An Introduction to Proteogenomics

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

Supplementary - 4
NGS: Sequencing By Synthesis II

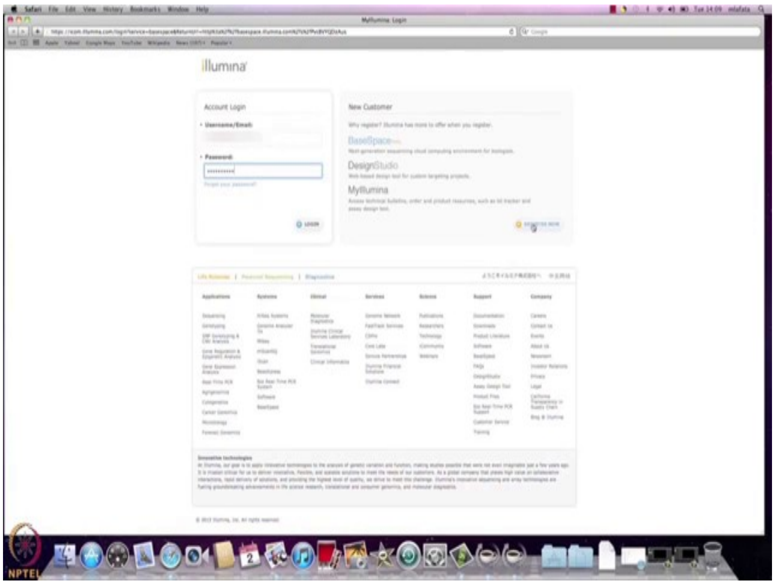
Welcome to MOOC course on Introduction to Proteogenomics. Today we are going to talk about Next Generation Sequencing, Sequencing By Synthesis. In this session Dr. Aarti Desai will give you a glimpse of BaseSpace sequence hub, which is Illumina cloud computing environment to analyze the sequencing data set.

She will mainly focus on how to search in public data, import data in personal dashboard and how to create your own project. Dr. Aarti will also explain you about the sharing of data among users and how this hub is playing a big role in collaborative platforms. Finally, she will also show you how to launch an analysis in BaseSpace with the help of multiple software already available there. So, let us welcome Dr. Aarti for today's lecture.

(Refer Slide Time: 01:34)

How to Open the Tumor/Normal Sequencing Data in BaseSpace

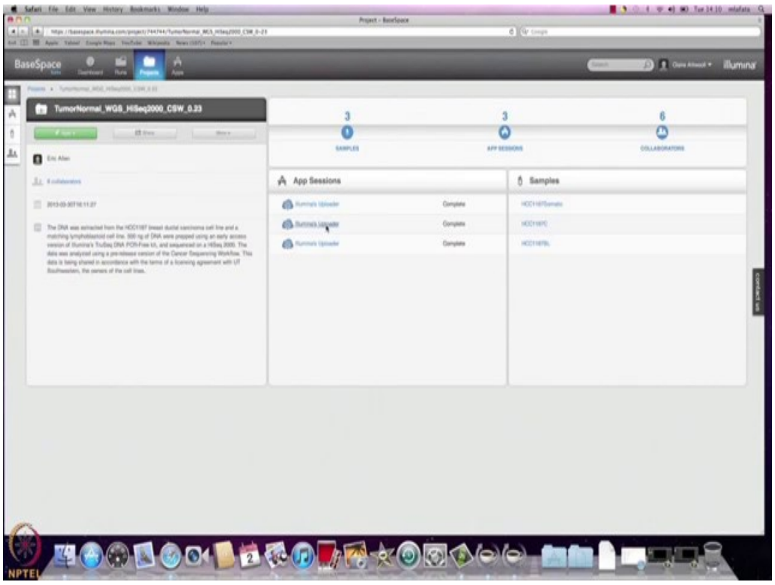
(Refer Slide Time: 01:37)



So, now we can actually move on to.

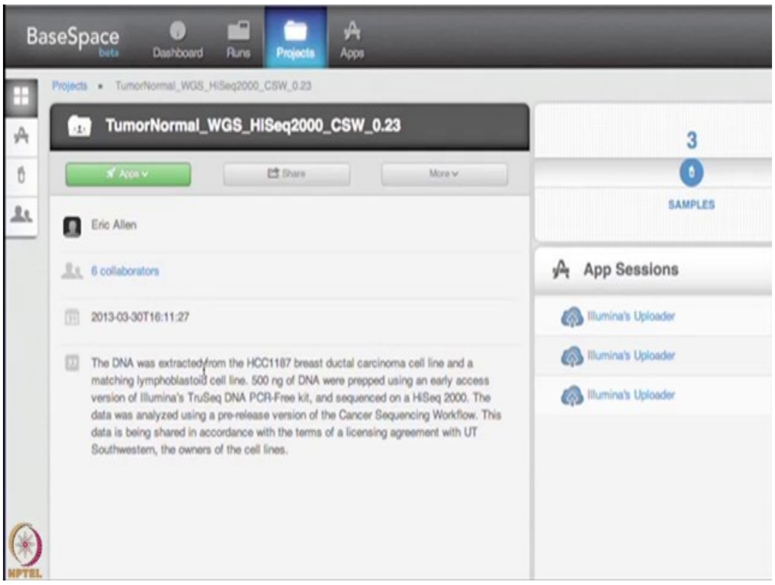
Clicking on the link will take you to the BaseSpace login page. Login or create an account by clicking register now, this will give you instant access. The first time you log into BaseSpace you may need to read and accept the user agreement.

(Refer Slide Time: 01:53)



After logging in you will be taken to the tumor normal BaseSpace project page.

(Refer Slide Time: 01:59)



TumorNormal_WGS_HiSeq2000_CSW_0.23

Eric Allen

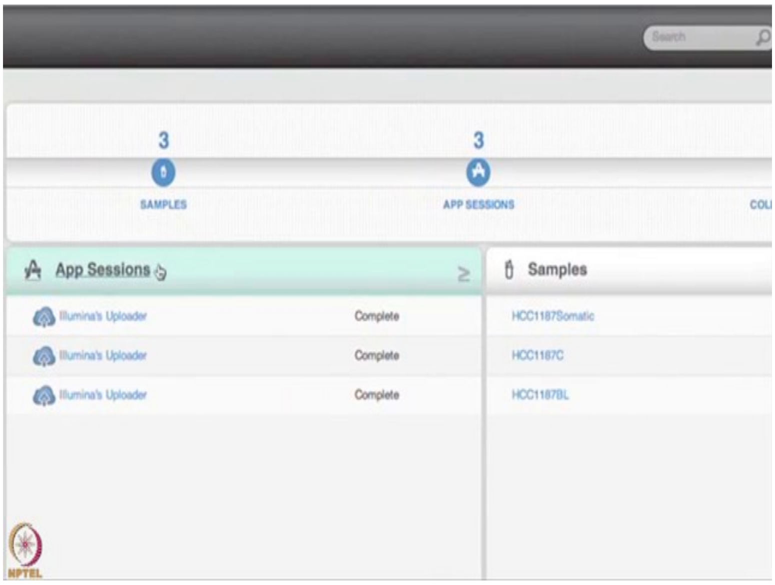
6 collaborators

2013-03-30T16:11:27

The DNA was extracted from the HCC1187 breast ductal carcinoma cell line and a matching lymphoblastoid cell line. 500 ng of DNA were prepped using an early access version of Illumina's TruSeq DNA PCR-Free kit, and sequenced on a HiSeq 2000. The data was analyzed using a pre-release version of the Cancer Sequencing Workflow. This data is being shared in accordance with the terms of a licensing agreement with UT Southwestern, the owners of the cell lines.

A project is a container for various analysis files. These include samples which have fast queues that result from runs and output files from running analyses. These are stored under app sessions. In contrast a run in BaseSpace corresponds to data from sequencing a single flow cell on an instrument. The project overview page contains a brief description of the dataset. You can navigate around using the quick launch buttons either using the buttons on the left tab, using the names or the direct links. These all link to the same place.

(Refer Slide Time: 02:32)

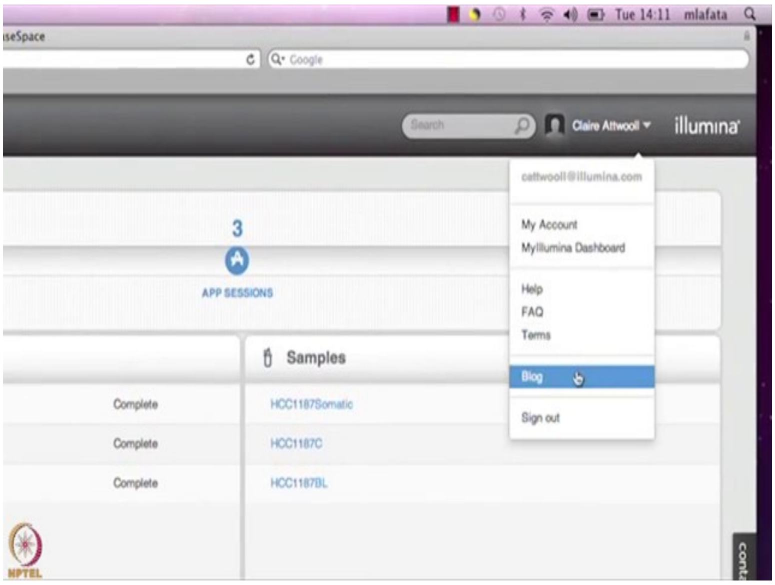


App Sessions	
Illumina's Uploader	Complete
Illumina's Uploader	Complete
Illumina's Uploader	Complete

Samples
HCC1187Somatic
HCC1187C
HCC1187BL

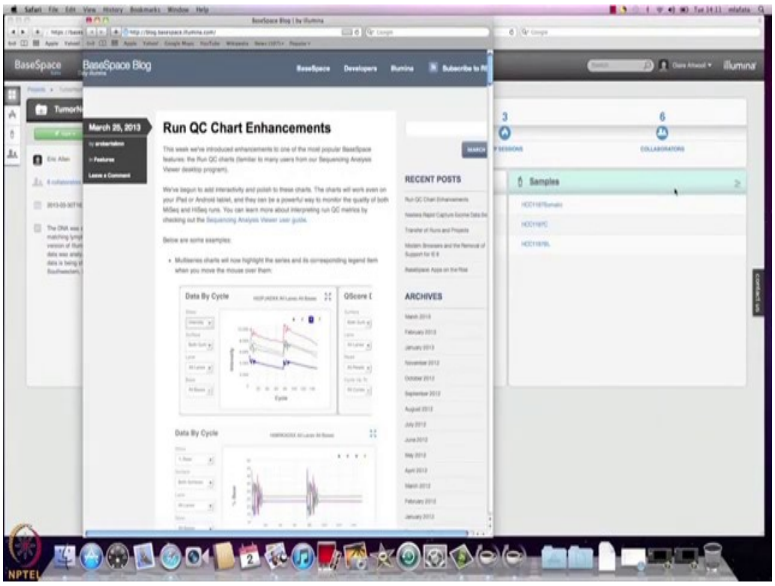
The collaborators button indicates who the run has been shared with. If you are the owner of a run or project you can change the name and transfer runs or projects to collaborators. The share button allows you to share your projects. Note that we cannot share this project since we are not the project owner.

(Refer Slide Time: 03:07)



You can also access the BaseSpace blog from this page.

(Refer Slide Time: 03:09)



Run QC Chart Enhancements

This week we've introduced enhancements to one of the most popular BaseSpace features: the Run QC charts (familiar to many users from our Sequencing Analysis Viewer desktop program).

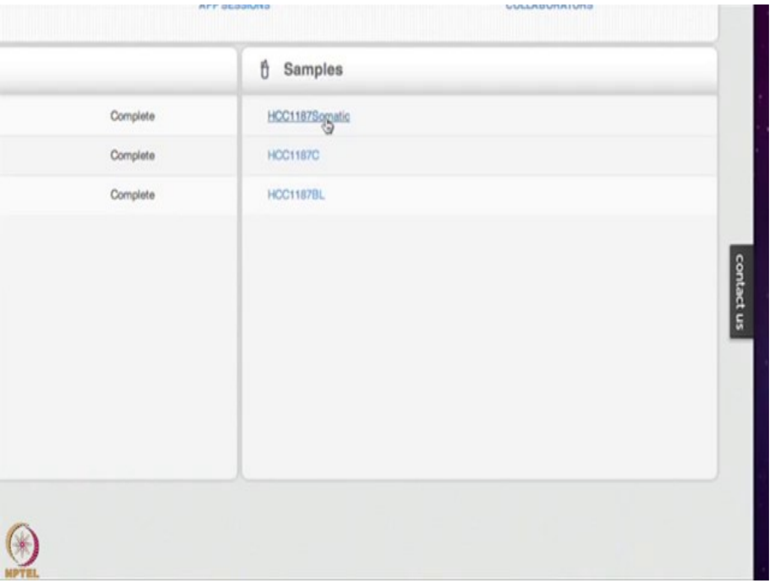
We've begun to add interactivity and polish to these charts. The charts will work even on your iPad or Android tablet, and they can be a powerful way to monitor the quality of both MiSeq and HiSeq runs. You can learn more about interpreting run QC metrics by checking out the Sequencing Analysis Viewer user guide.

Below are some examples:

- Multiseries charts will now highlight the series and its corresponding legend item when you move the mouse over them.

The blog provides useful updates and information on BaseSpace.

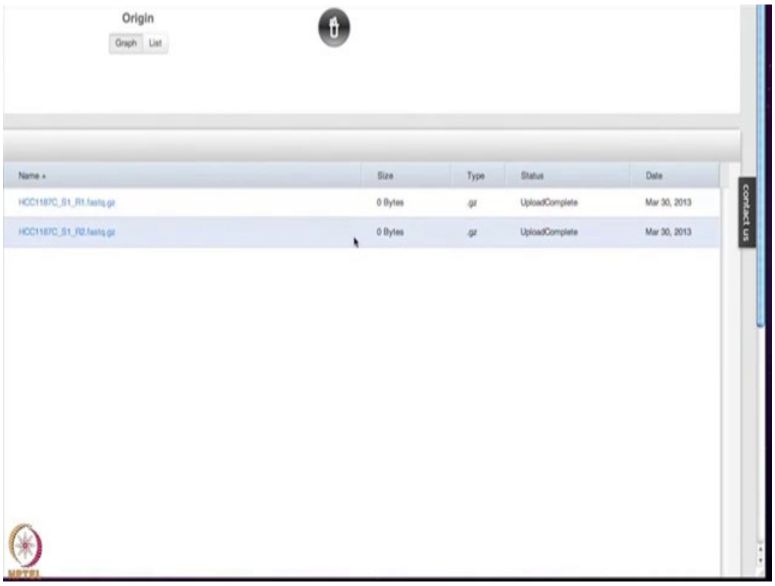
(Refer Slide Time: 03:18)



	Samples
Complete	HCC1187Somatic
Complete	HCC1187C
Complete	HCC1187BL

There are three samples for this dataset HCC 1187 C is the cancer sample from the breast ductal carcinoma cell line. HCC 1187 BL is the matching than for blastoid cell line from the same individual representing a normal sample. HCC 1187 Somatic represents the subtracted data. The normal and tumor samples are compared and matching sequences removed leaving only the tumor normal differences.

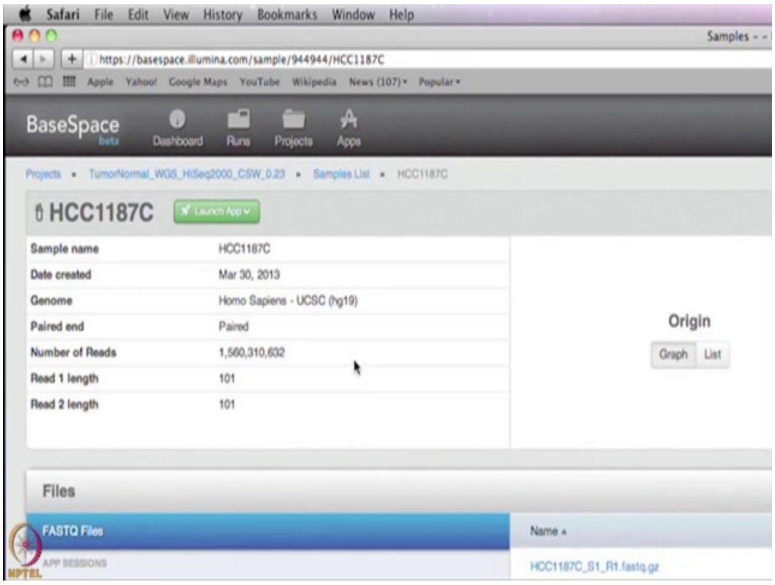
(Refer Slide Time: 03:53)



Name ▴	Size	Type	Status	Date
HCC1187C_S1_R1.fastq.gz	0 Bytes	.gz	UploadComplete	Mar 30, 2013
HCC1187C_S1_R2.fastq.gz	0 Bytes	.gz	UploadComplete	Mar 30, 2013

Inside the samples hyperlinks you will find fast q files, these will usually be full fast q files, but in this case empty fast q files were uploaded as placeholders.

(Refer Slide Time: 04:01)



Safari File Edit View History Bookmarks Window Help
Samples - -
https://basespace.illumina.com/sample/944944/HCC1187C
Apple Yahoo! Google Maps YouTube Wikipedia News (107) Popular
BaseSpace beta
Dashboard
Runs
Projects
Apps
Projects • TumorNormal_WGS_HiSeq2000_CSW_0.23 • Samples List • HCC1187C
Launch App
Origin
Graph
List

HCC1187C

Sample name	HCC1187C
Date created	Mar 30, 2013
Genome	Homo Sapiens - UCSC (hg19)
Paired end	Paired
Number of Reads	1,560,310,632
Read 1 length	101
Read 2 length	101

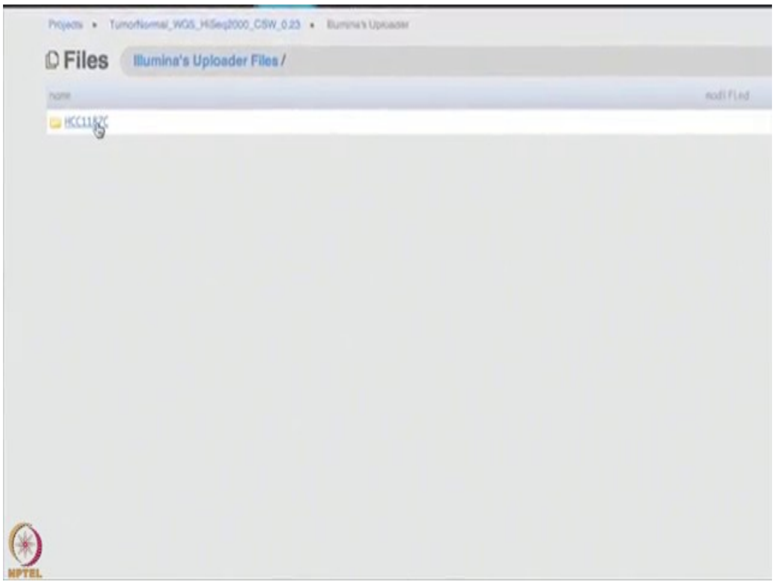
Files

FASTQ Files
APP SESSIONS
Name

HCC1187C_S1_R1.fastq.gz

You will also find the information about the samples including the read length, the total number of reads and whether this was a single read or paired end run. Inside the app sessions there is a folder for each analysis performed. Here they are named Illumina's uploader as they were uploaded directly by Illumina. The naming convention should be more informative in the future, but here the sample names correspond to the app sessions. So, you can use these sample names as a guide.

(Refer Slide Time: 04:35)

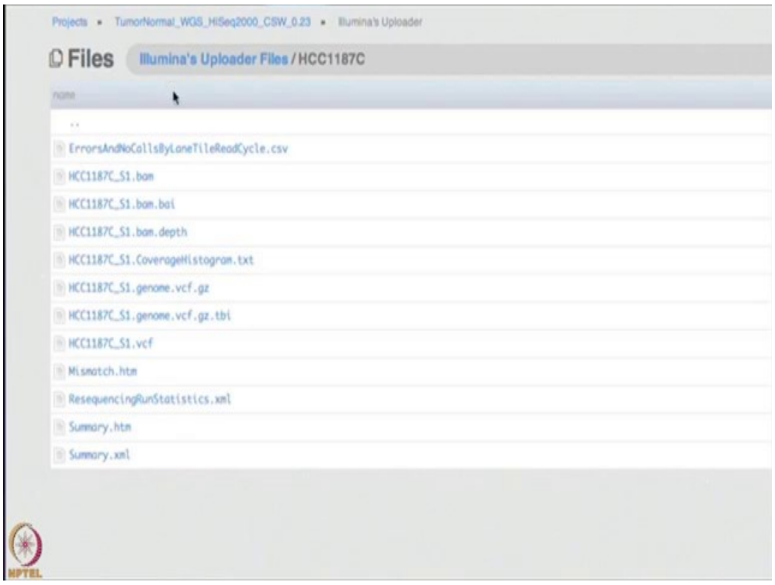


Projects • TumorNormal_WGS_HiSeq2000_CSW_0.23 • Illumina's Uploader
Files
Illumina's Uploader Files /
name
modified

HCC1187C

Clicking on the hyperlinks will also show you the folder name.

(Refer Slide Time: 04:41)



Projects • TumorNormal_WGS_HiSeq2000_CSW_0.23 • Illumina's Uploader

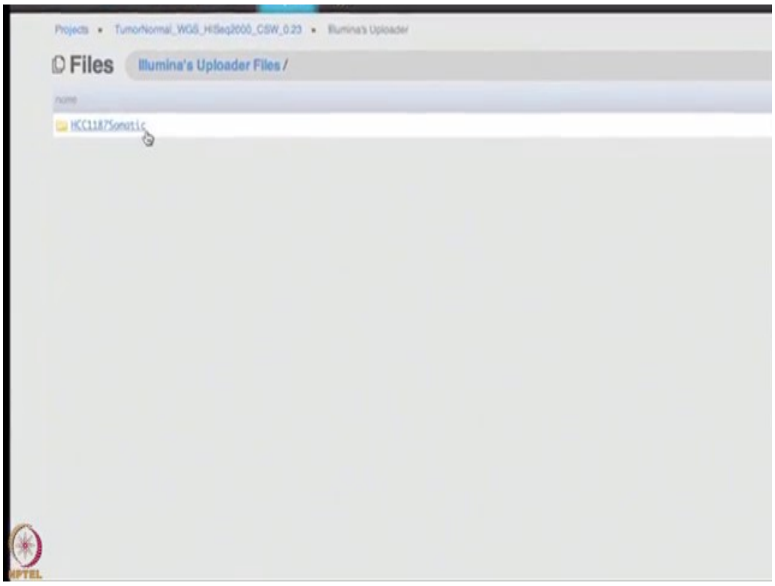
Files Illumina's Uploader Files / HCC1187C

name

..
ErrorsAndNoCallsByLaneTileReadCycle.csv
HCC1187C_S1.bam
HCC1187C_S1.bam.bai
HCC1187C_S1.bam.depth
HCC1187C_S1.CoverageHistogram.txt
HCC1187C_S1.genome.vcf.gz
HCC1187C_S1.genome.vcf.gz.tbi
HCC1187C_S1.vcf
Mismatch.htm
ResequencingRunStatistics.xml
Summary.htm
Summary.xml

Inside the cancer and bloodline folders you will find spam files vcf and genome vcf files and there will usually be a summary report for the analysis.

(Refer Slide Time: 04:58)



Projects • TumorNormal_WGS_HiSeq2000_CSW_0.23 • Illumina's Uploader

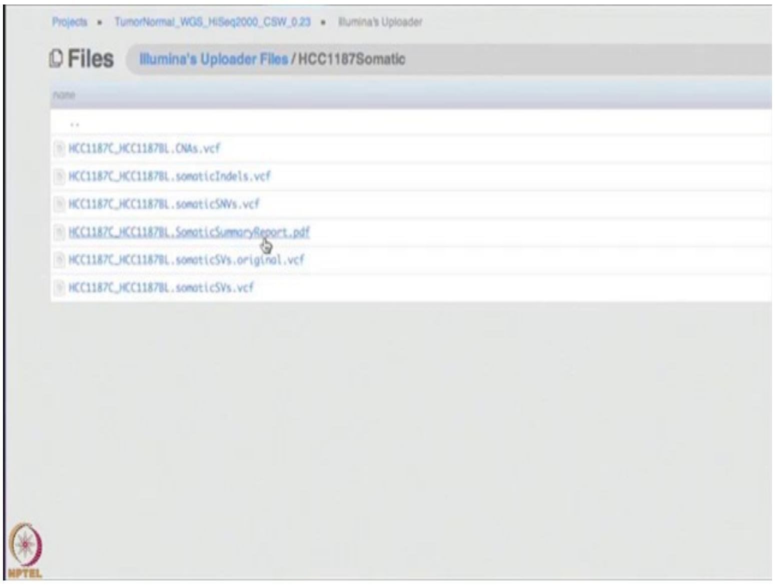
Files Illumina's Uploader Files /

name

HCC1187Somatic

Inside the somatic folder of vcf files from subtracted data these show the variance between the tumor normal data.

(Refer Slide Time: 05:01)



Projects • TumorNormal_WGS_HiSeq2000_CSW_0.23 • Illumina's Uploader

Files Illumina's Uploader Files / HCC1187Somatic

name

..

HCC1187C_HCC1187BL.CNAs.vcf

HCC1187C_HCC1187BL.somaticIndels.vcf

HCC1187C_HCC1187BL.somaticSNVs.vcf

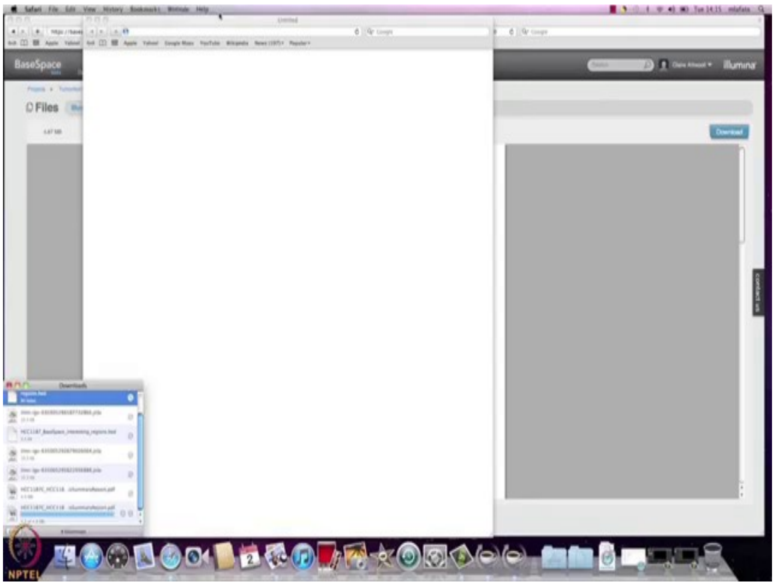
HCC1187C_HCC1187BL.SomaticSummaryReport.pdf

HCC1187C_HCC1187BL.somaticSVs.original.vcf

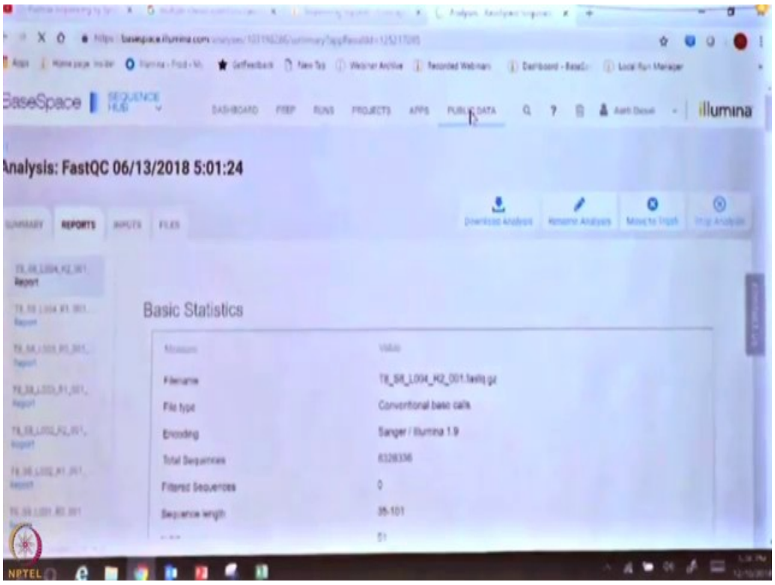
HCC1187C_HCC1187BL.somaticSVs.vcf

There is also the somatic summary report. You can open the somatic summary report within BaseSpace or download it directly to your computer.

(Refer Slide Time: 05:15)



(Refer Slide Time: 05:29)



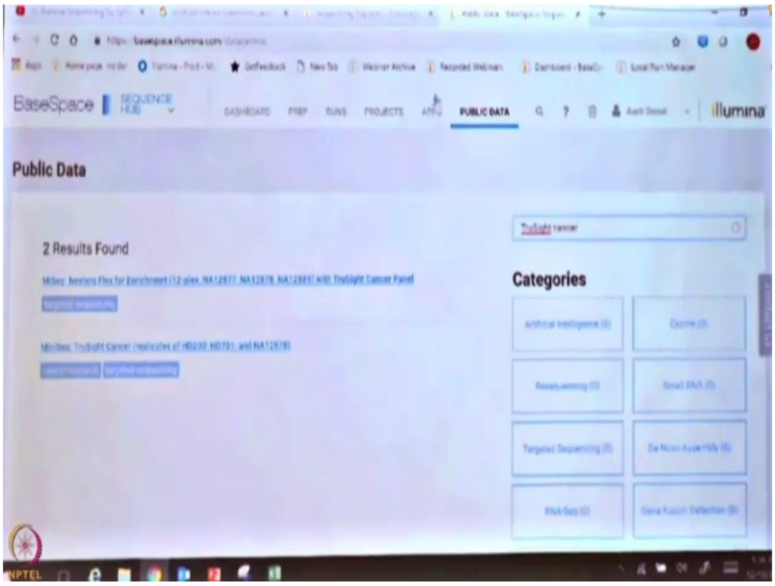
Analysis: FastQC 06/13/2018 5:01:24

Basic Statistics

Measure	Value
Filename	TX_S8_L004_R2_001.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	8328336
Filtered Sequences	0
Sequence length	35-101
[illegible]	51

So what I want you to do is, once you are into your account. I hope everybody has a BaseSpace account and you are able to log into your BaseSpace account ok.

(Refer Slide Time: 05:44)



Public Data

TruSight cancer

2 Results Found

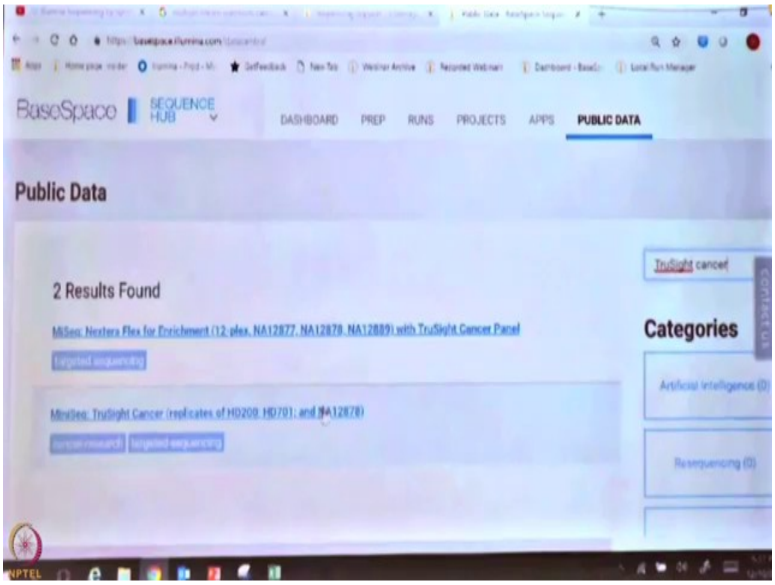
MiSeq: Nextera Flex for Enrichment (12-plex, NA12877, NA12878, NA12891) with TruSight Cancer Panel

targeted sequencing

MiniSeq: TruSight Cancer (replicates of HD200, HD701, and NA12878)

Go to public data and in here search for TruSight cancer. Yeah, go to public data here TruSight cancer its there on your screen ok. TruSight T R U S I G H T cancer. Now click on the second project that is the MiniSeq TruSight cancer.

(Refer Slide Time: 06:27)



Public Data

TruSight cancer

2 Results Found

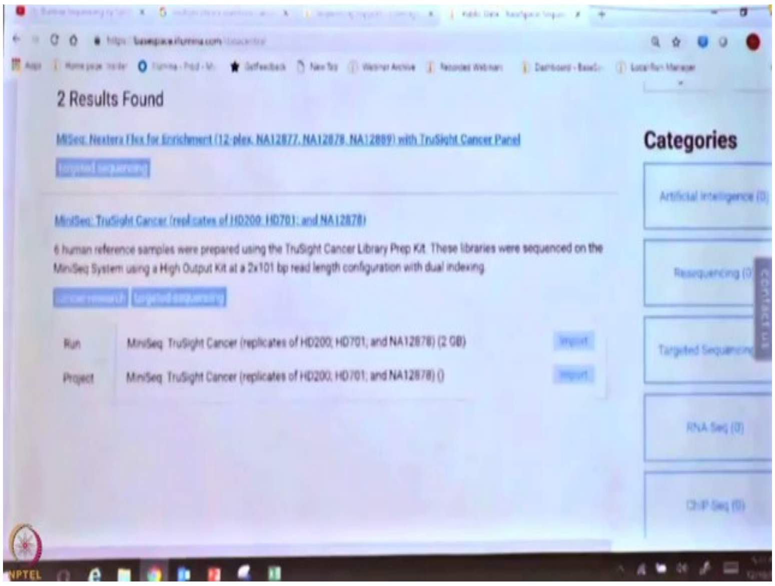
MiSeq: Nextera Flex for Enrichment (12-plex, NA12877, NA12878, NA12889) with TruSight Cancer Panel

MiniSeq: TruSight Cancer (replicates of HD200, HD701; and NA12878)

So, do you have this MiniSeq TruSight cancer project you all saw that.

Click on that.

(Refer Slide Time: 06:36)



2 Results Found

MiSeq: Nextera Flex for Enrichment (12-plex, NA12877, NA12878, NA12889) with TruSight Cancer Panel

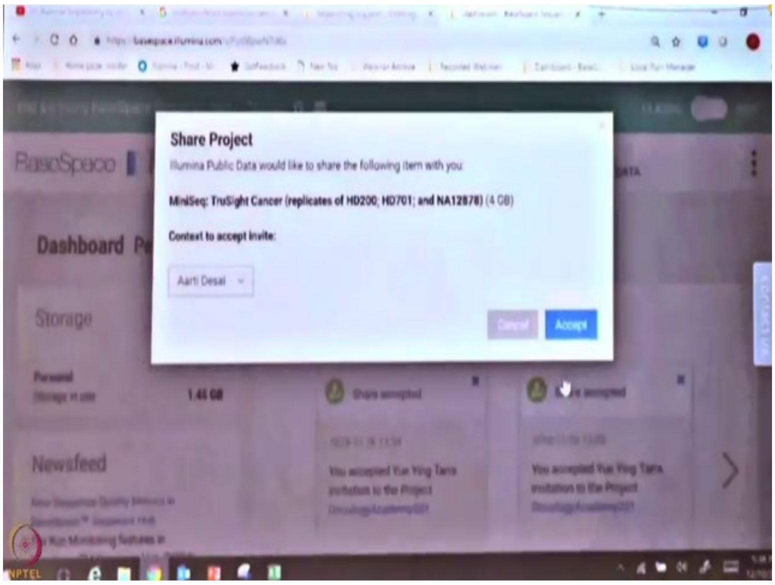
MiniSeq: TruSight Cancer (replicates of HD200, HD701; and NA12878)

6 human reference samples were prepared using the TruSight Cancer Library Prep Kit. These libraries were sequenced on the MiniSeq System using a High Output Kit at a 2x101 bp read length configuration with dual indexing.

Run	MiniSeq: TruSight Cancer (replicates of HD200, HD701; and NA12878) (2 GB)	Import
Project	MiniSeq: TruSight Cancer (replicates of HD200, HD701; and NA12878) ()	Import

When you click on that it will expand and you will see two options there; one is run and there is project. What I want us to do is to import the project. So, click on import.

(Refer Slide Time: 06:57)



Share Project

Illumina Public Data would like to share the following item with you:

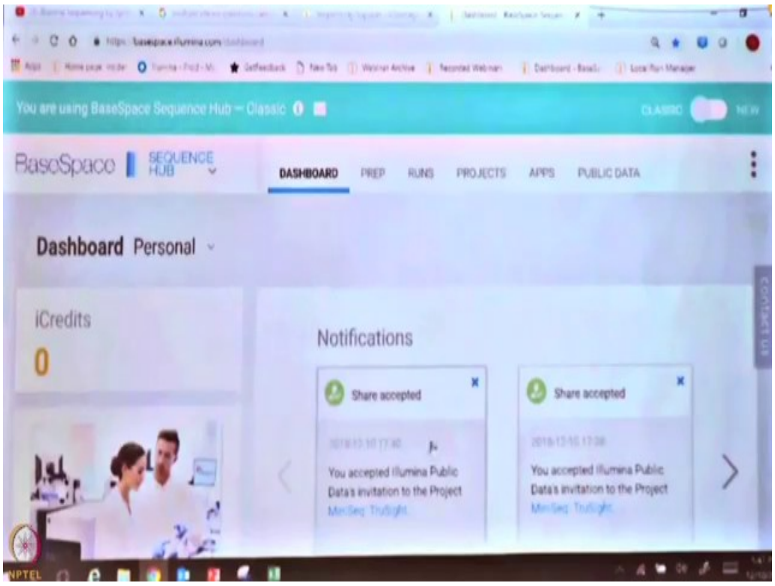
MiniSeq: TruSight Cancer (replicates of HD200; HD701; and NA12878) (4 GB)

Context to accept invite:

Aarti Desai

So, when you click on import. Did you guys see import or shall I go back? Here everybody with me so far ok. So, when you click on import you will get this pop up that says you know I want to share a project with you, accept it.

(Refer Slide Time: 07:17)



You are using BaseSpace Sequence Hub – Classic

Dashboard Personal

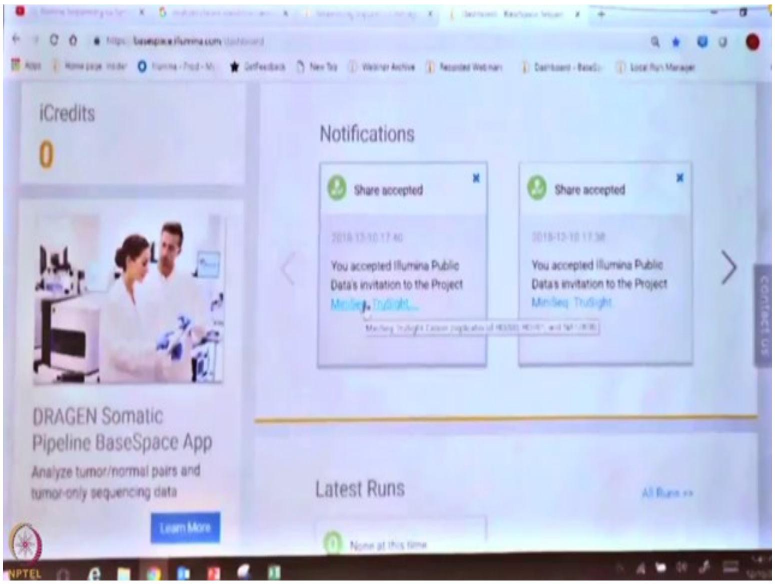
iCredits 0

Notifications

Share accepted
You accepted Illumina Public Data's invitation to the Project MiniSeq: TruSight

Share accepted
You accepted Illumina Public Data's invitation to the Project MiniSeq: TruSight

(Refer Slide Time: 07:17)



iCredits
0
DRAGEN Somatic Pipeline BaseSpace App
Analyze tumor/normal pairs and tumor-only sequencing data
Learn More
Latest Runs
All Runs »
None at this time

Notifications

Share accepted

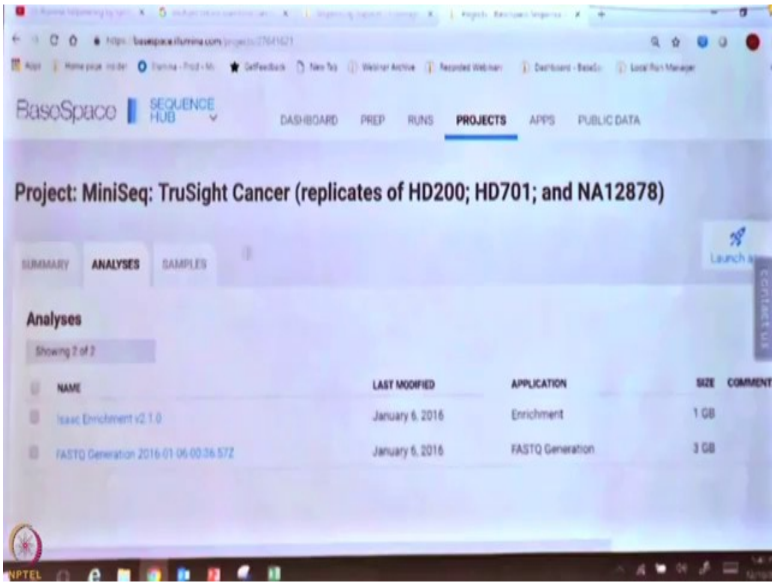
You accepted Illumina Public Data's invitation to the Project MiniSeq: TruSight...

Share accepted

You accepted Illumina Public Data's invitation to the Project MiniSeq: TruSight

So, once you accept it you should see it in your notifications that you now have this project MiniSeq TruSight cancer.

(Refer Slide Time: 07:38)



Project: MiniSeq: TruSight Cancer (replicates of HD200; HD701; and NA12878)

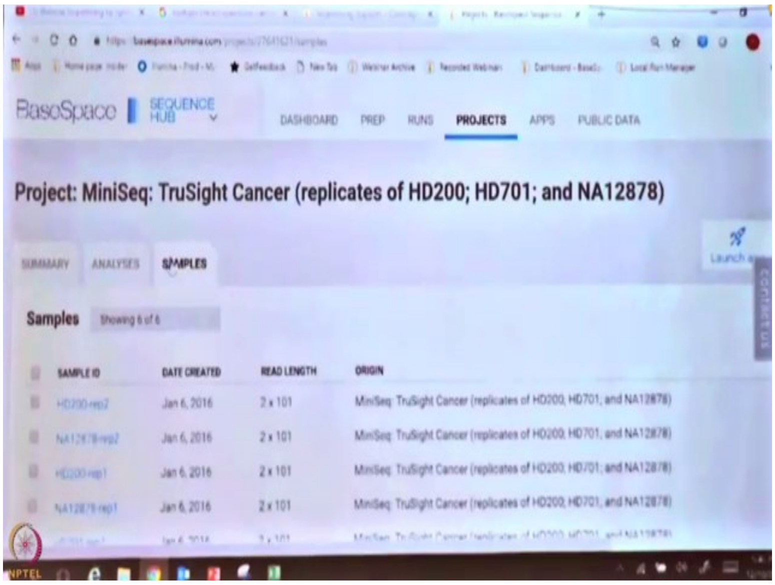
Analyses

Showing 2 of 2

NAME	LAST MODIFIED	APPLICATION	SIZE	COMMENT
Isaac Enrichment v2.1.0	January 6, 2016	Enrichment	1 GB	
FASTQ Generation 2016-01-06 00:36:57Z	January 6, 2016	FASTQ Generation	3 GB	

So, what I want you to do is click on that, you know click on this notification that says MiniSeq TruSight cancer. So, just click on that. Once you do, go to the samples tab. There is a tab here that says samples, go to the samples tab.

(Refer Slide Time: 07:52)

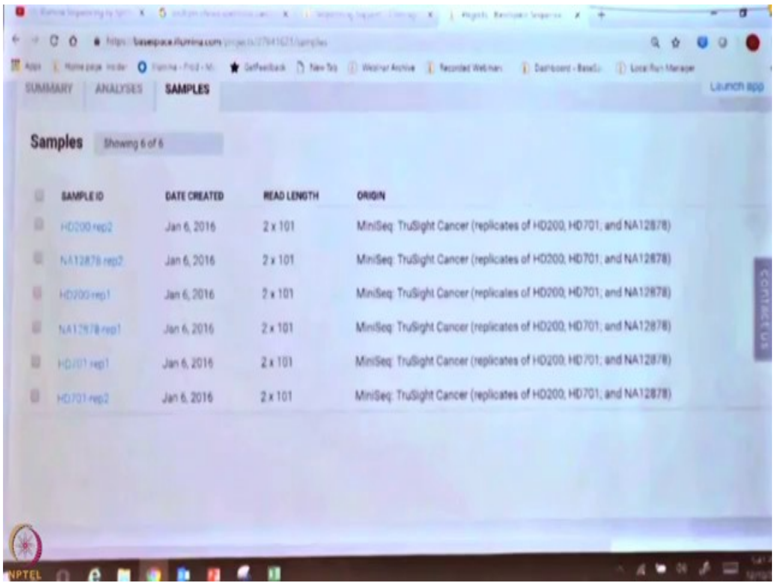


Project: MiniSeq: TruSight Cancer (replicates of HD200; HD701; and NA12878)

Samples Showing 6 of 6

SAMPLE ID	DATE CREATED	READ LENGTH	ORIGIN
HD200-rep2	Jan 6, 2016	2 x 101	MiniSeq: TruSight Cancer (replicates of HD200; HD701; and NA12878)
NA12878-rep2	Jan 6, 2016	2 x 101	MiniSeq: TruSight Cancer (replicates of HD200; HD701; and NA12878)
HD200-rep1	Jan 6, 2016	2 x 101	MiniSeq: TruSight Cancer (replicates of HD200; HD701; and NA12878)
NA12878-rep1	Jan 6, 2016	2 x 101	MiniSeq: TruSight Cancer (replicates of HD200; HD701; and NA12878)

(Refer Slide Time: 08:55)

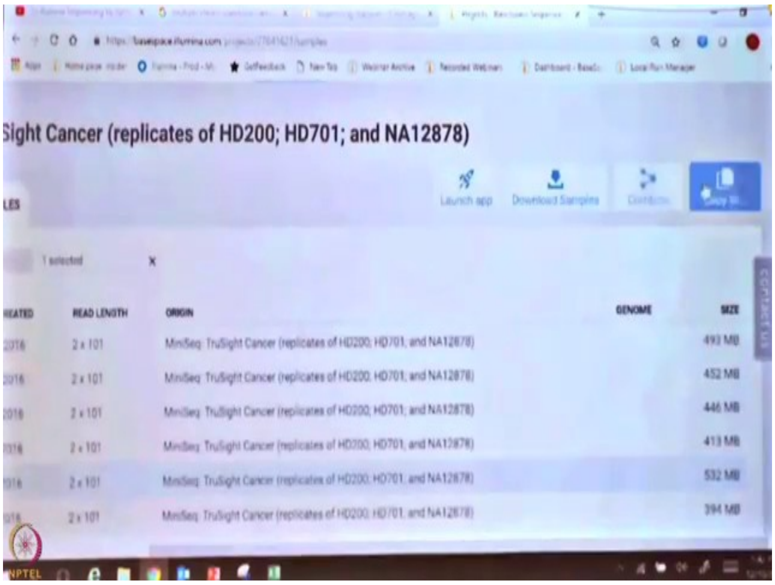


Samples Showing 6 of 6

SAMPLE ID	DATE CREATED	READ LENGTH	ORIGIN
HD200-rep2	Jan 6, 2016	2 x 101	MiniSeq: TruSight Cancer (replicates of HD200; HD701; and NA12878)
NA12878-rep2	Jan 6, 2016	2 x 101	MiniSeq: TruSight Cancer (replicates of HD200; HD701; and NA12878)
HD200-rep1	Jan 6, 2016	2 x 101	MiniSeq: TruSight Cancer (replicates of HD200; HD701; and NA12878)
NA12878-rep1	Jan 6, 2016	2 x 101	MiniSeq: TruSight Cancer (replicates of HD200; HD701; and NA12878)
HD701-rep1	Jan 6, 2016	2 x 101	MiniSeq: TruSight Cancer (replicates of HD200; HD701; and NA12878)
HD701-rep2	Jan 6, 2016	2 x 101	MiniSeq: TruSight Cancer (replicates of HD200; HD701; and NA12878)

Now, let us take one of the samples this guy, the second last one which says HD 701 rep 1. So, these are, this is a reference sample from Horizon, it is a company that provides reference RNA DNA samples, so that you can check the quality of the data, quality of the panels that you have developed and so on and so forth. So, it is essentially QC data right, but it is very easy to look at which is why I wanted to show this to you. So, you select this and what you do, I should be able to copy yeah. So, do copy.

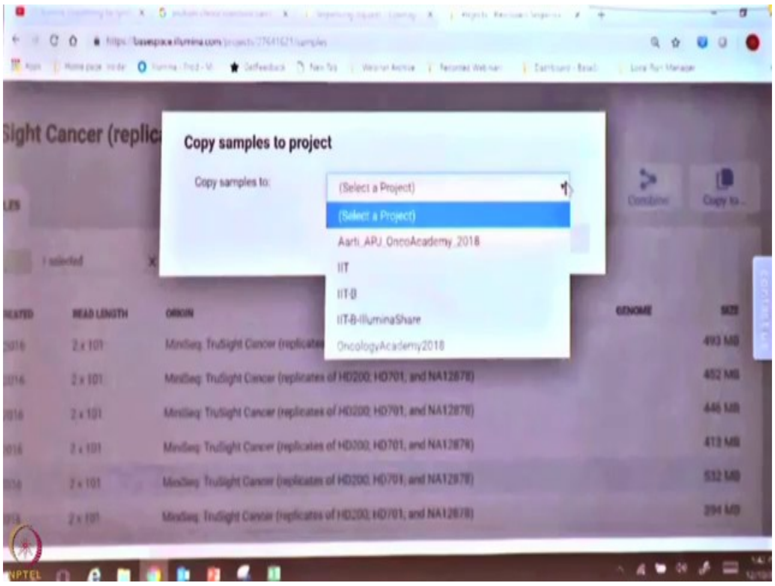
(Refer Slide Time: 08:37)



Sight Cancer (replicates of HD200; HD701; and NA12878)

CREATED	READ LENGTH	ORIGIN	GENOME	SIZE
2016	2 x 101	MiniSeq: TruSight Cancer (replicates of HD200; HD701; and NA12878)		493 MB
2016	2 x 101	MiniSeq: TruSight Cancer (replicates of HD200; HD701; and NA12878)		452 MB
2016	2 x 101	MiniSeq: TruSight Cancer (replicates of HD200; HD701; and NA12878)		446 MB
2016	2 x 101	MiniSeq: TruSight Cancer (replicates of HD200; HD701; and NA12878)		413 MB
2016	2 x 101	MiniSeq: TruSight Cancer (replicates of HD200; HD701; and NA12878)		532 MB
2016	2 x 101	MiniSeq: TruSight Cancer (replicates of HD200; HD701; and NA12878)		394 MB

(Refer Slide Time: 08:37)



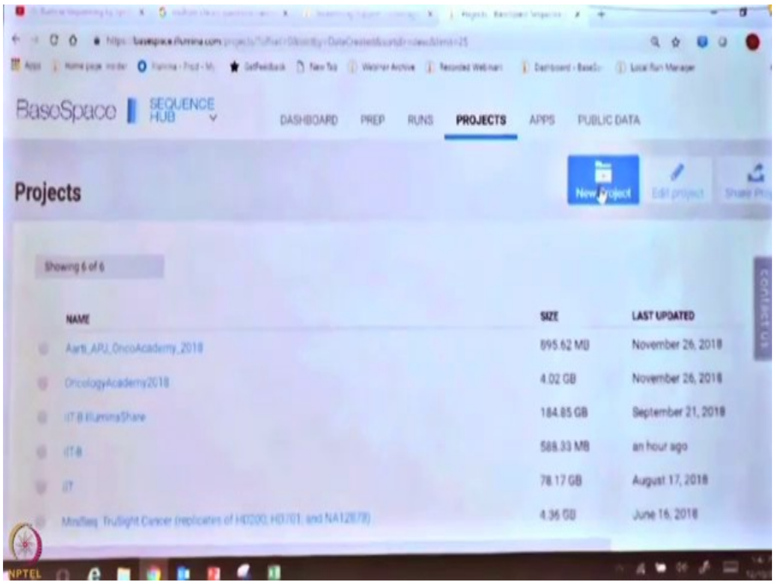
Copy samples to project

Copy samples to: (Select a Project)

- (Select a Project)
- Aarti_APJ_OncoAcademy_2018
- IIT
- IIT-B
- IIT-B-IlluminaShare
- OncologyAcademy2018

And do you guys have any projects; no ok. So, sorry my bad I missed a step. Let us exit this, cancel, let us cancel this ok.

(Refer Slide Time: 08:57)



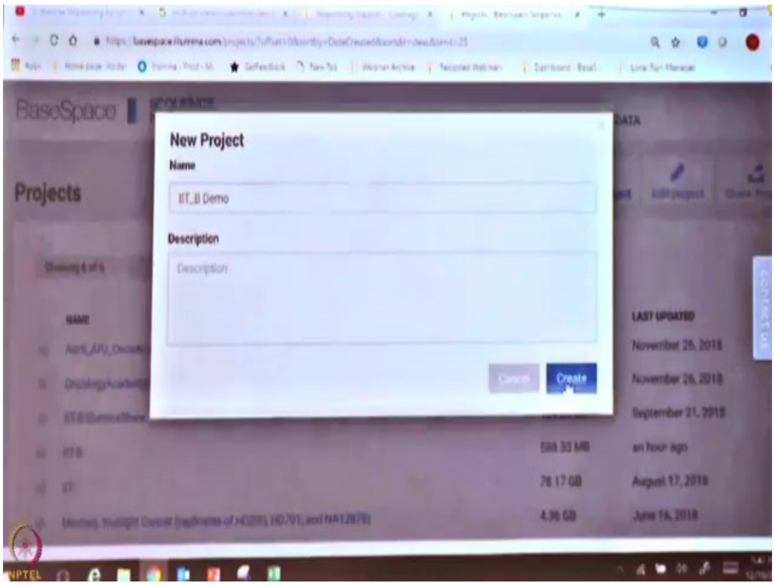
Projects

Showing 6 of 6

NAME	SIZE	LAST UPDATED
Aarti_APU_OncoAcademy_2018	895.62 MB	November 26, 2018
OncologyAcademy2018	4.02 GB	November 26, 2018
IIT B IlluminaShare	184.85 GB	September 21, 2018
IIT-B	588.33 MB	an hour ago
IIT	78.17 GB	August 17, 2018
MiniSeq: TruSight Cancer (replicates of HD200, HD701, and NA12878)	4.36 GB	June 16, 2018

Go back to projects, go back to projects ok. Let us exit that for a moment go back to projects. Is everybody in projects? Yeah, so just create a project.

(Refer Slide Time: 09:09)



New Project

Name

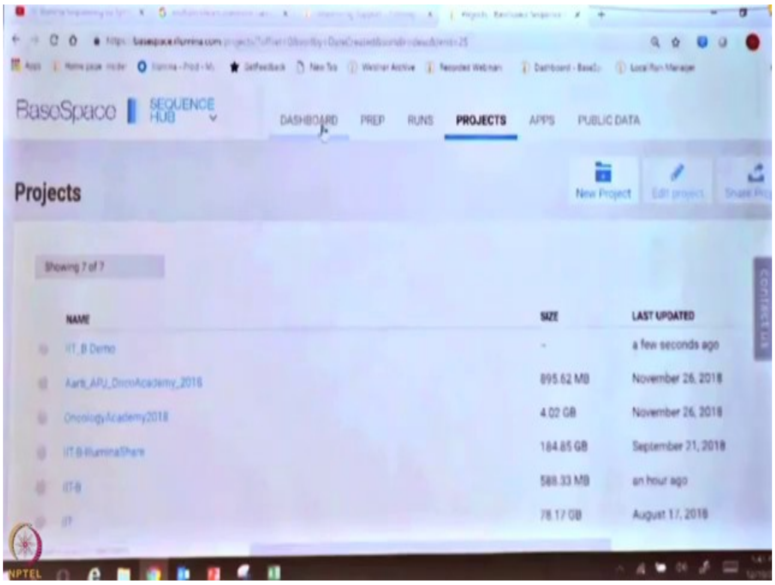
IIT_B Demo

Description

Description

You know name it whatever you guys want to name it. I just want to call this IITB Demo very creative. So, you can name the project whichever way you want.

(Refer Slide Time: 09:21)



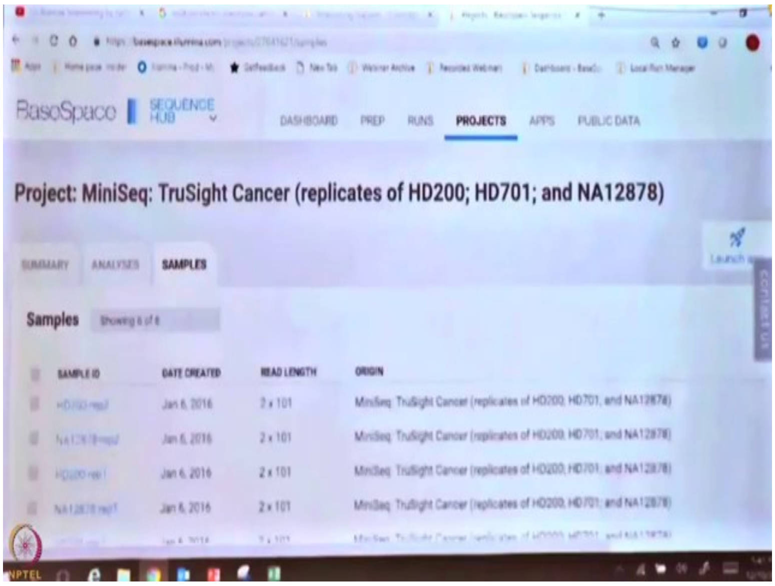
Projects

Showing 7 of 7

NAME	SIZE	LAST UPDATED
IIT_B Demo	-	a few seconds ago
Aarti_APJ_OncoAcademy_2018	895.62 MB	November 26, 2018
OncologyAcademy2018	4.02 GB	November 26, 2018
IIT-B-IlluminaShare	184.85 GB	September 21, 2018
IIT-B	588.33 MB	an hour ago
IIT	78.17 GB	August 17, 2018

So, it will create a project. Do you now have a project? Yes ok. Now, let us go back to the dashboard, let us go back to the dashboard now. Again click on the MiniSeq TruSight cancer project.

(Refer Slide Time: 09:45)



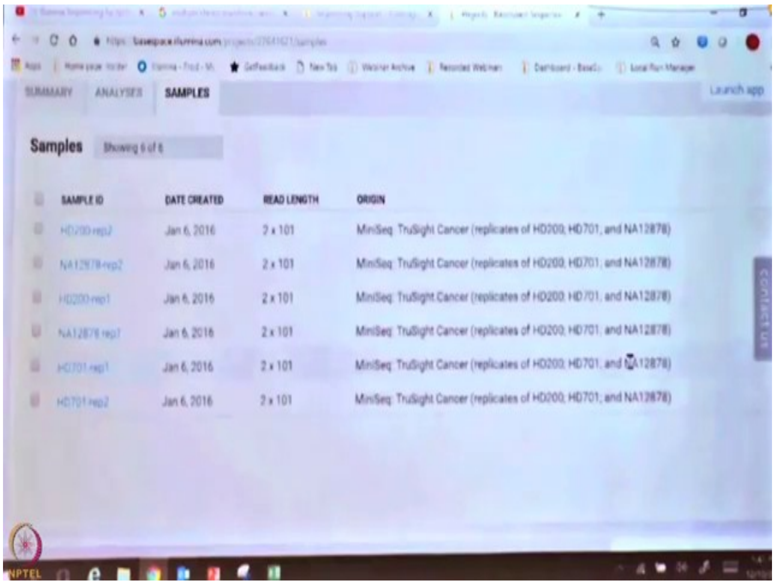
Project: MiniSeq: TruSight Cancer (replicates of HD200; HD701; and NA12878)

Samples Showing 6 of 6

SAMPLE ID	DATE CREATED	READ LENGTH	ORIGIN
HD200-rep2	Jan 6, 2016	2 x 101	MiniSeq: TruSight Cancer (replicates of HD200; HD701; and NA12878)
NA12878-rep2	Jan 6, 2016	2 x 101	MiniSeq: TruSight Cancer (replicates of HD200; HD701; and NA12878)
HD200-rep1	Jan 6, 2016	2 x 101	MiniSeq: TruSight Cancer (replicates of HD200; HD701; and NA12878)
NA12878-rep1	Jan 6, 2016	2 x 101	MiniSeq: TruSight Cancer (replicates of HD200; HD701; and NA12878)
[illegible]	Jan 6, 2016	2 x 101	MiniSeq: TruSight Cancer (replicates of HD200; HD701; and NA12878)

Go to samples.

(Refer Slide Time: 09:49)

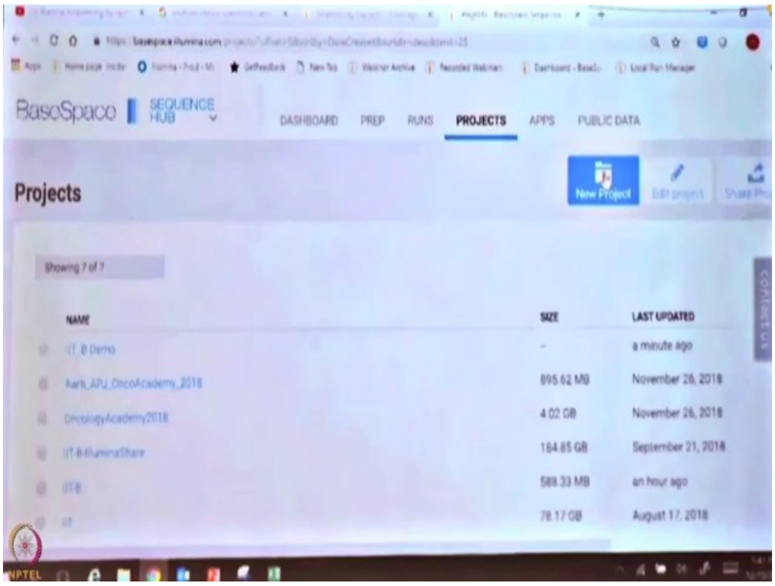


Samples Showing 6 of 6

SAMPLE ID	DATE CREATED	READ LENGTH	ORIGIN
HD200-rep2	Jan 6, 2016	2 x 101	MiniSeq: TruSight Cancer (replicates of HD200, HD701, and NA12878)
NA12878-rep2	Jan 6, 2016	2 x 101	MiniSeq: TruSight Cancer (replicates of HD200, HD701, and NA12878)
HD200-rep1	Jan 6, 2016	2 x 101	MiniSeq: TruSight Cancer (replicates of HD200, HD701, and NA12878)
NA12878-rep1	Jan 6, 2016	2 x 101	MiniSeq: TruSight Cancer (replicates of HD200, HD701, and NA12878)
HD701-rep1	Jan 6, 2016	2 x 101	MiniSeq: TruSight Cancer (replicates of HD200, HD701, and NA12878)
HD701-rep2	Jan 6, 2016	2 x 101	MiniSeq: TruSight Cancer (replicates of HD200, HD701, and NA12878)

Lost, sorry I will go step by step. So, you created a project? No, so ok. Do you see this projects here? Yeah.

(Refer Slide Time: 10:02)



Projects

Showing 7 of 7

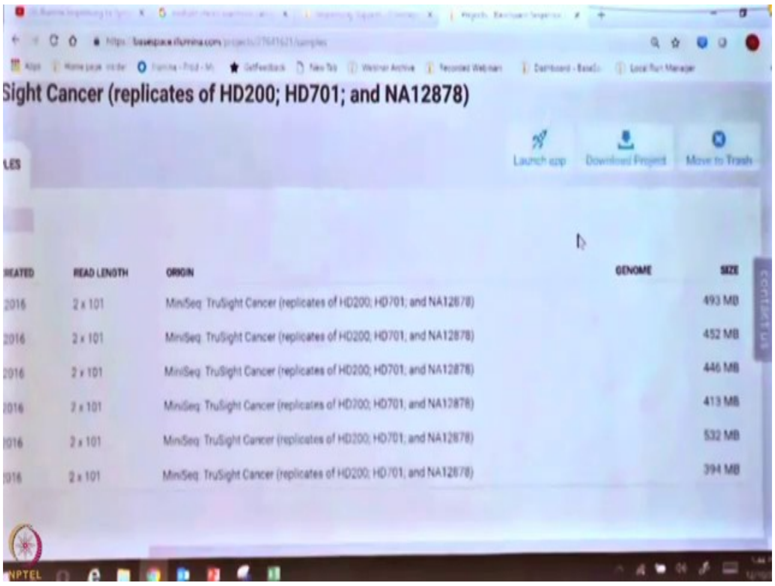
NAME	SIZE	LAST UPDATED
IIT_B Demo	-	a minute ago
Aarti_APU_OncoAcademy_2018	895.62 MB	November 26, 2018
OncologyAcademy2018	4.02 GB	November 26, 2018
IIT-B-IlluminaShare	184.85 GB	September 21, 2018
IIT-B	588.33 MB	an hour ago
IIT	78.17 GB	August 17, 2018

If you click on that there is a like a file icon which says new project ok. So, click on that and give a name to the project. Everybody has created the projects; yeah ok. So, let us go back to the dashboard into the TruSight cancer project, go to samples.

And go to this HD 701 rep 1 that is the sample that we want to copy. So, if you scroll to your right on the top, there is something that you see copy to, that menu option that says copy to.

So, you can copy this, now you should have a project the one that you just created and copy there ok.

(Refer Slide Time: 11:01)

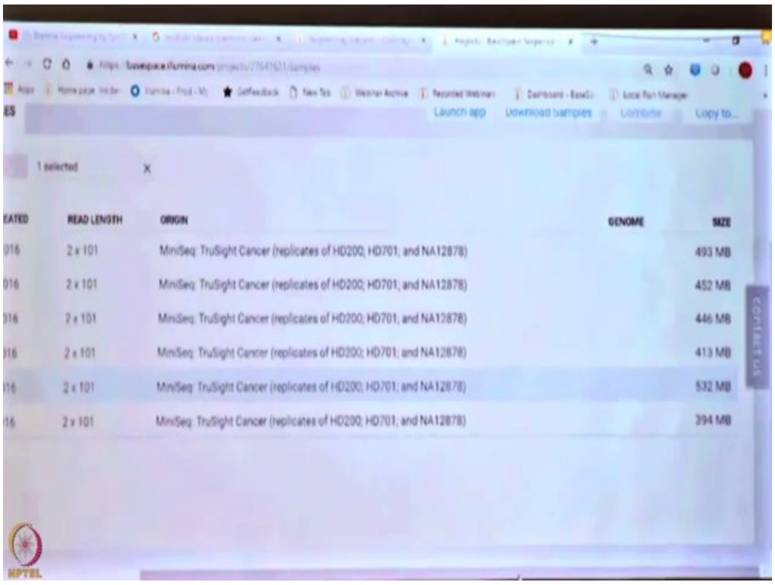


Sight Cancer (replicates of HD200; HD701; and NA12878)

EATED	READ LENGTH	ORIGIN	GENOME	SIZE
2016	2 x 101	MiniSeq: TruSight Cancer (replicates of HD200; HD701; and NA12878)		493 MB
2016	2 x 101	MiniSeq: TruSight Cancer (replicates of HD200; HD701; and NA12878)		452 MB
2016	2 x 101	MiniSeq: TruSight Cancer (replicates of HD200; HD701; and NA12878)		446 MB
2016	2 x 101	MiniSeq: TruSight Cancer (replicates of HD200; HD701; and NA12878)		413 MB
2016	2 x 101	MiniSeq: TruSight Cancer (replicates of HD200; HD701; and NA12878)		532 MB
2016	2 x 101	MiniSeq: TruSight Cancer (replicates of HD200; HD701; and NA12878)		394 MB

The reason we are doing this is, I wanted to show you how data sharing works on BaseSpace right.

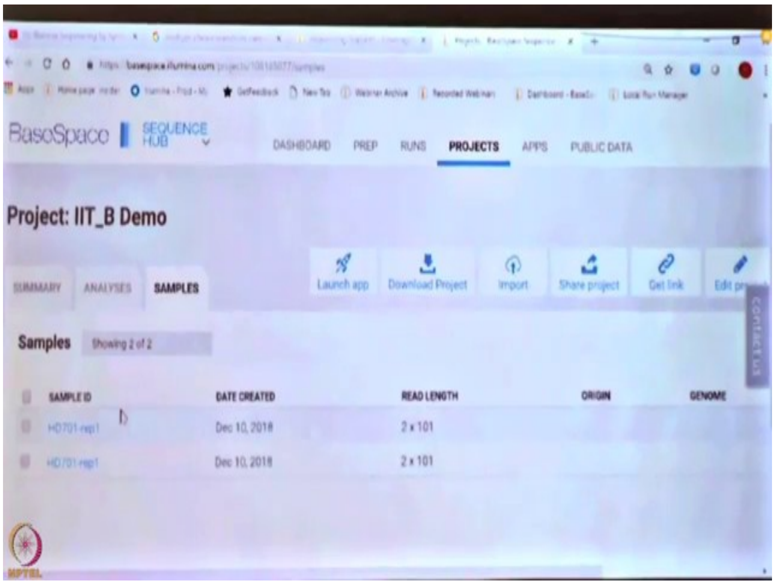
(Refer Slide Time: 11:11)



EATED	READ LENGTH	ORIGIN	GENOME	SIZE
016	2 x 101	MiniSeq: TruSight Cancer (replicates of HD200; HD701; and NA12878)		493 MB
016	2 x 101	MiniSeq: TruSight Cancer (replicates of HD200; HD701; and NA12878)		452 MB
016	2 x 101	MiniSeq: TruSight Cancer (replicates of HD200; HD701; and NA12878)		446 MB
016	2 x 101	MiniSeq: TruSight Cancer (replicates of HD200; HD701; and NA12878)		413 MB
16	2 x 101	MiniSeq: TruSight Cancer (replicates of HD200; HD701; and NA12878)		532 MB
16	2 x 101	MiniSeq: TruSight Cancer (replicates of HD200; HD701; and NA12878)		394 MB

Like I said this is not a very very critical function in the application, the only thing we wanted to show you is how you can share data across different users, because this is meant to be a collaborative platform. If there are any public datasets that you want to analyze, this is how you would import those datas into your own workspace, so that you can run your own analysis ok. That really was the crux and what else I wanted to show you was how to launch an app ok. So, let us say now am just sake of time going to copy this one particular file on to the project that I have just created. Now, let me go back to projects, let us go back to projects.

(Refer Slide Time: 11:59)



Project: IIT_B Demo

Samples Showing 2 of 2

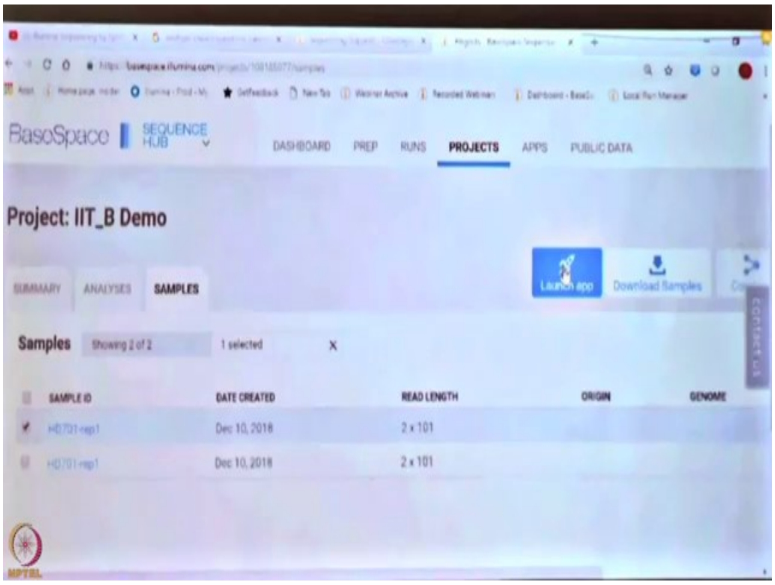
SAMPLE ID	DATE CREATED	READ LENGTH	ORIGIN	GENOME
HD701-rep1	Dec 10, 2018	2 x 101		
HD701-rep1	Dec 10, 2018	2 x 101		

And in this whatever project that you have just created, if you are able to copy the file, copy that particular sample it should be there ok. I have, I think I have copied it twice, so you see two times, but if you have been able to create the project, select the sample and copy it to the project that you created, you should be able to see a sample in your project. So, what I wanted to show you now is launching an analysis.

Once you have your data into BaseSpace how do you launch an analysis, and I just said that there are multiple apps or applications that are available on BaseSpace. So, this particular data was generated for a panel known as your TruSight cancer ok, this is a panel that is available from Illumina that is used yes.

So, this data has been generated for a panel known as TruSight cancer that is available from Illumina. it is a panel that is designed to detect germline mutations ok. So, what we are going to do now, is once I select the sample that I want to run the analysis for, I am going to click on this launch app.

(Refer Slide Time: 13:11)



Project: IIT_B Demo

Samples Showing 2 of 2 1 selected ×

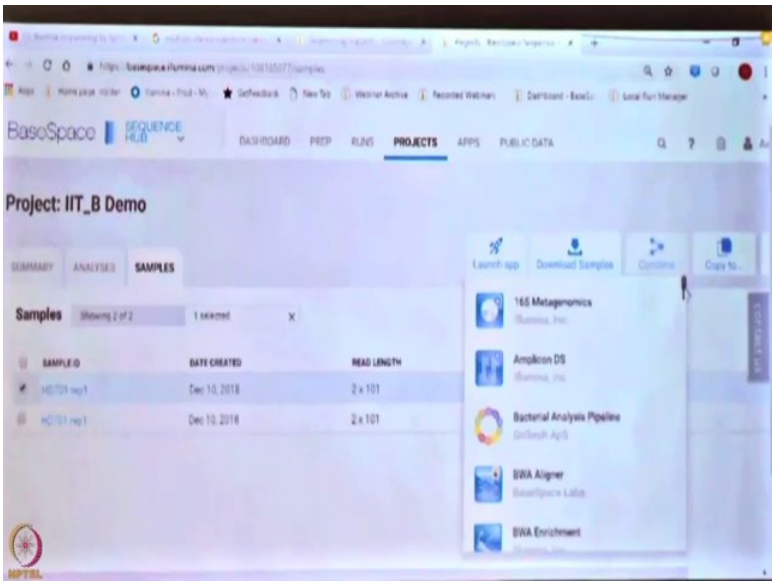
	SAMPLE ID	DATE CREATED	READ LENGTH	ORIGIN	GENOME
✓	HD701-rep1	Dec 10, 2018	2 x 101		
□	HD701-rep1	Dec 10, 2018	2 x 101		

Everybody or at least most of you are with me so far yeah. You have a project.

Student: Yes, we have…

But those of you who can see this, select the file and click on launch app.

(Refer Slide Time: 13:27)



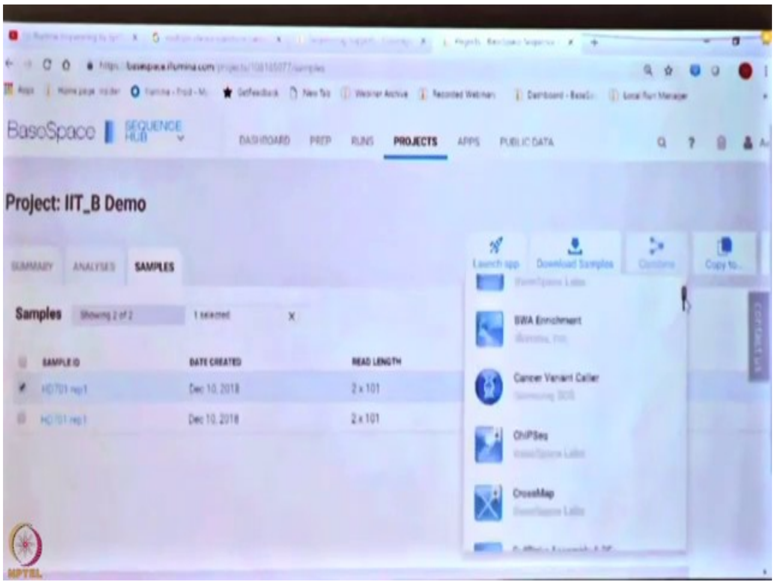
Project: IIT_B Demo

Samples Showing 2 of 2 1 selected ×

	SAMPLE ID	DATE CREATED	READ LENGTH
✓	HD701-rep1	Dec 10, 2018	2 x 101
□	HD701-rep1	Dec 10, 2018	2 x 101

When you click on launch app, the options that are available ok, let me just do this.

(Refer Slide Time: 13:36)

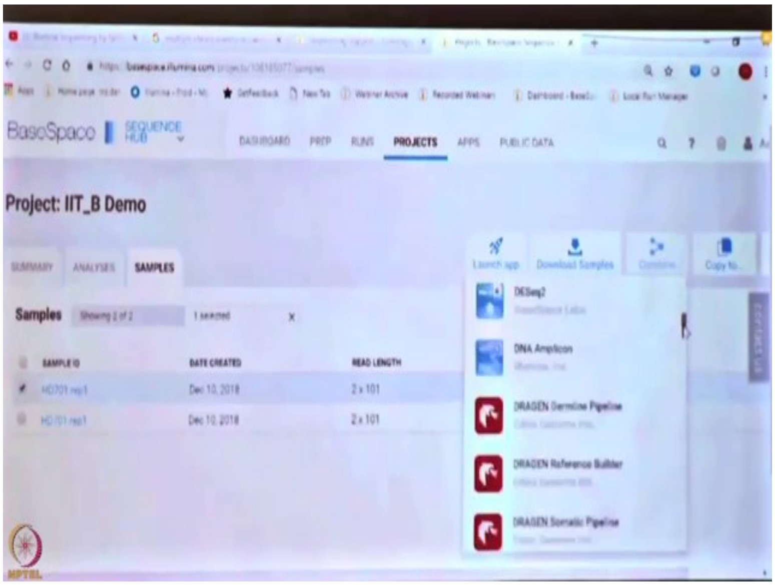


Project: IIT_B Demo

Samples Showing 2 of 2 1 selected

	SAMPLE ID	DATE CREATED	READ LENGTH
✓	HD701 rep1	Dec 10, 2018	2 x 101
□	HD701 rep1	Dec 10, 2018	2 x 101

(Refer Slide Time: 13:38)

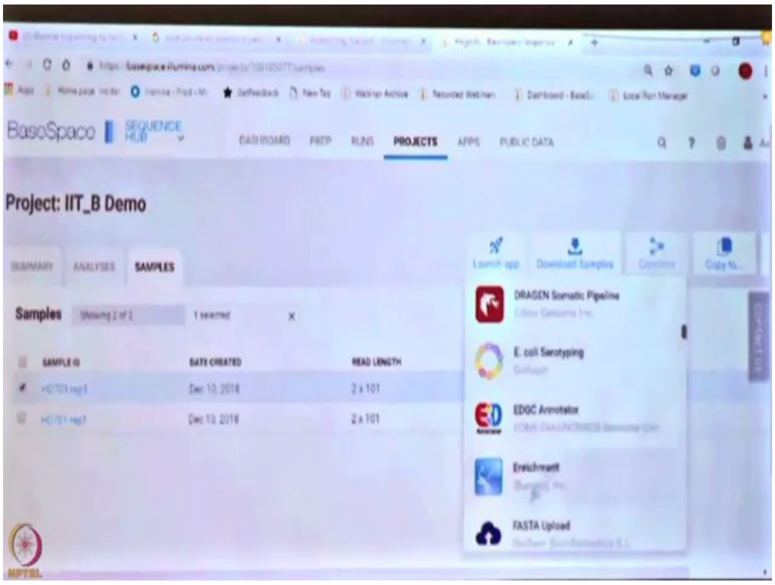


Project: IIT_B Demo

Samples Showing 2 of 2 1 selected

	SAMPLE ID	DATE CREATED	READ LENGTH
✓	HD701 rep1	Dec 10, 2018	2 x 101
□	HD701 rep1	Dec 10, 2018	2 x 101

(Refer Slide Time: 13:39)

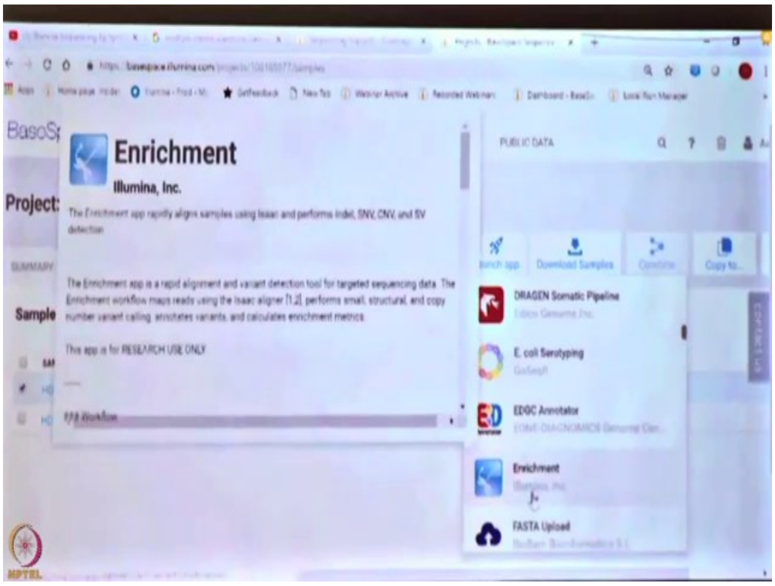


Project: IIT_B Demo

SAMPLE ID	DATE CREATED	READ LENGTH
HD701 rep1	Dec 10, 2018	2 x 101
HD701 rep1	Dec 10, 2018	2 x 101

There are multiple apps that are available, depending again on the analysis that you want to do ok, because this is a cancer panel that you know we have done in enrichment on.

(Refer Slide Time: 13:50)



Enrichment

Illumina, Inc.

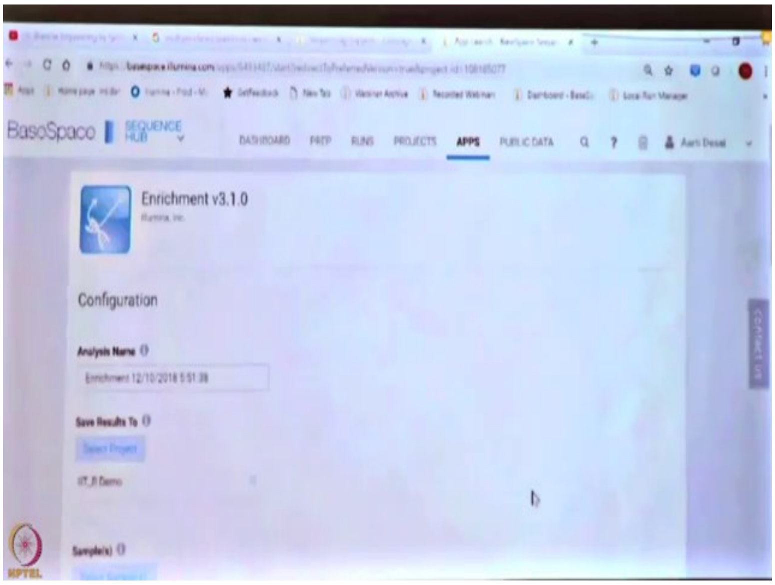
The Enrichment app rapidly aligns samples using Isaac and performs indel, SNV, CNV, and SV detection

The Enrichment app is a rapid alignment and variant detection tool for targeted sequencing data. The Enrichment workflow maps reads using the Isaac aligner [1,2], performs small, structural, and copy number variant calling, annotates variants, and calculates enrichment metrics.

This app is for RESEARCH USE ONLY

I can choose an enrichment app ok. Now what this will do is, know it will run what is known as variant calling ok, it will call mutations, it will call insertions, small indels ok.

(Refer Slide Time: 14:11)



Enrichment v3.1.0

Illumina, Inc.

Configuration

Analysis Name

Enrichment 12/10/2018 5:51:38

Save Results To

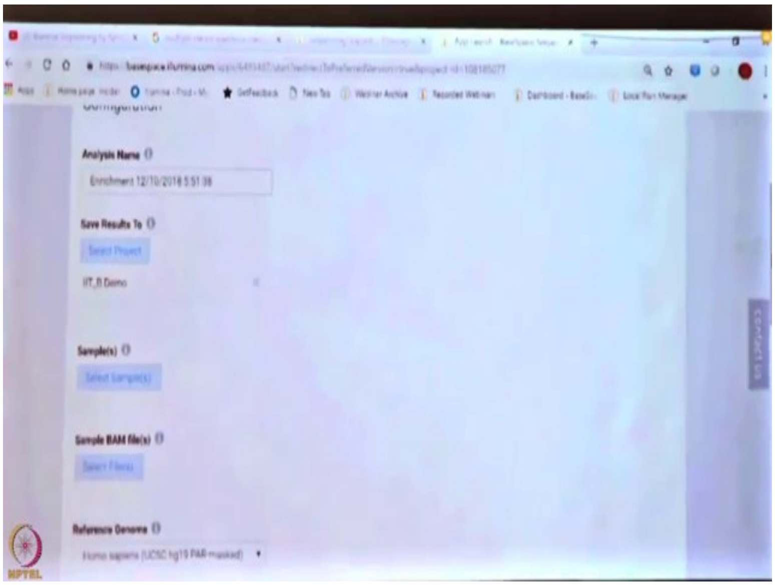
Select Project

IIT_B Demo

Sample(s)

And, then I can select the enrichment app. So, what I did was I selected a file from my project, I clicked on launch app, from the launch app I selected the enrichment app, because the data that I am using for the workshop today is from a cancer panel TruSight cancer. And, we want to run an app that will give me variant calls you know mutations and small insertion deletions ok, and these are again you know our third party apps.

(Refer Slide Time: 14:38)



Analysis Name

Enrichment 12/10/2018 5:51:38

Save Results To

Select Project

IIT_B Demo

Sample(s)

Select Sample(s)

Sample BAM file(s)

Select File(s)

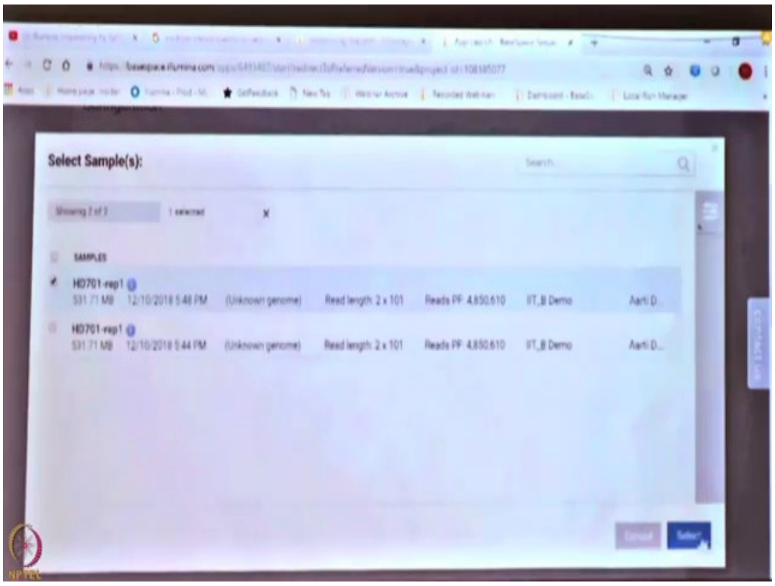
Reference Genome

Homo sapiens (UCSC hg19 PAR-masked)

See the whole idea behind BaseSpace is to make these apps which are traditionally you know developed as command line tools, available to the end users. Because, if they are in the command line form unless you know if you are good if you are, unless you are good at some level of you know scripting it is hard for you to use it. And, majority of the apps that are developed for NGS data analysis are command line apps, because they are developed by advanced bioinformaticians ok.

So, there is an analysis name you know which is the name that you want a unique name that you want to give to your analysis, you know that is again your call where do you want to save your results. So, you by default it will pick up the same project from which you have picked up the input data files. If you want to save it somewhere else you can do so.

(Refer Slide Time: 15:30)



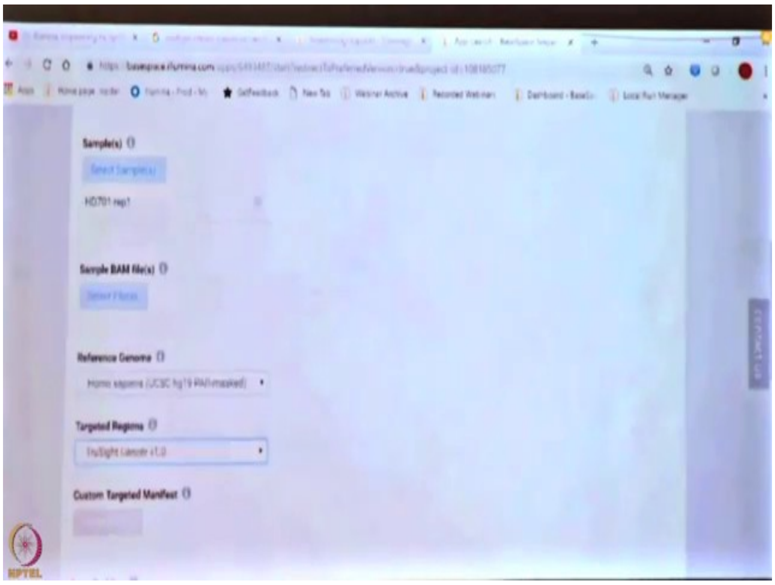
Select Sample(s):

Showing 2 of 2 | 1 selected

	SAMPLES						
☑	HD701-rep1 531.71 MB	12/10/2018 5:48 PM	(Unknown genome)	Read length: 2 x 101	Reads PF: 4,850,610	IIT_B Demo	Aarti D...
□	HD701-rep1 531.71 MB	12/10/2018 5:44 PM	(Unknown genome)	Read length: 2 x 101	Reads PF: 4,850,610	IIT_B Demo	Aarti D...

Select samples. So, you can go to the project, select any sample that you want to run, select ok.

(Refer Slide Time: 15:38)



Sample(s)

Select Sample(s)

HD701 rep1

Sample BAM file(s)

Select File(s)

Reference Genome

Homo sapiens (UCSC hg19 PAR-masked)

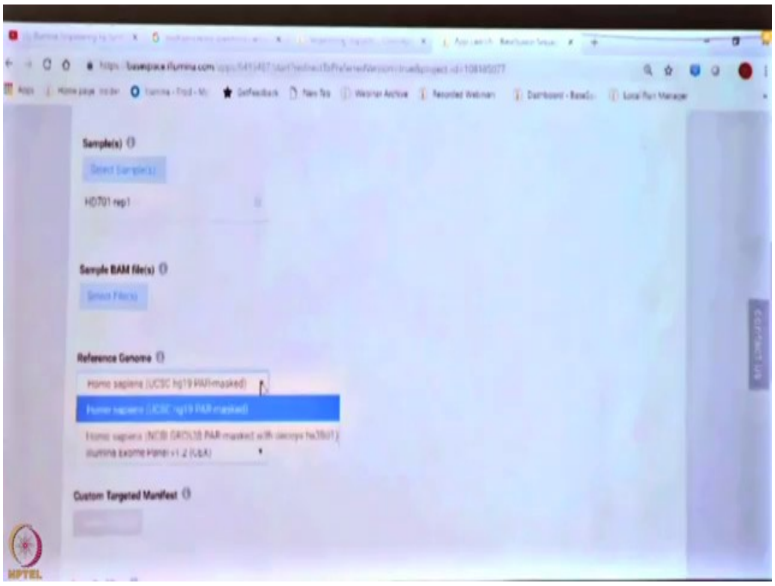
Targeted Regions

TruSight Cancer v1.0

Custom Targeted Manifest

If there are BAM file so, BAM files are nothing, but aligned data files that are generated by aligners.

(Refer Slide Time: 15:48)



Sample(s)

Select Sample(s)

HD701 rep1

Sample BAM file(s)

Select File(s)

Reference Genome

Homo sapiens (UCSC hg19 PAR-masked)

Homo sapiens (UCSC hg19 PAR-masked)

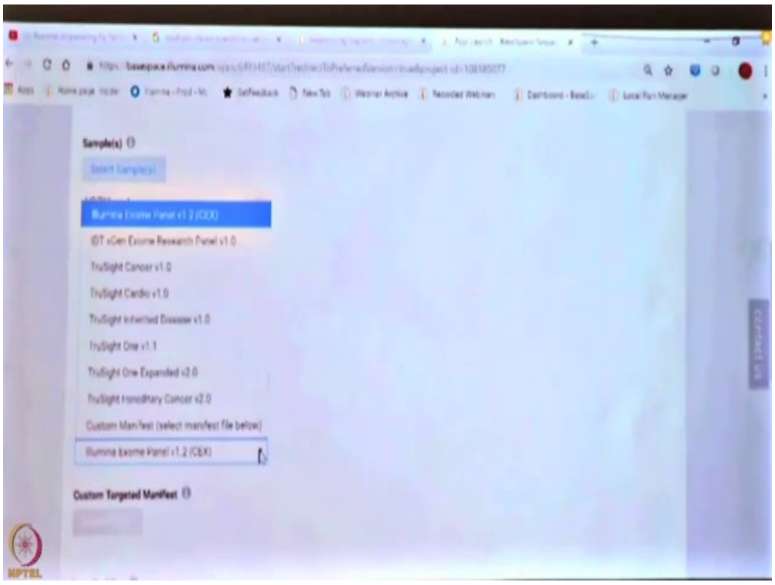
Homo sapiens (NCBI GRCh38 PAR-masked with decoys hs38d1)

Illumina Exome Panel v1.2 (U&A)

Custom Targeted Manifest

So, you can select that, reference genome that you want to use for your analysis, because if you remember from the video I showed you, once you generate the read data in order for you to do analysis, you have to map it back to the reference genome right and in this case we are dealing with human genome data, not unknown species. So, we will have a reference genome against which we will map our data files.

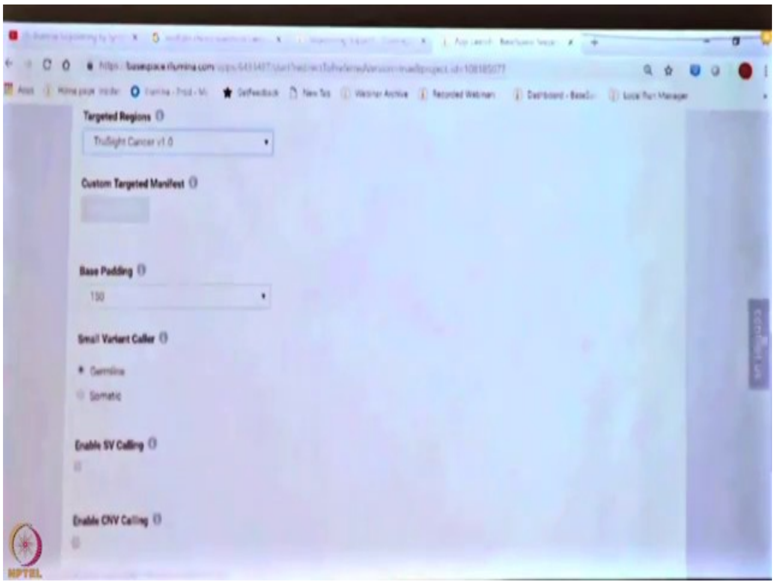
(Refer Slide Time: 16:15)


https://basespace.illumina.com
Sample(s)
Select Sample(s)
Illumina Exome Panel v1.2 (CEX)
IDT xGen Exome Research Panel v1.0
TruSight Cancer v1.0
TruSight Cardio v1.0
TruSight Inherited Disease v1.0
TruSight One v1.1
TruSight One Expanded v2.0
TruSight Hereditary Cancer v2.0
Custom Manifest (select manifest file below)
Illumina Exome Panel v1.2 (CEX)
Custom Targeted Manifest
NPTEL

And you can also choose the targeted panel, so that was used. So, in this case as I told you this is a TruSight cancer data right. So, what I will do is, I will choose this TruSight cancer, version 1 region file. So, essentially this will define for the application what are the regions in which it should look for variant calls right. So, it is significantly shorten the time it takes for analysis, otherwise the app may end up scanning the entire genome which may result in two things. One is extremely long analysis time and second some of these, because we have repetitive regions in the genome, homologous regions in the genome, you know it you may end up with random alignments which will give you false information right.

So, the targeted, the target files essentially help you in streamlining your analysis ok. And now let us hope that it works ok, I am just going to check that I have actually done everything that I am supposed to ok. I think there are more things.

(Refer Slide Time: 17:33)



Targeted Regions

TruSight Cancer v1.0

Custom Targeted Manifest

Base Padding

150

Small Variant Caller

● Germline

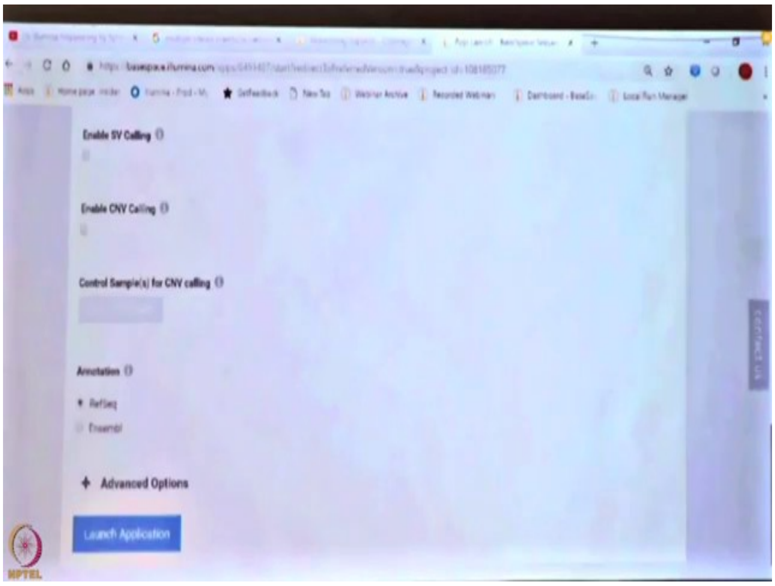
○ Somatic

Enable SV Calling

Enable CNV Calling

If you continue scrolling down, you will see that there are certain other options; for example, whether this is for germline or for somatic. I told you earlier the TruSight cancer panel is meant for germline ok. Germline variant detection so, that is what I will use.

(Refer Slide Time: 17:54)



Enable SV Calling

Enable CNV Calling

Control Sample(s) for CNV calling

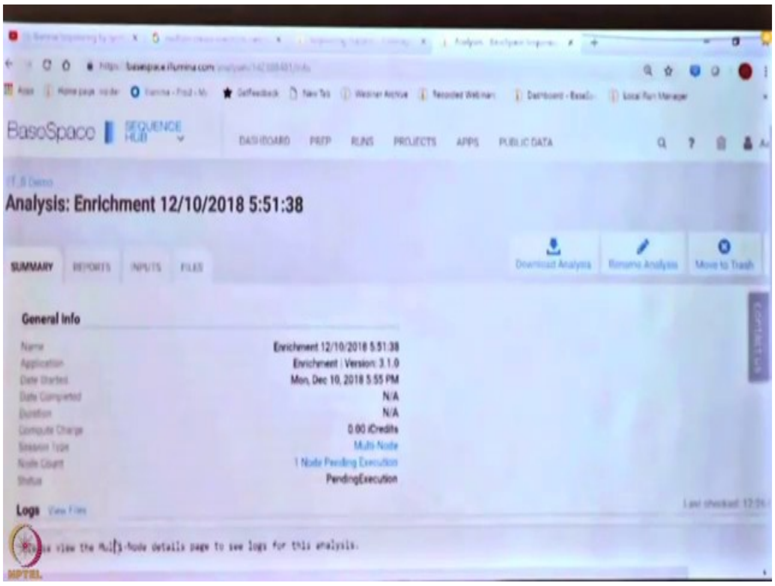
Annotation

● RefSeq

○ Ensembl

We are going to really leave all the other parameters as default right. There are many-many options, because most of these algorithms come with multiple options that you can use to tweak the way the analysis is performed and the kind of results that are reported. We recommend for our new users, basic users to use a default analysis as you become more comfortable with the data that you are handling, you can play around with the parameter options. And, I am going to I think I have done everything yup. So, I am going to launch the application.

(Refer Slide Time: 18:28)



Analysis: Enrichment 12/10/2018 5:51:38

General Info

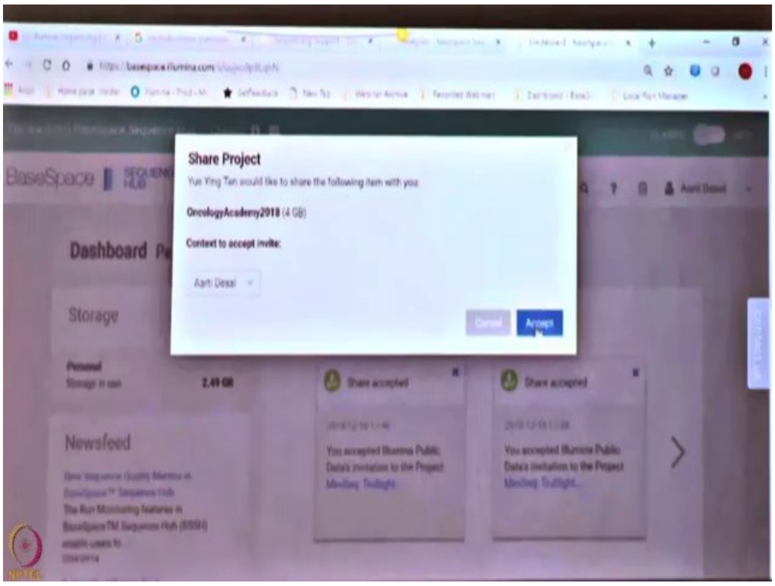
Name	Enrichment 12/10/2018 5:51:38
Application	Enrichment | Version: 3.1.0
Date Started	Mon, Dec 10, 2018 5:55 PM
Date Completed	N/A
Duration	N/A
Compute Charge	0.00 iCredits
Session Type	Multi-Node
Node Count	1 Node Pending Execution
Status	PendingExecution

Logs View Files

Please view the Multi-Node details page to see logs for this analysis.

So, when I do this, what the application should do is take the input files that I have given and start running for the analysis.

(Refer Slide Time: 18:40)



Share Project

Yue Ying Tan would like to share the following item with you:

OncologyAcademy2018 (4 GB)

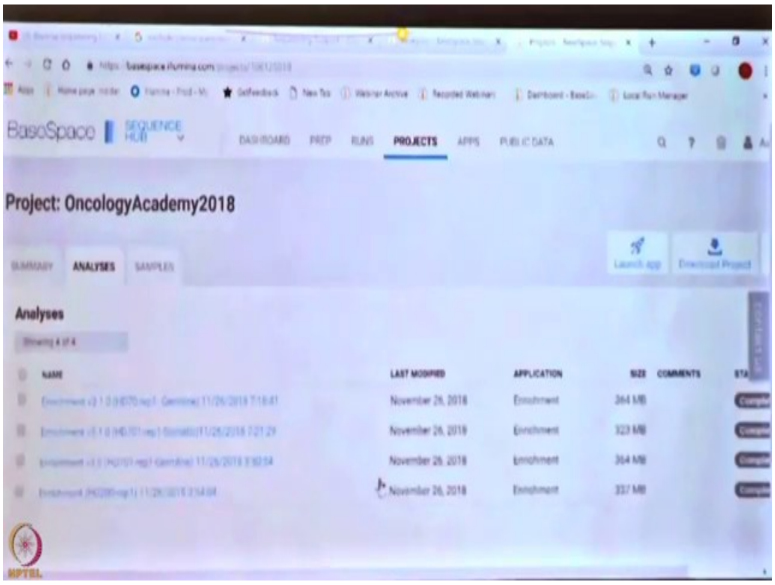
Context to accept invite:

Aarti Desai

Cancel
Accept
BaseSpace SEQUENCE HUB
Dashboard
Storage
Personal
Storage in use
2.49 GB
Newsfeed
Share accepted
Share accepted
You accepted Illumina Public Data's invitation to the Project
You accepted Illumina Public Data's invitation to the Project

This analysis will take a little bit of time, you will get a small pop up; accept it.

(Refer Slide Time: 18:53)



BaseSpace SEQUENCE HUB

DASHBOARD PREP RUNS PROJECTS APPS PUBLIC DATA

Project: OncologyAcademy2018

SUMMARY ANALYSES SAMPLES

Launch app Download Project

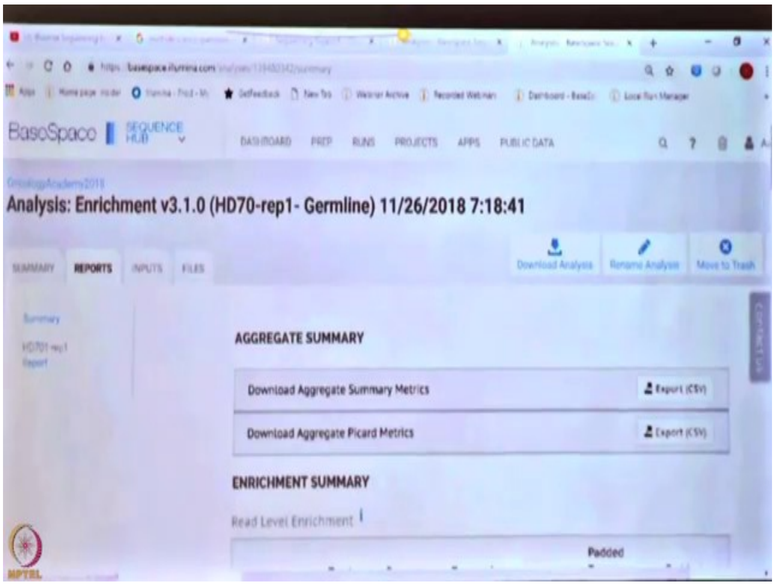
Analyses

Showing 4 of 4

NAME	LAST MODIFIED	APPLICATION	SIZE	COMMENTS	STA
Enrichment v3.1.0 (HD70-rep1- Germline) 11/26/2018 7:18:41	November 26, 2018	Enrichment	364 MB		Comple
Enrichment v3.1.0 (HD701-rep1 Somatic) 11/26/2018 7:21:29	November 26, 2018	Enrichment	323 MB		Comple
Enrichment v3.1 (HD701-rep1 Germline) 11/26/2018 [illegible]	November 26, 2018	Enrichment	364 MB		Comple
Enrichment (HD200-rep1) 11/26/2018 [illegible]	November 26, 2018	Enrichment	337 MB		Comple

Now, you should have one more notification saying that you know you have accepted data from one of my colleagues, and in that let us look at. Yeah let us look at the first analysis.

(Refer Slide Time: 19:18)



BaseSpace SEQUENCE HUB

DASHBOARD PREP RUNS PROJECTS APPS PUBLIC DATA

OncologyAcademy2018

Analysis: Enrichment v3.1.0 (HD70-rep1- Germline) 11/26/2018 7:18:41

SUMMARY REPORTS INPUTS FILES

Download Analysis Rename Analysis Move to Trash

Summary

HD701-rep1 Report

AGGREGATE SUMMARY

Download Aggregate Summary Metrics Export (CSV)

Download Aggregate Picard Metrics Export (CSV)

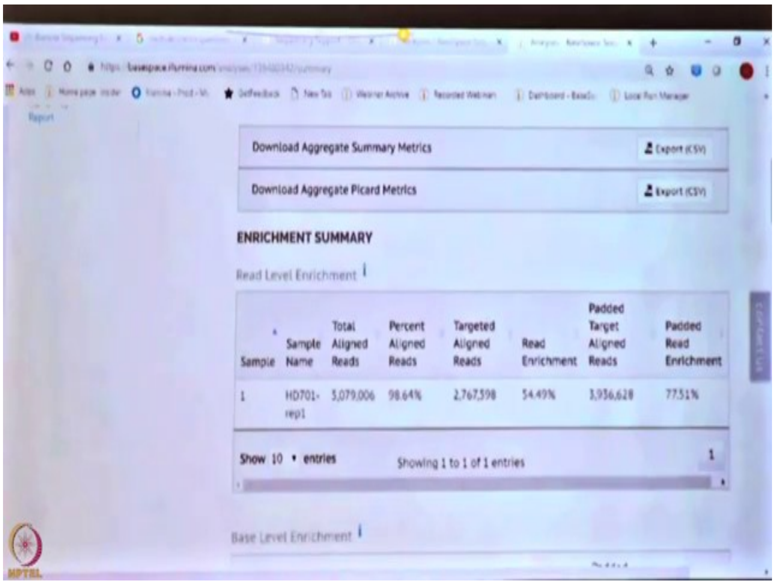
ENRICHMENT SUMMARY

Read Level Enrichment

Padded

So, just click on it, click on the first analysis. You will see four analysis in the, you know once you put in that address ok.

(Refer Slide Time: 19:27)



Download Aggregate Summary Metrics — Export (CSV)

Download Aggregate Picard Metrics — Export (CSV)

ENRICHMENT SUMMARY

Read Level Enrichment

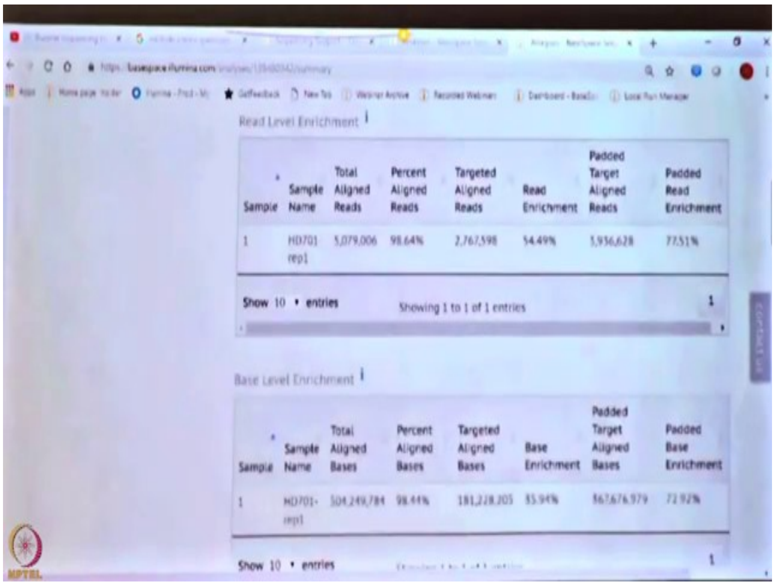
Sample	Sample Name	Total Aligned Reads	Percent Aligned Reads	Targeted Aligned Reads	Read Enrichment	Padded Target Aligned Reads	Padded Read Enrichment
1	HD701-rep1	5,079,006	98.64%	2,767,598	54.49%	3,936,628	77.51%

Show 10 entries Showing 1 to 1 of 1 entries 1

Base Level Enrichment

So, this is essentially going to be the output of the analysis that we just started. So, this is the output of the analysis that we just started ok.

(Refer Slide Time: 19:40)



Read Level Enrichment

Sample	Sample Name	Total Aligned Reads	Percent Aligned Reads	Targeted Aligned Reads	Read Enrichment	Padded Target Aligned Reads	Padded Read Enrichment
1	HD701 rep1	5,079,006	98.64%	2,767,598	54.49%	3,936,628	77.51%

Show 10 entries Showing 1 to 1 of 1 entries 1

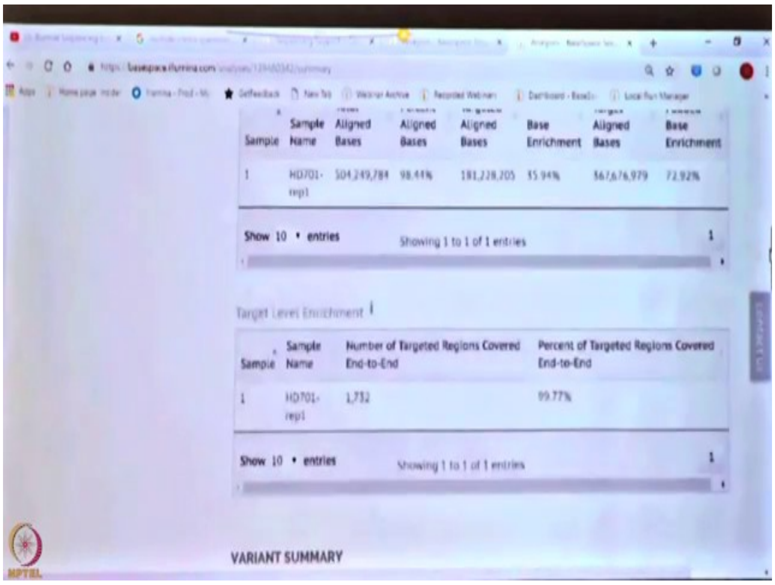
Base Level Enrichment

Sample	Sample Name	Total Aligned Bases	Percent Aligned Bases	Targeted Aligned Bases	Base Enrichment	Padded Target Aligned Bases	Padded Base Enrichment
1	HD701-rep1	504,249,784	98.44%	181,228,205	35.94%	367,676,979	72.92%

Show 10 entries 1

And what this is showing you is some metrics of the data that we have analyzed. So, you can see that for this particular sample, more than 98 percent reads aligned back to the reference genome ok. Again which means majority of the data is use able. More than 98 percent of the data, read data that was generated for this particular sample is mapping back to the reference genome, which means it is usable.

(Refer Slide Time: 20:14)



Sample	Sample Name	Aligned Bases	Aligned Bases	Aligned Bases	Base Enrichment	Aligned Bases	Base Enrichment
1	HD701-rep1	504,249,784	98.44%	181,228,205	35.94%	367,676,979	72.92%

Show 10 • entries Showing 1 to 1 of 1 entries 1

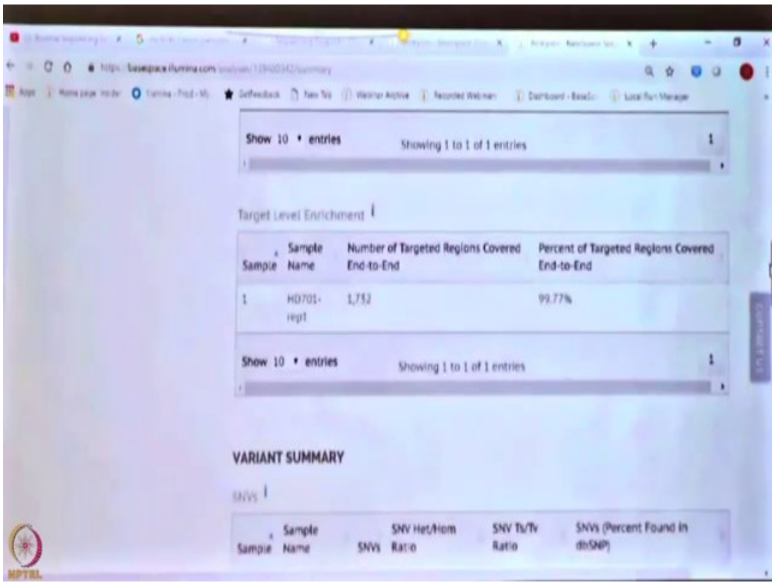
Target Level Enrichment

Sample	Sample Name	Number of Targeted Regions Covered End-to-End	Percent of Targeted Regions Covered End-to-End
1	HD701-rep1	1,732	99.77%

Show 10 • entries Showing 1 to 1 of 1 entries 1

VARIANT SUMMARY

(Refer Slide Time: 20:16)



Show 10 • entries Showing 1 to 1 of 1 entries 1

Target Level Enrichment

Sample	Sample Name	Number of Targeted Regions Covered End-to-End	Percent of Targeted Regions Covered End-to-End
1	HD701-rep1	1,732	99.77%

Show 10 • entries Showing 1 to 1 of 1 entries 1

VARIANT SUMMARY

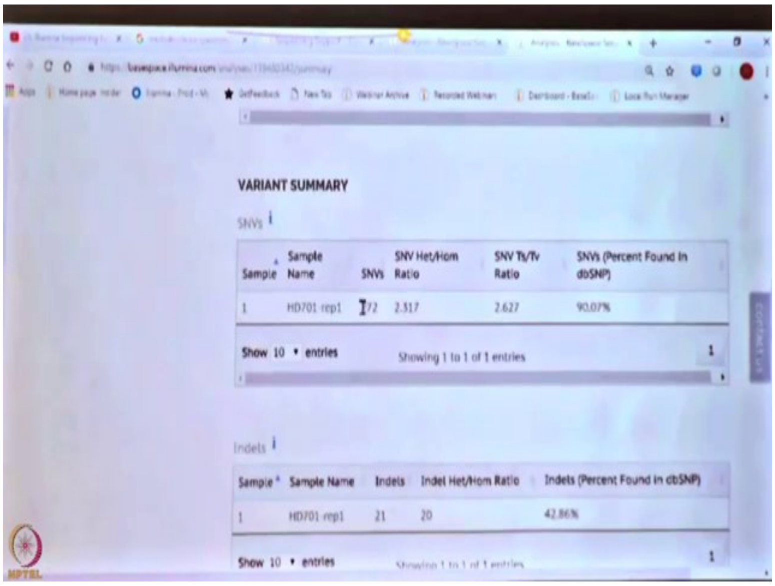
SNVs

Sample	Sample Name	SNVs	SNV Het/Hom Ratio	SNV Ts/Tv Ratio	SNVs (Percent Found in dbSNP)

So, it gives you information like read level enrichment, base level enrichment, target level enrichment. So, these are all really quality metrics that you want to use to make sure of two things; one is that you read data is of high quality and b you have very very specific sequence data that has been generated in. So, if you see here the target level enrichment is close to 100 percent 99.77 percent, which means that you have, the sequences that you have are from your target region.

Majority of the sequences that you have are from your target region, if this number is low which means that you have off target sequence data in your file ok. So, there may be something wrong with the way the data was generated. You know there may be something wrong in the way the library was prepared and so on and so forth. So, you can use this to make sure that your workflow was sort of you know error free so, to speak.

(Refer Slide Time: 21:15)



VARIANT SUMMARY

SNVs

Sample	Sample Name	SNVs	SNV Het/Hom Ratio	SNV Ts/Tv Ratio	SNVs (Percent Found in dbSNP)
1	HD701 rep1	[illegible]72	2.317	2.627	90.07%

Show 10 entries   Showing 1 to 1 of 1 entries   1

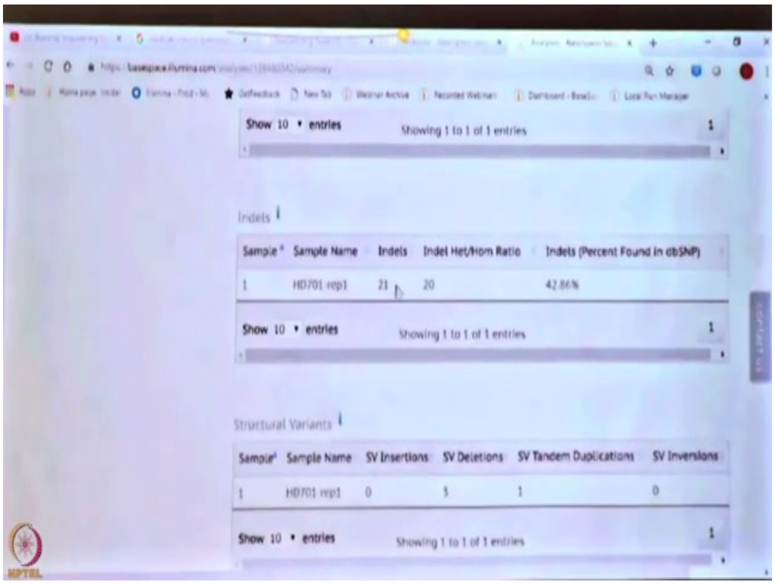
Indels

Sample	Sample Name	Indels	Indel Het/Hom Ratio	Indels (Percent Found in dbSNP)
1	HD701 rep1	21	20	42.86%

Show 10 entries   Showing 1 to 1 of 1 entries   1

What it will also give you is the number of SNVs; structural variations ok. Sorry SNV the Single Nucleotide Variations or one based changes that were identified in this particular sample.

(Refer Slide Time: 21:37)



Show 10 • entries | Showing 1 to 1 of 1 entries | 1

Indels

Sample	Sample Name	Indels	Indel Het/Hom Ratio	Indels (Percent Found in dbSNP)
1	HD701-rep1	21	20	42.86%

Show 10 • entries | Showing 1 to 1 of 1 entries | 1

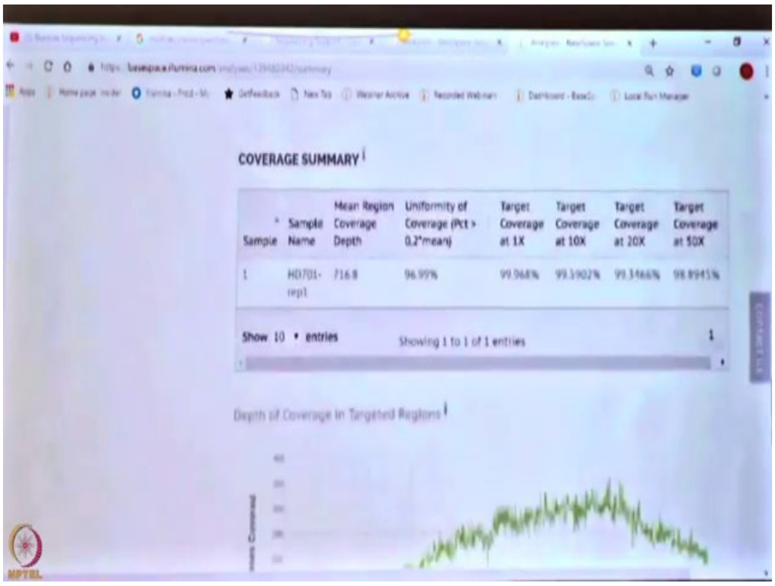
Structural Variants

Sample	Sample Name	SV Insertions	SV Deletions	SV Tandem Duplications	SV Inversions
1	HD701 rep1	0	5	1	0

Show 10 • entries | Showing 1 to 1 of 1 entries | 1

It will also give you the number of indels; that is insertions and deletions that were called in this particular sample.

(Refer Slide Time: 21:51)



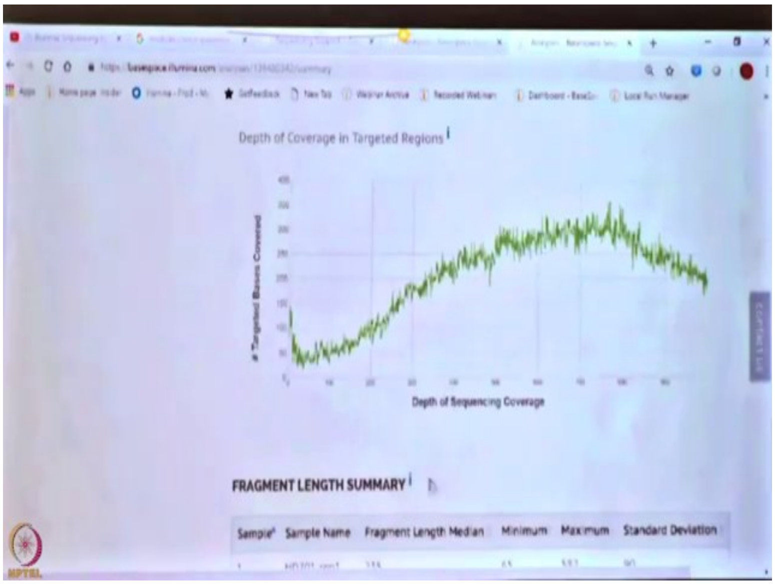
COVERAGE SUMMARY

Sample	Sample Name	Mean Region Coverage Depth	Uniformity of Coverage (Pct > 0.2*mean)	Target Coverage at 1X	Target Coverage at 10X	Target Coverage at 20X	Target Coverage at 50X
1	HD701-rep1	716.8	96.99%	99.968%	99.5902%	99.3466%	98.8945%

Show 10 • entries | Showing 1 to 1 of 1 entries | 1

Depth of Coverage in Targeted Regions

(Refer Slide Time: 21:52)


Depth of Coverage in Targeted Regions
# Targeted Bases Covered
Depth of Sequencing Coverage

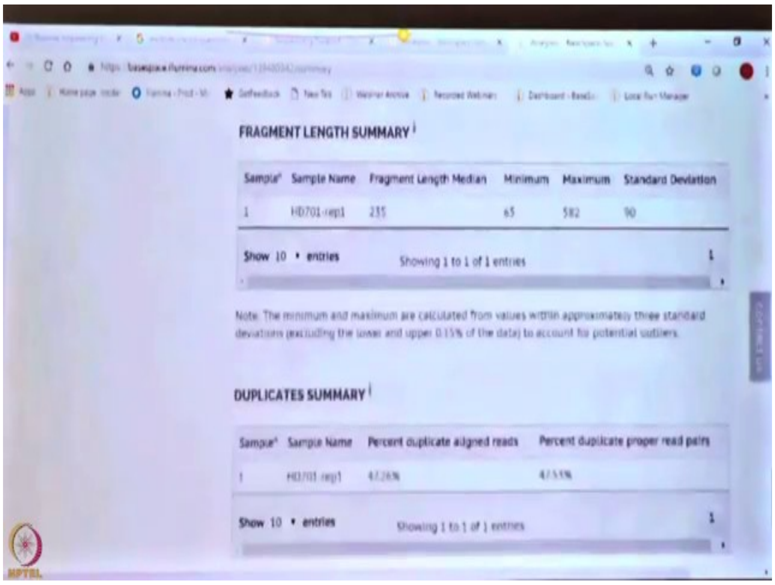
FRAGMENT LENGTH SUMMARY

Sample	Sample Name	Fragment Length Median	Minimum	Maximum	Standard Deviation

It gives you coverage summary and also the depth of coverage of the targeted region. So, what you see here on the horizontal axis is the depth of sequencing coverage. So, I think Mukesh talked about the x coverage rate. So, when you are doing next generation sequencing you are actually sequencing, just 2 minutes; you are sequencing every base multiple times and then depending on the application you are running your depth of sequencing can be as low as 30 x.

So, as Mukesh mentioned for whole genome sequencing, we are really looking at 30 x sequencing. For somatic mutations you may want to sequence as deep as 5000 x 10000 x depending on the frequency of the mutation right, if it is a rare variant and you know in cancer samples, especially if it is a heterogeneous tissue sample you may have to sequence deeply. Liquid biopsy is another example where you have to sequence deeply, because you are really trying to identify those cancer DNA cell free or you know CT CS in your blood which will have you know DNA from your normal cells right. So, the depth of sequencing in this case is very high. And the median fragment length and so on and so forth.

(Refer Slide Time: 23:06)



FRAGMENT LENGTH SUMMARY

Sample	Sample Name	Fragment Length Median	Minimum	Maximum	Standard Deviation
1	HD701-rep1	235	65	582	90

Show 10 entries Showing 1 to 1 of 1 entries 1

Note: The minimum and maximum are calculated from values within approximately three standard deviations (excluding the lower and upper 0.15% of the data) to account for potential outliers.

DUPLICATES SUMMARY

Sample	Sample Name	Percent duplicate aligned reads	Percent duplicate proper read pairs
1	HD701-rep1	47.26%	47.55%

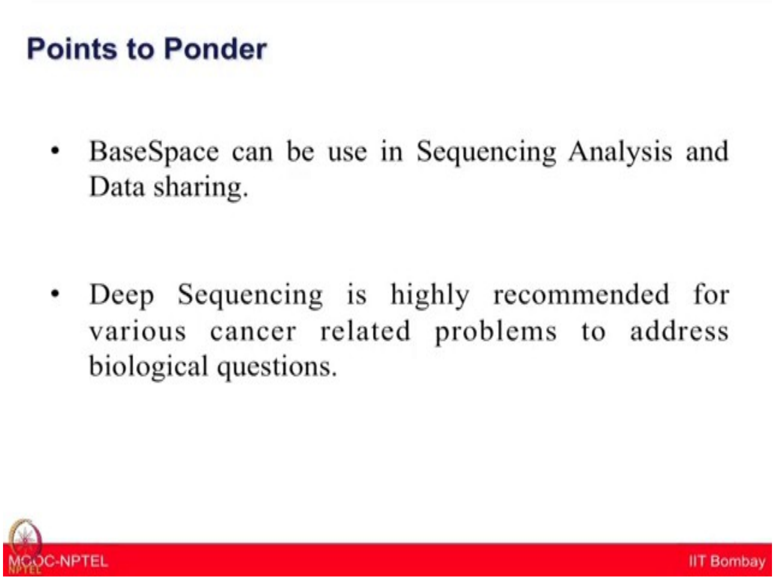
Show 10 entries Showing 1 to 1 of 1 entries 1

So, you are really getting a lot of matrix, you can also actually somewhere it is not shown here, but see the specific mutations that are called. So, if you download some of this data you will be able to actually see the specific variations that are called. You know we have looked at NGS data generated on the Illumina platform. We have seen how we can share data amongst you know our collaborators, run some analysis and looked at what the output, you know may look like depending on the application that you have run.

Since you all now have BaseSpace accounts there are many public data sets that are available right. So, if you go to the public data section on BaseSpace you will see that there are thousands of data sets that are available. So, you know you can look at those at your free time and you know reach out to us if you have any questions. Hope you guys found this session useful.

Thank you.

(Refer Slide Time: 24:01)

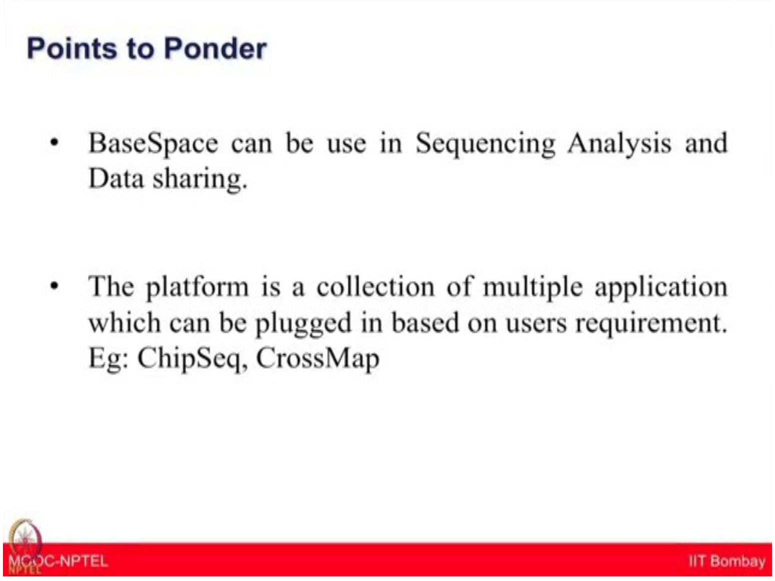


Points to Ponder

- BaseSpace can be use in Sequencing Analysis and Data sharing.
- Deep Sequencing is highly recommended for various cancer related problems to address biological questions.

MOOC-NPTEL IIT Bombay

(Refer Slide Time: 24:10)



Points to Ponder

- BaseSpace can be use in Sequencing Analysis and Data sharing.
- The platform is a collection of multiple application which can be plugged in based on users requirement. Eg: ChipSeq, CrossMap

MOOC-NPTEL IIT Bombay

I hope today's session by Dr. Aarti Desai was really informative, where you learnt how BaseSpace can be used in sequencing analysis, data sharing and how data output looks like. She also showed that this platform is a collection of multiple applications and demonstrated you a single application; that is variant calling. She gave a detailed information of parameters, launching of the app and how to run it. I suggest you to play with other applications available there and you will find them really interesting.

And, Dr. Desai also told you that you will learn more when you will play with the parameters of each application with new data set. I would like to mention you there is a large amount of data set available which is publicly accessible on various portals; such as The Cancer Genome Atlas or TCGA, and there are various ways one could download those data set and by using these tools try to understand and analyze the data.

So, more and more you are playing with these tools and you are familiar with the software features, you can then make use of large amount of public accessible data set from various large genome sequencing projects. In the next supplementary lecture we will learn about IonTorrent a bench top, next generation sequencing technology by Dr. Atima Agarwal.

Thank you.