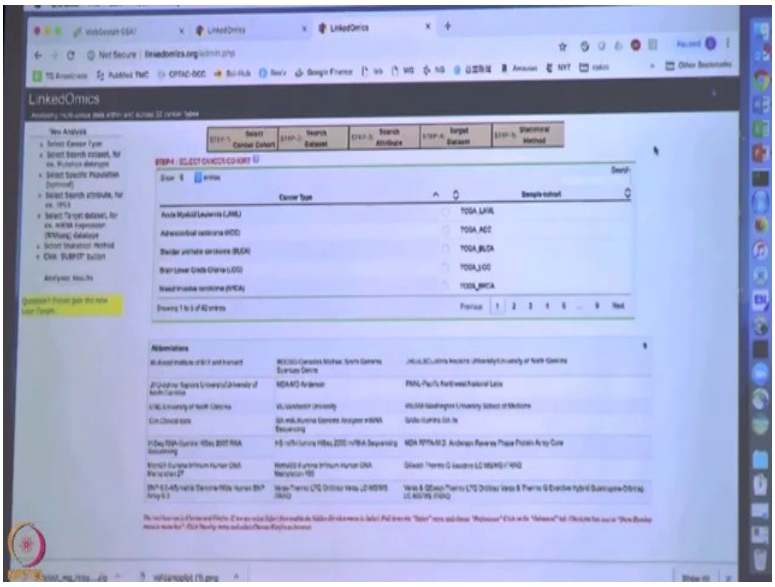
Introduction to Proteogenomics
Dr. Sanjeeva Srivastava
Dr. Bing Zhang
Department of Biosciences and Bioengineering
Baylor College of Medicine
Indian Institute of Technology, Bombay

Lecture – 59
Linked Omics (Part II)

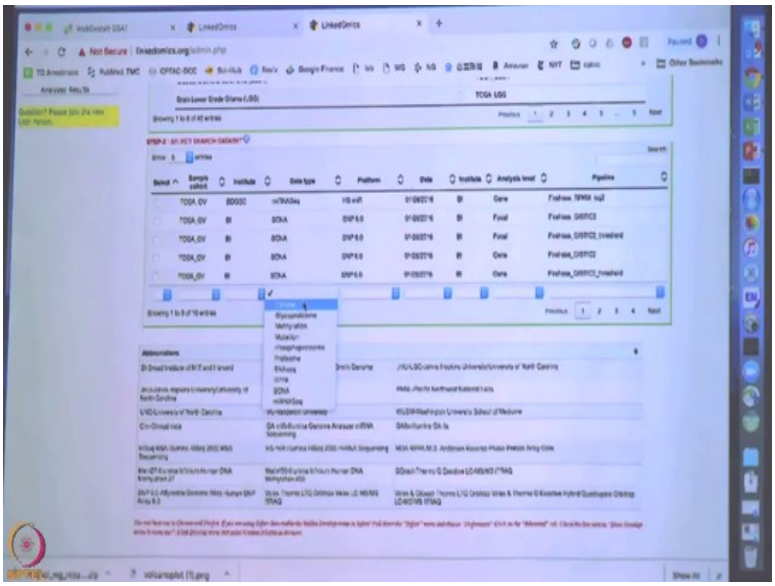
Welcome, to MOOC course on Introduction to Proteogenomics. In the previous lecture, Dr. Bing Zhang provided you the capabilities of an online tool Linked Omics. In today's lecture, you will be exposed to the various steps involved in analyzing large dataset using linked omics. So, let us welcome Dr. Bing Zhang for his last lecture.

(Refer Slide Time: 00:51)



So, let us go to the new analysis and then there a few steps you need to follow in order to perform this analysis let us start with proteomics data. So, basically we want to ask which proteins in ovarian cancer associated with poor prognosis and which proteins associated with poor prognosis in ovarian cancer. In order to do this first you have to identify the ovarian cancer data set to do the analysis on ovarian cohort. So, you can browse, but let us do a search bearing.

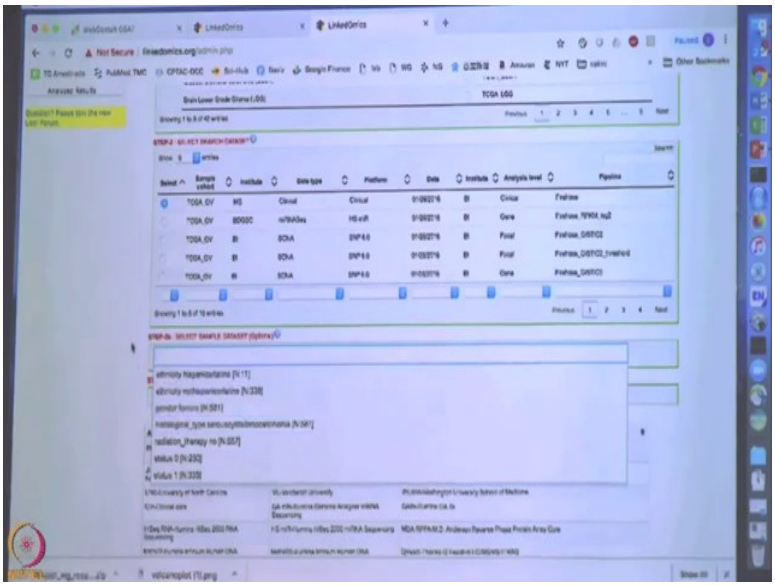
(Refer Slide Time: 01:35)



And then you can identify this TCGA underscore OV that is the cohort we want to look at basically it is a TCGA ovarian cohort and then you click on that and then it will give you the next step select search data set. So, as I mentioned the go of the linked omics is allow you to go from any attribute. For example, you can go from the mRNA or microRNA or mutation or proteomics to any other attribute, but here we have a fixed question we are interesting survival right.

Survival is a type of clinical data. So, in within the step 2 we want to define the search data set and the search attribute.

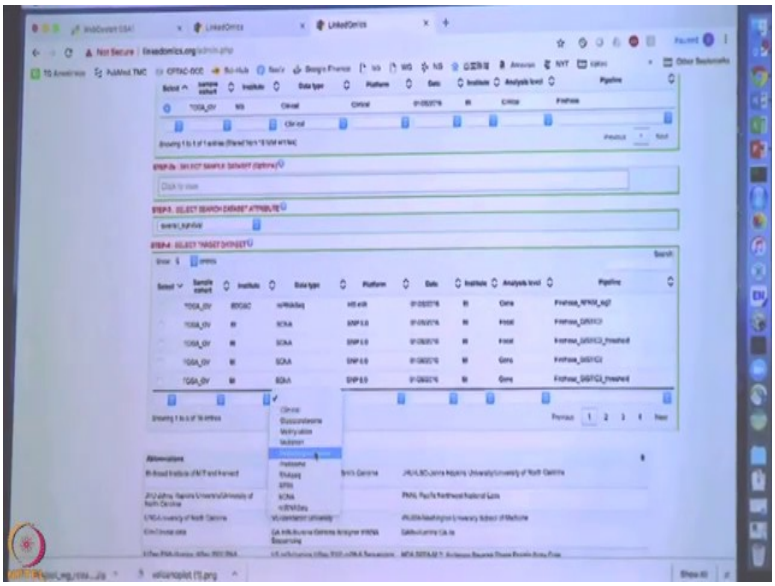
(Refer Slide Time: 02:28)



So, here we want to start with clinical data let us say you for clinical data and there is only one type. Let us say you select the clinical data and now you have step 2b as an option, but not say if you are only interested in for example, a certain stage of ovarian cancer or maybe you are only interested in a certain subtype of the cancer and then you can do the analysis for the subset of tumors.

But, today let us just say we do the analysis for the whole cohort and then we select the search within the clinical data right the multiple types of clinical information as the platinum resistance data or overall survival. So, today we are going to select the overall survival.

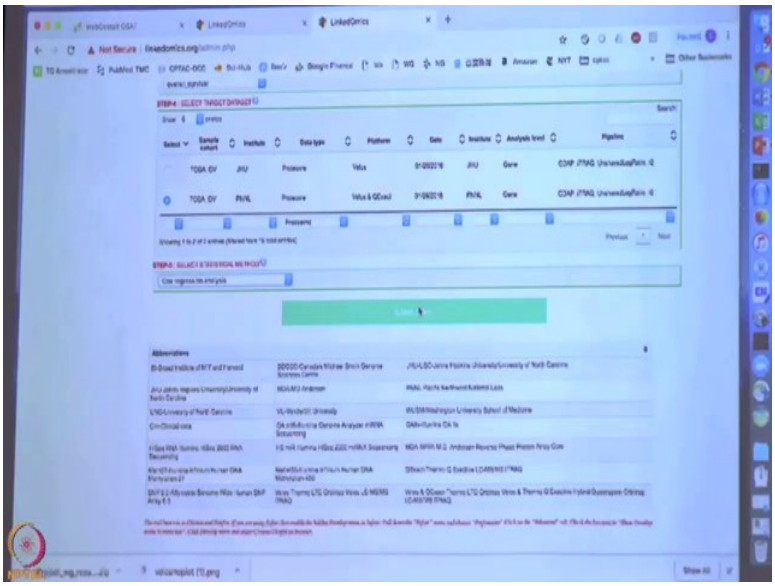
(Refer Slide Time: 03:26)



Now, basically you define the what to use as a query attribute basically it is within the TCGA ovarian cohort I am interested in the overall survival and then I want to ask and which proteins are associated with overall survival right. So, then on the step 4 the data type let's select the proteome. So, for this cohort both John Hopkins university and the PNNL they both did proteomics data generation for kind of overlapping subsets of the cohort.

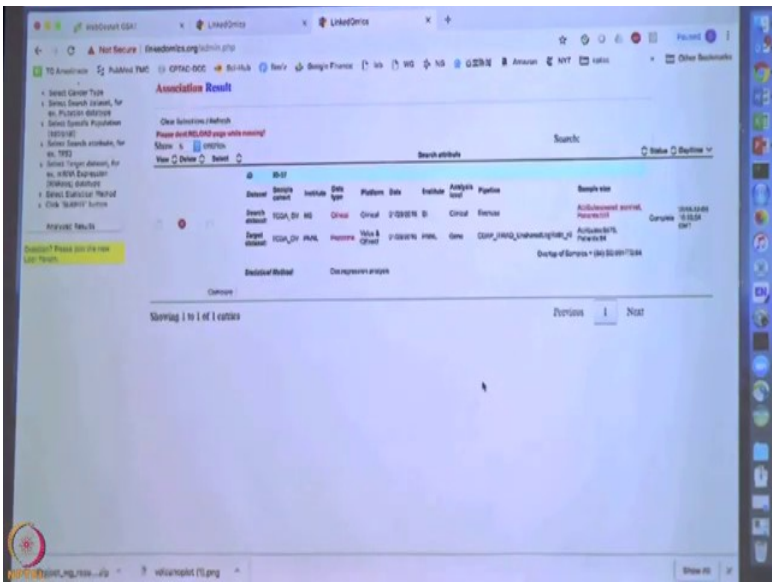
But, today let us just say we choose a PNNL data. You click on the Select next to the PNNL data, but here you after we finish this proteomics analysis you can come back and then you can select for example, RNA seq and then you can correlate the overall survival to RNA seq or you can select a copy number and then you can correlate survival to copy number. But, now we are going to do the proteomics only.

(Refer Slide Time: 04:42)



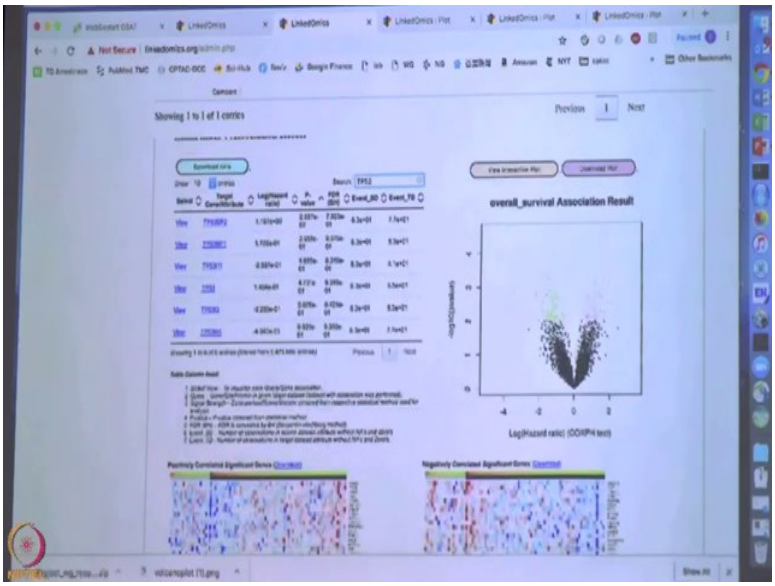
And, the another step 5 let us select Cox regression. So, this is for survival analysis and then Submit Query.

(Refer Slide Time: 04:51)



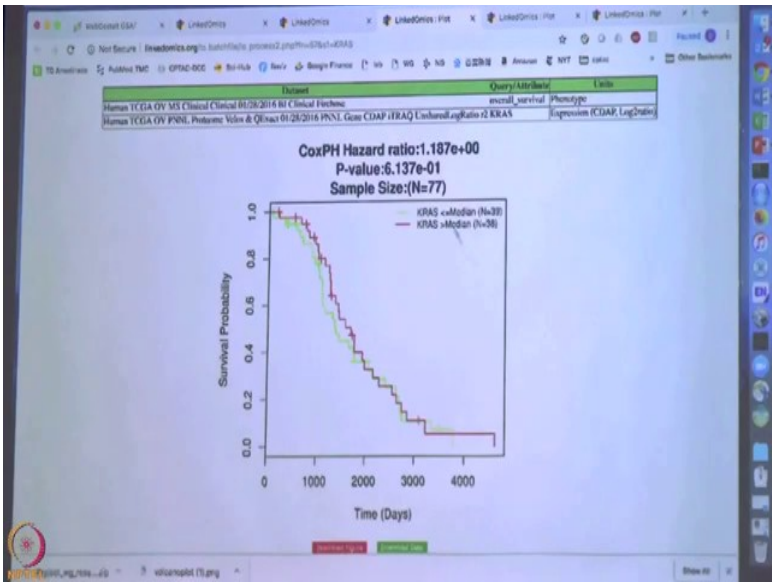
Any query that has been queried by other people because some other people already done this analysis and the result is saved. So, that is why it is so quick and now you get this and then you click on the view.

(Refer Slide Time: 05:05)



And, you can also search the gene you are interested in. For example, KRAS or some other genes let us say ovarian cancer, what gene could be interesting anyone has any suggestion on which gene you want to search for? Maybe let us look at the KRAS.

(Refer Slide Time: 05:31)

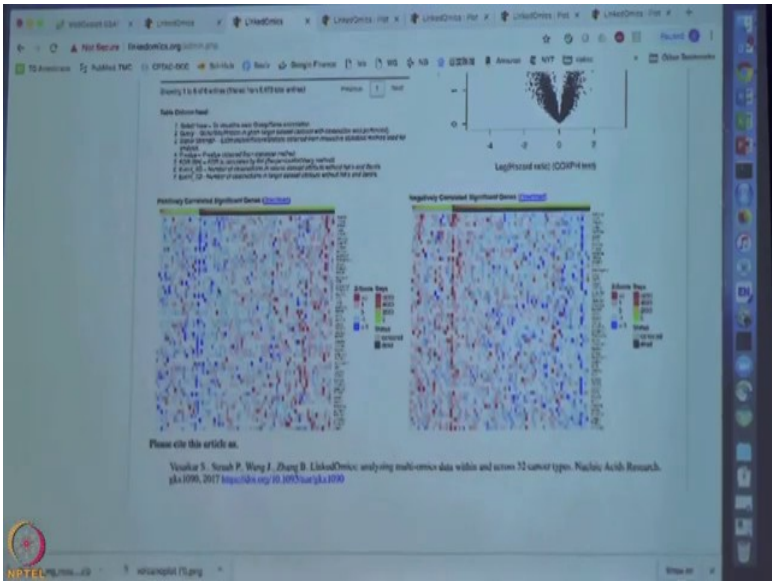


It looks like it is not significant. The P-value is 0.6.

Student: TP53.

TP53 it is not significantly either, but is apparently there is little bit of trend like higher expression is associated with a poor survival.

(Refer Slide Time: 06:15)

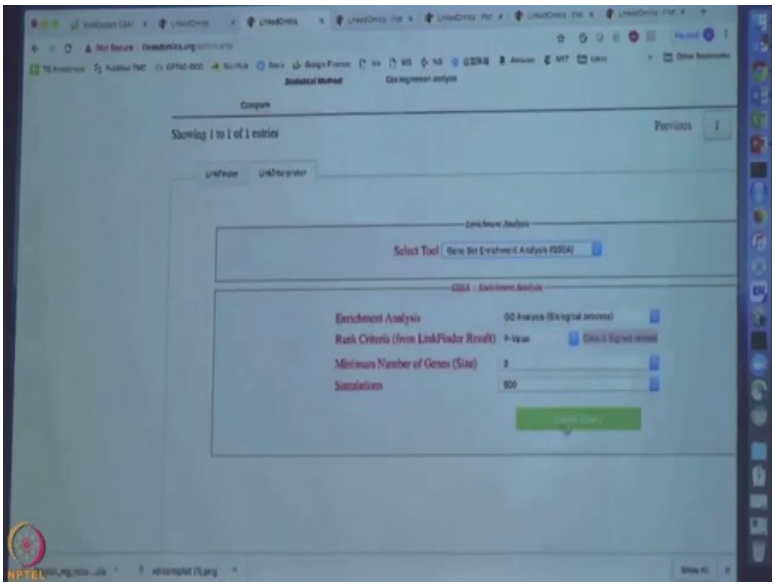


And you can also go to the bottom of this page and this will show you the top 50 genes.

I do not know 50 or 25 genes that are most positive in your correlated with survival or most correlated with survival. So, we found this to be useful for in.

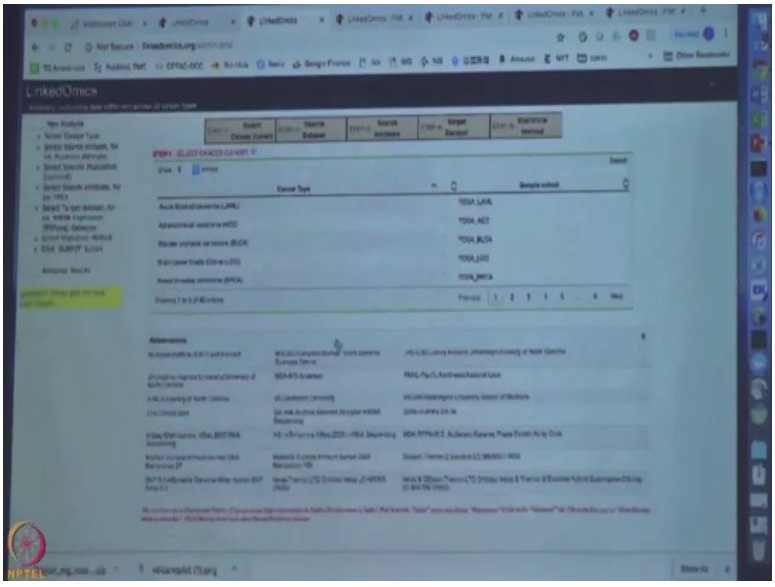
Yes, get a quick a view of the top genes and the detailed data. It is binary data it is. Subtype of one versus other subtype or and then you just have black next this two parts right. But, for survival data it is only you have test event and the still alive or dead. Right, but then you also have the time of survival or. So, let us me the first this part at the bottom shows you these are probably that. People who have died and these are still alive and then within each group and then this is a survival time. So, some most of these are censored data right most of them have not died. So, it is censored, but these are the ones which actually test events. So, basically two group of people and then within each group you have mean survival time from 0 to 9.

(Refer Slide Time: 07:57)



So, I think and then if you want to do some pathway analysis for this you can click on the Link Interpreter and then you can do ORA or GSEA similar to what we have done before, but with if you submit the query and basically it will give you the WebGestalt report and we have done that before. So, let us say we do not need to do that now rather I think we can go back to perform a new analysis.

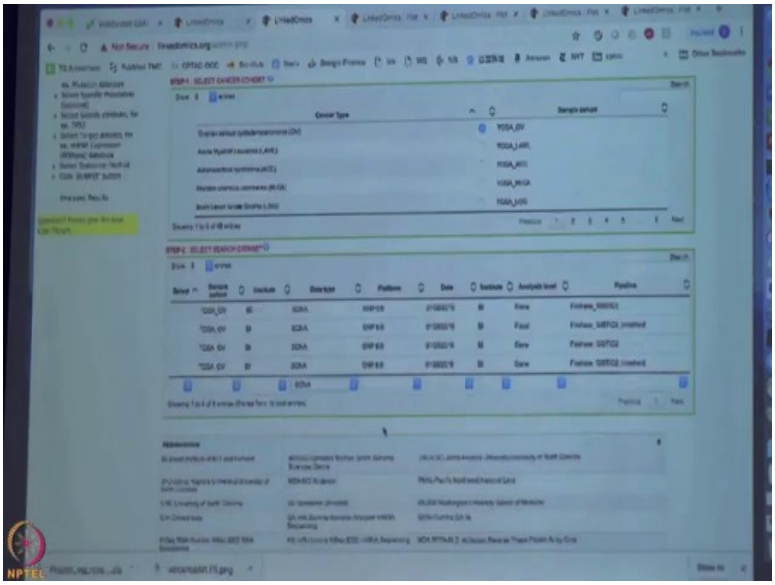
(Refer Slide Time: 08:22)



And, the in this time you still do the same analysis, but you ask a slightly different question you ask which copy number change is associated with survival in this ovarian cohort and the after that you do another query which RNA in the RNAseq data which mRNA abundance is correlated with survival. And, after you get results from all those three platform and I can show you how to do show you how to do the link compare.

So, you can go back to new analysis and the repeat what we have done, but just change the target set target data set.

(Refer Slide Time: 09:11)



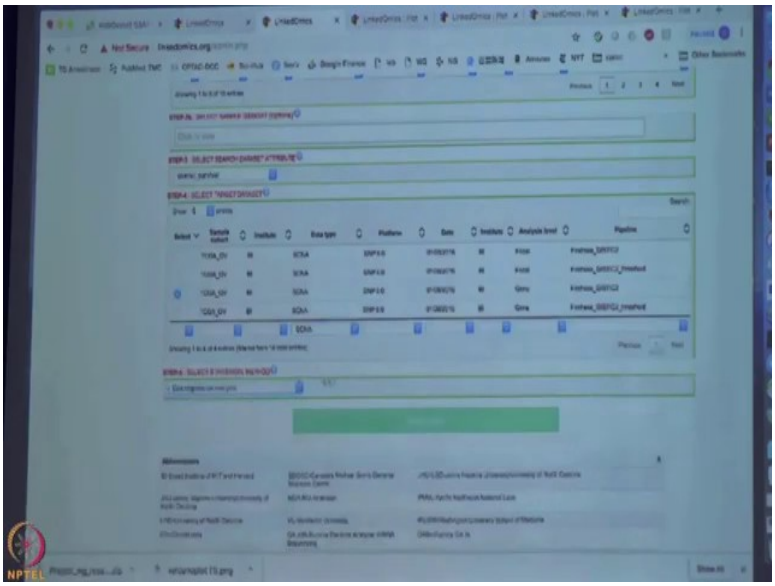
The copy number is SCNA and another for if you filter the data type for SCNA and then you get multiple four rows right you because the results can be we have results at the both focal level and the gene level and then at the gene level number the data can be stretched coded or without stretch code. In this case we pick the one without stretch code pick the third one.

Yeah, We pick the SCNA gene level Firehose GISTIC2 that is the algorithm used to do the analysis.

Ohh, Sorry in the at the very top you select clinical, I am sorry.

We should select clinical and it is in the target data set that is so many copy number.

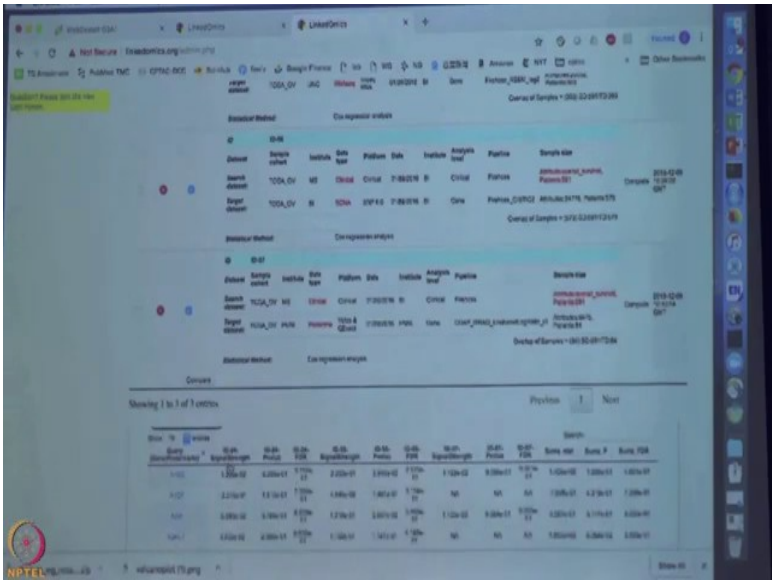
(Refer Slide Time: 10:17)



It is step 4 step 4 Select copy number.

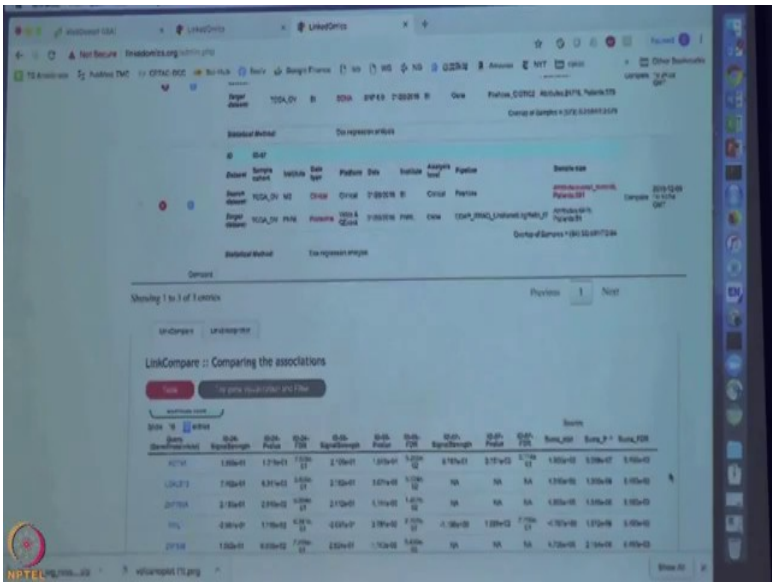
And, then you can also view the results for copy number the same way ok.

(Refer Slide Time: 10:43)



Let us say if you were able to get two or three results from this for example, if you get all the three RNA seq, copy number and the proteome in this OV data set associated this survival. Now, you can do the link compare and what you do is you select on this the third column this select column you select the let us say all the three of them and then you click on compare. Now, you have this table with all the results.

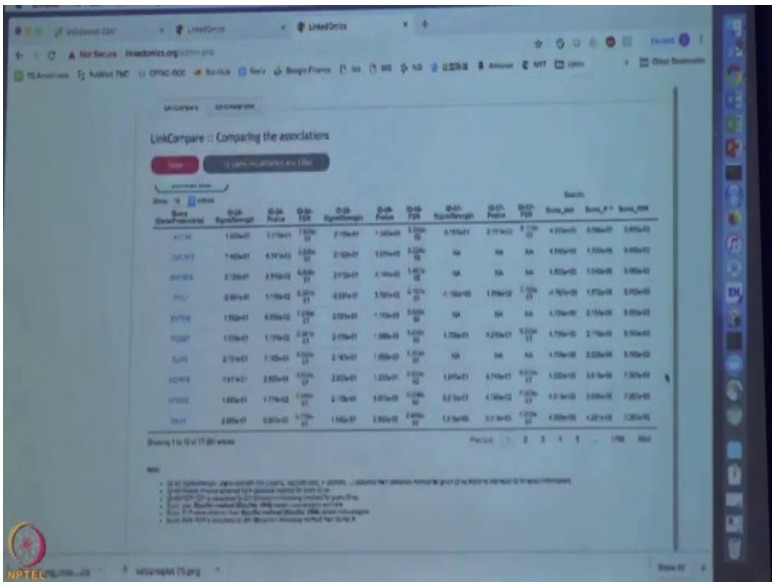
(Refer Slide Time: 11:32)



So, basically in the first few columns you have the three results from the three platforms lets say this is ID 24 correspond to this ID 24. So, this is the results from RNA seq and the then you have ID 56 this correspond to the results from copy number and ID 57. This is a result from proteomics.

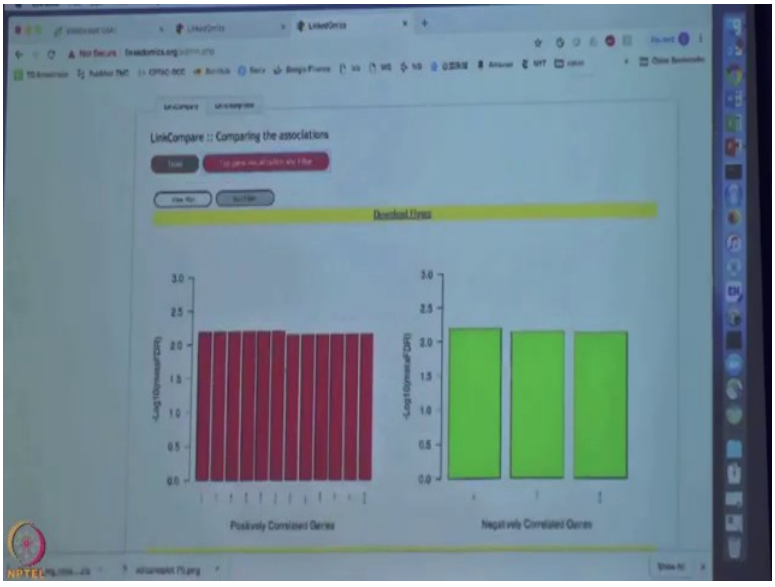
So, basically each of this you have the signal strengths meaning the signal ratio and then you have the P-value FDR right. And, then here we use the sumZ which is a meta-analysis method try to summarize a P-value from this three analysis and then you get to sumZ statistic and then sumZ P-value and sum FDR. So, let us sort the results based on the P-value.

(Refer Slide Time: 12:24)



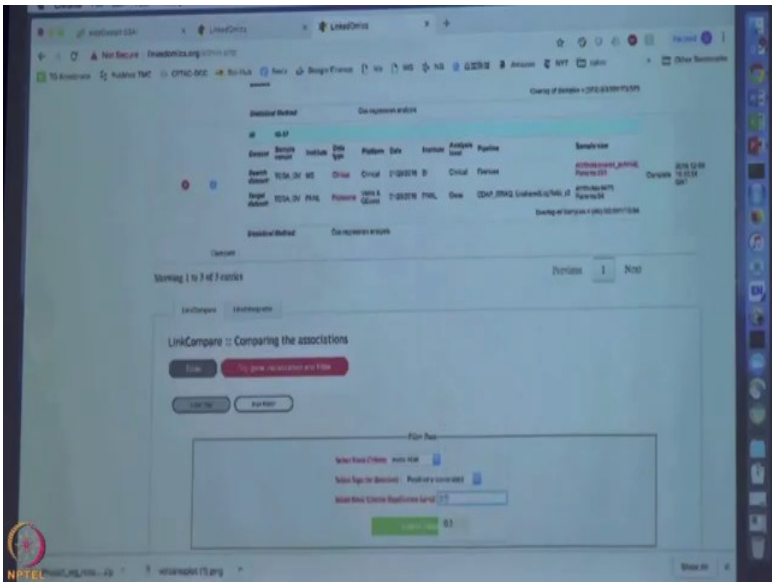
And, as you can see now before for individual platform we get very few significant genes right. In proteomics we did not get any gene that passed the FDR cutoff but now we can see after you integrate straight platforms the a lot more significant genes than individual platform. This indicate although each platform do not give you the signal strong enough to pass a cut off, but maybe all of them point to the same direction that gave you more confidence. So, when you put the result together and you get the enrichment.

(Refer Slide Time: 13:07)



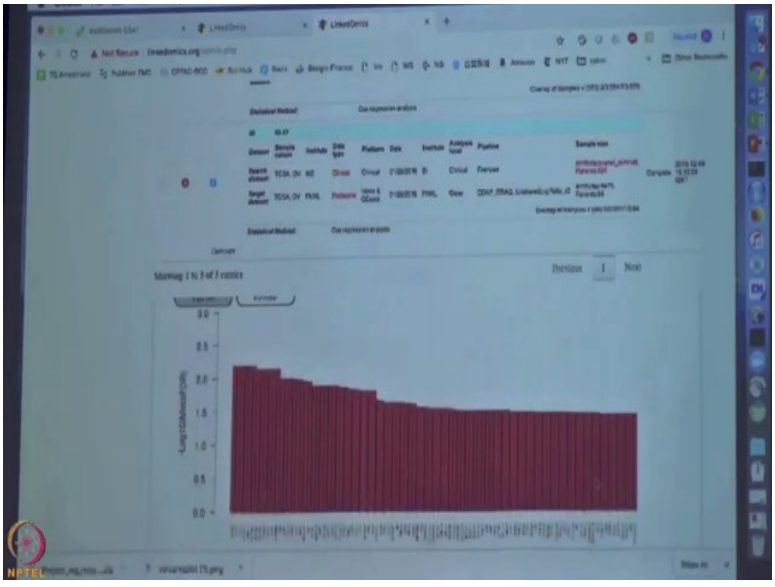
And then you can look at the top genes to see the minus log and meta FDR meaning this is integrated FDR and these are the top genes positive genes and the negative genes. And, then you can also have the result in the heat map.

(Refer Slide Time: 13:38)



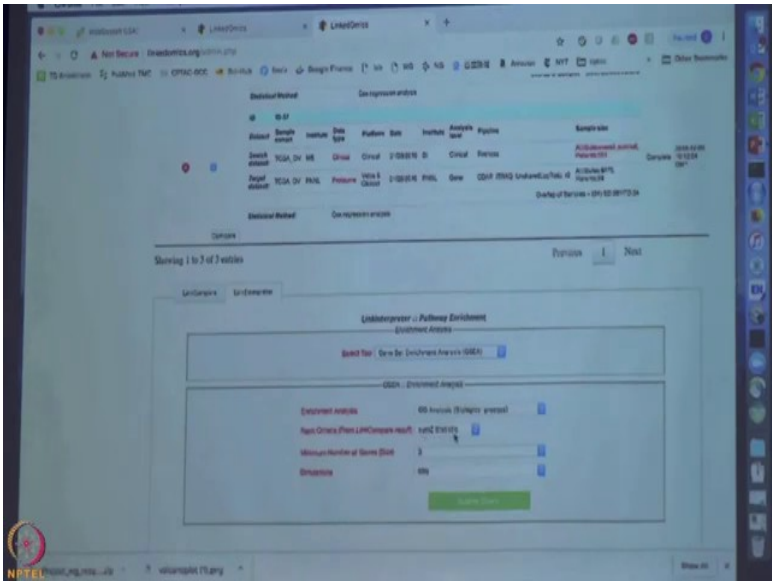
You can also redo the filtering and for example, you run the filter rather than setting the FDR as 0.05 let us say if you make this to 0.1 submit will give you more genes to look at. So, basically you can adjust this to get the genes that pass your cutoff. For example, now if we relax the FDR and then we get more genes to look at.

(Refer Slide Time: 14:03)



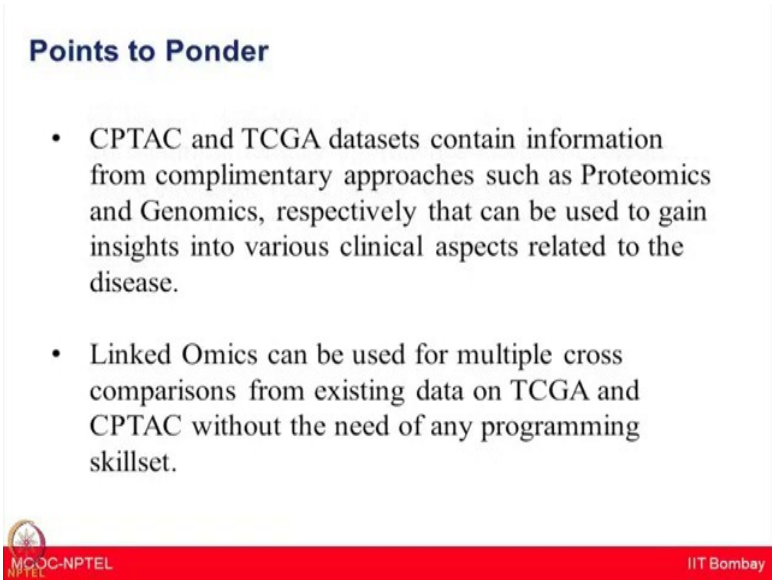
And, the all these results also can be downloaded and saved by click on the download figures.

(Refer Slide Time: 14:24)



So, again and the you can go to the Link Interpreter and for example, select GSEA and then now you can submit this sumZ statistic as rank metric to rank the genes and then the analysis will be perform the enrichment analysis will be performed against this summarize the statistic rather than individual ones. Again, the result will be the same I mean very similar to what we saw before for the GSEA in webGestalt

(Refer Slide Time: 14:59)



In today's session you got a demonstration about how the three modules of linked omics work using ovarian cancer TCGA dataset as an example. It also shown how the tool can be used to generate survival information from a target gene. Using data from RNA seq, copy number variation and proteome level information on the effect of target gene on survival was demonstrated.

Finally, using the third module of the tool it was demonstrated how the results appear and it can be interpreted. In conclusion, we hope that now you have a fair bit of idea how you can use these available tools and use them for your own research.

Thank you.