Next Generation Sequencing Technologies: Data Analysis and Applications

Single Molecule Real Time (SMRT) Sequencing

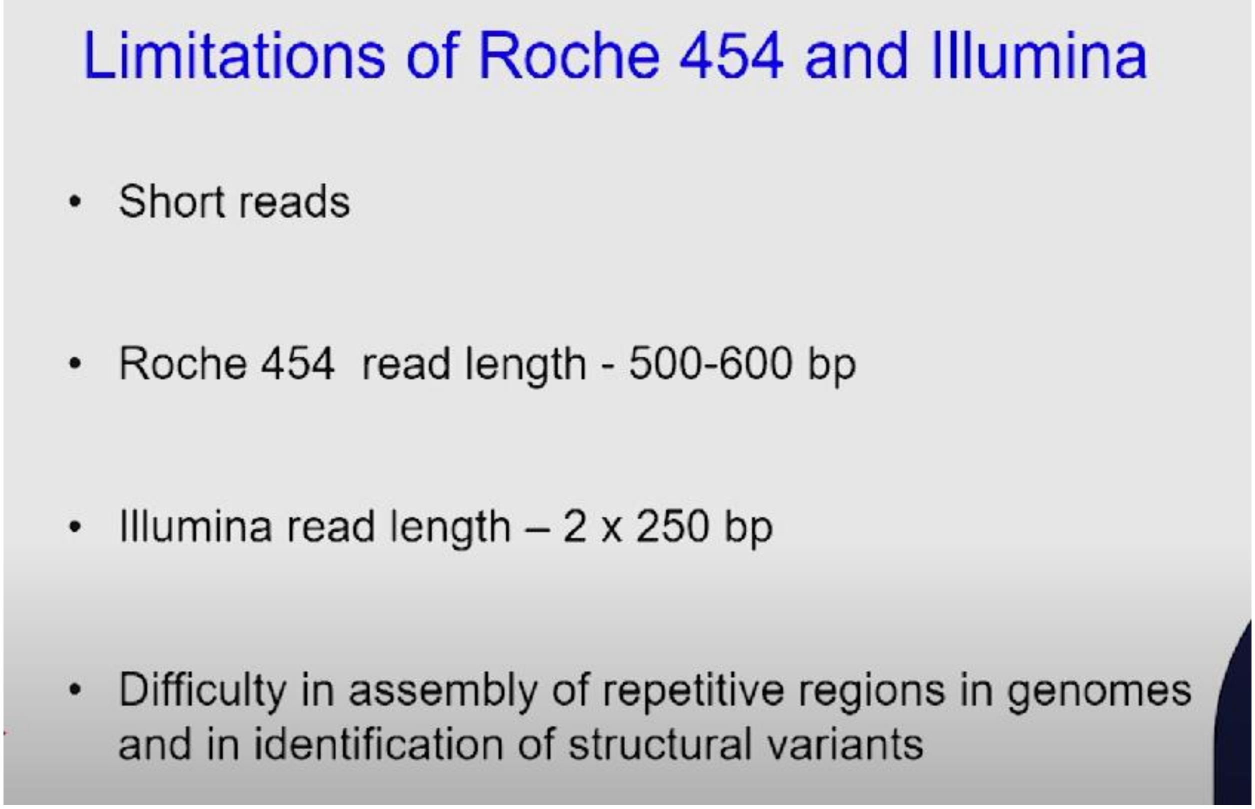
Dr. Riddhiman Dhar, **Department of Biotechnology**
Indian Institute of Technology, Kharagpur

Good day, everyone. Welcome to the course on Next Generation Sequencing Technologies, Data Analysis, and Applications. In the last few classes, we have been discussing different next-generation sequencing technologies, such as 454 sequencing and Illumina. So, today we will be discussing one more technology, which is called single-molecule real-time sequencing, or, in short, SMRT sequencing. And we will see how this method is different compared to the last two that we have discussed so far. So, let's begin.

So, this is the topic for today's class. We will discuss the single-molecule real-time sequencing method developed by Pacific Biosciences. These are the keywords that we will come across. So, we will have zero-mode waveguides, or ZMWs; we will see real-time sequencing quite often; and finally, we will also see something called Hi-Fi reads in the presentation.

So, just to briefly recap what we have discussed so far, we have talked about 454 sequencing, and 454 depends on pyrosequencing, right? So, if you remember the principle of sequencing, we have this DNA polymerase, and when a base is added to the growing strand, there is the release of pyrophosphate. Now, this pyrophosphate is finally converted to ATP, which, in the presence of the luciferase enzyme and luciferin, will give rise to a signal of light. So, this light is detected by the detectors that are present, and then the base call is OK. And if you remember, we pass one base at a time ok? So, one type of dNTP we give and we see if there is any signal or not, and this process happens at a very large scale across the picotiter plate, ok, and that is how we get this high-throughput sequencing. Now, one of the first things you probably remember is that we need to amplify the DNA, right? When you prepare the library, we have to do some amplification in 454; this is the emulsion PCR, which will generate multiple copies of the same DNA molecule on the bead. And the average signal for all the synthesis that is happening on a single bead is detected as a pulse of light. So, the assumption is right that all the synthesis across these molecules in a bead on a bead is happening simultaneously, and the signal comes together right simultaneously, and that will be detected.

We also talked about Illumina sequencing by the synthesis method. This method also relies on synthesizing the complementary strand, and here we have the reversible termination method. So, there is one-by-one right addition of bases and detection of signals. Again, there is an amplification step required, right? So, what do I mean by amplification? You need to again clonally propagate right through bridge amplification, and we generate clusters again. These are copies of the same molecule ok? So, we have seen that there are similarities between these two methods. We have an amplification step for the library before we do the sequencing, and both rely on the sequencing by synthesis method. Now, what are the limitations of 454 and Illumina?



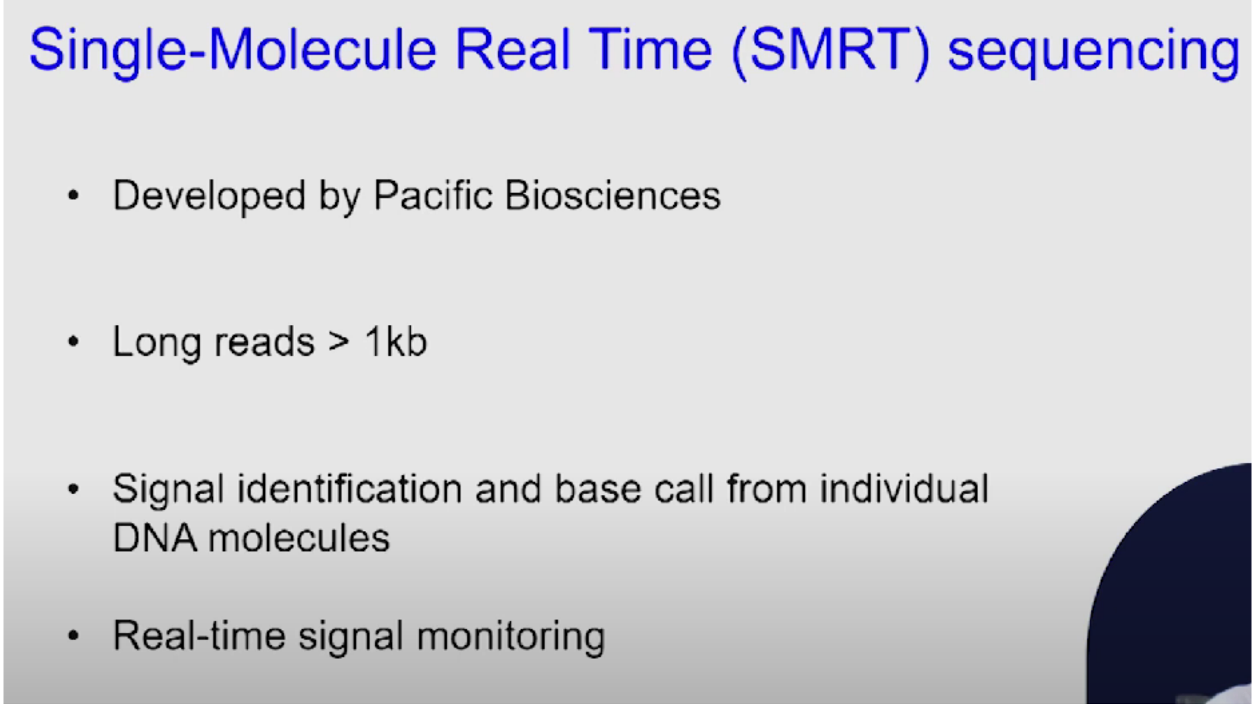
Limitations of Roche 454 and Illumina

- Short reads
- Roche 454 read length - 500-600 bp
- Illumina read length – 2 x 250 bp
- Difficulty in assembly of repetitive regions in genomes and in identification of structural variants

There are general limitations, so one of the things we have seen is that we get short reads. So, for 454 we get about 500 to 600 base pair reads, whereas for Illumina its maximum is 2 times 250 with the current technology. So, this is actually quite short if you think in terms of the genome sizes, right? So, if you are thinking about the human genome, this is 3 times 10 to the power of 9, and compared to that, these fragments are very small, which means if you want to assemble the human genome or such a big genome, you would have to assemble a huge number of reads. So, billions of reads will be required before you can do this assembly. Now, there is another problem: there are a lot of repetitive regions in our genome and in other genomes as well. So, if you have a very short read, even with paired sequencing, you will have a difficult time actually assembling these repetitive regions correctly. So, this is one limitation of both of these technologies. On top of that, we will have structural variations.

So, we will talk about what structural variations are in the next classes. When we talk about variant calling etcetera, you will understand that, and especially for resolving the structural variants these short reads are not really suitable, and we would like to have really long reads. So, longer reads will be much better. So, if you compare Sanger sequencing, we have 1 kB, but can we get at least 1 kB? And if we get more than that right, that is something that would be very helpful for genome assembly. So, as I said, there is difficulty in the assembly of repetitive regions in genomes and the identification of structural variants with these very small reads. So, can we get long reads in NGS right? This is one of the questions that we can ask, and what will happen is that if we can get long reads right, this will actually help in the genome assembly process overall, so it will be less challenging because we can then have a lower number of reads to assemble the genome.

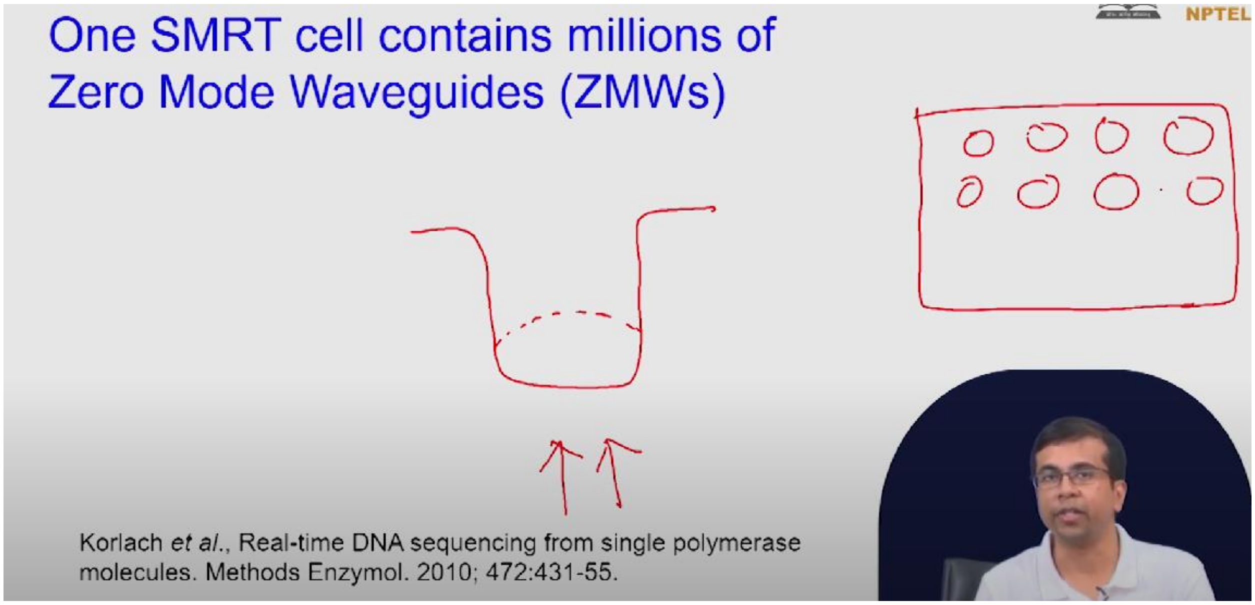
In addition, we can resolve these repetitive regions quite effectively, and we can also identify the structural patterns. So, today we will talk about one such technology that actually allows us to generate these long reads, and this is called single-molecule real-time sequencing, or SMRT sequencing in short. This was developed by Pacific Biosciences, and we get long reads greater than 1 kB and sometimes we also get reads of 5 or 10 kB in length. And the other advantage of this method is that we have signal identification and base calls from individual DNA molecules. So, as we go into the principle, you will see that instead of these clusters or these B, the emulsion PCR will generate a lot of these clones of a molecule, and then do the signal identification and base calling based on the average signal of these clusters. Here, we actually identify these bases from a single DNA molecule. So, this identification happens from individual DNA molecules, which means we have higher resolution, and on top of that, we also have real-time signal monitoring. So, because we are actually following the process of sequencing for individual DNA molecules, we can actually identify when the base addition is happening.



Single-Molecule Real Time (SMRT) sequencing

- Developed by Pacific Biosciences
- Long reads > 1kb
- Signal identification and base call from individual DNA molecules
- Real-time signal monitoring

If you think about Roche or Illumina in comparison right where we are looking at average signal, we are not following individual base addition events because there will be slight differences in time when this base addition happens on the Illumina platform. Between these different molecules, there will be slight differences. And we have also talked about this phasing problem, or we will talk about this phasing problem (pre-phasing etc.) later on, and we will see that this will actually lead to the degradation of the signal later on. So, when you talk about quality of data, you will see this problem coming up with Illumina data, especially OK. So, these are two advantages of this method, along with the long read that we get. So, the question is now: how does this sequencing happen, and how do we actually get this long read? So, we will talk about this principle now. So, sequencing is done on something called SMRT cells, like a flow cell or a picotiter plate. Here, we have these SMRT cells right where we can do the sequencing. So, it is kind of like a box-like structure where this sequencing would happen.

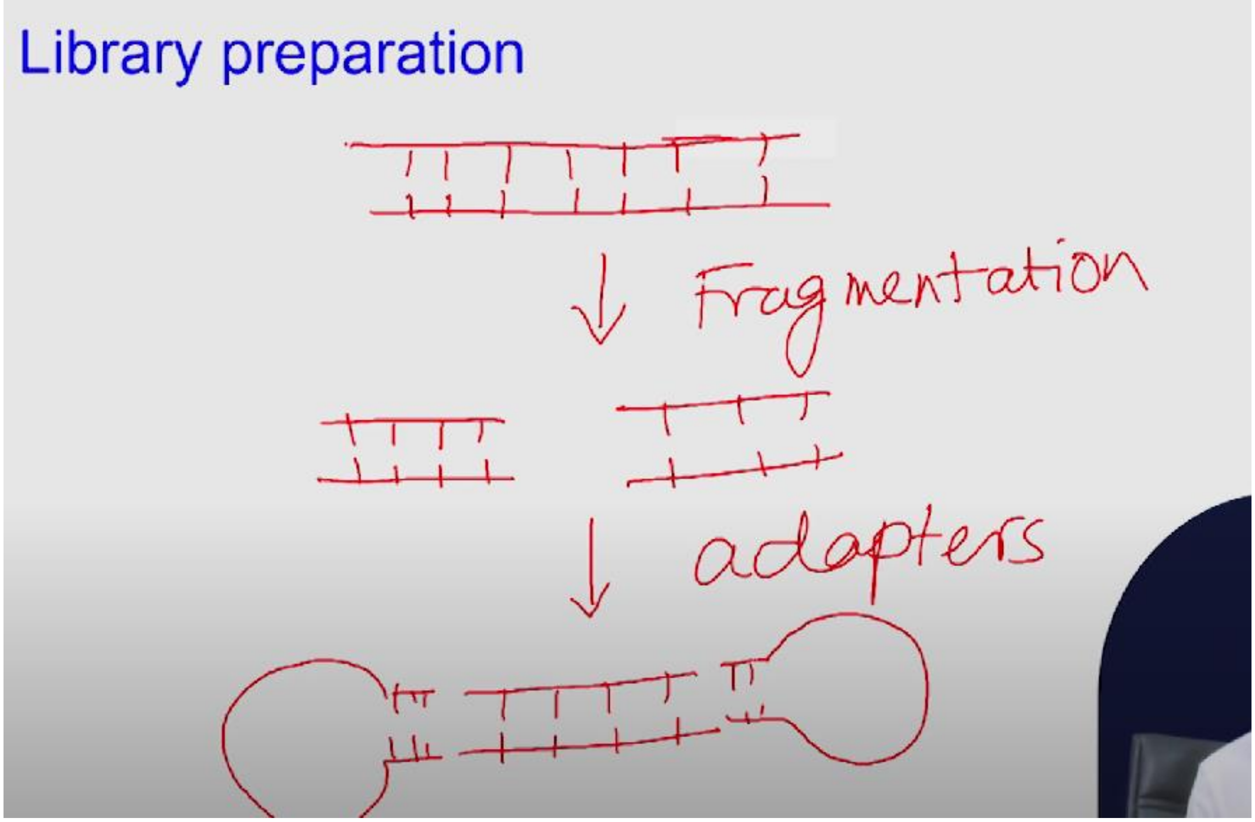


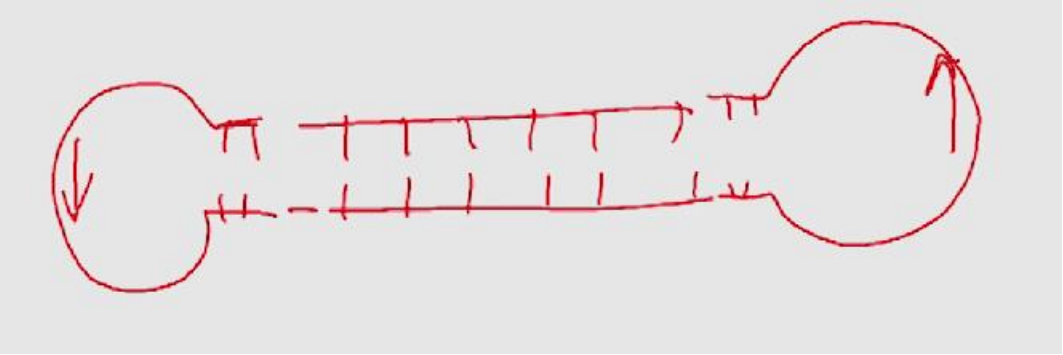
One SMRT cell contains millions of Zero Mode Waveguides (ZMWs)

Korlach *et al.*, Real-time DNA sequencing from single polymerase molecules. Methods Enzymol. 2010; 472:431-55.

So, you can look up search and look up for these images, and you will find quite a few pictures of these SMRT cells, ok. So, inside this we have something called zero mode waveguides, ok. So, one SMRT cell contains millions of these zero-mode waveguides, and you can look at this reference to actually see how these zero-mode waveguides look. So, I will just draw a kind of schematic of this right. So, it is kind of like a well, or something like this, but the dimensions are so small that if you shine light from one side, the light cannot pass through the well. So, what will happen is that only the lower part of these wells will be illuminated, and the light will not pass through the well because the dimensions are so small. And in the SMRT cell, we have millions of such zero-mode waveguides. So, you can imagine, right? If you kind of imagine in your mind right here, you have this, and we are looking from top right. So, we are looking at all these wells, and you have millions of them in the SMRT cell, ok? So, which means we can do this sequencing for millions of these molecules in parallel inside one SMRT cell? So, what happens is that sequencing actually happens inside these ZMWs. So, one of the terms I will just introduce is that we will see these terms: SMRT 1M and SMRT 8M. You will see this. So, it simply tells us how many ZMWs are present inside the SMRT cell. So, 1M means 1 million ZMWs per SMRT cell, and 8M is 8 million ZMWs per SMRT cell. So, as I showed you right, these sequencing reactions actually happen here inside these wells, and especially, to be more specific, they actually happen in this region, which is illuminated. So, now how does this process occur? So, we will go to that in a moment. First, we talk about library preparation, and then we will go into the sequencing part. So, for library preparation, we have these genome sequences right. So, this is the double-stranded

DNA in the genome, and the first step is fragmentation, ok? So, we generate these small fragments from the genome, and this is required because, again, we cannot sequence the full genome at one go. We need to generate these fragments, but in this case, in the case of SMRT sequencing, these fragments could be bigger compared to Illumina or Roche because we can now have reads of 1 kB or even more. So, we can have these bigger fragments generated, but nonetheless, we need this fragmentation step. So, once we have this fragmentation, we need to add the adapters, and one of the things you will probably notice later is that this method does not require any amplification. So, because we are sequencing single molecules, we do not need to generate clusters like Illumina right or the on-bead amplification that we see in 454 ok. So, the adapters are very special here. So, we have something like hairpin adapters, and they kind of look like this, and we can have two different adapters on two sides, ok? So, it kind of looks like this dumbbell shape, right? So, we have these hairpin adapters. The special part is that you have this double-stranded part right here. This is the double-stranded part of this adapter, and then you have the single-stranded part right here. So, these are called SMRT bell hairpin adapters. So, we kind of get these dumbbell-shaped molecules after the adapter addition. So, what actually happens is that if you kind of get this dumbbell shape, this is the adapter part.

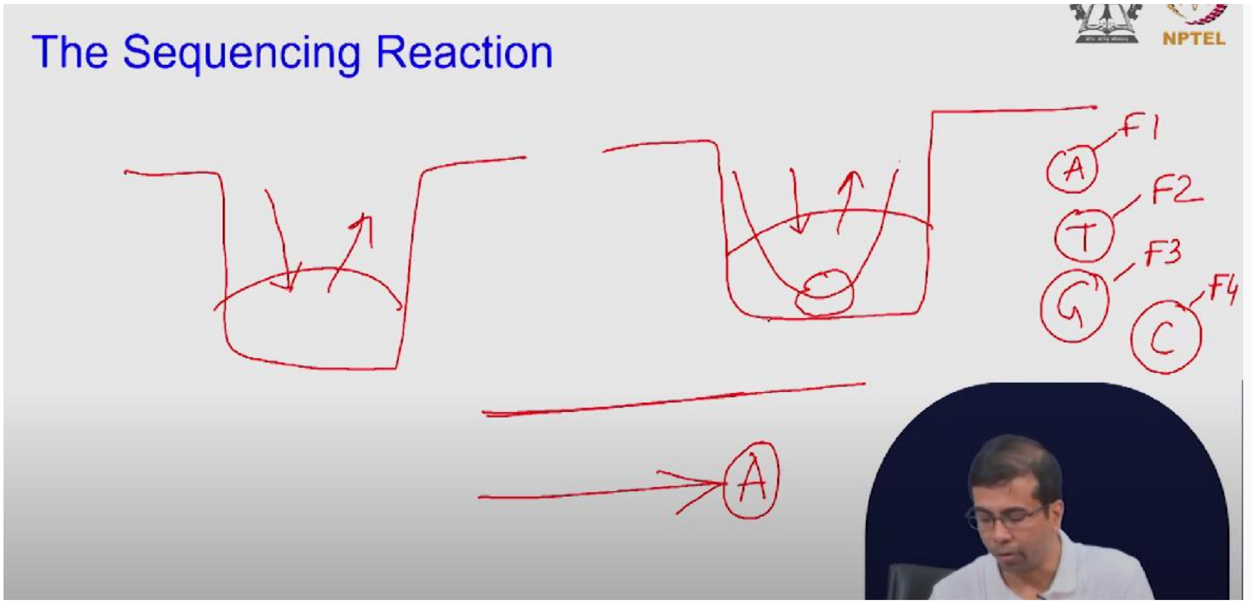

Library preparation



You said in the single-stranded part here that you can have primer binding sites, right? This is where you can have your primer bind and then sequence the molecules. So, you can probably now guess that this is a sequencing by synthesis method. You have the primer, you have polymerase, and you have dNTPs, which will then synthesize the complementary strand, and by synthesis in the synthesis process, it will identify the bases. So, let's now go into the sequencing process itself. How is this sequencing actually happening? So, for this synthesis, we have DNA polymerase, but we also have special dNTPs.

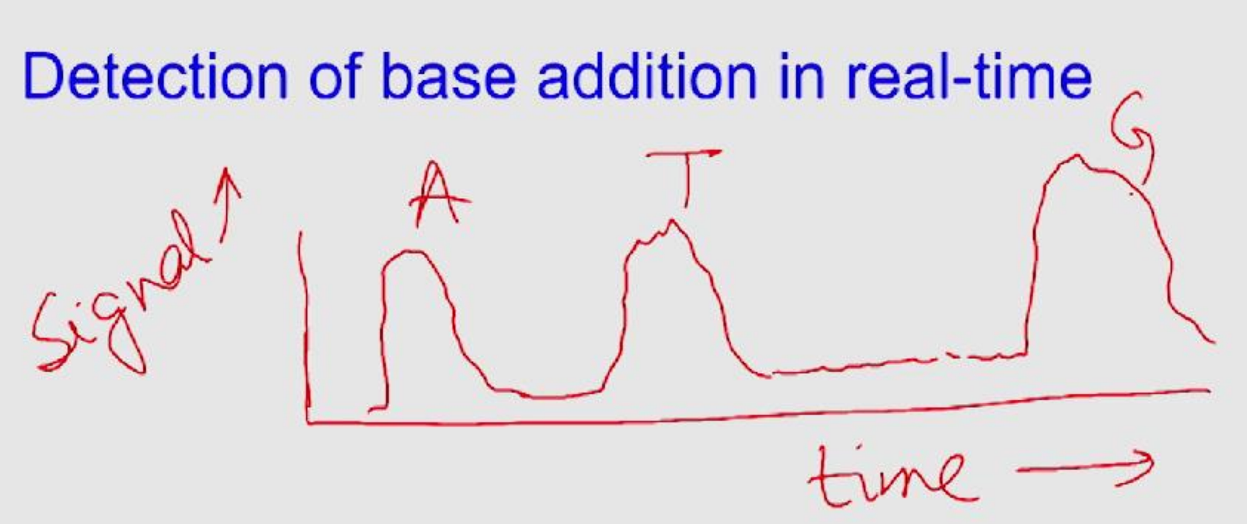
So, in these dNTPs, we have fluorophores, and each dNTP will have a unique fluorophore. So, dATP will have, let us say, a red color, and dGTP will have blue etcetera, etcetera. So, each type of DNTP will have a unique fluorophore. So, which can then be identified by a detector? So, the fluorescent signal can be detected ok.

So, in this dNTP, there are fluorophores linked to the 5 prime phosphate groups of the bases ok, and what happens is that during the synthesis process, these fluorophores are released right when the phosphate comes out, the pyrophosphate comes out right during the synthesis process, and along with that, the fluorophore also comes out ok. So, you can now visualize in your mind that this synthesis is happening and the fluorophores are coming one by one. So, now let us look at this in much more detail, right? So, as I mentioned, we have the ZMWs right, and this is where the sequencing is happening right. The lower part is illuminated, and inside this ZMW, you have the DNA polymerase sitting at the bottom of the well, and it then binds to the DNA. The DNA molecule comes in right, and of course, you would have to ensure that the right polymerase gets only one type of DNA molecule. This can be optimized again through trial and error. Now, once this DNA molecule comes in here right along with the adapters, you can now start the synthesis process, ok? So, now you remember right, we have these dNTPs that are labeled, right? Some sort of labeling is there, and these are also unique, and F1, F2, F3, and F4 are these unique fluorophores. Right there, these are different fluorophores that can be detected. Now, what happens is that, as you can imagine, we have the synthesis happening. So, here is your template and here is the primer, and we give all these fluorophores together, and the polymerase then selects the right one based on the complementary sequence. So, let us say you have this addition of A here, and during this addition of A, the fluorescence will come out, the fluorophore will be released, and the detector will detect these fluorescences.

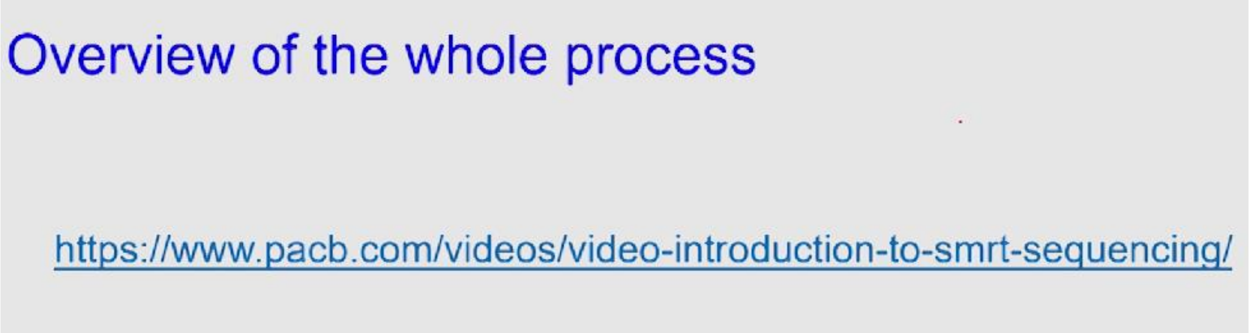


The Sequencing Reaction

And what will happen next is that we give, then get the next one right, we get to the next base, and maybe next it is G that is attached right, and this fluorophore will then come out and it will be detected by the optical detector, ok? So, this is happening one by one, right? You have this addition of base, then you have the release of the fluorophore and the detection of the signal. So, this fluorophore gives us the identity of the base. Now, one of the questions you might ask right now is: Since fluorophores are also present in the well, why do they not give any signal? So, these bases are floating around before they are added to the growing strand, okay? why do they not give a signal? Now, when these base bases are free, free dNTPs are moving in and out here, right? So, what happens? They are coming in and out right away, very randomly, ok? So, maybe I can draw this separately. If you think about this region, we have these dNTPs coming into the illuminated region and going out right there. This is random motion. So, they are coming in and out very, very fast. So, the time that they spend in this illuminated region where the fluorescence signal can be detected is very short. Only when they are binding to the growing strand right, the polymerase holds them for long enough for the detector to detect them. So, this ensures that we are detecting only the bases that are being added to the growing strand. And this way we can get the signal that is coming out in real time, right? So, you can now see why this is called single-molecule sequencing because the signal comes from the synthesis of the complementary strand of a single DNA molecule that is coming here and binding to the polymerase, where the synthesis is happening. And in addition you can see we can detect these signals in real time because the synthesis is happening at the single-molecule level and we are observing those individual base addition events. So, compared to Illumina and 454 this will give us higher resolution, and you will see this kind of real-time sequencing will have some other applications. We will discuss that later on when we actually talk about those applications. So, I hope this is clear how the sequencing reaction happens. This is sequencing by synthesis, where we have the fluorophores being added and the fluorescence signal being released, which is then detected by the optical detector. So, this detection of base addition happens in real time, as I have just mentioned.


Detection of base addition in real-time

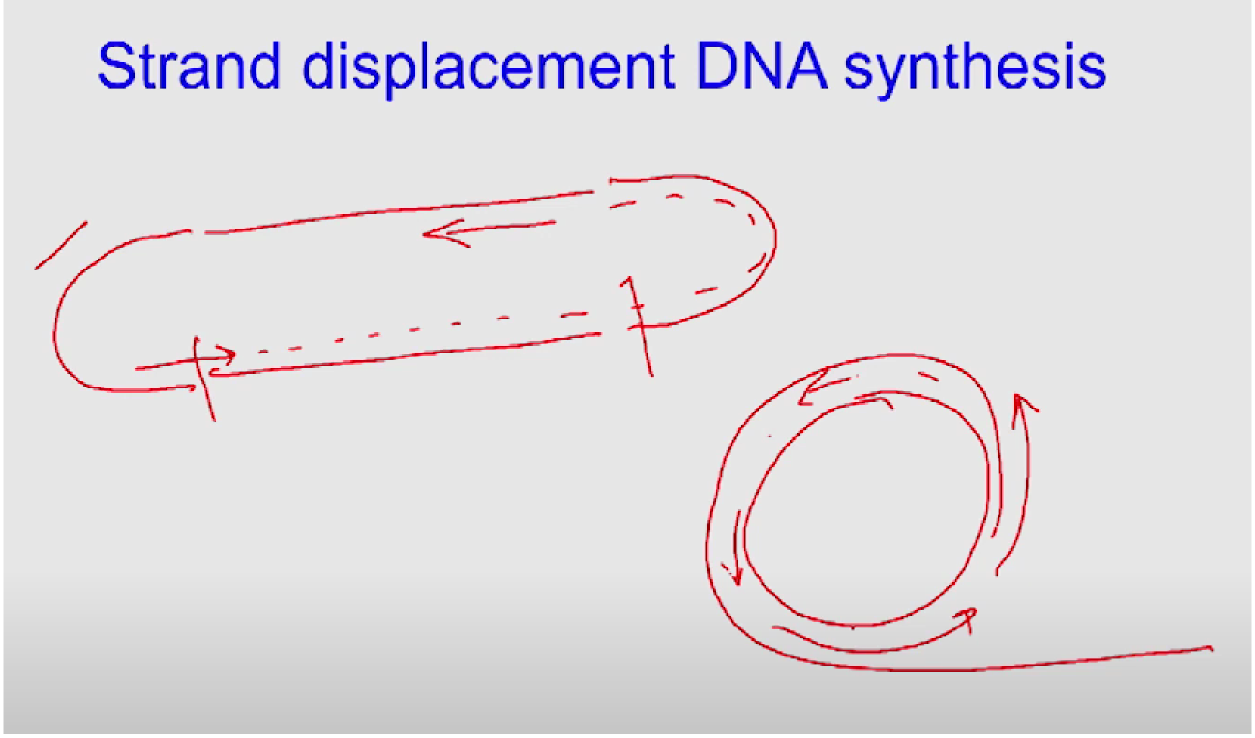
So, if you have the detector, you can probably see right if you kind of draw this right. So, this is time versus signal on the y axis. o, the detector will then detect, OK, there is this addition of A again. We have different colors, right? So, I am just drawing by one color, right? Maybe then you have the addition of T, right? Maybe after a certain time we have the addition of G, and so on, right? So, you are detecting these signals, and this data is then also stored by the machine. So, we can actually retrieve this kind of data, and this can also give us some valuable information, as we will see later on. Now, when we are detecting this base addition in real time, we can actually generate some of the statistics, and that is what we will talk about now.



Overview of the whole process

https://www.pacb.com/videos/video-introduction-to-smrt-sequencing/

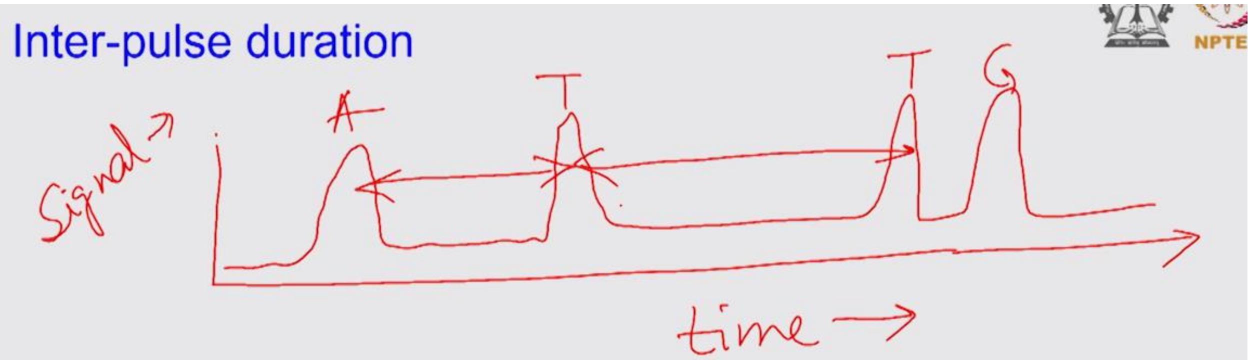
Before we go there, here is a link where you can actually see the animation of the whole process, and you will probably get a really good idea of how this process is happening. So, please look up and see the animation, okay? So, one of the things this method can do is something called strand displacement DNA synthesis. So, if you have used a specific polymerase you can actually do something like strand displacement DNA synthesis. So, what exactly is this strand displacement DNA synthesis? As I mentioned right here, you can remember right right we have this kind of shape, right? These are the adapters here, pin 1 and here.



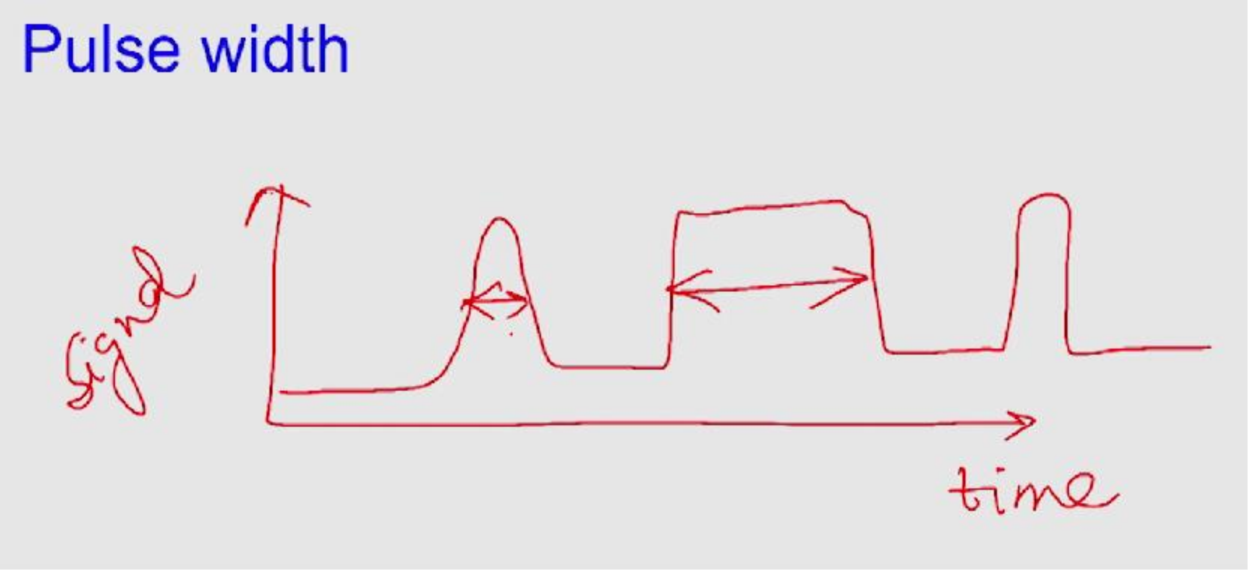
Strand displacement DNA synthesis

So, I am just now separating them out, and you can have this sequencing process going in this direction, going here, and also continuing right. So, you can operate this AH sequencer in such a way that you get just long reads, ok? So, you start from here and end up somewhere here, and you get the full sequence of this DNA fragment, and this fragment could be quite long. You can have 1 kB, 2 kB, etcetera, and you can get the full sequence. Another way you can actually run this sequence is through something called strand displacement DNA synthesis. So, depending on the polymerase, what you can do is simply So, if you consider this a circular molecule, you can simply continue to synthesize these molecules over and over again. So, if you just imagine this here and convert it to a circle, So, you can just keep on synthesizing the same molecule over and over again, ok? Why would you do that? So, we will talk about that in a moment. Why why we want to do this? Why do you want to sequence the same molecule multiple times? So, but there is this capability where you can actually do this strand displacement DNA synthesis.

So, if you go over the same sequence multiple times, ok. So, since you are observing the signals in real time, there are two parameters that you can measure, and again, this will actually have some implications that we will see later on. So, one is called inter-pulse duration ok.

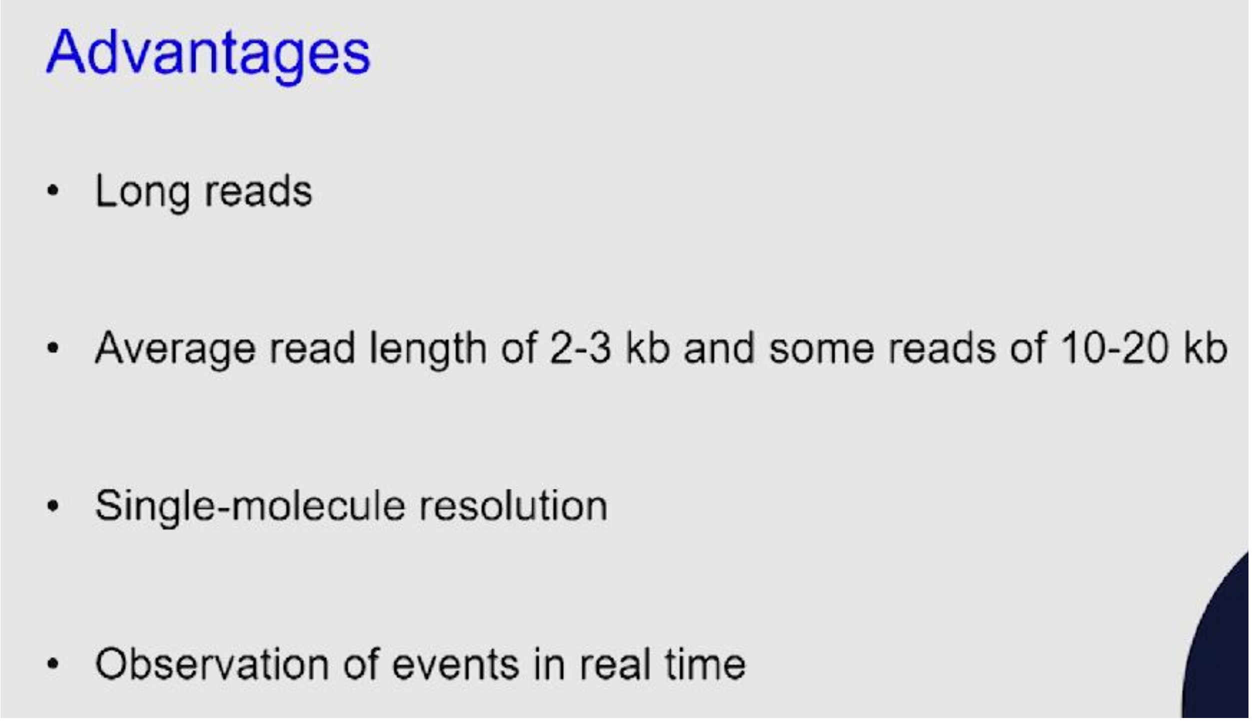


So, this is the signal that we are detecting. If we can imagine it again, we have time and a signal, and let us say we are detecting these different signals. So, you can imagine this is what is happening in real time, and let us say this is maybe T. This is also T. This is also G. So, maybe this is what is happening. So, one of the parameters that you can measure is the time between two pulses or two signals. And you might ask, right, what is the importance of these? As you will see, we can use the signal for specific applications. This specific time difference will actually tell us something about the DNA. So, we will talk about that later, but this is something that is quite important: we can measure this interpulse duration because we are observing the sequencing process in real time.


Pulse width

There is another parameter that we can also measure, which is something called pulse width. So, this is again if we draw the time versus signal. So, this is what is meant by pulse width. So, what you will observe, right? If you kind of look at the real data, you will observe these pulses of different widths, ok? And again, the question is, why is this important? So, I will tell you that this kind of signal can then be combined with inter-pulse duration, which can help us understand the modifications to DNA better. So, we will talk about this in much more detail later on, but just to be aware that these parameters can be measured correctly and can be stored in the data, which will be very useful for different types of analysis.

So, there are two parameters we have talked about: one is inter-pulse duration, and the other is pulse width. So, I hope this is clear right what inter pulse duration and pulse width is ok. Now, what are the advantages of SMRT sequencing?



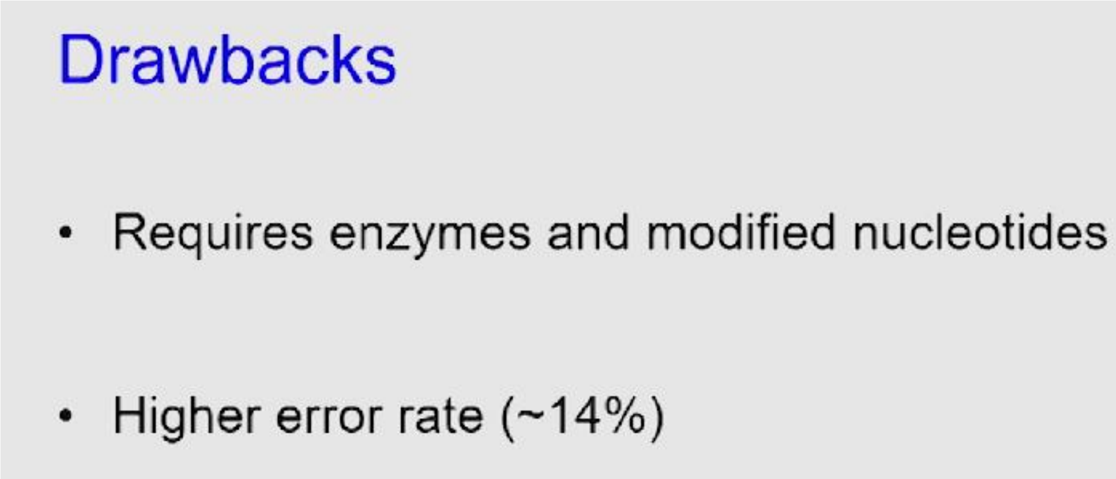
Advantages

- Long reads
- Average read length of 2-3 kb and some reads of 10-20 kb
- Single-molecule resolution
- Observation of events in real time

So, the first advantage is the long reads that you get, and this is something that kind of comes out favorably for this method compared to Illumina or 454 right. So, you get really long reads, and that kind of thing will be helpful for the genome assembly process. And we get an average read length of 2, 3 kB, and some reads may be quite big right at 20 kB, but at least we get more than 1 kB right.

So, which is more than double the number of existing sequencing technologies? We have single-molecule resolutions, right? So, as we have seen, we are getting this data from a single DNA molecule. We are synthesizing complementary strands of a single DNA molecule, and this process is happening across multiple wells and multiple ZMWs. We have discussed that, and millions of these ZMWs right now are getting parallel signals, which will give us sequences for millions of molecules, but at single-molecule resolution.

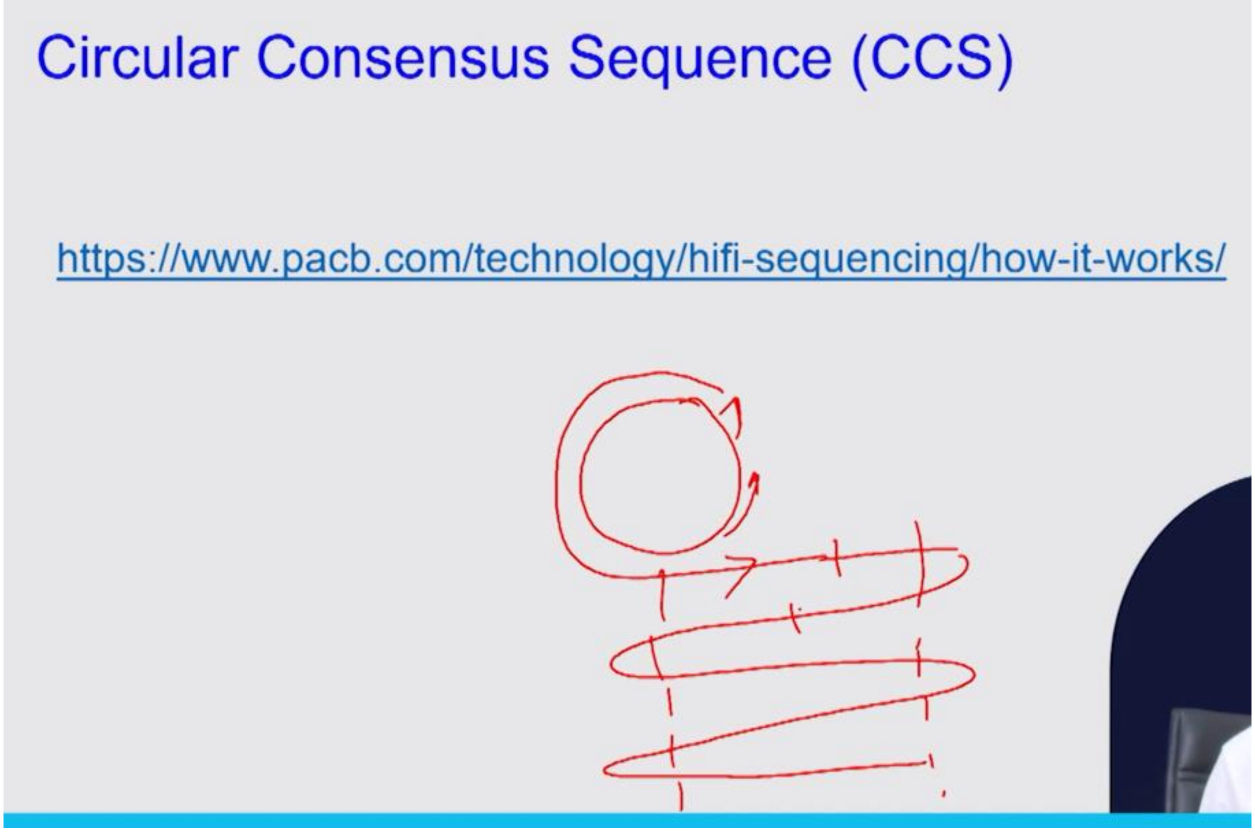
So, it is incredibly powerful. And we also have observations of events in real time, and we have seen that we can measure these two parameters, inter-pulse duration and pulse width, which will have some applications that we will talk about later. Whatever the drawbacks, as we have seen, we require enzymes and modified nucleotides. So, it means the running cost would be a bit high, but probably very similar to Illumina. Again, in Illumina, we have modified nucleotides and polymerase. The major drawback is that it has a higher error rate, which is about 14 percent.



Drawbacks

- Requires enzymes and modified nucleotides
- Higher error rate (~14%)

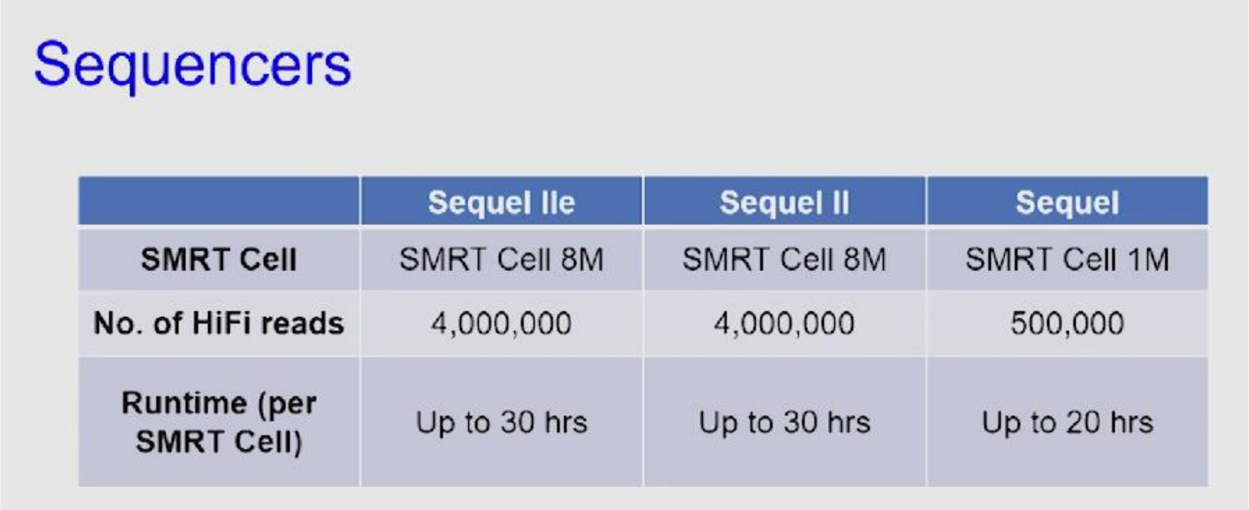
So, this again varies; it is something that has been measured, but it is significantly higher than Illumina. And why is that the case? So, what we think is that because this polymerase is exposed to light, it kind of affects the proofreading activity of the polymerase. So, to avoid this issue of a higher error rate, there is a method that has been developed called the circular consensus sequence or CCS. So, what actually happens is that we get something called strand displacement synthesis. So, we do this strand displacement synthesis, and we get something called high phi. So, what is a high phi? So, this is high-fidelity reading, ok? So, as I have just mentioned, we do strand displacement. So, what is happening is that we are sequencing the same molecule over and over again, right? So, we are getting just data from this right we are getting such like for the same region we are getting this multiple reads mapping to the same region and this will reduce the error rate right. So, you can imagine if you are sequencing the same region multiple times and in one call we have an error, and in another call we see a mismatch here compared to the reference sequence. So, we can see that we can say, OK, these two are probably sequencing errors because if it is a true mutation, it should be appearing in all the reads that we get right.



Circular Consensus Sequence (CCS)

https://www.pacb.com/technology/hifi-sequencing/how-it-works/

So, we are kind of sequencing the same region over and over again, and we are calling the consensus right. So, all the errors can then be taken care of by this process, ok? So, again, here is a link where you can probably see how this method works, and we will get a better idea of this through the animation. And the advantage is that hi-fi reads are now long, as well as accurate. So, this is something that is kind of game-changing because, as you have seen so far, the short read sequencing methods are very accurate. So, that's good, but they have some drawbacks, ok, and to counter that, this wrong read long read sequencing was developed, but it had a higher error rate, ok. And to circumvent that, if you can then make sequence long reads right and also get really accurate data, that is the ideal situation, right? So, this is something that is very helpful and useful for all sorts of purposes, ok? So, just to give you some idea, So, how many reads do you get? How many hi-fi reads can you get for this asymmetry cell? So, this is from Pacific Biosciences specifically.



Sequencers

	Sequel IIe	Sequel II	Sequel
SMRT Cell	SMRT Cell 8M	SMRT Cell 8M	SMRT Cell 1M
No. of HiFi reads	4,000,000	4,000,000	500,000
Runtime (per SMRT Cell)	Up to 30 hrs	Up to 30 hrs	Up to 20 hrs

So, you have Sequel 2e, Sequel 2 systems is the name of the sequencers and the SMRT cell that you can use, and Sequel 2e and Sequel 2 are the recent modern ones, and you can get up to 4 million reads from them. So, here are the references that you have used for this class. And to summarize, we have introduced single-molecule real-time sequencing, which is based on the detection of fluorescent signals from the synthesis of DNA that is happening inside ZMW. So, these are incredibly small wells where the polymer is immobilized and the synthesis process happens. So, this is something that is actually also in contrast with the 454 and Illumina. o, as you have seen in both of those methods the DNA is immobilized on the picotiter plate or on the flow cell, and the polymerase comes and does the synthesis. Here in SMRT, it is actually the opposite, where the polymerase is immobilized at the bottom of the wells, the DNA molecule comes, and the sequencing reaction happens. And what we have seen allows us to observe sequencing in real time and at the level of single molecules which actually has some real benefits for different applications, and we will talk about that later. And from these real-time observations, we can derive certain parameters, which we also talked about, and we can utilize those parameters for some applications. And what we have also introduced is that the original long read sequencing actually has a very high error rate, whereas the circular consensus reads, the CCS reads, actually deliver long read data, but also with high accuracy. These are called hi-fi reads or high-fidelity reads, and these will have very important applications in genome analysis. Thank you.