

Lecture-35
Next-Generation Sequencing Technology- MiSeq System

Welcome to Mooc course on applications of Interactomics using genomics and proteomics technologies in the last few lectures, we are discussing about one of the revolutionary technology next generation sequencing technology where you are given the concepts in 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 lecture series today we have Mr. Rahul Solanki a senior field application scientist from premise Life Sciences who will talk to us about Illumina next generation sequencing workflows. So, let me welcome Mr. Rahul Solanki for his lecture.

Today I'd say about protein sequencing what exactly the sequencing is, to know the sequence of amino acids if I speak about DNA sequencing right? So why we need it why it is needed so why you need to identify gene, study mutations so ultimately all the functions which are related to the protein ultimately those are coded by gene Right? it's correct. So, start I'll start with the Sanger though we don't have the Sanger sequencer with us but just to clarify the basic things I'll start with the Sanger so, there's the principle what you do is let's 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 (2:22)

Automated Capillary Sequencing: 1986

The diagram illustrates the automated capillary sequencing process. It starts with a DNA template being synthesized with a mixture of labeled and unlabeled nucleotides. The resulting DNA fragments are separated by capillary electrophoresis. The fluorescence of the fragments is detected by a laser, and the data is processed by a computer to generate a sequence 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

illumina

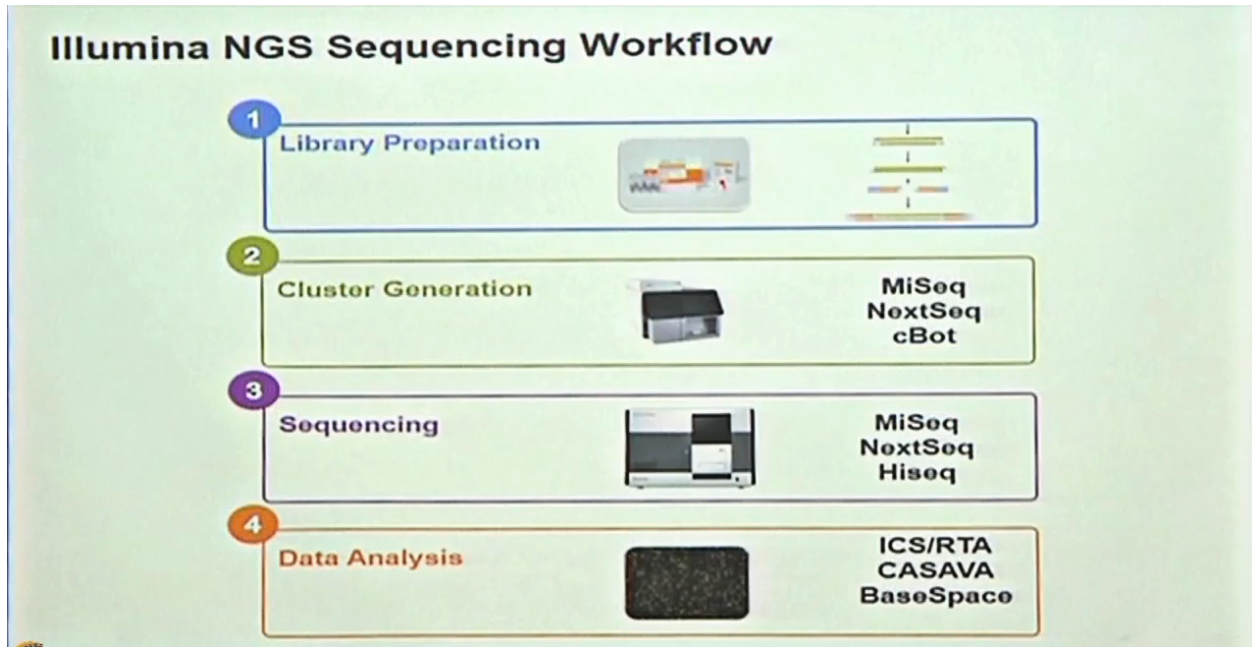
So, can you see the stretch of sequences the label, so those are ddNTPs which are labeled with a fluorescent. So let's say the first one is A added Right? so that polymerase can't extended why because the

base added is ddNTPs it can't it lacks a OH group Okay? 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 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 labeled 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 per one smallest one Okay? So through a agarose gel then top most sequence will be yeah! it will pass away first Okay? So, let's 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's the basic of Sanger sequencing so what is the maximum a base pair which we you can sequence using Sanger good result quality data, housing 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? But why we need NGS now the point is why the NGS is required Great? So, the right answer is throughput. Any idea about how long the first human genome took to get it sequenced, it took ten years, and almost how much million dollar was invested, a billion dollars how many mam almost three billion dollars, was invested for I first human genome to be sequence. So, what is the length of human genome Great? it's a common question of see a internet I mean who appeared in the night Right? Okay? I hope you are writing so it's three billion base pair three into ten to the power nine 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 three billion base pair three into ten to the power of nine, and the maximum length of gene you can sequences thousand base pair so three billion divided by thousand, how much it is how many reactions you need to carry out, how many, three million so three million Sangha reaction you need to carry out to complete the human genome sequence, and the sequence you will be getting will be covered 1x coverage.

So, even if you carry out one three million reactions also, you will be covering your bases how many times one time so the coverage will be 1x. So, if you wish to sequence a human genome do a which to invest in years more no so with the help of NGS technologist you can sequence fifty human genomes in a span of forty eight hours, how many fifty human genome in a span of forty eight hours.

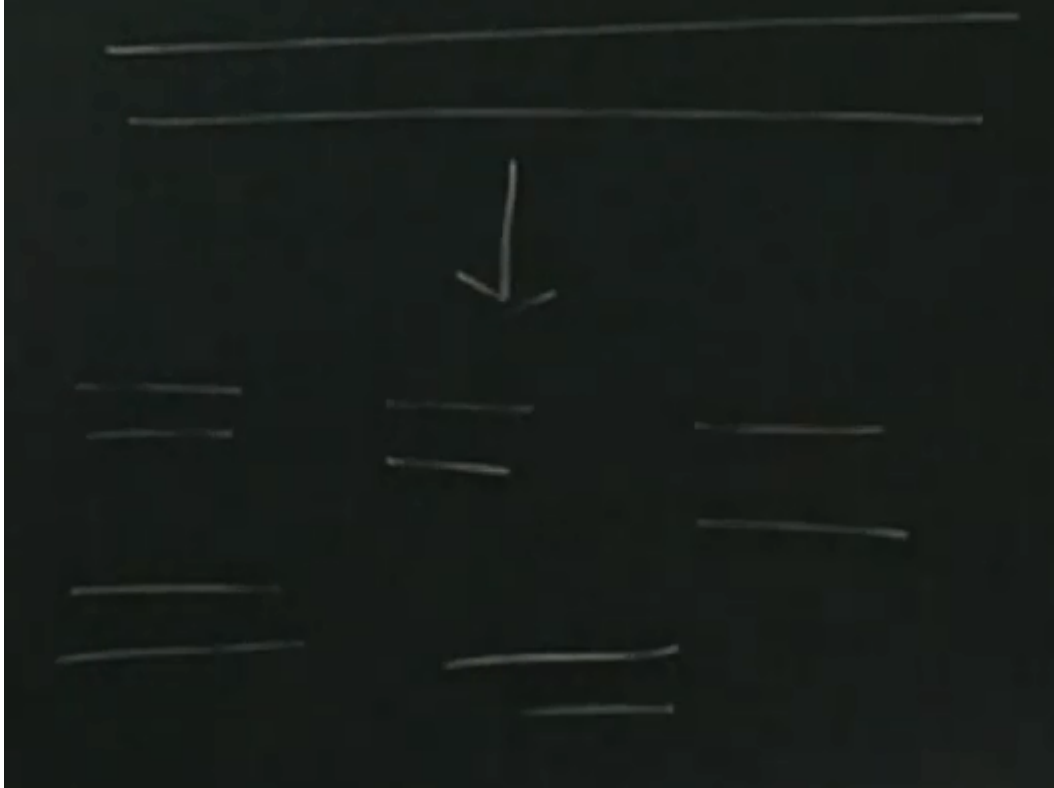
Refer Slide Time (5:56)



So, the key idea remains the same the principle remains the same what we do is let's say you have a this human genome let's say so what we do initially is we fragment this human genome into very small fragments, and just like cloning we don't need clothing at all in the case of NGS what we do here is, so I'll be covering all the parts one by one the very first step involved in all the sequencer our library preparation which is followed by cluster generation then comes is the sequencing and final is the data analysis part Okay? So first is the library preparation it's common for all the platforms though the sequencer are different but the concept remains the same, so what is library first of all in terms of sequencing if I say what is the library in Sanger what do you do I you clone it that vector contains the known sequences already it contains the primer for what the sequencers are no one.

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 and the unknown Right? Correct. what you will do? Where is the chalk so again the same question am, I asking?

Refer Slide Time (7:23)



let's see if there is a human genome in the Sanger what we do in the genome humans use on big genes what we do initially the first step is we fragmented into very small fragments let's say human genome so six hundred base pair will chop it down into various millions of fragments but these are unknown. So to get them sequenced what we need Right? you remember now so what we will do is initially we will use a polymerase is same so the tendency of polymerase is what it does is if you have this many fragments so initially we converted into blunt end it, all ends in tubes intended 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 simple ligation step. So, those fragments will be converted into the fragment which will be ligated with the adapters.

So, in library preparation the ultimate goal is,

Refer Slide Time (8:21)

Next-Generation Sequencing



Massively Parallel Sequencing >100x-1000x

```

AAAAACGAGAGTCTAGCAGCTTCTCATCAGGAGGA
AAAGCAGAGTCTAGCAGCTTCTCATCAGGAGGA
AACCCAGAGTCTAGCAGCTTCTCATCAGGAGGA
ACCGAGTCTAGCAGCTTCTCATCAGGAGGA
ACCGAGTCTAGCAGCTTCTCATCAGGAGGA
CCAGAGTCTAGCAGCTTCTCATCAGGAGGA
GAGTCTAGCAGCTTCTCATCAGGAGGA
CTAGCAGCTTCTCATCAGGAGGA
TAGCAGCTTCTCATCAGGAGGA
AGCAGCTTCTCATCAGGAGGA
CCCTTCTCATCAGGAGGA
AGCTTCTCATCAGGAGGA
CTTCTCATCAGGAGGA
ATCAGGAGGA
TCAGGAGGA
GAGGAGGA
GAGGAGGA
AAGCTTCTCATCAGGAGGA
AAGCTTCTCATCAGGAGGA
AAGCTTCTCATCAGGAGGA
AAGCTTCTCATCAGGAGGA
AAAAACGAGAGTCTAGCAGCTTCTCATCAGGAGGA
    
```

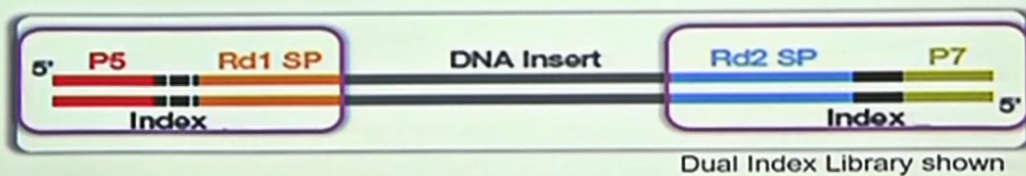
NGS provides "digital" data, each read is analysed independently and is quantitative

Overlapping reads are aligned together, resulting in quantitative and high confidence variant calling

the first of all we start with the DNA fragment what we do is we fragment it into various lengths of fragments let's say six hundred base pair, then we repaired the ends now it is converted into blunt ended. And finally, we like it the adapters. So, what are adapters, adapters are the switch of DNA which for which the sequences are known already known Right?

Refer Slide Time (8:43)

No matter the input, all libraries end up looking similar

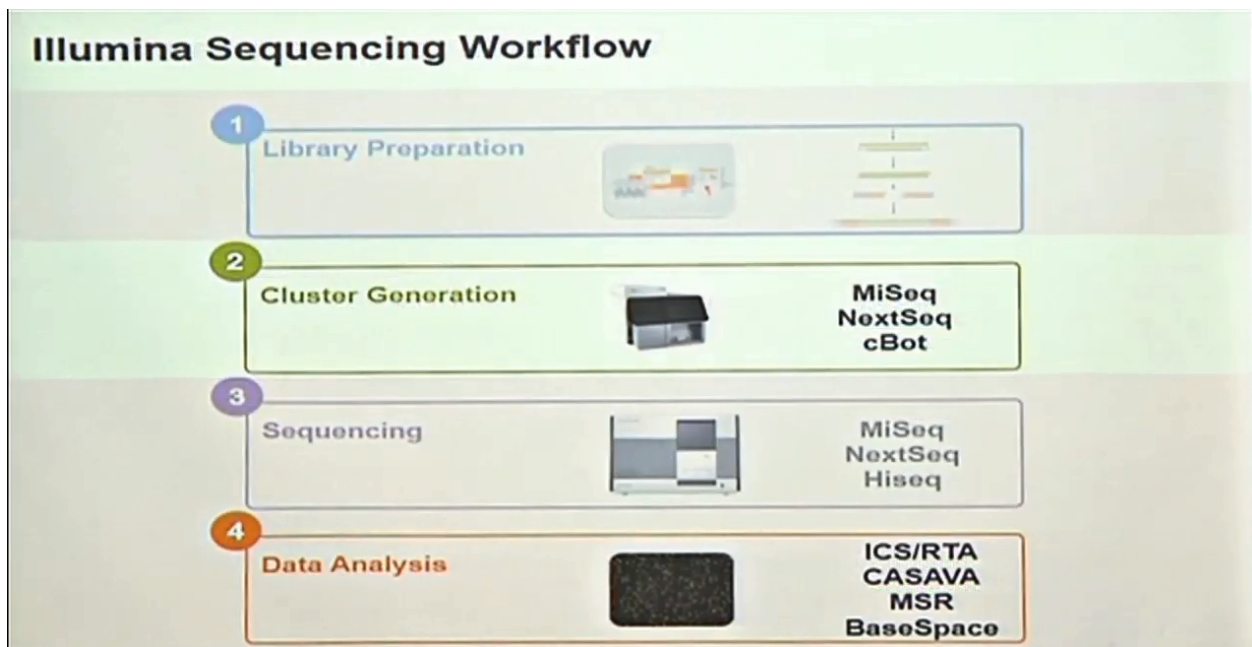


The aim of the sample prep step is to obtain nucleic acid fragments with adapters attached on both ends

So, this is how the library looks like finally, it will contain a DNA insert which you wish to sequence in the middle and of now you can see there are the two regions Rd1, Rd2 and these are essentially the

adapters. We have two kinds of adapters P5 and P7 Alright? is it clear. So I'll tell you what is the function of all there is the sequence unknown sequence Aright? here you can see they're two regions here our read one sequencing primer will bind so this is called as read one sequencing primer binding site flanked by both the ends are indices what are indices what are index these are barcodes, so what is the function of barcode so in NGS what you do? is you you can sequence as I said using noocity given sequence fifty genomes all together. So ultimately with the help of barcode you can identify which sequences belong to which of the samples am I clear Okay? 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 Right?


Refer Slide Time (10:01)



So, there's 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 light, you can see there two flow cells this HiSeq flow cell which is meant for sequencing human genome, and there's my MiSeq for targeted sequencing especially. So, it has eight lanes, it contains eight lanes and it has a commonly.

Refer Slide Time (10:27)

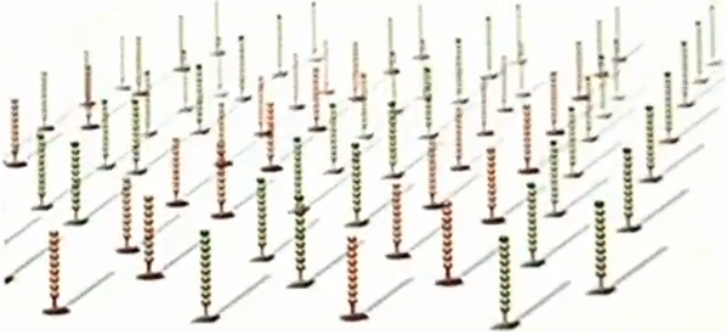
What is a Flow Cell?



Cluster generation occurs on a flow cell

A flow cell is a thick glass slide with channels or lanes

Each lane is randomly coated with a lawn of oligos that are complementary to library adapters

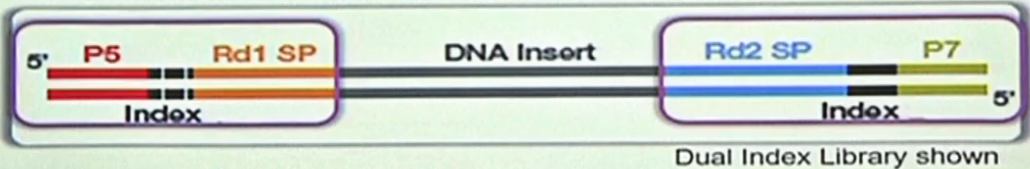


illumina

why we call it as a flow cell because the sequencing happens with the help of flow of reagents into this flow cell. Okay? So, the flow cell contains the lawn of oligos nucleotides.

Refer Slide Time (10:39)

No matter the input, all libraries end up looking similar



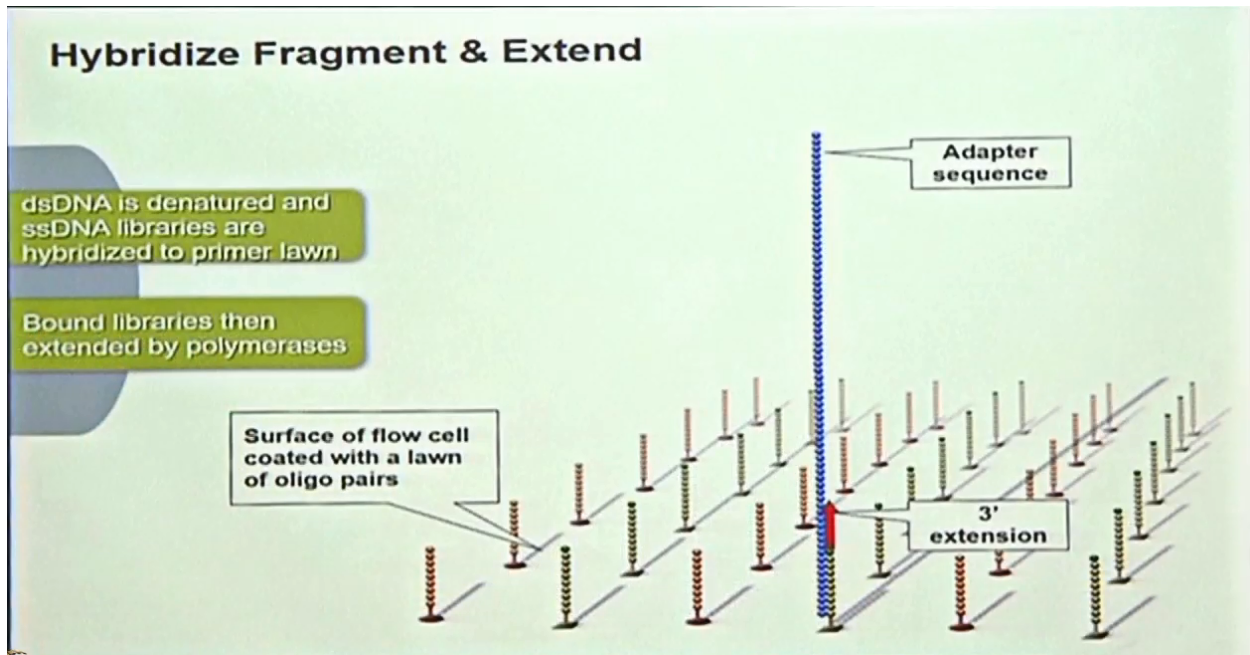
Dual Index Library shown

The aim of the sample prep step is to obtain nucleic acid fragments with adapters attached on both ends

So, essentially, we have two kinds of oligos, one oligo will be complementary to P5 region and one will be complementary to 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

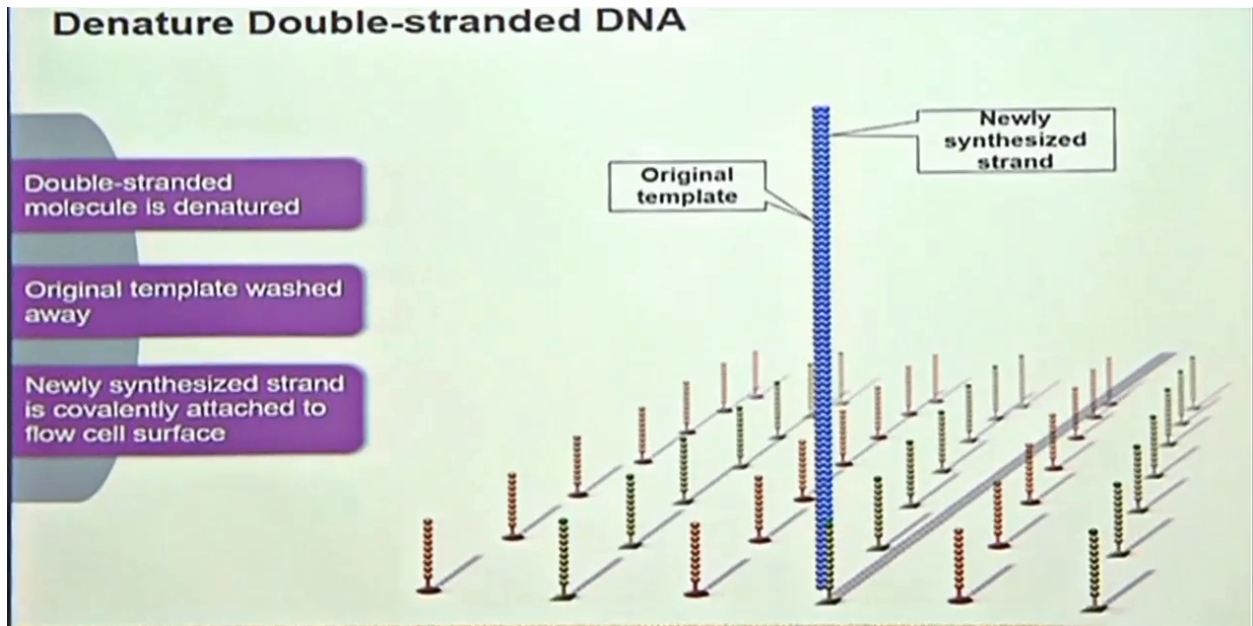
surface of flow cell Right? So, once you have denatured read this Lawns are complementary to those P5 and P7th Right? Already I have deleted it.

Refer Slide Time (11:07)



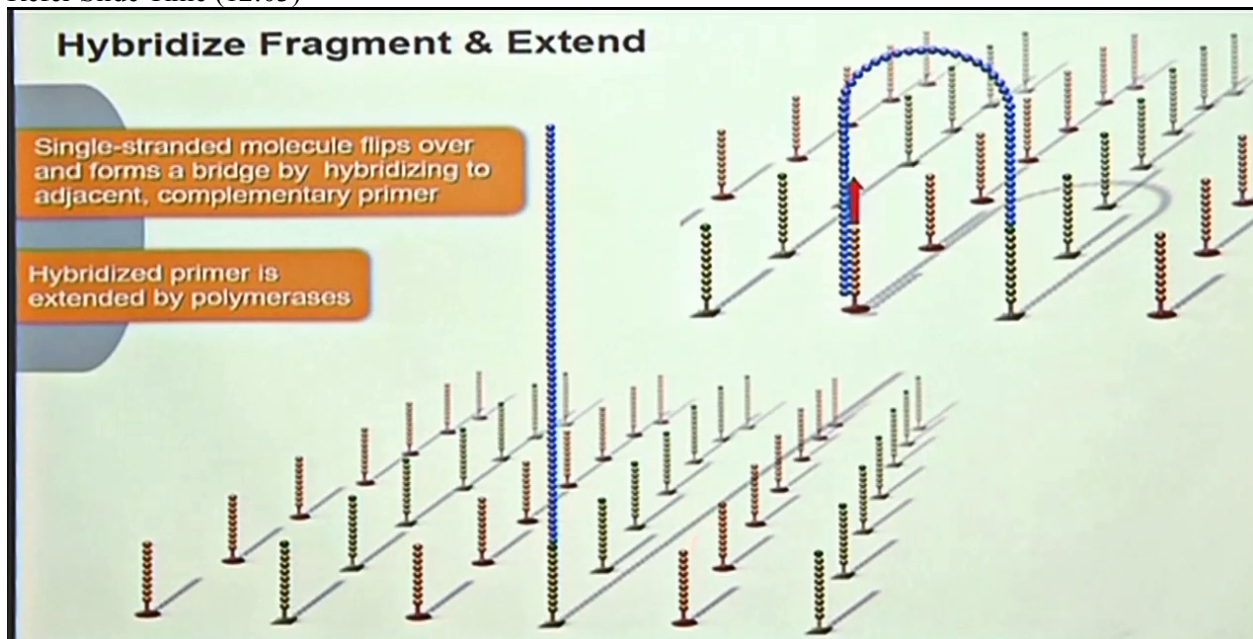
So, one of the fragment will go and by into the flow cell Right? In this region you can see how it binded here in the cases, with the help of hydrogen bonding simply because this region are complementary you can see this is covalently bound to the surface of flow cell, and this is because this region is complementary it will go and bind with the help of hydrogen bonding Okay? And we extend this ends with the help of DNA polymerase.

Refer Slide Time (11:37)



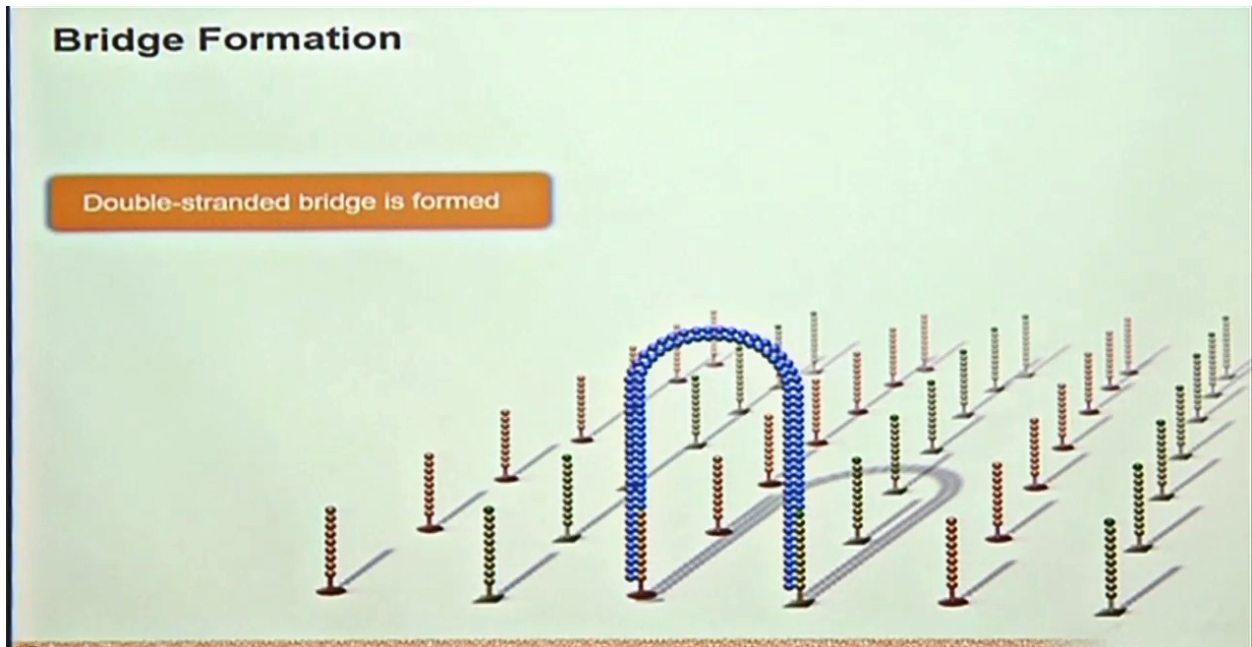
So, you can see you will be getting a structure like this then what we do is we retain this original fragment because all simply it is bound covalently to the surface of the cell, and because this strand is bounded by hydrogen bonding these are weak bond, so we denature it and we wash away this original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

Refer Slide Time (12:03)



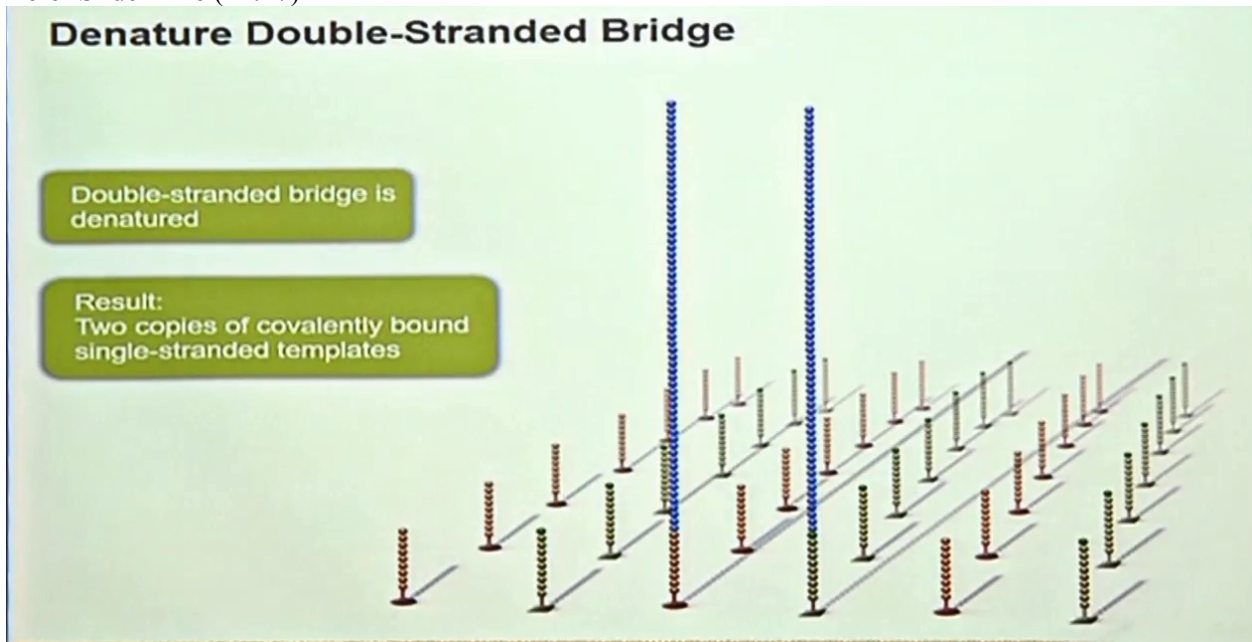
Then what you can see is this is left over here and since designed also this one this end and this end it is also complementary to this one you go it will slip over here again we will add DNA polymerase you will see a bridge kind of structure. Okay?

Refer Slide Time (12:19)



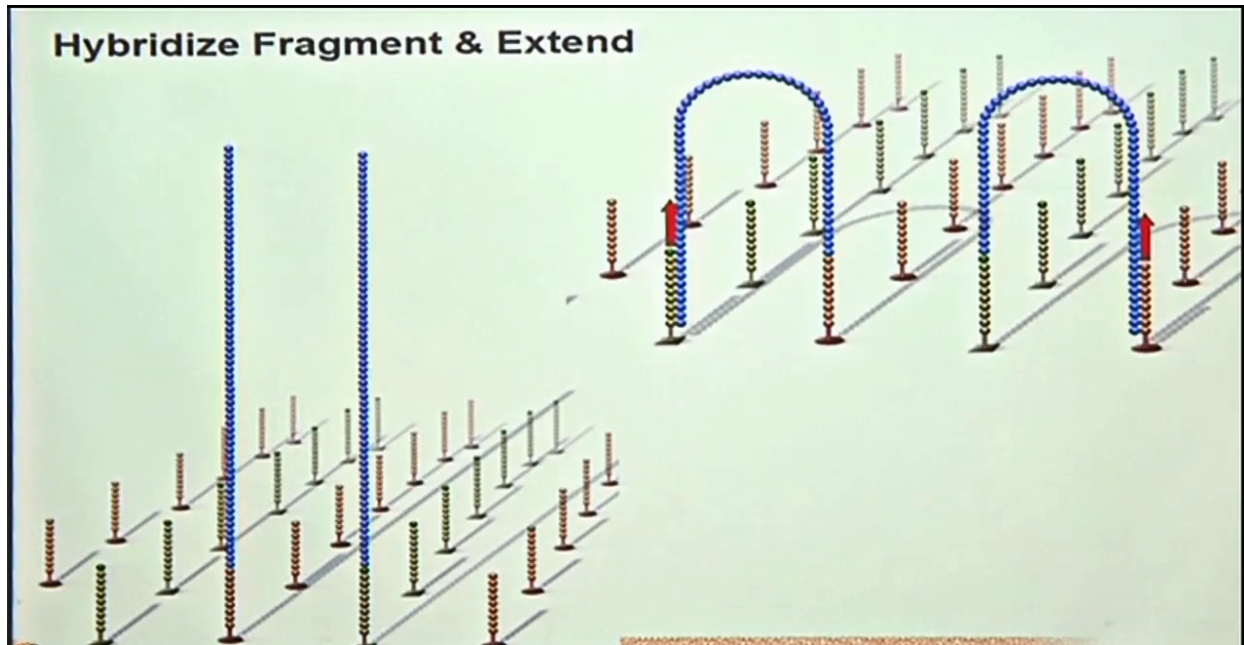
Then again, we will denature it and both this friend because they are bound to the surface covalently you will see a two stranded over here.

Refer Slide Time (12:27)



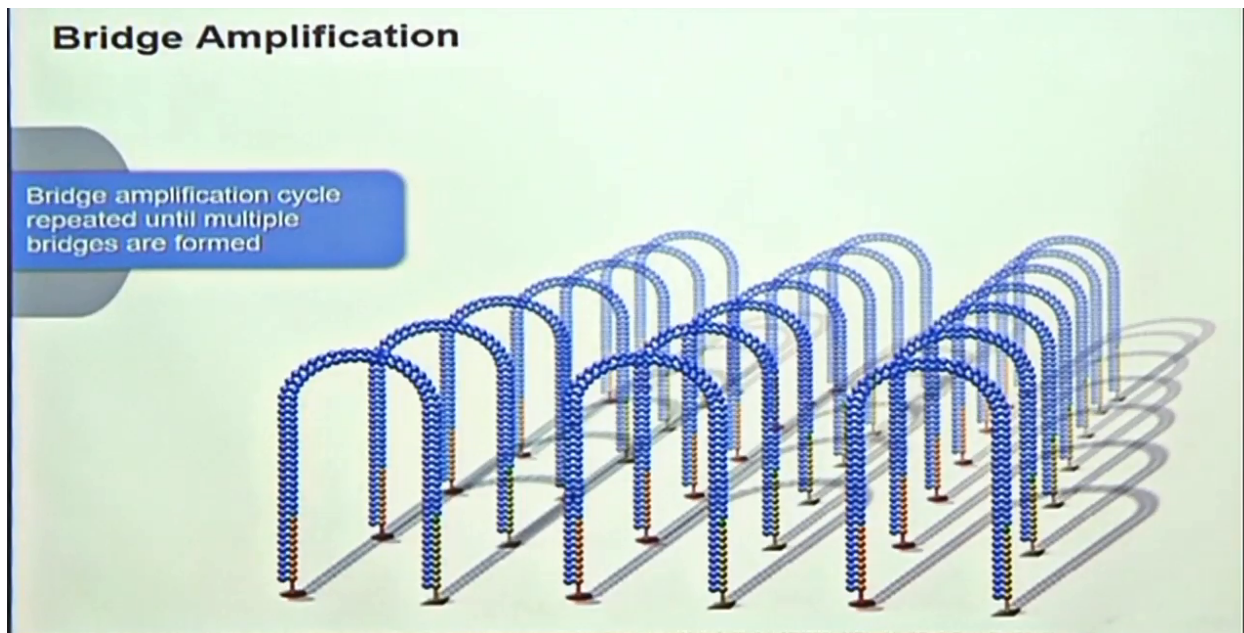
And this process will be repeated for millions of time,

Refer Slide Time (12:31)



and ultimately on the surface of flow cell

Refer Slide Time (12:35)

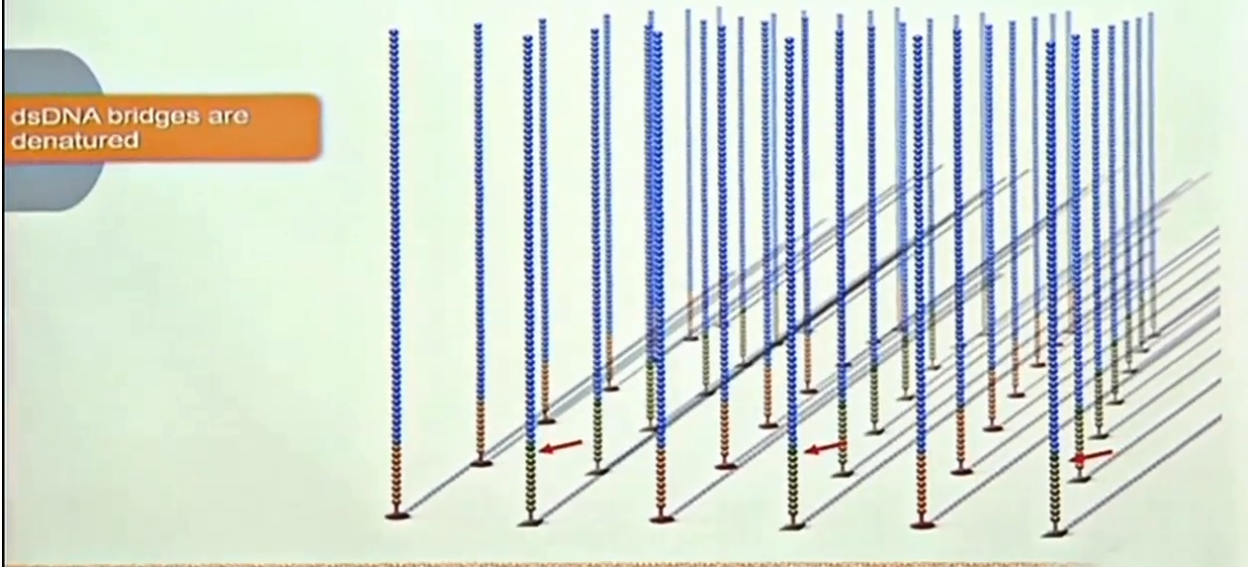


what you can see here as will be having a millions of fragment on the surface of flow cell Okay?

Refer Slide Time (12:41)

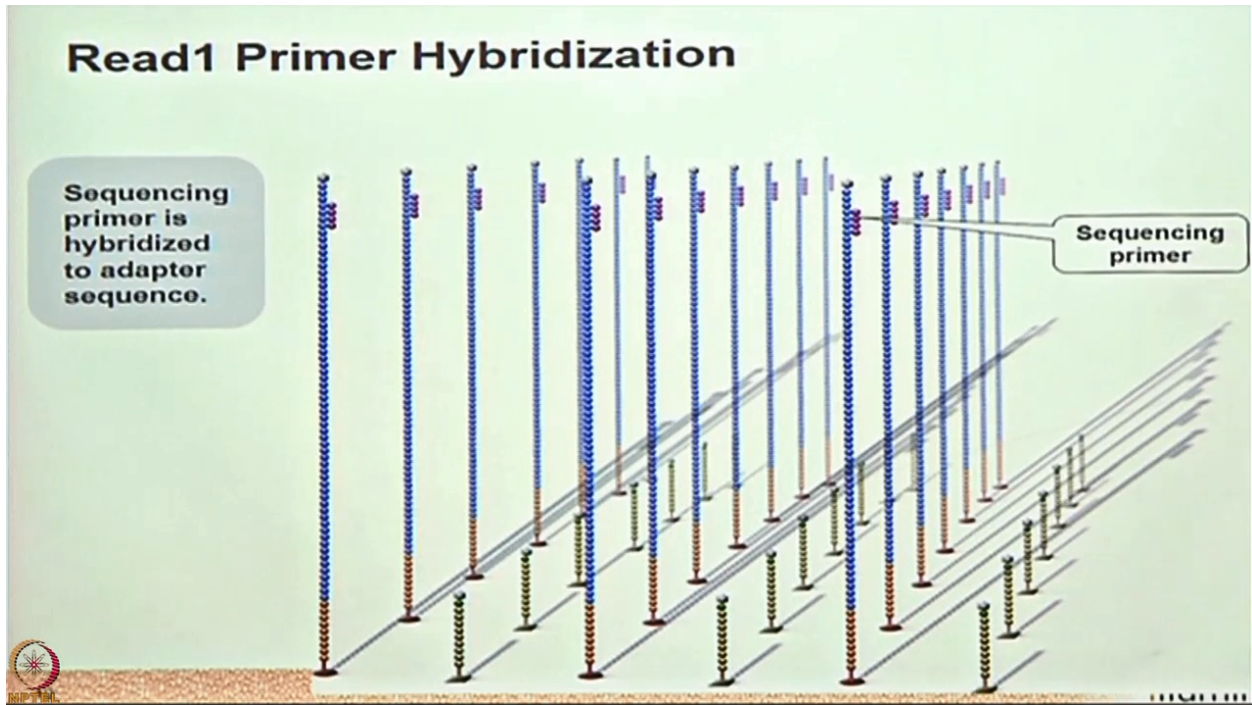
Linearization

dsDNA bridges are denatured



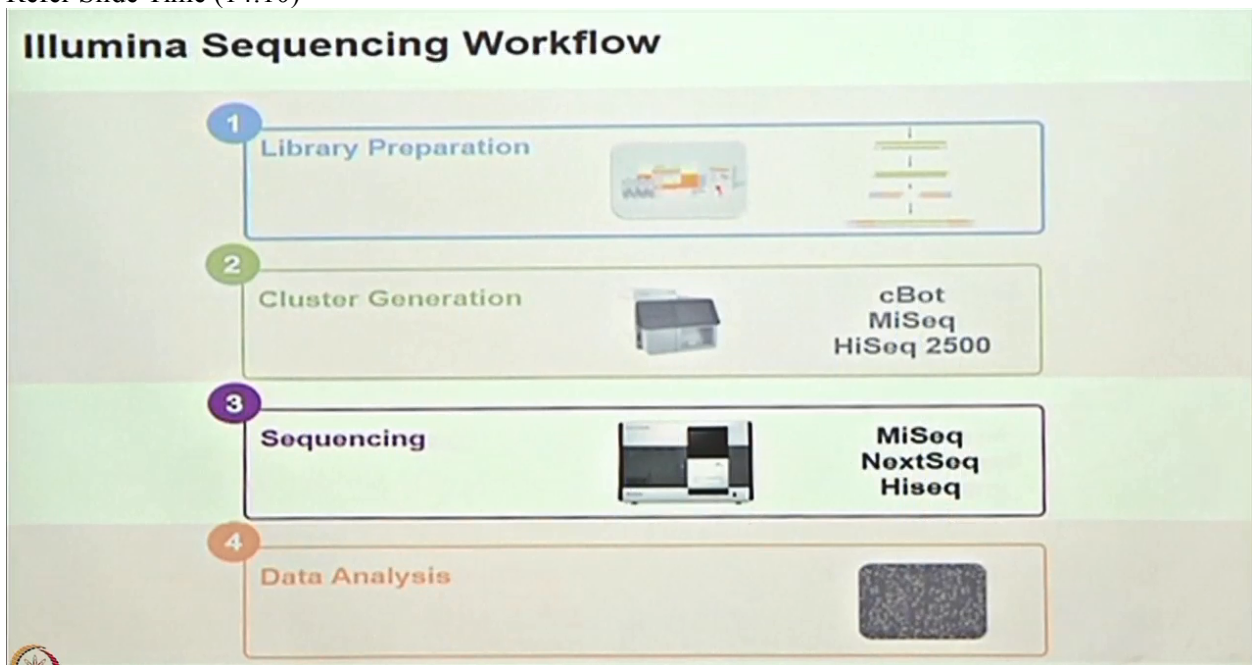
So, again we do these denature all the fragments so essentially you can see it contains two strands which strands it contains forward, and reverse all both strand it contains so what we do is initially we cleave the reverse strand and we carry out the sequencing for forward strand. Okay? So, now on the flow seal you can see which strand is retained hmm Alright? So, again there is a question what should I do to prevent again if I don't 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 go into this primer what should I do? Good. So, what I'll do next is I'll block the all free ends so that it won't flip Okay?

Refer Slide Time (13:43)



So, now you are done with your sequencing part I mean cluster generation is over 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 one sequencing primer. So, on the flow cell now you flow read one sequencing primer so you can see that primer go and bind over there then we are said to go for the sequencing.


Refer Slide Time (14:10)



Refer Slide Time (14:11)

Sequencing By Synthesis (SBS)

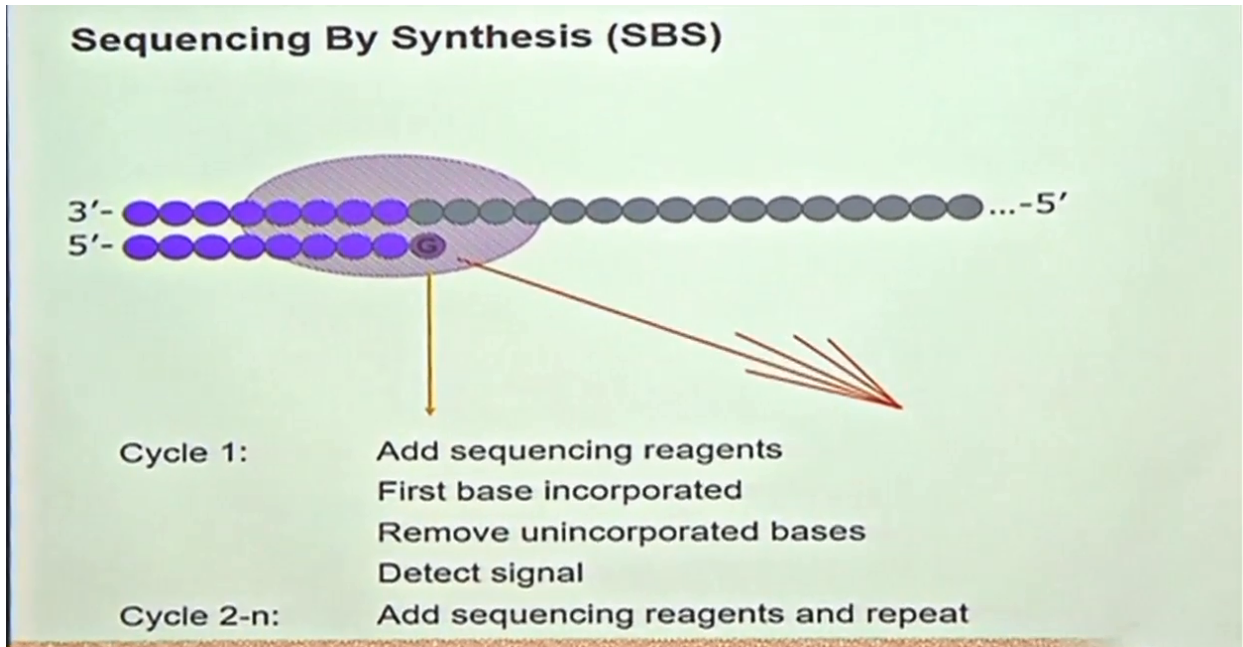
Cycle 1:	Add sequencing reagents
	First base incorporated
	Remove unincorporated bases
	Detect signal
Cycle 2-n:	Add sequencing reagents and repeat



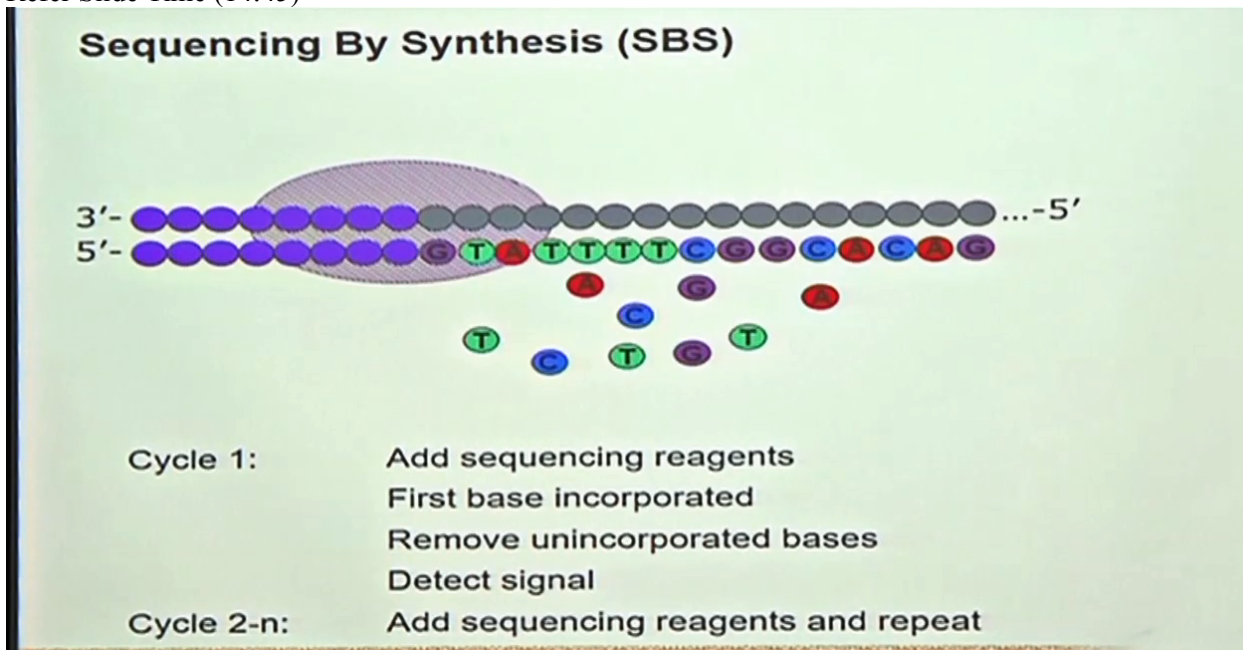
With Illumina what we do is we essentially mimic the Sanger sequencing just like natural phenomenon.

So, we have four kind of bases so let's say there's your strand we've added a primer this you can see all these bases are labeled with a fluorophore 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 said to go for the Nik cycle.

Refer Slide Time (14:37)



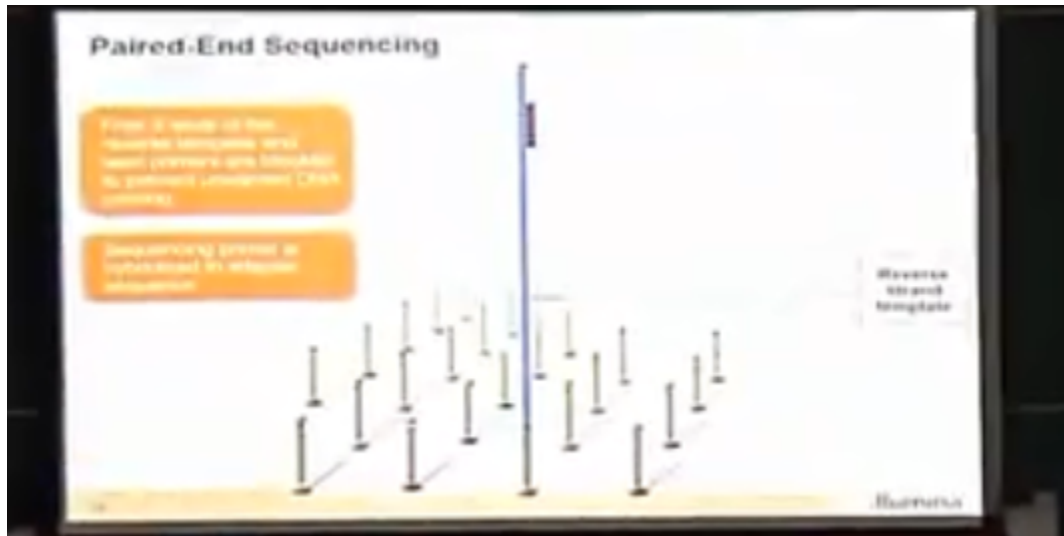
So, there is direct detection of the basis, so this let's say you are running a sample of three hundred cycle
Refer Slide Time (14:45)



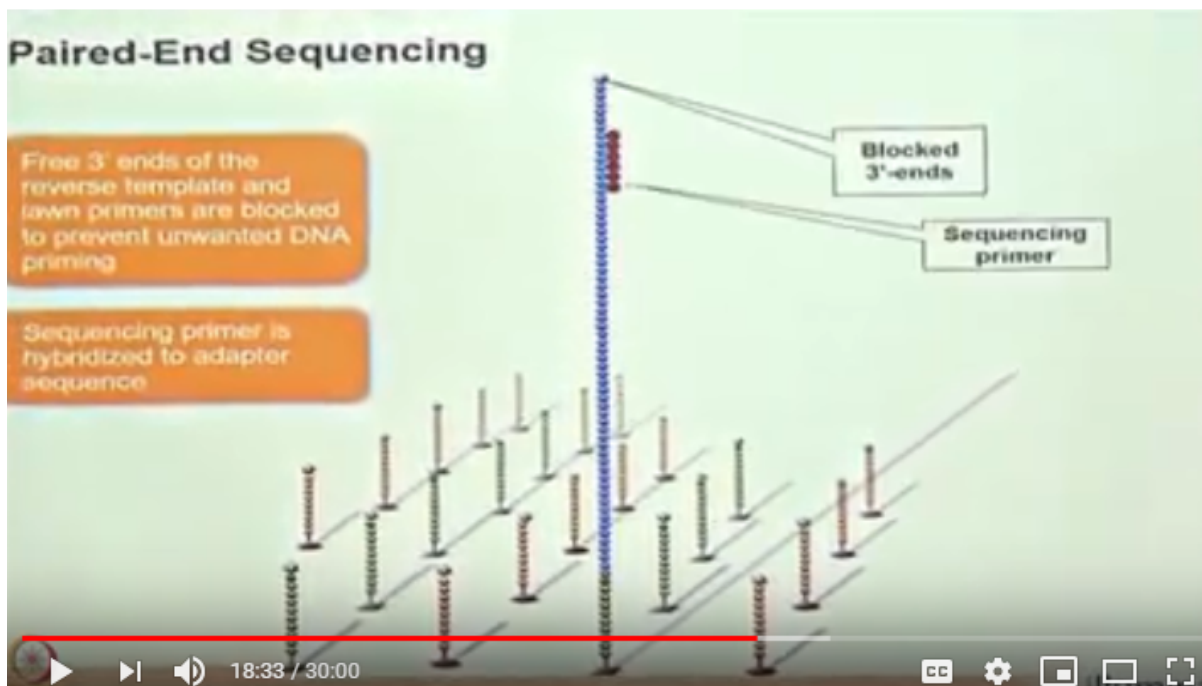
so, 300 base will go and bind this is your one fragment Okay? So, this is how if you are going for 300 cycle 300 basis will be added and we are done with the sequencing part. So, what is the key difference here in we essentially mimic the natural phenomenon there is no change phosphate or something the best part is if you detect any by secondary methods like phosphate or

something let's say. So, what will be the hurdle there? if one if one base is added let's say there is one call for A in a cycle. So, signal will be if two A are added the signal will be double but, if you have multiple A or multiple G over there. So, you can't resolve the signal. So, so, in the terms of NGS in the Sangha. How you detect the quality of the NGS data? By looking at the peak you get a peak you remember you data if the peaks are sharp your wall 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 30scores. So, a Q 30 score means is one error in thousand base pair added. So, the accuracy rate is 99.9 percent okay? If I say about you 30 what is Q 30. So, we define our run as let's say you performed a run which contains a equalized genome. So, we'll define 80% of the basis sequence were above Q 30values getting? So, the error rate in that 80% base calls for 99.9 percent not error rate accuracy rate the error rate was 0.1% or am I clear? So, what exactly the paid in sequencing is essentially you remember as mam said he initially we generate both fragments if you wish to sequence only forward strands. What you will do is you will cleave away the reverse prime and you will sequence the only forwards right the period in sequencing what you do is you regenerate both this trend 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 this strand off stretch of DNA you have to 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 paid in what we do is very regenerate we D block it remember? So, both the strand will be generated this time will clear away the forwards friend and we are to sequence the reverse strand. So, once we are done with the read one okay. So, let's say my read length was I have chosen 300 base pair heel in. So, there's your read one product and that this is your sequencing primer and from here how many bases were added let's say it was a three hundred cycle run. So, three hundred bases will be added over here those are detected and then what we do is we do measure this product the product of read one sequencing primer will denature it and, we are set to go for period sequencing what needs to be done is we will denature it again we will do block this design clear? So, we will deep block this and again it will slip because we have D block then and again you can see multiple clusters will be generated but, this time what we are going to do is, we are going to leave away the forward strand initially what we did we had cleaved away and we have gone for the sequencing of good okay? . So, where the cleave Aveda original forward strand and this how we are ready for sequencing for the second. So, this time which primer will go in mind

Refer Slide Time: (18:20)



RD two you remember the structure of library there was a d2 region. So, there the read one sorry, read two primer will go and bind and just like read one we are said for the sequencing right



Refer Slide Time: (18:33)

Sequencing By Synthesis 2nd Read

3'- ...-5'

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

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

~~Cycle 2-n. Add sequencing reagents and repeat~~

18:37 / 30:00

all the reagent will flow this is how your sequence just like the first one but, the difference is here you are sequencing the reverse strand

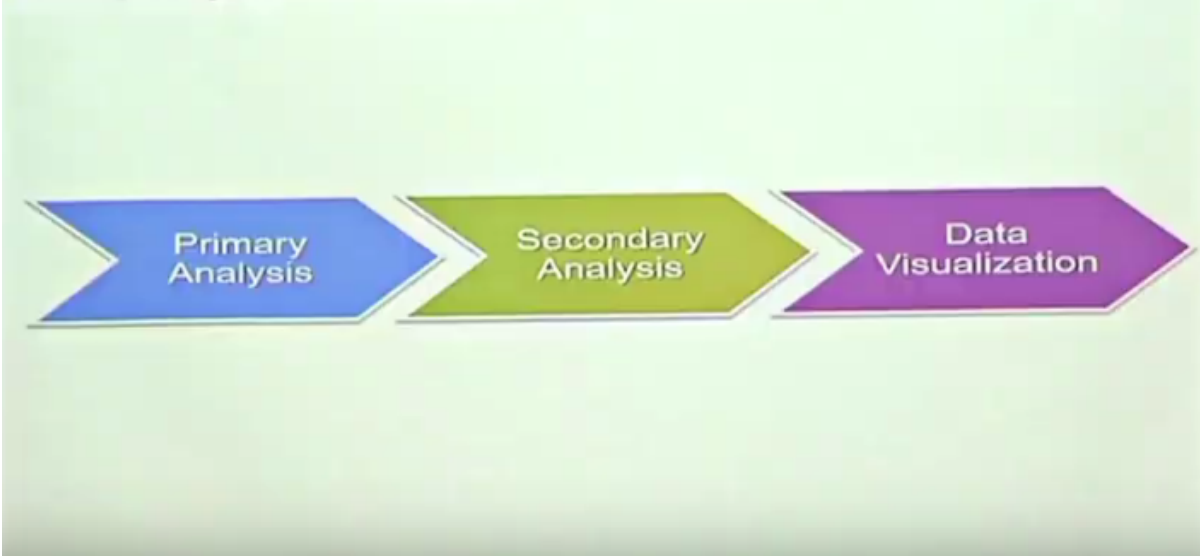
Refer Slide Time: (18:41)



this is all about the paid in sequencing. So, in all our platforms

Refer Slide Time: (18:44)

Data Analysis Overview



there's the simple logic. So, once we are done with the sequencing the final stages are data analysis okay? 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 the 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: (21:11)

Analysis Type	Software	Outputs
Sequencing	ICS/RTA	Images/TIFF files
Primary Analysis	ICS/RTA	Intensities Base Calling
Secondary Analysis	BaseSpace M/Sig Reporter CANVA	Alignments and Variant Detection

. 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's 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 a

real-time analysis software. So, it extract your intensities from the bases and convert those intensities into base calls. So, it will be getting? dot BC l5on 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 few files and finally using this machine tertiary analysis al. So, 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 past few files great . So, I will play a small video for you guys. So, I hope it will clarify everything's the same thing which I have shown. So, it will be shown sequentially

Video Start Time: (22:32)



clustering is a process wherein each fragment molecule is iso thermally amplified the flow cell is a glass slide with lanes each lane is a channel coated with a lawn composed of two types of oligos goes hybridization is enabled by the first of the two types of oligos goes on the surface this oligos go 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 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 ala goon the flow cell polymerases generate the complementary strand forming a double-stranded bridge this bridge is denatured resulting in two single-stranded copies of the molecule that are tethered to the flow cell the process is then repeated over and over and occurs simultaneously for millions of clusters resulting in clonal amplification of all the

fragments after bridge amplification the reverse strands are cleaved and washed off leaving only the forward strands the three prime ends are blocked to prevent unwanted priming sequencing begins with the extension of the first sequencing primer to produce the first read with each cycle fluorescently tagged nucleotides compete for addition to the growing chain 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 the number of cycles determines the length of the read the emission wave length 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 after completion of the index read the read product is washed off and the three prime ends of the template is d protected the template now folds over and binds the second al ago on the flow cell index to is read in the same manner as index one index to read product is washed off at the completion of this step polymerase Asst extend the second flow cell all ago forming a double-stranded bridge this double-stranded DNA is then linearized and the three prime ends blocked the original forward strand is cleaved off and washed away leaving the reverse strand read to begins with the introduction of the read to sequencing primer as with read one the sequencing steps are repeated until the desired read length is achieved the read to product is washed away 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 for each sample reads with similar stretches of base calls are locally clustered forward and reverse reads are paired creating contiguous sequences these contiguous sequences are aligned back to the reference genome for variant identification the paired end information is used to resolve ambiguous alignments

Video End Time: (26:42)



Refer Slide Time: (26:43)

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

all right! So, I'm sure by now you are very clear and convinced about the magic which next generation sequencing platforms have done for us the place the Icarus II they speed the cost what one could accomplish by this platform was not even possible to think ten 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 fraud
NGO sit 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 net 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 through put scalability and speed the
engine technology enables researchers to perform a wide variety of applications and Esteli balanced car
systems at a level which was never possible before today I hope you have learnt about the basics of
engines 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'll continue our discussion in the next lecture as well
thank you