An Introduction to Proteogenomics

Dr. Sanjeeva Srivastava
Dr. Kelly Ruggles
Department of Bioscience and Bioengineering
Indian Institute of Technology, Bombay
New York University

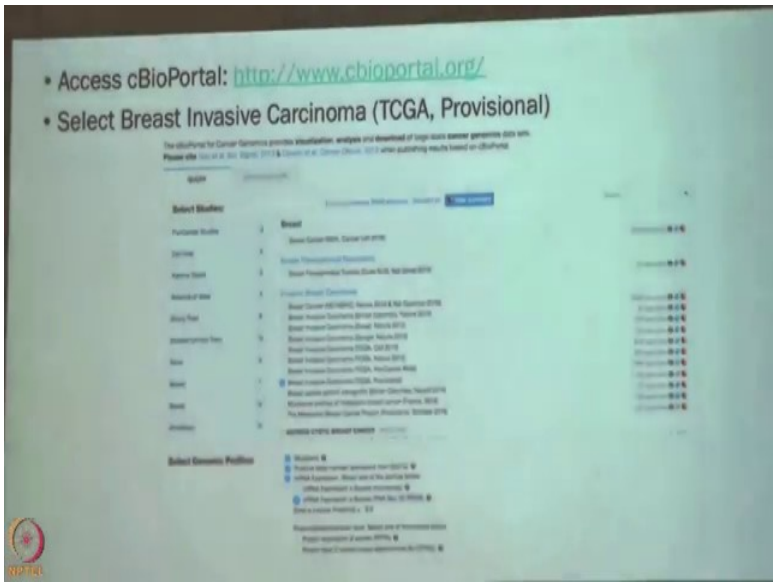
Lecture - 07
Introduction to Genomics
cBioPortal

Welcome to MOOC course on Introduction to Proteogenomics. Mutations and different genes, lead to different cancers such as the BRCA1 mutation, BRCA2 mutation and HER2 mutations, they lead to breast cancer while mutation and isocitrate dehydrogenase gene like IDH mutation lead to brain cancers. Therefore, understanding the modalities for efficient treatment of the cancer; it become very important to know about the mutations which are relevant to a particular tumor type.

In this slide, Dr, Kelly Ruggles today going to talk about how to use some of the online available tools like cBioPortal for accessing gene mutations and its expression from the published data sets. Dr. Kelly Ruggles will take the cases study of breast cancer and 6th most common oncogenic mutations in the clinical conditions. She will show how one could access the published data set and understand the correlation between various genes and their mutations.

So, let us welcome Dr. Kelly Ruggles for her lecture today.

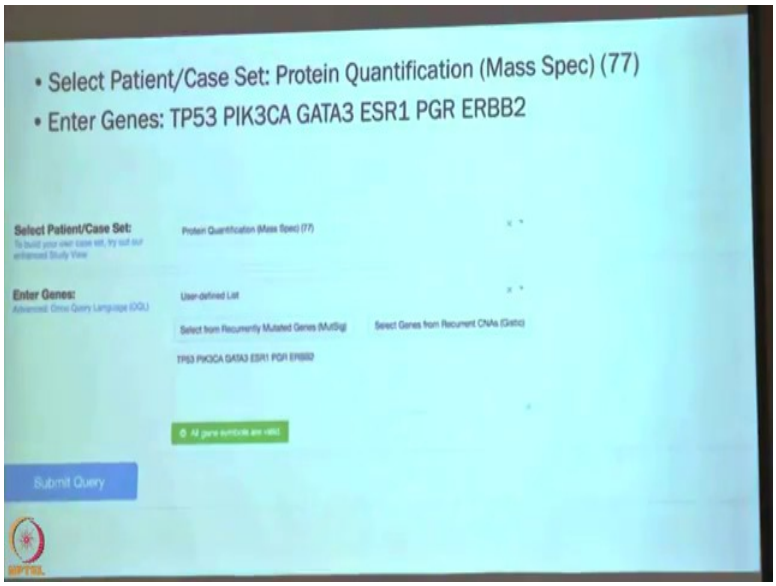
(Refer Slide Time: 01:47)



You should have this slides if you, that is the about. So, if you go to this cBioPortal.org that is the first step. So, everyone go there and then you are going to want to a select the specific data type, it is called the breast invasive carcinoma TCGA provisional dataset. So, if you look in, there should be a something that looks kind of like this, this you have to scroll down a bit. Also, make sure you click on these mRNA expressions the score, there is a sub panel at the bottom, that has mutations, copy number and then everyone look at mRNA expression in additions all of these.

Yeah make sure, you click on the TCGA provisional and not one of the other ones, because our data is not any other ones, it is not easy to pull out. So, once you get to you go down a bit and you will see this select patient case set.

(Refer Slide Time: 02:39)



So just everyone see that, go down. Yes, select patient case set and you want to a select this protein quantification mass back. There should be 77 samples that you see.

So, there is this there is and then there is 6 genes you will put in to in a gene section. So, it is these 6 genes TP53, PIK3CA, GATA3, ESR1, PGR and ERBB2.

Student: asks question

Yeah, you type in the gene symbols than that.

Student: That whole thing.

Just these.

Student: asks question. P53?

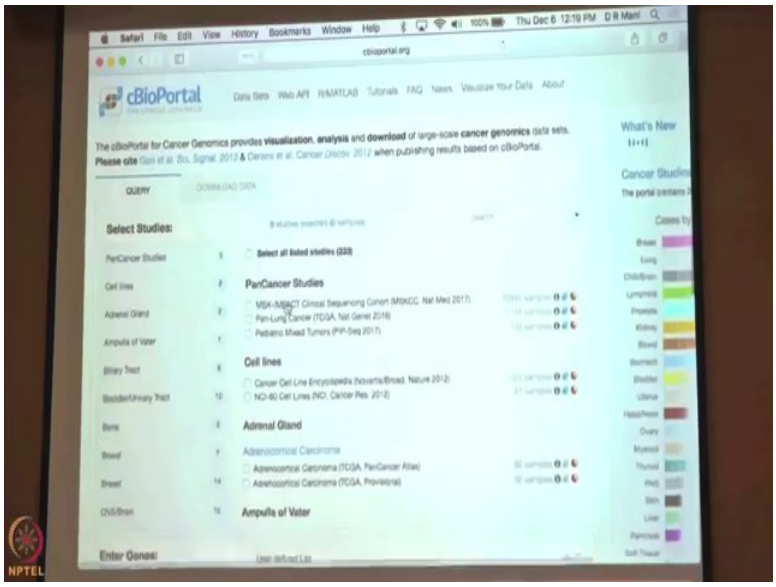
Yeah.

Student: Here getting one protein only of 290 proteins.

(Refer Slide Time: 03:21)

We want 6 proteins in that fields ok.

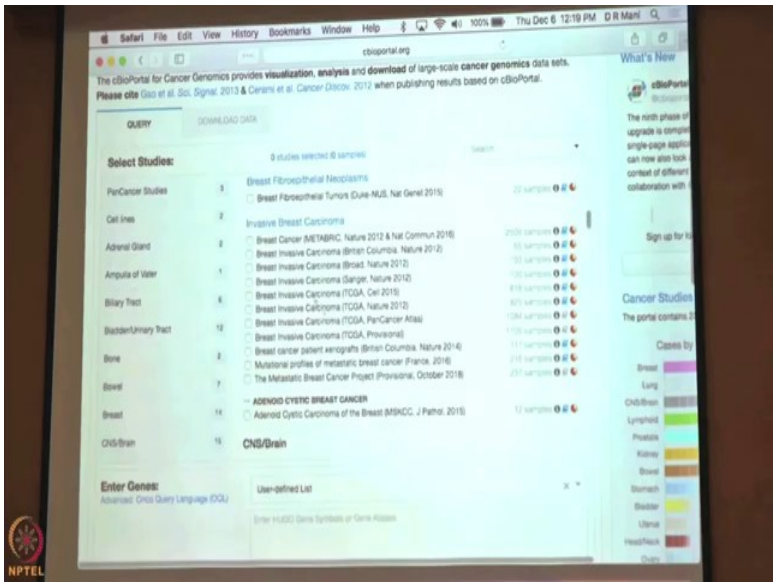
(Refer Slide Time: 03:26)



So, we are going start over for anyone he still has not found it.

Dr. Mani: (Refer Time: 03:30) we can pause and say sir, everyone I could get to the CbioPortal page webpage.

(Refer Slide Time: 03:35)



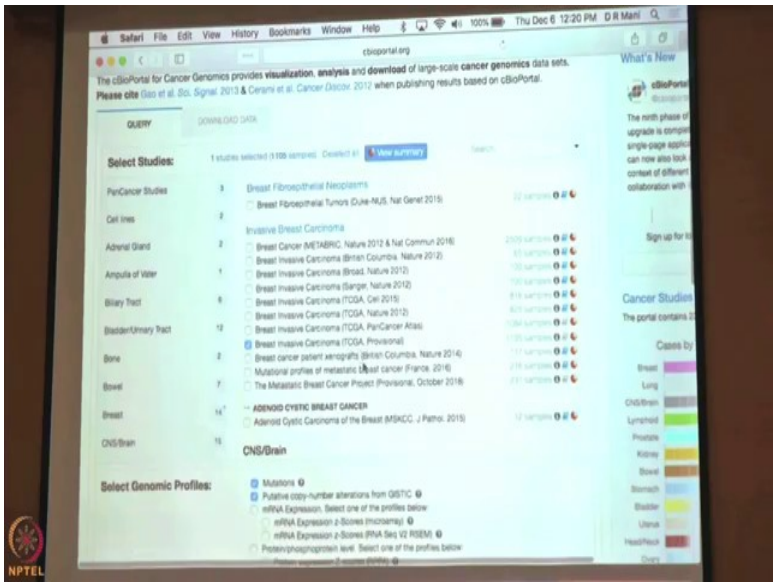
Dr. Mani: Any one has the problem raise your hand.

Student: asks question

No do not change anything yeah, so.

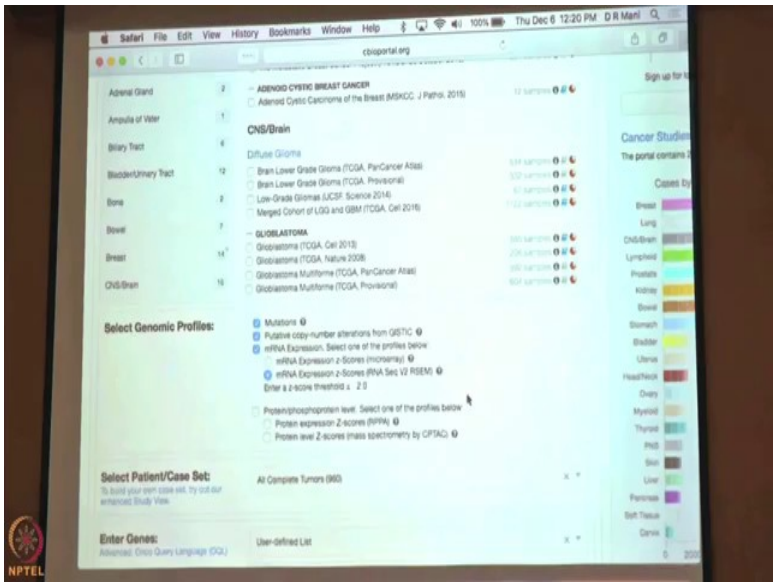
Here is the page if you go down, you click this breast invasive carcinoma.

(Refer Slide Time: 03:56)



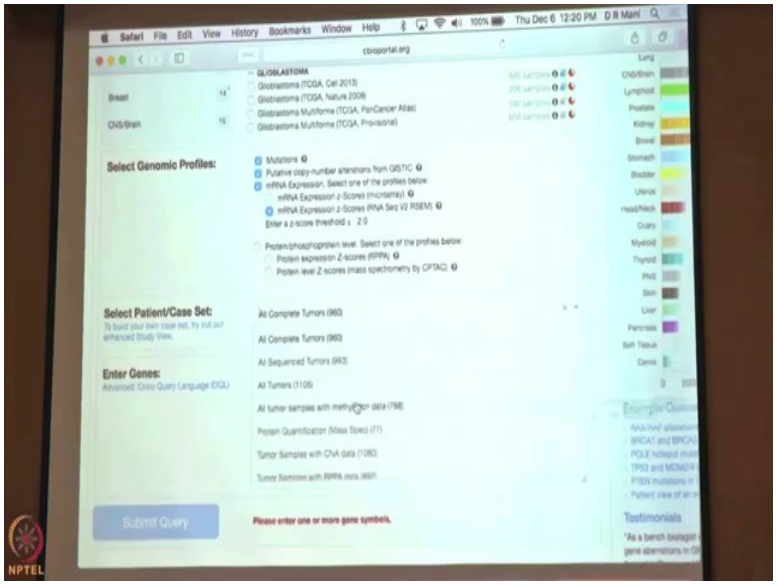
And, then there is the select genomic profiles.

(Refer Slide Time: 03:59)



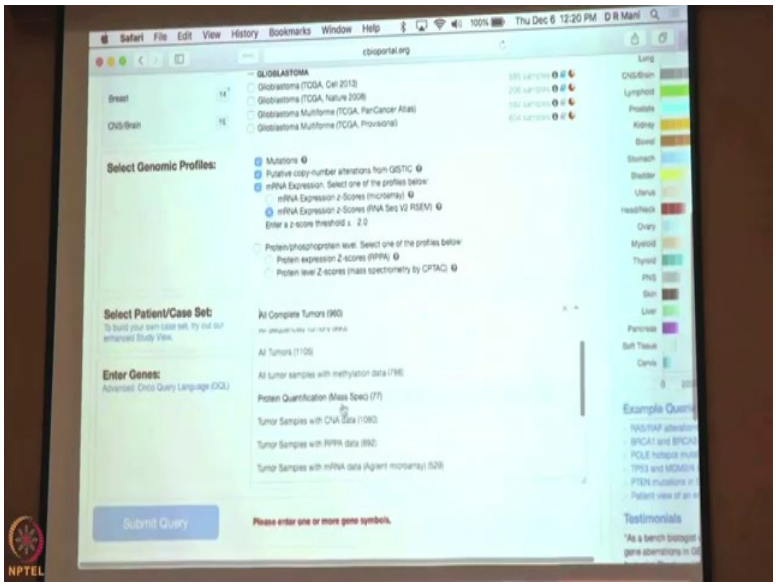
And you click on mRNA expression, the second mRNA expression, which is the RNAseq.

(Refer Slide Time: 04:12)



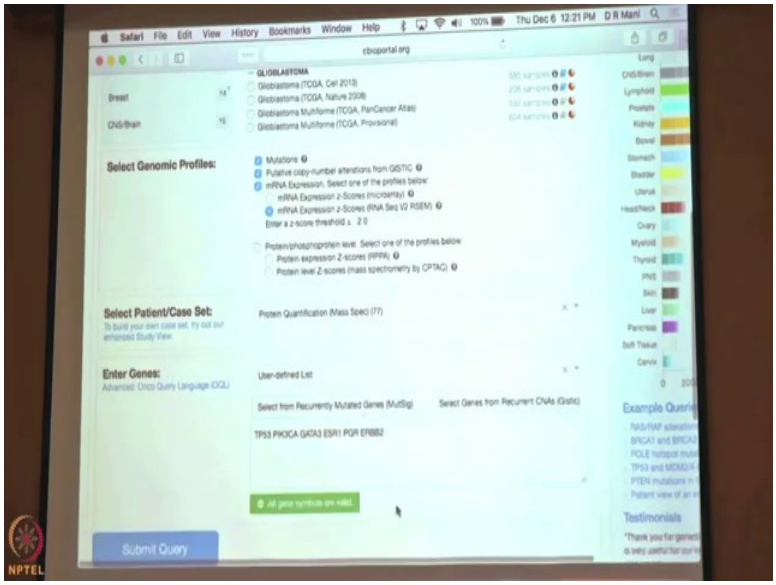
For this patient case set you want so, find the one that says.

(Refer Slide Time: 04:19)



Here, protein quantification mass spec. So, this is the 77 samples we used for other proteomics CPTAC analysis and then we are going to type in.

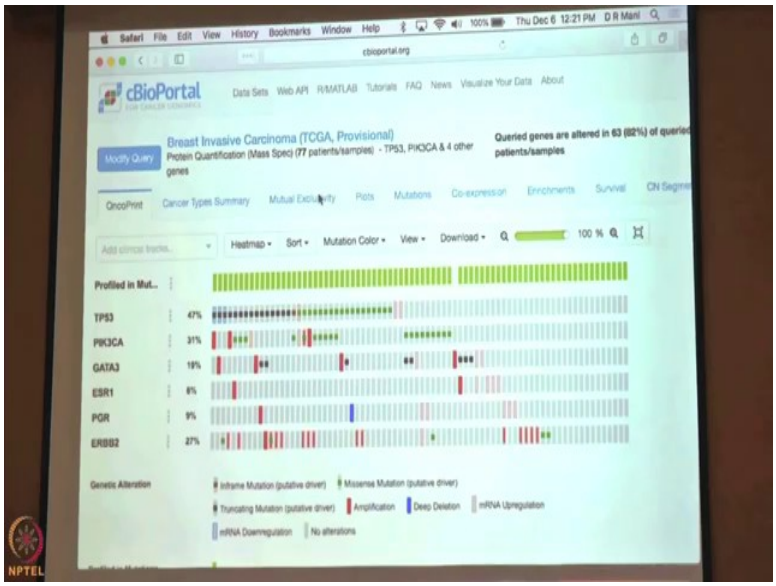
(Refer Slide Time: 04:29)



So, these 6 genes were the same genes that we focused on figure 1 B which is why I wanted to have us look at them in a little more detail. So, that is TP53, PIK3CA, GATA3, ESR1, PGR and ERBB2. So, you should be able to enter all of those in and submit the query.

Student: Last two genes.

(Refer Slide Time: 05:03).



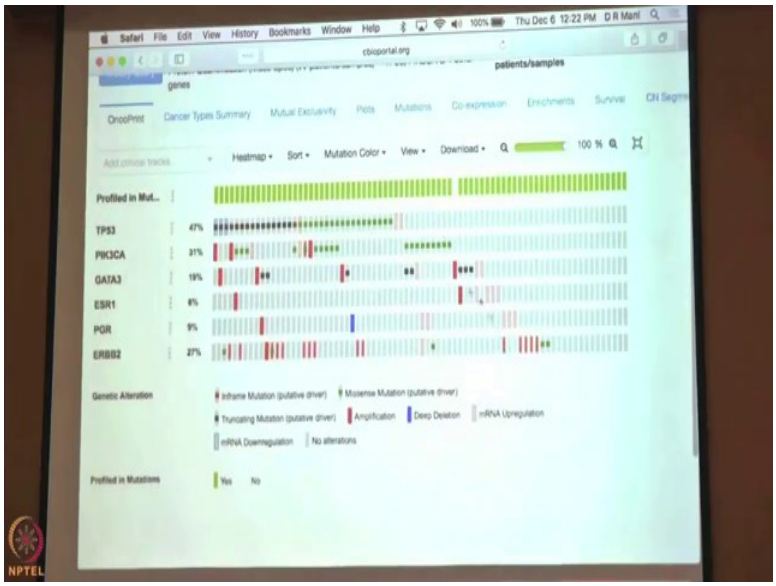
Last two genes and PGR ES.

Student: PGR.

ESR1, PGR and ERBB 2, they are up here now, if you need them. Yeah and you should get this. So, how many people were able to get to this? Awesome, killing it guys great ok. So, what you will see here is what this shows you is the mutation status. So, each of these columns is a sample.

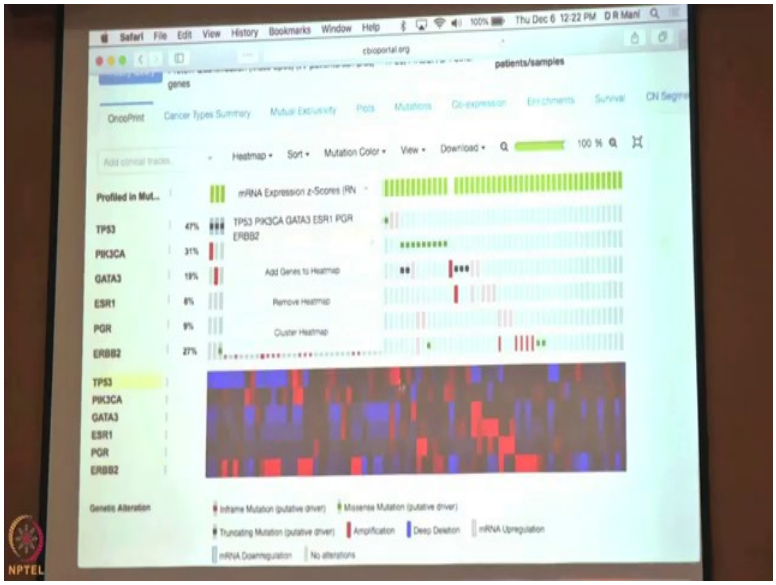
And, then each of these just shows what mutation status based on this here you see for each of the different genes.

(Refer Slide Time: 05:44)



So what we want to do is we do not have mRNA on here, we just have our mutations. So, what we want to do is you go to heat map.

(Refer Slide Time: 05:54)



So, find this heat map and click down on it and you say add genes to heat map and it will give you the gene RNA mRNA expression for each of these genes below the mutation status.

Student: One step back

One step behind.

Student: Yeah. how did we get it?

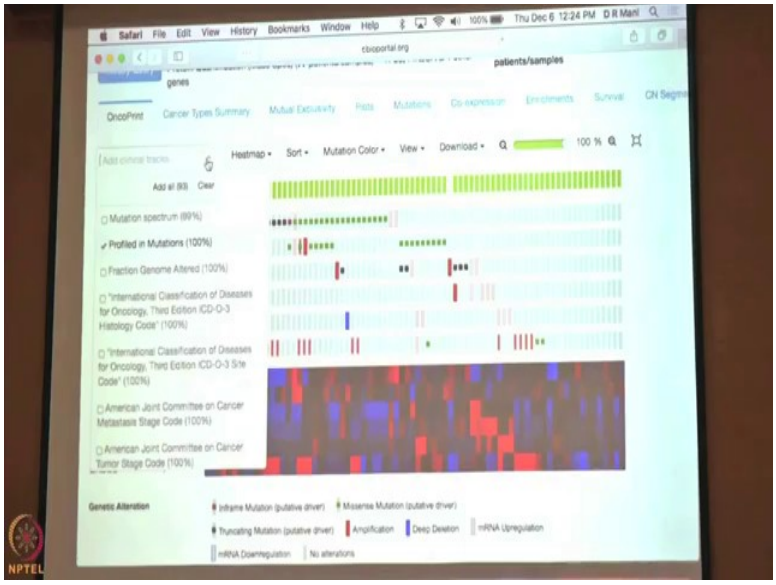
How did we get it? Heatmap; click down on heatmap upon this arrow here.

Student: ok

And, then click on add genes to heatmap and it will add them at the bottom.

And then we just have to do one more thing and then you, it is like all exploring and clicking things and looking at plots. So, one more, we are going to add a clinical track. So, if you look here. At the clinical tracks.

(Refer Slide Time: 06:49)



So, then what the clinical tracks add information about ER PR status HER2 So, anything that is like survival anything that is in the clinical data you add with the clinical track. So, I am just going to have you add one clinical track, let us do HER2 stat. let me see what we looked at here yeah do HER2. So, type in the clinical track IHC HER2 and it will come up and you can click on it and it will add that track to your data.

(Refer Slide Time: 07:28)

Student: Anything now.

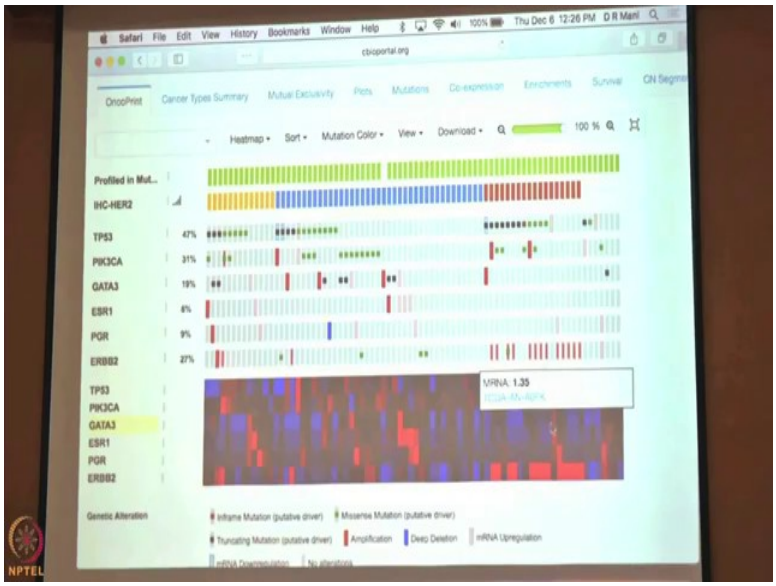
Here.

Is anyone having issues?

Student: no

That I can try good, great. So, what you can do here is you can sort by so, if you click on this the 3 dots next to IHC you can sort so, if you sort A to Z, it will sort the samples by their HER2 level and HER2 is one of the pam 50 genes, it is one of the genes that is used for prognosis.

(Refer Slide Time: 07:53)



It is associated with the HER2 subtype. So, what I have to just show here was that, you can do this. So, if you pull up the Mertins et al. figure 1 B that we had, there was you could see that the ERBB 2 and the HER2 phenotype or HER2 subtype, which is this these reds were very much associated which makes sense, because ERBB 2 is the gene that is actually the same as the protein HER2. So, it makes sense that these are related. So, you can kind of just by using this website, you can get very similar plots that, to the plots that we actually got and showed in that manuscripts.

Student: it is a the percentage of all those mutation is different, I mean is the same what do you have it ok.

Yeah.

Student: Yeah.

Student: But if you see the 27 percent the map is different, than what you have a achieved it? So, if you see the 27th, the last.

Yeah you have 27 right.

Student: Yeah, but if you see it here.

It is because I sort it.

Dr. Mani (Background): So, each by.

Student: So, the.

So, if you sort, if you go to the 3 dots with the IHC.

Dr. Mani: It is the different types of.

Go here click, click on the 3 dots.

Dr. Mani: yeah thats ok Next try this one, these guys.

Student: these one

Oops, no the 3 dots next to it.

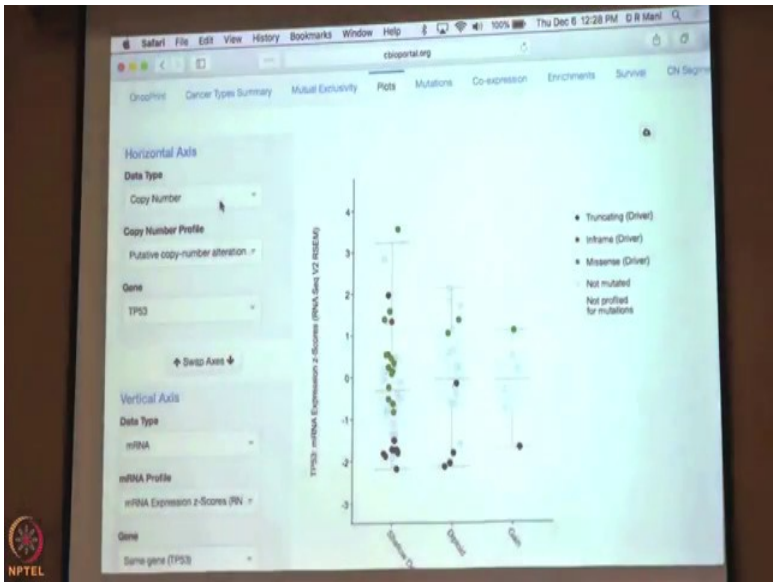
Student: 3 dots no, this one.

Student: Got it.

And you can sort.

A to Z so, ok so, once we have all this, what I wanted you guys so, look at was, if you go to the plots, there is this plots right here you can play around with how copy number of one gene affects the mRNA of another gene or how the mutation status of one gene impacts the mRNA.

(Refer Slide Time: 09:39)



So, for example, if you look at mutation of TP 53.

(Refer Slide Time: 09:56)

An mRNA of TP 53 you see, if there is a missense mutation or a truncating mutation or if it is wild type; it changes the expression of TP 53, based on what kind of mutation you have. So, you can what I would do right now is if you have gotten to this point play around with this data. So, you can see how different data types impacts or how different yeah omics data impacts the other level of omics data for the same gene and then between genes. So, just spend some time playing around with that.

Student: So, one is like the pictorial which I am getting is different, from the pictorial which you had shown. So, what is it?

It is just, because I sorted it, I will show you.

Student: Ok.

So, if you go here and you add in, the clinical tract IHC HER2.

Student: Ok.

So, then it has HER2 status, it adds it and.

Student: So, 1 minute so.

Yeah.

Student: First, we have selected that proteomics with the 77 datasets.

Yeah.

Student: and then now, we are saying select to only for whom we have them the IHC.

Yeah, well this is actually just showing what the results of the IHC are so, you can see whether or not HER2 is up or down in those samples.

Student: Accordingly, and then we can sort it out ok.

Yeah. So, then you can sort it I will show you if you click here sort A to Z.

Student: Ok.

And then it sorts it.

Student: Ok other thing, what you mean by here truncated mutation, because here I can see many yeah.

So, truncated means that there was a STOP codon that was introduced; so, it made the protein shorter.

Student: Ok no, that truncated I know, but it is showing only in the sample, where.

These are it so, each column is a sample.

Yeah. So, there is a lot of truncating mutations.

Student: Yeah.

Associated with ERBB 2, let us start over.

So, if you are starting from the cBioPortal and you have nothing worked.

No, it kept my cBioPortal, here we go.

Ok so, you are starting from the home site.

You are able to see studies, select studies here and you are able to see you can scroll down and see all these different cancer types right. So, if you go to the breast, you scroll down to get the breast.

There should be an invasive breast carcinoma header and under which you will see a whole bunch of different studies.

Right.

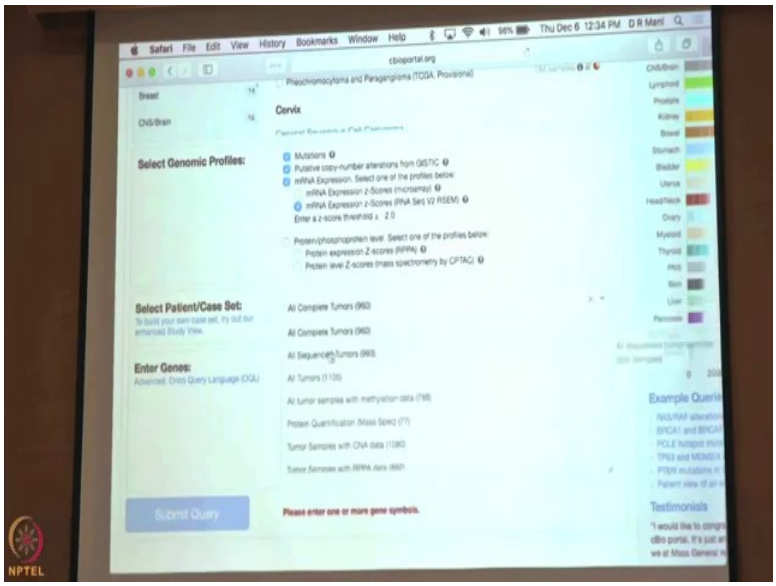
At this point you can click on the breast invasive carcinoma or the TCGA provisional.

Is everyone found that that was not able to find that before?

Ok.

So, that is good we found our samples.

(Refer Slide Time: 12:56)

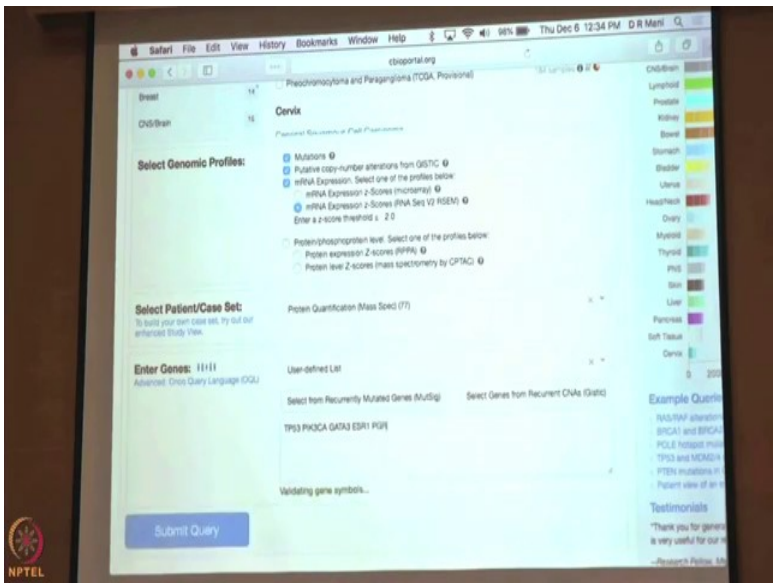


So, now, you can move down and look at the genomic profiles, we have mutations, copy number and then add in the mRNA expression, the RNA seq, which is the second one here. Do not change anything with the threshold, that is fine.

So, for the patient case set.

You go through and you will see different patient case sets and we are going to pick the protein quantification mass spec 77 case set.

(Refer Slide Time: 13:23)



And then we are going to selects a certain genes from our data set and that is going to be the following.

TP 53.

PIK 3 CA.

GATA 3, ESR 1, PGR and ERBB2. So, submit the query at this point. Ok so now, you should see this.

Student: Yeah

If you do not see this.

We see this great ok. So, what this is different each again, each column is a patient and it shows different mutation status for each of the genes, based on the color here you can see that there is a key here to see.

So, the other thing we wanted to do is add in the mRNA expression as that as the underneath, as the heatmap. So, you go to this heatmap tab and you set click, add genes to heatmap and it should populate that.

Yes good.

Were you able to get this? You weren't, ok. For some reason it does yeah.

Student: If I have my own data, you can put your own data on here as well yes, ok. So now, you should have your heatmap. The other there is one other thing to do and that is to add a clinical track. So, what you do here is you just type in IHC HER2 and that should add in another row of clinical data and that just indicates what level the HER2 expression was at when they did their immunohistochemistry and you can sort any of these. So, if you click on these 3 dots you can sort the tracks. So, if you want to sort by HER2 you can do that, if you want to sort by TP 53 expression you can sort by that, if you want to sort by ERBB 2 you can sort by that.

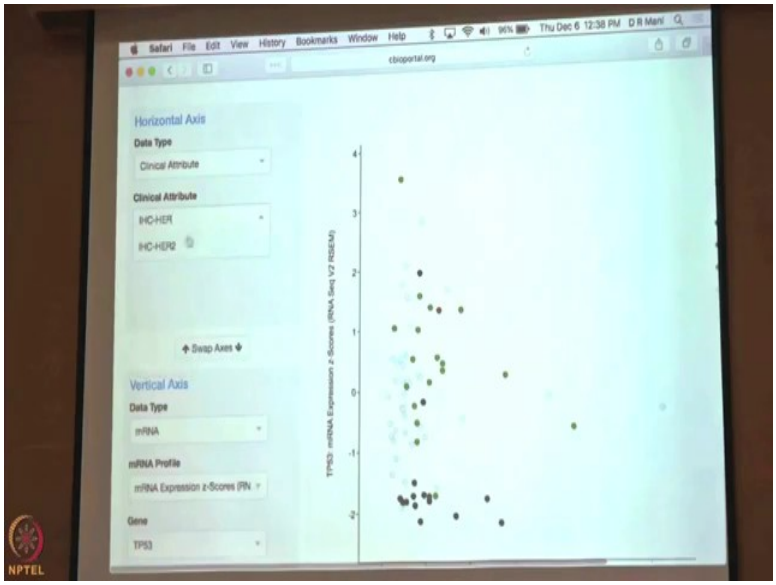
So, you have a lot of control here, if you want to play around with the data and any kind of you can pick any of these datasets you can pick all of the datasets and you can, we could play around with the data quite a bit.

And then so, the other thing, I was just wanted to point out to everyone was this plots. So, if you hit the plots tab you are able to do more quantitative assessments. So, that was more qualitative, you can just see how everything is related looking at these heatmaps. But, if you want to see how does one level of omics impacts another level of omics or how does one gene impact a second gene, you can come here and you can. So, here we have the copy number of TP53 versus the mRNA of TP 53. So, if there is a deletion or a normal or an amplification, you can see how that impacts the mRNA expression. You can change this to mutation and that is one of the ones I was wanted you all to look at. So, I guess we could look at one specific thing within this. So, let us look at how.

I mean picking one of my examples.

So, if you go into, you can change the data type. So, you could put in clinical attribute for example.

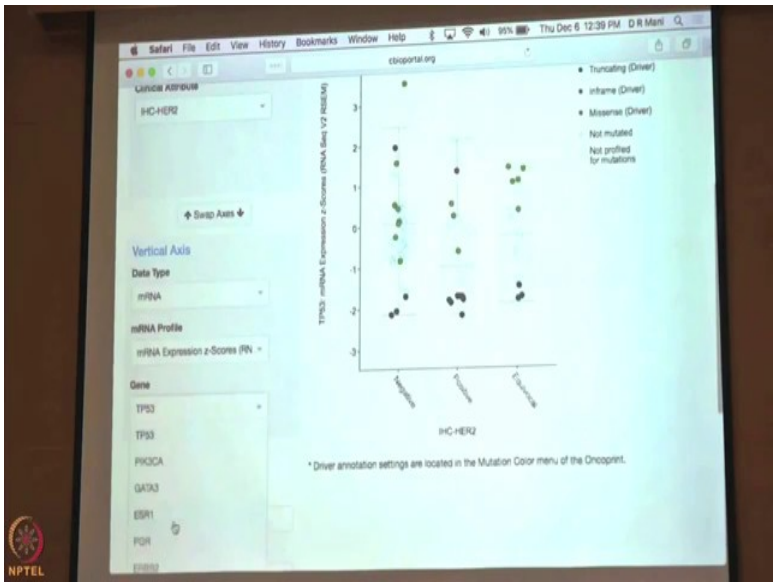
(Refer Slide Time: 16:54)



So, and we put in the HER2 status as their clinical attribute on the last one. So, you can say.

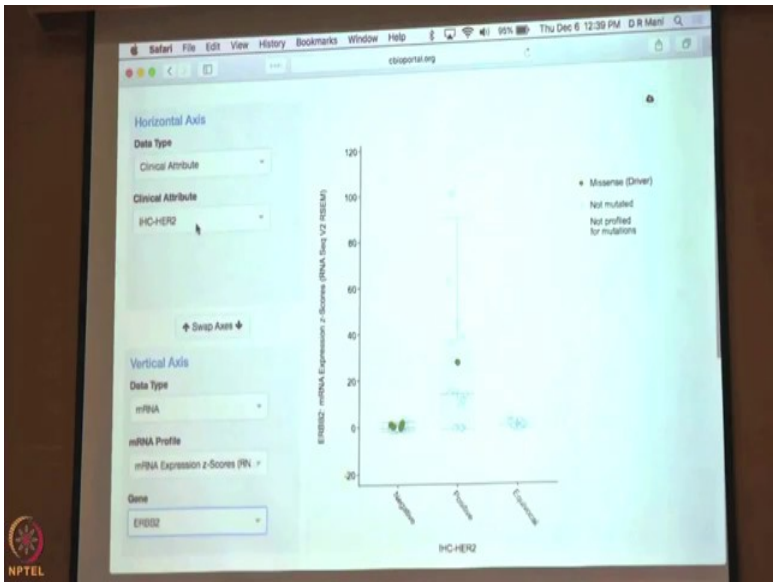
IHC HER2 here right. So, you can get HER2 status.

(Refer Slide Time: 17:04)



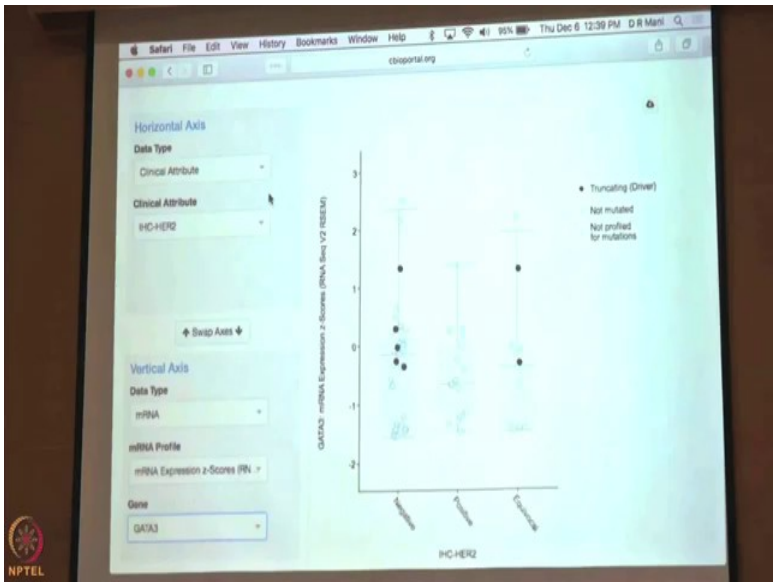
And then for every different genes. so, if you do HER2 status let us say with ERBB 2.

(Refer Slide Time: 17:13)



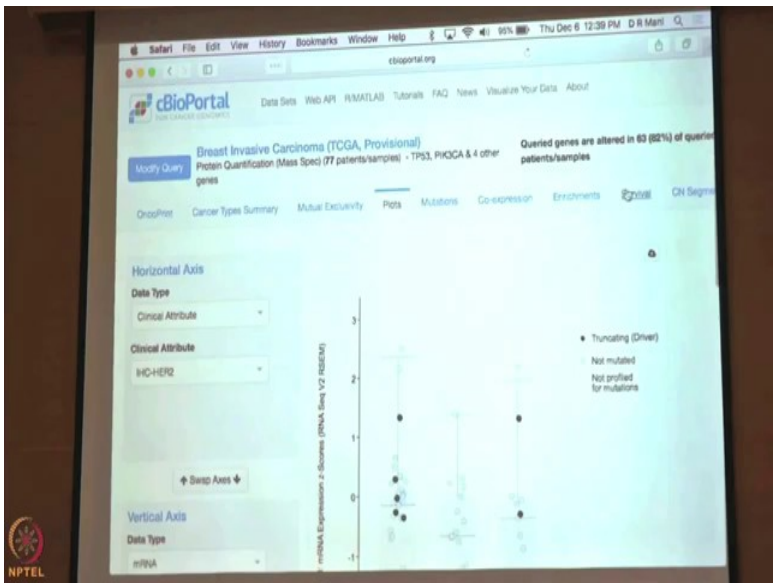
So, you have clinical attribute a IHC HER2. You have negative, positive and then you have, you can look at the different levels of let us say GATA 3, you can look and see how it changes with HER2 status.

(Refer Slide Time: 17:28)



So, I do not think we are going have time to go through specific examples here, but I just wanted you all to at least get to the point where you could play with the data within the portal. And, I mean sort of know that this data is available and that you can ask your own questions of the data and really make some nice plots and do some actual statistics within the act, the within those side.

(Refer Slide Time: 17:55)



You can also give survival analysis. There is a lot of, there you could play around with this all day and really find some neat ways of looking at your data at this data without

having to generate your own data, which is nice. And you know, when we put in those 6 genes you can also just, there is default where you just say highly mutated genes and it will just pick with them for you.

So, it just depends on what you are looking for, if you are just looking at exploring the data you do not have to say anything. But, if you are really interested in this set of genes you can look at all of these datasets and see if your gene has changed in certain at certain omics levels. So, I think it is really, they did a real I think they did a really good job. I had nothing to do with this. So, I am not patting myself on the back.

Student: So, like in which format we can submit the data so, we can get this kind of output that is more important

So, yeah I have to look and play around with it. I have all of the data, I have ever needed they have already put on there for me, but I know that you can upload your own data.

So, I can look into that and then we can talk about it. Yeah, I know that you can.

Student: What we can do with VCF file.

It is probably at it is going to be expression level data like VCF file; so, really processed data.

Student: So, sample.

Now, like raw data.

Student: What these sample, the mass spec data is also collected.

The mass spec data yeah, they it should be in there as well, I did not want to go into that, because we have not done proteomics yet, but yeah.

Student: right

Yeah.

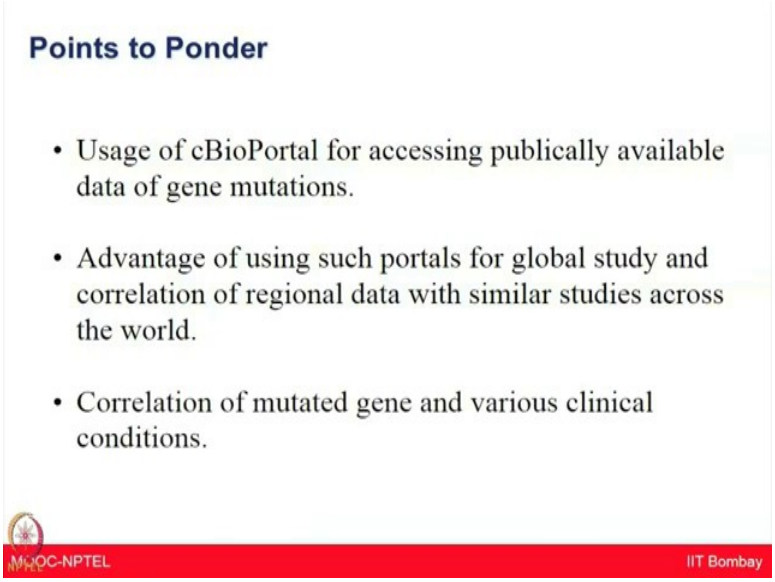
Student: I have a sample from, you know responder, non-responder, survival like that. So, I wanted to draw this graph; so, I would like to know what type of data format.

This is probably easier, I think there is easier ways of doing that then uploading it to this and I can send those to you, they exist.

Student: Ok.

This might be harder than its worth.

(Refer Slide Time: 19:42)



So in conclusions, today you have learnt how one could study the mutation patterns, in various cancer types in the global population. And, there are online tools a lot of information available which could be leveraged, utilized to first get a very good idea about possibilities of gene mutations for a given tumor type. In the next lecture, we are going to have another speaker; Dr. Bing Zhang, who will talk about the correlation of variations, in genotype, gene expression and its phenotype.

Thank you.