An Introduction to Proteogenomics

Dr. Sanjeeva Srivastava
Dr. Aarthi Desai
Department of Biosciences and Bioengineering
Illumina India
Indian Institute of Technology, Bombay

Supplementary - 3
Sequencing by Synthesis I

Welcome to MOOC course on Introduction to Proteogenomics. Next generation sequencing has really seen large applications especially in the clinical settings. It is a really good idea to catch up on use of new NGS platforms. In this slide, we have invited industry experts to provide you the hands on session how to use the latest NGS technologies.

Today, Dr. Aarti Desai from Illumina will provide you a brief lecture on sequencing technology, especially sequencing by synthesis. She will also give you an introduction of how sequencing technology actually works. She will talk about two key concepts, sequencing by synthesis and paired end sequencing. Finally, she will show you how to open an account in base space account and they proceed for the hands on session in the next lecture. So, let us welcome Dr. Aarti Desai from Illumina.

Before we actually get started with the hands on session, I wanted to show you guys a short video that recapitulates the Illumina sequencing technology. Mukesh did a great job of explaining all the platforms that we have and the key oncology applications and the panels that are currently available from Illumina but we are not sure on you know whether the concept of reads, the read length, pair end sequencing, depth of sequencing if all of that is you know known to everybody. So, before we start the actual hands on session which is going to be short what we would like to do is you know just give you a brief understanding of how the Illumina sequencing technology actually works.

(Refer Slide Time: 02:29)

Sample Prep

Cluster Generation

Sequencing

Data Analysis

And this there are plenty of videos available on YouTube; we have just picked one of the you know the one that really quickly and easily demonstrates how the sequencing technology works.

(Refer Slide Time: 02:34)

Sample Prep

Sample preparation begins with extracted and purified DNA.

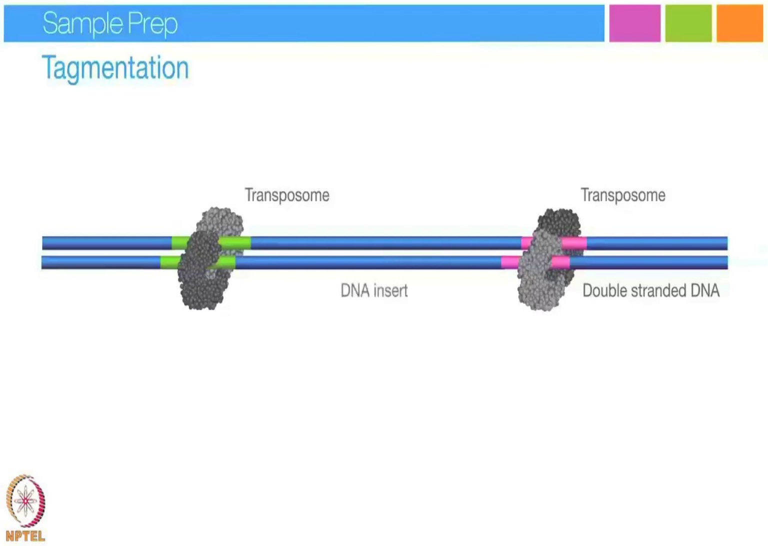
(Refer Slide Time: 02:42)

Sample Prep

Nextera

The first step in Nextera sample preparation is Tagmentation.

(Refer Slide Time: 02:44)

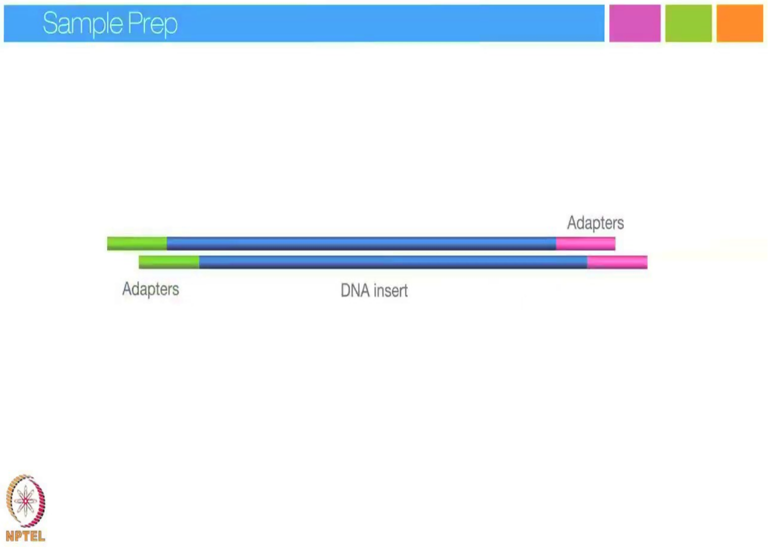


Sample Prep

Tagmentation

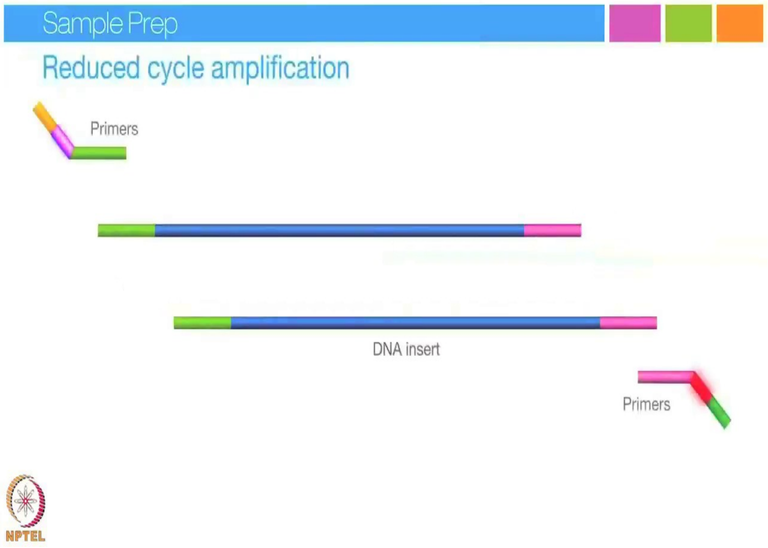
During tagmentation, transposomes simultaneously fragment and tag the input DNA with adapters.

(Refer Slide Time: 02:52)



Sample Prep

(Refer Slide Time: 02:54)

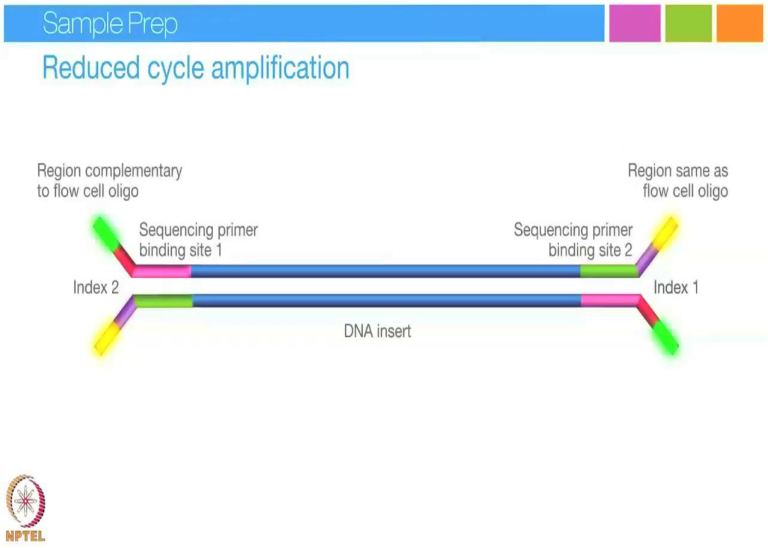


Sample Prep

Reduced cycle amplification

Once the adapters have been ligated reduced cycle amplification adds additional motifs.

(Refer Slide Time: 02:59)



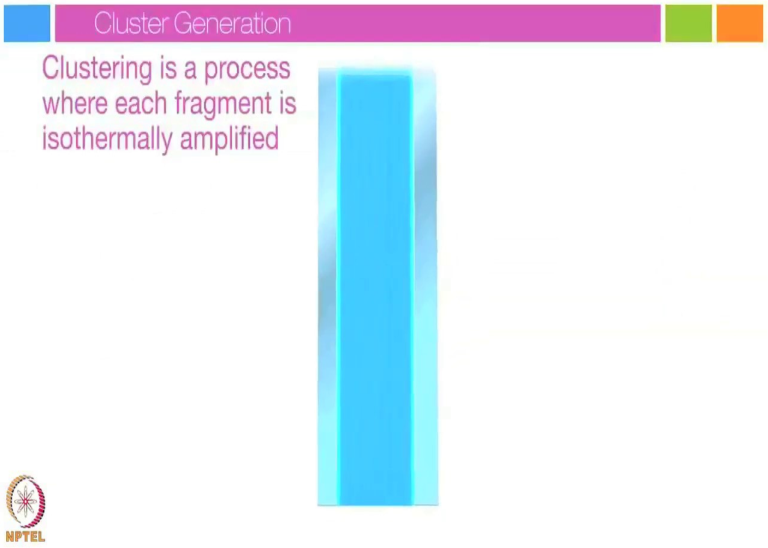
Sample Prep

Reduced cycle amplification

Region complementary to flow cell oligo
Region same as flow cell oligo
Sequencing primer binding site 1
Sequencing primer binding site 2
Index 2
Index 1
DNA insert
NPTEL

Such as this sequencing primer binding sites, indices and regions that are complementary to the flow cell oligos.

(Refer Slide Time: 03:08)

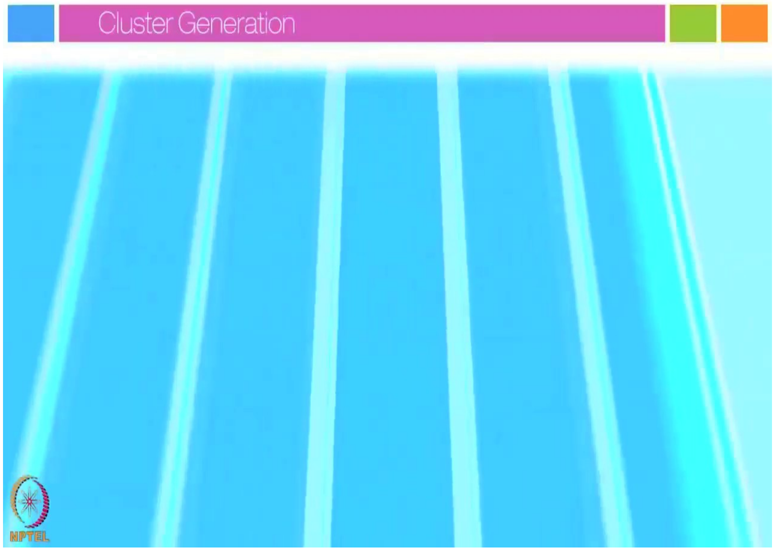


Cluster Generation

Clustering is a process where each fragment is isothermally amplified

Clustering is a process wherein each fragment molecule is isothermally amplified.

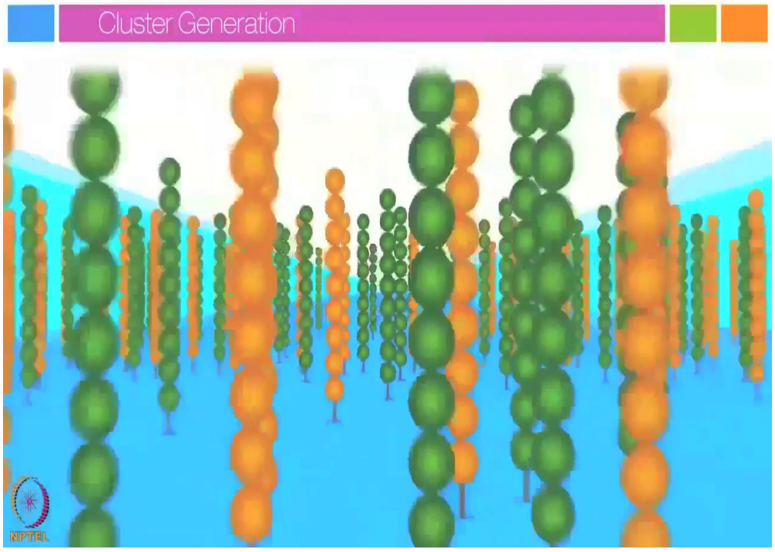
(Refer Slide Time: 03:15)



Cluster Generation

The flow cell is a glass slide with lanes.

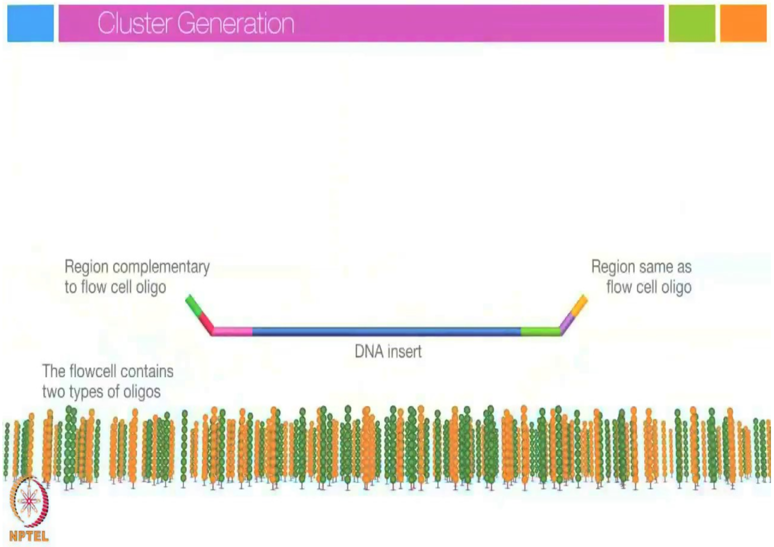
(Refer Slide Time: 03:17)



Cluster Generation

Each lane is a channel coated with a lawn composed of two types of oligos.

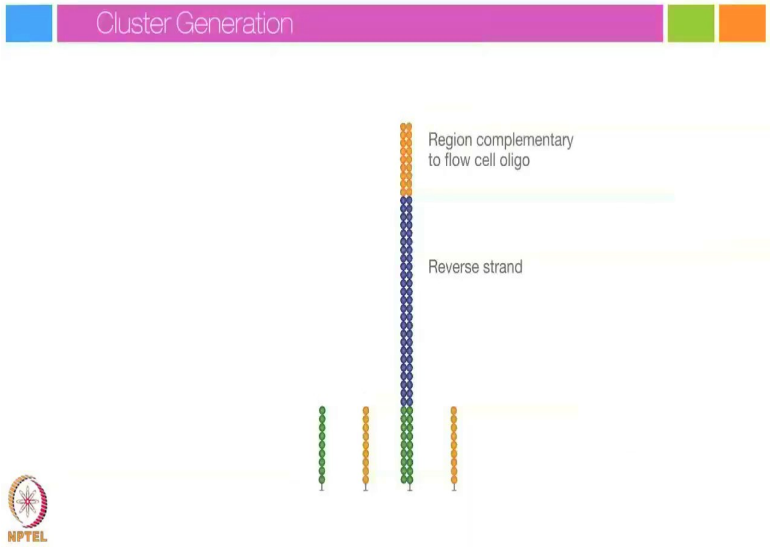
(Refer Slide Time: 03:23)


Cluster Generation

Region complementary to flow cell oligo
Region same as flow cell oligo
DNA insert
The flowcell contains two types of oligos
NPTEL

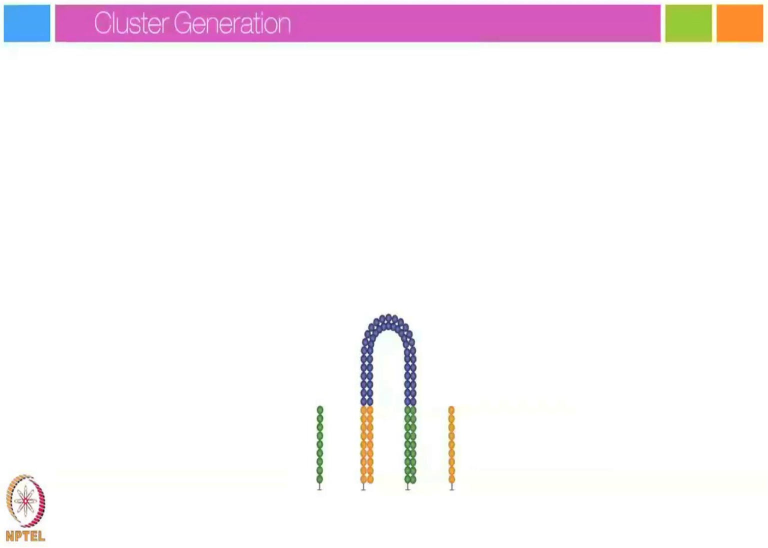
Hybridization is enabled by the first of the two types of oligos on the surface.

(Refer Slide Time: 03:26)


Cluster Generation

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 and the original template is washed away.

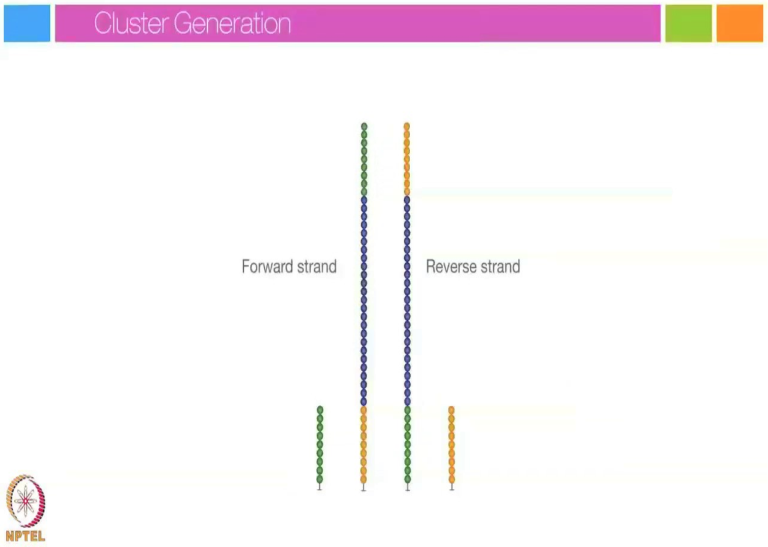
(Refer Slide Time: 03:41)



Cluster Generation

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: 04:01)



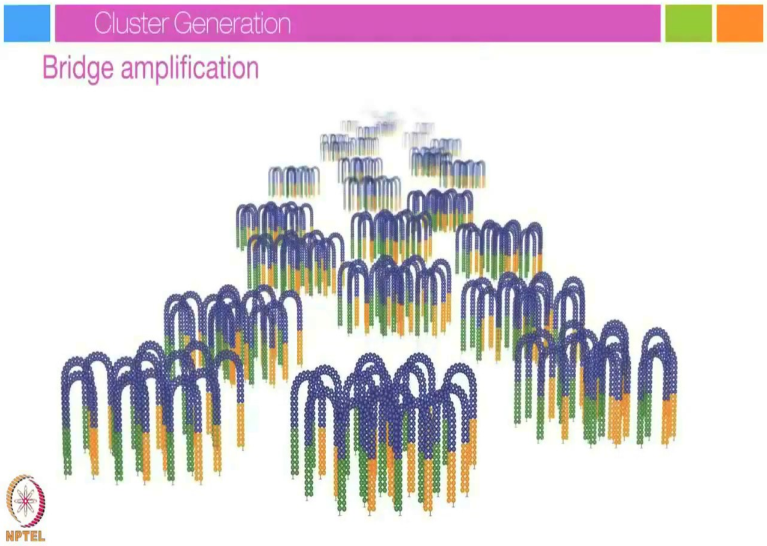
Cluster Generation

Forward strand

Reverse strand

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

(Refer Slide Time: 04:07)

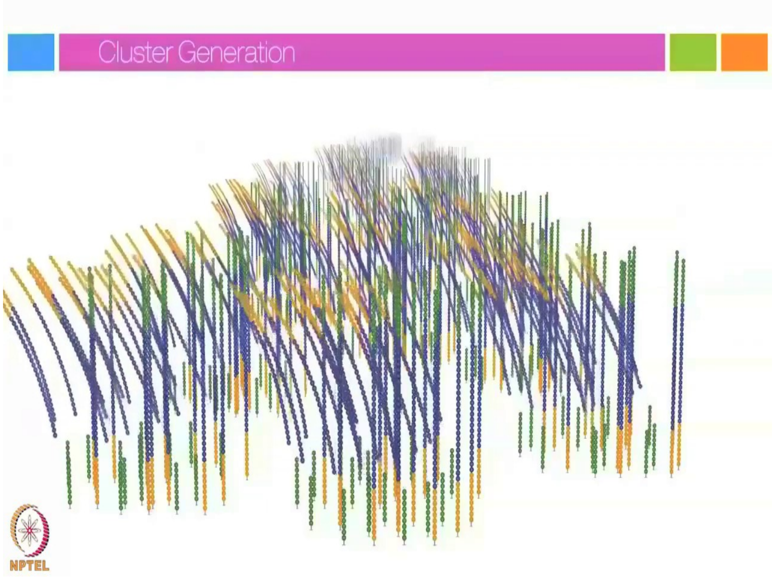


Cluster Generation

Bridge amplification

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

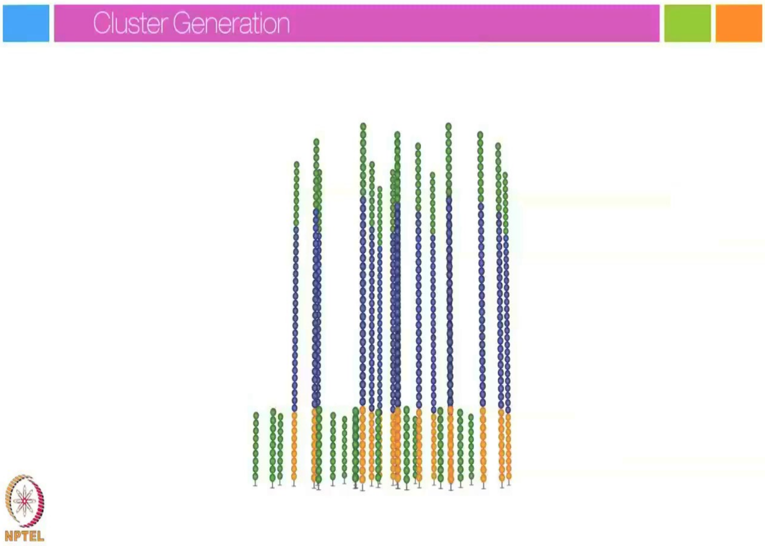
(Refer Slide Time: 04:19)



Cluster Generation

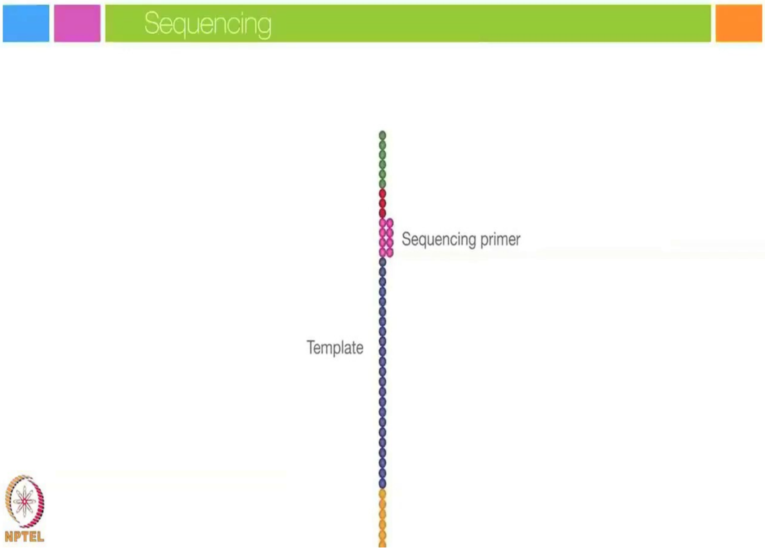
After bridge amplification the reverse strands are cleaved and washed off, leaving only the forward strands.

(Refer Slide Time: 04:25)


Cluster Generation

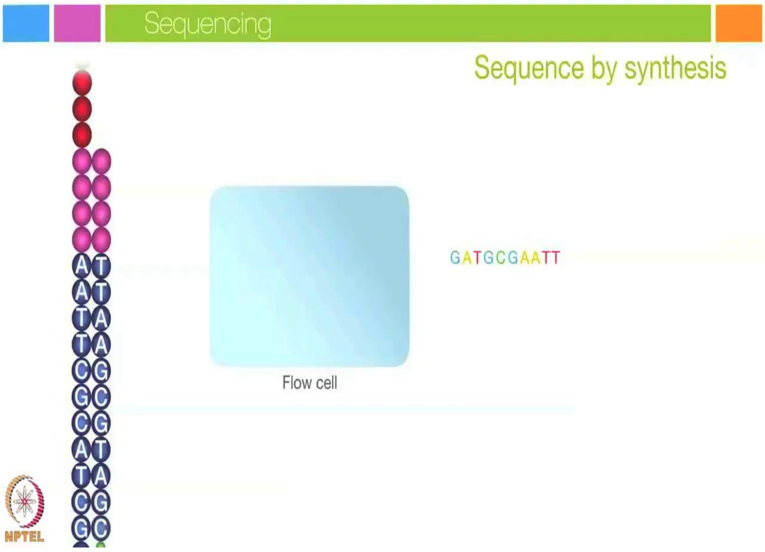
The three prime ends are blocked to prevent unwanted priming.

(Refer Slide Time: 04:33)


Sequencing

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

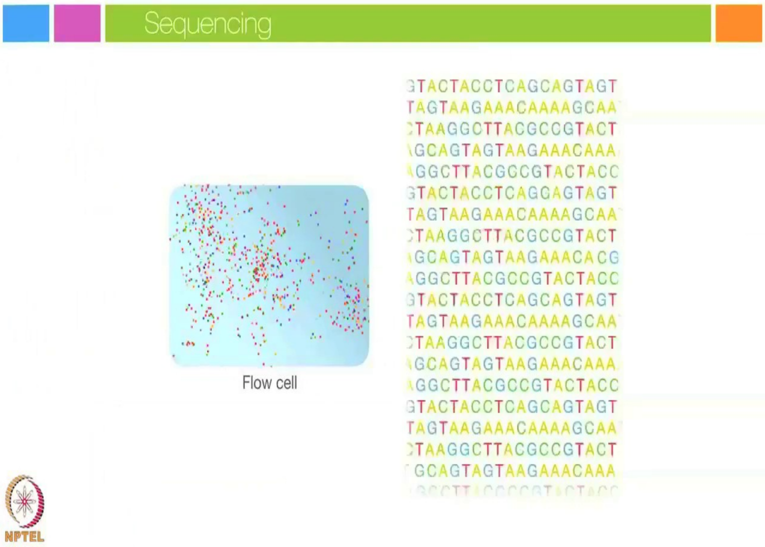
(Refer Slide Time: 04:41)



With each cycle four 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.

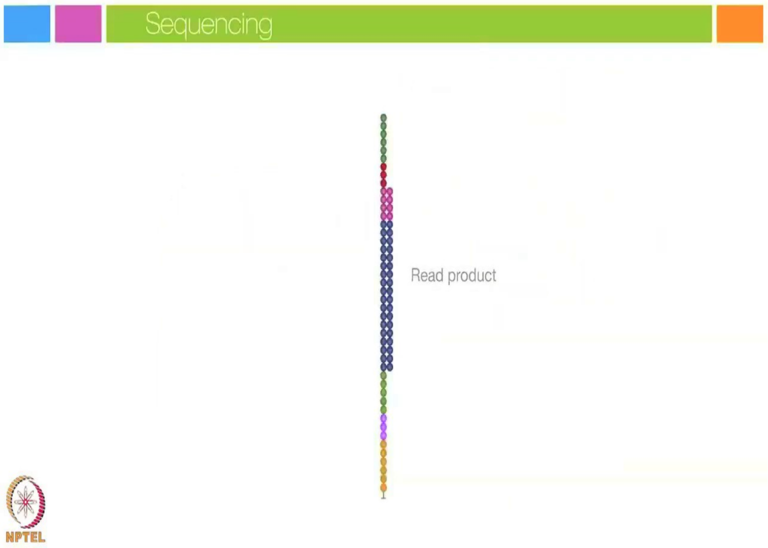
The number of cycles determines the length of the read. The emission wavelength along with the signal intensity determined the base call, for a given cluster all identical strands are read simultaneously.

(Refer Slide Time: 05:19)



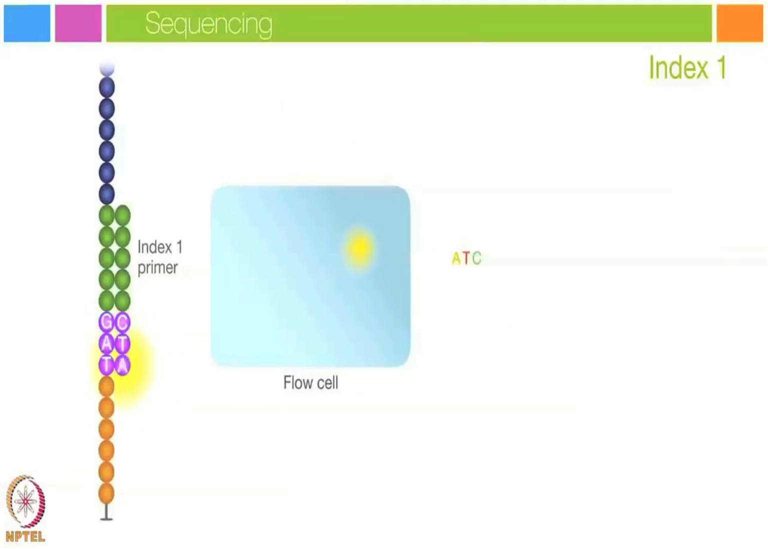
Hundreds of millions of clusters are sequenced in a massively parallel process. This image represents a small fraction of the flow cell.

(Refer Slide Time: 05:29)



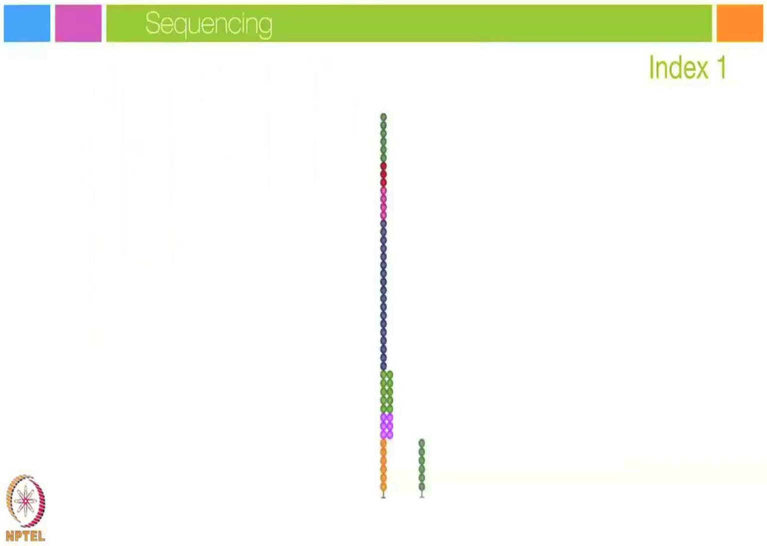
After the completion of the first read, the read product is washed away.

(Refer Slide Time: 05:36)



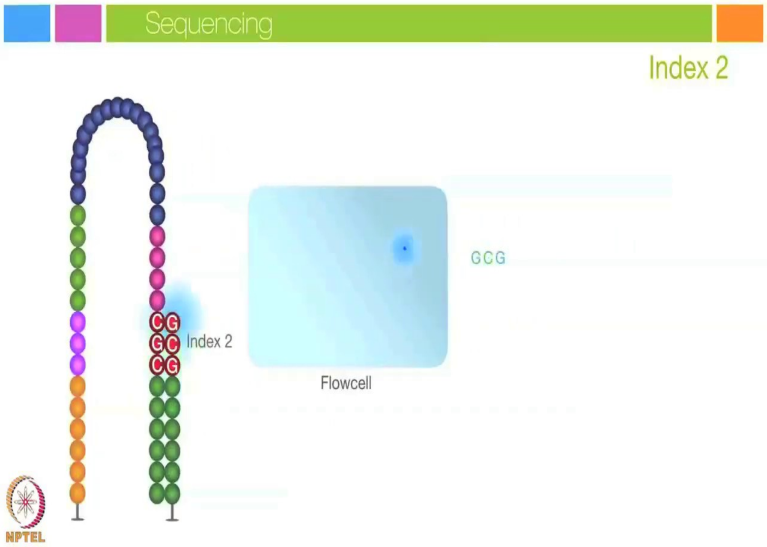
In this step, the index one read primer is introduced and hybridized to the template. The read is generated similar to the first read.

(Refer Slide Time: 05:45)



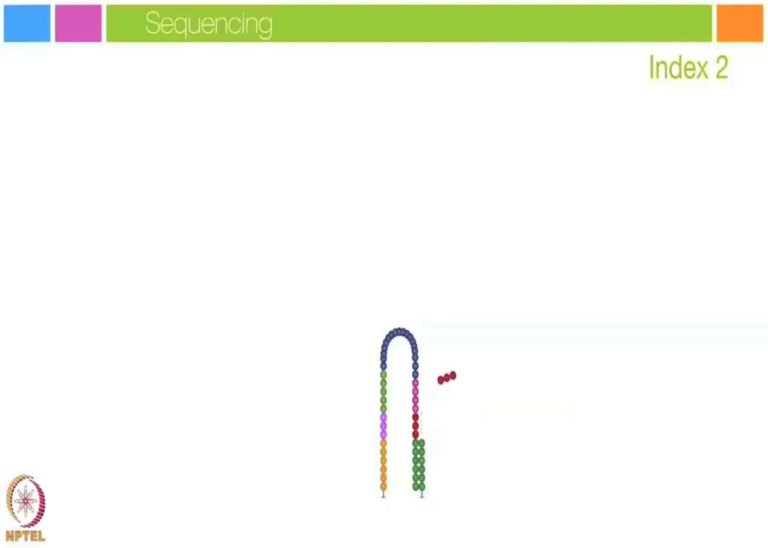
After completion of the index read, the read product is washed off and the three prime end of the template is de-protected.

(Refer Slide Time: 05:52)



The template now folds over and binds the second oligo on the flow cell. Index 2 is read in the same manner as index 1.

(Refer Slide Time: 06:03)

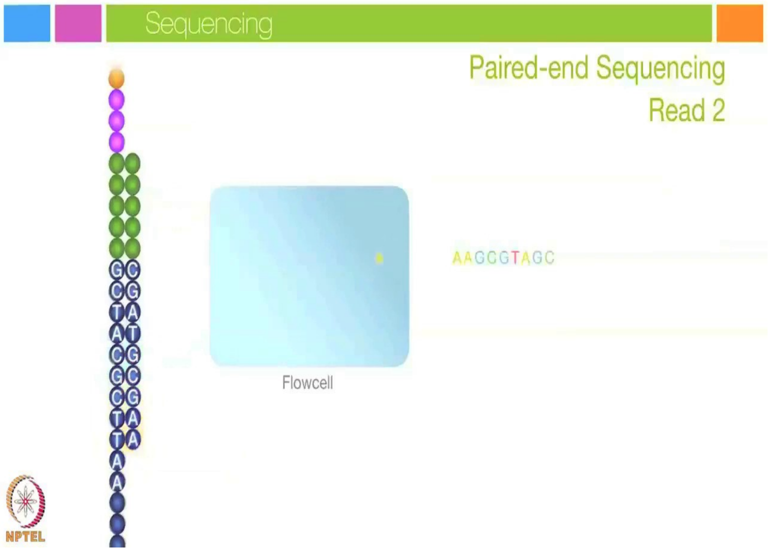


Sequencing

Index 2

Index 2, read product is washed off at the completion of this step.

(Refer Slide Time: 06:08)



Sequencing

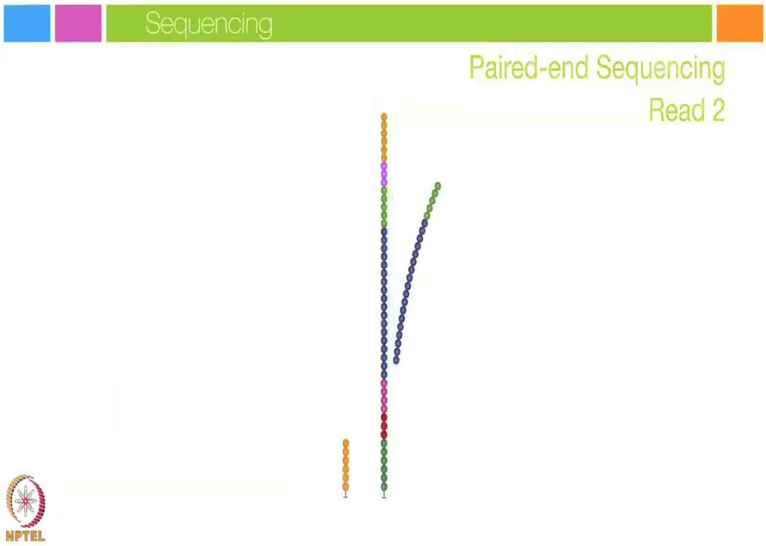
Paired-end Sequencing

Read 2

Polymerases extend the second flow cell oligo 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.

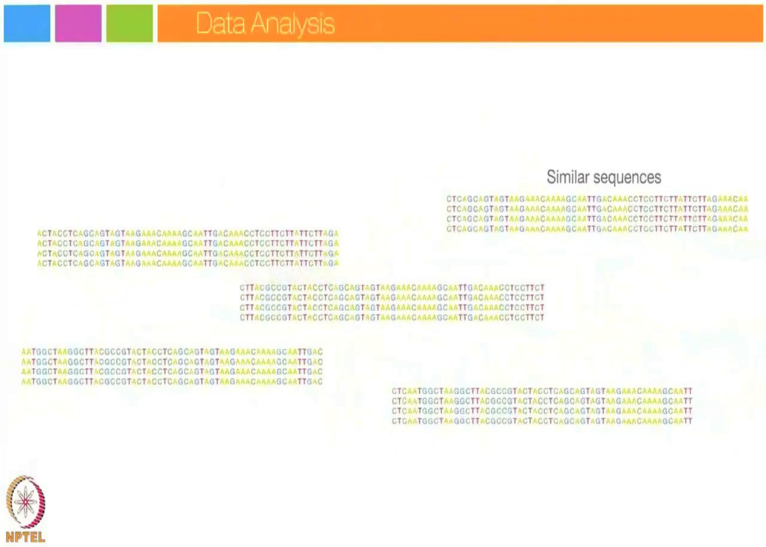
Read 2 begins with the introduction of the read 2 sequencing primer. As with read 1, the sequencing steps are repeated until the desired read length is achieved.

(Refer Slide Time: 06:35)



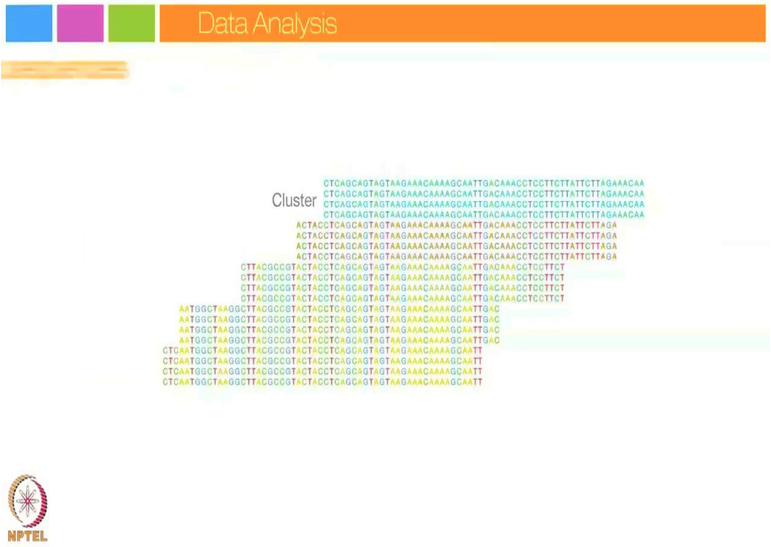
The read 2 product is washed away. This entire process generates billions of reads representing all the fragments.

(Refer Slide Time: 06:45)


Data Analysis
Similar sequences
CTCAGCAGTAGTAAGAAACAAAAGCAATTGACAAACCTCCTTCTTATTCTTAGAAACAA
CTCAGCAGTAGTAAGAAACAAAAGCAATTGACAAACCTCCTTCTTATTCTTAGAAACAA
CTCAGCAGTAGTAAGAAACAAAAGCAATTGACAAACCTCCTTCTTATTCTTAGAAACAA
CTCAGCAGTAGTAAGAAACAAAAGCAATTGACAAACCTCCTTCTTATTCTTAGAAACAA
ACTACCTCAGCAGTAGTAAGAAACAAAAGCAATTGACAAACCTCCTTCTTATTCTTAGA
ACTACCTCAGCAGTAGTAAGAAACAAAAGCAATTGACAAACCTCCTTCTTATTCTTAGA
ACTACCTCAGCAGTAGTAAGAAACAAAAGCAATTGACAAACCTCCTTCTTATTCTTAGA
ACTACCTCAGCAGTAGTAAGAAACAAAAGCAATTGACAAACCTCCTTCTTATTCTTAGA
CTTACGCCGTAGTACCTCAGCAGTAGTAAGAAACAAAAGCAATTGACAAACCTCCTTCT
CTTACGCCGTACTACCTCAGCAGTAGTAAGAAACAAAAGCAATTGACAAACCTCCTTCT
CTTACGCCGTACTACCTCAGCAGTAGTAAGAAACAAAAGCAATTGACAAACCTCCTTCT
CTTACGCCGTACTACCTCAGCAGTAGTAAGAAACAAAAGCAATTGACAAACCTCCTTCT
AATGGCTAAGGCTTACGCCGTACTACCTCAGCAGTAGTAAGAAACAAAAGCAATTGAC
AATGGCTAAGGCTTACGCCGTACTACCTCAGCAGTAGTAAGAAACAAAAGCAATTGAC
AATGGCTAAGGCTTACGCCGTACTACCTCAGCAGTAGTAAGAAACAAAAGCAATTGAC
AATGGCTAAGGCTTACGCCGTACTACCTCAGCAGTAGTAAGAAACAAAAGCAATTGAC
CTCAATGGCTAAGGCTTACGCCGTACTACCTCAGCAGTAGTAAGAAACAAAAGCAATT
CTCAATGGCTAAGGCTTACGCCGTACTACCTCAGCAGTAGTAAGAAACAAAAGCAATT
CTCAATGGCTAAGGCTTACGCCGTACTACCTCAGCAGTAGTAAGAAACAAAAGCAATT
CTCAATGGCTAAGGCTTACGCCGTACTACCTCAGCAGTAGTAAGAAACAAAAGCAATT
NPTEL

Sequences from pooled sample libraries are separated based on the unique indices introduced during the sample preparation.

(Refer Slide Time: 06:52)

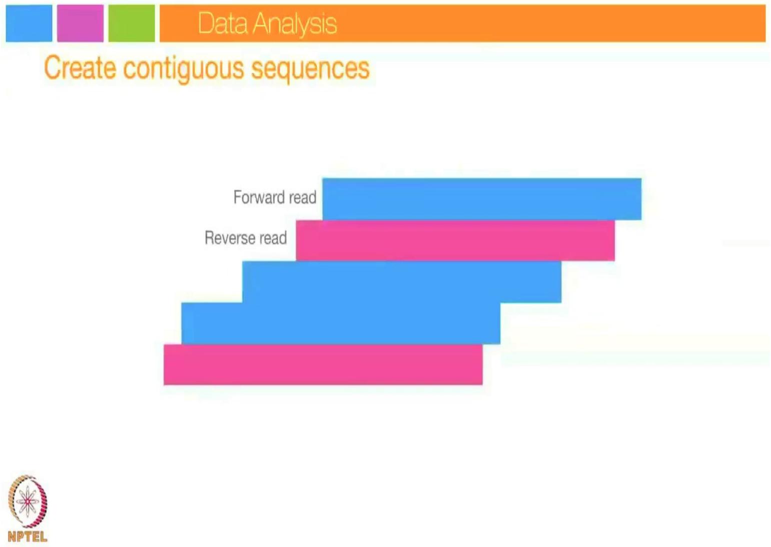


Data Analysis

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

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

(Refer Slide Time: 07:00)

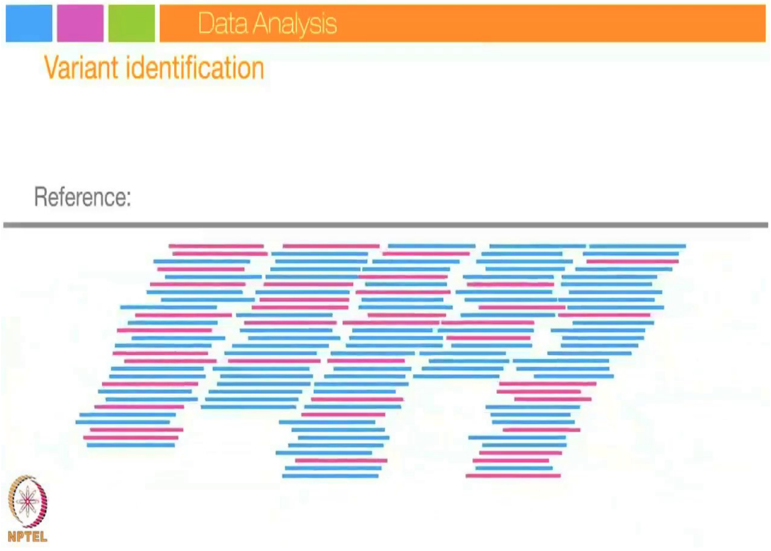


Data Analysis

Create contiguous sequences

Forward and reverse reads are paired, creating contiguous sequences.

(Refer Slide Time: 07:06)



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

So, two key concepts were sort of the take away of that video, one is the chemistry that is used for generating data on the Illumina sequencers, what is known as sequencing by synthesis, right. So, as you saw we actually add one base at a time and record that base. So, we are literally reading one base at a time right, which is why we have very high accuracy in our data set.

The second one was the paired end synthesis, right. So, we are essentially using the same fragment of DNA that we are using in our library prep, to read it from two ends. We read it from the forward end and then we read it from the reverse end, so which is why a lot of the Illumina data that you will see will have two reads for every fragment it is called R1 and R2, so read 1 and read 2 and as was alluded to in the video what that gives you is it gives you again very high confidence in the data that you are generating, primarily because the fragments are short. The fragments are about you know 150 to 200 base pairs and when you read them from both ends you have overlap, right.

So, your; your the chances of you reading one base you know more than or rather twice every time you read it is very high, right. So, you have again very high confidence in the data that you generate and because again you are reading it from both ends it is very useful for certain applications like translocations or you know deletions or insertions, then so on and so forth, because the distance between the two reads is fixed, right. So, every time you map it back to the reference genome if there is any deviation from this fixed length you can infer that there may be a gen a structural variant in that particular region of the genome, right.

So, it is a very very powerful way of sequencing genomic DNA and as you can imagine, it is the leading provider of sequencing technology globally today. More than 90 percent of the data that is available in public databases comes from Illumina technology and it is not only true for research but it is also very much true for clinical applications, ok.

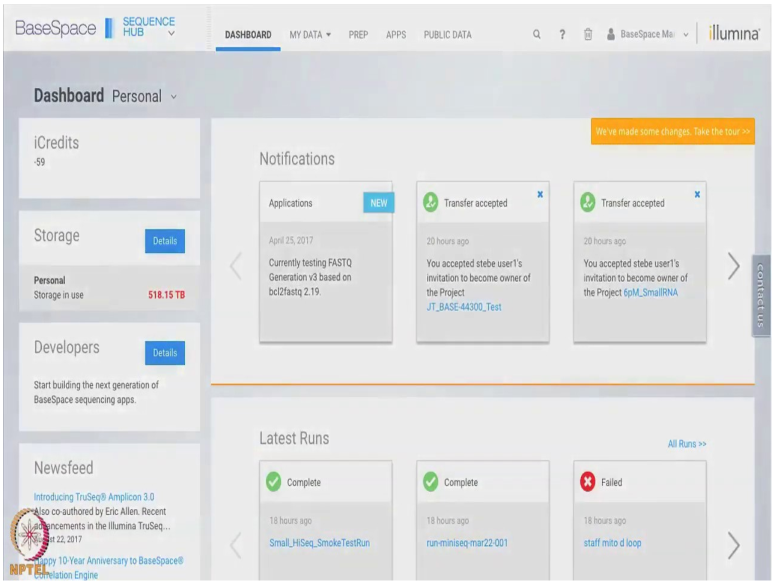
We ended the video on the base space application. So, I wanted to take you guys to base space if I can.

(Refer Slide Time: 09:38)

BaseSpace
SEQUENCE HUB
illumina
NPTEL
For Research Use Only. Not for use in diagnostic procedures.

Base space sequence hub is a luminous cloud based next generation sequencing platform that performs automated sample to result workflows for your lab.

(Refer Slide Time: 09:49)



Dashboard Personal

iCredits
-59

Storage Details
Personal
Storage in use 518.15 TB

Developers Details
Start building the next generation of BaseSpace sequencing apps.

Newsfeed
Introducing TruSeq® Amplicon 3.0
Also co-authored by Eric Allen. Recent advancements in the Illumina TruSeq...
...st 22, 2017
...ppy 10-Year Anniversary to BaseSpace® Correlation Engine

We've made some changes. Take the tour >>

Notifications

Applications NEW
April 25, 2017
Currently testing FASTQ Generation v3 based on bcl2fastq 2.19.

Transfer accepted
20 hours ago
You accepted stebe user1's invitation to become owner of the Project JT_BASE-44300_Test

Transfer accepted
20 hours ago
You accepted stebe user1's invitation to become owner of the Project 6pM_SmallRNA

Latest Runs
All Runs >>

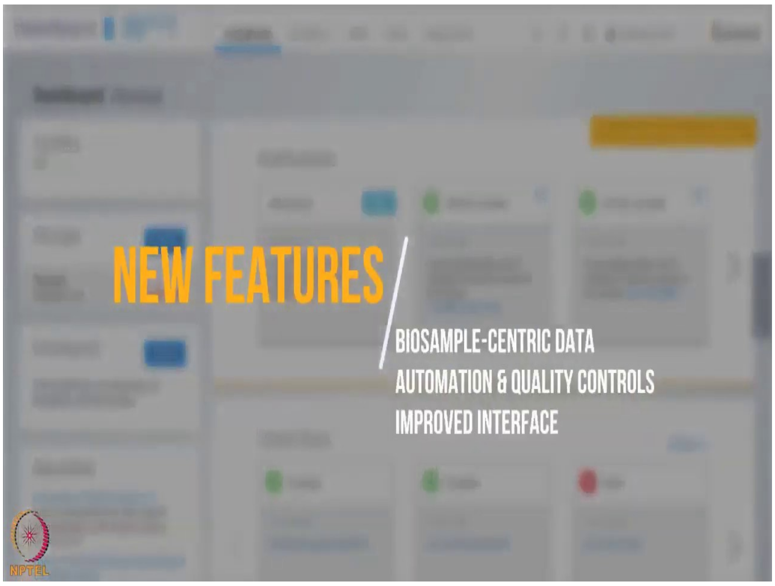
Complete
18 hours ago
Small_HiSeq_SmokeTestRun

Complete
18 hours ago
run-miniseq-mar22-001

Failed
18 hours ago
staff mito d loop

We recently released a number of new features designed to enhance your laboratories efficiency.

(Refer Slide Time: 09:55)

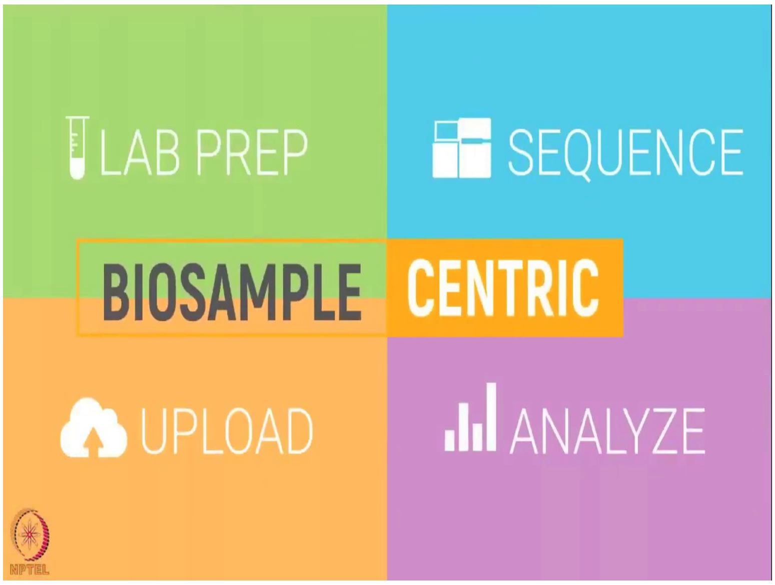


NEW FEATURES

BIOSAMPLE-CENTRIC DATA
AUTOMATION & QUALITY CONTROLS
IMPROVED INTERFACE

Including, a new biosample-centric data model that enables easy tracking of all bio sample activity from lab preparation through analysis delivery, new automation and quality control features to streamline the efficiency and consistency of your workflows, and an improved interface that helps you access your data and perform functions more quickly.

(Refer Slide Time: 10:16)



LAB PREP

SEQUENCE

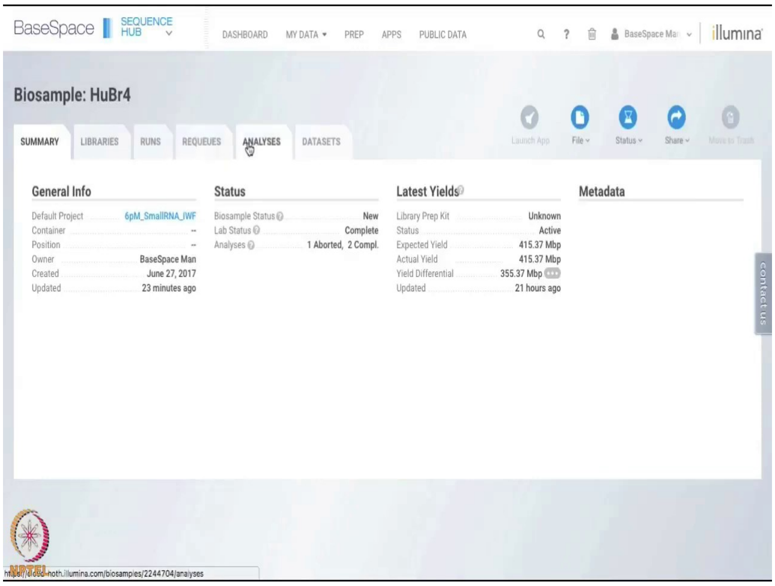
BIOSAMPLE CENTRIC

UPLOAD

ANALYZE

The new bio-sample centric data model that you easily track all bio-sample activity from lab preparation to sequencing to processing and upload enough data to the cloud and analysing results.

(Refer Slide Time: 10:30)



Biosample: HuBr4

General Info

Default Project 6pM_SmallRNA_IWF
Container --
Position --
Owner BaseSpace Man
Created June 27, 2017
Updated 23 minutes ago

Status

Biosample Status New
Lab Status Complete
Analyses 1 Aborted, 2 Compl.

Latest Yields

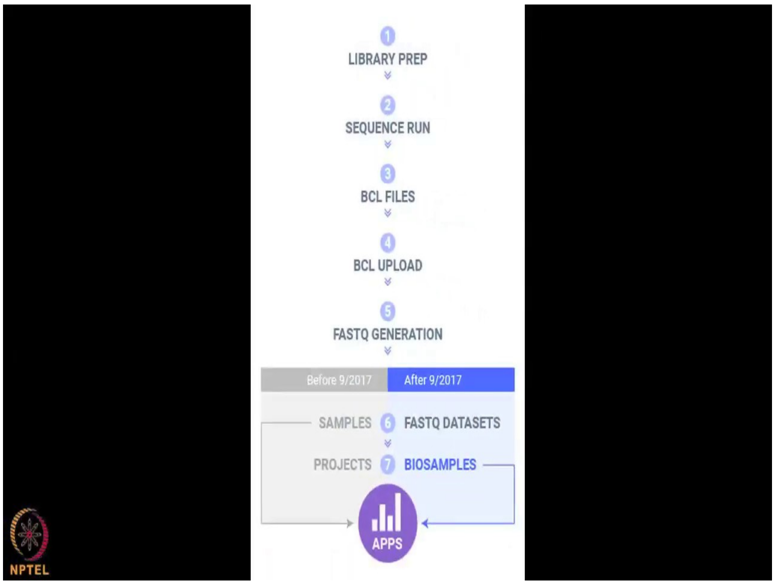
Library Prep Kit Unknown
Status Active
Expected Yield 415.37 Mbp
Actual Yield 415.37 Mbp
Yield Differential 355.37 Mbp
Updated 21 hours ago

Metadata

https://[illegible]noth.illumina.com/biosamples/2244704/analyses

Bio-samples support data aggregation which can be linked to multiple libraries, runs, requeues, analysis and can have multiple data sets associated with them.

(Refer Slide Time: 10:41)



1 LIBRARY PREP

2 SEQUENCE RUN

3 BCL FILES

4 BCL UPLOAD

5 FASTQ GENERATION

Before 9/2017 | After 9/2017

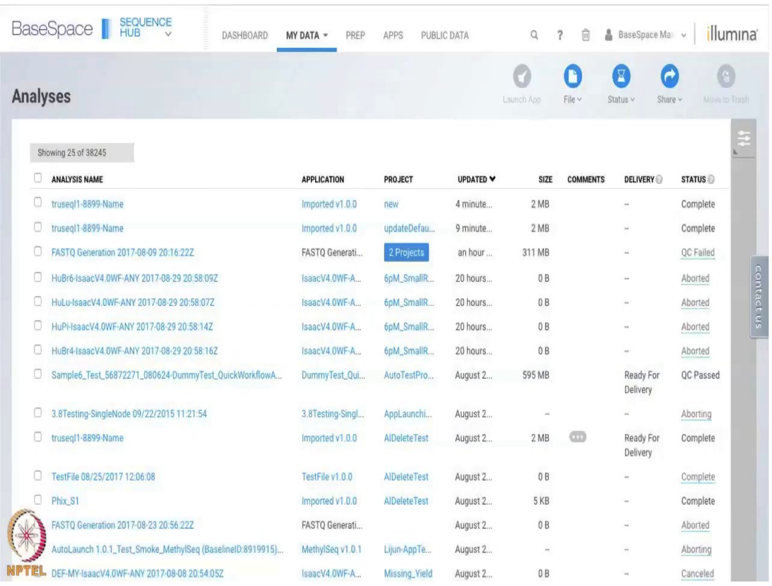
SAMPLES 6 FASTQ DATASETS

PROJECTS 7 BIOSAMPLES

APPS

FASTQ data sets have replaced samples and are now stored inside of bio-samples. Your existing samples have already been converted, so that you can use the new file types as inputs on launching apps.

(Refer Slide Time: 10:53)

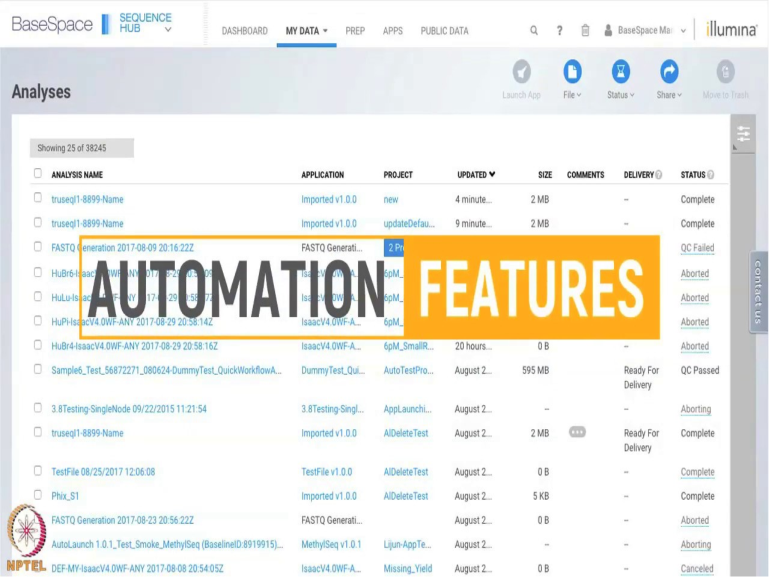


Analyses

Showing 25 of 38245

ANALYSIS NAME	APPLICATION	PROJECT	UPDATED	SIZE	COMMENTS	DELIVERY	STATUS
truseql1-8899-Name	Imported v1.0.0	new	4 minute...	2 MB		–	Complete
truseql1-8899-Name	Imported v1.0.0	updateDefau...	9 minute...	2 MB		–	Complete
FASTQ Generation 2017-08-09 20:16:22Z	FASTQ Generati...	2 Projects	an hour ...	311 MB		–	QC Failed
HuBr6-IsaacV4.0WF-ANY 2017-08-29 20:58:09Z	IsaacV4.0WF-A...	6pM_SmallR...	20 hours...	0 B		–	Aborted
HuLu-IsaacV4.0WF-ANY 2017-08-29 20:58:07Z	IsaacV4.0WF-A...	6pM_SmallR...	20 hours...	0 B		–	Aborted
HuPi-IsaacV4.0WF-ANY 2017-08-29 20:58:14Z	IsaacV4.0WF-A...	6pM_SmallR...	20 hours...	0 B		–	Aborted
HuBr4-IsaacV4.0WF-ANY 2017-08-29 20:58:16Z	IsaacV4.0WF-A...	6pM_SmallR...	20 hours...	0 B		–	Aborted
Sample6_Test_56872271_080624-DummyTest_QuickWorkflowA...	DummyTest_Qui...	AutoTestPro...	August 2...	595 MB		Ready For Delivery	QC Passed
3.8Testing-SingleNode 09/22/2015 11:21:54	3.8Testing-Singl...	AppLaunchi...	August 2...	–		–	Aborting
truseql1-8899-Name	Imported v1.0.0	AlDeleteTest	August 2...	2 MB	…	Ready For Delivery	Complete
TestFile 08/25/2017 12:06:08	TestFile v1.0.0	AlDeleteTest	August 2...	0 B		–	Complete
Phix_S1	Imported v1.0.0	AlDeleteTest	August 2...	5 KB		–	Complete
FASTQ Generation 2017-08-23 20:56:22Z	FASTQ Generati...		August 2...	0 B		–	Aborted
AutoLaunch 1.0.1_Test_Smoke_MethylSeq (BaselineID:8919915)...	MethylSeq v1.0.1	Lijun-AppTe...	August 2...	–		–	Aborting
DEF-MY-IsaacV4.0WF-ANY 2017-08-08 20:54:05Z	IsaacV4.0WF-A...	Missing_Yield	August 2...	0 B		–	Canceled

(Refer Slide Time: 10:56)



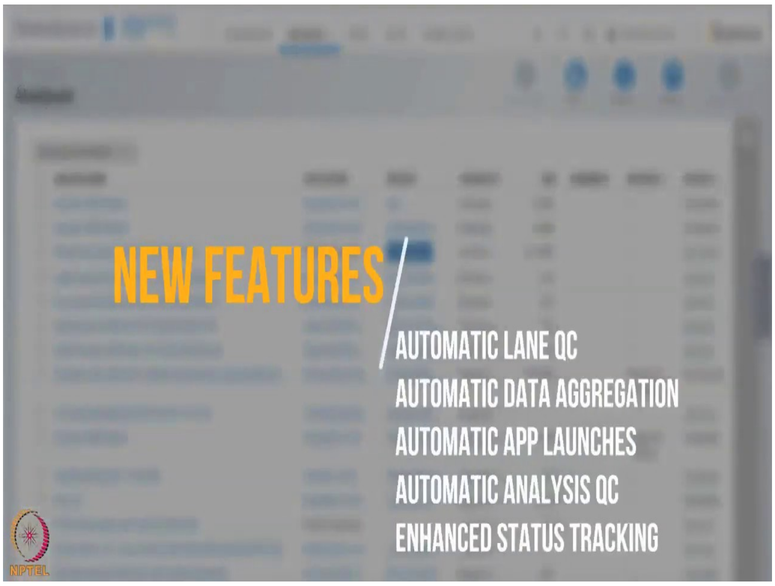
Analyses

Showing 25 of 38245

ANALYSIS NAME	APPLICATION	PROJECT	UPDATED	SIZE	COMMENTS	DELIVERY	STATUS
truseql1-8899-Name	Imported v1.0.0	new	4 minute...	2 MB		--	Complete
truseql1-8899-Name	Imported v1.0.0	updateDefau...	9 minute...	2 MB		--	Complete
FASTQ Generation 2017-08-09 20:16:22Z	FASTQ Generati...	2 Pr	[illegible]	[illegible]		[illegible]	QC Failed
HuBr6-Isaac[illegible]	Isa[illegible]	6pM_	[illegible]	[illegible]		[illegible]	Aborted
HuLu-Is[illegible]	Isa[illegible]	6pM_	[illegible]	[illegible]		[illegible]	Aborted
HuPi-IsaacV4.0WF-ANY 2017-08-29 20:58:14Z	IsaacV4.0WF-A...	6pM_	[illegible]	[illegible]		[illegible]	Aborted
HuBr4-IsaacV4.0WF-ANY 2017-08-29 20:58:16Z	IsaacV4.0WF-A...	6pM_SmallR...	20 hours...	0 B		--	Aborted
Sample6_Test_56872271_080624-DummyTest_QuickWorkflowA...	DummyTest_Qui...	AutoTestPro...	August 2...	595 MB		Ready For Delivery	QC Passed
3.8Testing-SingleNode 09/22/2015 11:21:54	3.8Testing-Singl...	AppLaunchi...	August 2...	--		--	Aborting
truseql1-8899-Name	Imported v1.0.0	AlDeleteTest	August 2...	2 MB	...	Ready For Delivery	Complete
TestFile 08/25/2017 12:06:08	TestFile v1.0.0	AlDeleteTest	August 2...	0 B		--	Complete
Phix_S1	Imported v1.0.0	AlDeleteTest	August 2...	5 KB		--	Complete
FASTQ Generation 2017-08-23 20:56:22Z	FASTQ Generati...		August 2...	0 B		--	Aborted
AutoLaunch 1.0.1_Test_Smoke_MethylSeq (BaselineID:8919915)...	MethylSeq v1.0.1	Lijun-AppTe...	August 2...	--		--	Aborting
DEF-MY-IsaacV4.0WF-ANY 2017-08-08 20:54:05Z	IsaacV4.0WF-A...	Missing_Yield	August 2...	0 B		--	Canceled

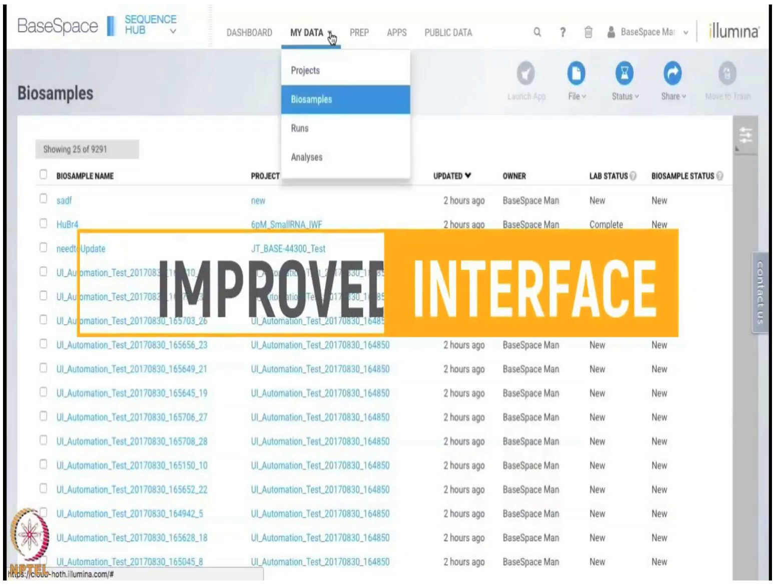
Base space sequence up includes new features that allow you to automate analysis workflows, reducing the time it takes to process samples and eliminating costly errors.

(Refer Slide Time: 11:05)



New features include automatic lane QC settings, automatic data aggregation, automatic app launches, automatic analysis QC settings, and enhanced status tracking.

(Refer Slide Time: 11:18)



Biosamples

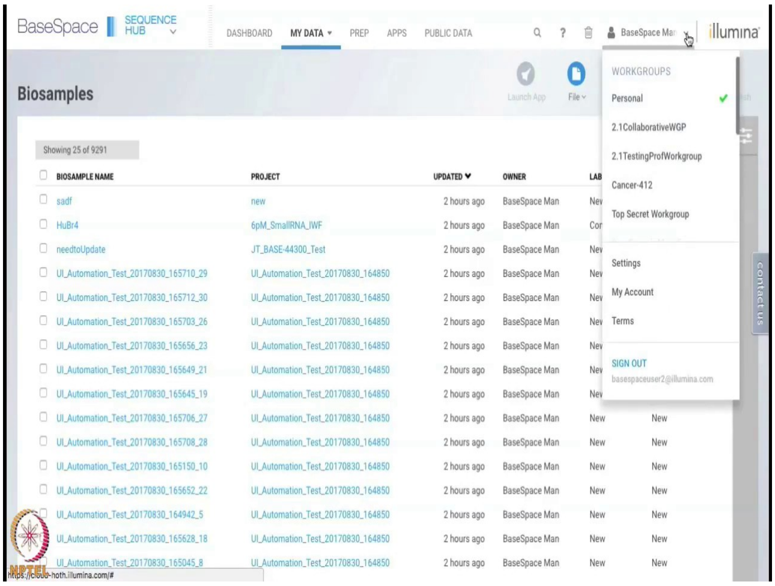
Showing 25 of 9291

BIOSAMPLE NAME	PROJECT	UPDATED	OWNER	LAB STATUS	BIOSAMPLE STATUS
sadf	new	2 hours ago	BaseSpace Man	New	New
HuBr4	6pM_SmallRNA_IWF	2 hours ago	BaseSpace Man	Complete	New
needtoUpdate	JT_BASE-44300_Test	[illegible]	[illegible]	[illegible]	[illegible]
UI_Automation_Test_2017083[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
UI_Automation_Test_2017083[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
UI_Automation_Test_20170830_165703_26	UI_Automation_Test_20170830_16485	[illegible]	[illegible]	[illegible]	[illegible]
UI_Automation_Test_20170830_165656_23	UI_Automation_Test_20170830_164850	2 hours ago	BaseSpace Man	New	New
UI_Automation_Test_20170830_165649_21	UI_Automation_Test_20170830_164850	2 hours ago	BaseSpace Man	New	New
UI_Automation_Test_20170830_165645_19	UI_Automation_Test_20170830_164850	2 hours ago	BaseSpace Man	New	New
UI_Automation_Test_20170830_165706_27	UI_Automation_Test_20170830_164850	2 hours ago	BaseSpace Man	New	New
UI_Automation_Test_20170830_165708_28	UI_Automation_Test_20170830_164850	2 hours ago	BaseSpace Man	New	New
UI_Automation_Test_20170830_165150_10	UI_Automation_Test_20170830_164850	2 hours ago	BaseSpace Man	New	New
UI_Automation_Test_20170830_165652_22	UI_Automation_Test_20170830_164850	2 hours ago	BaseSpace Man	New	New
UI_Automation_Test_20170830_164942_5	UI_Automation_Test_20170830_164850	2 hours ago	BaseSpace Man	New	New
UI_Automation_Test_20170830_165628_18	UI_Automation_Test_20170830_164850	2 hours ago	BaseSpace Man	New	New
UI_Automation_Test_20170830_165045_8	UI_Automation_Test_20170830_164850	2 hours ago	BaseSpace Man	New	New

https://cloud-hoth.illumina.com/#

The updated interface provides quick access to all your data from a single place while the new action toolbar contains new and approved app functions like requeues, QC status changes, workflows and collaboration tools.

(Refer Slide Time: 11:31)



basespaceuser2@illumina.com

Biosamples

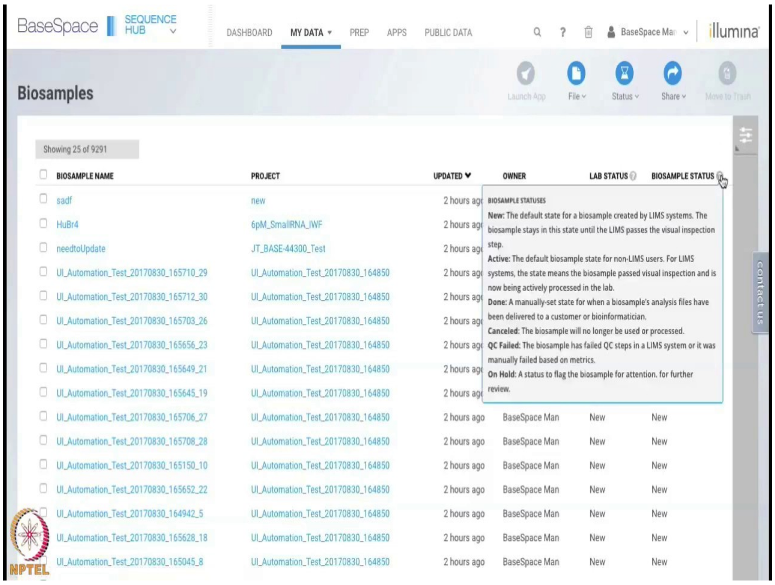
Showing 25 of 9291

BIOSAMPLE NAME	PROJECT	UPDATED	OWNER	LAB	
sadf	new	2 hours ago	BaseSpace Man	Nev	
HuBr4	6pM_SmallRNA_IWF	2 hours ago	BaseSpace Man	Cor	
needtoUpdate	JT_BASE-44300_Test	2 hours ago	BaseSpace Man	Nev	
UI_Automation_Test_20170830_165710_29	UI_Automation_Test_20170830_164850	2 hours ago	BaseSpace Man	Nev	
UI_Automation_Test_20170830_165712_30	UI_Automation_Test_20170830_164850	2 hours ago	BaseSpace Man	Nev	
UI_Automation_Test_20170830_165703_26	UI_Automation_Test_20170830_164850	2 hours ago	BaseSpace Man	Nev	
UI_Automation_Test_20170830_165656_23	UI_Automation_Test_20170830_164850	2 hours ago	BaseSpace Man	Nev	
UI_Automation_Test_20170830_165649_21	UI_Automation_Test_20170830_164850	2 hours ago	BaseSpace Man	Nev	
UI_Automation_Test_20170830_165645_19	UI_Automation_Test_20170830_164850	2 hours ago	BaseSpace Man	Nev	
UI_Automation_Test_20170830_165706_27	UI_Automation_Test_20170830_164850	2 hours ago	BaseSpace Man	New	New
UI_Automation_Test_20170830_165708_28	UI_Automation_Test_20170830_164850	2 hours ago	BaseSpace Man	New	New
UI_Automation_Test_20170830_165150_10	UI_Automation_Test_20170830_164850	2 hours ago	BaseSpace Man	New	New
UI_Automation_Test_20170830_165652_22	UI_Automation_Test_20170830_164850	2 hours ago	BaseSpace Man	New	New
UI_Automation_Test_20170830_164942_5	UI_Automation_Test_20170830_164850	2 hours ago	BaseSpace Man	New	New
UI_Automation_Test_20170830_165628_18	UI_Automation_Test_20170830_164850	2 hours ago	BaseSpace Man	New	New
UI_Automation_Test_20170830_165045_8	UI_Automation_Test_20170830_164850	2 hours ago	BaseSpace Man	New	New

https://cloud-hoth.illumina.com/#

Access your work groups and review your compute and storage usage from the account menu.

(Refer Slide Time: 11:37)



Biosamples

Showing 25 of 9291

BIOSAMPLE NAME	PROJECT	UPDATED	OWNER	LAB STATUS	BIOSAMPLE STATUS
sadf	new	2 hours ag			
HuBr4	6pM_SmallRNA_IWF	2 hours ag			
needtoUpdate	JT_BASE-44300_Test	2 hours ag			
UI_Automation_Test_20170830_165710_29	UI_Automation_Test_20170830_164850	2 hours ag			
UI_Automation_Test_20170830_165712_30	UI_Automation_Test_20170830_164850	2 hours ag			
UI_Automation_Test_20170830_165703_26	UI_Automation_Test_20170830_164850	2 hours ag			
UI_Automation_Test_20170830_165656_23	UI_Automation_Test_20170830_164850	2 hours ag			
UI_Automation_Test_20170830_165649_21	UI_Automation_Test_20170830_164850	2 hours ag			
UI_Automation_Test_20170830_165645_19	UI_Automation_Test_20170830_164850	2 hours ag			
UI_Automation_Test_20170830_165706_27	UI_Automation_Test_20170830_164850	2 hours ago	BaseSpace Man	New	New
UI_Automation_Test_20170830_165708_28	UI_Automation_Test_20170830_164850	2 hours ago	BaseSpace Man	New	New
UI_Automation_Test_20170830_165150_10	UI_Automation_Test_20170830_164850	2 hours ago	BaseSpace Man	New	New
UI_Automation_Test_20170830_165652_22	UI_Automation_Test_20170830_164850	2 hours ago	BaseSpace Man	New	New
UI_Automation_Test_20170830_164942_5	UI_Automation_Test_20170830_164850	2 hours ago	BaseSpace Man	New	New
UI_Automation_Test_20170830_165628_18	UI_Automation_Test_20170830_164850	2 hours ago	BaseSpace Man	New	New
UI_Automation_Test_20170830_165045_8	UI_Automation_Test_20170830_164850	2 hours ago	BaseSpace Man	New	New

BIOSAMPLE STATUSES

New: The default state for a biosample created by LIMS systems. The biosample stays in this state until the LIMS passes the visual inspection step.

Active: The default biosample state for non-LIMS users. For LIMS systems, the state means the biosample passed visual inspection and is now being actively processed in the lab.

Done: A manually-set state for when a biosample's analysis files have been delivered to a customer or bioinformatician.

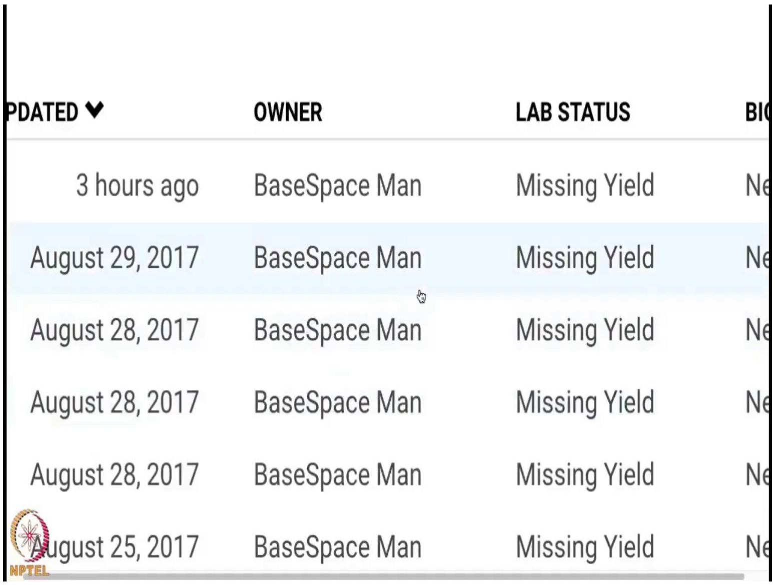
Canceled: The biosample will no longer be used or processed.

QC Failed: The biosample has failed QC steps in a LIMS system or it was manually failed based on metrics.

On Hold: A status to flag the biosample for attention. for further review.

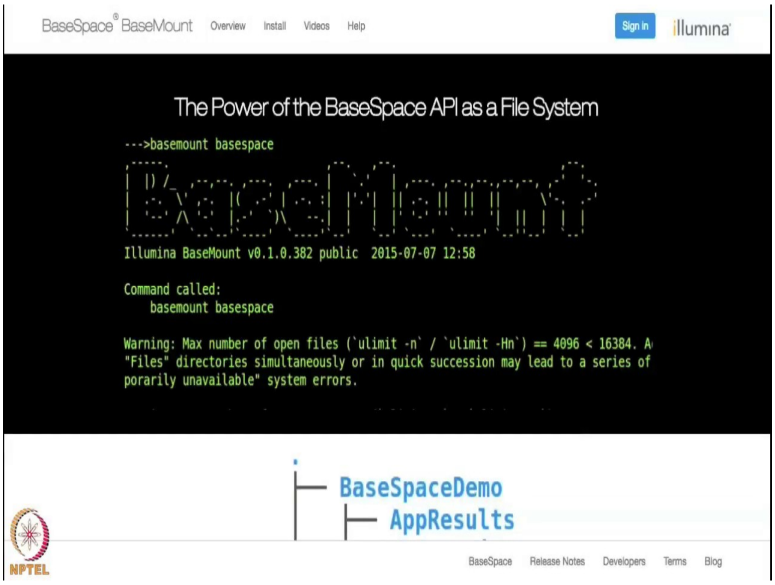
Inline tooltips help you understand what is occurring with your data and the enhanced filters widget has been added to more places, letting you get to the data you care about more quickly.

(Refer Slide Time: 11:47)



PDATED	OWNER	LAB STATUS	BI
3 hours ago	BaseSpace Man	Missing Yield	Ne
August 29, 2017	BaseSpace Man	Missing Yield	Ne
August 28, 2017	BaseSpace Man	Missing Yield	Ne
August 28, 2017	BaseSpace Man	Missing Yield	Ne
August 28, 2017	BaseSpace Man	Missing Yield	Ne
ugust 25, 2017	BaseSpace Man	Missing Yield	Ne

(Refer Slide Time: 11:49)



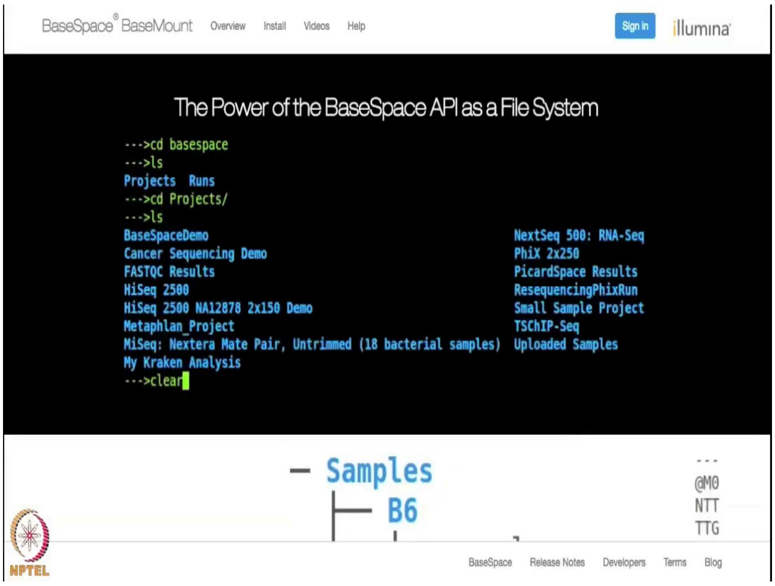
The Power of the BaseSpace API as a File System

```
--->basemount basespace

Illumina BaseMount v0.1.0.382 public  2015-07-07 12:58

Command called:
    basemount basespace

Warning: Max number of open files (`ulimit -n` / `ulimit -Hn`) == 4096 < 16384. A
"Files" directories simultaneously or in quick succession may lead to a series of
porarily unavailable" system errors.
```

```
.
├── BaseSpaceDemo
│   ├── AppResults
```

The API base space cli and base mount tools provide access to your data from a command line interface.

(Refer Slide Time: 11:51)



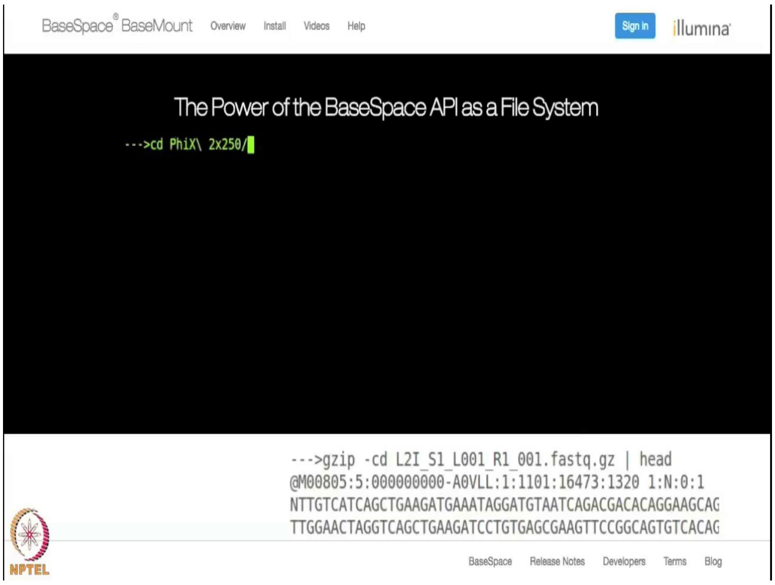
The Power of the BaseSpace API as a File System

```
--->cd basespace
--->ls
Projects  Runs
--->cd Projects/
--->ls
BaseSpaceDemo                                             NextSeq 500: RNA-Seq
Cancer Sequencing Demo                                    PhiX 2x250
FASTQC Results                                            PicardSpace Results
HiSeq 2500                                                ResequencingPhixRun
HiSeq 2500 NA12878 2x150 Demo                             Small Sample Project
Metaphlan_Project                                         TSChIP-Seq
MiSeq: Nextera Mate Pair, Untrimmed (18 bacterial samples) Uploaded Samples
My Kraken Analysis
--->clear
```

```
─ Samples                    ---
  ├── B6                     @M0
                             NTT
                             TTG
```

And have been improved to facilitate more advanced integration and automation.

(Refer Slide Time: 11:57)

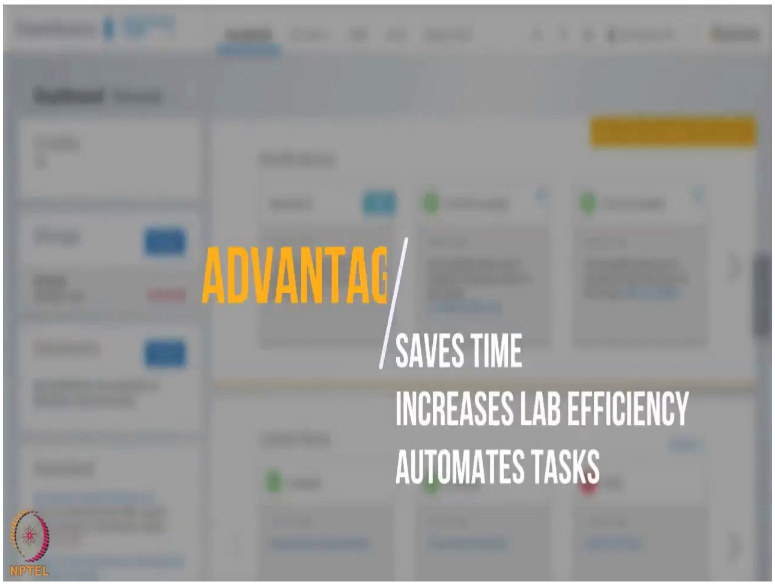


The Power of the BaseSpace API as a File System

```
--->cd PhiX\ 2x250/
```

```
--->gzip -cd L2I_S1_L001_R1_001.fastq.gz | head
@M00805:5:000000000-A0VLL:1:1101:16473:1320 1:N:0:1
NTTGTCATCAGCTGAAGATGAAATAGGATGTAATCAGACGACACAGGAAGCAG
TTGGAACTAGGTCAGCTGAAGATCCTGTGAGCGAAGTTCCGGCAGTGTCACAG
```

(Refer Slide Time: 11:59)



With these new enhancements base space sequence up takes your work from sample to result more quickly, more efficiently and with more control than never before.

(Refer Slide Time: 12:09)

BaseSpace SEQUENCE HUB

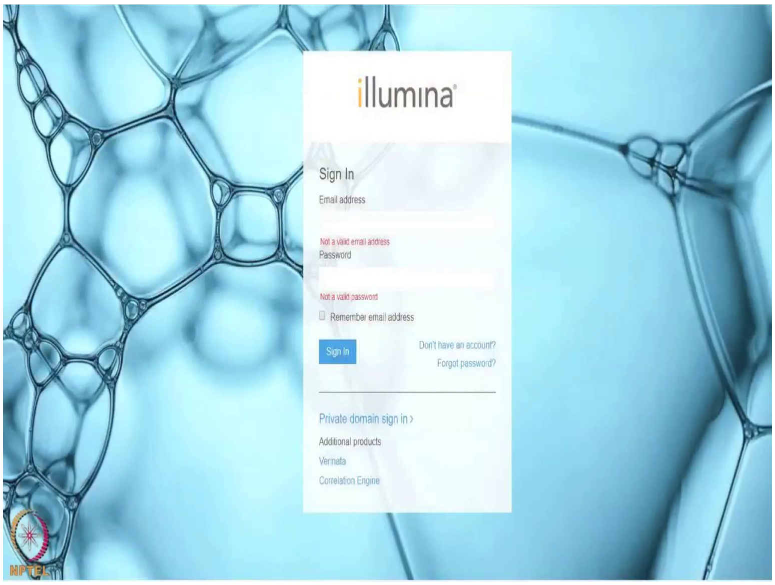
www illumina.com/basespace

For Research Use Only. Not for use in diagnostic procedures.

illumina
NPTEL

To learn more about base space sequence of and all of our new features please visit http://www.illumina.com/basespace.

(Refer Slide Time: 12:23)



Sign In

Email address

Not a valid email address

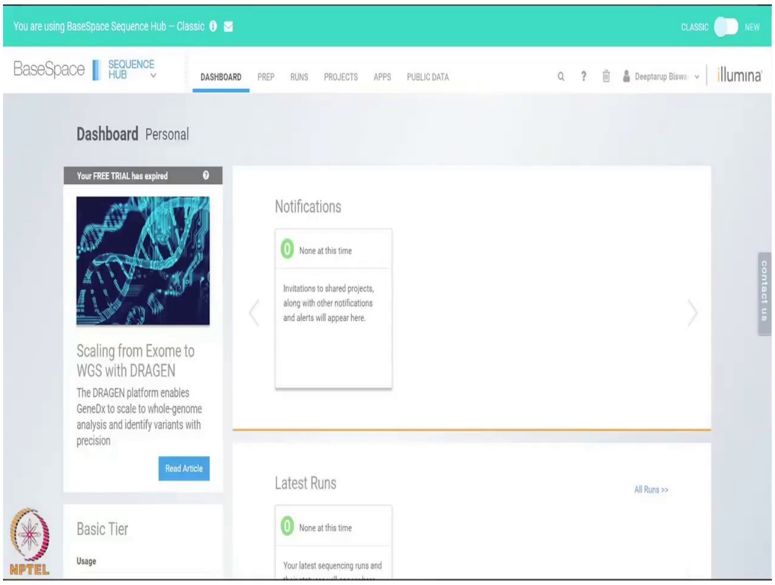
Password

Not a valid password

□ Remember email address

Find it, ok. Here it is. So, base space is nothing but a cloud application that is developed by Illumina. So, this is hosted on Amazon Web Services, AWS and it is a free application, in the sense anybody can access the application. Some of the apps on base space are paid which means that you have to pay for using those apps but primarily base space is freely available.

(Refer Slide Time: 13:11)



You are using BaseSpace Sequence Hub – Classic

CLASSIC NEW

BaseSpace SEQUENCE HUB

DASHBOARD PREP RUNS PROJECTS APPS PUBLIC DATA

Deeptarup Biswa

illumina

Dashboard Personal

Your FREE TRIAL has expired

Scaling from Exome to WGS with DRAGEN

The DRAGEN platform enables GeneDx to scale to whole-genome analysis and identify variants with precision

Read Article

Notifications

0 None at this time

Invitations to shared projects, along with other notifications and alerts will appear here.

Latest Runs

All Runs >>

0 None at this time

Your latest sequencing runs and

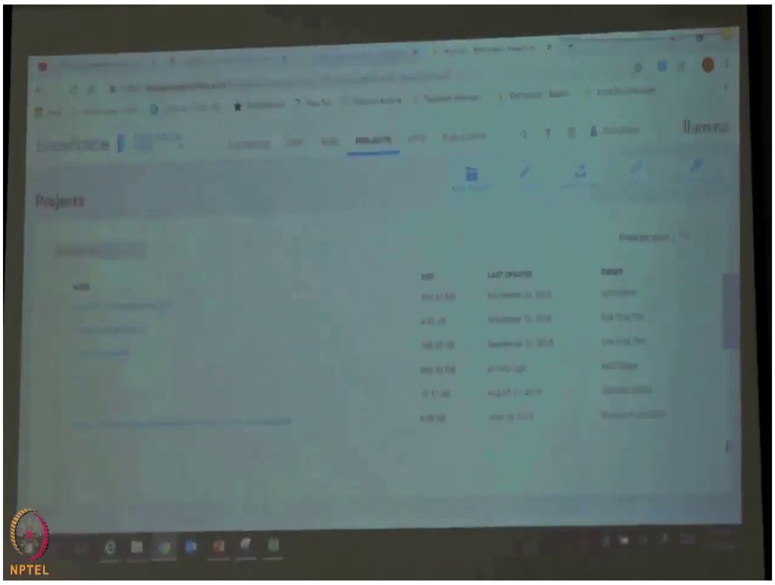
Basic Tier

Usage

contact us

What I want you guys to do is log into your base space account and refer to the hand out that was given to you but before we actually start doing the hands on session according to the handout, I wanted to show one very interesting property of the Illumina data that is something that Mukesh again touched upon during his presentation that is the quality score.

(Refer Slide Time: 13:33)

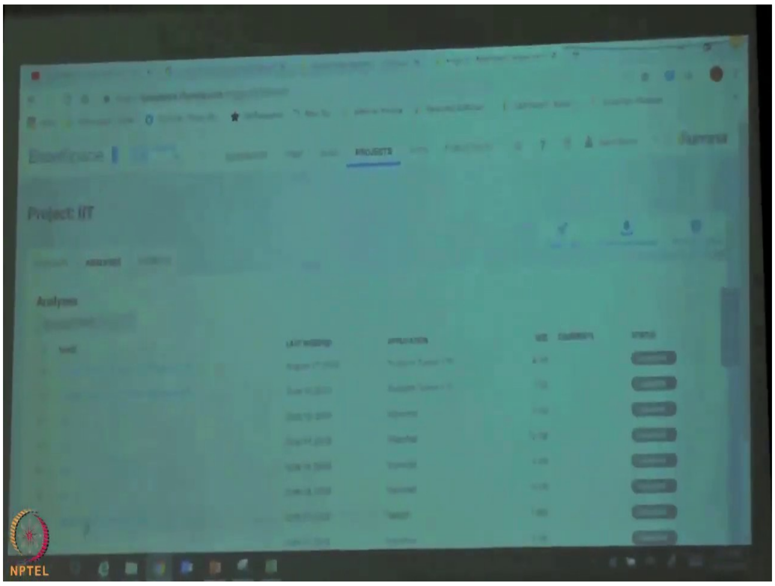


The very very high quality data that is generated on the Illumina platform. So, he talked about Q30, right. Q30 is nothing but a Phred score. So, those of you who have used sanger sequencing will be aware that Phred score is a quality score that is assigned to a base call that is made by any sequencer and it is nothing but the confidence that the caller has in the base that it has called, right.

So, it is a probability. So, when we say 99.99 percent, you know we are 99.99 percent sure that the base that we have called, let us say if we call it an a, it is going to be an a, right. So, as Mukesh said that the error rate is going to be 1 in 1000s. So, we are going to be wrong 1 in every 1000 bases that are read.

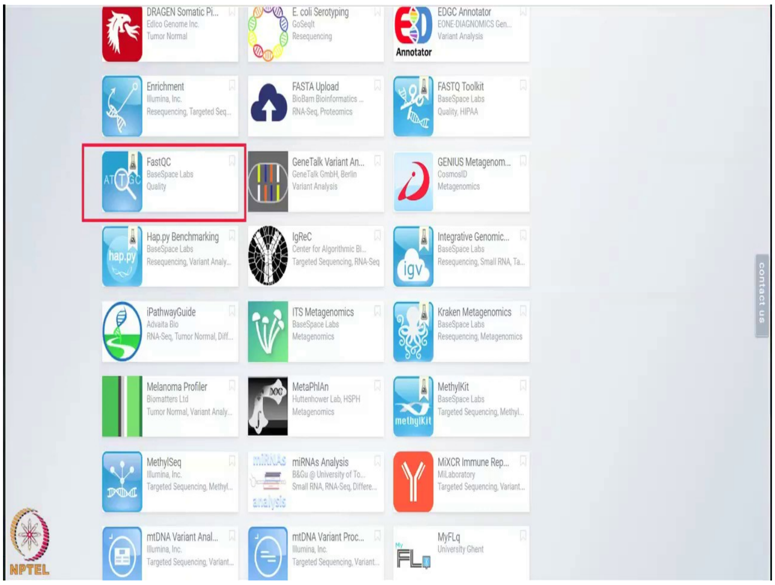
So, you can imagine because the read lengths, you know that we generate from our platforms are no more than 600 bases and our error rate is 1 in a 1000. So, the chances of there being an error in the data that we generate are very very small. This is a IIT data. So, we had run a project for one of the PIs here, and you guys do not have access to this.

(Refer Slide Time: 14:45)



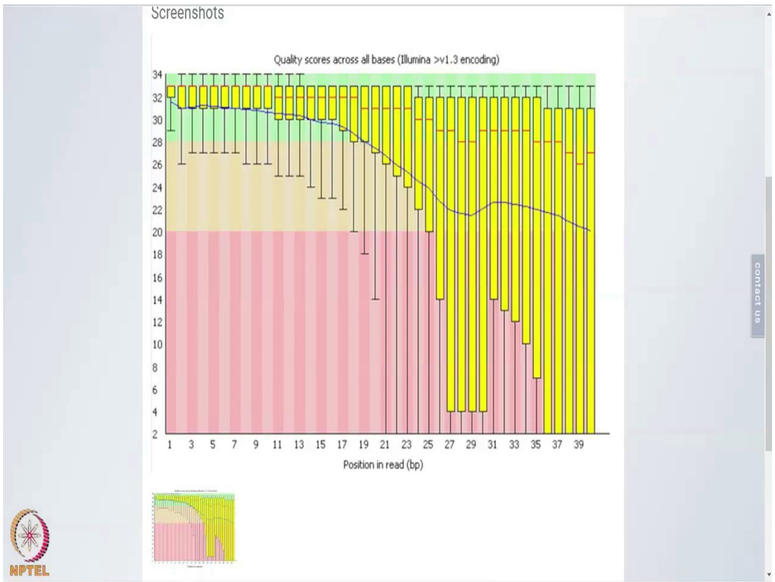
So, I am just going to show you the data because I really wanted to show you the quality of the data that gets generated.

(Refer Slide Time: 15:21)



So, on base space when you have some time and if you are interested there are multiple apps that are available. So, apps are nothing but small widgets that are created either by Illumina or by third party, you know researchers, companies that are supporting data analysis on the Illumina platform and they are you know made available. So, based on the application you want to or you are working on, you can choose the appropriate app and run the analysis. What I wanted to show you today was data that is generated from an application known as FastQC, ok.

(Refer Slide Time: 15:27)



So, FastQC is an application. It is an open source application. I forget which university it came out of but this has been around for at least 8 or 9 years now, very very widely used to evaluate the quality of the data that is generated by sequencers.

So, the way to read this data is horizontally 1 to 99. So, this is a 100 base pair read, 100 base read, right. We talked about the size of the fragment, the size of the read that you generate. So, when we say read it means it is the contiguous output that is generated by the sequencer. So, in this particular instance we are looking at a 100 base pair read data that is generated. So, the Illumina platforms can generate as small as 36 base pair reads and as long as 600 base pair reads, ok.

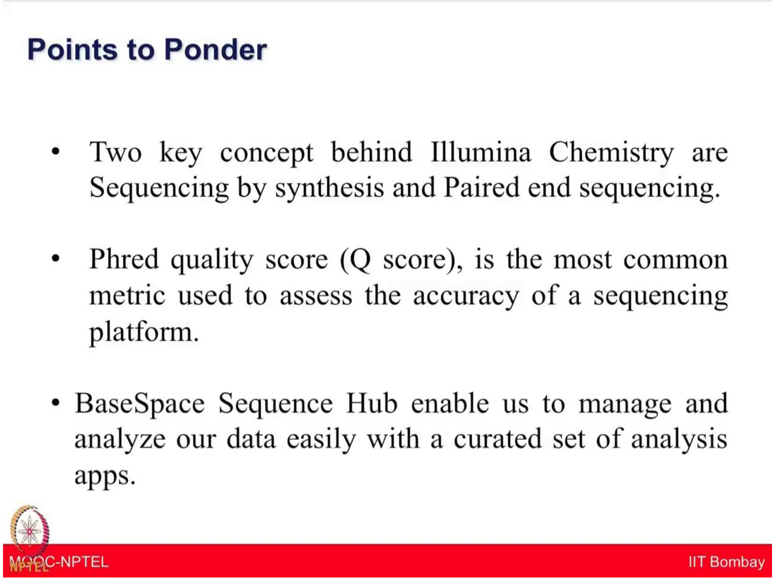
The y axis rather the vertical axis shows you the Q score, ok. So, this is again a measure of the quality of the data that is generated on the Illumina sequences and as you can see we are literally touching the ceiling of the scale that is available. So, the Phred score ranges from Q10 to Q40, ok. So, Q40 being 100 percent accuracy Q10 is I think 1 in a 10, 1 in 10 error rate that is the way to interpret it.

So, you can see that for majority of the length of the read our Q scores are very very well above Q30, right. So, essentially what that means is all the data that you are generating practically all the data that you are generating is usable. You do not have to throw out or filter out any data because it is low quality and this is very very critical in clinical applications primarily because you want to make sure that any data that you generate is of high quality, right; because, what are you looking for when you are generating sequence data; you are looking for differences from the reference genome, right, you are looking for differences from the reference genome which can be in the form of single base variations, right and what are errors in sequencing generally; they are single base variations, right.

So, you want to make sure that whatever variations you are calling basis of which you may be taking some clinical decisions, one have to be accurate, right. You have to be 100 percent sure that the base that you are calling as a variant is actually a variant, right. So, this is where something like this becomes extremely critical and you can pick up any data.

This is actually the data that we are generated here itself in house rather for IIT, Bombay. So, these are actually patient samples, these are tumour samples. So, these are not even you know very very well maintained cell lines or blood samples which is where you all pretty much expect to get high quality data. These are tumour samples. So, again you can see that on real biological samples you get very high quality data.

(Refer Slide Time: 18:45)



Points to Ponder

- Two key concept behind Illumina Chemistry are Sequencing by synthesis and Paired end sequencing.
- Phred quality score (Q score), is the most common metric used to assess the accuracy of a sequencing platform.
- BaseSpace Sequence Hub enable us to manage and analyze our data easily with a curated set of analysis apps.

MOOC-NPTEL IIT Bombay

In today’s lecture, we learnt about NGS technology platform especially how Illumina chemistry work. Dr. Desai also talked about the importance of Phred score and how it gives the idea about our data quality. She also showed how a real data from biological sample actually looks like and how to read their data. So, I hope all of you have opened an account in base space which is available free. Please open an account and get ready for Dr. Aarti Desai next hands on session which will be based on the base space account.

In the next hands on session, she will take you a journey where you can use varous datasets from your own experiments or publicly available data sets, analyse them and make meaningful insights from their data.

Thank you.