Next Generation Sequencing Technologies: Data Analysis and Applications

Introduction

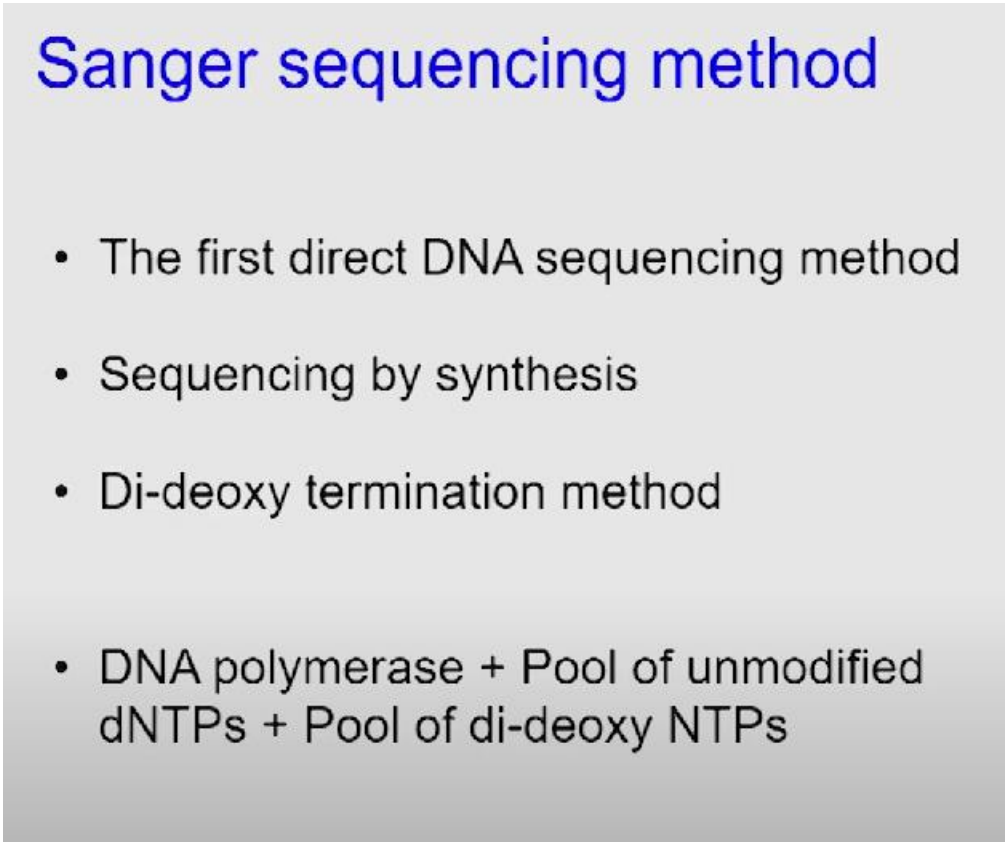
Dr. Riddhiman Dhar, **Department of Biotechnology**,
Indian Institute of Technology, Kharagpur

Good day, everyone. Welcome to the first class of the course. I am going to talk about the next generation of sequencing technologies, data analysis, and applications. In this class, we will introduce some of the methods that were used before the next generation of sequencing technologies came about. This will help you understand how next-generation sequencing technologies have impacted the fields of genomics, transcriptomics, metagenomics, and epigenomics. And following that, we will also introduce next-generation sequencing, and in subsequent classes, we will talk about these technologies one after another.

So, let's begin, ah. This is the agenda for today's class. So, we will start with the Sanger sequencing method. So, this is the first direct sequencing method that was available.

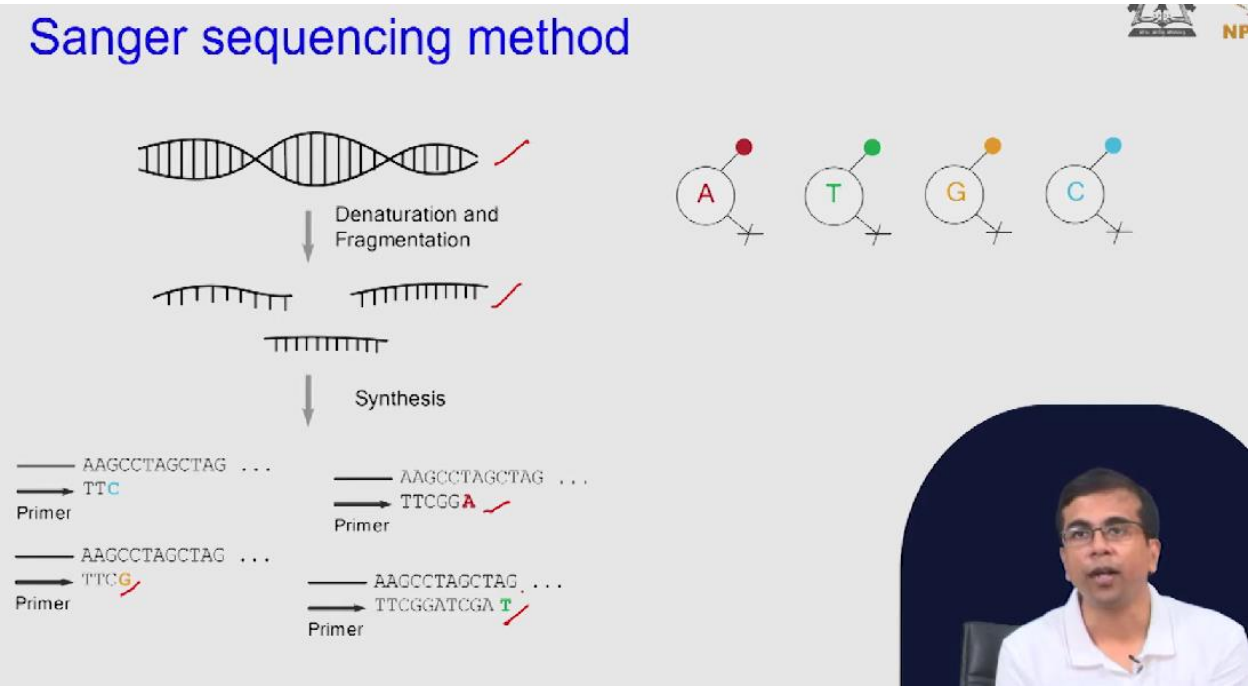
We will talk about microarray technology. This is a high-throughput method that actually allows some degree of mutation identification and gene expression analysis. Then we will introduce next-generation sequencing. The idea of next-generation sequencing will not go into the methods in this class. And then finally, we will talk about some applications so that you understand how these next-generation sequencing methods have actually impacted our lives. Ok. So, let us start with Sanger sequencing. So, these are the keywords you will come across: di-deoxy termination reaction, sequencing by synthesis, DNA probes, and high-throughput sequencing. So, Sanger sequencing is the first direct sequencing method that was proposed in 1977. And for that, there was a Nobel Prize because this method had so much impact, which allowed us to actually first look at the bases of DNA molecules. So, we could read these bases now, one by one, ok? So, how does this method work? So, this is a sequencing by synthesis method, and that means we need to synthesize the complementary strand of a molecule that we want to sequence. So, imagine I give you a DNA molecule and say, Let us find out the sequence of that molecule. So, what we have to do is synthesize the complementary strand of that molecule, right? So, for that, we need polymerase, the dNTPs, right? You know what kind of process we need to do. So, we need these molecules to actually synthesize the complementary strand. Now, compared to normal

amplification and normal synthesis, Sanger sequencing utilizes something called the di-deoxytermination method. So, how does this di-deoxy termination work? So, it works by having certain modified nucleotides, which will stop the reaction. So, we will illustrate this bit in this class.



So, what we have is the DNA polymerase, right? We need to synthesize the complementary strand, and we need primers, right? So, that will bind, and then we will synthesize the complementary strand; we will have a pool of unmodified dNTPs. So, these are normal dNTPs that we use in the lab. So, we will have dNTPs, which means dATPs, dGTP, dCTP, and dTTP, plus we have a pool of di-deoxy NTPs. So, these are the di-deoxy terminators, right? Once they are incorporated into the synthesized strand, this will block further synthesis, ok? So, further extension would be blocked, and then there will be some other process that will be done, ok? Now, what should be the ratio of these unmodified dNTPs to the di-deoxy dNTPs? There are specific ratios that we have to follow, again depending on the length of the molecule that we are sequencing. So, again, there is a specific ratio that we follow in this process. So, just to illustrate this now, So, these are the di-

deoxy NTPs. Ah, I have just put them in some schematics. So, we have 4 types: A,T,G, and C.So, each of them ah, they are tagged ah by some fluorescent color. So, they have all these different colors, ah? You can see orange, green, yellow and blue, ah? So, this is again a schematic, of course; these are not the exact colors that are given, but they are different. So, that is the point. So, what we do is, if we have genomic DNA, So, the first step is to denature and fragment the DNA. So, we cannot synthesize or sequence the full genome, right? If we are, let's say, talking about the human genome, we cannot directly sequence the full genome in one go. So, we have to do this in a step-by-step manner.



So, what we do is fragment the DNA, denature it, and fragment it. So, we get single-stranded fragments, right? So, we are representing the single-stranded fragments, and here we get multiple such fragments. Now, we can take one such fragment, and we can do the Sanger sequencing. So, that is how this process works, okay? So, what we do is imagine we have this strand and we have this primary binding site for which, ah, we know the sequence right. So, we have to know a part of the sequence to design this primer, and then we synthesize the complementary strand.

So, you can see this complementary strand being synthesized. Now, what happens is that when these normal dNTPs are incorporated, the normal ah, extension is going on, and you are getting a

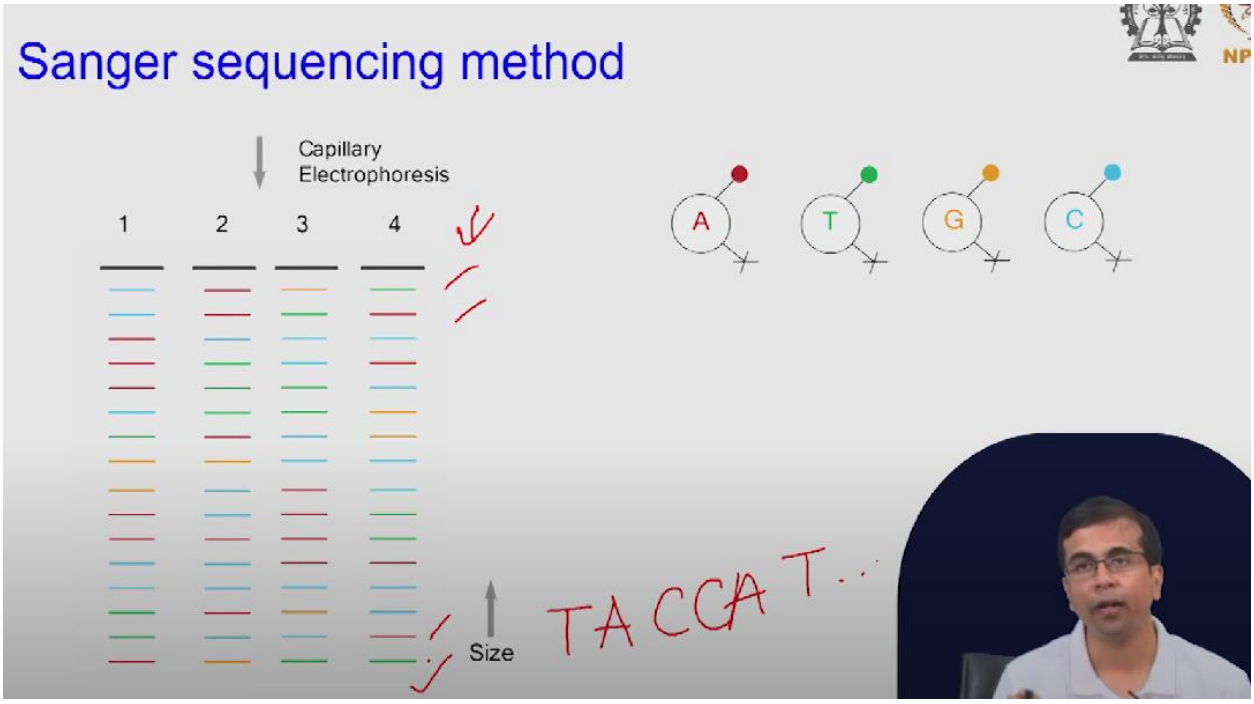
growing strand of DNA, but the moment a modified dNTP is incorporated, then the synthesis stops. So, here you can see the right we have this TTC and so this is C is a modified one right and it will stop the extension, ok? And similarly here we have A. So, it will stop the extension. Here we have J, right? Ah, this will stop the extension. Here we have T. It will stop the extension.

Now, this is again by chance, right? In some cases, we will get these ddNTPs right, but most of the time, we will get normal NTPs and DNTPs right. So, this is where the ratio of dNTP and ddNTP plays a very important role. So, this has been optimized through trial and error. Now, for the moment, this stops, right? What you can do now is take these different fragments and try to separate them out and read the signal, ok? So, you can imagine that if you have this fragment of, let us say, 100 base pairs, you will generate a lot of these fragments of different sizes right with normal dNTPs, but at the end, one modified NTP or ddNTP is ok.

So, what we do is then take these samples and run them through capillary electrophoresis. This is like your DNA gel, but of course, running through a capillary is not just the agarose; it is a different electrophoretic medium. And here you can see 4 samples running 1 2 3 4 right, and what will happen is that you get this fragment generated from each sample right of different sizes; they will be separated out based on size. So, the smaller fragments will run faster through this electrophoretic medium and the bigger fragments will run slower right. So, if you look from so, if you are running in this direction, right? So, after some time, what you will see is that the smaller fragments will come down here, right, and the bigger fragments will remain here.

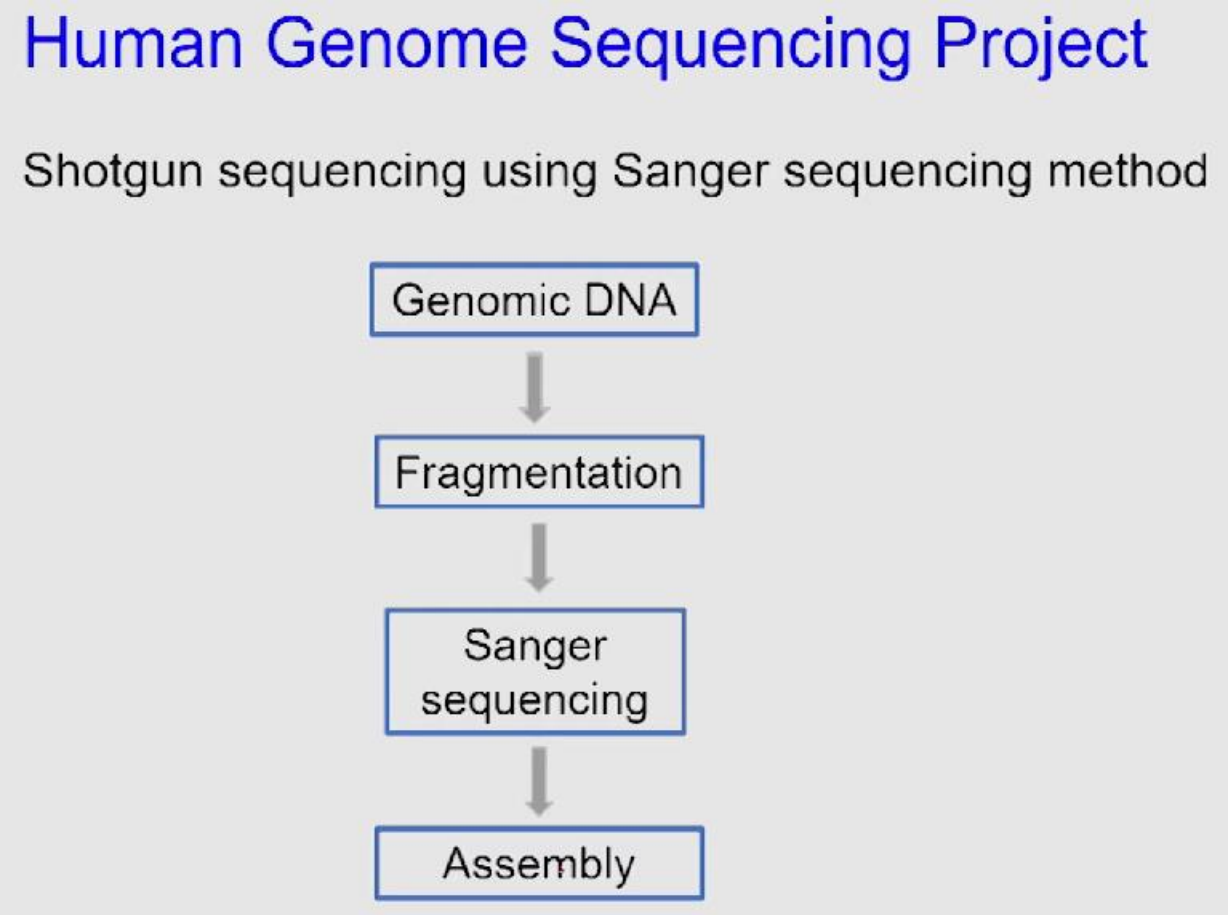
Now, what can you do once these fragments are separated? We can now read the fluorescence signal of each of these fragments. So, let's imagine the fluorescence signal here is green, the lowest fragment here is green, and the next one is red, so on the right. So, we are reading these signals, and by reading this signal, we know what the base is that is present at the end of that fragment. So, the in this fragment we know the green.

So, green is T-right. So, what we will get is T, then what we will get is a red, then you have blue, 1 blue 2 blue, right? So, 2 Cs, then again red, a, then green T, and so on.



So, for each lane, we can now read this and figure out what the sequence of these fragments is because these two neighboring fragments differ by just one nucleotide difference because there is nothing in between them. So, this is how the sequencing is done right. So, we can now read the sequence from this process, ok? So, this is the method that was actually applied for doing the human genome sequencing project, right? So, that started quite a while ago, and they employed a method called shotgun sequencing, but they used sanger sequencing to actually sequence the genome. This was a very important project because we thought that by looking at the genome sequence, we would understand the functions of genes, which genes are responsible for which types of diseases, etcetera. So, that is why this was a very important project at that point, ok? So, to do so, they took genomic DNA from an individual. So, they isolated genomic DNA, did the fragmentation right, and generated a lot of these fragments. So, you can imagine, right? So, the human genome size is 3 times 10 to the power of 9. So, with 3 billion bases, you can get an enormous number of fragments, and you can start sequencing these fragments. So, again, some fragments will be present multiple times, and you can process each of these fragments for sequencing. And once you get these fragments what you have to do is generate this full assembly of the genome. So, the fragments are only part of the genome, right? You need to assemble the genome, right? So, our genome is organized in terms of chromosomes, right? So, I have chromosomes 1, 2, etcetera. So, we have to actually build the assembly and say, "Okay, this is

chromosome 1, this is chromosome 2, and so on.

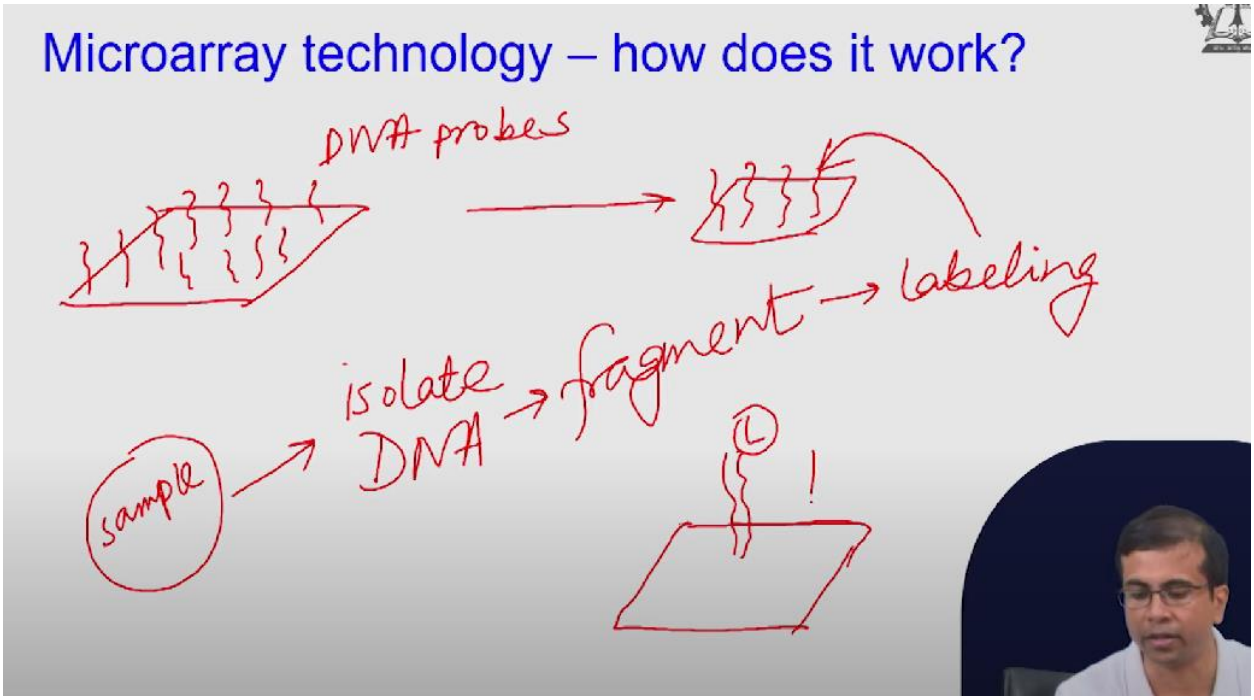


As you can imagine, we are sequencing these DNA fragments one by one. So, in parallel, maybe you can do 96 or so at most. So, it will take an enormous amount of time to actually sequence all those fragments, and that is what actually happened. So, this was a very time-consuming project, and it was very expensive, ok? Because of the sequencing, you have to process these fragments one at a time, and you cannot parallelize this process. So, this was completed. The human genome sequencing project was completed in 2003, and it took about 13 years to sequence just one genome. So, that is a very, very long time, and the cost was huge; it was US dollars in terms. So, it was 3 billion at that time. So, you can imagine how much money actually went into just sequencing one genome. So, you can probably now understand the limitations of Sanger sequencing if you want to do genome sequencing or something like that. So, this is where the limitations came into play, right? So, the first limitation of sanger sequencing is that it has low throughput. So, you cannot process too many DNA fragments at once. So, we saw that if sequencing is done for one sample, you can maybe get 96 samples at once in the machine, right? So, this is a very low-throughput

method. To generate these fragments in the lab, you have to clone them and then do the sequencing reactions, right? This is the dideoxy termination reaction for each of these fragments, and then run through capillary ectoposies. So, this is a very, very tedious process, and it takes a lot of time.

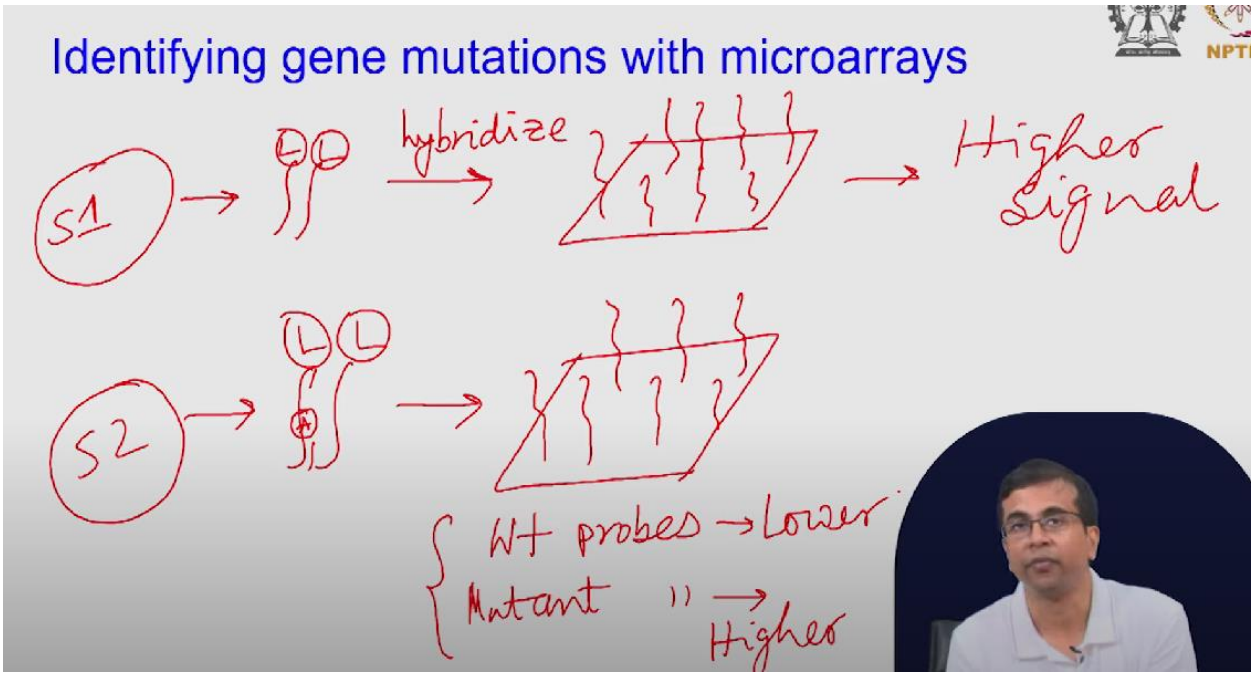
So, when you apply this, as we have just seen, when you apply sanger sequencing to a very large-scale genome project, this becomes very expensive, and it takes a lot of time. So, this is probably not the way you want to sequence not just human genomes but also other species right around us—plant genomes, etc.—crop genomes. So, this is this is this is a big limitation of the sanger sequencing project, ok? So, what have people thought about? Maybe we can speed up this by using something called a microarray. Of course, this is not the solution to sequencing new genomes, but you can at least identify mutations, do genotyping, and also do some sort of gene expression study using this microarray technology. So, what microarray technology does is use something called DNA probes, which are present on solid surfaces. So, these DNA probes are designed so that they are complementary to the sequences in a genome, okay? So, we will discuss this in a moment. So, what happens is that you have these probes, you take genomic DNA, and you hybridize, and by looking at this hybridization process, you can identify whether there are mutations or genetic variation in your sample compared to the reference that is present in the DNA probe. The limitation is that the sequence has to be known, right? So, if you want to synthesize and design these DNA probes, you have to know the sequence of the sample that you are working with. So, if you are working with the human genome, you have to know the sequence first before you can design these probes. So, you see this problem immediately, right? So, you cannot sequence new genomes, right? So, you cannot work with genomes for which you have no sequence data. Nevertheless, people use this method quite a bit for genome typing and gene expression studies because it was a high-throughput method that was available this time. So, how does it work? So, we will just illustrate this now. So, as I mentioned, you have a solid surface, ok, and on top of that, you have these DNA probes. You have many of them, right? I am just drawing a few, but you have 100,000 of them, ok. So, again, you can also control how many you can have. So, you can have between 100,000 and millions, ok? So, you can have these probes that are complementary to some region of the genome sequence that you are working with, ok? And what you do now, if you imagine this is your sample, you have the cell right or some cells, like clusters of cells, you isolate DNA fragments right. We have seen this in the case of Sanger sequencing, and you label this with some sort of

fluorescent label right. So, you can read the signal OK, and then you add this to this array OK. So, after leveling, you add these DNA fragments to the array, and there is hybridization.



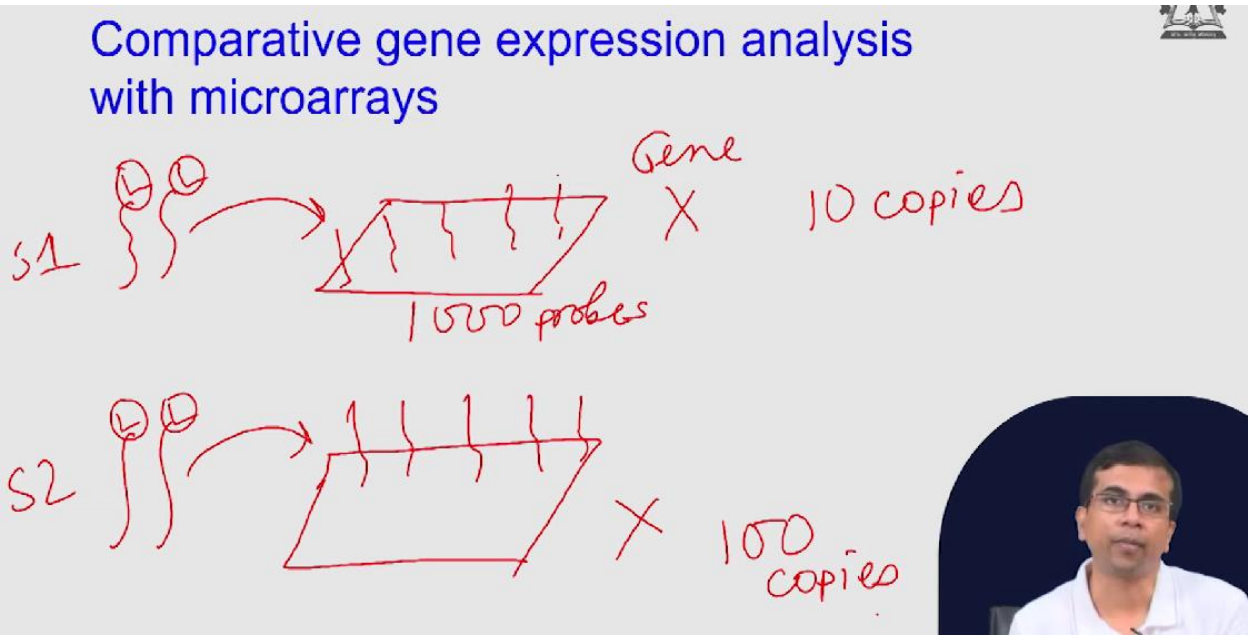
So, what does it mean, right? So, if you have the complementary sequence right, you will see the formation of double strands. So, double-stranded DNA molecules will form, and with this formation right, these levels will be retained on the array right. So, this is the process of hybridization, right? So, you have these level molecules that are now allowed to bind to these probe DNA probes here, and some of these fragments will form these double-stranded DNA molecules on this chip or solid surface, ok? So, these double-stranded molecules will be there, and they will carry the levels right.

So, you have this double-stranded molecule formation, and they have this level L right. So, these levels can then be read by a detector. So, there is a detector that will actually read these levels correctly and identify the signal.



So, how do you actually then use this method to identify mutations? So, what you will see, let us say, let us imagine these two samples, right? So, you have sample 1 and sample 2, and you want to identify which sample contains some mutations in the DNA. So, we have these DNA fragments right at the levels that we have generated from both samples, and then they are hybridized against these solid surfaces. So, these DNA probes are there. So, you do this hybridization. Now, imagine this sample S2; it has a mutation here, right? So, let us say there is a mutation here, ok, in this position. So, what will happen is that since sample 1 is right these DNA molecules are perfectly complementary to the DNA probes on the array. So, there will be strong hybridization, okay, because there is perfect complementarity. So, this will give a very strong signal compared to that because there is a mutation, which will disturb this hybridization process because the mutation is not complementary to the part before and after this mutation, which is complementary to the probe, but this mutation is not. So, this will disturb the hybridization process, which means you will get a lower signal. So, you will get a higher signal here right? In case you have stronger hybridization, you will get a higher signal. Now, you can also do something else, right? So, in these arrays, when you are designing this DNA probe, you can guess, ok, maybe we will see this mutation ok and design probes that are complementary against this mutation ok. So, in the array, you have both types of probes, one complementary to the wild-type sequence and the other set of probes that are complementary to a mutant sequence. So, you can have these probes: wild-type probes and mutant

probes. So, what you will see is that if there is a mutation in the sample, this mutant probe will give you a higher signal and the wild-type probe will give you a lower signal. So, now, if you compare this, So, you have both types of probes in the microarray, and you are comparing these S 1 and S 2 right. S 1 will give you a higher signal for a wild-type probe and a lower signal for a mutant probe, because there will be weak hybridization with the mutant probe, and in the case of S 2, you will see exactly the opposite. So, you will get a lower signal for a wild-type probe; you will have weaker hybridization and you will see a stronger signal or a higher signal for a mutant probe because there is very strong hybridization. So, in this way, you can take a guess and design probes that will help you identify gene mutations or genotypes. So, this is what researchers did with microarray technology. You can also study gene expression using these microarrays. So, we will illustrate this very briefly. So, again, we have this sample 1 right. I am just drawing here again, and they are now bound to these levels, which gives us signals, ok? And in microarray technology what you will do is you will. So, for gene expression analysis, you will identify or measure the expression levels of genes between these 2 samples. So, what we will do is isolate RNA, then convert that to cDNA, and hybridize that cDNA to the microarray. So, this microarray is now designed with oligos and probes that are complementary to mostly cDNA sequences, not against the whole genome but against only the gene sequences, because the imidinase will come from the coding sequence or the gene word. So, we design this probe so that it is only complementary to the cDNA. So, we use the same set of probes against these 2 samples, and we hybridize them. This is the hybridization process again. So, once we go for hybridization, what will happen? So, let us imagine we are measuring the expression level of a single gene, X. So, let us say probes against this gene X are right in the microarray chip or on the array. So, these probes are present in multiple copies—maybe in hundreds or thousands of copies. So, what will happen?



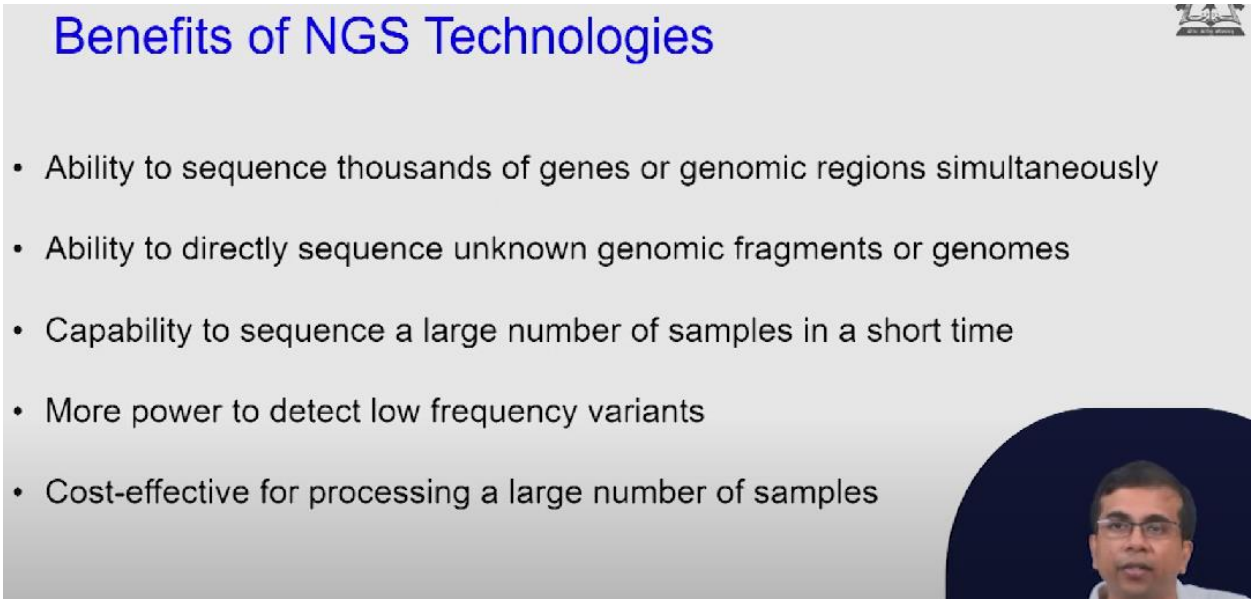
Let us say in this sample 1 that gene X is expressed at a level of only 10 copies, and in sample 2, you have only 10 copies. So, here you have only 10 copies of this mRNA produced and in this sample 2, you will get 100 copies of the mRNA produced. So, in the first hybridization you will see only 10 copies. So, in the probe array, you have 1000 such probes against this gene X, ok? Now, if only 10 copies are produced, not all of these probes will be occupied or hybridized; only a fraction of them will be hybridized. So, the signal will be proportional to this number of copies. In the case of sample 2, you have a higher number of copies, 100 copies. So, a larger fraction of these probes will be hybridized by the sample. So, as you can see now, because there are a larger number of copies, what will happen in the case of sample 2 is that you will get a stronger signal, ok? And this way, it will tell us, ok, in sample 2, you have a higher expression of this gene X compared to sample 1, right? So, this can now be repeated for multiple genes, right? So, as I said, you can have millions of probes. So, you can do this for 5000 genes or 10000 genes, and you can measure the expression levels of these genes and compare the expression levels of genes between these 2 samples. So, as you can probably now imagine, you need these different types of arrays for doing these two different types of analysis. So, you have something called a snp array that will help you identify these mutations or genotypes, and you also need something called an expression array that has to be designed for gene expression analysis. These arrays would have to be organism-specific, as, as you have seen, they require these DNA probes, for which we need to know the organism sequence, the genome sequence, and that these sequences are different. So, we need

these organism-specific arrays. And you can also design customized arrays. You can have your DNA molecules that you want to probe right here that you want to see or that you want to design as probes. You can also do this customization and design your own array and you can print those probes on these chips. So, with this discussion, you can probably now understand the drawbacks of microarray technology.

So, it only works with known sequences; you cannot sequence anything that is unknown to you. So, you first need to figure out this genome sequence or trans cDNA sequence before you can use microarray technology. It is not capable of identifying new genes or sequencing new genomes; this is obvious, as you have seen, and this is an indirect method, right? So, we are not directly looking at the sequence of these DNA molecules; we are just looking at the hybridization process, which is dependent on a lot of parameters. So, we have not discussed all the details of the conditions in which we do this hybridization.

So, these conditions are also very important in determining the hybridization efficiency. So, if you have a slight modification or fluctuation in this hybridization condition, that can tremendously impact this hybridization efficiency which means the signals will be affected by this hybridization. So, this leads to reproducibility issues. So, if you are doing this experiment on different days at different times and there are slight fluctuations in your sample preparation or the salt concentration is slightly higher or lower, this can lead to different signals, which can lead to reproducibility issues. So, this actually led to the emergence of next-generation sequencing technologies.

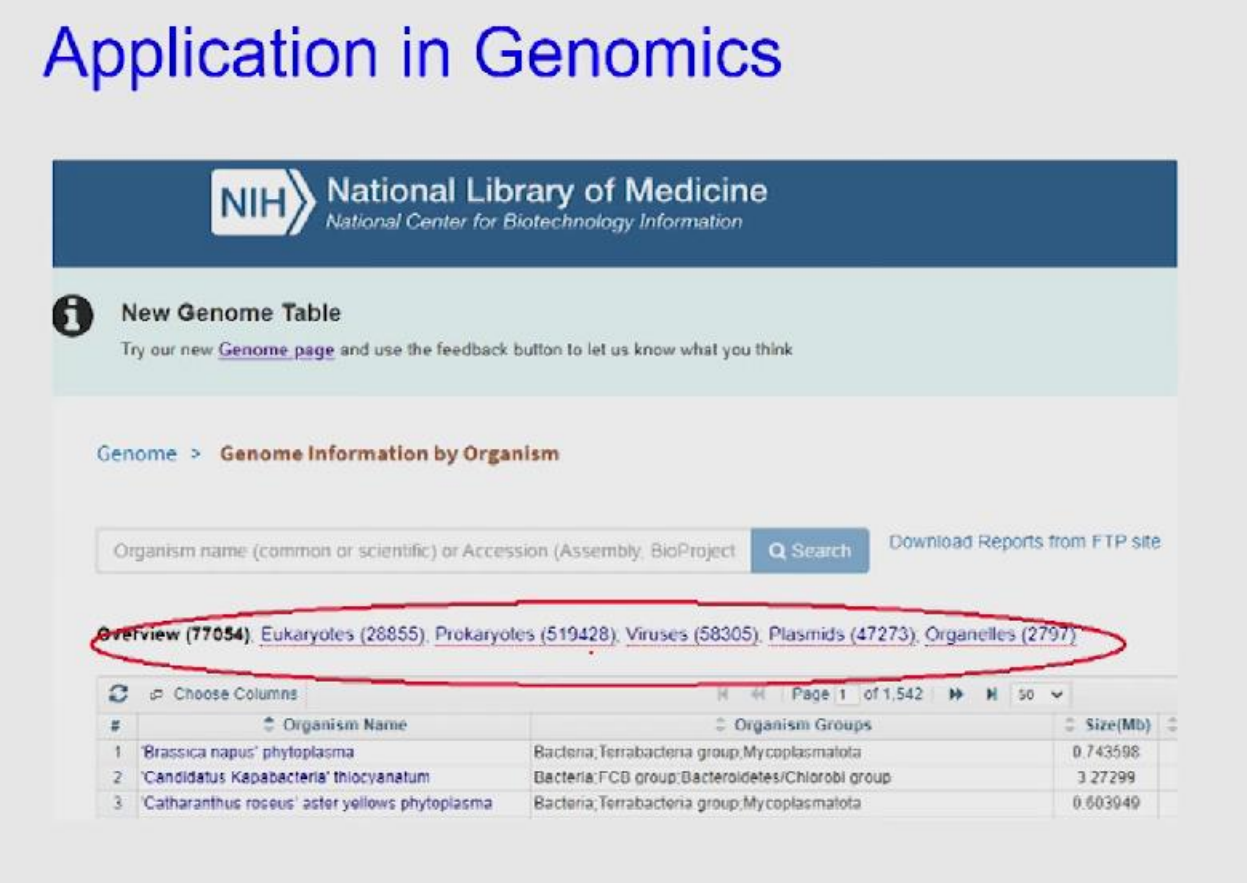
So, what was needed was a direct sequencing method. As you have seen, microarray is an indirect method, and that is actually not good enough. So, we needed a direct sequencing method, and it should be high-throughput, accurate, and reproducible. There should not be too much error when you are sequencing these DNA molecules. It should be cost-effective for large-scale projects, right? And how do you define this cost-effectiveness? The target was said to be that the human genome should be sequenced for, ah, 1000 dollars or less, ok? So, we are there; actually, we have already reached that goal, and the goal is to actually take this cost down further. So, what are the benefits of NGS technology?



So, NGS technology is simply scaling up the whole process of sequencing right and making it high throughput right. So, what it does is sequence 1000s of genes or genomic regions simultaneously instead of just one gene fragment that we have seen in Sanger sequencing. It can, ah, sequence unknown genomic fragments because this is a direct sequencing method that does not rely on hybridization. It can process a large number of samples in a very short period of time. So, you do not have to wait 10–13 years to get one human genome; you can do this now in 7–10 days.

You have more power to detect low-frequency variance. So, we will see what this low frequency variance means when you actually go into the topic. So, these are actually genomic ah, genomic variations that are present in population ah, and are present in ah, low frequency or ah, they are there in the population. And this is why these methods are also cost-effective for processing a large number of samples. So, coming to the application, So, NGS technologies have had a huge impact on these different fields of genomics, transcriptomics, and epigenomics.

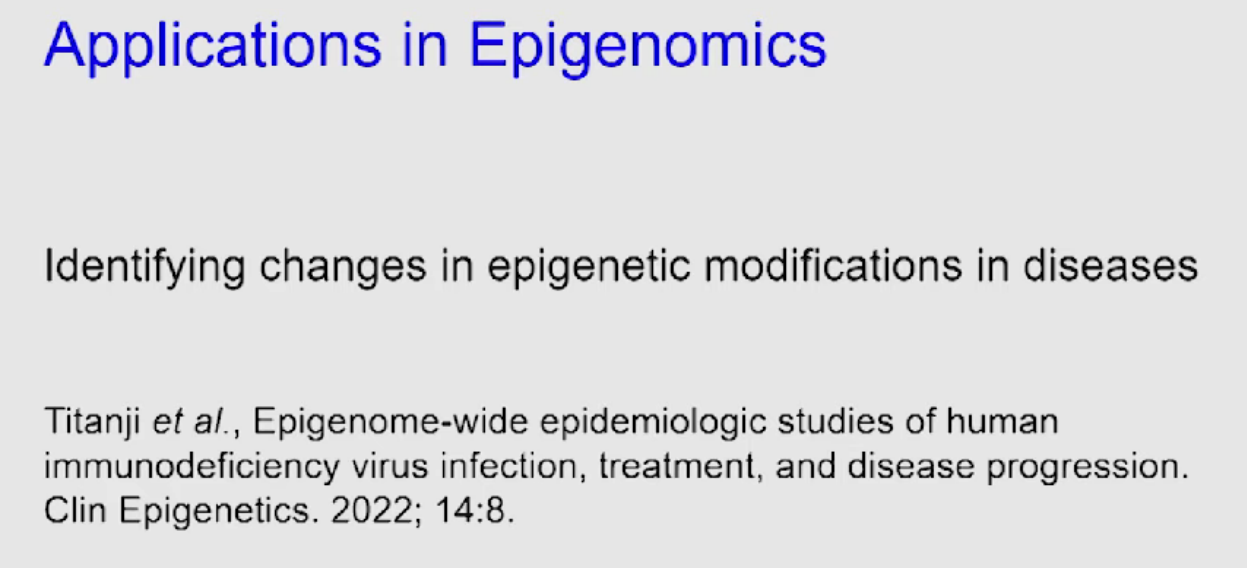
So, I will just very briefly mention and of course, ah, you are welcome to go and read a bit more through the references. So, here are some statistics that I will give you: here is the number of eukaryotes or genomes that have been deposited in this set. You can see these numbers right here. So, eukaryotes, prokaryotes, and viruses, right? So, this has seen tremendous growth in the last few years.



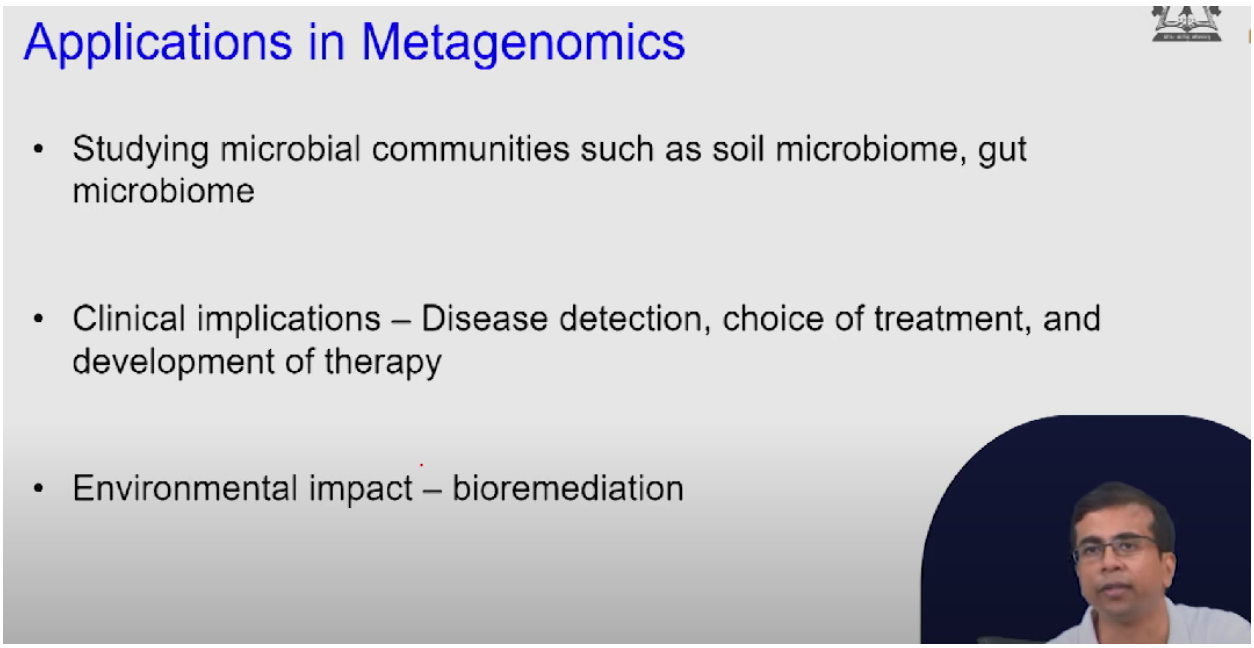
So, in the last 10–15 years we have, we have seen, ah, 300 genomes ah, being deposited per month ah, to these databases. So, so this is a direct contribution of the next generation sequencing because it has sped up the process of genome sequencing quite significantly compared to single sequencing. It can be applied in transcriptomics, and we are seeing this application across many different fields. One of the most important studies in transcriptomics is to look at disease samples and try to understand how gene expression changes between disease and normal samples. So, this can help us gain insights into the disease, process, etc. So, this is very important, and this has actually been done for, ah, many diseases now. So, what happens in transcriptomics is that we measure the expression level of all genes, and we sequence all mRNA molecules. In transcriptomics cells, of course, we do not directly sequence mRNA. There are specific processes that we go through that allow us to discover novel transcripts, and there are probably transcript variations that can happen from normal to disease, samples, and we can identify those, and we can also see something called alternative splicing. So, we can look at different spliced variants through this application of NGS in transcriptomics, which was earlier very difficult to do with microarray because you did not

know what kind of spliced variants could appear. We could have taken a guess again and designed probes, but that was a very tedious process. So, this has been applied to NGS for disease states in cancer, and there is a whole bunch of data. For different types of cancer, we have this transcriptomic data.

So, it is called the Cancer Genome Atlas Project. So, here is a link where you can actually look into some of these data sets, explore the data sets and ah, you can also read papers ah, looking at the differences between cancer versus normal samples. This has been applied in epigenomics. So, the epigenome, or epigenetics, is the study of modifications to the DNA without any change in sequences. So, this has actually become a very important field now where we can apply nystereogen sequencing technologies to identify these DNA modifications at a very rapid pace. So, and again, this is very important for many human diseases, and we know, for example, that in cancer, changes in the epigenome ah are the ah, key ah, and important drivers of disease, as well as determining ah, whether a therapy works or not.

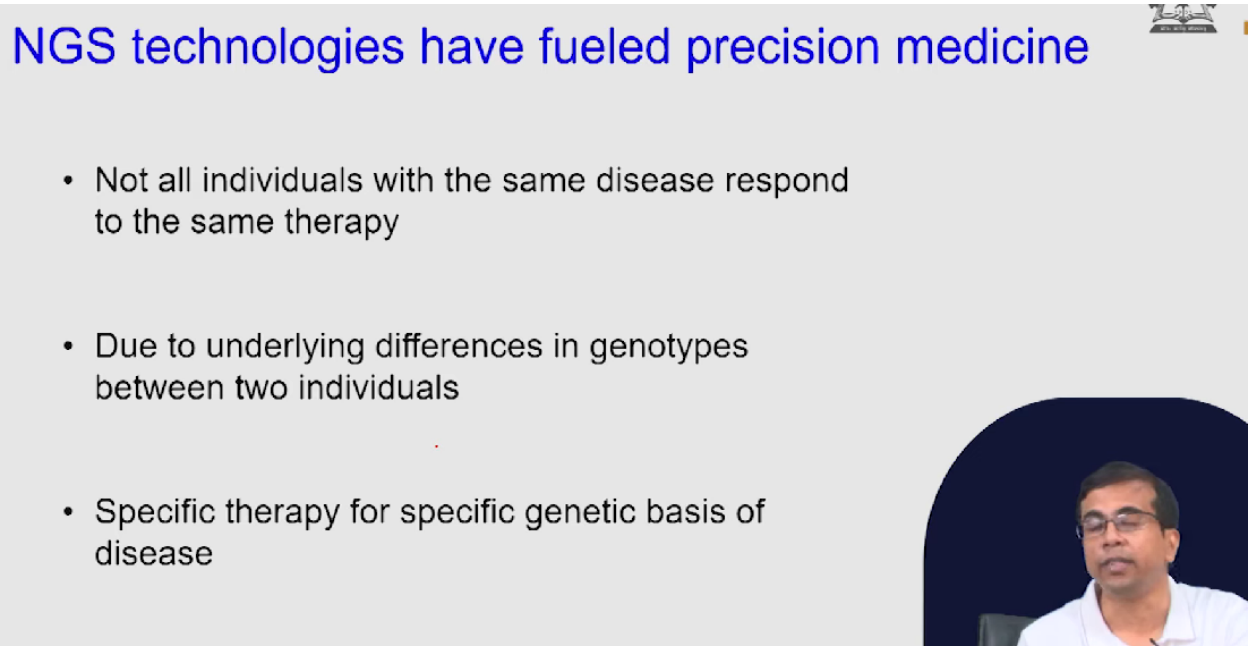


So, here is a reference that you can study, right? So, we are talking about ah, the ah, epigenome studies ah, across human ah, diseases, and finally, these methods have been applied in something called metagenomics.



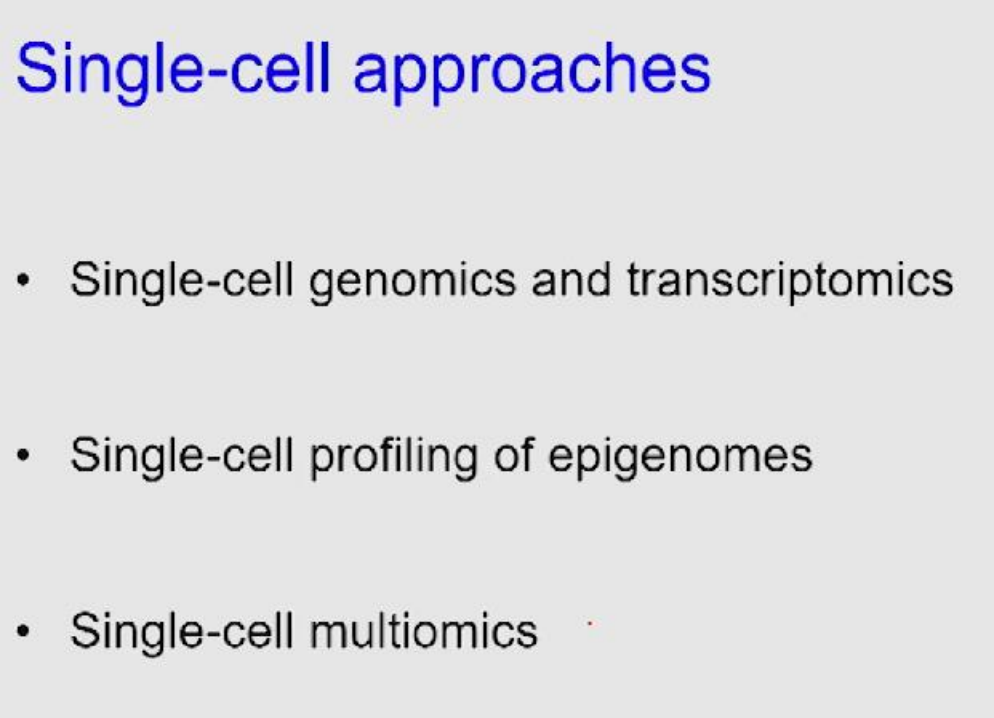
So, again, this is a study of the microbial community. So, for example, we have something called the soil microbiome or gut microbiome. So, we have these microbial species that are living inside our body, and we can study them, and this has very important clinical implications. For example, for disease detection, you can specify or develop novel therapies that can prevent certain types of

diseases. So, probiotics are a direct outcome of this. You can also use this metagenomic study for ah, environmental bioremediation, where you can design communities of microbes to clean up certain pollutants such as plastics, ah, hydrocarbons, etc. Now, with this technology, through the application of genomics and transcriptomics, they have something called precision medicine. So, what you can do is study patients in much more detail than just the symptoms or some traditional diagnostics. We can now actually study the genomes, the genomic changes, and the transcriptomic changes, and we understand that even two people with the same disease may not respond to the same therapy. So, not only all individuals will respond to the same therapy, and this does not work, and this has been seen a lot in cancer, and this is because there are differences in genotypes in 2 individuals, as well as how the disease is progressing, which we can now study through analysis of ah, NGS ah, data, and we can then design specific therapy for a specific genetic basis of disease.



So, this has been, ah, termed precision medicine, right? So, we do not give the same treatment to all patients with the same disease; we can identify subtypes of the genetic basis and specify specific treatment for each patient. Now, again, with all the data that is coming in from all these fields, we can also devise machine learning approaches. So, this is where the ah, data science field also marges with ah, biotechnology and ah, biological sciences. So, we have data sets in genomics, transcriptomics, and epigenomics, and we can take this data and apply machine learning models to utilize it for classification and prediction tasks. This could be, for example, disease subtype classification. I have given you an example here, something like cancer subtype classification from

transcriptome data. So, there are a lot of advantages and benefits that we have seen from NGS technologies, but there are also some limitations that we can probably address. So, one is that it is not as cost-effective for processing the small number of samples. Now, this will probably not be a major problem as the price or cost of these technologies is going down. Now, the second ah, problem is the time. It is very time-consuming right if you are processing a small number of samples, but then you can probably go through the Sanger sequencing route and in addition, with the newer advancements, we are actually making this ah, ah, steps ah, much simpler (the diabetic preparation, etc.) so that ah, we take less time. And these technologies have benefited so much that we can now take this single cell approach where you can actually apply this ah, genomics transcriptome, epigenomics to the study of single cells instead of a full organism or a whole tissue ah, and this has actually given us tremendous insight into the ah, biology of the workings of these biological systems.



These are the references that we have followed for this class, to conclude. So, next session, sequencing technologies have provided really high-throughput, accurate, and cost-effective solutions for direct sequencing. We have seen that they have actually accelerated the pace of new genome sequencing to a great degree.

And finally, we have talked about the applications across diverse fields and how they have provided tremendous benefits in clinical settings. Thank you.