Next Generation Sequencing Technologies: Data Analysis and Applications

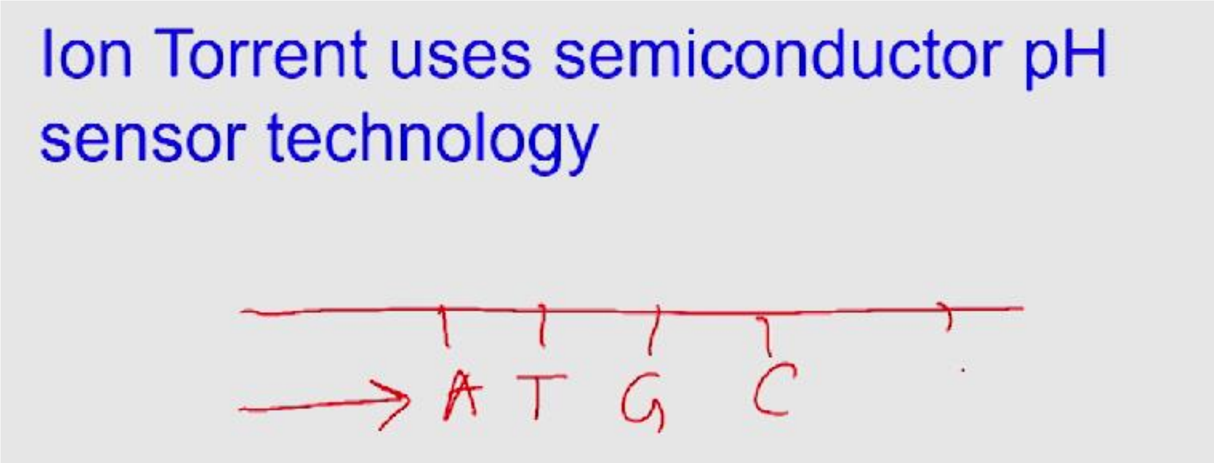
Ion Torrent and Nanopore Sequencing

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

Indian Institute of Technology, Kharagpur

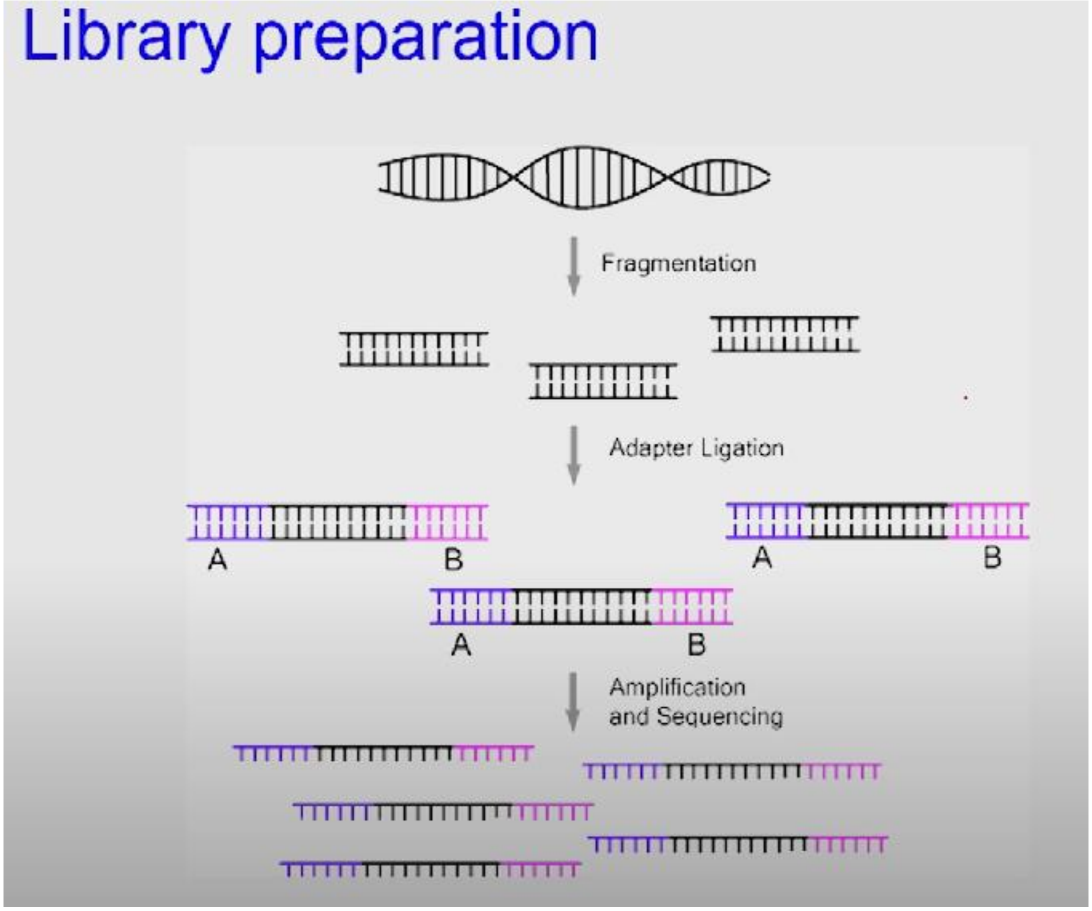
Good day, everyone. Welcome to the course on Next-Generation Sequencing generation sequencing technologies: Data Analysis and Applications. In the last few classes, we have introduced some of the recent NGS technologies, so we have talked about Roche 454, we have talked about Illumina, and then in the last class, we have talked about SMRT sequencing. In today's class, we will discuss two sequencing technologies in brief. So, one is called ion torrent sequencing, and the other is nanopore sequencing. So, that is the agenda for today's class. So, we will talk about the ion torrent sequencing platform, we will talk about the principles of how this sequencing works, and then we will talk about the most recent sequencing method, which is called Oxford Nanopore Technology. So, these are the keywords that we will come across. So, one is called a semiconductor pH sensor. We will see if this is relevant for the ion torrent.

We will talk about nanopores and pico-ampere current. So, let us start with the ion-torrent sequencing method. So, ion torrent uses semiconductor pH sensor technology, ok? So, what happens is that there is a semiconductor sensor that can detect any change in H+ ion concentration through a change in current. So, this method also relies on synthesizing the complementary strand of a DNA molecule. So, again, we have to imagine this is our genome fragment, and we have the primer binding site here and the bases are being added continuously, ok? So, this requires polymerase and dNTPs, but here we need just normal dNTPs; no modified dNTPs are required. So, what is the principle that is actually used by this ion torrent method? So, we will discuss this in a moment. So, here is the library preparation.



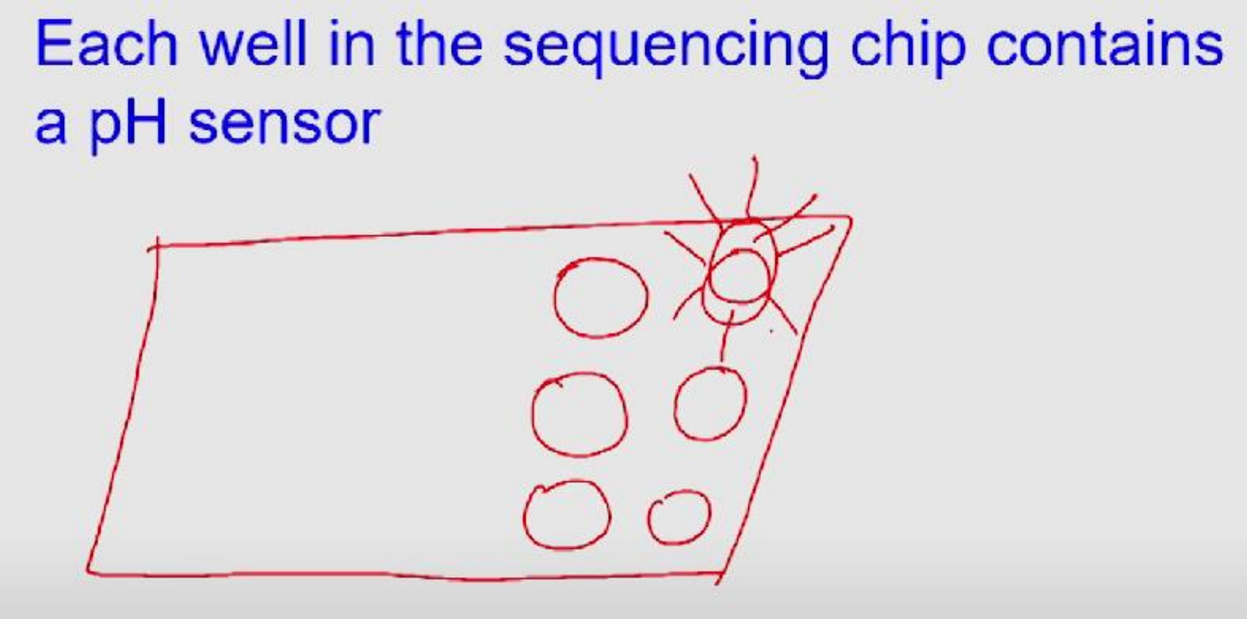
Ion Torrent uses semiconductor pH sensor technology

So, it actually is very similar to what we have seen in 454 and Illumina. So, we have the genome here and we fragment it again into small pieces and then we add the adapters A and B. This could be different, and again we amplify these fragments by some method, and we do the sequencing.



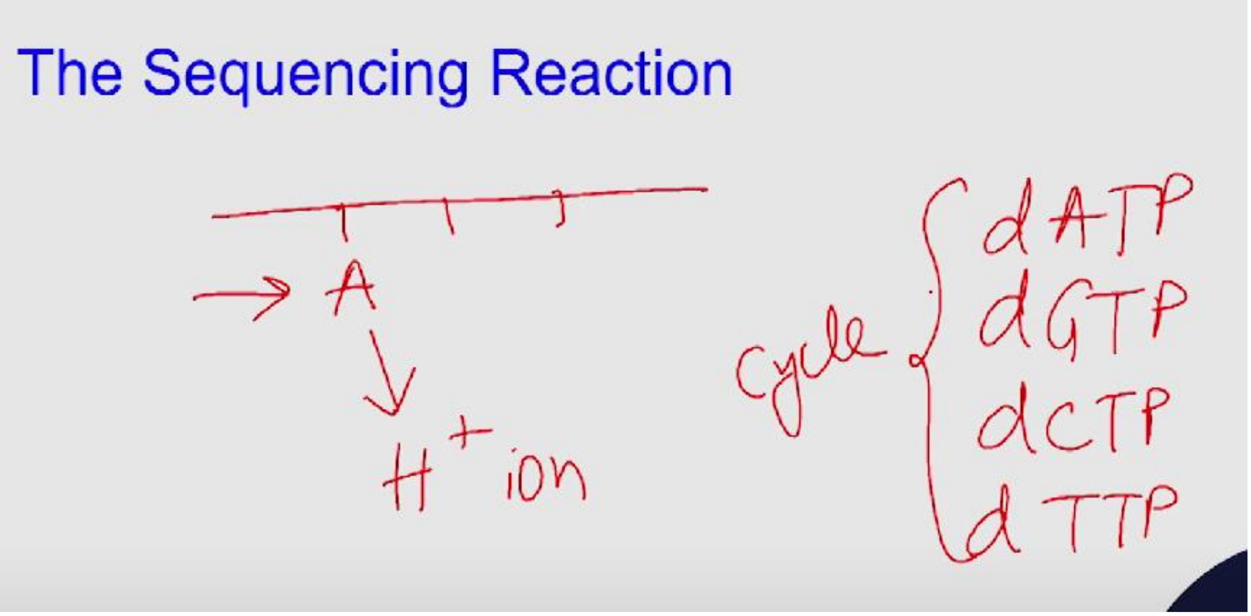
Library preparation

So,what happens in library preparation is that here we again follow a bead-based amplification that we have seen in 454 right. So, if you remember from the 454 class, we talked about these beads where you have some oligos or probes that are already attached and the molecule DNA molecules come there again, they kind of prime right, they kind of form this double strand, and the complementary strand is synthesized right. So,through emulsion PCR we actually generate multiple copies of the same DNA fragment on a bead. So, we follow the exactly same procedure here, right? We have again these beads, we have emulsion PCR, and we get multiple copies of the DNA molecule, and ideally, if you remember from our discussion in 454 class, each bead should contain the right copies of a single molecule, but sometimes this might not happen, and of course, those will give out confusing signals, which will then be filtered out later on during the data processing. So, we have the emulsion PCR step, and we again get this amplification of the DNA fragments.



Each well in the sequencing chip contains a pH sensor

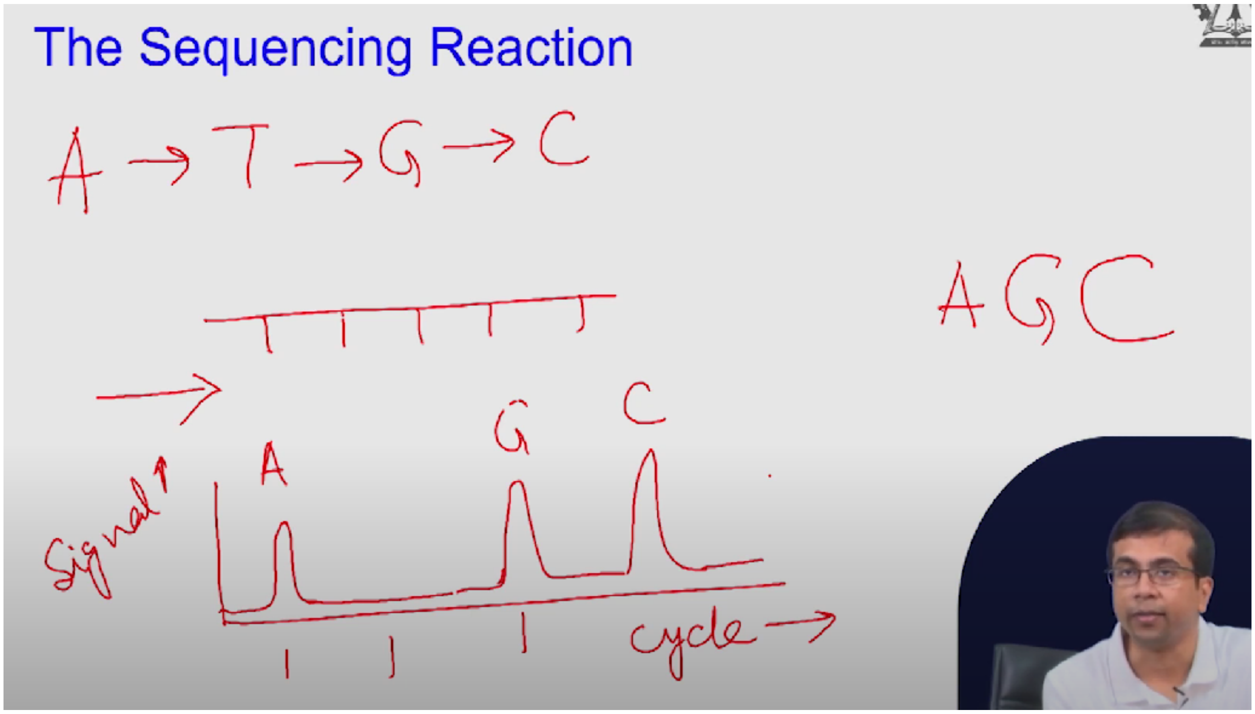
Now, what happens is that the sequencing happens in this kind of picotiter plate that we have seen in case 454. So, in these picotiter plates we have these again; we have these wells, and in these wells, we have the beads. So, these beads will come here; you will see them on these wells, and each well can accommodate only one bead. We remember that each bead will have these kinds of DNA molecules attached to it, and each well will have a pH sensor at the bottom of the well. So, as you can imagine, this method relies on the detection of H+ ions. So,the change in the H+ ion concentration. Now, how does this H+ ion change them? So, this is something that is very important, right? So, when you have this DNA again and here is a primer, you have the polymerase, and it is adding this base, ok? So, the moment one base is added, what you get is an H+ ion release, ok? So, every base addition will lead to the release of the H+ ion, ok? So, if you look at the polymerization reaction, you will see this H+ ion coming, and the detector then detects this H+ ion. So, the moment it detects an H+ ion, it will say, "Okay, there is this addition of some base ok, and this way we can do the base curve. Now, you might ask, right, how does it differentiate between different bases and different dNTPs, whether it is A, T, or C, and how does it detect that? If you remember, like 454, we supply dNTPs one by one. We do not put all dNTPs at once; we start with one dNTP, then give the other, then the next one, and the last one. So, we have four going-on cycles, ok? So, this process is again the same: we do not put all dNTPs at once; we start with one dNTP, then give the other, then the next one, and the last one. So, we have four going-on cycles, ok? So, this process is again the same as 454, right? So, we have dATP, dGTP, dCTP, and dGTP, and this forms a cycle, right? So, one after another, this gives us one cycle, and you can again repeat this process as many times as you want, and you get a certain length of the reads in this data.



The Sequencing Reaction

So, as you can imagine, we can actually get very similar numbers as 454. So, as I just mentioned, we provide dNTPs in cycles and in each cycle, only one type of dNTP is provided, so ok. So, again, let us look at the sequencing reaction. So, maybe we can have a better understanding, and let us say we are following this kind of cycle, right? So, we are providing A, then T, then G, and then C, ok, and we have this DNA strand, ok, and this is the primer, and synthesis is happening, ok, and what we are doing is this is the cycle that is going on, right, and here is the signal y axis, ok.

So, this signal comes from the H+ ion, right? Not light here; it is an H+ ion signal, ok. So, what we do, let us, so now, you can imagine what we are doing. We are adding first dATP, ok, and we see is there any signal? Is there any H+ ion detected by the pH sensor? Let us imagine there is ok and we say ok there is this addition of A then we let us say give E and so here is what we have added then we give T right and we see ok there is no signal ok which meansT has not been added right. So, the first phase is A, then we add G, and we see that there is a signal. So, this means the sequence is A G, now ok, then we add C right, and we see ok, there is a signal.



The Sequencing Reaction

So, then we see and repeat this process, right? So, now, you will have A OK again, and you will see if there is a signal, then D, G, C and so on. We can continue as many times as you want, of course, but there is a limit because of the polymer that we are using. Now, one of the things you probably realize that, again, like 454, you have a washing step, right? So, you give one polymer, one dNTP, right? You see, wait for the signal, then wash away everything and again give the next dNTP. So, it requires a washing step like 454 because we are giving one dNTP at a time.

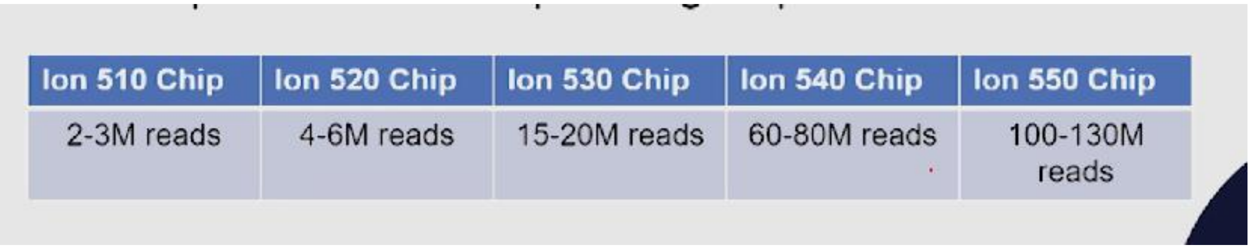
So, you need this washing step, ok? Now, you can look at this video, which will give you an overview of the whole process. This is an animation, and you will get a better idea of the whole process. So, what are the advantages of this method? So, this method gives us a very similar read length of 454. So, it is about 400 to 700 base pairs. It has reasonably high accuracy.

Overview of the whole process

https://www.youtube.com/watch?v=zBPKj0mMcDg

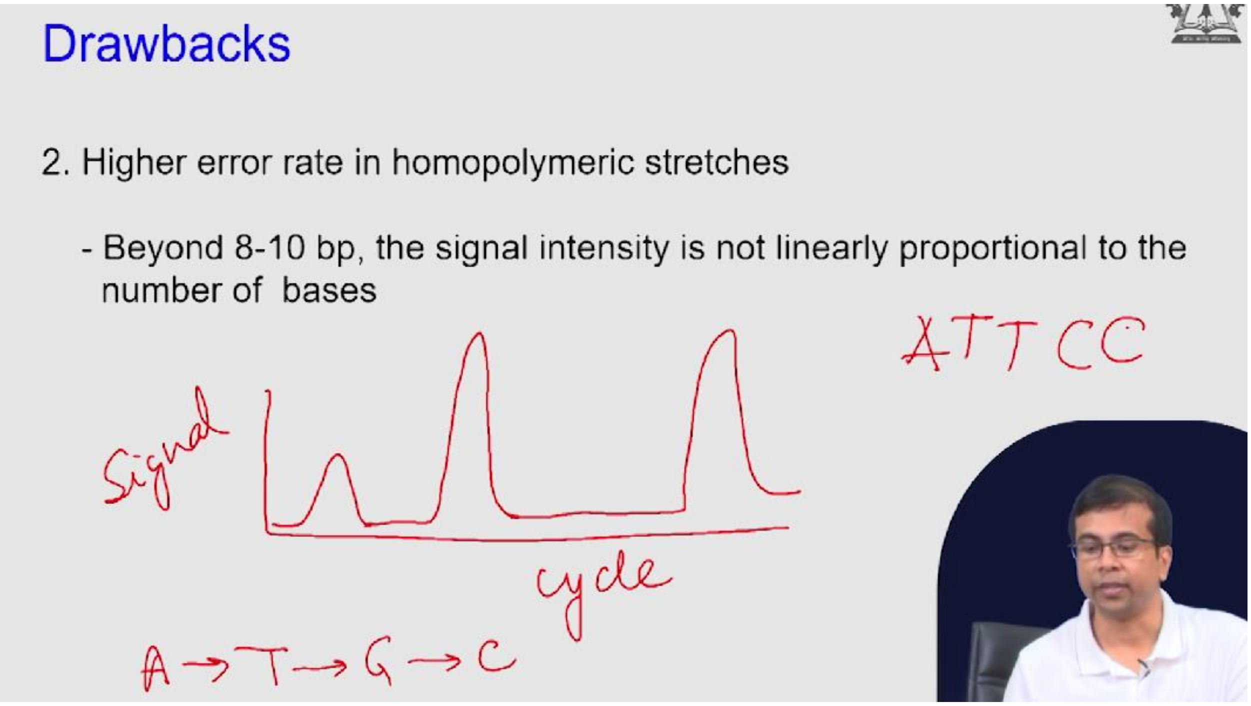
So, as you probably remember, 454 was also reasonably accurate, and it is about 99 percent accurate that you can get. So, which is good, it is also quite economic and compared to the competition right because you do not need any optical detection units, which could be a bit expensive and require kind of a higher capital cost for that machine, whereas here you need less capital cost. So, the machine is not so expensive because you do not have any optical detection units; you have a semiconductor-based pH sensor, which is more economical. And also, we do not need any modified dNTP, right? So, this also reduces the operating cost of the system.

What are the drawbacks, right? So, one of the first drawbacks is that you have lower throughput compared to the Illumina platform. So, one of the things you will notice is that the maximum we can get is 100 to 130 million reads, and this is much less compared to the Illumina data that we have seen right now. We have seen some statistics for Illumina, where you can get billions of reads, and here it is only about 100 to 130 million at most. And it also varies from the chip sequencing chip that we have used to the chip that you will use here, right? So, again, it depends on whether you can see some of the chip names and the throughput that will be created from these chips.



Ion 510 Chip	Ion 520 Chip	Ion 530 Chip	Ion 540 Chip	Ion 550 Chip
2-3M reads	4-6M reads	15-20M reads	60-80M reads	100-130M reads

Again, depending on your application, you will pick the right one, which will be most suitable for you. So, one of the things you can probably notice is that this is probably not really suitable for large genome sequencing projects. So, if you want to do, let's say, a huge genome sequencing project for large genomes, you probably would not choose this method. But on the other hand you can think that if you are working with a small number of samples and small genomes, this is probably the ideal and most cost-effective method that you can use. There is another drawback right now, and this is actually shared with 454. So, one of the things you probably now remember is that in 454 we have this signal that is proportional to the number of bases being added. So, if I can again draw this signal versus cycle and we are following this ATGC cycle right and imagine, let us say we have first A, we are detecting OK. So, we have detected A; now we see T, and what we see is that the signal is much higher, ok? So, which means we have probably added two T's here, ok? So, the signal that the sensor detects should be proportional to the number of bases being added.

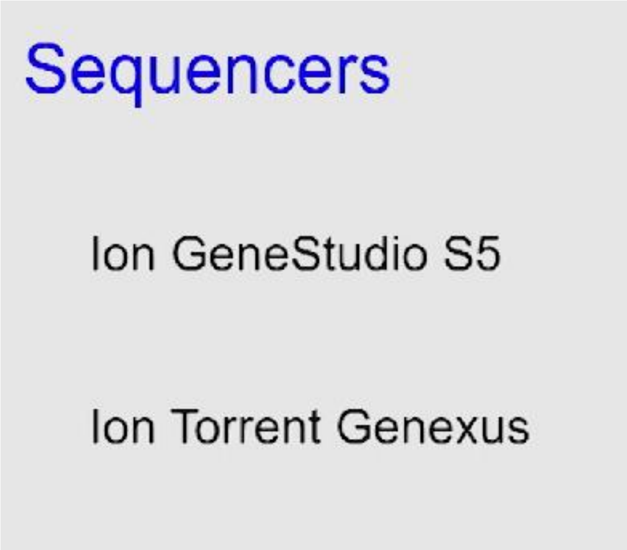


Drawbacks

2. Higher error rate in homopolymeric stretches

- Beyond 8-10 bp, the signal intensity is not linearly proportional to the number of bases

So, you can imagine that if you have two bases, you have two H+ ions that are being released. If you have three bases added, you have three H-plus ions. So, the signal that will be detected by the sensor should be proportional to the number of bases being added. So, then we have G, maybe we miss G, and then again for C, maybe we see a higher signal, and we can say we have two T's and two C's added here, ok. Now, the problem is that beyond 8 to 10 base pairs, this signal intensity is not proportional to the number of bases. As we have seen in the case of 454, this tends to saturate OK, which means that beyond these 8–10 base pairs, the homopolymeric stretch regions will not be able to accurately determine the length of the stretch right and will also have more indels, insertions, and deletions that will be detected OK. So, this is a problem where we are not doing the synthesis of this complementary strand one base at a time. So, we see that this is not there in Illumina, this is also not there in PACbio SMRT sequencing, but here we have this problem again because the principle is very similar right like 454, but I mean the signal detection is different where we in 454 detect light here we are detecting the change in H plus ion right, but we are using unmodified dNTPs right and we are allowing addition of multiple bases at time ok. So, these are some of the sequence and names that that you are that are available again; they vary based on their throughput, but the principle is the same. So, you have ion GeneStudio S5 or ion-tolerant gene access, ok?

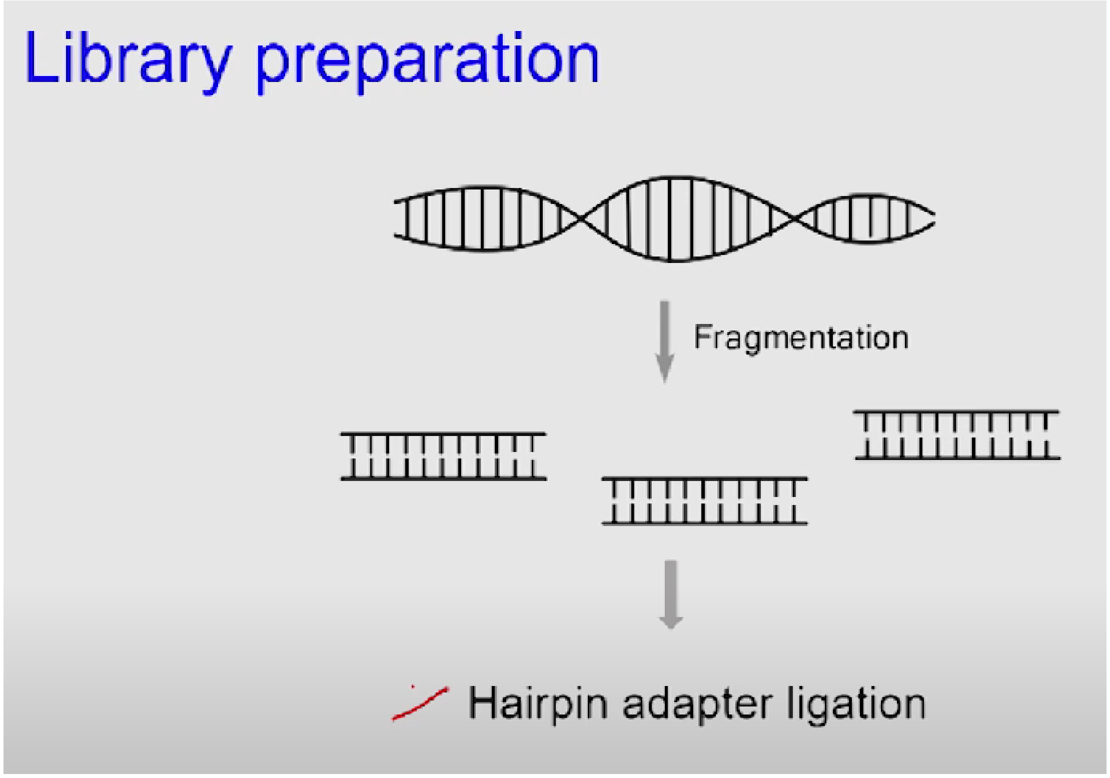


Sequencers

Ion GeneStudio S5

Ion Torrent Genexus

So, let us now move on to the final sequencing technology that we will discuss. It is called Oxford Nanopore Technology, or, in short, ONT. You will see this term in many places. So, this is a method that actually relies on direct DNA sequencing. All the methods that we have discussed so far, starting with 454 Illumina, then SMRT, and then ion torrent, all require the synthesis of a complementary strand of the DNA molecule. So, this method does not do that. Okay, this method directly sequences DNA molecules, and this is what makes it incredibly powerful. So, it means it simplifies everything, right the library preparation, etcetera. You do not have to do any amplification, you do not have to generate any clusters, etcetera. You do not have to have any polymerase in there to synthesize the complementary strand, ok? And this means that this method is likely to be more economical than all the other methods that you have discussed. It also means this method can sequence really long reads; it can generate really long reads because you are not limited by the processivity of the polymerase that you are using for the synthesis process. So, let us now move on to this method and see how it actually achieves all this right. How does it do the direct DNA sequencing? So, the library preparation, as I said, is very simple, right? So, we have this genome that is fragmented into smaller regions again. Although this method can sequence really long reads it cannot sequence the full genome at once. Of course, we would like to do that, but that is still not possible.

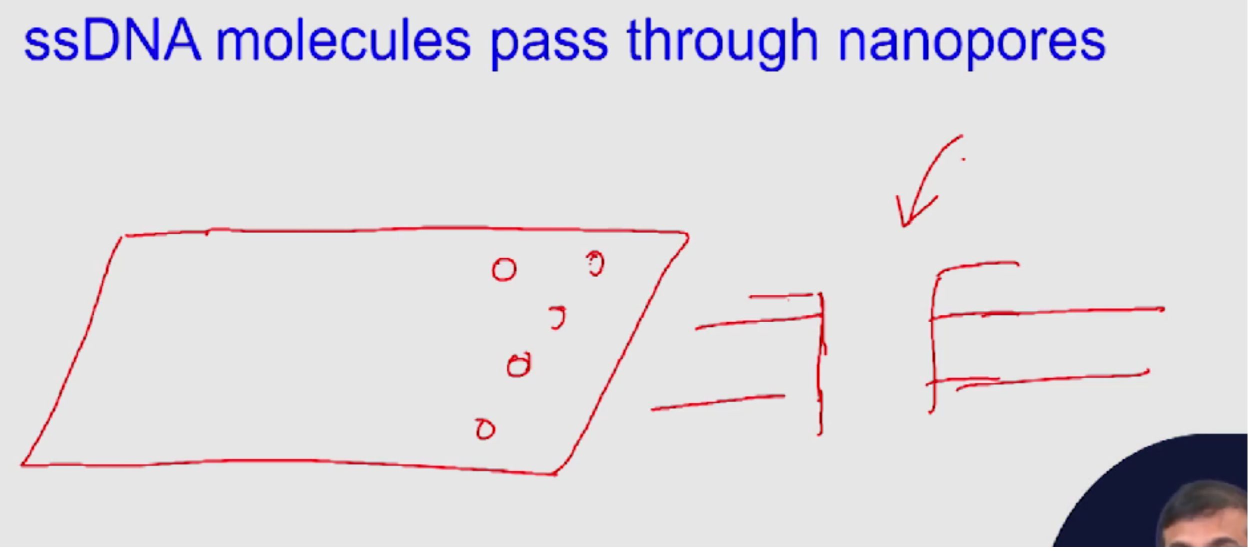


Library preparation

Fragmentation

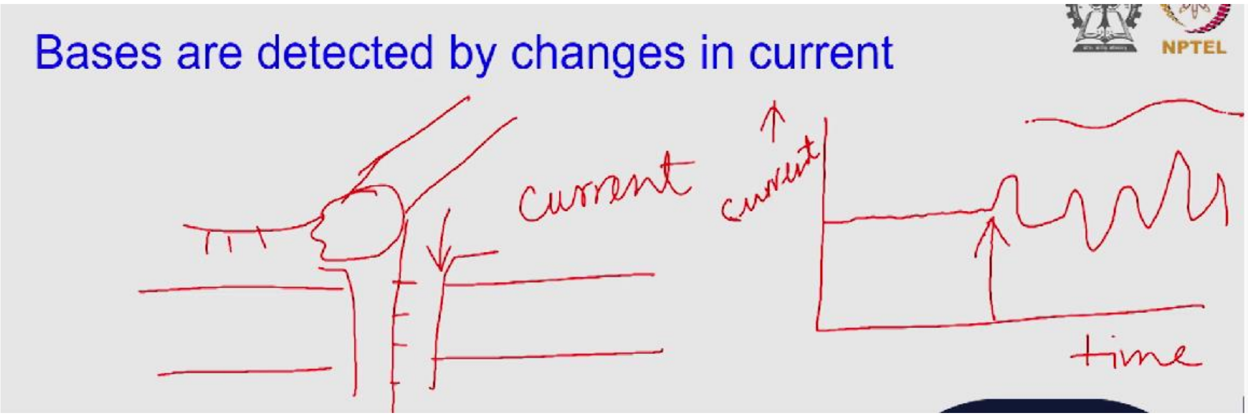
Hairpin adapter ligation

So, you need to fragment it into smaller fragments, right? You need to get smaller regions, but they can be much larger than all the methods that we have discussed so far. So, in Illumina, Antorin, or 454, where you need smaller fragments, you can have quite large fragments which can then be sequenced. And then we have this hairpin adapter ligation. So, it is a very simple step; there is no amplification required; you simply add this adapter, and then you can start the sequencing. So, what happens here? So, is there something like this? So, you have a nanopore-like structure. So, you can imagine a flow cell, right? So, the sequencing happens inside the flow cell. You need something like this, either a semi-cell flow cell we have seen or picotiter plates. So, here you have a flow cell where the sequencing happens, and on this flow cell we have millions of these very small pores, which are called nanopores. So, they are actually a small structure that is embedded in a kind of matrix, and these nanopores are made of proteins. So, these are very small dimensions, ok? So, only one strand of DNA molecule will pass through this. So, single-stranded DNA molecules pass through the nanopores okay, and when they are passing through these nanopores, they are being sequenced okay.



ssDNA molecules pass through nanopores

So, we will see in a moment how this happens. So, as I have just mentioned, you have the matrix, and you have these nanopores. So, these are the nanopore structures where you have these very small holes or pores right and you have a protein right that will separate out this double-stranded molecule; it will just direct one strand through this right and the other strand will be going outside ok? So, only one strand will be passing through this, right? This is a protein that will do this job, ok? Now, how do you actually detect these bases? As you said, this is direct DNA sequencing. So, how do you actually detect this? So, it turns out there is a current that is applied across this nanopore. So, it is a pico-ampere current, right? So, a very small amount of current is applied through this nanopore.



Bases are detected by changes in current

Now, when these DNA bases are passing through the nanopore, they actually lead to some change in the current OK, and this change is again dependent on the sequence of DNA that is passing through the nanopore. So, you can imagine this situation, right? So, if you are drawing this time versus current, So, the y axis is current, right? This is in pico ampere.

So, very small current. So, let us say we have this kind of constant level of current that is passing through the nanopore, and then at this point we start passing the DNA molecule through the nanopore. So, what will happen is that because of these specific sequences, they will modulate or change the current ok, and we will start seeing this kind of peaks and valleys some sort of change ok. Again, this change is dependent on the sequence. So, by looking at the change in current here, we can identify what kind of change we are seeing. We can then identify the bases that are passing through the nanopore, and there is a specific characteristic of this specific set of sequences that will induce certain changes in the current. Now, as you can probably imagine, this is quite a difficult process. So, identify which bases are passing through the nanopore from the current signal. So, one of the questions you might have, ok, this is detected by single base, but the answer is actually no. This is actually detected by a pattern or combination of bases that are passing through right. So, a combination of bases will have different patterns, and you can find these unique patterns from these combinations, ok? So, I hope this is clear as to the general principle of how nanopores will detect DNA bases. Here are the two links where you can actually see some animation that will make the whole process clear to you, and we will get a better understanding. So, please have a look.

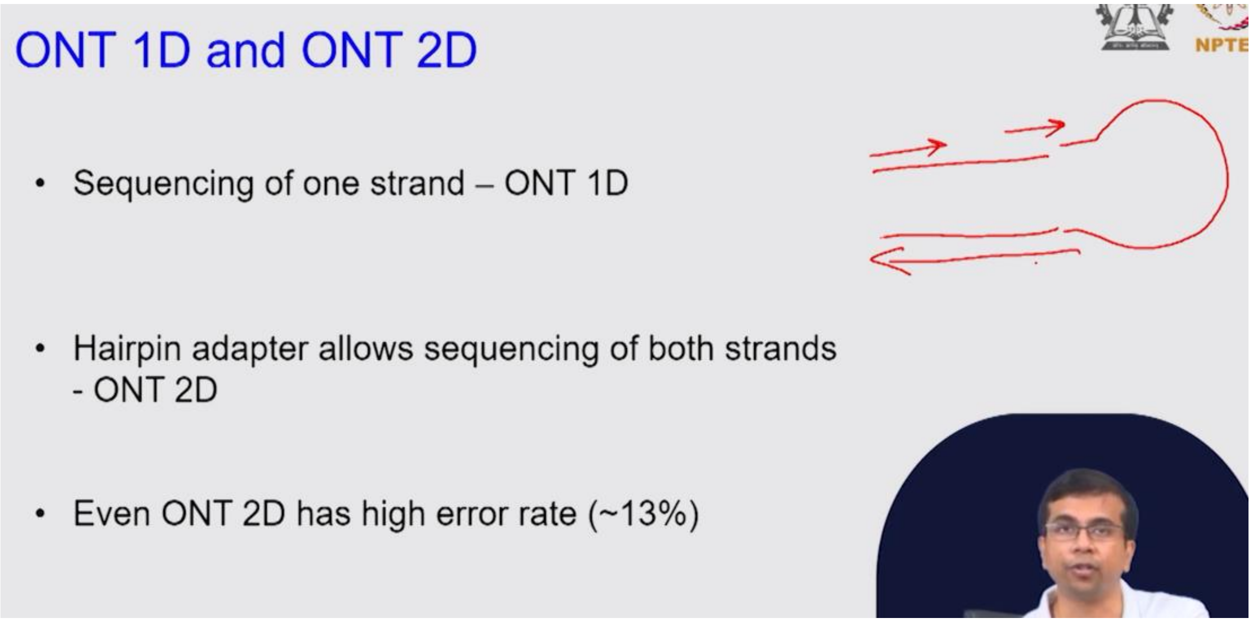
Overview of the whole process

https://nanoporetech.com/support/how-it-works

https://nanoporetech.com/applications/dna-nanopore-sequencing

So, what are the advantages? So, one advantage is that this is a really long, rich sequencing because you are not limited by the processivity of the polymerase; you are detecting that DNA sequence directly based on the change in current. And you can imagine you can continue this process as long as you want, right? There is no polymerase required that actually is doing the synthesis, and it will stop at a certain point. So, you can keep on passing DNA molecules through the nanopore, right? So, the ideal is that is what we want, and that is where we are going, right, where you can probably sequence the full genome in one go. So, we get long, rich sequencing. So, this helps in genome assembly, and really long reads will also help in identifying the long structural variants, the big ones, that are present in the genome. So, we can achieve a read length of 10 to 100 kB. You can imagine, right? You can see that compared to the other meters this is actually quite long. In addition, as you have seen with SMRT, we have real-time observation of the current, the change in current that is happening right now, that is the signal, and we have real-time observation of what is happening to the current with respect to time, and this data is stored in the data file that comes from Nanopore. So, when you talk about these data formats you will see we have this data in there, and this data is stored, and that can be quite useful for specific applications. In addition, this is also a single-molecule sequencing right. So, we are getting the signal from a single molecule, and like SMRT, we are sequencing one molecule, and each nanopore will get a unique DNA molecule. So, this is something again really advantageous because we can now identify what the changes in the DNA molecule are just beyond the basis. What if there are some other modifications or some specific characteristics that can be identified because we are now observing the single molecules in real time? So, you can see these two classes of methods now, right? So, we have like 454; we have Illumina; we have Ion-Torren; we need an amplification step; and we get the average signal out right. So, the detectors are detecting the average signal from multiple clones of a DNA molecule, right? So, that is why the amplification step is there. In the case of SMRT and in the case of nanopore, what you see now is that there is no amplification step; we have detection of signals from single molecules, and we have real-time signals. So, this is really valuable for different applications. And one of the biggest advantages of nanopore sequencing is that it is really economical. As you can see, you do not require any optical detection devices, you do not need any polymerase, and you do not need any DNTP, which means the cost comes down a lot. So, compared to other methods where you need this synthesis process, even if you compare now with SMRT, which is a single-molecule real-time sequencing that requires the synthesis of the complementary strand compared to all the methods, this is the most economic and cost-effective compared to all sequencing technologies. So, this is why this method will get popular and will probably be the sequencing of the future. Now, one of the questions you might ask is: So, when we discuss that Illumina is probably the dominant platform today, most sequencing happens in Illumina.

So, the question is, why do we not shift to nanopore sequencing? Why are we not doing sequencing in nanopore because it is really great? We have a long reach, we have low cost, etcetera. Everything is in favor of this method. So, why aren't we switching? So, here comes a drawback, right? So, the error rate is actually very high, okay? So, if you compare it to Illumina or other methods like 454 or iron taurine, this error rate is quite high. So, it is about a 20 percent error rate, okay?



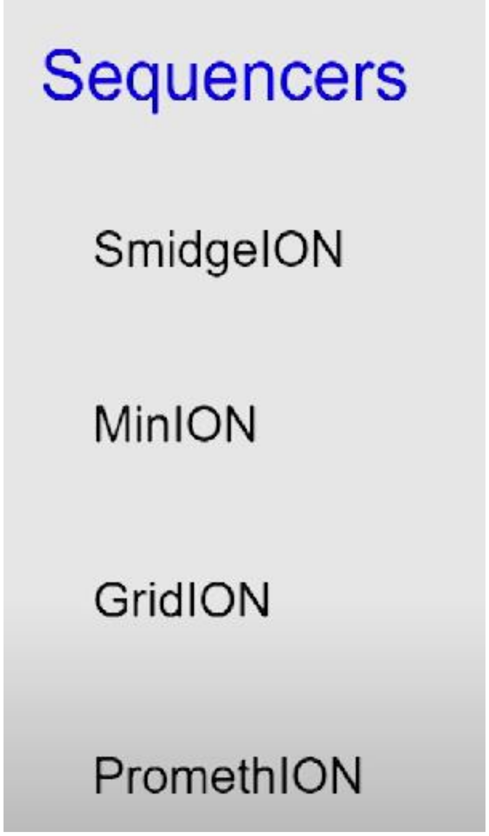
ONT 1D and ONT 2D

- Sequencing of one strand – ONT 1D
- Hairpin adapter allows sequencing of both strands - ONT 2D
- Even ONT 2D has high error rate (~13%)

Now, the good thing is that you can actually improve upon this error rate by using a method called ONT2D. So, we have ONT, which stands for Oxford Nanopore Technology in short, right? So, ONT1D is right where we actually have the DNA molecule, as you can imagine. So, as we have mentioned again, we have this hairpin adapter. So, we can sequence only one side of this because these are complementary strands. So, double-stranded DNA, right? So, we can sequence pass only one side of the DNA molecule, right? This protein that is present at the entry of the nanopore can pass only one side, and we can detect this sequence. So, that will be ONT1D. Now, what you can do is, because we have this hairpin adapter we can also sequence the other side. So, as you can see, you can kind of move these and move back and forth, right? We can go through this, then we can again sequence the other side ok? Now, again, this will help in improving the accuracy because we are getting sequence from the same DNA fragment; these are complementary strands, but the same part of the same fragment.

So, we will get more sequence data, and this again helps us in calling the consensus sequence right. So, this hairpin adapter allows us to sequence both strands, and this is called the ONT2D, right? But what we have seen and what researchers have seen is that even the ONT2D has a high error rate of about 13 percent. So, what can we do about this? So, because this 13 percent error rate is very high, if we are interested in identifying mutations, genotyping, etcetera, So, what has been happening in recent years is that there have been improvements in accuracy, and one of these comes from the increased number of detectors being added to the nanopore.

So, earlier, in the nanopore method, there was only one detector per nanopore. So, we can imagine this situation: you have the flow cell, you have these nanopores, and each nanopore has only one detector. Now, there are two detectors that are slightly separated from each other right now and are detecting this change in current. So, they are being affected by a larger combination of bases, right? So, as you increase the uniqueness of the bases, you can actually get a better correlation with the signal change or the current change that is happening. So, a larger number of detectors actually leads to higher accuracy. So, these are some of the sequences that are available. So, again, these are based on different throughput levels, but the principle of sequencing is the same.



Sequencers

SmidgeION

MinION

GridION

PromethION

So, the first one is SmidgeION; it is a very small sequencer. You can actually connect it to your smartphone and do the sequencing and get the data on your smartphone, and then you have this sorted according to the throughput. So, you have a minimum ion, a grid ion, and brometium. So, what about the potential of the nanopore technique? So, one thing you can see right here is direct

DNA sequencing, and perhaps this can have many more applications just beyond DNA sequencing. So, one of the first things we see is that we can actually do direct RNA sequencing. So, we will talk about this in the transcript of the analysis, right? So, when you do the transcript sequencing, we want to measure the expression levels of genes. We first have to convert the RNA that we isolate into cDNA before we can apply any of this earlier sequencing technology. But with nanopores, you can actually directly sequence RNA right here; there is no amplification, etcetera, no primary binding, etcetera required. So, you can simply set up the process for sequencing RNA directly.

So, one of the things that will change is that this RNA bases the signal; the current signal change will be different, and that will have to be taken care of again when we are detecting the bases. When you are doing the base calls through the change in current, we need to have a program that will identify those RNA bases, ok? But that can be a solvable problem, and we can do that. So, we have seen recent improvements in accuracy, and it has actually reached about 99 percent, and hopefully it will go even further now. It is real-time sequencing, and we have seen that this is direct detection of nuclei modifications we can do. Again, we will talk about this in much more detail because of the real-time nature of the signal, and we are also detecting signals from single molecules.

And we have, I have told you right, we can do RNA sequencing, but it also turns out that we can also sequence proteins. Now there have been some reports and some work from some research groups that show that we can actually sequence proteins directly using nanopore technology. Now, one of the things you can now visualize right in your mind because the principle is so simple that there is no synthesis required is that the principle is based on a change in current due to the chemical nature of the molecules or atoms that are passing through it. So, you should ideally be able to do this protein sequencing as well because of the principle right? So, proteins again, different amino acids will generate different patterns, or the combination of amino acids will generate different patterns for the current that is flowing through the nanopore. So, recently, this has been applied and has been shown to identify even a single change in amino acid in a stretch of protein. So, here is the reference where you can see the application of nanopore sequencing for protein sequencing and can also identify this difference in a single amino acid. So, please go ahead and have a look at the paper, and we can now help you understand the potential of this Oxford nanopore technology. So, here are the references that we have used for this class, and to summarize, we have talked about two sequencing methods in this class. So, the first one was ion-torrent sequencing. The ion torrent utilizes a semiconductor pH sensor, and it detects changes in H+ ion concentration when the sequencing reaction is going on. So, when you synthesize the complementary strand of a DNA molecule, there is a release of H+ ions every time a base is added, and this change in H+ ion concentration is detected by the sensor, which then calls the base OK. And this happens in the cycle that we have discussed: we do one dNTP, we add one dNTP at a time, and we observe for signals. And we have seen that we get reads that are comparable in size to Sanger sequencing, but we have again the problem that we have seen earlier: a higher error rate in the homopolyclinic stretch region. So, we get more indels in those regions. We then talked about the nanopore sequencing technology, and this is a technology that is entirely different from the other technologies that we have discussed.

So, this method directly sequences DNA molecules, and the detection happens through the change in current. So, again, a stretch of or a combination of bases will change the current in a unique way compared to other combinations, right? So, this detection actually helps in direct identification. And this is real-time detection right at the single-molecule level, and it is very cost-effective because we do not have any polymerase or dNTP involved. But what we have seen is that we have a higher error rate, and this is being addressed now by the addition of more detectors and by doing something like 22D right or consistent sequencing, which will actually help in better accuracy. And we have also discussed right here that this is a method that has a lot of potential for the future because it is cost-effective. It is doing that sequencing by direct detection. So, it can be applied for RNA sequencing as well as protein sequencing, and I have given you some references for further reading. Thank you.