

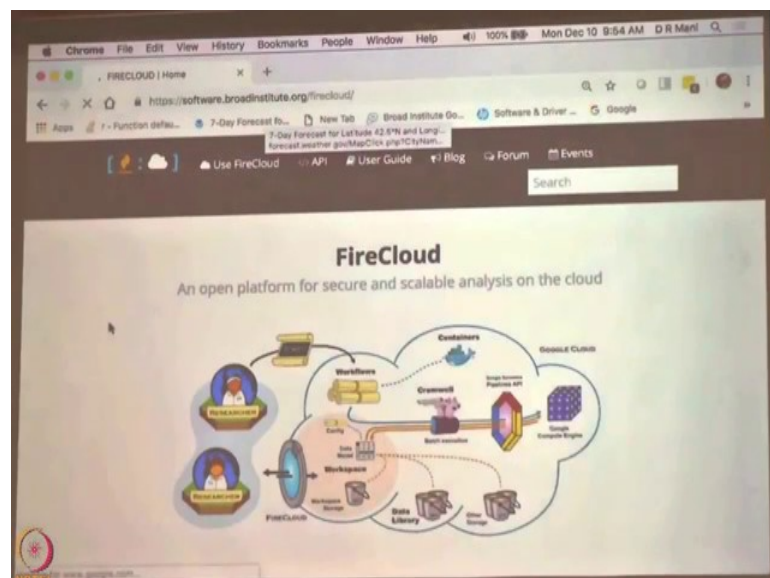
**Introduction to Proteogenomics**  
**Dr. Sanjeeva Srivastava**  
**Dr. D. R. Mani**  
**Department of Biosciences and Bioengineering**  
**Indian Institute of Technology, Bombay**  
**Broad Institute of MIT and Harvard**

**Lecture - 28**  
**Firecloud for Data analysis**  
**(Demo Session)**

Welcome to MOOC course on Introduction to Proteogenomics. While, it is not very difficult to generate big datasets currently using mass spectrometry or NGS platforms, but data storage, data processing, data analysis is not very easy. And, in this light cloud computing has provided some hope in which way big data can be handled. In this light the last lecture Dr. Mani has given you introduction about FireCloud. From the last lecture it was clear that FireCloud allows users to set up the data analysis using customized workflows across multiple hardware platforms.

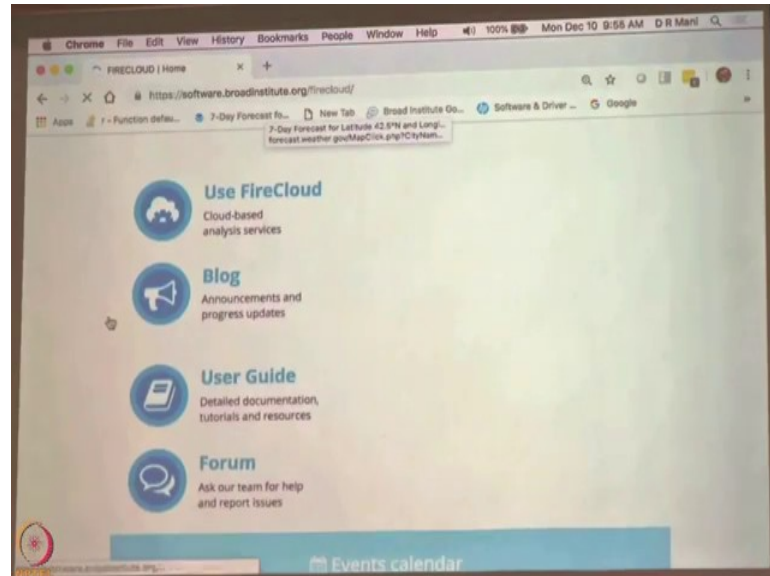
It also allows the users to use several paths for running the pipeline which includes the direct invocation using command line or through any other pipeline service. Today Dr. Mani is going to give you a demonstrations for using FireCloud and in which way you can analyze your data even it is very big dataset using this platform. So, let us welcome Dr. Mani for today's demonstration session.

(Refer Slide Time: 01:37)



So, if you type fire cloud dot org, it will go to the homepage site which kind of shows the same picture that I have showed you 4 times, but it will also have these others.

(Refer Slide Time: 01:49)



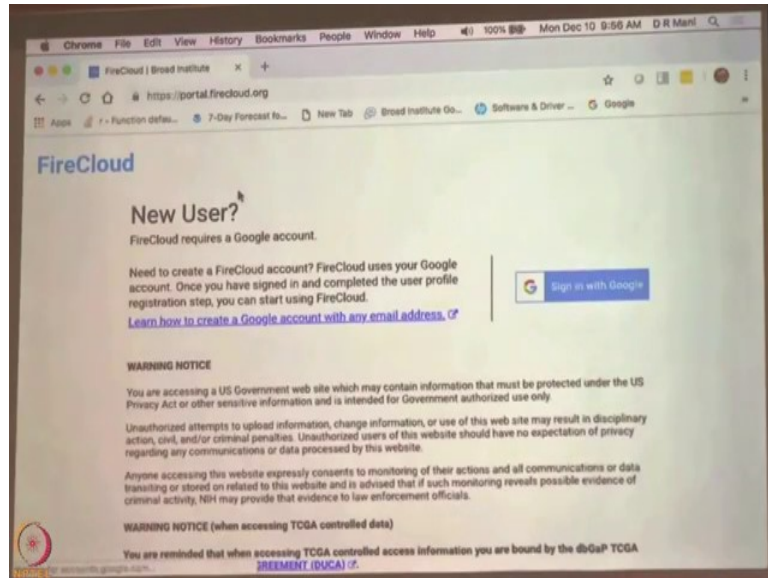
So, if you want to start using FireCloud you click on use FireCloud, but I want to point out the other things that are on the page. So, there is a user guide so, if you click on that it will give you a lot of information on how to use FireCloud. There is also a user guide for WDL which you can access separately. And so, a lot of documentation has been reasonably written for this. There is a blog that announces new updates or when the system is not working or there are maintenance and things like that; the blog has like updates that you can follow, but the cool thing is the forum.

So, the people who actually created FireCloud and who are working on all the algorithms respond to questions if you have any. So, you go to the forum, I think to use the forum you have to log in. So, you have to create a user ID and login, but once you login you can see what others have asked and what responses you have. And, like David was saying yesterday a lot of the questions that you encounter when you start using it, somebody has already encountered. And, then there will be an answer that you can quickly look up.

But, if there isn't and your question is unique to your problem and you are encountering some bug, that you need to have them fix; then you can post on the forum and then somebody will respond usually within a few hours. Well maybe here it might be half a

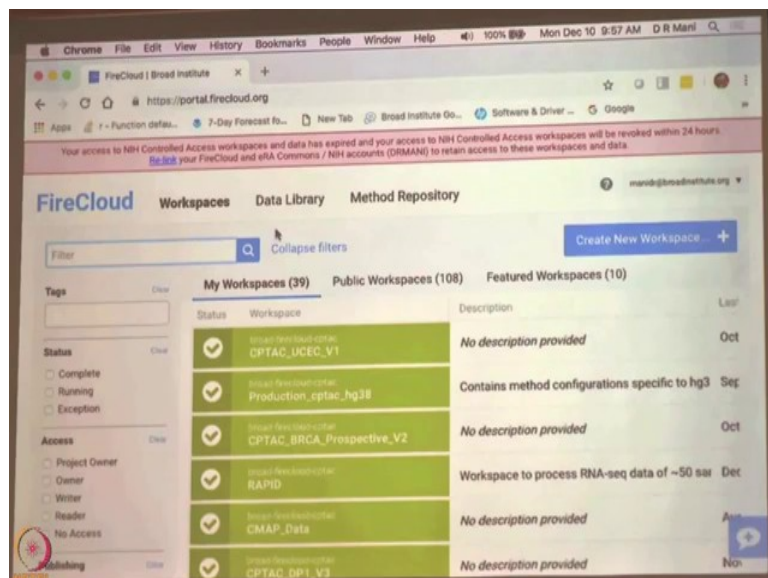
day because, when you post now they are all sleeping. And so when you click on use FireCloud, it goes to the FireCloud main page, this is a portal page.

(Refer Slide Time: 03:30)



If you type portal dot firecloud dot org you would end up at this page directly and so, you can see you have to login to use it.

(Refer Slide Time: 03:36)

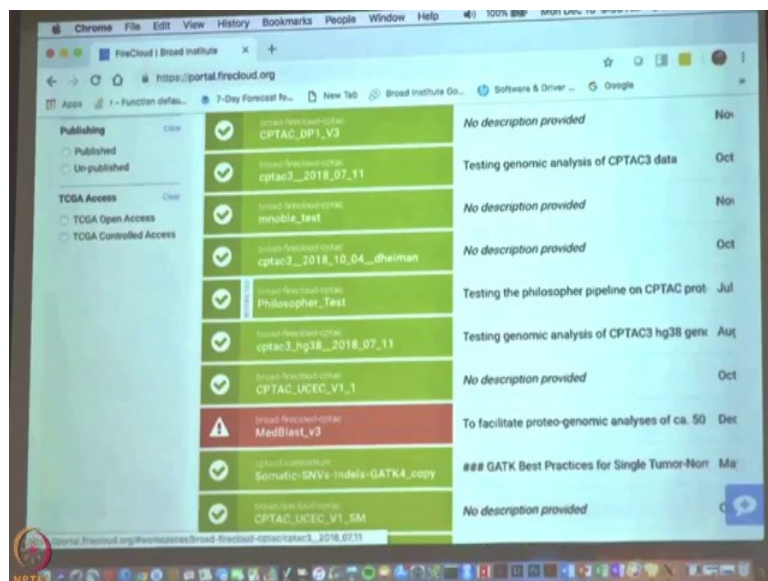


You can use your Google ID to login, I think mine is already set up so it just went in and I think you have to use only a Google ID to login; I do not you can use anything else. So, the main page looks like this and if you read the top red bar that comes up as a warning

for me, it is saying my NIH controlled access data permission has expired. So, controlled access data has the identifiable data with genomic SNPs that I just mentioned previously; in order to access that data you need special permission from the NIH. So, you have to say why do I want access to it and what project am I going to use that data for. And so, you have to apply and then they will look at your application and say ok, you are now authorized to access this data.

So, when I click on this relink, it will take me to a login page where if I log in correctly it knows I am authorized and then it will enable me to access those data. But, after a month the authorization will expire and you have to reauthorize yourself in FireCloud, the NIH authorization is usually valid for a year. So, this is the kind of security I was talking about, if you downloaded the data to your institute, when your institute is responsible for making sure that people who access the data are authorized to access it. But, here FireCloud implements a uniform process for everybody and you do not have to worry about it now.

(Refer Slide Time: 05:11)



And, then you can see three main tabs here; one says so one says workspaces, the other says data library and the third one says method repository. So, workspace is where you are doing your analysis. So, it has a set of methods, it has your data and it has your configuration for the methods whatever inputs you want to give for your methods. The

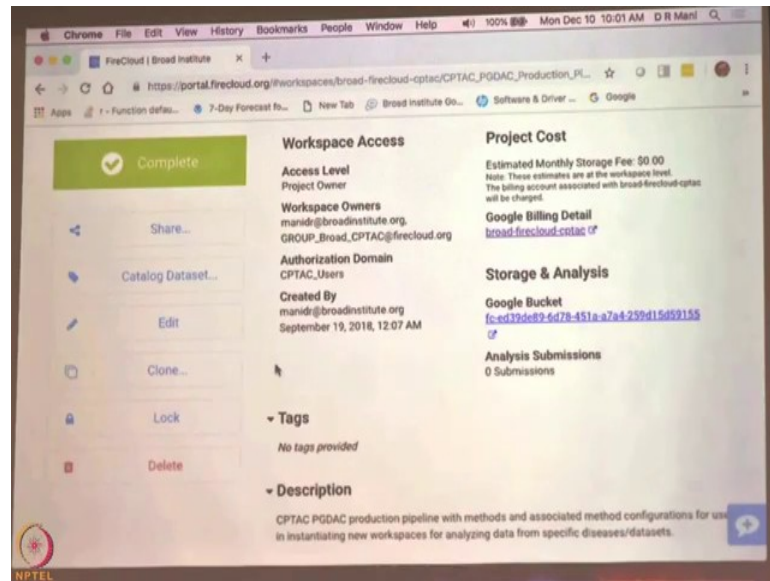
data library has publicly available data. So, the CPTAC data is available, the genomics, TCGA data is available; all those data are available in the data library.

And, you can use it for your analysis or you can just look at it or you can download it, but if you have to download it I think you need an account because, it costs money. Then finally, there is the method repository which has methods that everybody has written. So, some methods are public then you can see those, some methods are shared. So, the method author will say, you can look at it and they will give you permission and you can look at it and some methods are private only the authors can see.

And, whoever writes the method can control what the access privileges are for that method. So, if it is a method you are still testing and you are not sure works very well, you would most likely keep it private. But, then once it is been tested and you are using it in a project in your group, then people in your group might be might have access to that method and ultimately when your paper is released you make it public. So, you can change the access anytime and you can control who has access to methods you have written.

So, I will pick a workspace and we can take a look at it. So, let see so, you see there is this workspace that is called CPTAC PGDAC production pipeline. So, that is our collection of pipelines for the proteogenomic data analysis all the CPTAC projects for which we do proteogenomics. And you can see that it says restricted, that is because only people at the broad currently have access to it.

(Refer Slide Time: 07:18)



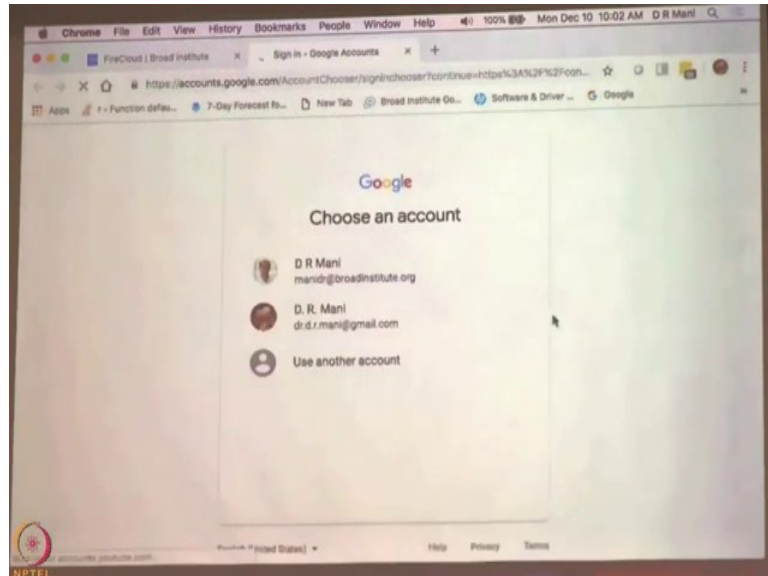
So, if you click on that this is the summary page. So, let say summary, this is the summary page for that workspace. So, you can see the workspace has data, it has analysis, it has notebooks, it has method configurations and when a process is running you can monitor what is happening to it. And, here on the left side it says that the last process run has completed successfully which is green. If it did not complete successfully, you will have a red triangle and the whole thing will be red.

You can share the workspace with others, you can edit it, you can make a copy of it and experiment with it if you want or you can just delete it or let us say you have published your paper and you do not want any changes made to this workspace; you can lock it. So, if you lock it nothing can be changed in the workspace, but people can copy it and do stuff with it and then it says who has access, who owns it, who are authorized users. So, this is all basically access related stuff and then here it is showing me how much it cost to run it.

So, you can see I think this one has no data in it, so, there is no monthly cost. If you had data files stored then they will be charging you some fee for storing the data on a continual basis. And then so, this is the billing detail, if you click on this it will show you how much you have used in the last whatever time frame, for storage, compute you can get a lot of details there, but the thing I wanted to point out is the Google bucket. So, every workspace has a Google bucket associated with it and this is the address of the

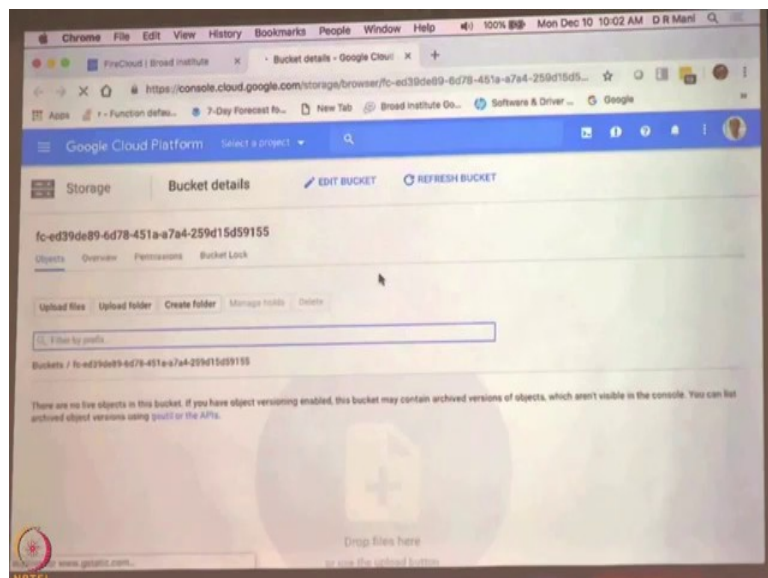
Google bucket associated with this workspace. So, if I click on it, actually, fine let me click on it.

(Refer Slide Time: 09:13)



So, accessing a bucket is again based on permissions, you have to have the permissions to access the bucket.

(Refer Slide Time: 09:27)

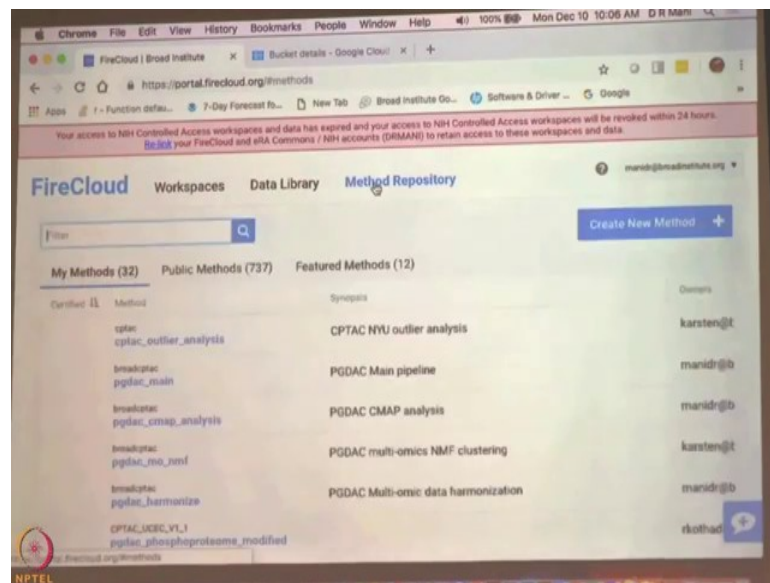


So, this is your bucket. So, basically it is saying there are no objects in this bucket. So, because this is a workspace we set up to just store methods; so, that when we need to do a real analysis we copy the methods from here into a new workspace, put the data in the

new workspace and do it because, every time you get different data or slightly modified data. So, you have breast cancer you need breast cancer data, you made some changes you have new version of breast cancer data, you want to apply it to ovarian cancer you have ovarian cancer data.

So, usually the data changes and with that the settings used to run the methods also change. And so, we keep the methods in this workspace and then copy it to a new workspace to apply it.

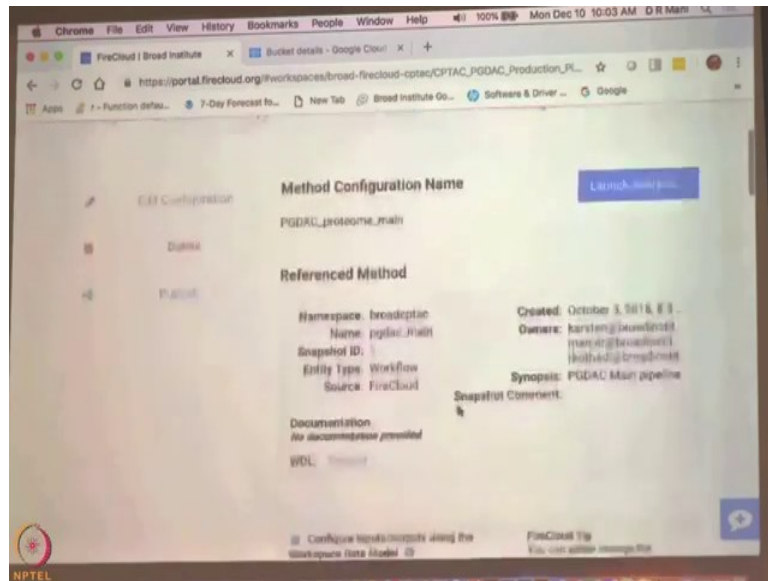
(Refer Slide Time: 10:13)



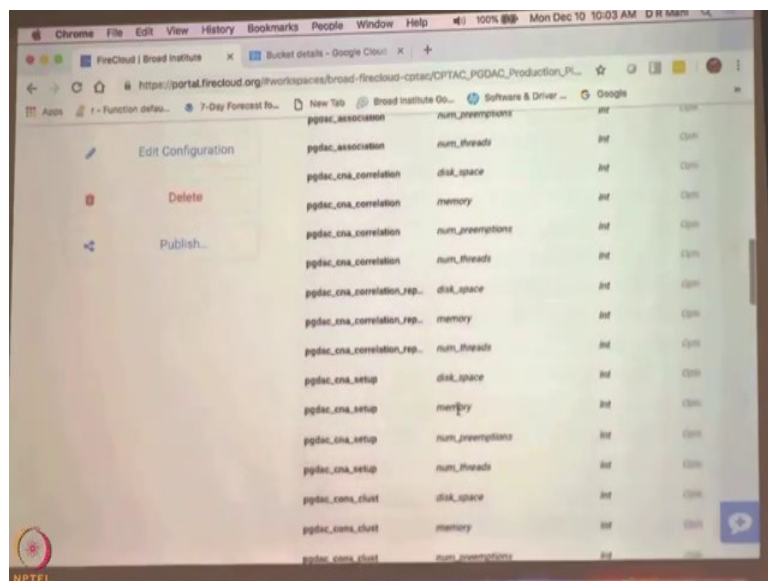
So, before going to actual analysis workspace I want to point out the methods that we have. So, here are the methods. So, each line here is one pipe line. So, you can see there is a main pipe line.



(Refer Slide Time: 10:26)

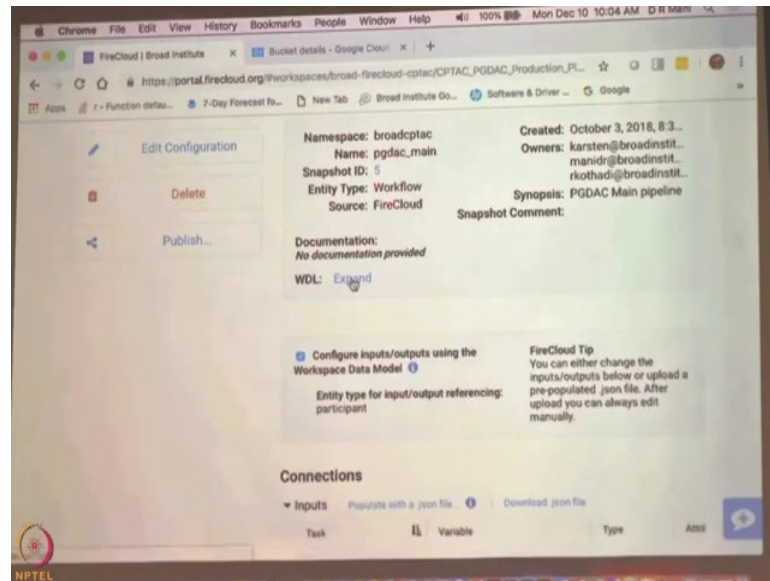


(Refer Slide Time: 10:35)



So, this is the basic pipeline that runs a lot of proteogenomics analysis and these are all the inputs to that, many of them are optional and do not have to specify anything, but some of them you have to specify as inputs. So, like the input data set for example, without that you cannot run an analysis. So, things like that you have to specify.

(Refer Slide Time: 10:54)



And, here the snapshot is the version of that method, if you made a modification and you uploaded the method again the snapshot will change to 6. So, it will keep track of all the changes you have made and you can revert back to old methods if you want to. So, let us say you have made some change and then you realize that was that actually introduced a bug and you do not want to use that snapshot, you can go back to an older one. And, when you click on the there is this thing called WDL and if you click on expand it will show you the actual workflow for that pipeline.

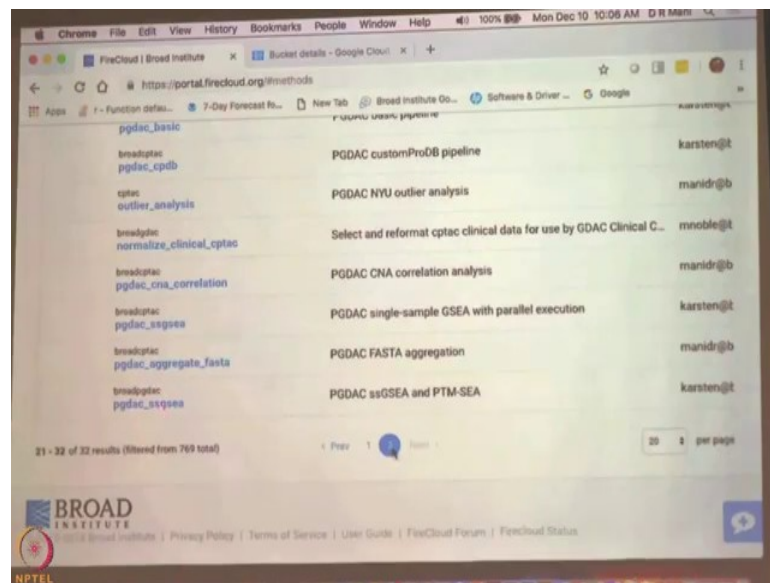
So, you can see there is a task called PGDAC association, it has some inputs and outputs then there is another task called CNA correlation report, CNA copy number analysis, consensus clustering; this is for like harmonizing data, normalizing mass spec data then a report for that. So, basically all the tasks that you would need in order to start with preprocess genomics data and preprocess proteomics data. And, then do like a proteogenomic analysis, replicating a lot of things that was reported in the breast cancer nature paper or are included in this pipeline.

So, I will actually show you and then so, these are all the task descriptions and at the end of it is the workflow. So, it can see it's as a workflow, main pipeline, those are all the inputs; some of them are optional, but others are required. And, then here you say you start by parsing the spectrum mill table, the output of that goes into normalization, output of that goes into report creation and so on. So, you just go in sequence in for how the

workflow goes, if there is no dependence you can execute them in parallel, ok. So, and the method repository has all the methods that we have generated available. So, these are all the pipelines and workflows.

So, there is outlier analysis from David Fenyo and Kelly Ruggles lab that is available here. This is the main pipeline, this is the connectivity map analysis that was also reported in the paper and there are many others. So, I think Karsten's PTM, GSEA is also somewhere here maybe on the next page yeah, ssGSEA.

(Refer Slide Time: 13:30)



So, all the methods are here and you can use any of them you can. So, this is the method repository, when you create your workspace you can import any methods you want and then put your data there and apply the methods to that data. So, let us look at one workspace where we have actually done an analysis. So, I will do the so, this is the new breast cancer data which is we call the prospective data set and you see how version 2, version 1 was had some differences and now we updated it. So now, here if I go to so, you can also see that I ran 20 ran this many 20 pipelines in this workspace, 2 of them aborted due to problems, 18 of them we are done, we are finished.

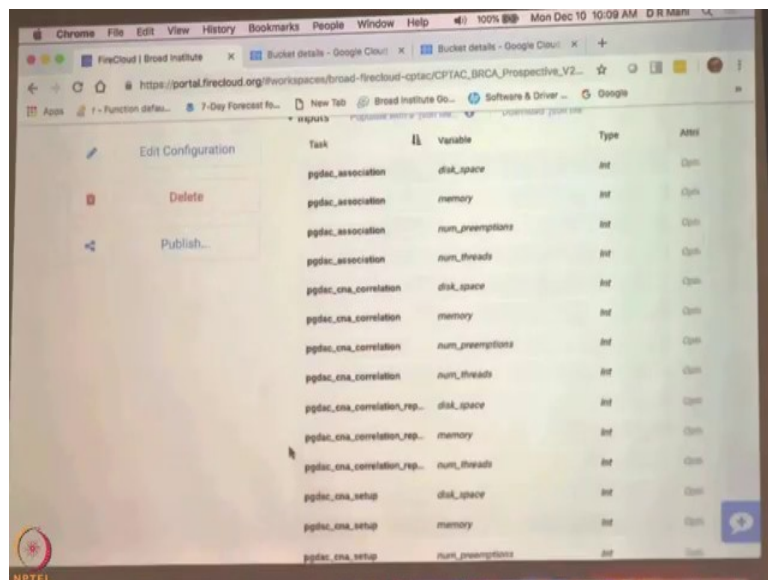
If I go to the Google bucket so now, you can see there are a lot of folders in the bucket and when you run a pipeline the output is written in a bucket that is named like this. So, this is like a long set of hexadecimal characters so, that you do not have like the same file for directory name or folder name. So, it creates a unique name every time with some

long set of, but you do not have to keep track of it. It will automatically know which folder was used for which task and then copy the right output to where it needs to go. So, but it will write all the outputs to the bucket that you have for the workspace, but the important thing to notice this thing called input here.

So, that is where all the input data is; so, if I click on that it will show what input data I have. So, I have RNAseq data, I have acetylome, phosphoproteome, proteome, I have some groups that I want to look at for a look for enrichment when I do my clustering. This is for the connectivity map analysis, this is the copy number data and then if you want to change some of the configurations of the pipeline, there is a configuration file you can use. The experiment design file is the most important one which associates samples with various annotations for the samples and also says which TMT run had a specific sample in which channel.

And so, this is basically all the data that you would need as input is here and in your workspace you would in the method configurations because, this is one that we are using for actual analysis the methods will be configured.

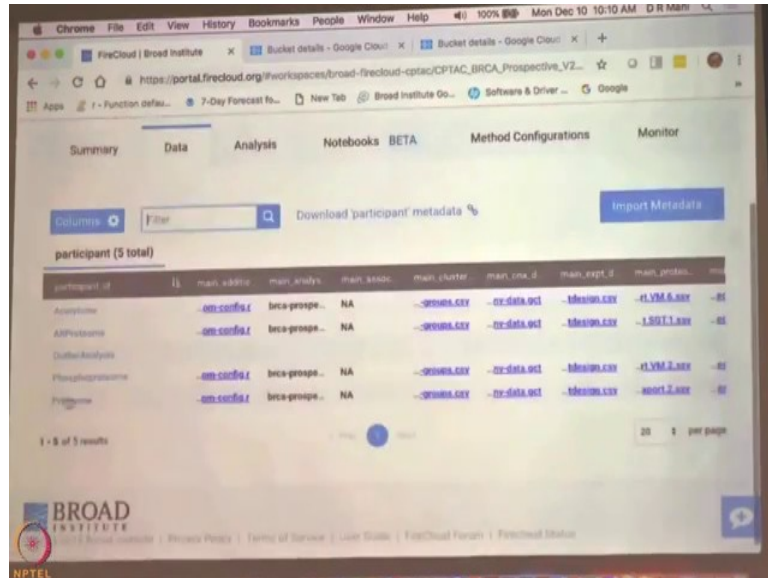
(Refer Slide Time: 16:20)



So, if you look at the proteome analysis pipeline you can see the optional things are not filled in, but if you look at; so, look at this analysis directory. So, this actually points to data in the metadata table. So, when it says this dot whatever I think it is getting cut off because the resolution is too low, but you can see that many of these are filled in with

values and where. So, some of them are actual numbers or strings like this one for example, is a string, but many of them start with this.

(Refer Slide Time: 16:59)



participant id	main_analysis	main_notes	main_cluster	main_file_1	main_file_2	main_protein		
Acetylation	-am-confid.r	bca-prospe...	NA	-swmsa.csv	-cr-data.csv	-Msaion.csv	-t.VM.6.csv	-81
ADPProteome	-am-confid.r	bca-prospe...	NA	-swmsa.csv	-cr-data.csv	-Msaion.csv	-1.SGT.1.csv	-81
GlycanAnalysis								
Phosphoproteome	-am-confid.r	bca-prospe...	NA	-swmsa.csv	-cr-data.csv	-Msaion.csv	-t.VM.2.csv	-81
Proteome	-am-confid.r	bca-prospe...	NA	-swmsa.csv	-cr-data.csv	-Msaion.csv	-sort.2.csv	-81

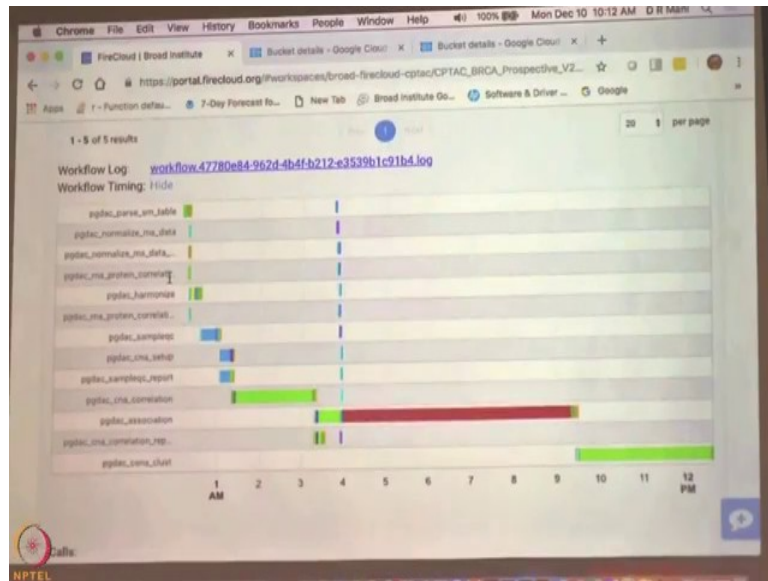
So, those are obtained from a metadata table, that kind of says for the proteome what are my input datasets and what are my input files that I need to use; for the phosphoproteome the table will specify what input files I need to use. So, for each of these you will have like a specification of inputs and when you run the method, it will ask you which participant should I use; the proteome, phosphoproteome which one? And, when you pick proteome it will get the right inputs, it will do the proteome analysis and the output will be also in one of these files. I think so there is a output table that is written out somewhere here or you can look at go to monitor, it will show you all the analyses that have been run.

(Refer Slide Time: 17:44)

Status	Method Configuration	Date	Time	Data Entry	User
View ✓ Done	cptac/CPTAC_outlier_analysis	October 13, 2018,	12:58 AM	OutlierAnalysis (participant)	manishj
View ✓ Done	cptac/CPTAC_outlier_analysis	October 12, 2018,	9:05 AM	OutlierAnalysis (participant)	manishj
View ✓ Done	cptac/CPTAC_outlier_analysis	October 12, 2018,	7:55 AM	OutlierAnalysis (participant)	manishj
View ✓ Done	cptac/CPTAC_outlier_analysis	October 12, 2018,	7:36 AM	OutlierAnalysis (participant)	manishj
View ✓ Done	cptac/CPTAC_outlier_analysis	October 12, 2018,	3:22 AM	OutlierAnalysis (participant)	manishj
View ✓ Done	breadptac/PGDAC_proteome_cmap_analysis	October 8, 2018,	9:28 PM	AllProteome (participant)	manishj
View ✓ Done	breadptac/PGDAC_proteome_cmap_analysis	October 8, 2018,	9:56 AM	Proteome (participant)	manishj
View Aborted	breadptac/PGDAC_proteome_cmap_analysis	October 6, 2018,	3:09 AM	Proteome (participant)	manishj
View Aborted	breadptac/PGDAC_proteome_cmap_analysis	October 5, 2018,	9:33 AM	Proteome (participant)	manishj
View Done	breadptac/PGDAC_proteome_cmap_analysis	October 3, 2018,	10:02 PM	AllProteome (participant)	manishj
View Done	breadptac/PGDAC_proteome_cmap_analysis	October 3, 2018,	10:02 PM	Proteome (participant)	manishj
View Done	breadptac/PGDAC_proteome_cmap_analysis	September 27, 2018,	12:44 AM	Phosphoproteome (participant)	manishj
View ✓ Done	breadptac/PGDAC_proteome_cmap_analysis	September 26, 2018,	10:21 PM	AllProteome (participant)	manishj
View ✓ Done	breadptac/PGDAC_proteome_cmap_analysis	September 26, 2018,	10:06 PM	AllProteome (participant)	manishj
View ✓ Done	breadptac/PGDAC_proteome_cmap_analysis	September 26, 2018,	6:44 PM	AllProteome (participant)	manishj
View Done	breadptac/PGDAC_proteome_cmap_analysis	September 26, 2018,	6:44 PM	Proteome (participant)	manishj
View Done	breadptac/PGDAC_proteome_cmap_analysis	September 26, 2018,	12:06 AM	Acetyome (participant)	manishj

So, this one for example, was a proteome analysis that we did, if we look at it will tell you what all happened in the proteome analysis. So, this was the proteome analysis, it will tell you what the inputs are. So, all the files and parameters that were used as input, it will tell you what the outputs are. So, there is one output file which is like a zip file that has all the outputs from the various tasks that were run. But, there are also reports that you have, that will show you in on a webpage; many times in an interactive web page on some of the results that were obtained. And, then these are all the tasks that were run, it is not in the order they were run, but these were all the tasks that were run and you can also see the timing.

(Refer Slide Time: 18:37)



So, it will show you which tasks were run when; so, you started off by doing the parsing the spectrum mill table, normalizing and then doing the normalization report, RNA protein correlation and so forth. And, you can see how long each one takes, the association analysis takes the longest and after that you do consensus clustering and then you are done. So, it gives you a visual picture of how the entire flow was executed and the data library basically has various data sets, I will not go into that now.

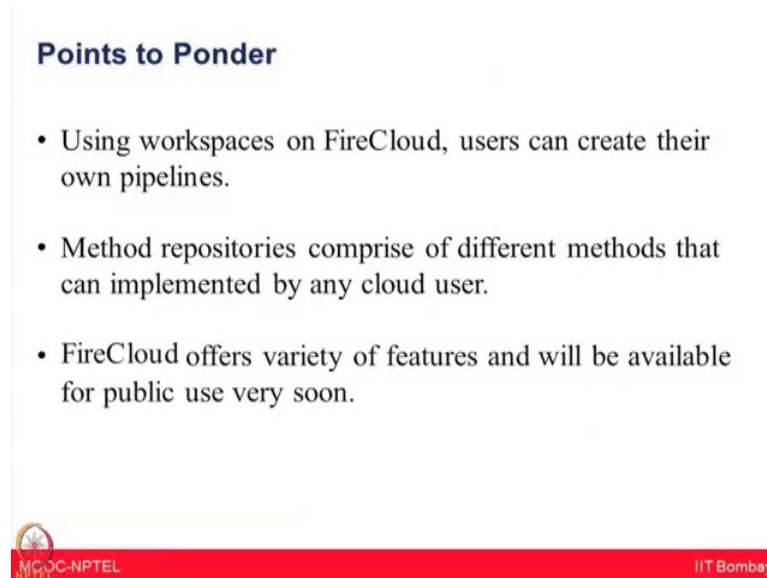
(Refer Slide Time: 19:11)

The screenshot shows the FireCloud Data Library interface. The URL is <https://portal.firecloud.org/#library>. The page displays a search for "Matching Cohorts" with 191 datasets found. The table below lists the cohorts with their names, phenotypes, and the number of subjects.

Cohort Name	Cohort Phenotype/Indi..	No. of Subj..	Consent Codes
ALSTDL_Paris_30xPCRFreeWGS_A...	NA	4	
Bipolar_Neuro_Fremer_GSA_MD	bipolar disorder	243	
BlazejMisek_SCCZ_FOC_GSA_MD	schizophrenia	445	HMB
CCDG_ASD_State_Sanders_Daily_W...	autism spectrum disorder	534	
CCDG_IBD_Daily_Cho_IBD_HMB_WGS	inflammatory bowel dise	253	DS:Inflammatory b
CCDG_IBD_Daily_Fugthaban_IBD_G...	inflammatory bowel dise	914	GRU
CCDG_IBD_Daily_McCusker_Abrou_J...	inflammatory bowel dise	856	
CCDG_IBD_Daily_McCusker_Targan...	inflammatory bowel dise	1565	GRU
CMG_BoneMass_MuscleDisease...	prostate disease	21	GRU
CMG_Broad_Histatransd_KidneyDis...	kidney disease	25	DS:kidney disease
CMG_Broad_Pierre_PresnalDisease...	retinal disease	35	GRU
CMG_Glioma_NeurologistDisease...	neurological disease	480	GRU
CMG_Histatransd_Kidney Disease...	kidney disease	77	DS:kidney dise

So, you can see that there is lot of data that you can use from there, I think that is kind of all I wanted to show.

(Refer Slide Time: 19:19)



**Points to Ponder**

- Using workspaces on FireCloud, users can create their own pipelines.
- Method repositories comprise of different methods that can implemented by any cloud user.
- FireCloud offers variety of features and will be available for public use very soon.

MOOC-NPTEL IIT Bombay

In today's demonstration session, you were given the concepts that in which way FireCloud could use, three main tabs; the workspaces, data library and method repository. The workspaces let the user create their own pipeline; the data library contains publicly available data that can be also used by the user. The method repository it contains the methods used by different users. These methods are also available for any user using cloud computing platforms, once you have submitted a job the user can now visualize the various steps involved including the input files and the output files.

I hope it at least gives you the glimpse of in which way the big data can be handled using cloud computing platforms. Some of these shown tools and platform currently are not available for the public usage, but hopefully Broad Institute will make it available online in some time. At least you should be aware that these are the good resources in which way you can analyze your data using cloud computing.

Thank you.