

Interactomics: Basics and Applications
Prof. Sanjeeva Srivastava
Dr. Rahul Solanki
Department of Biosciences and Bioengineering
Indian Institute of Technology, Bombay

Lecture – 50
Next-Generation Sequencing Technology- MiSeq System

In the last few lectures we are discussing about one of the revolutionary technology Next Generation Sequencing Technology where you are given the concepts starting from the basic to the latest advancement happening in this area. And we are very fortunate to have some of the very leading industries and their application scientist directly sharing their experience with you.

So, in this light in today's lecture series, today we have Mr. Rahul Solanki a senior field application scientist from Premas Life Sciences who will talk to us about Illumina next generation sequencing workflows. So, let me welcome Mr. Rahul Solanki for his lecture today.

Let us say about protein sequencing what exactly the sequencing is.

Student: (Refer Time: 01:18).

To know the sequence of amino acids.

Student: Amino acids.

If I speak about DNA sequencing.

Student: (Refer Time: 01:18).

Right. So, why we need it? Why it is needed?

Student: To identify the genes.

So, why you need to identify gene?

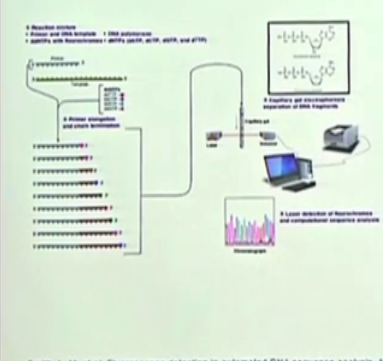
Student: To study the mutations and (Refer Time: 01:30).

Study mutations. So, ultimately all the functions which are related to the protein ultimately those are coded by gene, right it is correct.

So, start, I will start with the Sanger though we do not have the Sanger sequencer with us, but just to clarify the basic things I will start with the Sanger. So, there is the principle what you do is let us say you have a sequence, you flow what you flow into the chamber? You have four kind of NTPs and one contains ddNTPs right.

(Refer Slide Time: 02:03)

Automated Capillary Sequencing: 1986



The diagram illustrates the automated capillary sequencing process. It starts with a DNA template being sequenced using a mixture of labeled and unlabeled ddNTPs. The resulting DNA fragments of varying lengths are separated by capillary electrophoresis. A laser detector then identifies the fluorescently labeled bases, and a computer system processes the data to produce a chromatogram.

- Mixture of unlabeled and P-labeled chain-terminating ddNTPs
- Incorporation of terminating ddNTP prevents further extension
- Varying length fluorescent DNA fragments generated
- DNA fragments separated by capillary electrophoresis
- Fluorescence detected by laser
- Each DNA fragment with the same length is detected at the same time

Smith, L. M. et al. Fluorescence detection in automated DNA sequence analysis. *Nature* 321, 674-679 (1986).
Image: http://en.wikipedia.org/wiki/Sanger_sequencing

So, can you see the stretch of sequences the labelled? So, those are ddNTPs which are labelled with a fluorophore. So, let us say the first one is A added right. So, that polymerase cannot extended, why? Because the base added is ddNTP, it lacks a OH group ok.

Then so on you will be having a different kind of fragments over here, then after that you run it through a capillary. So, capillary is nothing but a gel right. So, what we do is after the sequencing is done we flow all the sequences through a capillary and you can see there is a detector, this all fluorophore is labelled with a fluorophore all the sequences last base. So, which will be the first base will be flowing through a gel that first one? The upper one, the smallest one ok. So, through agarose gel that top most sequence will be?

Student: (Refer Time: 02:56).

Yeah, it will pass away first ok. So, let us say it contains A, so you will get a signal for A. Similarly, let us say the second base is G in the second strand, so G will be called right and so on there is the basic of Sanger sequencing. So, what is the maximum a base pair which we you can sequence using Sanger good result quality data?

Student: 1000 base pairs.

1000 base pair, correct. So, see Sanger is still a matter of choice when you wish to sequence the gene and ideal length of a gene and human is almost around a kb correct?

Student: Yes.

But why we need NGS? Now the point is why the NGS is required?

Student: Throughput.

Great. So, the right answer is throughput, right. Any idea about how long the first human genome took to get it sequenced?

Student: (Refer Time: 03:54) 10 years; 10 years.

It took 10 years and almost how much million dollar was invested?

Student: Billion dollars.

A billion dollars. How many mam? Almost 3 billion dollars.

Student: (Refer Time: 04:07).

Was invested for a first human genome to be sequence. So, what is the length of human genome?

Student: 3 into 10 raise to 9, 3 into 10 raise (Refer Time: 04:18).

Great, it is a common question of CSIR NET. I mean who appeared in the NET, I guess [FL]. I hope you are writing. So, its 3 billion base pair, 3 into 10 to the power 9. So, you just think if you wish to sequence a human genome how many reactions of Sanger you need to carry out? The human length of human genome is?

Student: 3.

3 billion base pair 3 into 10 to the power of 9 and the maximum length of gene you can sequences 1000 base pair. So, 3 billion divided by?

Student: 1000.

1000 how much it is?

Student: (Refer Time: 04:47).

How many reactions you need to carry out?

Student: (Refer Time: 04:51) 3 million.

How many?

Student: (Refer Time: 04:52) 3 into 10 to the power.

3 million; so, 3 million Sanger reaction you need to carry out to complete the human genome sequence and the sequence you will be getting will be covered it one x coverage. So, even if you carry out one 3 million reactions also you will be covering your bases how many times?

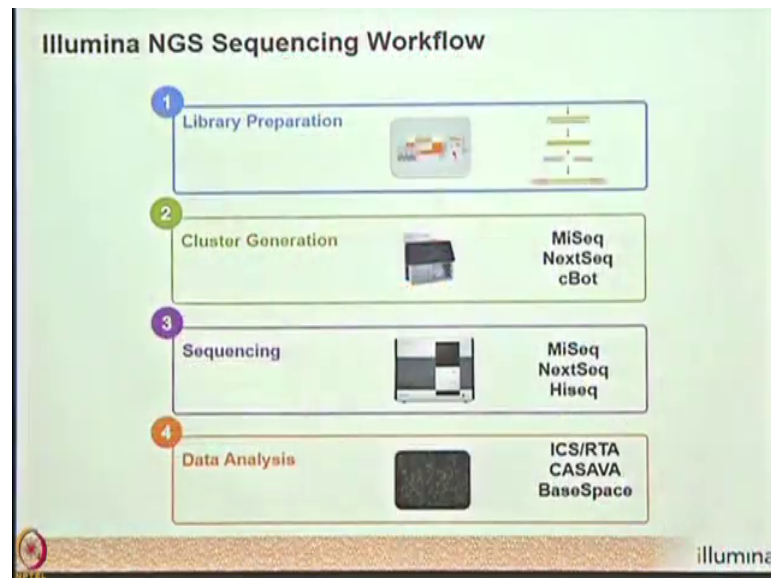
Student: One sir.

One time. So, the coverage will be?

Student: 1.

1 x. So, if you wish to sequence a human genome do you wish to invest 10 years more? No. So, with the help of NGS technologies, you can sequence, 50 human genomes in a span of 48 hours. How many? 50 human genome in a span of 48 hours.

(Refer Slide Time: 05:37)



So, the key idea remains the same, the principle remains the same what we do is let us say you have a this human genome let us say. So, what we do initially is we fragment this human genome into very small fragments and just like cloning, we do not need cloning at all in the case of NGS, what we do here is.

So, I will be covering all the parts one by one, the very first step involved in all the sequencer are library preparation which is followed by cluster generation, then comes is the sequencing and final is the data analysis part ok. So, first is the library preparation its common for all the platforms though the sequencer are different, but the concept remains the same. So, what is library first of all?

Student: (Refer Time: 06:22) collection of things.

In terms of sequencing if I say what is the library?

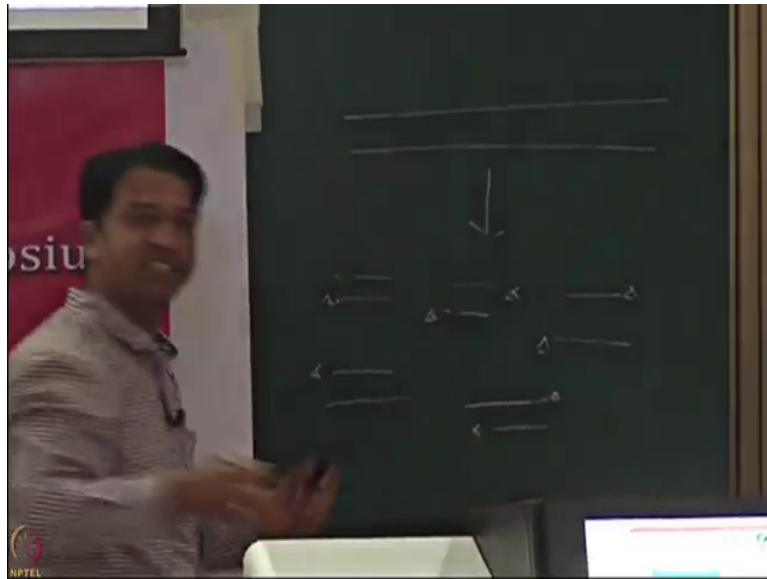
Student: (Refer Time: 06:31).

In Sanger what do you do? I you clone it.

Student: (Refer Time: 06:34).

That vector contains the known sequences, already it contains the primer for what the sequencers are know. So, using the primers you can sequence your gene. In the case of NGS what we do? We chop down the entire human genome into smaller fragments, but those are unknown right correct, what you will do? Where is a chalk? Again the same question I am asking.

(Refer Slide Time: 07:03)



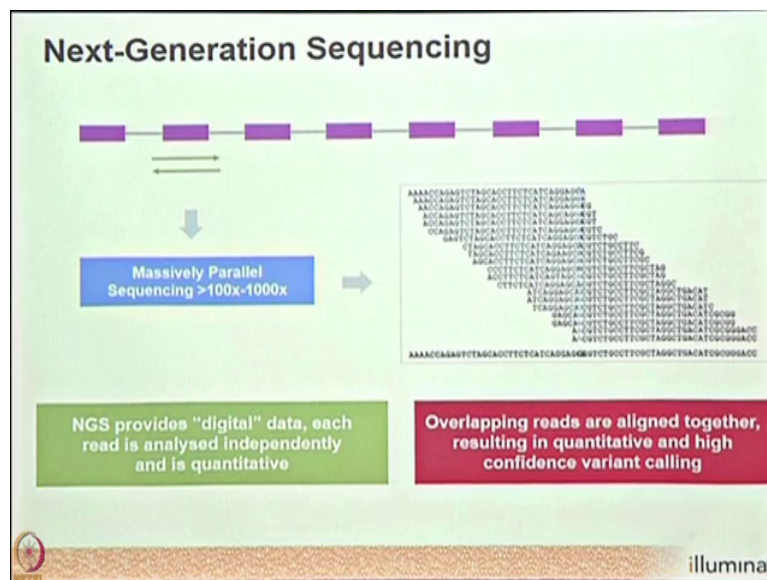
Let us say this is a human genome. In the Sanger what we do any genome, human genome big genes what we do? Initially the first step is we fragment it into various small fragments, let us say a human genome. So, 600 base pair will chop it down into various millions of fragments, but these are unknown. So, to get them sequenced what we need?

Student: Adapter ligation.

Right, you remember now. So, what we will do is initially we will use a polymerase same. So, the tendency of polymerase is what it does is if you have a these many fragments. So, initially we converted into blunt ended all the ends in to blunt ended you will add a polymerase. So, the tendency of polymerase is it will add A to all the ends right you know and then we have adapter which contains a T over there, there will be a simple ligation step.

So, those fragments will be converted into the fragment which will be ligated with the adapters.

(Refer Slide Time: 08:01)



So, in library preparation the ultimate goal is the first of all we start with the DNA fragment, what we do is we fragment it into various lane of fragments let say 600 base pair, then we repair the ends. Now it is converted into blunt ended and finally, we ligate the adapters. So, what are adapters?

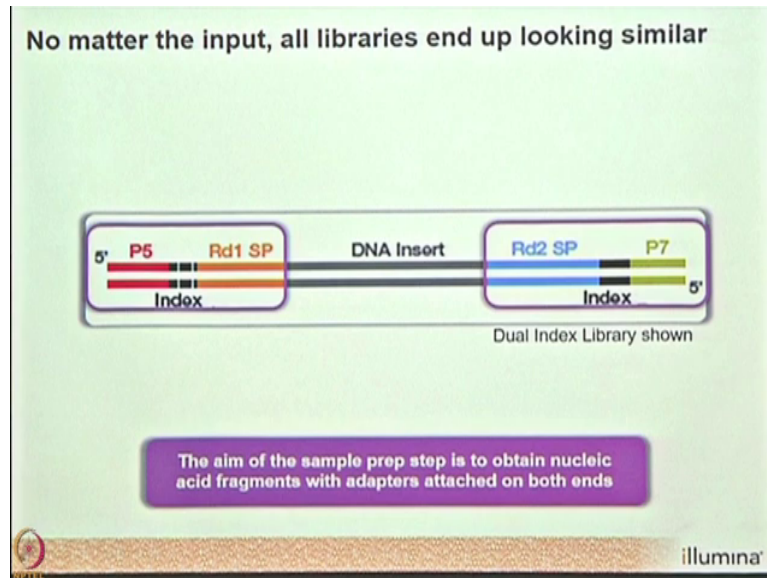
Student: (Refer Time: 08:16).

Adapters are the switch of.

Student: (Refer Time: 08:19) adapt.

DNA which for which the sequences are known already known right.

(Refer Slide Time: 08:25)



So, this is how the library looks like finally, it will contain a DNA insert which you wish to sequence in the middle or now you can see there are the 2 regions Rd1, Rd2 and these are essentially the adapters. We have 2 kinds of adapters; P5 and P7, alright, is it clear.

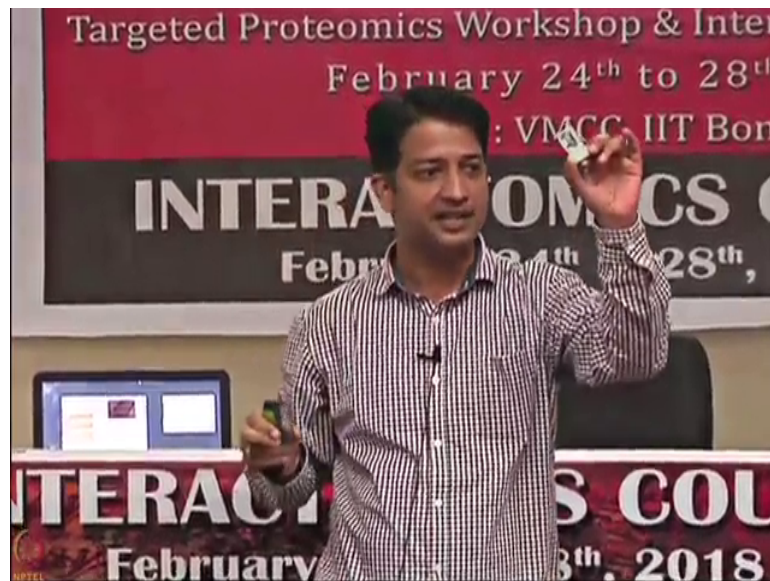
So, I will tell you what is the function of all. This is the sequence unknown sequence all right, here you can see there are two regions here our read 1 sequencing primer will bind. So, this is called as read 1 sequencing primer binding site; flying by both the ends are indices. What are indices? What are index? These are barcodes. So, what is the function of barcode?

Student: To identify the samples.

So, in NGS what you do is you can sequence as I said using NovaSeq you can sequence 50 genomes all together. So, ultimately with the help of barcode you can identify which sequences belong to which of the samples, am I clear ok. So, this is how the library looks like, the ultimate goal of library preparation is to ligate the adapters at both the ends of all the DNA fragments, clear great.

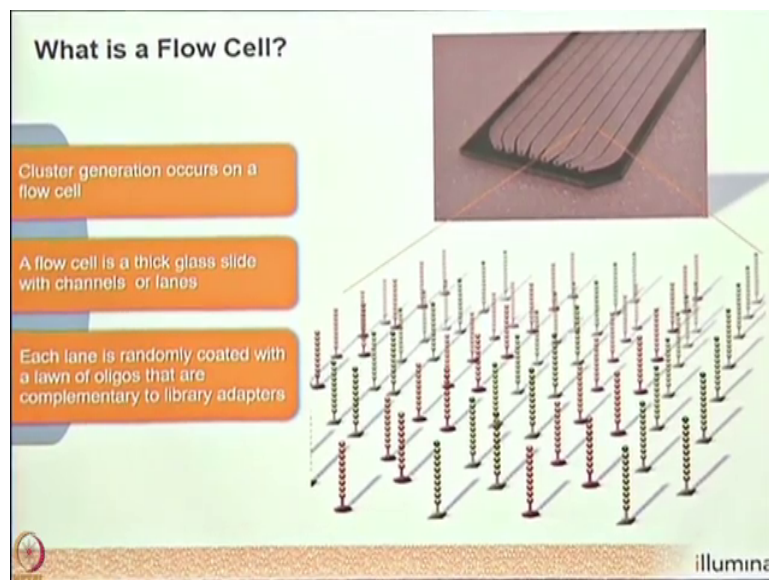
So, there is all about library preparation, coming on to the cluster generation. So, now, what is cluster generation? So, essentially all the process happens on the flow cell right; flow cell is not nothing but a glass slide.

(Refer Slide Time: 09:58)



You can see there are two flow cells: this is Hi C flow cell which is meant for sequencing human genome and there is Mi C flow cell for targeted sequencing especially.

(Refer Slide Time: 10:08)



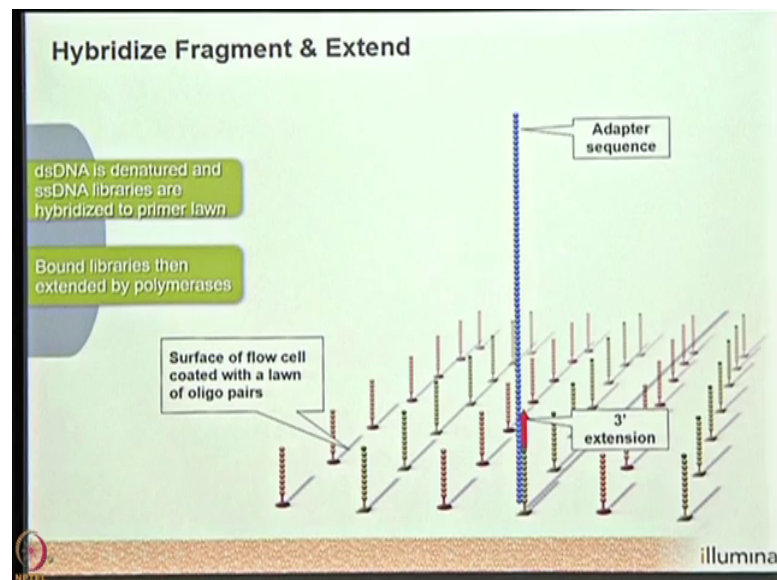
So, it has 8 lanes, it contains 8 lanes and it has a common lane. Why we call it as a flow cell because the sequencing happens with the help of flow of reagent into this flow cell ok. So, the flow cell contains the lawn of oligonucleotides. So, essentially we have two kinds of oligos; one oligo will be complementary to P5 region and one will be complementary to P7.

Student: P7.

P7 region. So, before loading into the flow cell what we do is we denature this double stranded DNA fragment into single stranded DNA fragment. So, that it can go and bind to the

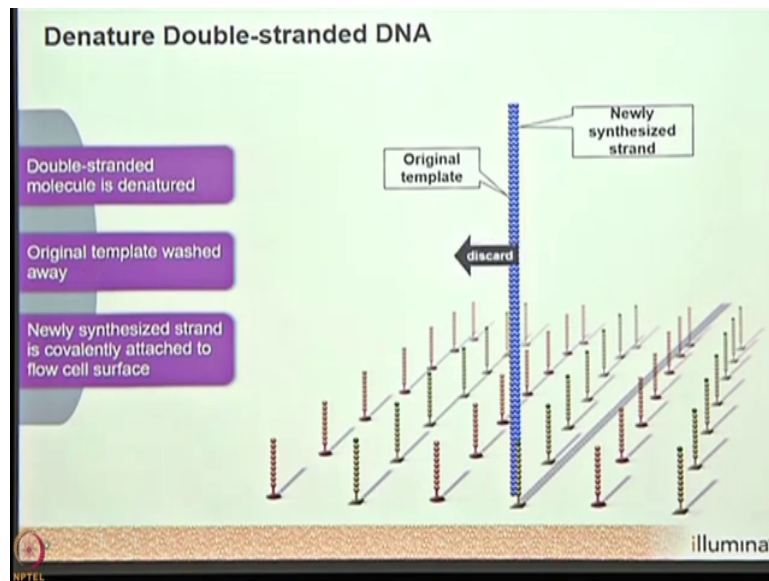
surface of flow cell right. So, once you have denatured it, these are complementary to those P5 and P7 that I have already denatured.

(Refer Slide Time: 10:48)

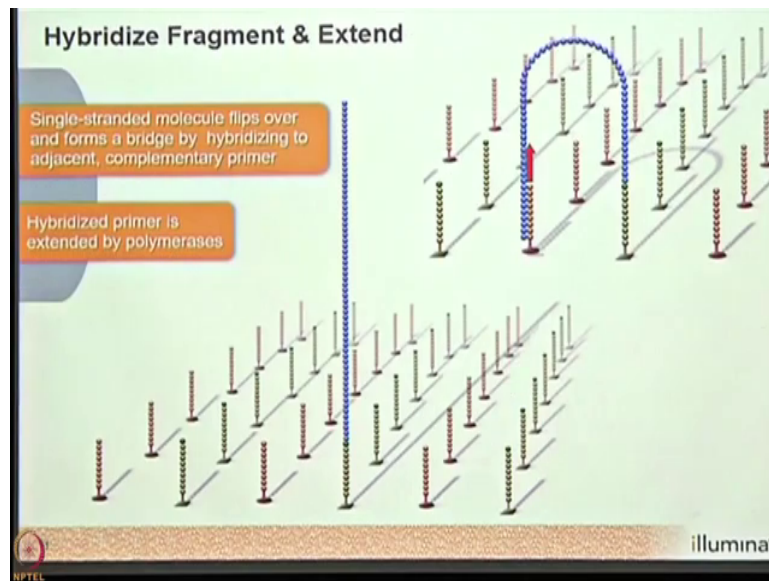


So, one of the fragments will go and bind to the flow cell right and this region you can see how it binds here any guesses? With the help of hydrogen bonding simply because these regions are complementary you can see this is covalently bound to the surface of flow cell. And this is because this region is complementary it will go and bind with the help of hydrogen bonding and we extend these ends with the help of DNA polymerase.

(Refer Slide Time: 11:17)



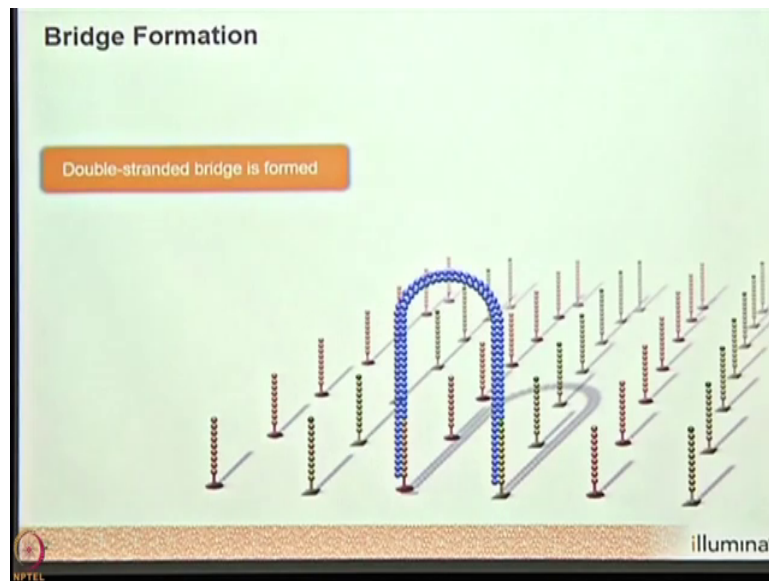
(Refer Slide Time: 11:43)



So, you can see you will be getting a structure like this. Then what we do is we retain this original fragment because simply it is bound covalently to the surface of flow cell and because this strand is bounded by hydrogen bonding these are weak bond. So, we denature it and we wash away this original template, we retain this one clear.

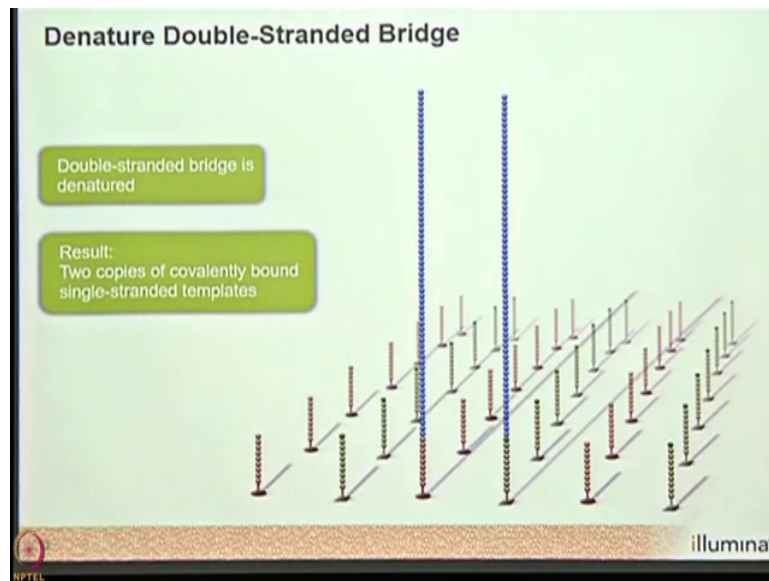
Then what you can see is this is left over here and since this end also this one this end and this end it is also complimentary to this oligo, it will flip over here again we will add DNA polymerase you will see a bridge kind of structure ok.

(Refer Slide Time: 11:59)



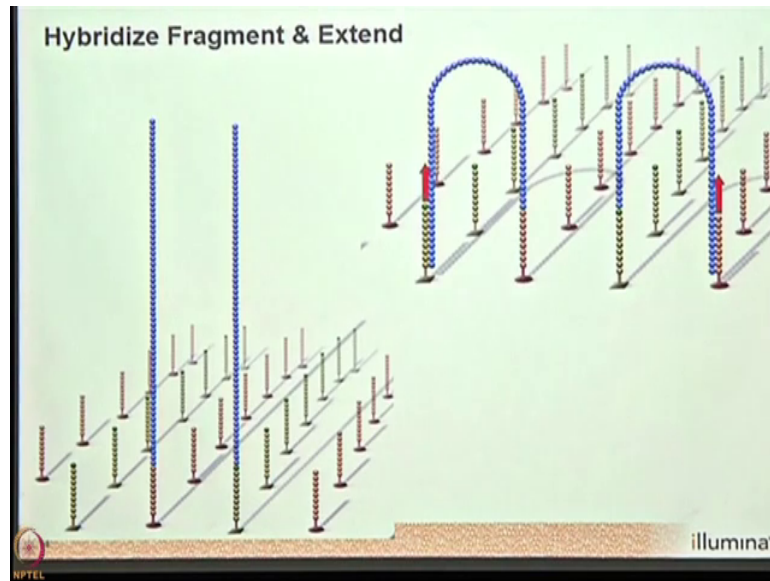
Then again we will denature it and both this strand because they are bound to the surface covalently.

(Refer Slide Time: 12:07)

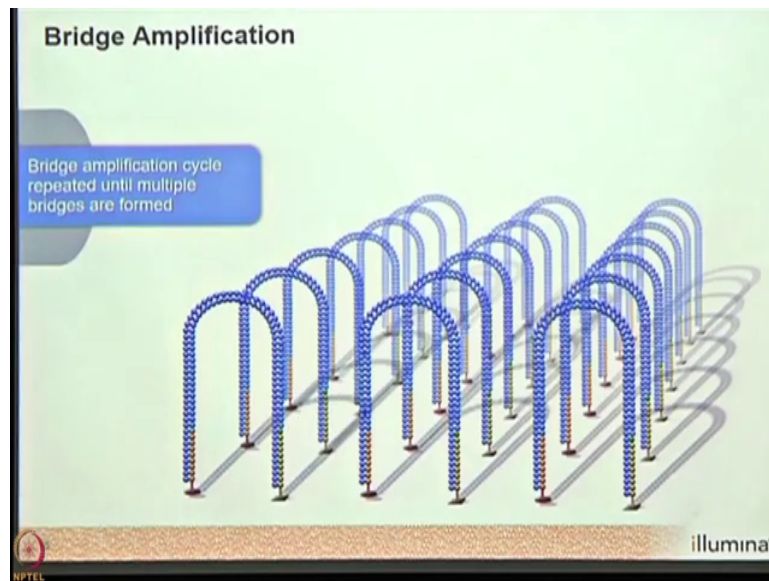


You will see a two strand over here and this process will be repeated for millions of time.

(Refer Slide Time: 12:11)

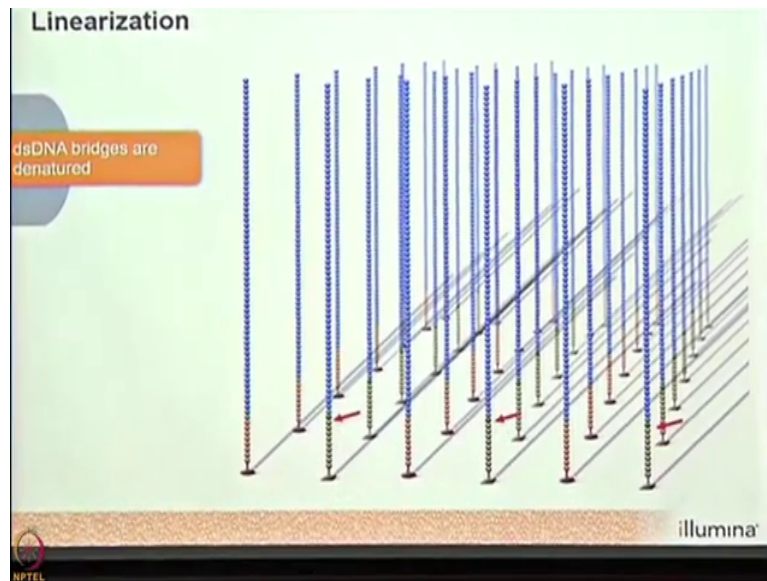


(Refer Slide Time: 12:14)



And ultimately on the surface of flow cell what you can see here as will be having a millions of fragment on the surface of flow cell ok.

(Refer Slide Time: 12:22)



So, again we denature all the fragments. So, essentially you can see it contains two strands which strand it contains? Forward and reverse.

Student: (Refer Time: 12:33).

All both the strand it contains. So, what we do is initially we cleave the reverse strand and we carry out the sequencing for forward strand ok. So, now, on the flow cell you can see which strand is retained?

Student: Forward strand.

All right. So, again there is a question, what should I do to prevent again, if I do not do a one step what will happen what is going to happen?

Again it will flip and bind to this surface. What should I do to prevent again it should not bind to this primer, what should I do?

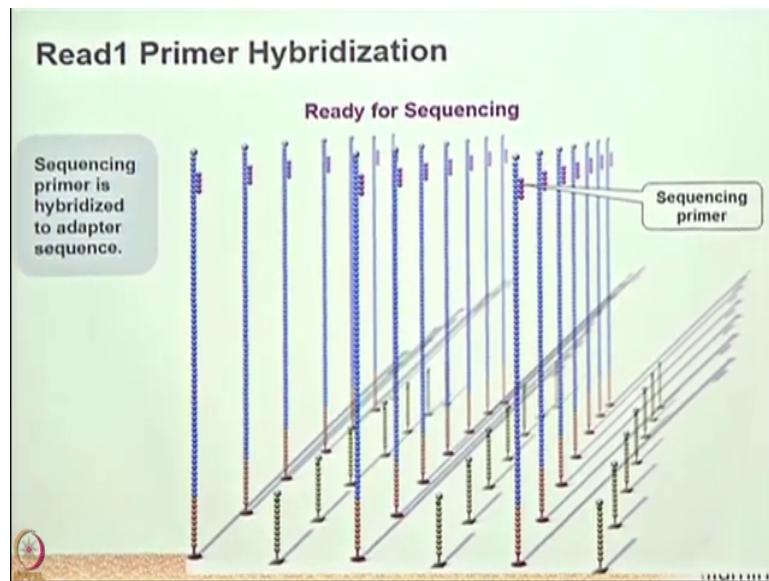
Student: Block the free ends (Refer Time: 13:11).

Good. So, what I will do next is I will block the all free ends. So, that it would not flip ok.

Student: Yes sir.

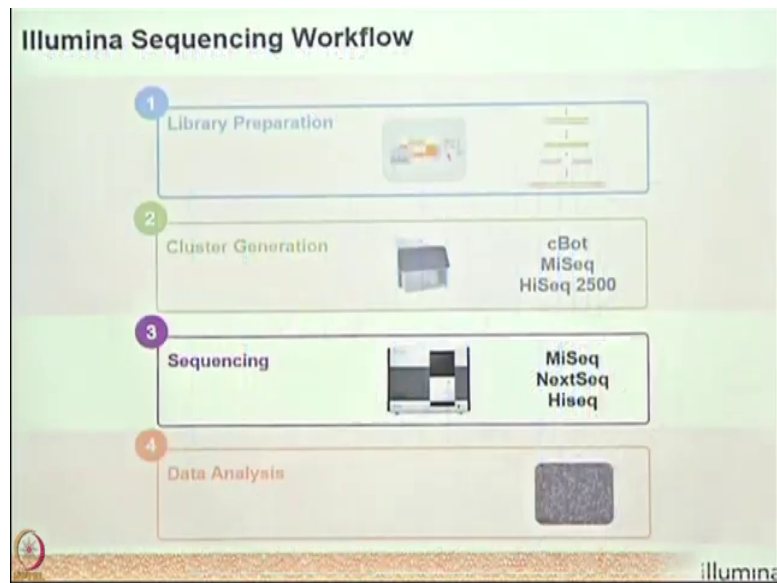
And now you are done with your sequencing part, I mean cluster generation is over.

(Refer Slide Time: 13:27)



Now, you are set to go for sequencing. So, you remember in the library I have shown you a region Rd1 region. So, which is meant for hybridization of read 1 sequencing primer. So, on the flow cell now you flow read 1 sequencing primer. So, you can see that primer go and bind over there then we are set to go for the sequencing.

(Refer Slide Time: 13:51)




(Refer Slide Time: 13:52)

Sequencing By Synthesis (SBS)

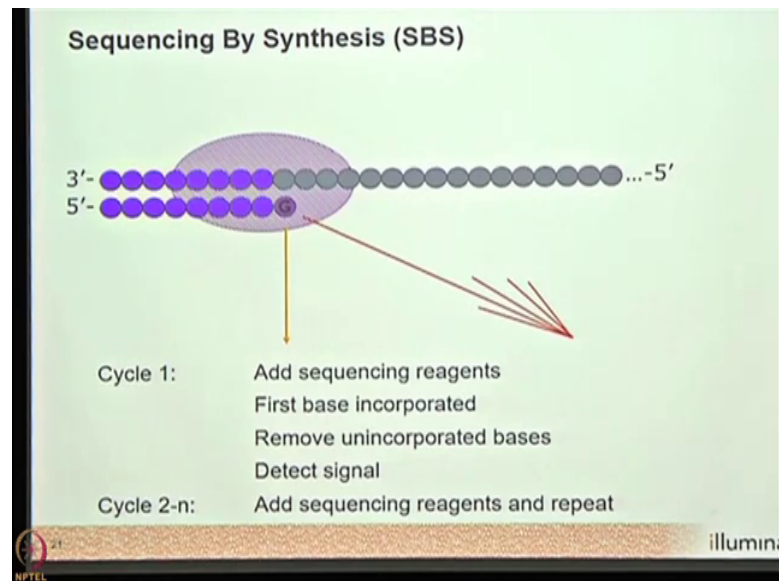
Cycle 1: Add sequencing reagents
 First base incorporated
 Remove unincorporated bases
 Detect signal

Cycle 2-n: Add sequencing reagents and repeat

 illumina

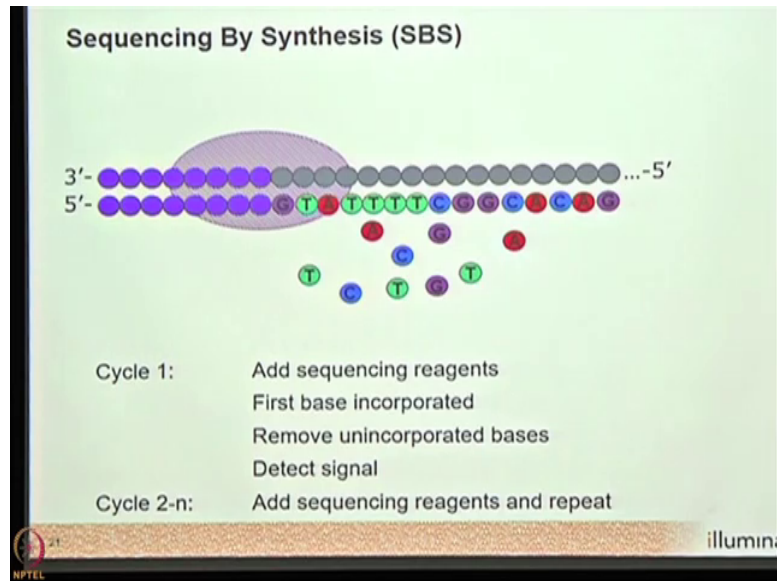
With Illumina what we do is we essentially mimic the Sanger sequencing just like natural phenomenon. So, we have four kind of bases. So, let us say this your strand we have added a primer this you can see all these bases are labelled with a fluorophore.

(Refer Slide Time: 14:13)



So, one at a time one base will go and bind to the sequence, it will be detected just like Sanger will wash away the unbound bases and then we are set to go for the next cycle.

(Refer Slide Time: 14:26)



So, there is direct detection of the bases. So, this let us say you are running a sample of 300 cycle. So, 300 base will go and bind, this is your 1 fragment ok. So, this is how if you are going for 300 cycle 300 bases will be added and we are done with the sequencing part.

(Refer Slide Time: 14:36)

Reversible Terminator Chemistry Advantages

- All 4 labeled nucleotides in 1 reaction - Mimicking the Natural Phenomenon
- No problems with homopolymer repeats
- Higher accuracy with most of the bases higher than Q30.
- Paired – End Sequencing

3'- ...-5'

5'- G T T T T C G C C G

illumina

So, what is the key difference here is we essentially mimic the natural phenomenon there is no change in phosphate or something. The best part is if you detect any by secondary methods like phosphate or something let us say. So, what will be the hurdle there? If 1 if 1 base is added let us say there is 1 call for A in a cycle. So, signal will be if 2 A are added the signal will be double, but if you have multiple A or multiple G over there. So, you cannot resolve the signal.

So, in the terms of NGS in the Sanger how you detect the quality of the NGS data? By looking at the peak you get a peak you remember your data if the peaks are sharp your call is good.

So, in the case of NGS we define the quality in the terms of Q values quality values. So, if I say Q 30 Illumina use Q 30 scores. So, a Q 30 score means is 1 error in 1000 base pair added. So, the accuracy rate is 99.9 percent ok.

If I say about Q 30, what is Q 30? So, we define our run as let us say you performed a run which contains a E coli genome. So, we will define 80 percent of the basis sequence were above Q 30 values, getting. So, the error rate in that 80 percent base calls were 99.9 percent not error rate accuracy rate the error rate was.

0.1 percent alright, am I clear. So, what exactly the patent sequencing is essentially you remember as mam said she initially we generate both fragments, if you wish to sequence only forward strand what you will do is you will cleave away the reverse strand and you will sequence the only forward strand. The parent sequencing what you do is you regenerate both this strand, in the second time what you do is you retain forward strand sorry you retain the reverse strand and you cleave away the forward strand in this manner you can sequence both the strand of stretch of DNA.

You have do one is forward one is reverse. So, initially what we do is we cleave away the reverse fragment; we retain we sequence the forward. During patent what we do is we regenerate we de block it remember. So, both the strand will be generated this time will cleave away the forward strand and we are going to sequence the reverse strand.

So, once we are done with read 1 ok. So, let us say my read lane was I have chosen 300 base pair read lane. So, there is your read 1 product and that this is your sequencing primer and from here how many bases were added? Let us say it was a 300 cycle run.

So, 300s bases will be added over here those are detected and then what we do is we denature this product, the product of read 1 sequencing primer will denature it and we are set to go for patent sequencing. What needs to be done is we will denature it again we will de block this end clear.

So, we will de block this end again it will flip because we have de block the end and again you can see multiple clusters will be generated, but this time what we are going to do is we are going to cleave away the forwards strand. Initially what we did?

Student: (Refer Time: 17:45).

We had cleaved away the?

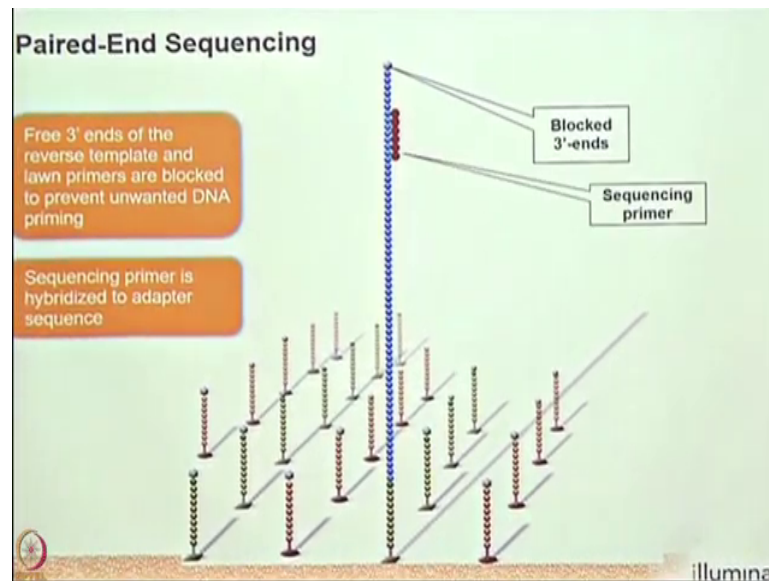
Student: Reverse strand.

And we have gone for the sequencing of?

Student: Forward strand.

Good ok. So, where we cleave away the original forward strand and this how we are ready for sequencing for the second. So, this time which primer will go and bind?

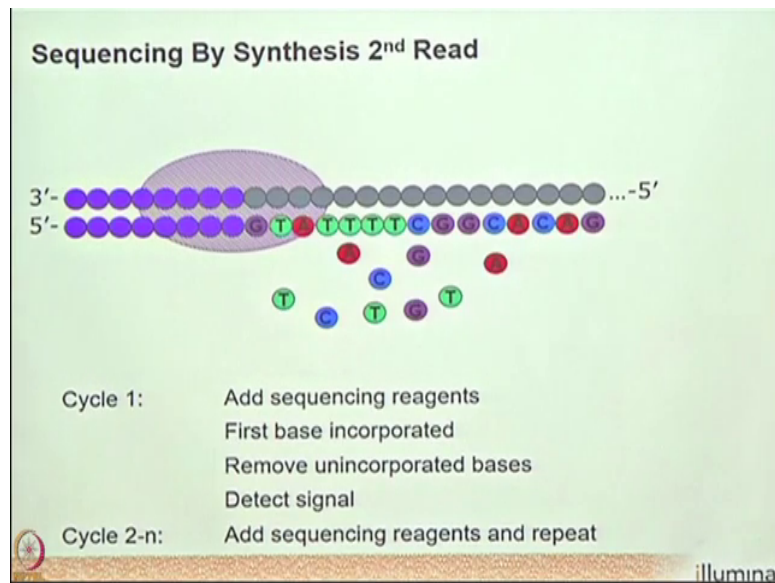
(Refer Slide Time: 18:01)



Student: Rc, Rd2.

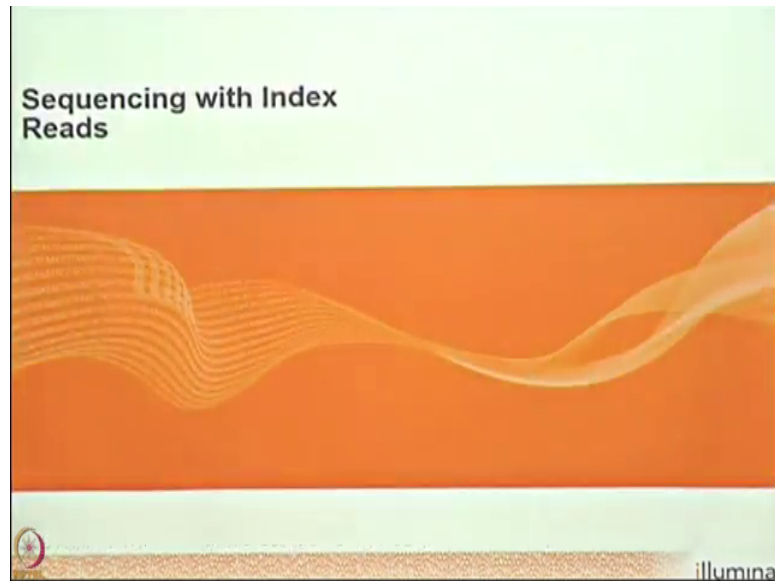
Rd2, you remember the structure of library there was a Rd2 region. So, there the read 1 sorry read 2 primer will go and bind and just like read 1, we are set for the sequencing right all the reagent will flow this is how you will sequence.

(Refer Slide Time: 18:14)



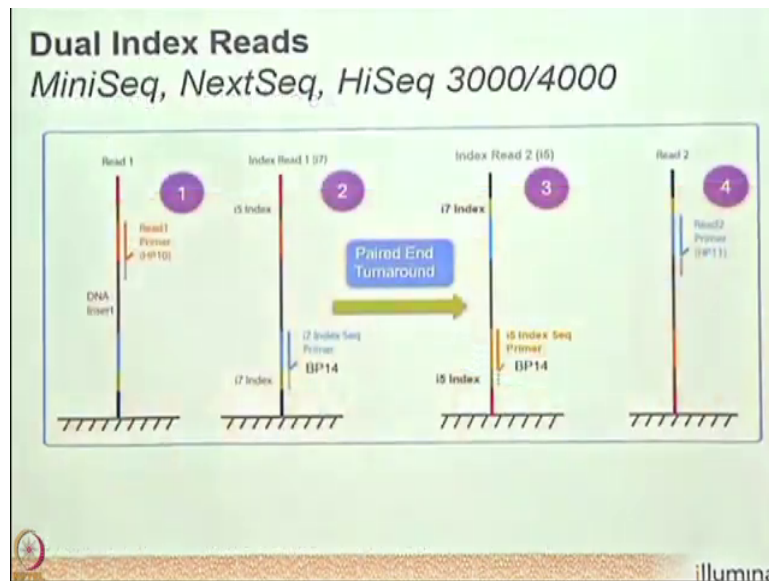
Just like the first one, but the difference is here you are sequencing the reverse strand this is all about the patent sequencing.

(Refer Slide Time: 18:21)



So, in all our platforms essentially what we do is we initially carry out the read 1 sequencing ok.

(Refer Slide Time: 18:25)



Then we go for indirect sequencing barcode sequencing which is essentially 6 to 8 base pairs long, then barcode 2 and then finally, read 2 ok. So, initially we sequence read 1. [FL] one question when I was in the Singapore during my training. So, one of candidate asked a question to trainer I mean he was also not able to answer.

So, one candidate asked me [FL] why you need separate read for indexes right for let us say if you have this DNA fragment and you can remember the structured. So, here is the indices. So, why you need this separate read you can directly read from here itself, why you need a separate read for that?

Student: (Refer Time: 19:17).

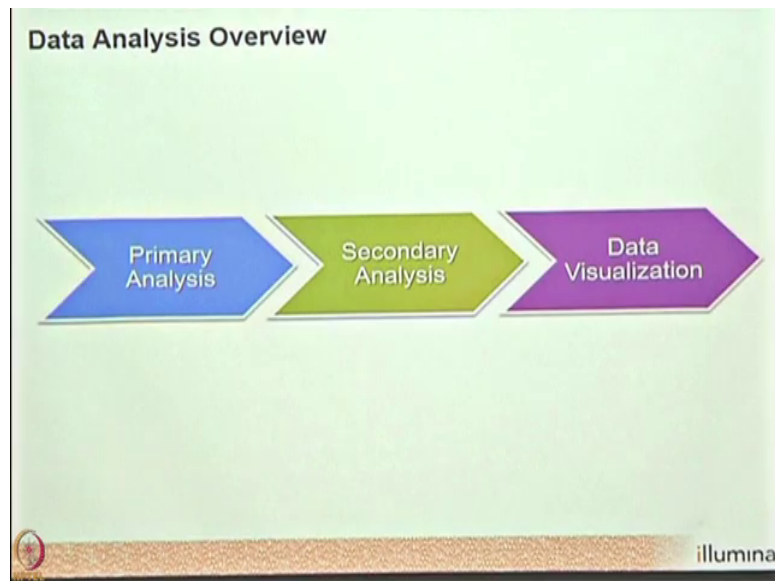
Makes sense? His question was logical getting, what I am telling is what we do is initially we sequence the read 1, then we go for read 2 sequencing sorry barcode 1 sequencing this barcode 2 sequencing and then you are read 2 finally, read 2.

Student: (Refer Time: 19:33).

Good. So, your catch is right. In Sanger if you see if you go beyond ideally speaking 6700 base pair your peaks will be not that sharp. So, as the read length will increase let us say I have a DNA instead of more than 600 base pair. So, as the read length increases, the quality may drop and for demultiplexing the samples from the pool I need very accurate data.

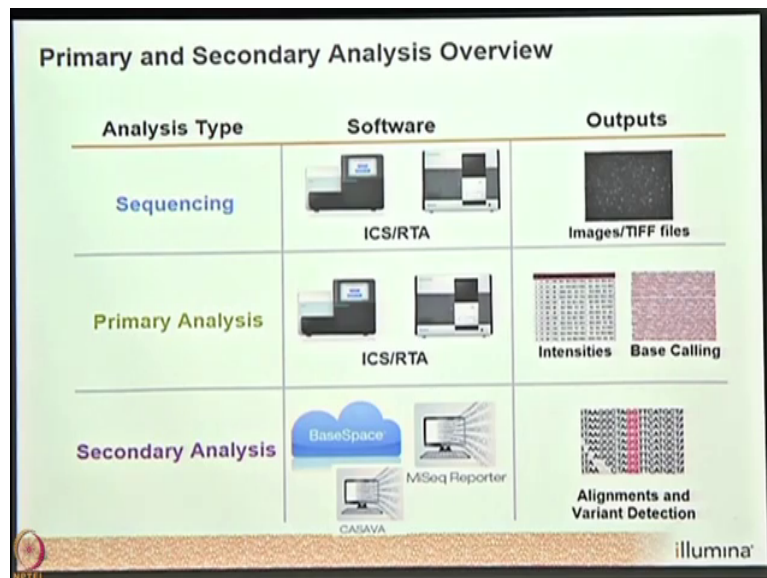
So, because the bar codes are only 8 base pair long. So, I should have very good data for demultiplexing the samples from the common pool correct. So, that is the reason why what we do is we sequence the read 1 primer sorry read 1 sequences, then barcode 1, barcode 2 and finally, we carry out what? Read 2 that is all there is the simple logic.

(Refer Slide Time: 20:28)



So, once we are done with the sequencing the final stage is the data analysis ok. So, there are three kinds of analysis; first one is the primary analysis. So, during primary analysis what happens is you know all the bases are tagged with a fluorophore, right. So, initially the camera will record the signals like this, it will be coloured though. So, red, green, yellow but these are all your bases called.

(Refer Slide Time: 21:51)



So, what, there is a software which is quite less real time analysis software. Once your bases are added those are detected in the real time right. So, once we are done with cycle one let us say, so you will get all the base calls file. So, during primary analysis it happens with the help of a software which is called as real time analysis software. So, it extract your intensities from the bases and convert those intensities into base calls. So, you will be getting dot BCL files on the board.

Once the dot BCL files are generated, the secondary analysis using my seek, again start on board itself. So, what it will do is it will generate those base calls files into fast queue files and finally, using this machine tertiary analysis also you can get the reports on board everything. So, which are the pathogenic mutations, which are the silent mutations every report you will get; what you need to understand is what is primary analysis.

During primary analysis the intensities of the base calls are extracted and you will get dot BCL files on the board during secondary analysis you will get all alignments and fast queue files right.

So, I will play a small video for you guys. So, I hope it will clarify everything, the same thing which I have shown. So, it will be shown sequentially.

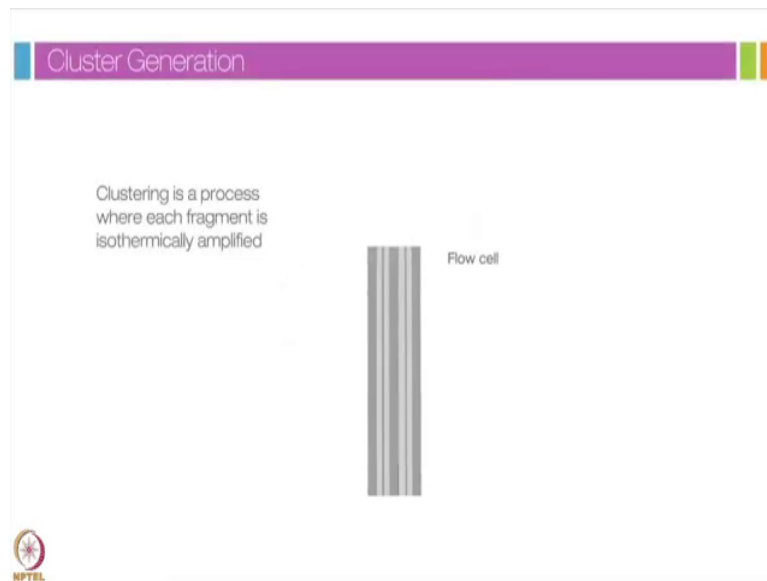
(Refer Slide Time: 22:14)



(Refer Slide Time: 22:17)

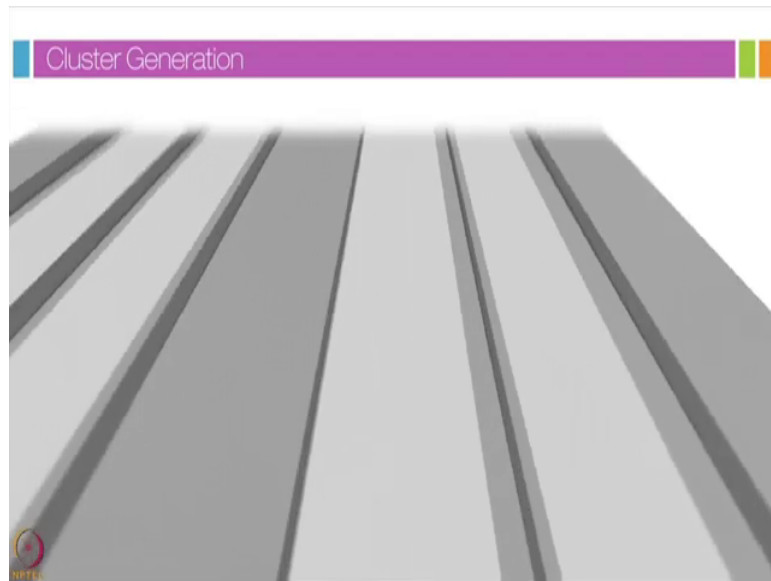


(Refer Slide Time: 22:20)



Clustering is a process wherein each fragment molecule is isothermally amplified. The flow cell is a glass slide with lanes.

(Refer Slide Time: 22:27)

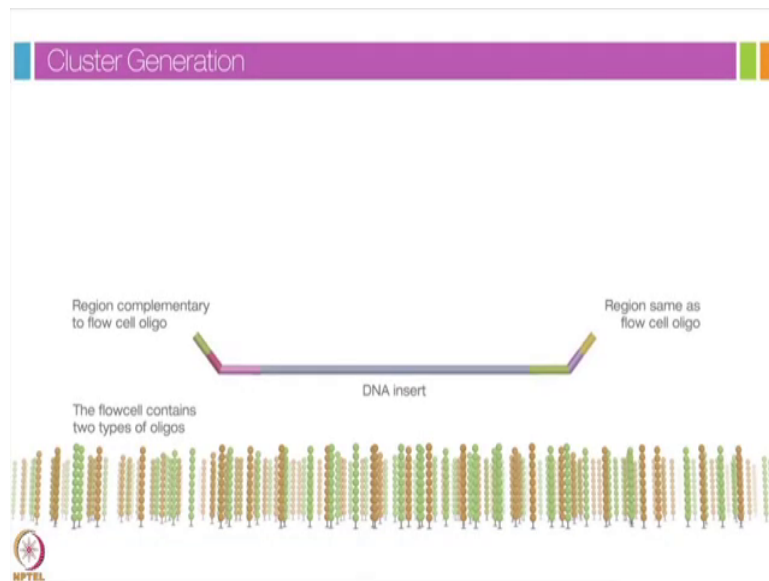


Each lane is a channel coated with a lawn composed of two types of oligos. Hybridization is enabled by the first of the two types of oligos on the surface.

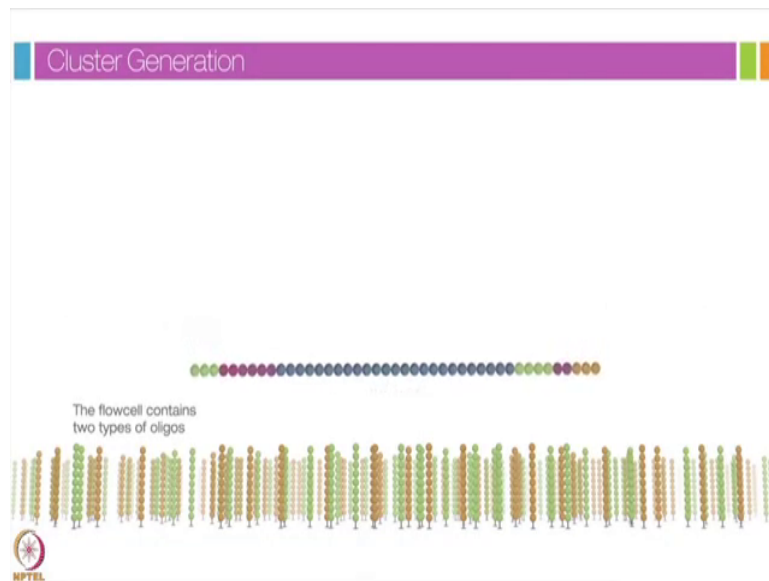
(Refer Slide Time: 22:30)



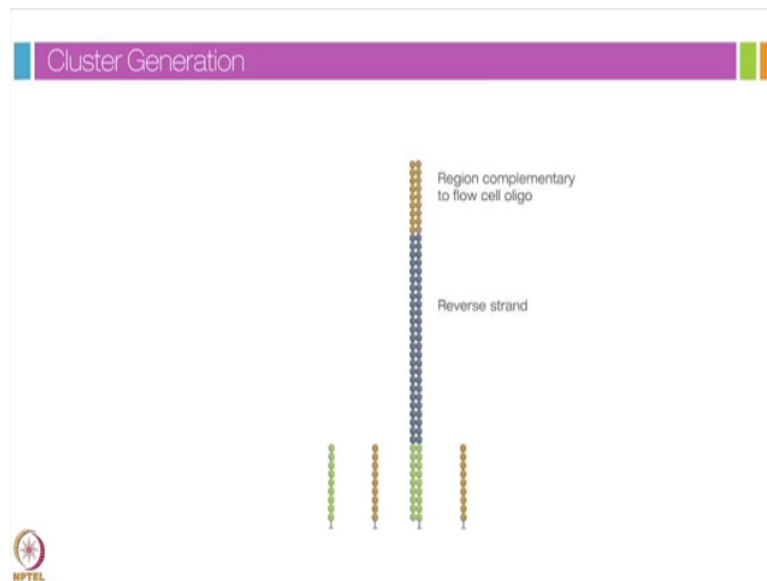
(Refer Slide Time: 22:34)



(Refer Slide Time: 22:38)

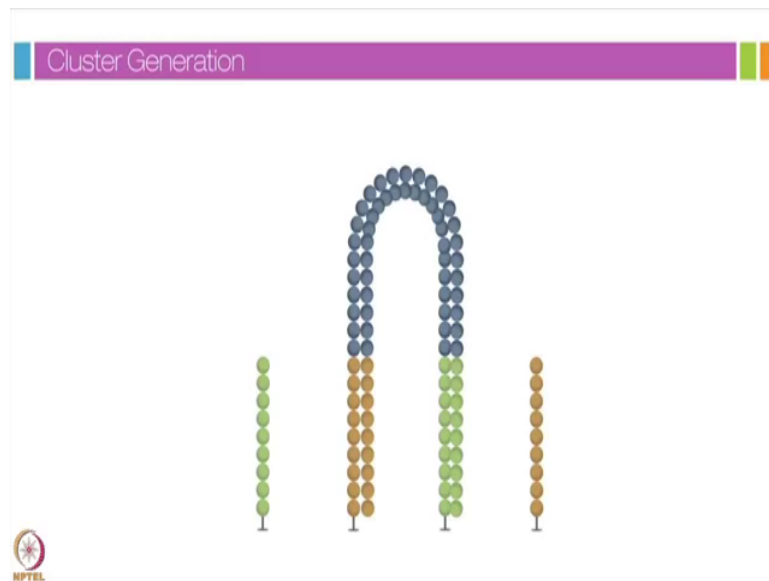


(Refer Slide Time: 22:39)



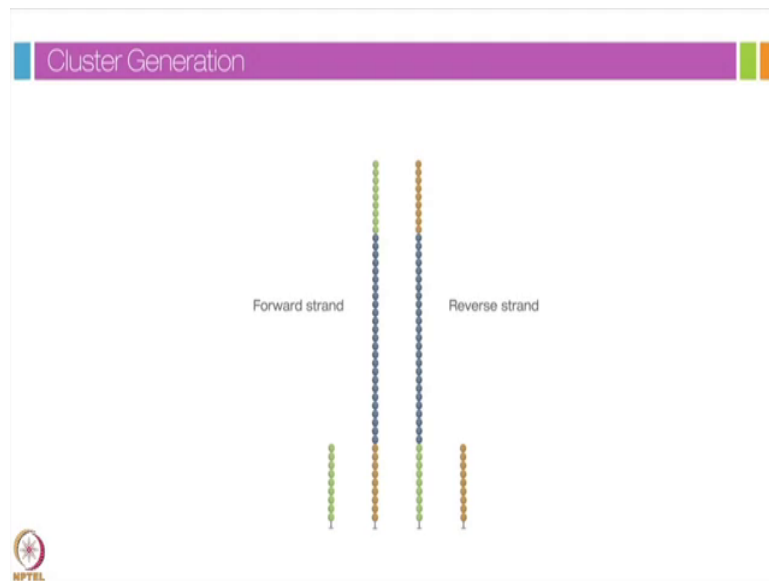
This oligo is complementary to the adapter region on one of the fragment strands. A polymerase creates a complement of the hybridized fragment. The double stranded molecule is denatured.

(Refer Slide Time: 22:51)



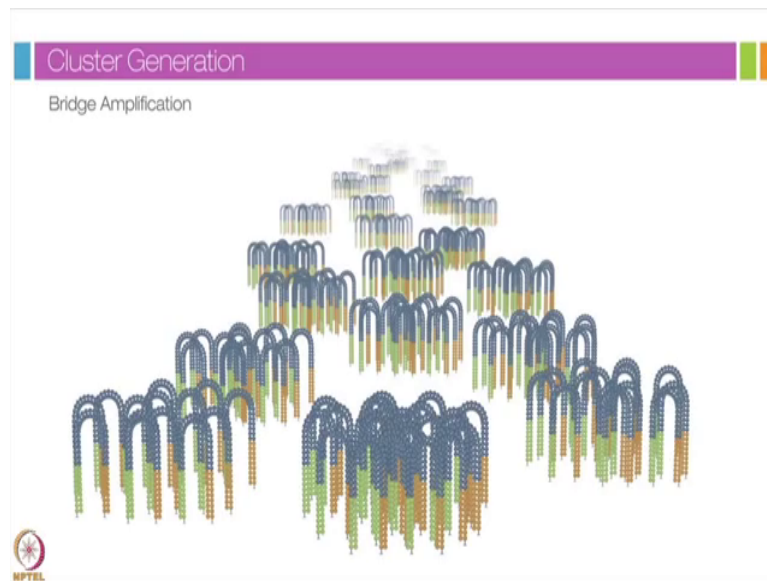
And the original template is washed away. The strands are clonally amplified through bridge amplification. In this process the strand folds over and the adapter region hybridizes to the second type of oligo on the flow cell. Polymerases generate the complementary strand forming a double stranded bridge, this bridge is denatured.

(Refer Slide Time: 23:13)



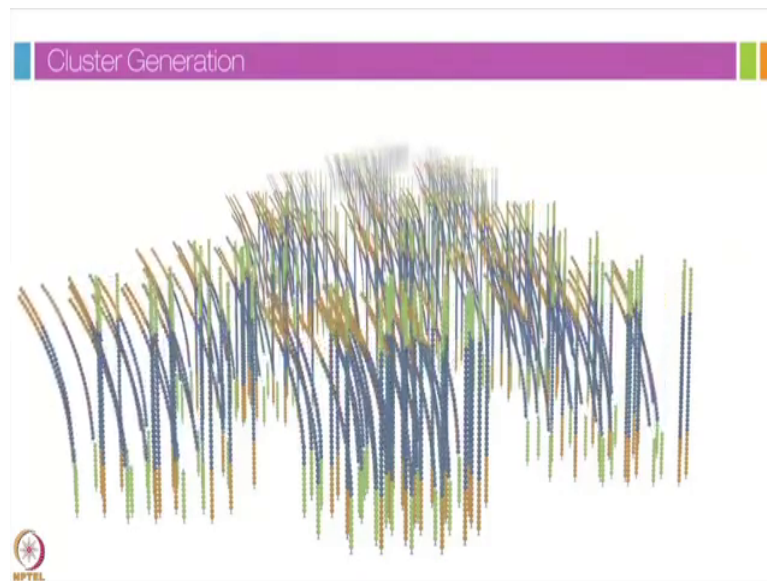
Resulting in 2 single stranded copies of the molecule that are tethered to the flow cell.

(Refer Slide Time: 23:19)



The process is then repeated over and over and occurs simultaneously for millions of clusters resulting in clonal amplification of all the fragments.

(Refer Slide Time: 22:30)

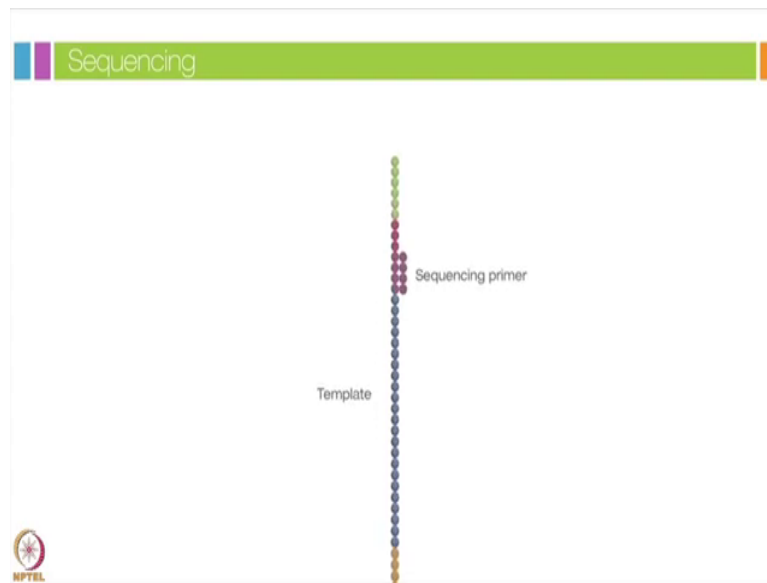


After bridge amplification the reverse strands are cleaved and washed off leaving only the forward strands. The 3 prime ends are blocked to prevent unwanted prime end.

(Refer Slide Time: 23:36)

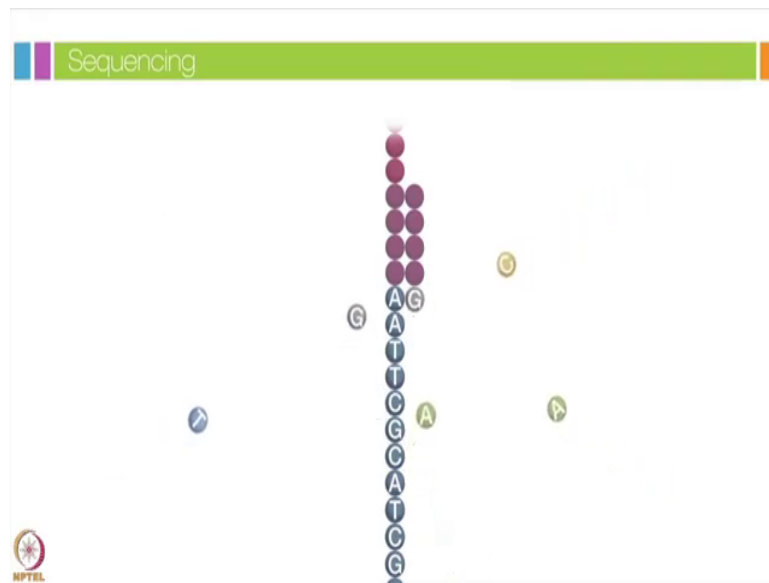


(Refer Slide Time: 23:43)



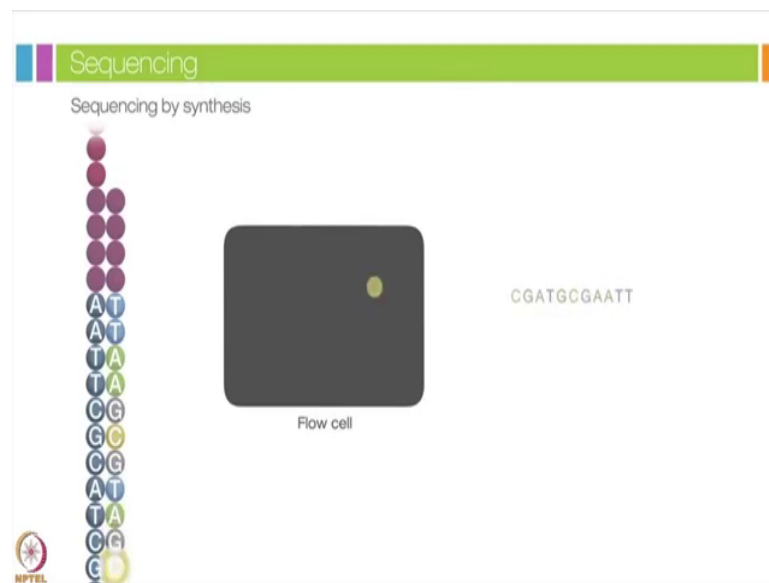
Sequencing begins with the extension of the first sequencing primer to produce the first read.

(Refer Slide Time: 23:51)



With each cycle fluorescently tagged nucleotides compete for addition to the growing chain.

(Refer Slide Time: 23:56)



Only one is incorporated based on the sequence of the template. After the addition of each nucleotide the clusters are excited by a light source and a characteristic fluorescent signal is emitted, this proprietary process is called sequencing by synthesis.

(Refer Slide Time: 24:29)

Sequencing



The slide features a green header with the word 'Sequencing'. Below the header, on the left, is a dark square labeled 'Flow cell' containing a dense field of small, multi-colored dots representing DNA clusters. To the right of the flow cell is a list of 20 DNA sequences, each on a new line. The sequences are: GTAGTAAGAAACAAAAGCA, GCTAAGGCTTACGCCGTAC, CAGCAGTAGTAAGAAACAA, AAGGCTTACGCCGTACTAC, CGTACTACCTCAGCAGTAG, GTAGTAAGAAACAAAAGCA, GCTAAGGCTTACGCCGTAC, CAGCAGTAGTAAGAAACAC, AAGGCTTACGCCGTACTAC, CGTACTACCTCAGCAGTAG, GTAGTAAGAAACAAAAGCA, GCTAAGGCTTACGCCGTAC, CAGCAGTAGTAAGAAACAA, AAGGCTTACGCCGTACTAC, CGTACTACCTCAGCAGTAG, GTAGTAAGAAACAAAAGCA, GCTAAGGCTTACGCCGTAC, TTGCAGTAGTAAGAAACAA, and AAGGCTTACGCCGTACTAC. In the bottom left corner, there is a small circular logo with the text 'HPTCL' below it.

Flow cell

GTAGTAAGAAACAAAAGCA
GCTAAGGCTTACGCCGTAC
CAGCAGTAGTAAGAAACAA
AAGGCTTACGCCGTACTAC
CGTACTACCTCAGCAGTAG
GTAGTAAGAAACAAAAGCA
GCTAAGGCTTACGCCGTAC
CAGCAGTAGTAAGAAACAC
AAGGCTTACGCCGTACTAC
CGTACTACCTCAGCAGTAG
GTAGTAAGAAACAAAAGCA
GCTAAGGCTTACGCCGTAC
CAGCAGTAGTAAGAAACAA
AAGGCTTACGCCGTACTAC
CGTACTACCTCAGCAGTAG
GTAGTAAGAAACAAAAGCA
GCTAAGGCTTACGCCGTAC
TTGCAGTAGTAAGAAACAA
AAGGCTTACGCCGTACTAC

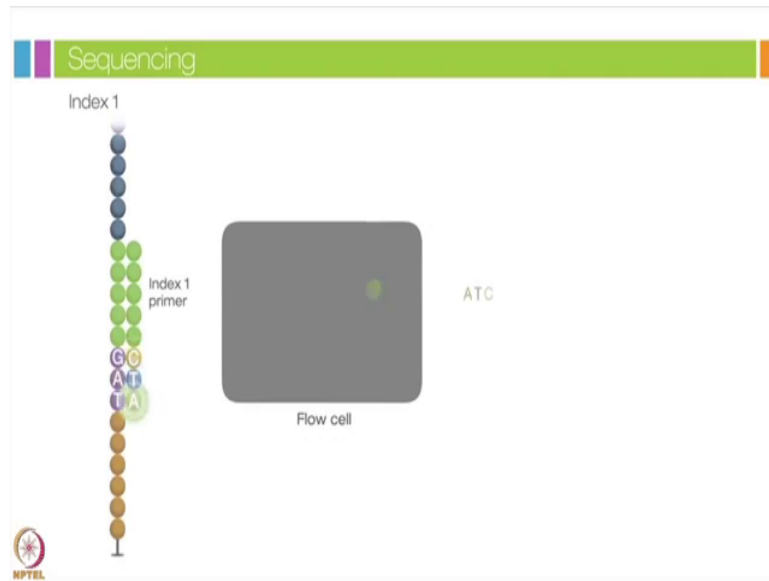
HPTCL

The number of cycles determines the length of the read. The emission wavelength along with the signal intensity determined the base call, for a given cluster all identical strands are read simultaneously. Hundreds of millions of clusters are sequenced in a massively parallel process; this image represents a small fraction of the flow cell. After the completion of the first read the read product is washed away in this step the index one read primer is introduced and hybridized to the template the read is generated similar to the first read.

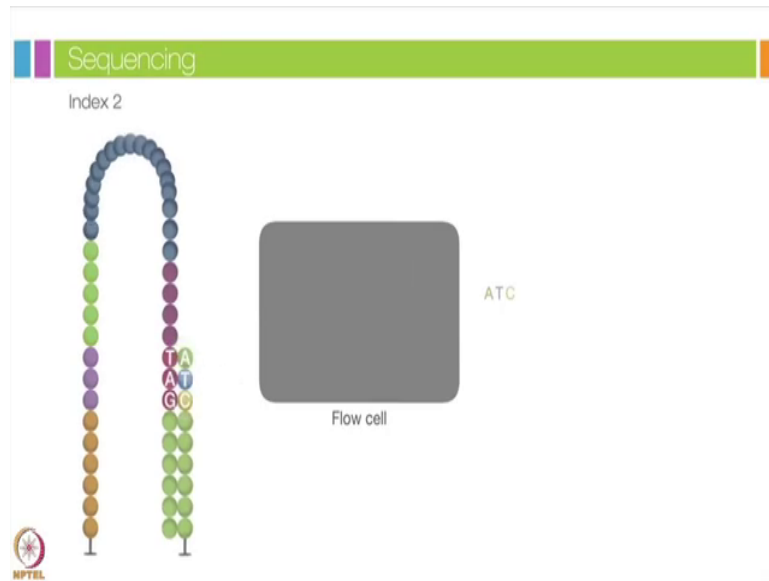
(Refer Slide Time: 24:40)



(Refer Slide Time: 24:44)



(Refer Slide Time: 25:03)



(Refer Slide Time: 25:14)



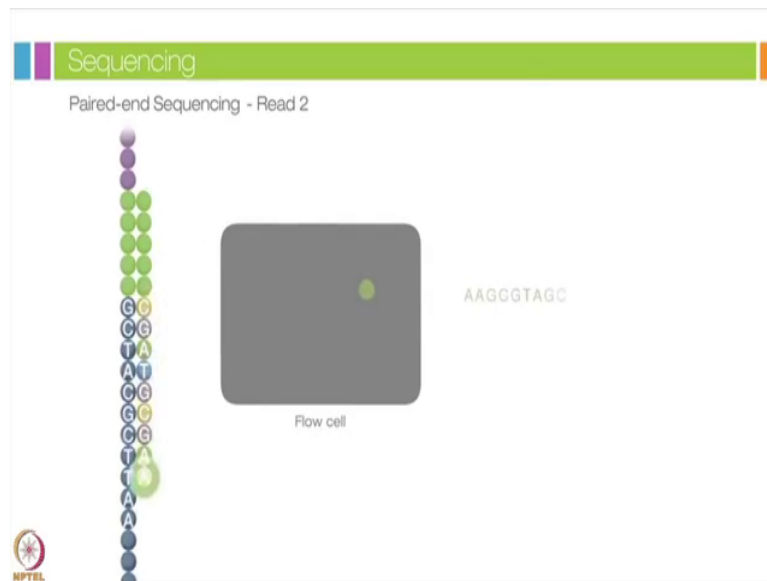
Index 2 is read in the same manner as index 1. Index 2 read product is washed off at the completion of this step. Polymerase extend the second flow cell oligo forming a double stranded bridge.

(Refer Slide Time: 25:24)



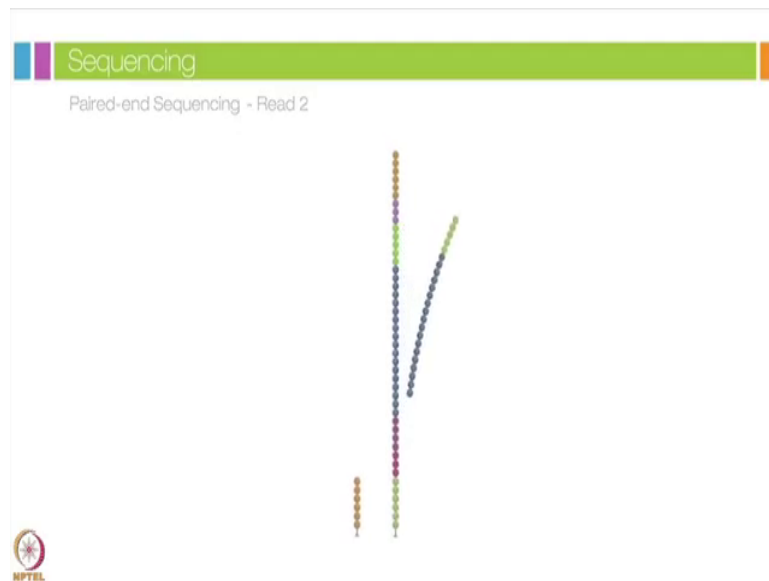
This double stranded DNA is then linearized and the 3 prime ends blocked, the original forward strand is cleaved off and washed away leaving the reverse strand.

(Refer Slide Time: 25:32)



Read 2 begins with the introduction of the read 2 sequencing primer. As with read 1 the sequencing steps are repeated until the desired read length is achieved.

(Refer Slide Time: 25:46)



The read two product is washed away.

(Refer Slide Time: 25:51)

Data Analysis



Flow cell

```
AAAAGCAATTGACAAACC  
GCCGTACTACCTCAGCAG  
GAAACAAAAGCAATTGAC  
GTACTACCTCAGCAGTAG  
GCAGTAGTAAGAAACAAA  
AAAAGCAATTGACAAACC  
GCCGTACTACCTCAGCAG  
GAAACAAAAGCAATTGAC  
GTACTACCTCAGCAGTAG  
GCAGTAGTAAGAAACAAA  
AAAAGCAATTGACAAACC  
GCCGTACTACCTCAGCAG  
GAAACAAAAGCAATTGAC  
GTACTACCTCAGCAGTAG  
GCAGTAGTAAGAAACAAA  
AAAAGCAATTGACAAACC  
GCCGTACTACCTCAGCAG  
GAAACAAAAGCAATTGAC  
GTACTACCTCAGCAGTAG
```



This entire process generates millions of reads representing all the fragments. Sequences from pooled sample libraries are separated based on the unique indices introduced during the sample preparation.

(Refer Slide Time: 25:58)

 Data Analysis

```
CTCAGCAGTAGTAAGAAACAAAAGCAATTGACAAACCTCCTTCTTATTCTTAGAAACAA
CTCAATGGCTAAGGCTTACGCCGTACTACCTCAGCAGTAGTAAGAAACAAAAGCAATT
ACTACCTCAGCAGTAGTAAGAAACAAAAGCAATTGAGAAACCTCCTTCTTATTCTTAGA
AATGGCTAAGGCTTACGCCGTACTACCTCAGCAGTAGTAAGAAACAAAAGCAATTGAC
CTTACGCCGTACTACCTCAGCAGTAGTAAGAAACAAAAGCAATTGACAAACCTCCTTCT
CTCAGCAGTAGTAAGAAACAAAAGCAATTGACAAACCTCCTTCTTATTCTTAGAAACAA
CTCAATGGCTAAGGCTTACGCCGTACTACCTCAGCAGTAGTAAGAAACAAAAGCAATT
ACTACCTCAGCAGTAGTAAGAAACAAAAGCAATTGACAAACCTCCTTCTTATTCTTAGA
AATGGCTAAGGCTTACGCCGTACTACCTCAGCAGTAGTAAGAAACAAAAGCAATTGAC
CTTACGCCGTACTACCTCAGCAGTAGTAAGAAACAAAAGCAATTGACAAACCTCCTTCT
CTCAGCAGTAGTAAGAAACAAAAGCAATTGACAAACCTCCTTCTTATTCTTAGAAACAA
CTCAATGGCTAAGGCTTACGCCGTACTACCTCAGCAGTAGTAAGAAACAAAAGCAATT
ACTACCTCAGCAGTAGTAAGAAACAAAAGCAATTGACAAACCTCCTTCTTATTCTTAGA
AATGGCTAAGGCTTACGCCGTACTACCTCAGCAGTAGTAAGAAACAAAAGCAATTGAC
CTTACGCCGTACTACCTCAGCAGTAGTAAGAAACAAAAGCAATTGACAAACCTCCTTCT
CTCAGCAGTAGTAAGAAACAAAAGCAATTGACAAACCTCCTTCTTATTCTTAGAAACAA
CTCAATGGCTAAGGCTTACGCCGTACTACCTCAGCAGTAGTAAGAAACAAAAGCAATT
ACTACCTCAGCAGTAGTAAGAAACAAAAGCAATTGACAAACCTCCTTCTTATTCTTAGA
AATGGCTAAGGCTTACGCCGTACTACCTCAGCAGTAGTAAGAAACAAAAGCAATTGAC
```

 NPTEL

(Refer Slide Time: 26:03)

Data Analysis

Local sequence clustering

Similar sequences

```
CTGAGTADTGTAGAAACAAAGCAATTGCAAACTCTCTCTATTCTAGAAACA  
CTGAGCAATGATAGAAACAAGCAATTGCAAACTCTCTCTATTCTAGAAACA  
CTGAGCAATGATAGAAACAAGCAATTGCAAACTCTCTCTATTCTAGAAACA  
CTGAGCAATGATAGAAACAAGCAATTGCAAACTCTCTCTATTCTAGAAACA
```

```
CTAGCGCTACTACTCTGAGCAATGATAGAAACAAGCAATTGCAAACTCTCTCT  
CTAGCGCTACTACTCTGAGCAATGATAGAAACAAGCAATTGCAAACTCTCTCT  
CTAGCGCTACTACTCTGAGCAATGATAGAAACAAGCAATTGCAAACTCTCTCT
```

```
AATGGCTAGGGCTTACGGCTACTACTCTGAGCAATGATAGAAACAAGCAATTGCA  
AATGGCTAGGGCTTACGGCTACTACTCTGAGCAATGATAGAAACAAGCAATTGCA  
AATGGCTAGGGCTTACGGCTACTACTCTGAGCAATGATAGAAACAAGCAATTGCA  
AATGGCTAGGGCTTACGGCTACTACTCTGAGCAATGATAGAAACAAGCAATTGCA
```

```
CTCAATGGCTAGGGCTTACGGCTACTACTCTGAGCAATGATAGAAACAAGCAATT  
CTCAATGGCTAGGGCTTACGGCTACTACTCTGAGCAATGATAGAAACAAGCAATT  
CTCAATGGCTAGGGCTTACGGCTACTACTCTGAGCAATGATAGAAACAAGCAATT  
CTCAATGGCTAGGGCTTACGGCTACTACTCTGAGCAATGATAGAAACAAGCAATT
```

RPTEL

For each sample reads with similar stretches of base calls are locally clustered.

(Refer Slide Time: 26:08)

Data Analysis

Create contiguous sequences

```

Forward read  CTGAGCAGTGTAGAAAACAAAAGCAATTGCAAAAGCTCCCTCTTATCTTAGAAAACA
                CTGAGCAGTGTAGAAAACAAAAGCAATTGCAAAAGCTCCCTCTTATCTTAGAAAACA
                CTGAGCAGTGTAGAAAACAAAAGCAATTGCAAAAGCTCCCTCTTATCTTAGAAAACA
Reverse read   ACTACCTCAGCAGTGTAGAAAACAAAAGCAATTGCAAAAGCTCCCTCTTATCTTAGAAAACA
                ACTACCTCAGCAGTGTAGAAAACAAAAGCAATTGCAAAAGCTCCCTCTTATCTTAGAAAACA
                ACTACCTCAGCAGTGTAGAAAACAAAAGCAATTGCAAAAGCTCCCTCTTATCTTAGAAAACA
                CTTAAGCCCTACTACTCTCAGCAGTGTAGAAAACAAAAGCAATTGCAAAAGCTCCCTCT
                CTTAAGCCCTACTACTCTCAGCAGTGTAGAAAACAAAAGCAATTGCAAAAGCTCCCTCT
                CTTAAGCCCTACTACTCTCAGCAGTGTAGAAAACAAAAGCAATTGCAAAAGCTCCCTCT
                AATGGCTAAGGCTTAAAGCGGTACTACTCTCAGCAGTGTAGAAAACAAAAGCAATTGC
                AATGGCTAAGGCTTAAAGCGGTACTACTCTCAGCAGTGTAGAAAACAAAAGCAATTGC
                AATGGCTAAGGCTTAAAGCGGTACTACTCTCAGCAGTGTAGAAAACAAAAGCAATTGC
                CTCAATGGCTAAGGCTTAAAGCGGTACTACTCTCAGCAGTGTAGAAAACAAAAGCAATT
                CTCAATGGCTAAGGCTTAAAGCGGTACTACTCTCAGCAGTGTAGAAAACAAAAGCAATT
                CTCAATGGCTAAGGCTTAAAGCGGTACTACTCTCAGCAGTGTAGAAAACAAAAGCAATT
                CTCAATGGCTAAGGCTTAAAGCGGTACTACTCTCAGCAGTGTAGAAAACAAAAGCAATT

```



Forward and reverse reads are paired creating contiguous sequences.

(Refer Slide Time: 26:11)

The slide is titled "Data Analysis" and has a subtitle "Create contiguous sequences". It contains a diagram with four horizontal bars of varying lengths and positions, arranged in a staircase pattern. The bars are colored in shades of blue and purple. The top bar is blue and is the longest. The second bar is purple and is shorter than the first. The third bar is blue and is shorter than the second. The bottom bar is purple and is the shortest. The bars overlap in a way that suggests a sequence of steps or a staircase. In the bottom left corner, there is a small circular logo with a star and the letters "HYTEL" below it.

(Refer Slide Time: 26:15)




These contiguous sequences are aligned back to the reference genome for variant identification. The paired end information is used to resolve ambiguous alignments.

(Refer Slide Time: 26:24)

Points to Ponder

- Basics of DNA sequencing through Sanger's sequencing method
- Basic steps involved in NGS sequencing method: (i) Library Preparation, (ii) Cluster Generation, (iii) Sequencing, (iv) Data Analysis
- The 'sequencing by synthesis' approach used by Illumina platform to perform high-throughput sequencing
- Utility of paired-end sequencing that provides the sequence of both forward and reverse strands



MOOC-NPTEL

IIT Bombay

All right. So, I am sure by now you are very clear and convinced about the magic which next generation sequencing platforms have done for us. The pace, the accuracy, the speed, the cost what one could accomplish by this platform was not even possible to a think 10 years ago.

So, really rapid advancement which have been made in this area are tremendous and now the major advantage one could see from this that many applications are directly reaching to the clinics. So, now, doctors are pretty much relying on sequencing technologies and their results for the patient care. And this itself conveys that a technology has reached to its robustness, this maturity to its accuracy to an extent that now it could be brought to the clinics and for the patient care.

So, now you are getting introduced to different type of platforms for NGS, it is entirely up to you to think about what are the pros and cons of each technology which technology offers you

what more superiority and advantages. But I must say all these technologies are very good it all depends on whether your aim was to do the whole genome sequencing, RNA sequencing only looking at the panel of the genes or what the exactly you want to address. Accordingly you can choose a platform there are many next generation sequencing, technologies are really at the advanced level and it entirely depends on you which platform you can choose.

Nevertheless just you know keep in mind that this NGS is a parallel sequencing technology which have really changed the way we have seen how to look inside at the genome level. And these applications have made tremendous revolution in the entire biological science and medical science area. With the ultra high throughput the scalability and speed, the NGS technology enables researcher to perform a wide variety of applications and study biological systems at a level which was never possible before.

Today I hope you have learnt about the basics of NGS is starting from the Sanger sequencing to the Illumina platform using sequencing by synthesis method. In the next class you will study another application of NGS using another leading technology platform and we will continue our discussion in the next lecture as well.

Thank you.