Next Generation Sequencing Technologies: Data Analysis and Applications

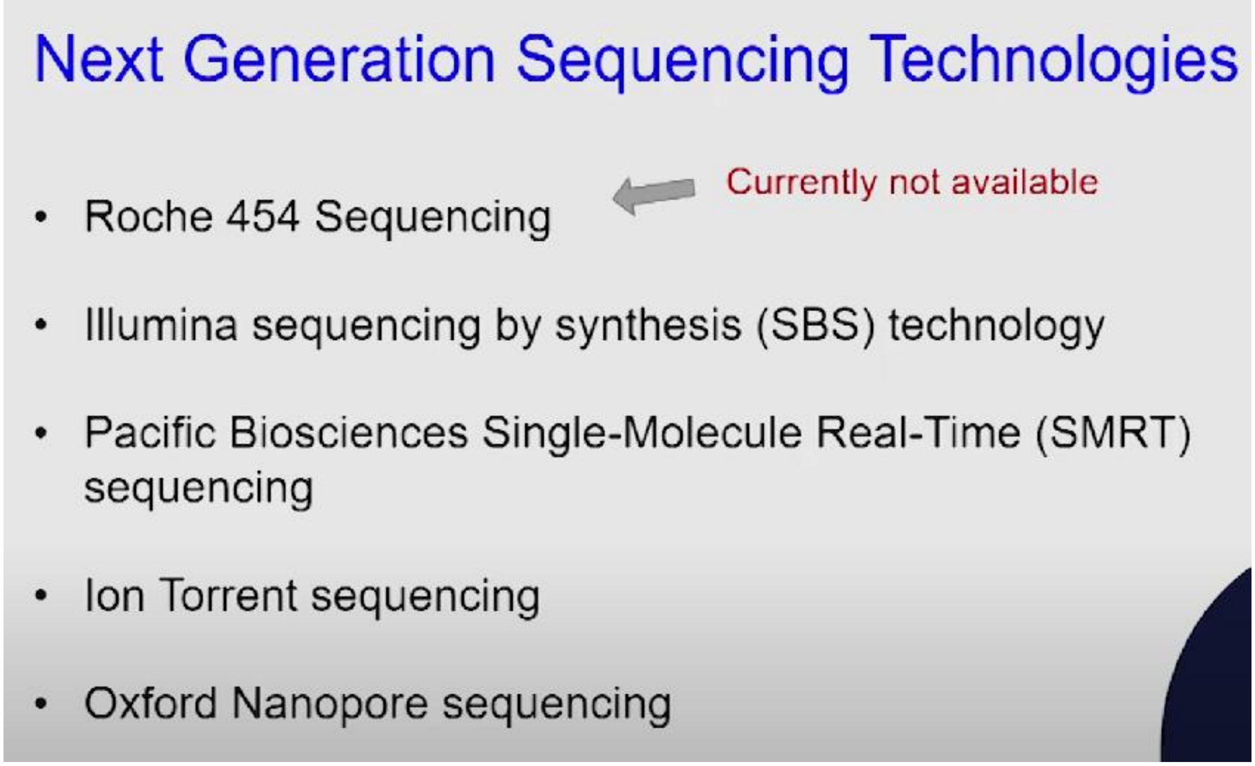
Next Generation Sequencing Technologies - 454 Sequencing

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 class, we talked about next generation sequencing, and from this class, we will now introduce the technologies themselves and discuss how these technologies have actually accelerated the pace of genome sequencing and how they have provided this high throughput sequencing. So, in this class, we will talk about one of the first next generation sequencing technologies, which is 454 sequencing. So, let us start in this class. We will talk about something called reads and what these reads are. We will talk about DNA library preparation and how you actually prepare DNA libraries for nucleotide sequencing.

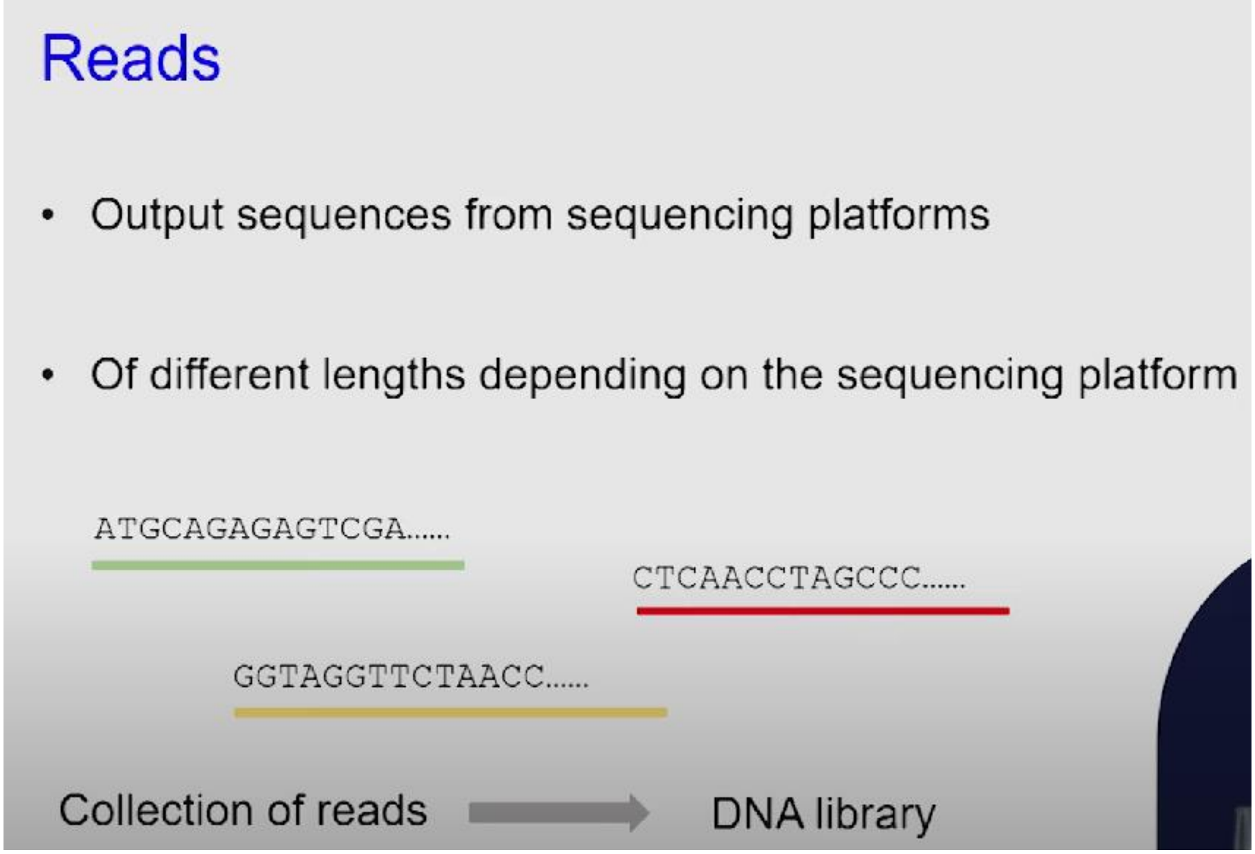
We will talk about adapters, and then we will talk about the Roche 454 sequencing method. These are the keywords that will come up: pyro sequencing, emulsion PCR, homopolymeric stretches. So, there are several next-generation sequencing technologies that are available, and these are some of them. The first one is Roche 454 sequencing, then we have Illumina sequencing by synthesis technology, we have single-molecule real-time sequencing from Pacific Biosciences, we have ion-durant sequencing, and finally, we have Oxford nanopore sequencing.



Next Generation Sequencing Technologies

- Roche 454 Sequencing ← Currently not available
- Illumina sequencing by synthesis (SBS) technology
- Pacific Biosciences Single-Molecule Real-Time (SMRT) sequencing
- Ion Torrent sequencing
- Oxford Nanopore sequencing

This Roche 454 sequencing is currently not available. So, if you want to do high-throughput sequencing, you cannot use this 454 platform; you will have to use any of the other 4 that are available today. Now, despite that, we will discuss the ROSE 454 sequencing because this was the first method of next-generation sequencing, and you will also find a lot of data in databases from this technology. So, it is good to understand how this data was generated before you can actually analyze those data sets. So, before we go into the technologies themselves, So, let us talk about what the readings are. So, these are actually output sequences that you get from sequencing platforms, ok? So, instead of saying fragments or etcetera, we will use this term throughout this course, and they can be of different lengths depending on the sequencing platform. Again, as we discuss sequencing platforms, we will talk about read lengths, etcetera, and here are some examples. You can have different sequences of these reads, and these collections of reads are actually coming from the DNA library.



Reads

- Output sequences from sequencing platforms
- Of different lengths depending on the sequencing platform

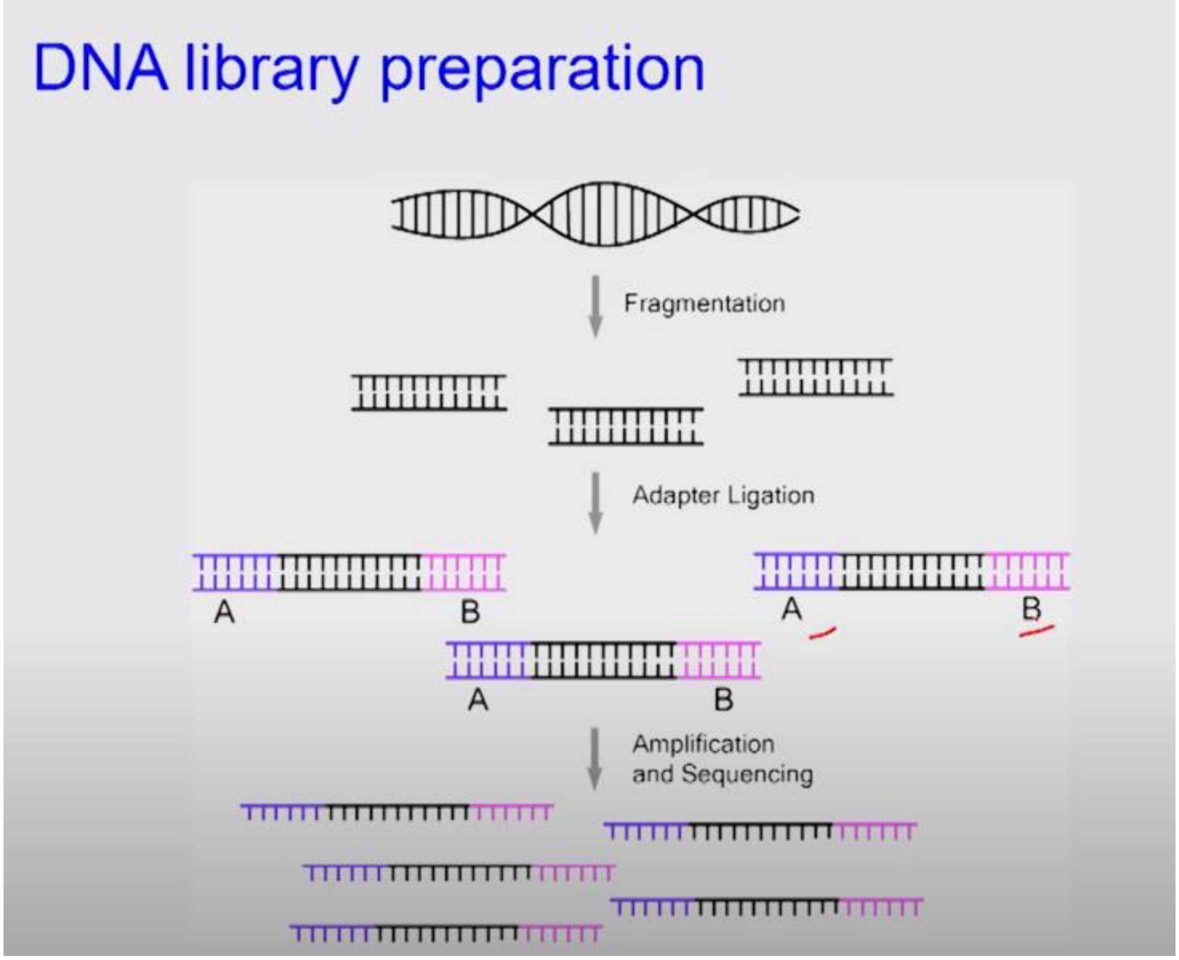
ATGCAGAGAGTCGA......

CTCAACCTAGCCC......

GGTAGGTTCTAACC......

Collection of reads → DNA library

So, you have this genomic DNA from which you prepare this DNA library, and that actually gives rise to these reads. So, how do you prepare this DNA library? So, here is a schematic of the full process. So, like Sanger, we start with the genomic DNA, and we generate these fragments okay because we cannot sequence the full genomic DNA in one go because our genome is quite big. Now, after you have fragmented this fragmented DNA, what you do is something called adapter ligation. These adapters are shown in two different colors here, as well as by the letters A and B.



DNA library preparation

Fragmentation

Adapter Ligation

A

B

A

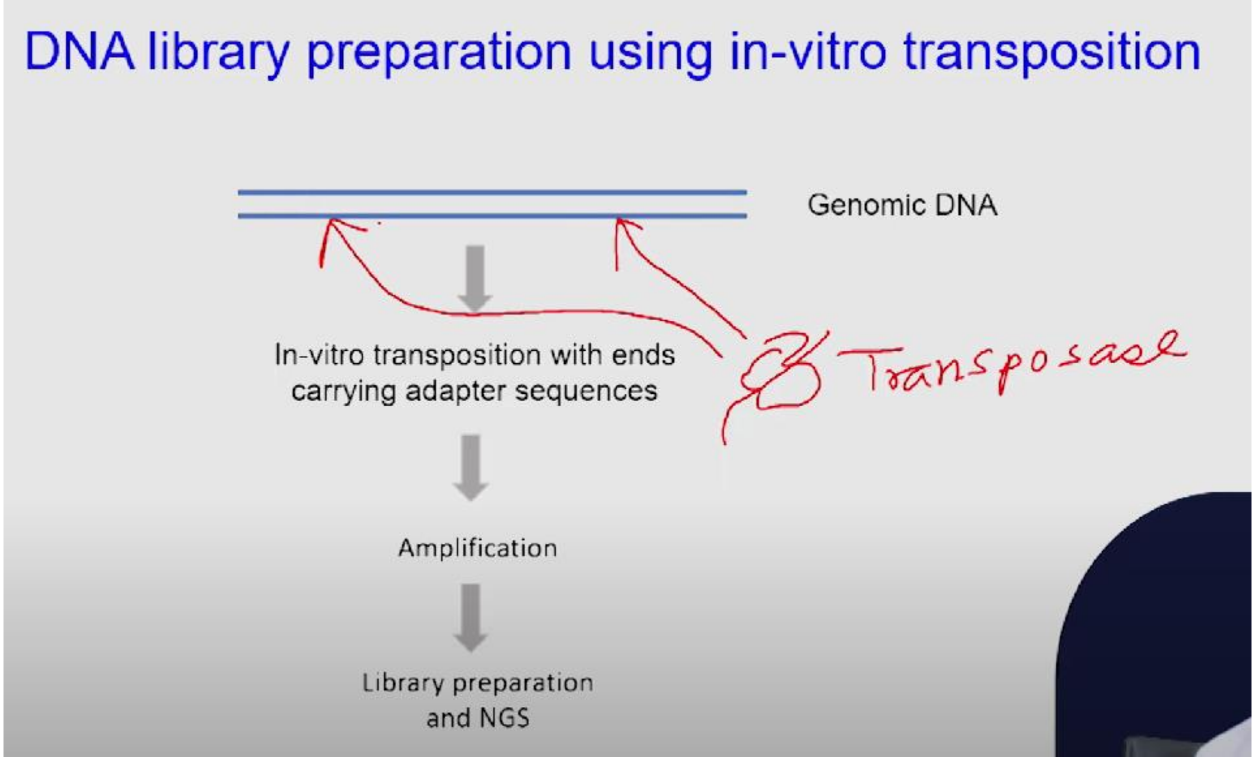
B

A

B

Amplification and Sequencing

You can see these letters A and B right there. These are usually two different adapters that are added to two different ends, and there is a specific reason, which you will probably realize when you talk about the methods of the sequencing. Now, in this case, we have added these adapters to these fragments, and then you have something called amplification. So, you amplify some of these fragments, and this is required for some platforms, whereas on other platforms you do not require amplification, and once you have done amplification, you do sequencing using these different sequencing platforms. Again, the sequencing method will be different, right? So, up to adapter ligation, this is kind of common across all platforms, but this amplification and sequencing step are platform-specific. So, some platforms may not require even amplification, okay? So, once you have this DNA library right, you can also prepare this library using in vitro transposition. So, this is a different method than what we just discussed, ok? So, I will just elaborate on this very briefly. So, you have genomic DNA, and instead of fragmenting it, what you do is something called in vitro transposition.



DNA library preparation using in-vitro transposition

Genomic DNA

In-vitro transposition with ends carrying adapter sequences

Amplification

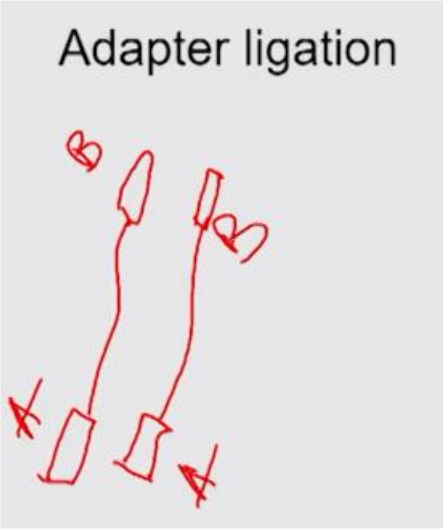
Library preparation and NGS

So, you have heard about these enzyme transposes, right? So, these actually carry some DNA fragments and insert them into the genomic DNA at random places, ok? So, once you have isolated genomic DNA, you do transposition in vitro in the test tube, and these sequences that are inserted contain adapter sequences. So, imagine you have this transpose enzyme right like this. So, this is your transpose, right? It carries some DNA fragments, and this is actually your adapted sequence.

So, what it does is kind of go and insert these adapter sequences at random places, ok? So, once you have these adapter sequences inserted in random places, you can amplify these fragments and go for NGS, right? So, you can have this library preparation step, which again could be specific to the platform, ok? So, you can use these two methods, and you can also do this transposes method. In some cases this probably might be the only solution that you have. So, what is the function of these adapter sequences that we just talked about? So, you have seen this right. So, why do you add these adapter sequences? So, these adapter sequences actually contain primer binding sites, right? So, as you have seen, in many cases, we need to do amplification of the DNA fragments. Now, how do you amplify these DNA fragments? We do not know the sequence of these fragments. For amplification, you need a primer binding site; you need polymerase right, which will start extension from this primer sequence. So, to generate this primer binding site, we actually add these adapter sequences. So, they contain some sequences that can be used as primer binding sites. In addition, these adapters can contain something called index sequences. So, these are again some base pairs that enable the mixing of multiple samples, right? So, you have some sequences that could be unique, and you can use these unique AH index sequences for different samples. Because you can generate a huge amount of data with this next-generation sequencing technology, you can sequence, let's say, 96 samples in one go. Now, how do you mix them and separate out this data? So, to do that, you use these index sequences, and you can have one unique index sequence per sample. So, if you are mixing 96 samples, each of them will have a unique index sequence, and this process is known as sample multiplexing, where we can mix multiple samples with these index sequences. And once the sequencing is done, we can then look at these index sequences and separate out the samples. So, this is actually known as demultiplexing. So, there are two terms: sample multiplexing and sample demultiplexing. So, let us now talk about the 454 sequencing method itself. So, this was the first NGS method that was developed; it was developed by a company called 454 Life Sciences in 2006 around that time. And in 2007, it was acquired by the company Roche, and it actually was closed down. This 454 sequencing was closed down in 2013.

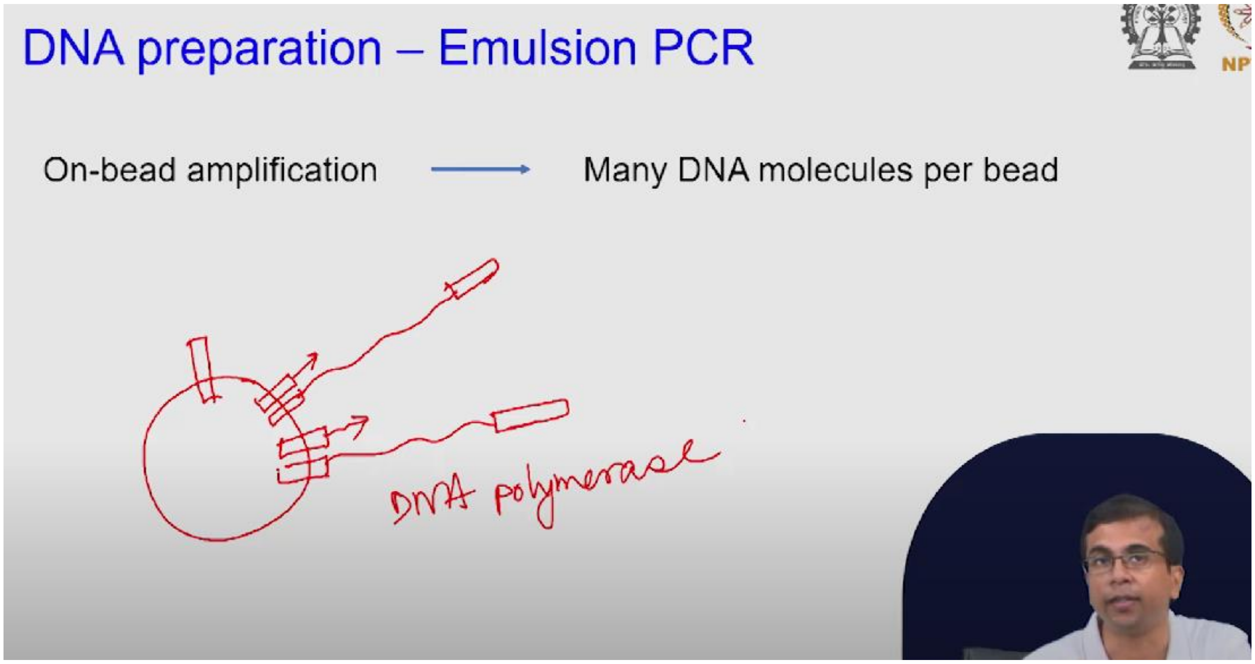
So, we will talk about why this was closed down. So, quickly, within 6 years, ah, this was closed, ok. So, how does this method work? How do you actually get the sequence data? So, 454 sequencing relies on something called pyrosequencing. So, you may or may not have heard about this pseudosequencing, and we will talk about what this method is in much more detail.

So, in pyrosequencing, 454 sequencing actually took this pyrosequencing method and massively paralyzed this sequencing process, ok, and again, we will see how this actually happened, ok. So, in pyro sequencing, given a fragment, you have to synthesize the complementary strand, like in single sequencing, you have to polymerase the NTPs, etcetera, and in this, you have to amplify the complementary strand, ok. And these bases are called right when these bases are called ah from these light signals that appear during the synthesis process. So, again in a moment, you will see how this happens, and that is why the term pyro comes from ok. So, one of the first things we do is adapter ligation. As we have mentioned, we have this DNA fragment right, and we can have adapters A and B right, and these two ends of these fragments are ok.



Adapter ligation

So, for DNA, we have these adapters ligated, and once we have these adapters, what we do in 454 sequencing is something called bead amplification. So, these are small beads on which we already have some DNA sequences, like probes, and these are complementary to these adapter sequences. So, what will happen is that some of these DNA molecules will come together and bind. So, one adapter will be complementary, right? So, they will form this double-stranded part, part ok.

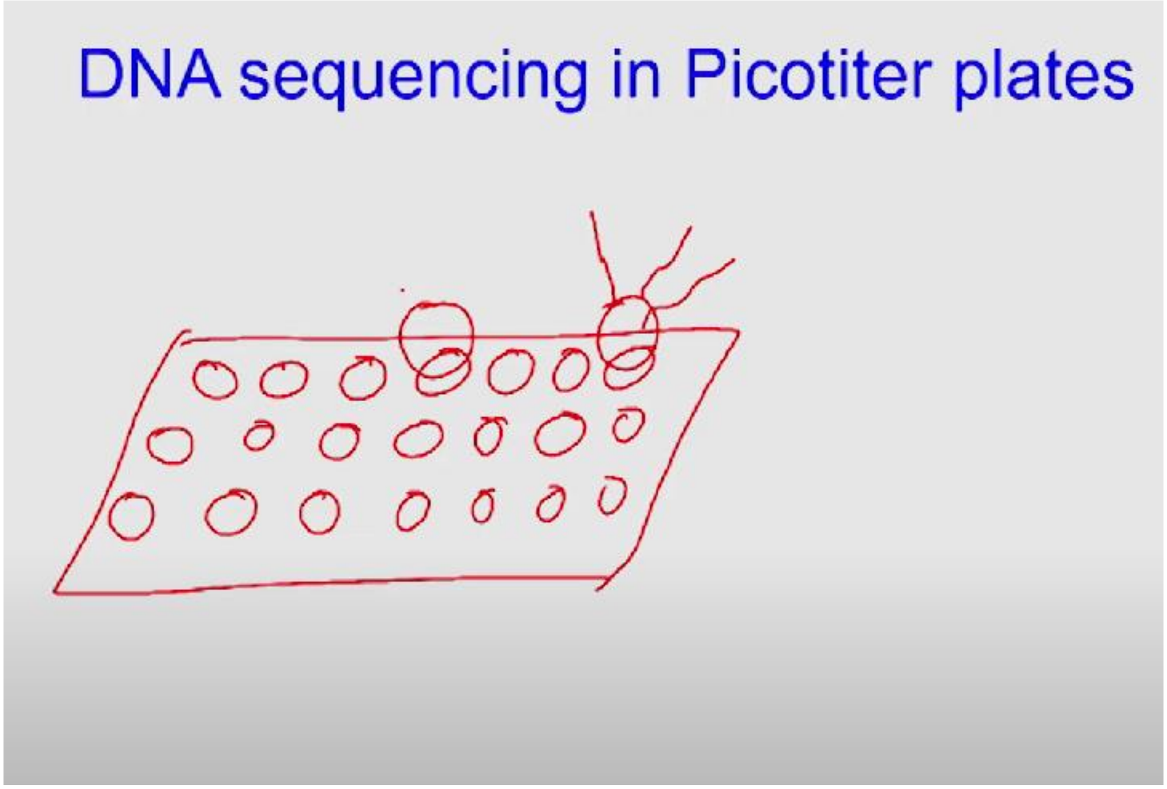


DNA preparation – Emulsion PCR

On-bead amplification → Many DNA molecules per bead

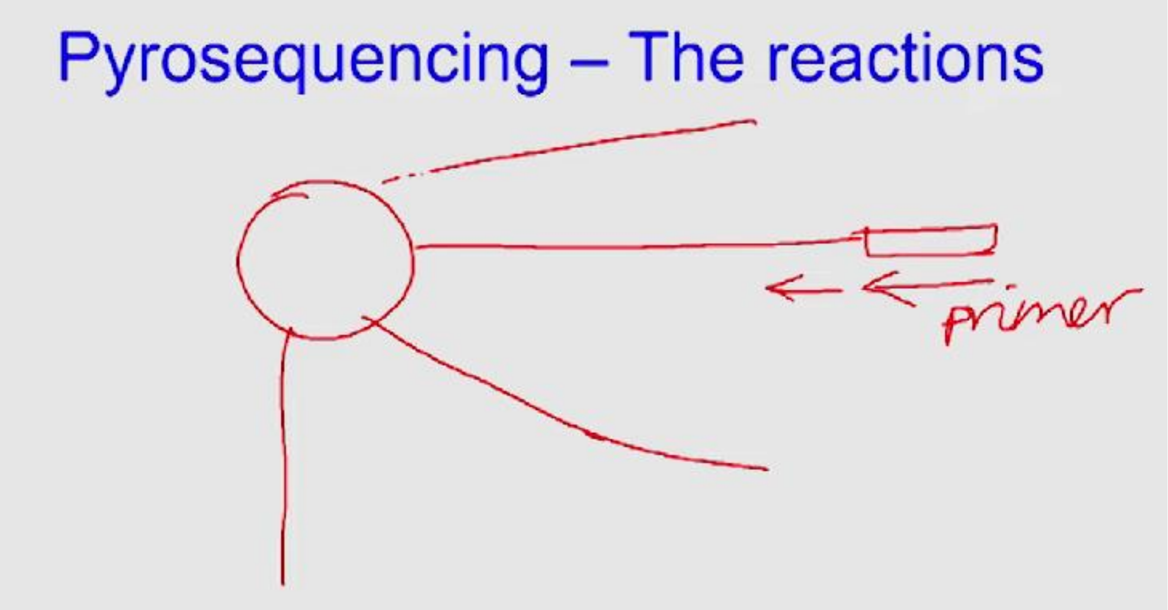
Now, once this is done, what you can do is synthesize it. So, you can have this extension right from here through DNA polymers. So, you can synthesize the complementary strand here, ok, and you can have this on bead amplification, ok. So, the idea behind bead amplification is that you are generating multiple copies of the same molecule. So, you are generating these copies in the bead from just one or two molecules, okay? Now, this bead amplification leads to the generation of many DNA molecules per bead, which will be used for sequencing. Now, this bead amplification happens in emulsion oil emulsion, right? So, each of these beads is separated from one another, right? So, you have one DNA molecule right and one bead, and there is amplification of that molecule within that bead right. So, only that type of molecule will be amplified within the emulsion, okay? So, you can now generate these different fragments with as many beads as you want. You can have billions of beads, and each of them will contain a unique fragment. So, how do we ensure there is this clonal amplification? As I mentioned, we start with one molecule, and we amplify that molecule on the bead. So, you have one data molecule per bead and per droplet, or this emulsion, and this is optimized through trial and error, but sometimes it happens that you cannot get one molecule; you might get two, three, or more, and in that case there would be some sort of signal mixing during the synthesis process, as we will see in a moment, and that signal has to be filtered right. So, we cannot use that data where you see this signal mixing, okay? So, the sequencing happens on picotiter plates.

So, these are plates like this, ok? So, here you have these wells, and there are 100,000 wells, right? So, more, and each well is designed in such a way that you have one bead per well, and this bead contains all these DNA molecules, which are clonal copies. So, copies are generated from just one molecule, right? So, you can have these beads across these 100,000 wells, and this is where the sequencing will happen.



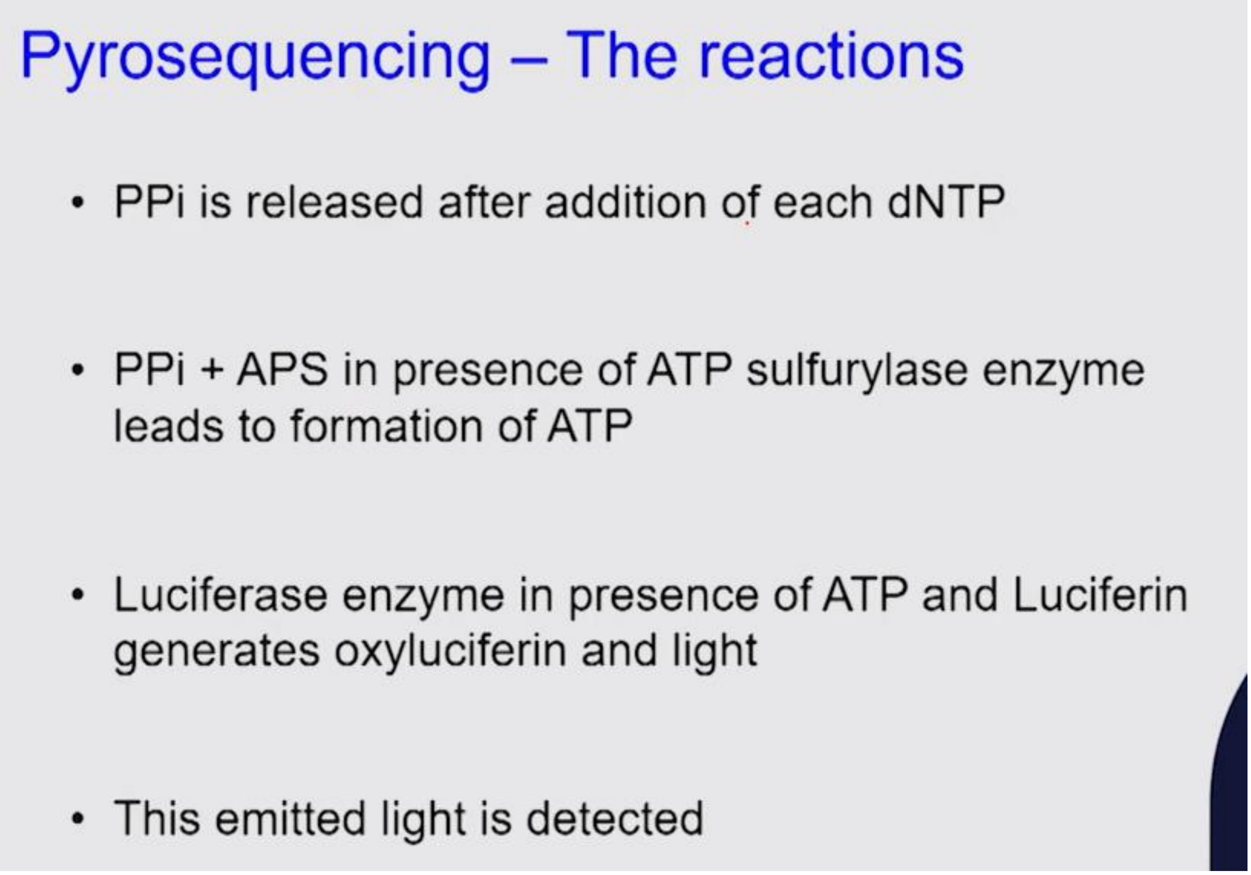
DNA sequencing in Picotiter plates

Now, you see how this whole process is parallelized, right? So, how do you actually make it high throughput? Right, you are sequencing 100,000 molecules at once instead of just one or two. So, as I said, with hundreds of thousands of wells on the plate, you can have more than one bead per well, and parallel sequencing happens in these wells, ok? So, what actually happens? What are the reactions that happen inside these wells? So, you have this bead with this molecule ok, and as of this, this is after the emulsion PCR, right? So, you have generated multiple copies of the same molecule, and we have these adapters here also, right? So, you can design primers that will bind here during the sequencing process and synthesize the complementary strand. So, for that, we need DNA polymerase and dNTPs.



Pyrosequencing – The reactions

So, what kind of dNTPs are these? These are normal dNTPs; there is no modification, ah, just they will come and bind based on the complementary base here, right? So, which base is present in this other strand? Based on that, they will be incorporated in this growing strand here, ok? So, in this process, there is a release of pyrophosphate after the addition of each dNTP. So, any dNTP, dATP, dTTP, dCTP, or dHTP. So, any type of dNTP will release pyrophosphate if there is an addition or incorporation of that base in the synthesizing strand in the growing strand. So, this pyrophosphate in the presence of APS right and the enzyme ATP sulphurylase leads to the formation of ATP ok. So, again, these APS and ATP sulphurylases are present in these wells; they are supplied with the ah media or the reagents there. In addition, you have luciferase enzyme in these wells, and in the presence of ATP and luciferin, these generate oxyluciferin and light ok. So, you can imagine this reaction now this you can visualize this.

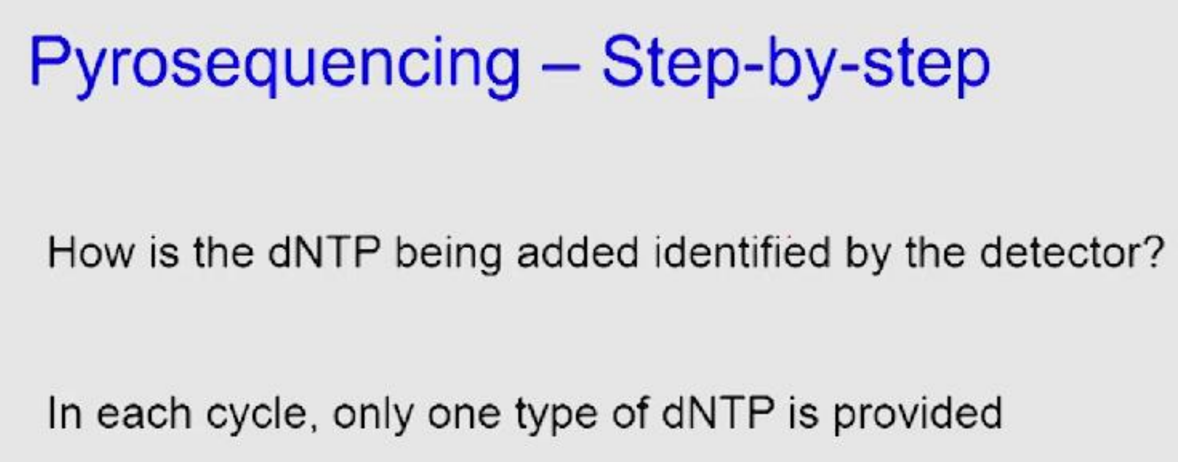


Pyrosequencing – The reactions

- PPi is released after addition of each dNTP
- PPi + APS in presence of ATP sulfurylase enzyme leads to formation of ATP
- Luciferase enzyme in presence of ATP and Luciferin generates oxyluciferin and light
- This emitted light is detected

So, you have the addition of one base in the growing strand right next to the complementary strand that is being synthesized by DNA polymerase. There is an addition of A; immediately there is some sort of chain reaction; right there is a pyrophosphate release that leads to the formation of NTP that is converted to oxyl luciferin and light by the luciferase enzyme. This is a very fast process, but there is a light signal that could be detected by an optical detector. Now, this is the principle right, and the detection happens by the light signal that is emitted because of all these reactions that are happening inside these wells, ok, and you have a detector that exists, ok.

So, step by step, let us look at this. So, one of the questions that you might have is: how do you actually identify which NTP is being added, whether it is A, G, T or C? So, what we do is actually add these dNTPs one by one. Okay, we do not add all 4 together; we start with one, then wash away that one, then add the next one, wash away that one, then add the next one, and so on. So, this happens one after another, ok? So, in each cycle, we provide only one type of dNTP.

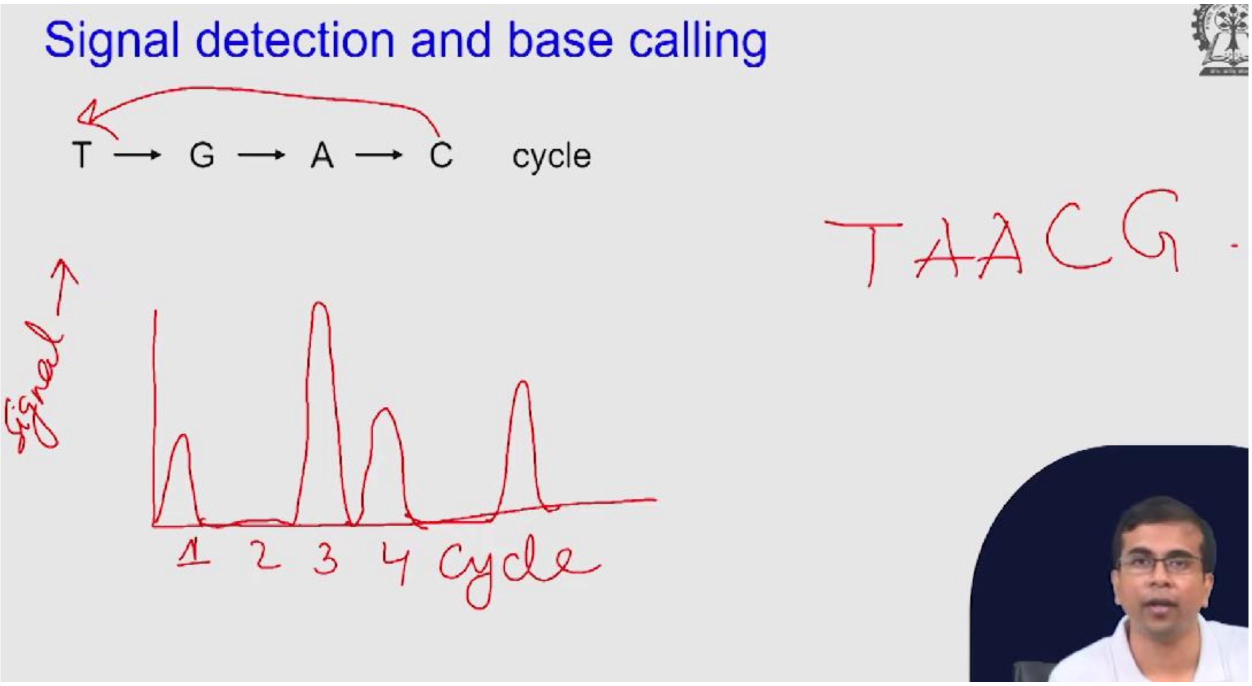


Pyrosequencing – Step-by-step

How is the dNTP being added identified by the detector?

In each cycle, only one type of dNTP is provided

So, these are called cycles. So, you have we are adding one type of dNTP. Let us say we provide A, we see if there is any signal, and then we wash away all this A, and then we add G. Let us say we again wait for a signal. If there is any signal or not, and then we wash away G, then we continue this process with C, T, etcetera, and then we start again with G. So, this is a continuous process that happens for many cycles, ok? So, let us imagine. Let us look at this to see what it looks like. So, how does the base call actually happen? So, this is your cycle number right? So, we have cycle 1 2 3 4 etcetera, and this is your signal intensity. The y axis is the signal, and we are following this cycle. Let us say T G A C. So, we first provide T, and we see whether there is any signal. Yes, there is some signal. So, the base is T right, we have identified the first base, next we give G right, we wash away T, we give G, is there any signal, no, there is no signal ok? So, there is no G base here; after T, the next base is A. is there any signal? Yeah, there is quite a bit of signal, and this signal probably comes from 2 A's. So, you can now imagine that if you add one base, you get one unit of light. Okay, let us say if you add two bases simultaneously, one after another, you have two complementary bases, and that will give you, let us say, two bases added.



Signal detection and base calling

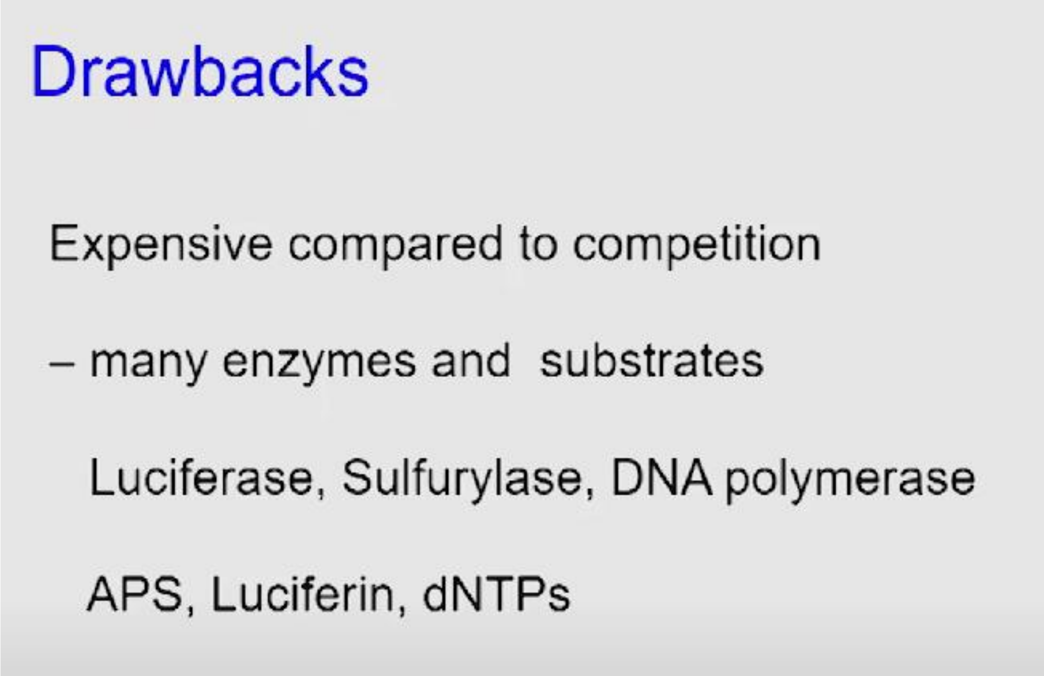
T → G → A → C cycle

So, what will happen? You will get 2 pyrophosphate. You get everything happening in double right. So, we should give you twice the signal intensity right? So, twice the intensity of light that you would get if you had just one base, ok? So, that is what we see right now. You can now calibrate the detector, saying, "Okay, if you see this much signal, this means you have an addition of two; if you see an even higher signal, there is an addition of three bases and so on."So, let us continue the cycle right next one. Let us say we see a signal here, ok, and we say ok, there is C, ok, and then we go back right, we go back to T again, and we see whether there is signal no.

So, no T for G. Yes, let us say there is a signal. So, G, and so on. So, we continue this process, right? You can run this for 100 cycles, 150 cycles, or as many cycles as you want, or before you start losing accuracy. So, this is the principle of how this sequencing method works. So, you now understand that there is this chain reaction that happens because of the release of pyrophosphate during the addition of base that actually translates into light, for which we can then measure this light intensity or the signal intensity, and through which we can determine the base that is being added. Now, here are the sequences that were actually present and were available when 4/5/4 sequencing was done.

So, these are again slightly different chemistries, slightly different processes, and slightly improved ones. There are also differences in throughput, depending on how many sequences you can get in one experiment. So, there was this difference because sometimes you need fewer samples or have fewer samples, and in some cases, you have a very high number of samples. So, depending on your needs, you would use this platform, ok? So, these are their names, and you could use them if you wanted to do 4 5 4 sequencing. What are the advantages of 4 5 4? So, this was the first method, as I mentioned. So, it has had tremendous popularity, right? So, there were many research work that has been done using this technology. The read length was about 500 to 600 base pairs and up to 1 kb in some cases, and this was longer than completion at that time. So, let us say, on average, it was around 500 to 600 base pairs and had reasonably high accuracy. So, about 99 percent accuracy.

So, that is quite good, right? So, you had only 1 percent error in the data, but it also had some drawbacks. So, we will talk about the drawbacks. So, it was actually expensive compared to the competition that actually came later on. So, this competition came from other technologies that we will discuss later. And why this was expensive is because it utilized many enzymes and many substrates for the sequencing reaction. So, as you have seen, we have DNA polymerase, but in addition, we have sulfurylylase, luciferase, and then different substrates like FPS, luciferin, and dNTPs. All these reagents are quite costly. And on top of that, in every cycle of reaction, we are washing away these enzymes and reagents. So, which means we need a lot more than that is actually required if we do this process continuously. So, we are losing a lot of these molecules and enzymes. So, that made this process very expensive compared to the competition that we will talk about a bit later, ok? And you had these washing steps, as I just mentioned, so that means we utilized a lot of these molecules.



Drawbacks

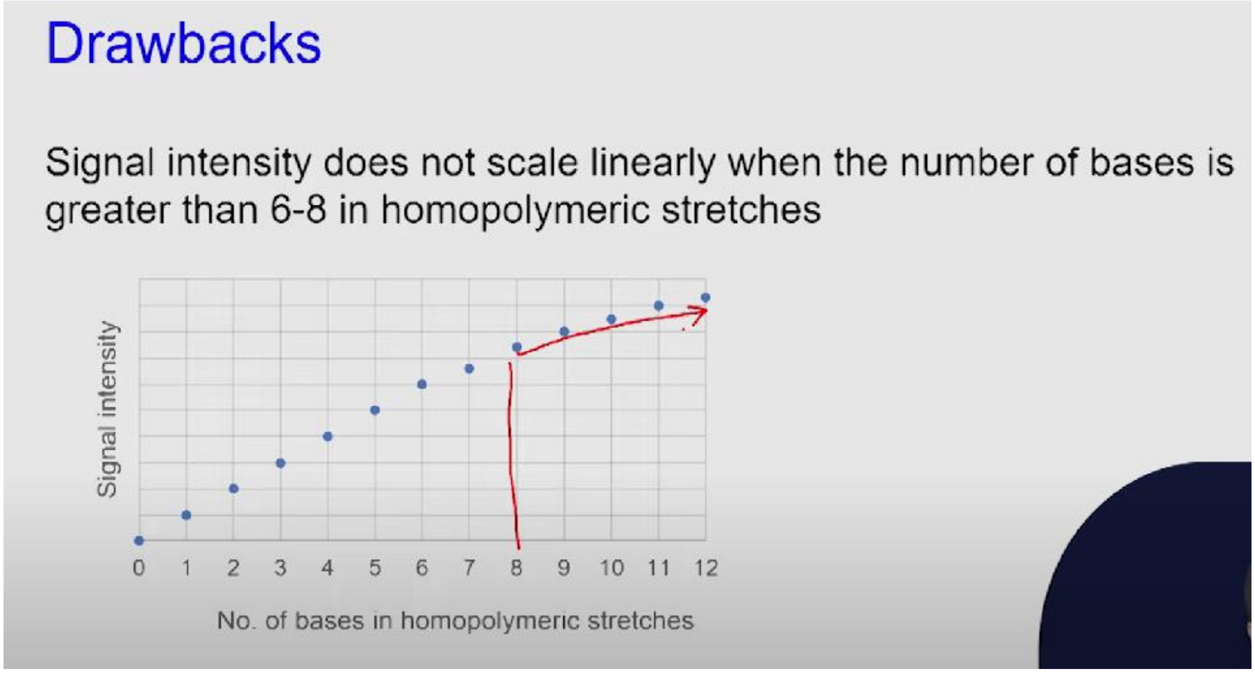
Expensive compared to competition

– many enzymes and substrates

Luciferase, Sulfurylase, DNA polymerase

APS, Luciferin, dNTPs

There was a bigger problem with 454: the signal intensity was not linearly proportional when the number of bases was greater than 6 to 8 in homopolymeric stretches. So, there are regions in the genome where you see a single base repeated multiple times. So, this is quite common across the human genome as well as other genomes. So, one base, like, for example, a is repeated, let us say 10 times 20 times right.



Drawbacks

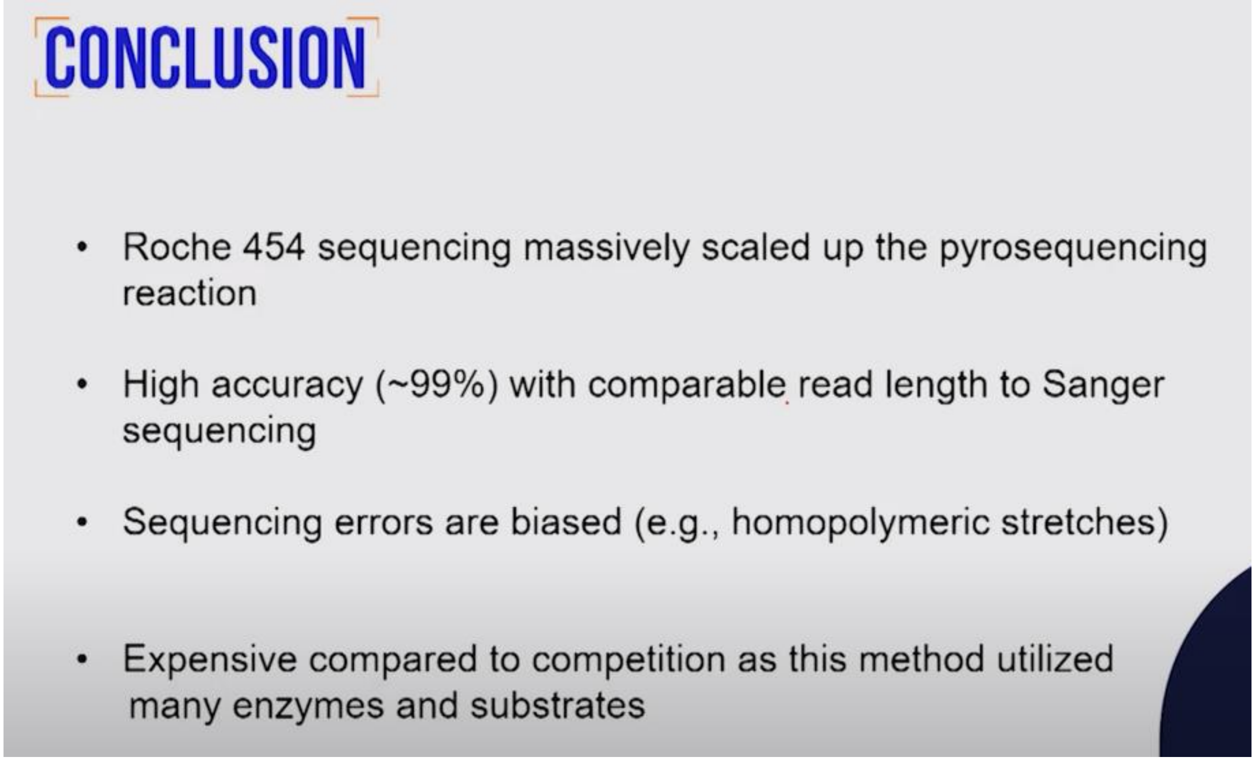
Signal intensity does not scale linearly when the number of bases is greater than 6-8 in homopolymeric stretches

So, you have these variations across the genomes; in some places, it may be 5 times, in some places, 8 times, in some places, 12 times, in some places, 20 times, and so on. These regions are called homopolymeric stretches. Now, what happens in 454? As you have seen, the signal intensity goes up. So, if you have the addition of just one base you have some signal intensity; if you have two bases, the signal intensity should be double; if you have three bases, the signal intensity should be triple; and so on. So, it should go up linearly. What we have seen and what the researchers have seen is that this signal was not proportional after a certain length. So, after you have reached 6 or 8 bases beyond that, the signal is not proportional to the number of bases anymore. So, which means you could not determine the exact length of the homopolymeric stretch? So, whether it is actually 10 or 11 or 12 or 14 that there is to be confusion, ok. So, which means this method gave a lot of sequencing errors around these homopolymeric stretches, especially insertions and deletions.

So, we will talk about the different types of variants that you can determine. So, what it means is that we got more errors in the sequencing data in these regions. So, here is a kind of schematic of the whole thing, right? As you can see, the x axis is the number of bases in homopolymeric stresses, as I just explained. And then you have the y axis, you have the signal intensity, and in the initial part up to 6/8, the signal intensity is linearly proportional to the number of bases in the homopolymeric stretch, but then after that, it slowly starts to saturate right.

So, as you go along, right as you go beyond these 12 bases, what you will see is that the sequencer will not be able to resolve this number of bases very accurately. So, whether it is 14/15 or actually 18 right, this will not be resolved. So, you will start seeing a lot more sequencing errors in these regions, ok? So, these are the references that you have utilized to summarize. So, 454 sequencing massively scaled up the pyrosequencing reaction we have, as we have just discussed right here. This scale-up process actually involves this one bead amplification.

So, we start with DNA fragments, and then we do something called emulsion PCR in this oil emulsion. So, what you have is something like micro reactors, and each micro reactor contains one bead and one DNA molecule coming from the genomic DNA, and this genomic DNA is then amplified because there are certain complementary sequences to the adapters already present on these beads. So, this leads to massive amplification, which we call emulsion PCR within these emulsions. So, at the end, what we get is something like a bead with multiple copies of the same DNA fragment, and we have millions of such beads that we have prepared in parallel from the genome fragment. Now, once we have these beads, what we do is put them on picotritor plates. So, there are multiple wells in 100,000; you can have more, and then each well will hold one bead. So, once this bead goes in, this will actually lead to the sequencing reaction. Now, of course, in this process, there will be some wells where you do not have any beads, and some wells may end up with two beads, or more. This can be optimized for a trial-and-error process. Once you have the beads in the wells, we can now start the sequencing reaction, and the sequencing reaction, as we have just discussed, depends on the pyrosequencing principle. So, pyrosequencing has been used before for DNA sequencing, but it was never scaled up to such a level. So, what happens in pyrosequencing is that you have sequencing by synthesis which means you are synthesizing the complementary strand of the DNA molecule that is present on the bead, and using DNA polymerase, the NTPs, etcetera, you start synthesizing the complementary strand. Now, once there is an addition of a base, whether it is A, T, G, or C there is a release of pyrophosphate, which then, through a series of reactions, is converted to a light signal. So, it does not depend on the base itself; you end up with a light signal, whichever the base is, and this light signal can be detected, which indicates there is an addition of a base to the complementary strand that is growing, and this light signal then acts as the detection right across these wells. So, if you can if you can visualize in your mind across the picotiter plate the detector is seeing this signals of light as the sequencing reaction is going on is seeing this signal ah, signals of light across these 100,000 wells in in the picotiter plate ok and from that is doing the base calls ok and we have talked about ah, how these ah, base calls actually are dependent on the number of a base additions this should scale linearly the signal should scale linearly with the number of bases that are added ok. This method is highly accurate as as has been observed it is about 19 percent with comparable read length to Sanger sequencing Sanger sequencing gave about 900 base pair to 1 kB read length this method gave us about 500 to 600 base pair at most and then what we have seen is that of course, this is not perfect there are some drawbacks.



CONCLUSION

- Roche 454 sequencing massively scaled up the pyrosequencing reaction
- High accuracy (~99%) with comparable read length to Sanger sequencing
- Sequencing errors are biased (e.g., homopolymeric stretches)
- Expensive compared to competition as this method utilized many enzymes and substrates

So, one of the drawbacks is that it is very expensive because it utilizes so many enzymes and so many reagents, and it cannot actually compete with the newer technologies that are much more affordable compared to this 454. So, this is the reason why 454 was closed down in 2013; it could not compete with the other methods. But we have also seen that these sequencing errors are biased. So, we have seen that there are homopolymeric stretches where this method does not perform that well and you get a higher error rate and we have just mentioned this is ah, more expensive compared to competition, and that is why ah, we ah, this method was stopped ah, but nevertheless, it had a huge impact; it actually allowed researchers to do a lot of fantastic work, and even today we will see a lot of data in databases around, and in case you ah, analyze this data, you need to keep these drawbacks in mind especially for example, if you are doing this looking at homopolymeric stretches, you need to be very careful with the analysis. Thank you.