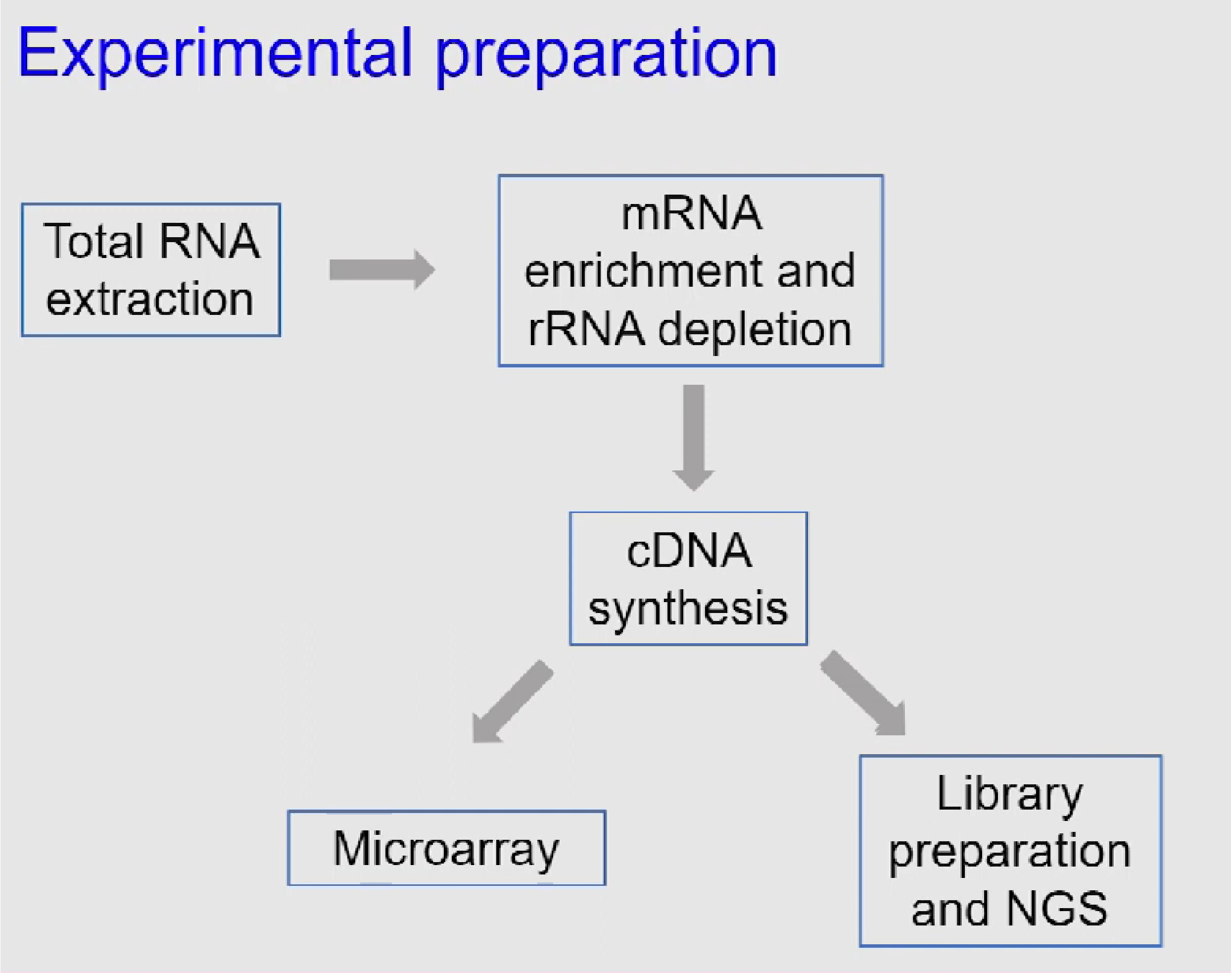


NGS Data analysis

So, we will talk about the transcriptome analysis now and we will follow that up with a discussion on read assembly for reference genomes and new reference genomes, and then we will talk about the epigenomic studies. So, we will start with the transcriptome analysis. So, transcriptome analysis is also called RNA sequencing, or RNA-seq, even though sometimes we are not directly sequencing RNA, as we will see in some of these methods and most of these experiments are done to identify differences in the expression levels of genes between samples or between different conditions. So, it could be disease versus healthy, or it could be some sort of treatment that is given to certain samples, and you want to compare treated versus untreated samples. So, these are the main goals of transcriptome analysis, and we used to do them using microarrays earlier.

So, what is the protocol? How do you do this? So, some of this is actually shared with the microarray protocol. So, what you have in the first step is total RNA extraction from the sample that you want to work with. So, for example, the tissue sample could be a cell sample, right? So, we do our RNA extraction, and there is something called mRNA enrichment because we want to measure the expression level of genes. So, we want to get rid of other RNAs that are present in the sample, for example, ribosomal RNA, etcetera.

So, when you isolate total RNA, you will see that most of the RNA is ribosomal RNA, which is the most abundant RNA in the cell. So, we want to get rid of that, and we perform a step called ribosomal RNA depletion and mRNA enrichment. So, this mRNA enrichment can be done in different ways. For example, we can use something like amplification using primers against the poly A tail in the case of eukaryotes, and we can amplify only the mRNA and reduce the concentration of ribosomal RNA. Now, if we use this amplification, of course, we need to keep in mind that that will introduce certain biases, which we will talk about in the next few classes. So, why is it important? Because these biases will come into the data analysis, right? So, we need to deal with those when we are analyzing this kind of data from RNAseq experiments, ok? The better method is rRNA depletion, and this is something that people use without amplifying the mRNA, but in some cases, if you have a very low amount of mRNA or total RNA, you would have to perform some amplification steps that will, of course, affect your results. So, once we have isolated these mRNA molecules, we do something called cDNA synthesis using reverse transcriptase, and again, this will introduce certain biases into our data. So, these biases we will have to deal with when you actually analyze the data. Once we have cDNA, we can proceed with this library preparation, which is adapter ligation. If you are going with the Illumina platform, you have to do bridge amplification, etcetera, or if you are going with the other platform, you have to perform some platform-specific steps, followed by next-generation sequencing.



Experimental preparation

Total RNA extraction

mRNA enrichment and rRNA depletion

cDNA synthesis

Microarray

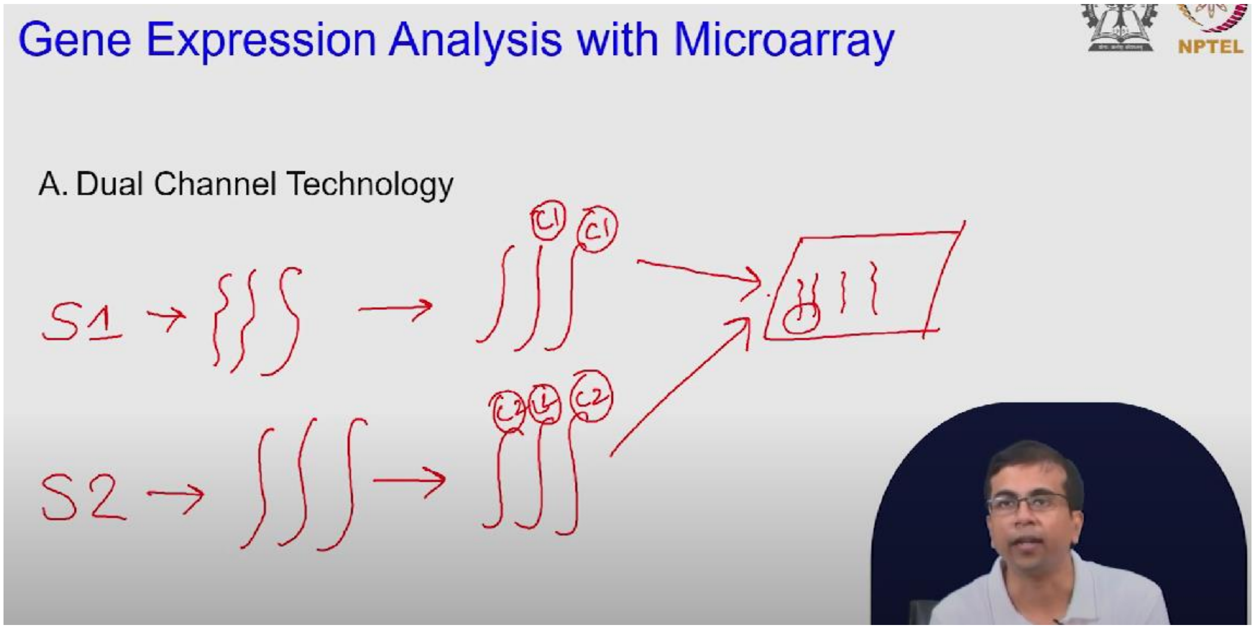
Library preparation and NGS

And if you are going for microarray, you have to prepare this cDNA for microarray experiments. It could be the addition of some sort of molecule or some sort of fluorescent molecule, etcetera, and then hybridization in microarray. So, what is the main goal of this kind of analysis? It is something called differential gene expression analysis. As I mentioned, we want to compare the expression levels of genes between these two different sets of samples that have been processed in exactly the same way. So, this is what we will discuss again: differential gene expression analysis. It will come again later on, and this is a comparison of two sets of samples. So, what I will do is now just briefly mention how it is done with microarrays, or how we used to do them earlier. For gene expression analysis with microarrays, we can have dual-channel technology or single-channel technology.

So, we will discuss what is dual-channel technology and what is single-channel technology. So, let us start with dual-channel technology first. So, as I mentioned, we do differential gene expression analysis where we have two sets of samples. In dual-channel technology, these two samples are labeled with two different dyes. So, you get two different fluorescent levels.

So, you can imagine this, right? So, you have sample 1 from which you get these RNA molecules, which are then converted to cDNA and you have sample 2 from which you also get RNA molecules that are converted to cDNA according to the flow diagram. Now, once they are converted to DNA, So, this one will get one type of tag. So, of one color, right? So, let us say this is color 1, and the second sample will get a tag of a different color. Let us call it color 2, right? So, this is usually a red-green coloration. So, c1 is usually red, and c2 is green. So, what then happens is that we mix these two samples and hybridize them against the microarray chip. So, in the microarray chip, if you remember.

So, we have these probes already attached. So, these cDNA molecules will go there and hybridize. Now, if you have higher expression of a gene in S1, what you will see is that color 1 will dominate right. So, if color 1 is red, you will see a red color coming from this region, and if the expression of a gene in S 2 is higher, then you will see color 2 dominate in the microarray. If these expressions are balanced, then you will see a balanced color.

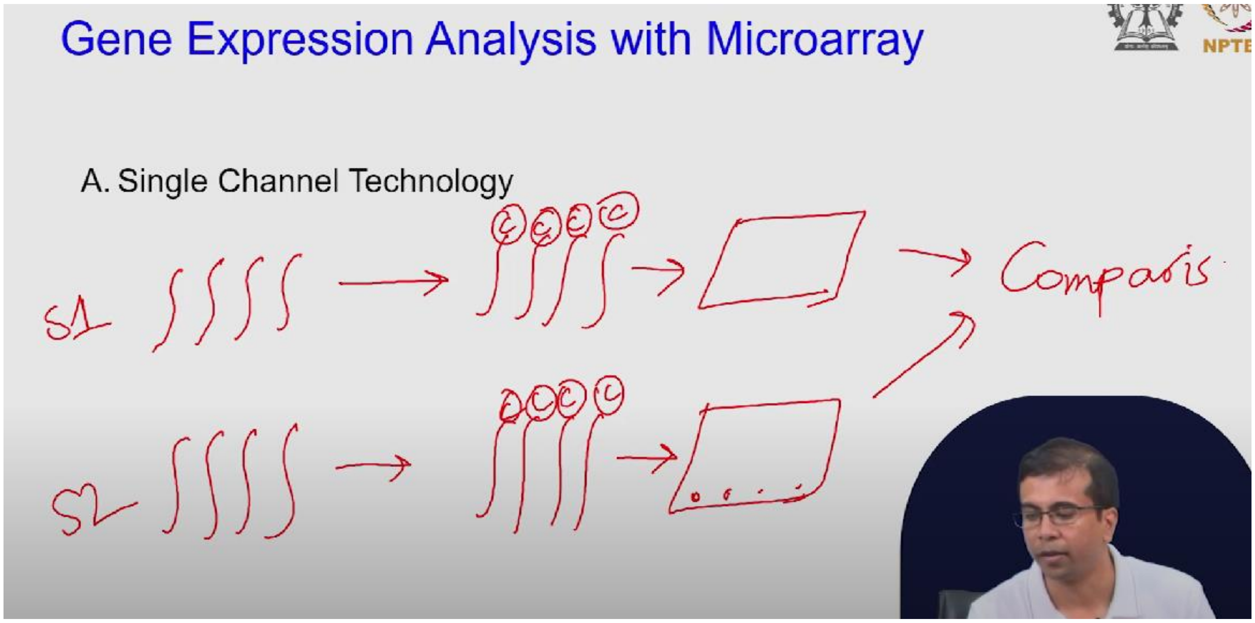


Gene Expression Analysis with Microarray

A. Dual Channel Technology

So, you can imagine that if you have red and green colors. So, if one sample is dominant, So, if you have S 2 dominating, you will get a green color, but if the expression is balanced between these two samples, then you will get a yellow color. Now, in single-channel technology, you do not have these two colors. So, what we do is take both of these samples.

So, again, if you have S 1 and S 2, they get the same kind of level, ok? So, the level is the same, and they are hybridized separately. So, there is no mixing, and they are hybridized separately on two different chips. What we do is measure the signal that is coming from these positions, and these different positions are measuring the expression levels of genes in these two samples. So, we get independent values, and then we can compare.

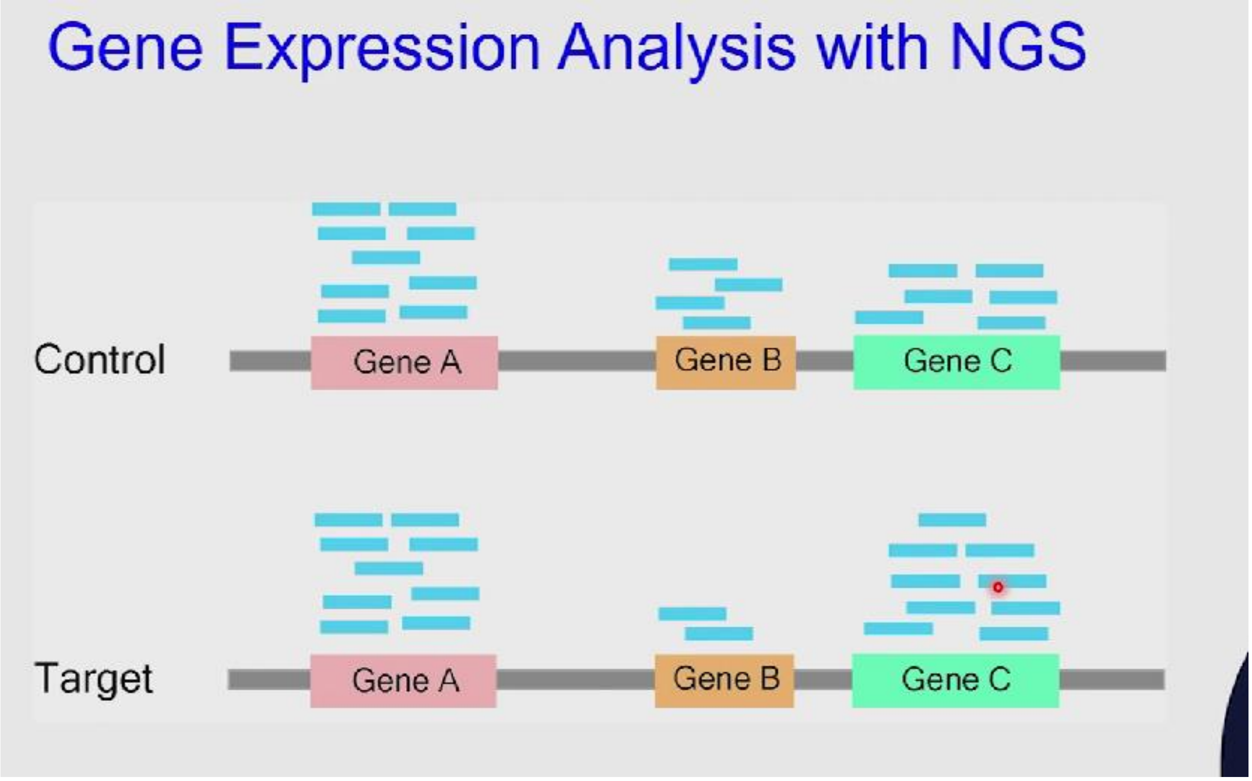

Gene Expression Analysis with Microarray

A. Single Channel Technology

So, this is the bioinformatic part where we actually compare the expression level, but there are several limitations. So, because microarrays rely on hybridization, we need to have prior knowledge of the transcriptome for oligo design. So, we need to design or synthesize these oligos on the chip. So, we need to know the transcriptome sequence right? So, genome sequence, gene sequence, mRNA sequence, etcetera, And the second major point is that we have you can detect only known features and you cannot detect any novel transcript.

For example, there is a novel transcript, or a splice variant that is expressed in the disease sample, but since we did not know about that, we did not design any oligos to be printed on that chip. So, in that case, we will not be able to identify that novel transcript in microarray experiments. The other limitation, of course, is that you can imagine that these transcript sequences are different from one organism to another, and then there are new genes in different organisms. So, you have to have organism-specific microarray chips, right? So, you would find disease serum, human array, etcetera.

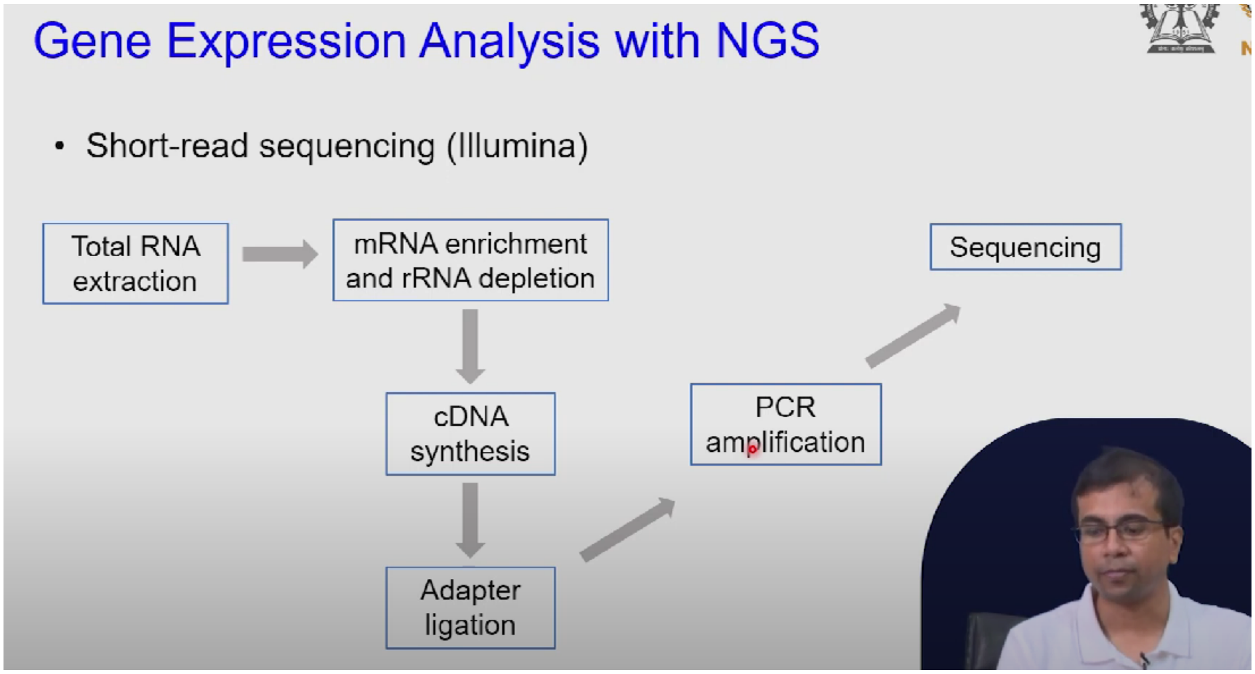
The other major limitation of microarray technology is its limited dynamic range. So, the expression level is detected by hybridization. So, you can have, let us say, for a gene, 1000 copies, maybe because of the space concepts, or because you want to represent all the genes that are expressed in a cell in a sample. So, you want to have all of them on the chip. So, for each individual gene, you can have, let us say, 1000 copies of oligos against which the cDNA can bind. Now, if the gene is expressed at a much higher level than 1000, So, it is producing maybe, let us say, 10,000 mRNA copies, right? So, would you be able to detect that? So, because you will reach that saturation because you have only 1000 copies of oligo, So, this means you are limited by this dynamic range, right? So, you cannot go very high because you will reach saturation; you also cannot go very low because the signal-to-noise ratio is low. So, again, you will see that there is a problem in detecting lowly expressed transcripts. So, if you have only one or two transcripts of a gene, you will probably not get any signal at all, or you will not be able to detect the signal that is generated. So, you will miss the expression level of these genes, which are expressed at very low levels, again and again. So, you cannot compare the expression level of these genes, which are expressed at a very low level, to whether they have changed between your normal and the disease sample. The other limitation, of course, is that you need a lot of RNA samples to actually do these hybridization experiments, and again, in some cases, they may not be available. So, because you would have to have a lot of starting material to get this large amount of RNA sample, and also because this method relies on hybridization, which could change depending on the hybridization conditions, If there are some fluctuations in the hybridization condition that can change the hybridization pattern, this could affect the signal, and again this could lead to reproducibility issues. So, from one experiment to the next, you might not get the same results. So, that is where the gene expression analysis with NGS comes in. The principles are very simple when you are analyzing with NGS: the number of counts or counts of reads that map to a gene is a proxy for the number of transcripts, and hence for the gene expression level. So, in microarrays, the hybridization signal is the number of copies that you have that will hybridize and that will generate some signal, and this signal is a good indicator of expression. In case of next generation sequencing or RNA sequencing, the number of or count of reads that we get mapping to each gene is the signal for the expression level, ok? So, something like this: if we present in a cartoon, you have gene A gene B gene C right and you have the control. For example, we have the normal, which is the control, and the target may be the disease, and in this case, you have gene A gene B gene C. We are comparing the expression level between these two samples, and if you just do what you do after sequencing, you map the reads, and most of the reads are supposed to map against the genes because they are coming from mRNA. And you see, for example, you compare this and you see if gene B in the target sample has a very low number of reads. So, you can probably tell that perhaps this gene B is showing low expression in this target sample compared to the control. And for gene C, you might see an increase in the number of reads that are mapping to gene C in this target sample. So, there is probably an increase in the expression of gene C in this sample.



Gene Expression Analysis with NGS

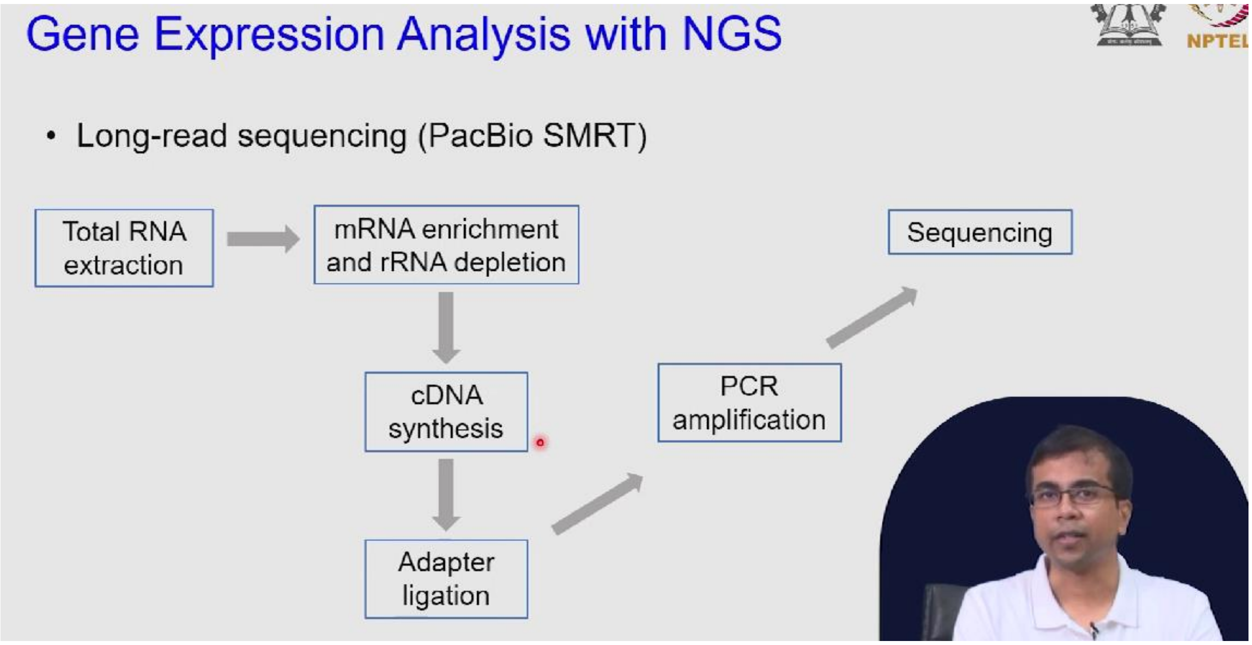
So, this is a very rough idea of course, you cannot just look at this count and say there is change in expression you have to have a very robust and rigorous statistical analysis that has to be done ok. So, and we will talk about this statistical analysis a bit later and the tools that can that actually will do this for you and we will actually do that hands on as well we will analyze this kind of data hands on and see the differential expressions right identify the gene that show change in expression between two different samples ok. Now, when you are doing this gene expression analysis you can do this with short read sequencing for example, Illumina platform you can do this with long read sequencing now with packed by SMRT or oxford Nanopore and you can do direct target sequencing using oxford Nanopore right.

So, if you remember the principle of the nanopore platform, you can actually directly pass RNA, and you can detect the signals and identify the bases, ok? So, let us talk about short-read sequencing. What would be the protocol in this case? So, you have the total RNA extraction; you get the mRNA enrichment and cDNA synthesis; followed by that, you have adaptor ligation; you have PCR amplification; this is the bridge amplification step that we need to perform; and then the sequencing is okay.



Gene Expression Analysis with NGS

- Short-read sequencing (Illumina)

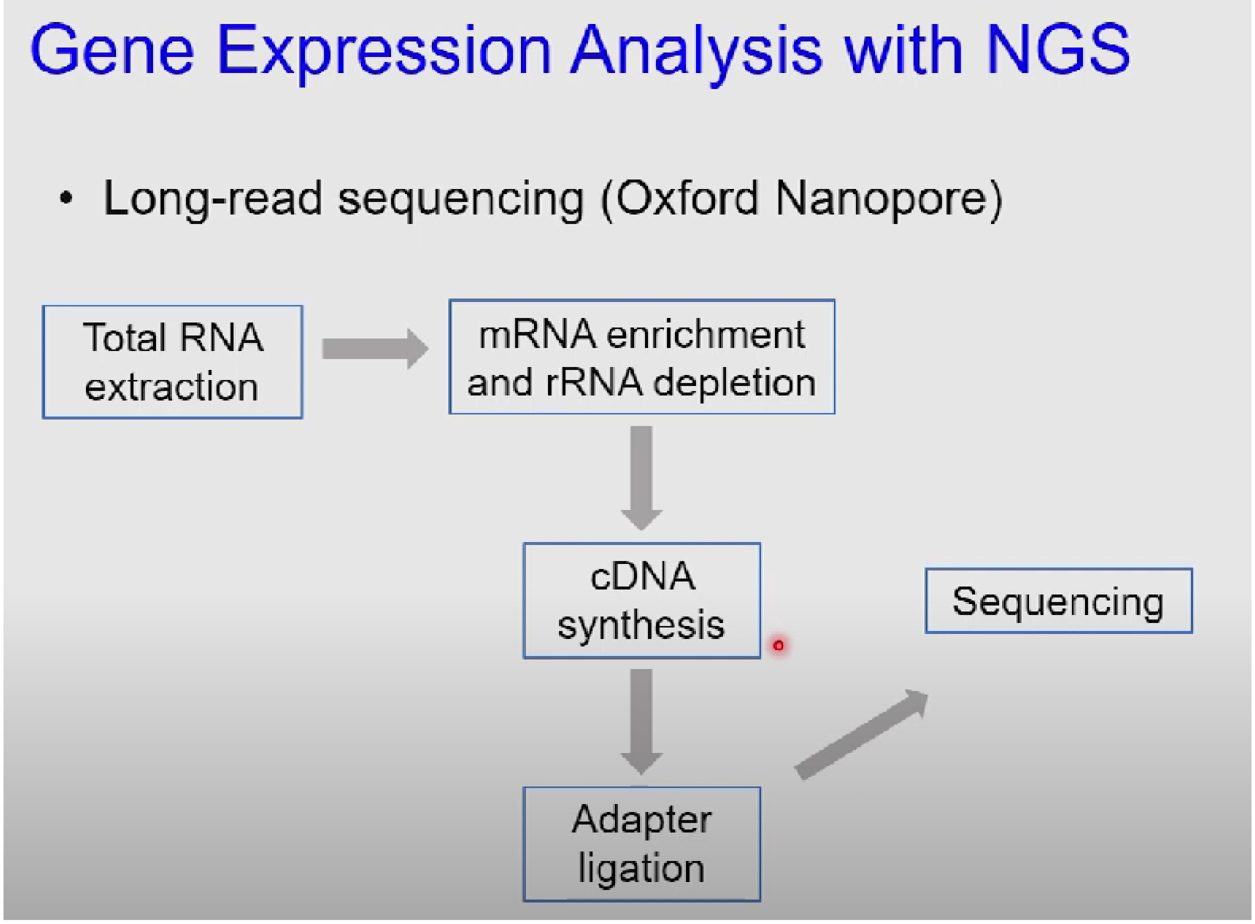


Gene Expression Analysis with NGS

- Long-read sequencing (PacBio SMRT)

So, you have the usual steps after cDNA synthesis, and of course, you cannot sequence this with RNA; you will have to synthesize cDNA. If you are going with long-read sequencing such as SMRT, the steps are pretty much the same: you have cDNA synthesis, adaptor ligation PCR amplification, and sequencing. If you want to go with Oxford Nanopore, you can skip the PCR amplification and directly pass the cDNA after adaptor ligation through the sequencer, and the current signal will change, which will indicate the base, and you can then do the mapping back to the reference sequence.

The direct RNA sequencing using Oxford Nanopore is actually much simpler, and as you can see, you can go from total RNA extraction to mRNA enrichment and then adaptor ligation and sequencing. So, this direct protocol might be quite useful in some cases because you are introducing a smaller number of amplification steps, which might give us better results. So, any amplification will have some bias, right? It will amplify certain genomic fragments much better than others because of their sequence preferences, GC content, etcetera. So, all those factors come into play, and that is why having fewer steps without amplification is actually very good for our analysis. So, what we do with this NGS is that we can do differential gene expression analysis, or DGE analysis, which we will do later and we will discuss this in much more detail, what kind of statistics we use, etcetera.



Gene Expression Analysis with NGS

- Long-read sequencing (Oxford Nanopore)

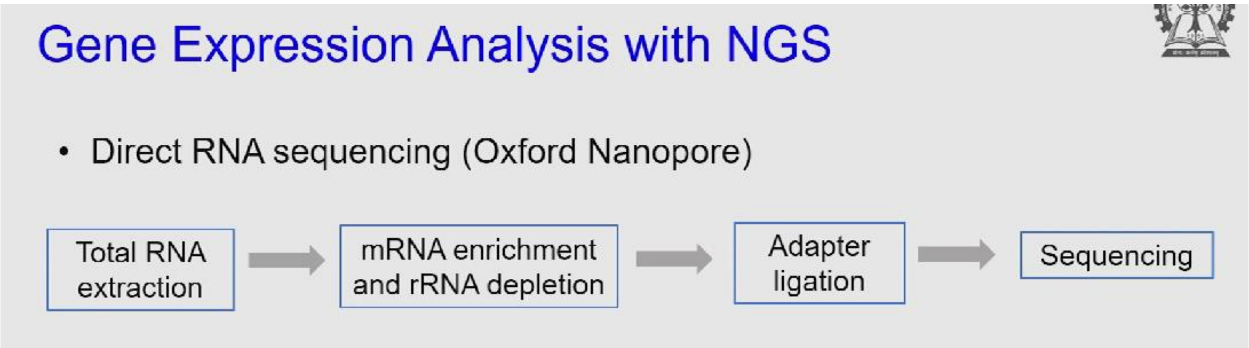
Total RNA extraction

mRNA enrichment and rRNA depletion

cDNA synthesis

Adapter ligation

Sequencing



Gene Expression Analysis with NGS

- Direct RNA sequencing (Oxford Nanopore)

Total RNA extraction

mRNA enrichment and rRNA depletion

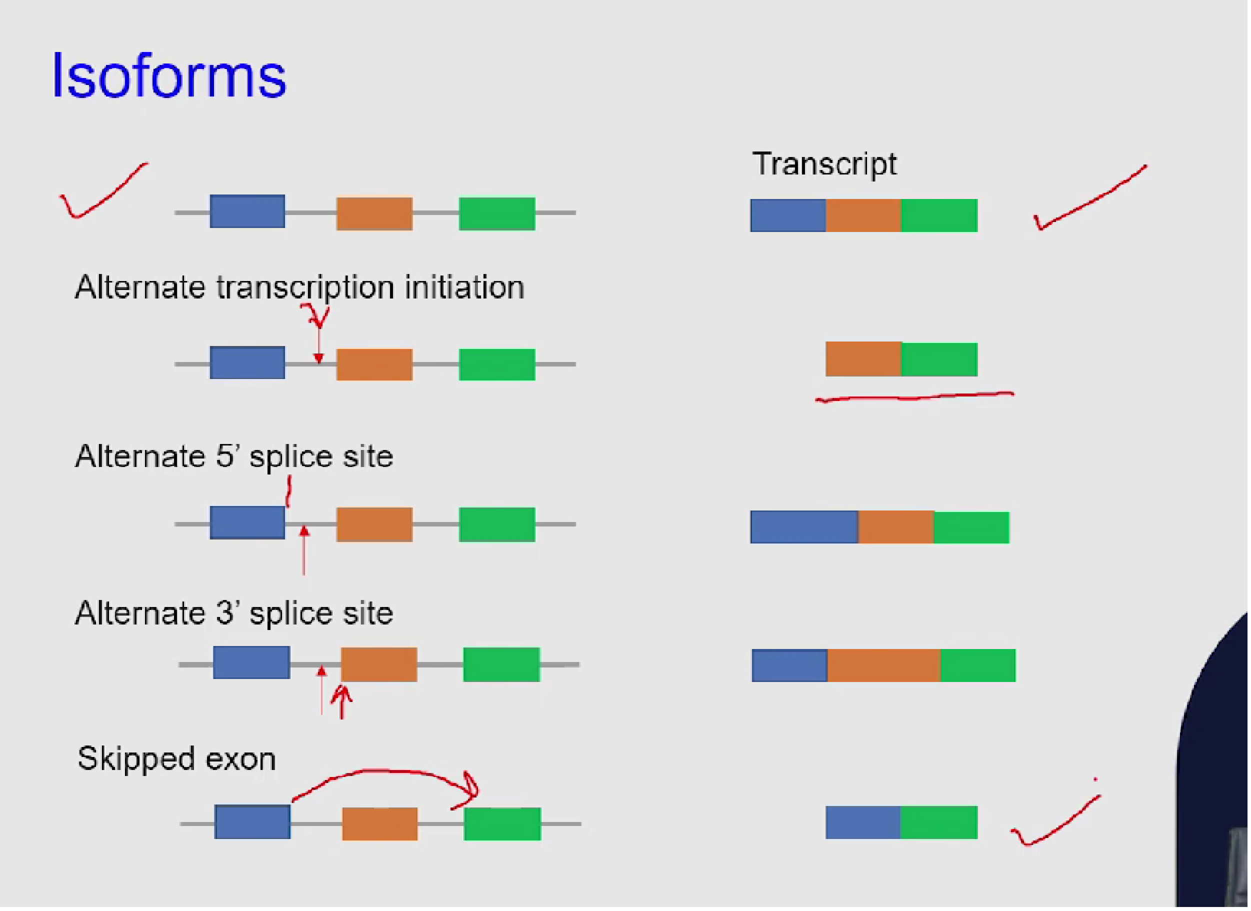
Adapter ligation

Sequencing

We can do something called identification of isoforms and alternative splicing events, of course; this was also possible in microarray experiments, but in that case, you would have to design those oligos again. You would have to think about what kind of alternative isoforms you might see and then design oligos against those, and this becomes very tedious. If you have to do this for, let's say, thousands of genes this becomes a very tedious process, whereas in NGS, you can do it very easily because this is a direct sequencing process. So, you can identify what isoforms you are getting.

So, we will talk about the different isoforms that are possible, and we can also study the role of non-coding RNA. So, we can quantify non-coding RNA expression levels, and we can study their role in expression regulation.

So, let us talk about different types of isoforms that could be present in a cell. So, I have drawn some scenarios, and there could be a few more. So, the first scenario here is the normal scenario, right? You have these three exons in blue, orange, and green, and they are combined together to give the transcript OK. So, this is the normal transcript here that you get with all the exons. Now, if you have these different types of processes that can generate these isoforms, So, one type of process is that you can have alternate transcription initiation, right? So, transcription initiation happens at a different place, ok? So, here may be the transcription initiation happens in this place, given by this red arrow, and what that means is that you will end up with only these last two exons, and the first exon will not be there. You can have other types of cases, right? For example, you can have alternate 5 prime splice sites, right? So, what you can have is that instead of splicing happening right at this boundary here, right at the end of this exon, it can be shifted a little bit, and this shift means you will include a larger region after the exon 1.



Similarly, you can have three alternate prime splice sites, right? So, splicing happens slightly upstream compared to the original splice site, and then again, you include a bit more region with exon 2, and finally, this is the classic case. So, you can have a skipped exon, right? So, maybe the second exon is not included; only the first and third exons are combined, and then they give right to this transcript. You can have other kinds of scenarios again, which you can find in the reference.

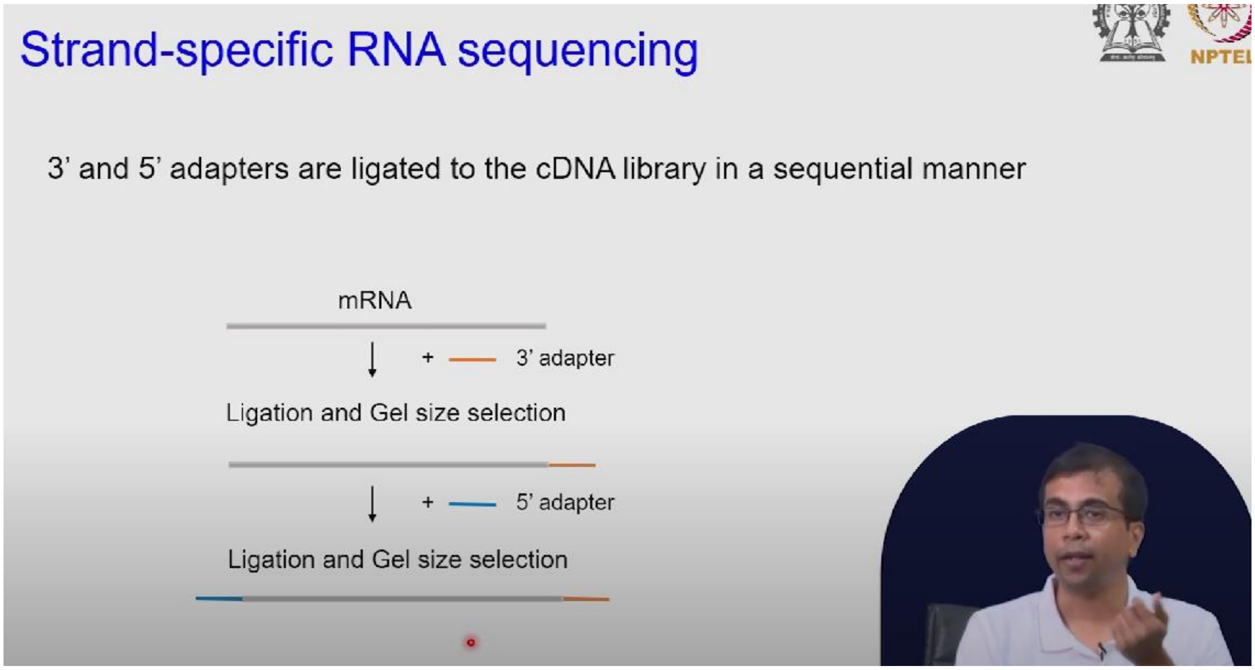
Advertices for shortlist sequencing are right because, as we know, elavida has really high throughput, so you get a large number of reads quite easily and are not very expensive. And we also know the biases and error profiles of the Illumina method, and we have an established data processing pipeline. We have already seen right when you are doing this variant analysis, some of these pipelines that are available, and most of them are developed for fastq files, right for the Elavida platform. What are the drawbacks then? So, why do not always use Elavida, or why do we need the log-read sequencing at all? So, one of the first steps is the multistep sample preparation. So, we have seen that for shortlist sequencing, we need this multistep preparation, we need this bridge amplification step, and this can introduce bias in the data. So, if this is a problem, then you also have limited isoform detection, right? So, you can imagine that if you have a long gene with multiple exons and you are working with short-read data, even with paired data, the data will not be able to identify all the isoforms if they are present. For example, if you have let us say alternative initiation sites or alternative splice sites, they may not be detected in the short-read data.

And in case you want to discover these isoforms, So, you may have to do this de novo transcript assembly for this transcript discovery again. Any assembly process right now becomes computationally more challenging. So, here comes the long read sequencing, right? So, that is, it can probably address some of these issues of short-read sequencing. So, long-read sequencing can sequence full-length transcripts, right? So, we do not have to worry about this assembly of short reads to get the full transcript sequence. So, because it is long read, you can sequence the full length, which makes the computational analysis simpler because you do not have to go through all this assembly process again. You can also discover these isoforms because it is long read. So, if it is like 1 k B, 2 k B, or even 5 k B sequences you are getting, you will probably discover most of the isoforms that are out there, so you do not have to worry about looking again, going to assembly, and then trying to find the isoforms. So, this is actually a big advantage of long-read sequencing. What are the drawbacks, then? So, one is, of course, the low throughput.

So, they have a lower throughput than the Illumina platform at this point, but of course, this might increase in the future. And also, you have multistep sample preparation in this process. If you are going for packed bioSMRT, you have to do the amplification again, and this can introduce bias. So, this is where the long redirect RNA sequencing comes in, ok? So, it retains the advantages of long-read sequencing and also addresses some of the drawbacks. So, as we have seen, it can sequence full-length transcripts, and computation analysis is very simple. It can also discover these isoforms.

In addition, it can detect some of the RNA modifications, because it is directly sequencing RNA. Any chemical modification on the RNA that will change the current signal is okay. And so, this information can also be collected during RNA sequencing. So, this is the advantage of direct sequencing, and it can also detect the polylent because it is directly sequencing the RNA. So, it is not through any amplification step. So, which can change or which can make this estimation process very difficult? What are the drawbacks, then? Why do we not do direct RNA sequencing? So, again, the throughput is still lower compared to the Illumina platform or short-read platforms, and in some cases, this might be a limiting factor because you can measure the expression levels of even the very lowly expressed genes, and in that case, you need a higher amount of data or a bigger amount of data. So, you need a lot more throughput. So, at least 30, 40, or 50 million reads are needed to actually get to that level where you can measure the expression levels of even very lowly expressed genes. So, in that case, you probably want to use something like a combination of long read and short read sequencing. So, because of this, you can combine the advantages of both of these methods.

Another drawback of direct RNA sequencing is that we still do not understand these biases and errors very well because we have been doing this method for a very short time, and more time is required to understand what kind of errors you can get and how to address them. Now there is something called strand-specific RNA sequencing. We have talked about RNA sequencing in general up to this point. So, there is something called strand-specific RNA sequencing. So, why do we need this strand-specific RNA sequencing? We have not talked about this strand-specific genome sequencing, so in that case, it was not really required.



Strand-specific RNA sequencing

3' and 5' adapters are ligated to the cDNA library in a sequential manner

mRNA
+ 3' adapter
Ligation and Gel size selection
+ 5' adapter
Ligation and Gel size selection
NPTEL

So, in the case of RNA sequencing, you can imagine a scenario, right? So, you have the genome, and two genes are on two different strands, and they still overlap. And in that case, right. So, you have these 2 genes that are opposite strands; they can overlap with each other, and they will produce RNA that will be processed, and then you are doing the RNA sequencing. And after the RNA sequencing, you get the data, you get some reads, and that maps to these overlapping regions. Now which of these 2 genes is producing these reads? So, this is something that is very difficult to figure out, right? If you have this kind of overlapping gene in two different strands in a genome, you cannot tell which gene is producing this mRNA, and you are getting the read data from that expression of that gene. So, this is a challenge, right? So, if you have reads mapping to the overlapping regions between two genes, the read can come from either of them, and we cannot determine this with normal RNA sequencing. And this is where you can actually do this strand-specific or stranded RNA sequencing, which can remove this ambiguity and measure the expression levels of these genes much more accurately because you can now resolve these reads much better.

So, how does this work? How do you do this strand-specific RNA sequencing? So, what we do is add these adapters, 3 prime and 5 prime, sequentially, right? These adapters are unique for 3 prime and 5 prime, and we can add them in a sequential manner with purification steps in between. So, now that we have an adapted sequence, we know we are getting the forward strand or the reverse strand data, right? So, this helps us in separating out this strand information from the read data. So, here is the brief schematic, right? So, we have the mRNA, which is added to the three prime adapters in orange, and then we do ligation and gel size selection.

So, you might ask why you need not get attached to the other side. There are ways to actually make sure that this will only be attached to one side. And this side can be prepared for that way, and then you have the 5 prime adapters added on this side again. You can specify right, you can control this process, you can regulate this process, and you do the final product with these 3 prime and 5 prime adapters, and these are unique ones. Now, once you sequence right, go through all the processes, and then sequence, you get the adapted sequence, and you know which strand you are getting the sequence from, and based on that, you can assign which strand is producing this mRNA and the read data, right? So, the one more thing that comes in RNA sequencing is called something called small RNA sequencing. So, this actually is a part of the transcriptome, right? So, you have normal mRNA, but you also have small RNA molecules inside the cell. So, this is something we know now, and this is small RNA sequencing, which means sequencing of small non-coding RNA molecules such as microRNA. Now, so far, we have talked about sequencing of mRNA, right mRNA enrichment, etcetera, amazomolar RNA depletion, etcetera, but we have not talked about this microRNA. So, how do you go about it? So, the first step is the same, right? So, we have the total RNA extraction from samples, and in this step, what we do is then do something called size selection for small RNA's only. So, we can separate out only the small RNA's that are of a certain size that you can say are okay. We will separate out may be only the 35 base pair up to 35 base pair RNA molecules or up to 50 base pair RNA molecules, and you can separate out these molecules add to the RNA sequencing ok. And once you do the RNA sequencing, you get the expression level of this microRNA or these non-coding RNA molecules. So, this information is very useful because we now know that these molecules are involved in expression regulation. So, knowing their expression would be very useful in understanding what is happening inside the cell.

So, we can now move on to the advantages of RNA sequencing after all this discussion. So, one of the first things we have seen is that it has the ability to detect novel transcripts and isoforms quite easily. We can combine short-feed and long-lead sequencing to actually identify all these variants very easily. The second is that we have a very large dynamic range compared to bicrylate because we are not limited by the number of oligos that are available for binding or hybridization. We can sequence if you have 10000 copies. We will sequence the sequence data, which will be proportional to that. So, this will give us this flexibility right? From very low expression to very high expression. And this means we can detect these rare and low-abundance transcripts quite easily. Now, coming to the challenges, of course, it is not perfect. There are a lot of challenges when you are dealing with this RNA sequencing data, and some of them can be addressed when we are doing the bioinformatic analysis. So, there are biases in the data, right? So, as we have mentioned, we have something called amplification bias, or selection bias. So, all these things come into play. This is something we will discuss in much more detail in the next few classes, and we will talk about the methods that can address some of these biases. It can be expensive and time-consuming RNA sequencing because you have to sequence a lot of samples to quite a bit of depth, as I mentioned, 30 to 40 million reads or even more if you want to identify these very lowly expressed genes. So, this could be expensive, and the preparation of the library could be time-consuming. One of the things that RNA sequencing still does not do very well is the transcription starts like mapping, and because of these amplification steps or even sometimes direct RNA sequencing, you do not get information about this transcription start site. And this is something that can also now be addressed by devising different ways of protocols, ok? Here are the references that we have used for this class, and to summarize, we have talked about transcript analysis through RNA sequencing, and we have compared against microRNA. We have seen that there are several advantages of RNA sequencing, which is why we do RNA sequencing these days more regularly than microRNA. We talked about short read, long read, and direct RNA sequencing. We talked about the advantages of each of these methods as well as the drawbacks. And perhaps we can combine these methods—short read, long read or short read direct RNA—to address some of the limitations that each of these methods has. So, finally, we can use RNA sequencing to detect isoforms. We talked about the different types of isoforms that are there and novel transcripts. And finally, we can also do small RNA sequencing, which can help us understand the role of small RNA molecules in gene expression regulation. Thank you.