Next Generation Sequencing Technologies: Data Analysis and Applications

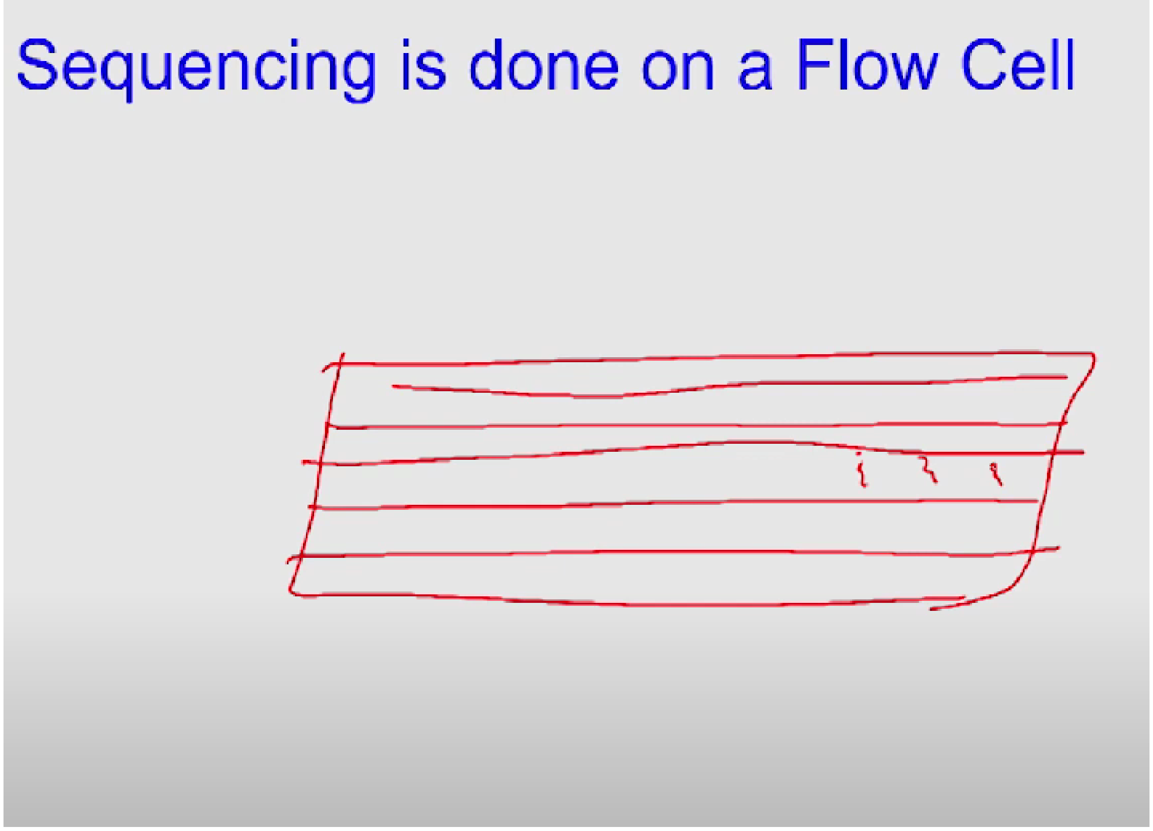
Illumina Sequencing By Synthesis (SBS)

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

Indian Institute of Technology, Kharagpur

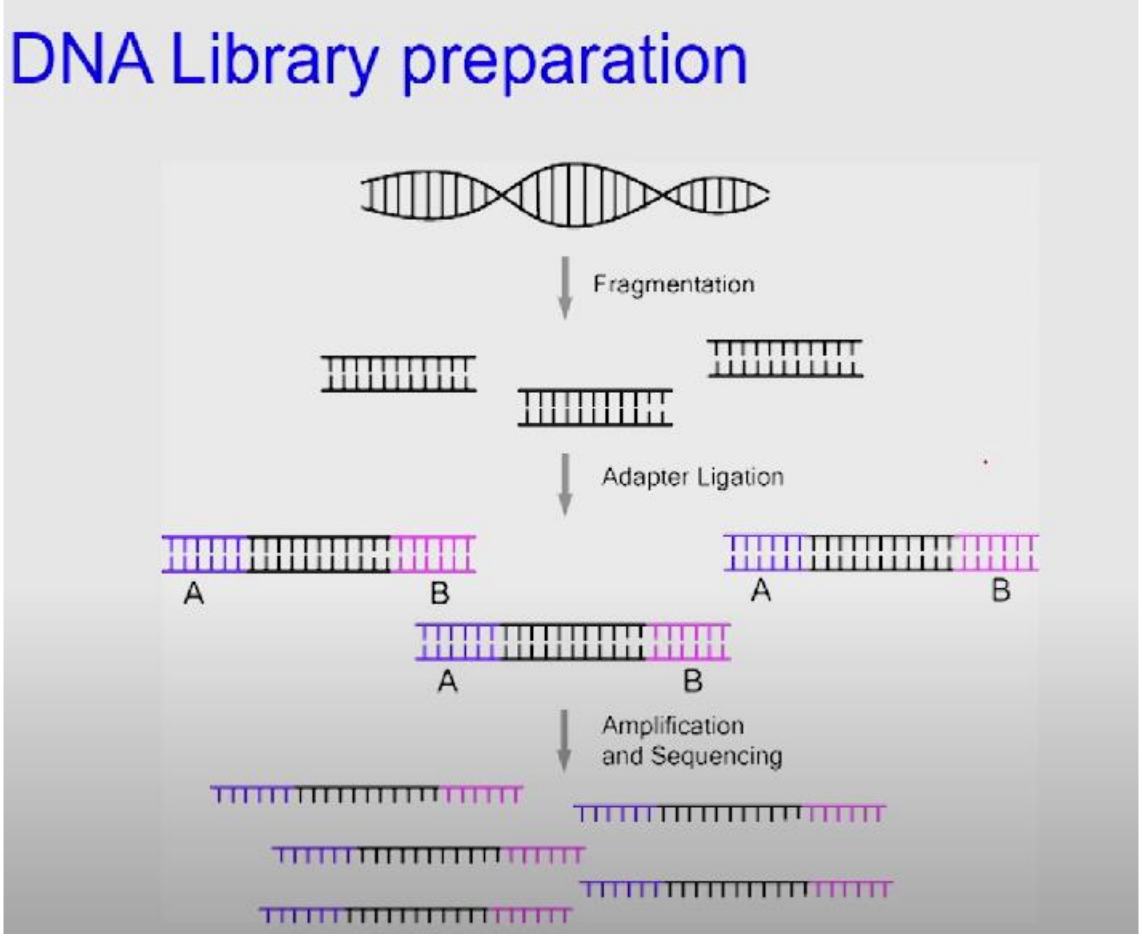
Good day, everyone. Welcome to the course on Generation Sequencing Technology, Data Analysis, and Applications. In the last class, we started our discussion on the first next-generation sequencing technology, which was Roche 454. So, in this class, we will continue our discussion on other next-generation sequencing technologies that are available today, and especially, we will be focusing on the Illumina sequencing by synthesis method. So, this is the agenda for this class. So, we will talk about this Illumina Sequencing by Synthesis technique and how it works.

We will talk about single-end and paired-end sequencing. So, let us start, ah. So, these are the keywords that you will come across. One is reversible termination, bridge amplification, and short- and long-lead sequencing. So, in Illumina sequencing, sequencing is done on a flow cell, not a picotiter plate, but it is a very similar idea. So, what you have there is something like a plate, and you can find a lens here. So, you will see this, ah. If you see a picture of this right, you will have these lanes, and where you can have these DNA molecules that will be bound in these lanes is okay. So, we will talk about this in more detail.



Sequencing is done on a Flow Cell

So, you have a plate-like structure, and then you have lanes inside which all these frequency reactions will happen. So, this is called a flow cell, okay? Now, come to the library preparation. So, this is the typical library preparation as we have already discussed. So, we have DNA fragmentation starting from the genomic sequence; we have adapter ligation adapters A and B that have been ligated to these fragments; and you have amplification and sequencing.

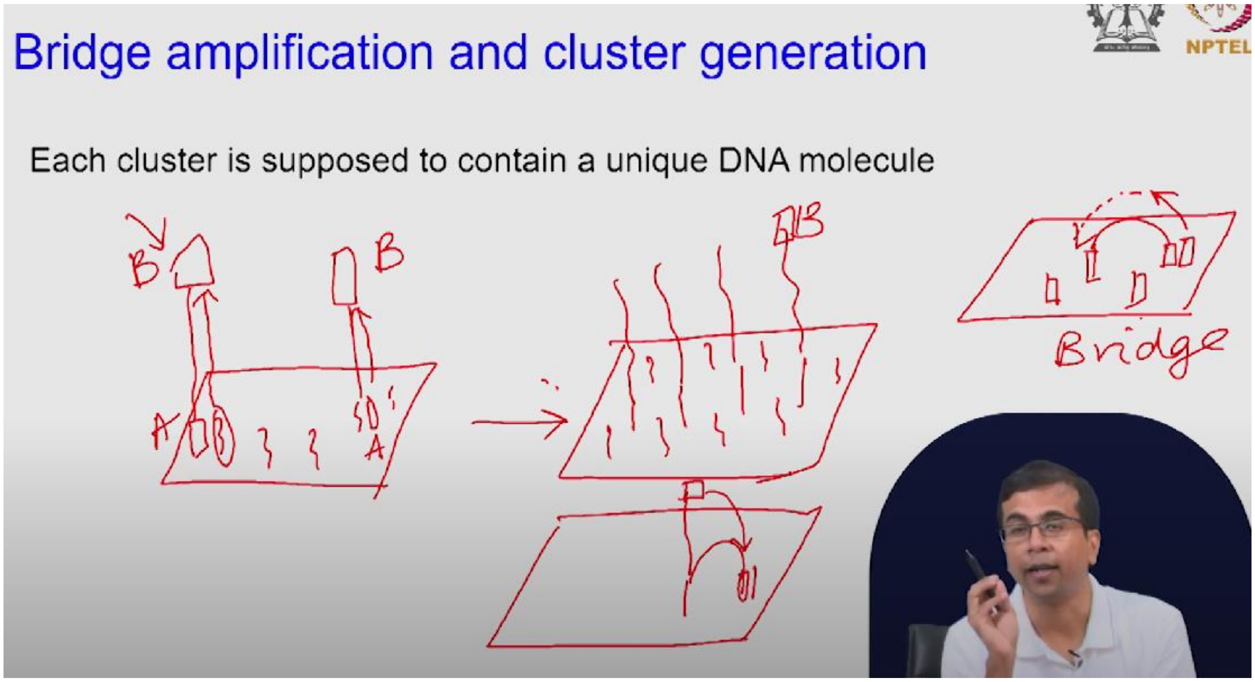


DNA Library preparation

Now, we will talk about this amplification step. As we have seen in Roche 454, this amplification step actually requires something called emulsion PCR. So, this is specific to 454 sequencing. In this method in Illumina, this method is called bridge amplification, and we will see in a moment why this is called bridge amplification. So, let us look at this right.

So, as you have seen, So, what happens in this flow cell inside this flow cell, you can imagine, is that you again have some DNA molecules that are complementary to these adapter sequences. So, like 454 AH, you have certain AH sequences of DNA probes that are complementary to adaptors. So, what happens now is that this DNA molecule with adapters will come and hybridize, right? So, here is again, let us say this will come and hybridize.

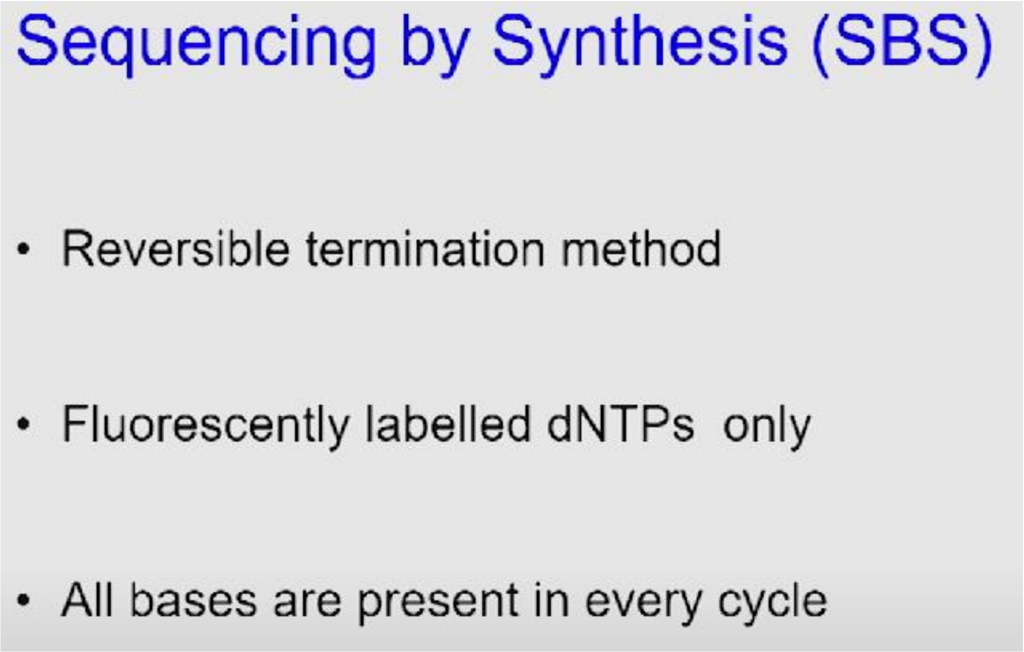
Now, this means you can generate or amplify these fragments using DNA polymers. You can then amplify these fragments using DNA polymers. So, ah, what will happen is that once this is amplified, these molecules can then be released. So, once these fragments are extended starting from this complementary sequence, you can release these fragments. So, what you have after the end of this process is something like this. So, we have these DNA fragments that are generated on this plate, and again, we can imagine you have millions of them on this plate. Now, what happens is that again, you need to calibrate the number of DNA molecules that you will allow to hybridize on this plate. In the next step, there is this bridge amplification. So, what is this idea? What is the idea behind this bridge amplification?



Bridge amplification and cluster generation

Each cluster is supposed to contain a unique DNA molecule

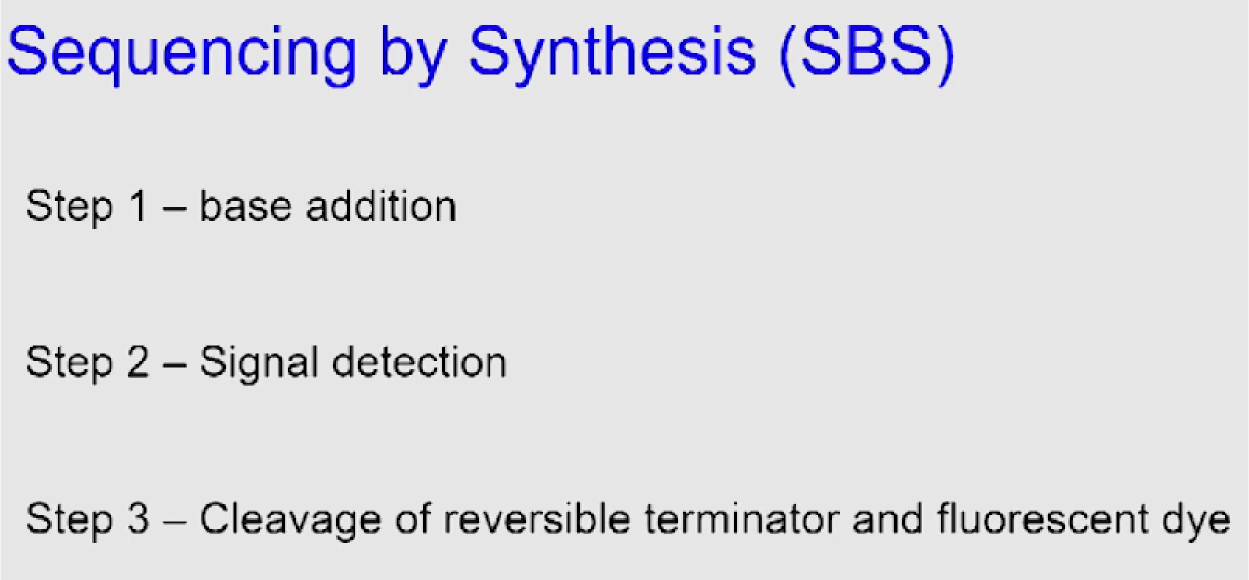
So, you have seen this emulsion PCR; this is a clonal amplification, right? So, a single molecule actually generates multiple copies inside these emulsions, ok? Why do you generate this many copies? Because of the signal detection that is happening, we need a lot more sequencing. So, these detectors in 454 or in Illumina cannot directly identify the signal from a single molecule. So, if you have clonal molecules, they will be synthesized, and the complementary strands to those molecules will be synthesized at the same time because they contain the same sequence, and that signal comes from that well in 454. Similarly, in Illumina, you need to generate these clonal molecules, which we call clusters. So, how do you generate these clusters? We do this through a process called bridge amplification. Now, what happens? We have these amplifications happening right here. We have generated these DNA fragments on this flow cell, but surrounding that, you will also have some DNA probes that will be complementary to the other part of this DNA fragment. So, remember, the other part contains this other adapter, right? So, if you imagine this is adapter A, this is adapter B. So, here, adapter A is adapter B, right? So, the other part contains adapter B. Again, we know the sequence of this adapter B. So, you can have these DNA probes or oligos on the flow cell that are complementary to this adapter. So, if that is the case, then what will happen in this scenario is that we will focus on one molecule just here. So, this will come and it will fold here and it will kind of show something like a bridge structure, ok? So, this will actually come and bend and will form something like a bridge structure. You can see this kind of structure right? So, this is the bridge structure. So, once you have this bridge structure here, you can again synthesize it, and you can amplify it using DNA polymerase and the dNTPs. This is called bridge amplification. Because you have this bridge-like structure, this is called the bridge amplification. So, the idea again is to generate multiple copies of the same clone. So, because you are generating this bridge, So, you can imagine that now that this happens, this molecule will bend, and if there is another complementary adapter here, it will bend and form this bridge, and there will be amplification again. This process is repeated right across all the adapter complementary sequences that are present in the flow cell. So, this is a very small region when this is happening, and you can imagine this is happening in parallel for millions of positions in the flow cell. So, each position will contain clonal copies of the same molecule, ok, and this forms the cluster, ok, and this cluster is where the sequencing reaction will take place, and the signals will come right, and the signal that will come will be the average signal, right, and this would be much stronger because the signal is coming from multiple molecules at a time and these molecules contain the same sequence. So, they are expected to give out the same signal during the sequencing process, okay? So, we have now generated these clusters, and we can now proceed to the sequencing step. So, here the sequencing step again is the sequencing by synthesis method, or SBS, which means we need to synthesize the complementary strand using DNA polymerase and dNTPs.



Sequencing by Synthesis (SBS)

- Reversible termination method
- Fluorescently labelled dNTPs only
- All bases are present in every cycle

Now, again, the dNTPs that are used in this process are not normal dNTPs; these are modified dNTPs, and in such a way that it allows for reversible termination. So, we have seen di-deoxytermination in the case of Sanger sequencing. Here we have reversible termination, which means once these bases bind by this modified basis to the complementary strand that is being synthesized, they will block the synthesis process, but with the subsequent wash and release step, they can be removed again. So, it is not an irreversible termination, right? These terminators their modifications are such that these termination sequences or termination groups can be removed from the dNTPs and the synthesis can go on. So, with this in mind, we also have something called fluorescently labeled dNTPs there. So, you have special dNTPs that contain these terminators as well as fluorescent labels. Again, each dNTP is labeled with a unique color. So, unique fluorescence and all these bases are present in every cycle of the sequencing reaction, ok? So, since each of the dNTPs has its own unique color, they can be identified very easily by the detector during the synthesis process. So, there are multiple steps. The first step is the base addition, which I will illustrate in a moment. So, where these dNTPs are added right and then the next second step is signal detection, where the detector is an optical detector that can read the signal, write the fluorescent signal that is present, and identify the base right. And then the third step is the cleavage of the reversible terminator and fluorescent dye.



Sequencing by Synthesis (SBS)

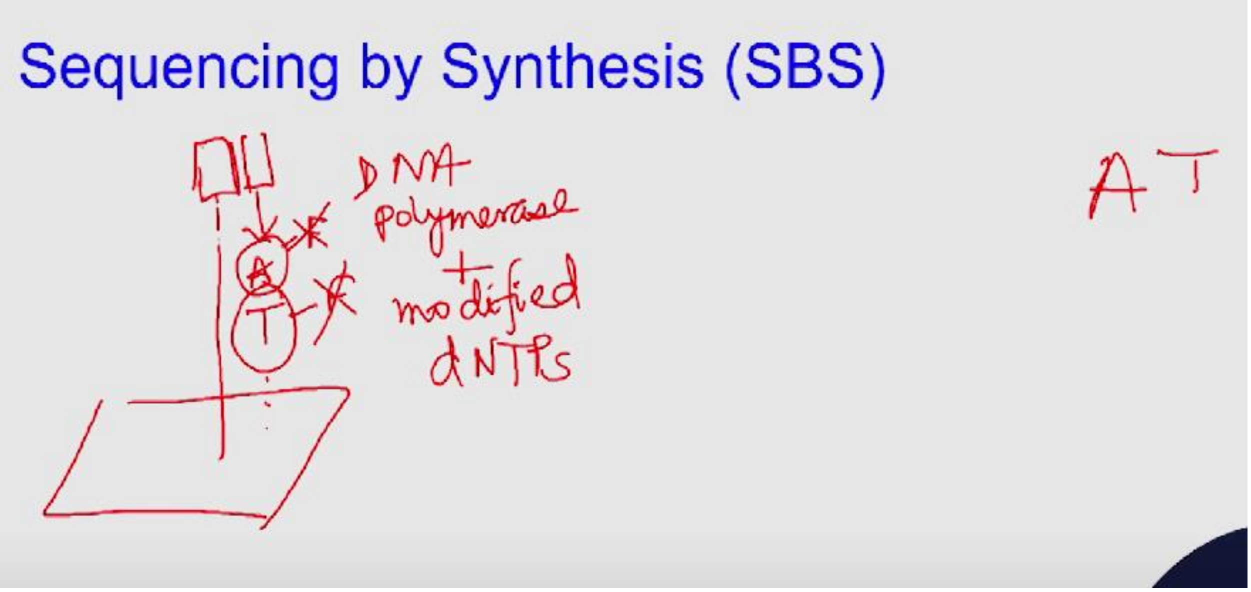
Step 1 – base addition

Step 2 – Signal detection

Step 3 – Cleavage of reversible terminator and fluorescent dye

So, the dye from the already-incorporated base is washed out, ok? So, let us draw this right. So, you can imagine now let us say you have this ah on the flow cell. We just focus on one single molecule, right? We have the adapter here, we have the primer right here, the primer is complementary to the adapter, let us say, and we do the synthesis in this direction, right? So, what we have is polymerase; we have DNA polymerase, and we have modified dNTPs. So, as I have just mentioned these are reversible terminators, right? So, what happens? Let us say there is an addition of A OK and it has a fluorescent signal F OK, and this may be a unique one that is again, if you can imagine writing this, let us say it is red fluorescence. So, once this base A is added, So, the synthesis process will stop, right? We are using this modified dNTP only, not normal dNTP. So, the synthesis process will stop, but there is a fluorescent signal that will be detected by the detector, and this will identify that there is an addition of A. So, the sequence then reads as A ok.

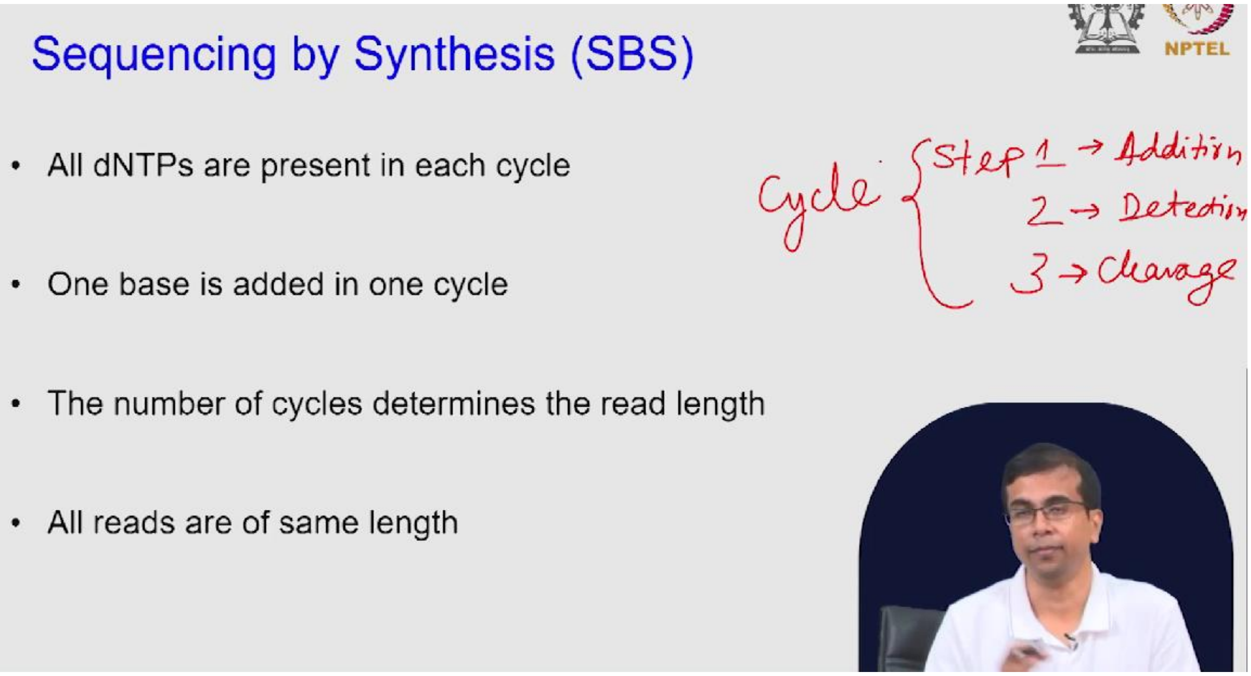
In the next step, this terminator and the fluorescence are washed out ok. So, these groups are cleared. So that they are washed out and there is no fluorescence. So, this is gone, now the terminator is also gone, and let us say the next phase is T, which is added ok, and again, there will be some fluorescence signal again, something different compared to A, but it will also stop the synthesis process and there will be a detection of signal. So, T then you have the washing step right here; this F is gone, there is cleavage, and also the terminator is gone, and the next phase will come in now, ok? So, this process will go on, right? So, you will have G.C., etcetera again, and all the bases are present. So, depending on the complementary sequence, these bases will be incorporated by the DNA polymers. Now, one thing you probably have noticed now is that this base addition is happening one base at a time. Okay, there is no multi-base addition because of this terminator. So, even if you have let us say multiple A's are present, they will not be incorporated at once; they will be incorporated one by one. So, the advantage then is that this will eliminate any problem with homopolymeric stretching. So, even if you have this very long homopolymeric stretch of 25 base pairs, 30 base pairs, or even longer, you have no issue determining the exact length because you are detecting signals ah for one base at a time, ok, and you can now imagine that this process is happening across millions of clusters, ok.



Sequencing by Synthesis (SBS)

So, we have just seen the process for one single cluster here, but this is happening across millions of clusters in the flow cell, and in parallel, you are sequencing millions of fragments and detecting these signals ah for one base at a time, ok. So, that is why this method is very accurate, right? So, you are getting this done in one base-at-a time process, and this makes this process very accurate. So, here are some key points that you need to remember. So, we have just mentioned that all GNTPs are present in each cycle, right? So, all of them are there, and whichever is found to be complementary is incorporated, and this is not an issue because the detector can detect each of these bases by their tags. So, because each base has its own tags, So, the optical detector is there, and it can actually identify this individual bases very easily.

And one point that we have seen is that one base is added in one cycle, ok? You do not have multi-base additions at once, and this is in contrast to 4/5/4 sequencing, where there are no limits. So, you can have 20 base additions, right? So, if you have 20 bases, they will be added one after another very quickly, and then the signal will be read out. So, which will be the total signal intensity, or the light intensity that we get here? We are getting a signal for one base at a time, and this makes for a very accurate sequencing process. What it also means is that the number of cycles will determine the read length. So, you are controlling the number of cycles of this base addition signal detection and cleavage, right? So, this is how we have talked about these 3 steps, right? So, we have steps 1 2 3 right here is the addition base addition we have detection and then you have the cleavage ok. So, we have these 3 steps in one cycle.



Sequencing by Synthesis (SBS)

- All dNTPs are present in each cycle
- One base is added in one cycle
- The number of cycles determines the read length
- All reads are of same length

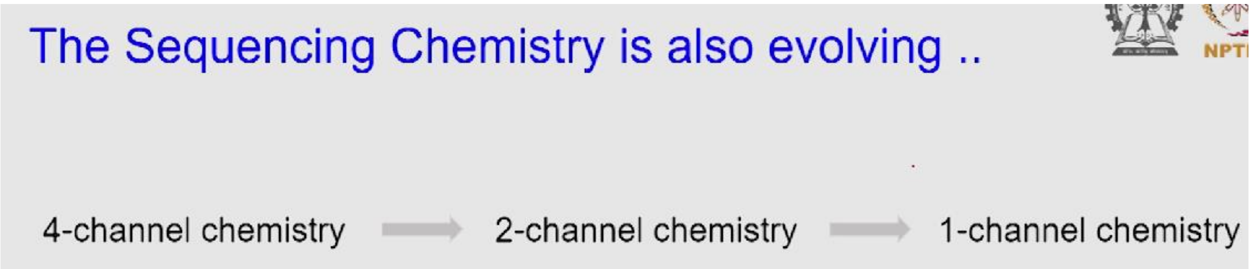
Cycle { Step 1 → Additin
2 → Detectin
3 → Cleavage

So, this constitutes a cycle, right? So, we have this 3-step addition detection cleavage. You can decide how many cycles you want by controlling the cleavage step. For example, if you do not want any more cycles, you just stop there, and there will be no more synthesis after that. So, the number of cycles that you allow will determine the read length that you get. So, in Illumina, it is typically 150, 200, or 250 at most. So, you can control between 50, 150, and 250. So, you can decide, ok, we will stop this after, let us say, 150 cycles. So, depending on the machine that you are using again, some machines can handle 250 cycles, and some machines will give only 150 cycles. The other outcome is that all reads are of the same length. So, you can imagine this process happening across these clusters—millions of them—and everywhere there is this ah-read-reading of one base at a time, and you are supplying this or doing this ah-cycle thing for each of these clusters in a similar manner. So, all reads that are generated in Illumina will be of the same length because of these reversible terminators. So, we understand how this actually impacts the AH sequencing process. So, in comparison to Roche 454 you will get reads of varying lengths right again because in some cycles you might have let us say 10 A's added and the next one you have 5 G's added, and this number is different from different fragments right. So, you get reads of different lengths, but in Illumina you get reads of the same length, ok. So, here is an overview of the whole process that you can see. This is a very nice animation of the whole process that you can see starting from the library process, like the cluster generation bridge amplification, then the sequencing, ok, and you will get a very nice idea about the whole process.

Overview of the whole process

https://sapac.illumina.com/science/technology/next-generation-sequencing/sequencing-technology.html

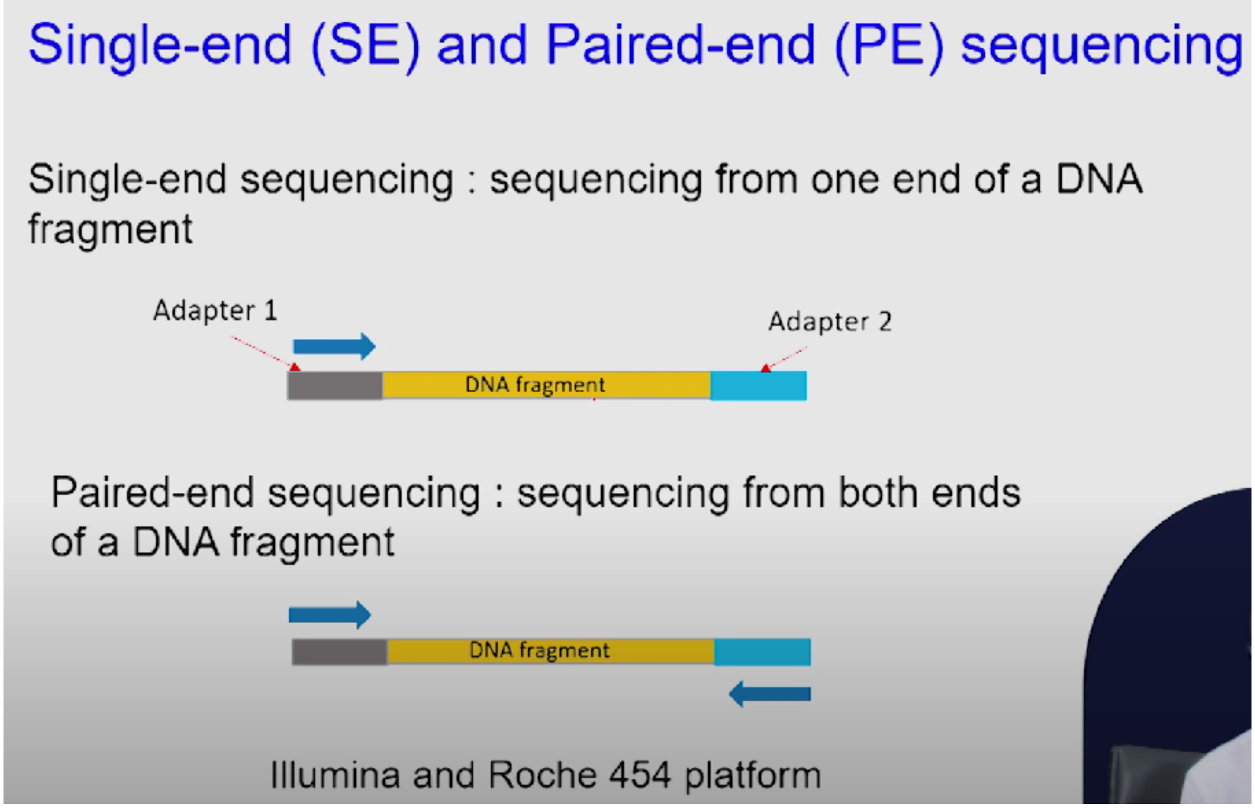
So, one thing I would also like to mention is that these sequencing methods are also changing very rapidly. So, here is one example of the sequencing chemistry in Illumina that is also evolving, starting with 4 channel chemistry, where you needed 4 fluorescent signals to detect these 4 bases. It has now come down to 1 channel chemistry. So, it needs only one signal to detect even four bases, right? So, this actually makes for even faster sequencing because, in 4 channel chemistry, you do not need to take 4 images at 4 different wavelengths to identify these 4 different fluorescences. So, for each cycle you have these image acquisitions, and that actually slows down the whole process. In 1 channel chemistry, you can take these images in just one channel. So, that is a much faster process for the sequencing.



The Sequencing Chemistry is also evolving ..

4-channel chemistry → 2-channel chemistry → 1-channel chemistry

So, as you can see, as the number of cycles increases, this actually improves in the time part. So, we can make these processes, the sequencing process, much faster. Now, one of the key ideas in Illumina is called paired-end sequencing. So, we will talk about that now. So, there is something called single-end sequencing and something called paired-end sequencing. So, in single-end sequencing, you have sequencing from one end of a DNA fragment, right? So, you can imagine this is a DNA fragment in yellow, and we have adapted these adapters, adapter 1 and adapter 2, and we can decide that we will just sequence from adapter 1, ok? So, we add primers in such a way that we only sequence in one direction. So, this is and will be your single-end sequencing, right? So, we are sequencing from just one end of the DNA fragment that we generated, but in Illumina as well as in 454 you can actually sequence from both ends. So, you can; it just requires these primers that are complementary to these adapters. So, this is, of course, not happening at once, right? So, what happens is that you should remember this synthesis process in Illumina, right? So, you have synthesis starting from one end and then this will continue towards the other end, or if you stop the number of cycles, that is, it will stop the sequence reaction, and then you can release the strand that has synthesized and you can start sequencing from the other strand, right? So, from the other side, okay.



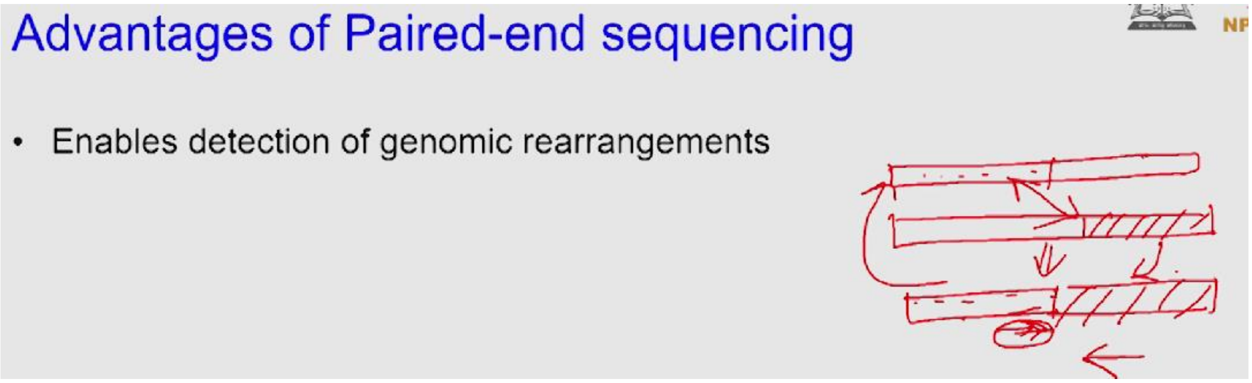
Single-end (SE) and Paired-end (PE) sequencing

Single-end sequencing : sequencing from one end of a DNA fragment

Paired-end sequencing : sequencing from both ends of a DNA fragment

Illumina and Roche 454 platform

So, this is how paired end sequencing will happen: you will sequence from both ends of the same DNA fragment, right one after the other. And this actually is very useful in many cases. As we go along, we will see why we actually prefer sometimes to do paired-end sequencing. So, one of the first advantages of paired-end sequencing is that it actually allows the detection of genomic rearrangements. So, what do you mean by genomic rearrangements? So, you can have something like this, right? So, let us say you have these two DNA fragments, or chromosomes, and maybe there is some sort of exchange between these two chromosomes. And this will give rise to the idea that we mark them in different ways. So, this will give rise to something like this, right? So, once you have this kind of chimeric chromosome, So, this will be something like this. Now, if you are sequencing and trying to identify the genomic changes that have happened in the sample, And if you employ, let us say, single-end sequencing, you sequence this part here, ok? If you are sequencing this part, you have no idea that this genomic rearrangement has happened with single-end sequencing. But if you are using paired-end sequencing and you have one part here and the other part here, then you are right. So, one part is coming from this fragment or this chromosome, and this part is coming from this fragment or this chromosome, right? So, this paired-end sequencing will help us in the detection of genomic rearrangements.



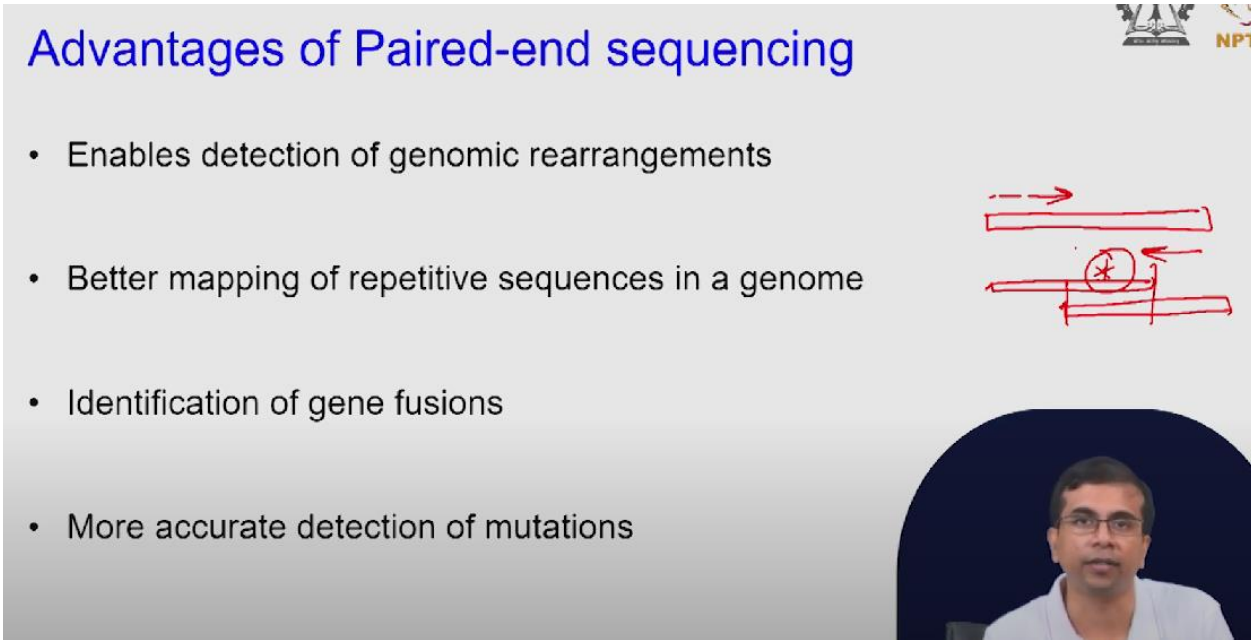
Advantages of Paired-end sequencing

- Enables detection of genomic rearrangements

We will see this in much more detail when we actually analyze these variants, etcetera, and how we actually detect these variations. We can also have better mapping of repetitive sequences in a genome. So, you can imagine right now that you have this genome sequence, and you often have this repeat sequence present. This is quite common in the human genome as well as in other genomes. So, you have these repeat sequences, which means they are identical in sequence. Now let us say you get some read data ok from, let us say matching this region, but you cannot be sure whether this read comes from this region or this region, right? They are identical in sequence, and you have this read, but you do not know whether this read was generated from this region of the genome or that region of the genome. So, this is something that is not possible if you are just doing single-end sequencing. Now imagine you are doing paired-end sequencing. So, you have one here, right, and the other part you are also sequencing, and that falls here ok? Now this pair right this is unique right. So, this region is unique, and there is no repeat sequence here. Since this region maps to this position, there is a match with this location of the genome. We can now say, "Okay, this read comes from this repeat only, right, not from this one," So, having this pair actually allows us to resolve this ah reads coming from repetitive sequences, ok? So, again, we will illustrate this in much more detail when you go into the analysis part.

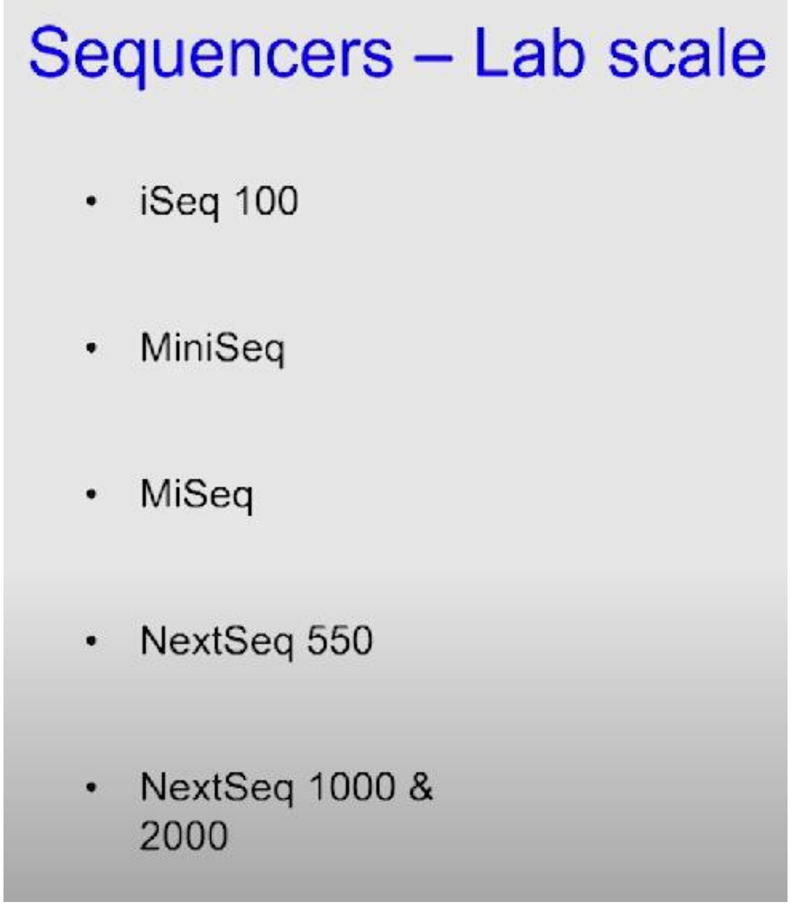
And then we can also identify gene fusions, as you have just seen for genomic rearrangements, which are very similar. So, we have gene fusions where two different genes fuse together in one place. Again, with sequencing only one part of this gene fusion, you have no idea that this gene fusion has happened. But if you have paired end sequencing, you will see one part coming from one gene and the other part coming from another gene, and this can give you an idea about this gene fusion. So, and then finally, we can also detect mutations more accurately. So, this is actually quite simple. So, imagine this fragment we are sequencing from both ends, and we generate this read. And if there is any overlap between these two parts, So, we are sequencing from both directions, and we do that sometimes also in Sanger sequencing, where we sequence a DNA fragment from both directions. And in this overlapping part, let us say we find there is a mutation here; there is a change, right from, let us say, one base to another.

But the other read or the pair here does not show that change. So, it will then tell us, OK, this is probably a sequencing error, right? So, although the sequencing error you have seen is actually low, this happens, right? So, here we can then identify that this is a sequencing error that has happened. So, in this way, we can actually double check the data that we are getting, and this helps us get a more accurate picture of the mutations.



Advantages of Paired-end sequencing

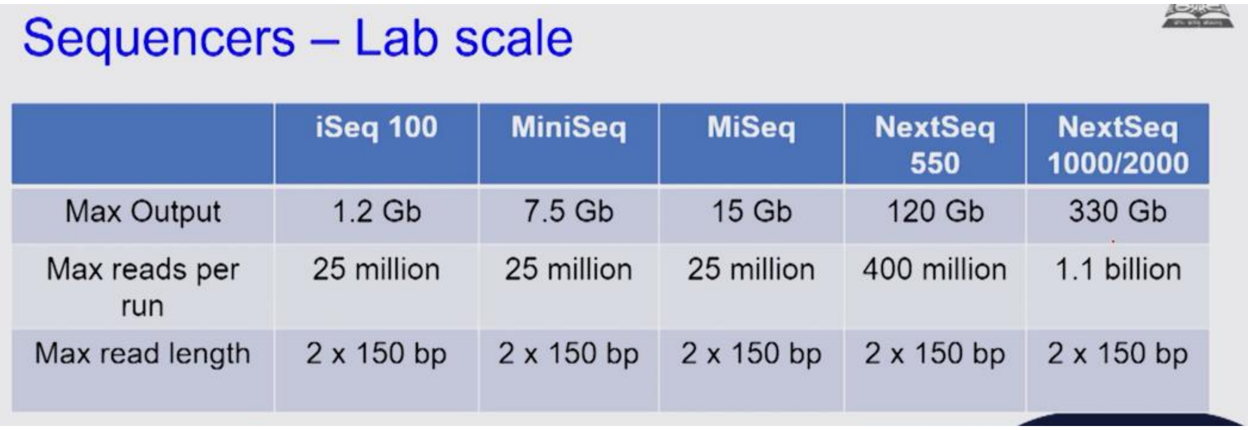
- Enables detection of genomic rearrangements
- Better mapping of repetitive sequences in a genome
- Identification of gene fusions
- More accurate detection of mutations



Sequencers – Lab scale

- iSeq 100
- MiniSeq
- MiSeq
- NextSeq 550
- NextSeq 1000 & 2000

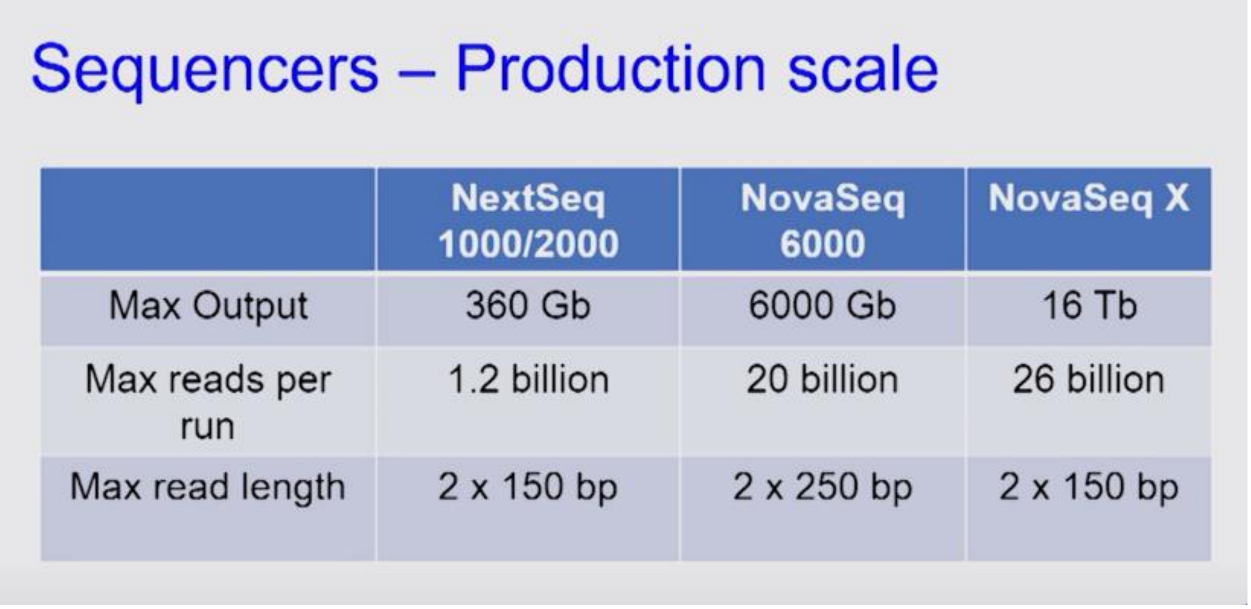
And this is something we do in Sanger sequencing as well. So, there are several sequences again, varying because of their throughput and the kind of capability that they have. So, these are some of the names you can have: MINISEC, MiSEC, etcetera; NextSeq sequencers; these are the principles of sequencing as they are. We follow the sequencing by synthesis method, but the only thing is that their throughput is different. So, here are some ideas about the throughput that you get from this machine. As you can see, they vary a lot, starting from 1.2 giga bases to 330 giga bases, going for NextSeq 1000 and 2000, depending on your requirements. The maximum read number is also given, and as you can see, you can get 25 million reads again. That is a huge number, right up to 1.1 billion on the NextSeq platform. So, these are really, really big numbers, which means you can actually work with really big genomes. The maximum read length that you can get from the sequencers is 1 2 into 150.



Sequencers – Lab scale

	iSeq 100	MiniSeq	MiSeq	NextSeq 550	NextSeq 1000/2000
Max Output	1.2 Gb	7.5 Gb	15 Gb	120 Gb	330 Gb
Max reads per run	25 million	25 million	25 million	400 million	1.1 billion
Max read length	2 x 150 bp	2 x 150 bp	2 x 150 bp	2 x 150 bp	2 x 150 bp

So, 2 means it has a paired end. So, you can get paired ends into 150 means at most 300 bases that you get from the sequencers, ok? Again, there are sequencer-specific limitations; there will be problems with signal detection, and there will be more sequencing errors if you go beyond this. So, these sequencers are most commonly used in small whole genome sequencing, targeted gene expression analysis where we want to look at the gene expression of only a specific number of genes, microRNA small RNA analysis, 16S metagenomic sequencing, and then also production scale sequencers. These are called Nexec 6000 and Novasek 6000, and these are the outputs that you get, and you can see that these are even more bases that you are getting out of the sequencers, going up to 16 terabases.



Sequencers – Production scale

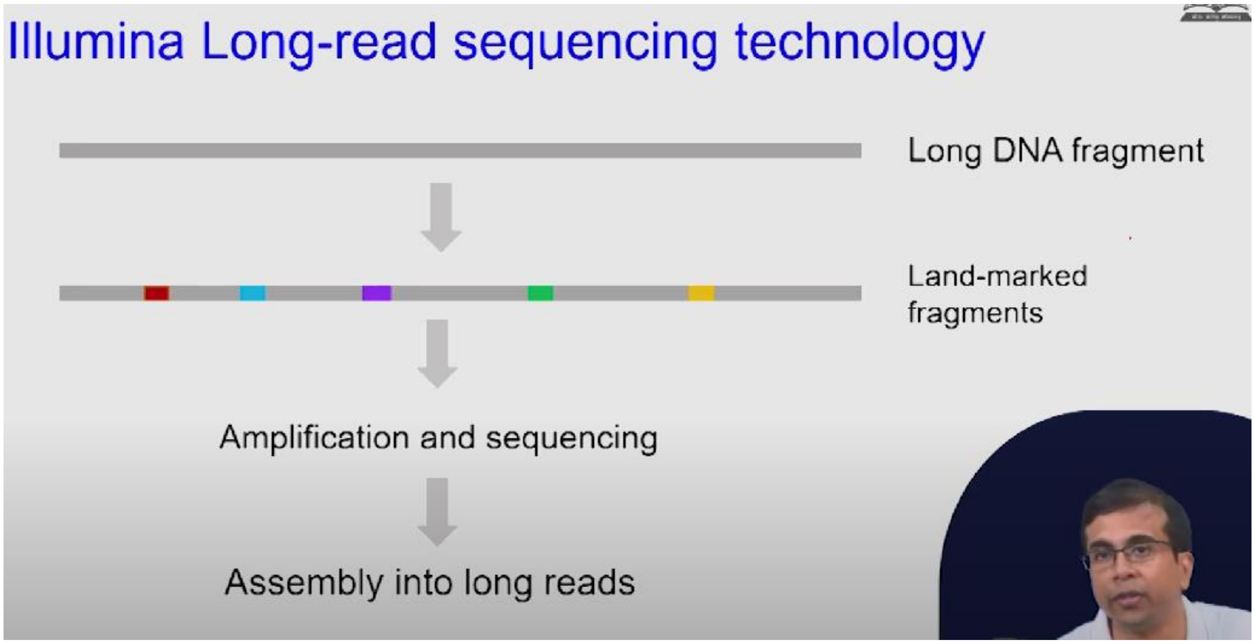
	NextSeq 1000/2000	NovaSeq 6000	NovaSeq X
Max Output	360 Gb	6000 Gb	16 Tb
Max reads per run	1.2 billion	20 billion	26 billion
Max read length	2 x 150 bp	2 x 250 bp	2 x 150 bp

And the maximum read number also increases proportionally as you get down to the billionth level, right? So, 20 billion, 26 billion, and the read length is again more or less the same, 2 into

150, but in the case of Novasek 6000, you can go up to 2 into 250, again paired end 250 base each. So, up to 500 base pairs and these production-scale sequencers have applications in large genome sequencing. So, we are working on big genomes, and you will use this sequencer for transcriptome analysis, where you want to look at the gene expression of all genes in the genome, not just targeted analysis. This can also be used for methylation, chromatin analysis, and single cell profiling. The advantages of the Illumina sequencing platform are cost-effective as compared to ROSE 454 and this has actually, ah, been the reason for success for Illumina. One of the reasons it is very accurate is, ah, even higher than 454 because this has addressed either the homopolymeric stretch as you have discussed.

So, it is about 99.9 percent accuracy that you get and really high throughput, as we have seen in the numbers, and this is suitable for large-scale sequencing. It is also very cost-effective, and that is what makes it the most widely used technology today because of all these points together. There are, of course, some drawbacks. So, this is still quite expensive compared to newer NGS techniques, which we will discuss later. This is because of the use of enzymes, etcetera. The capital cost is quite high because you have these optical detectors, and they have to be very sensitive. So, the cost of the instrument is actually quite high, high is the running cost because you are using enzymes and modified nucleotides for the sequencing process.

It is not cost-effective for a small-scale experience because, as you can see, the numbers are really large for the number of reads that you are getting. So, if you have only two samples and do not want too much data, then this is not the method that you will use for your application, and one of the major drawbacks is that we get only short reads, which are up to two into 250 base pairs. So, to address this, Illumina has developed something called long-read sequencing technology. So, ah, this addresses the short read ah, issue and gets reads of length 5 to 7 kB, and this is generated by something called ah, landmark ok.



Illumina Long-read sequencing technology

Long DNA fragment

Land-marked fragments

Amplification and sequencing

Assembly into long reads

So, these long DNA fragments are actually first landmarks and then sequences. So, this is the schematic of, ah, something like this, right? So, you have a long DNA fragment, and you have landmarked them, which are shown in different colors here in this region, using some sort of tagmentation or tagging right again using something like a transpose and so on. And then you can amplify these fragments. You can sequence these fragments separately, and then you can assemble them because of these tags that you have. You can identify, ok, which fragment will align with which fragment, and which fragment comes after which fragment. So, you can then assemble them into long reads.

So, these are the references that we have utilized for this class. So, to summarize, we have seen that Illumina sequencing utilizes the synthesis method and is based on reversible termination technology. So, this ah, actually allows for ah, very accurate sequencing, and here the base calling and detection are based on optical detection of the fluorescence signal that is out there, and again, this ah, chemistry is evolving with time. We get very, very highly accurate data, but the reads are actually short, and of course, to address that, Illumina has developed something called long read sequencing, and that ah, what comes with the limitation of shortage, but again, this requires multi-step preparation, which is quite complex. We have seen you need this landmarking process, etcetera. That is it for this class. Thank you. .