

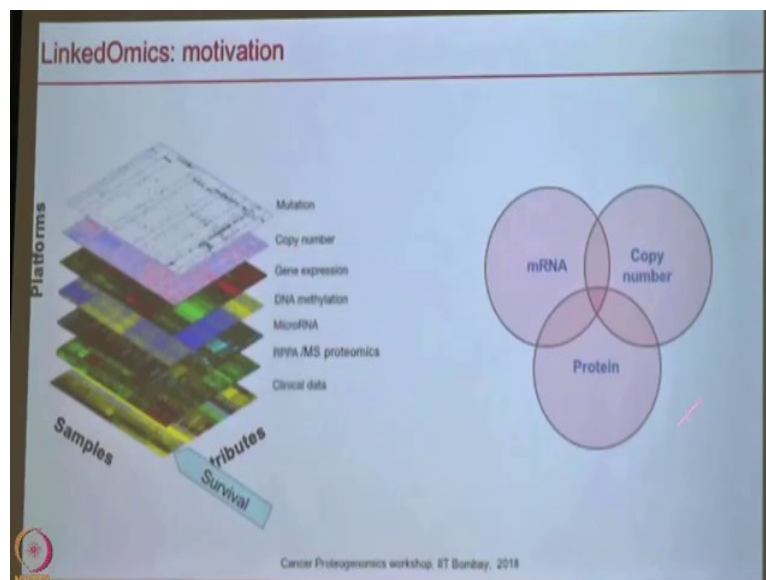
Introduction to Proteogenomics
Dr. Sanjeeva Srivastava
Dr. Bing Zhang
Department of Biosciences and Bioengineering
Baylor College of Medicine
Indian Institute of Technology, Bombay

Lecture – 58
Linked Omics (Part I)

Welcome to MOOC course on Introduction to Proteogenomics. In the previous lectures, we have learnt about how data is generated from various omics technologies. The amount of data is very huge and it is very challenging to make meaningful insights from the big datasets. To understand the mechanisms at multiple levels data visualization tools make the job easier.

For example, a tool can help a researcher find the correlation between a gene with its mRNA or protein or even micro RNA. In today's lecture, we will take a look at Linked Omics which is an online tool that helps in visualization and correlation of multi omics data set. So, let us welcome Dr. Bing Zhang for today's lecture.

(Refer Slide Time: 01:19)



So, first I will just give you a brief introduction to the motivation and the basic functions in linked omics what we can do there. So, this is the tool basically you try to bring the data and the tool together like WebGestalt and it is just the tool right you have to provide you your

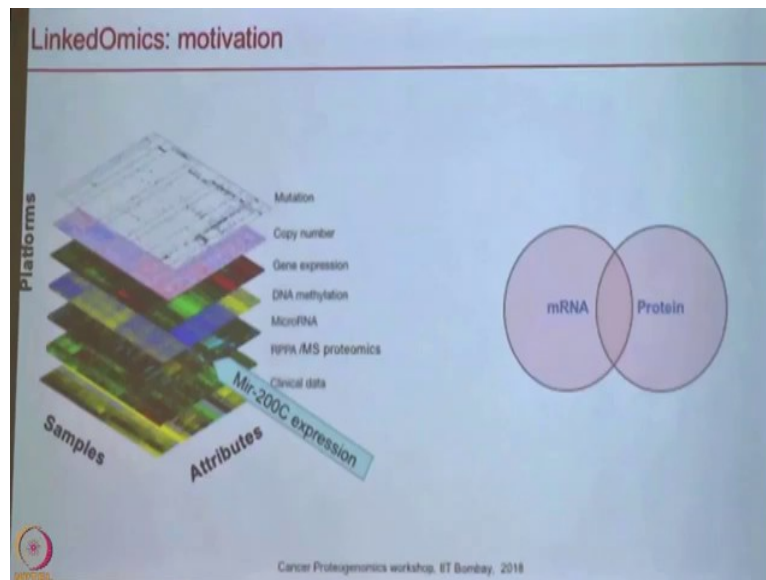
own data about results in order to do analysis. But here I think the motivation for this project is that in recent studies makes the TCGA and CPTAC has produced a huge amount of data, and for us especially and for some of you who are work in the cancer area, and this provide very important resource for us to explore. But for ordinary biologists who do not know how to program it is not easy to get access to the data and also do the analysis.

What we want to do here is to try to develop tools centered around this data resource and then allow everyone to access and use a data. So, then the question is of course, with, so the huge amount of data you can do a lot of sense right, but we are asking what are the most typical questions like biologists will want to ask about this data set.

Of course, one question a lot of people are interesting is about survival. Let us say if you have survival data and then you of course, want to ask and it maybe which mRNA associated with survival and also maybe you also want to do the analysis that is a copy number want see any copy number change associated with survival or you can also do this as the protein level to say which protein chains associated with survival.

So, you can do this analysis separately, but sometimes we also want to compare the results maybe you want to prioritise some macros and then you want to say do I see RNA genes that are commonly associated with survival at different omics level , meaning copy number change RNA and the protein all associated with survival, right. Of course, you might be able to identify some unique sense, make gene whose protein is associated with survival, but mRNA and the protein are not.

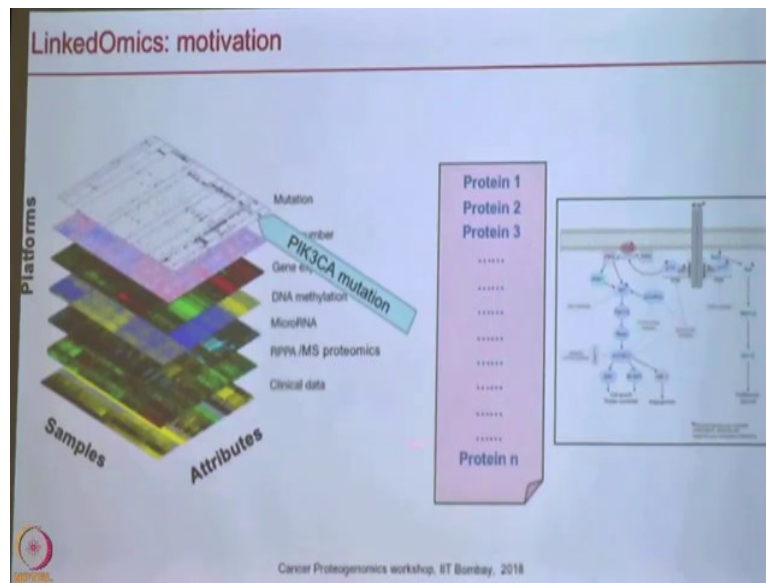
(Refer Slide Time: 03:52)



And the not only survival, and if you are interested in the biology, let us say if even if you are not interested in cancer, but you are interesting a particular microRNA let us say microRNA200C and then you want to see what are the target genes of micro RNA 200C, right.

We know micro RNA might inhibit gene expression through mRNA decay or inhibit translation right, then we can correlate the micro RNA expression with mRNA always protein and this can give us protein. So, mRNAs that are negatively correlated with micro RNA and those could be the potential candidate for targets of the micro RNA.

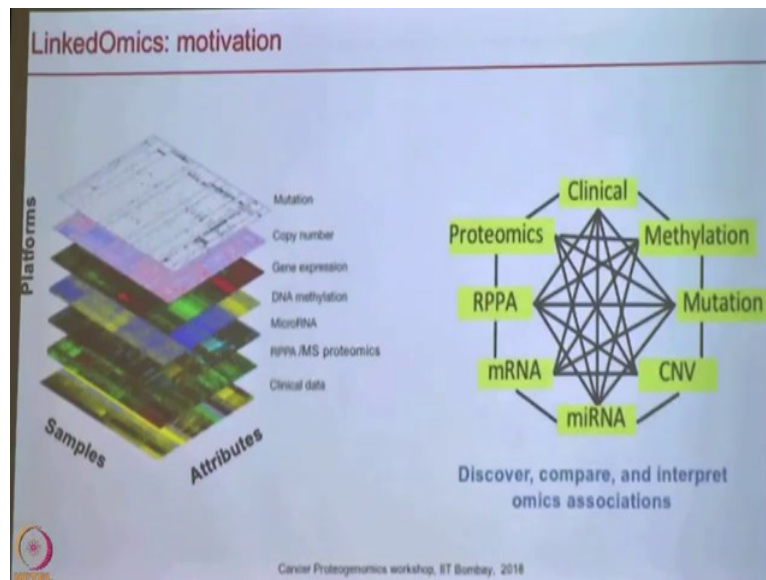
(Refer Slide Time: 04:42)



And then you might be interested in a mutation. Let us say PIK3CA mutation and then you may want to see: what is transcriptomic or proteomic consequence of the mutation, right. Let us say we have this mutation maybe we want to ask in the proteome or even the phosphoproteome, which changes are associated with this mutation.

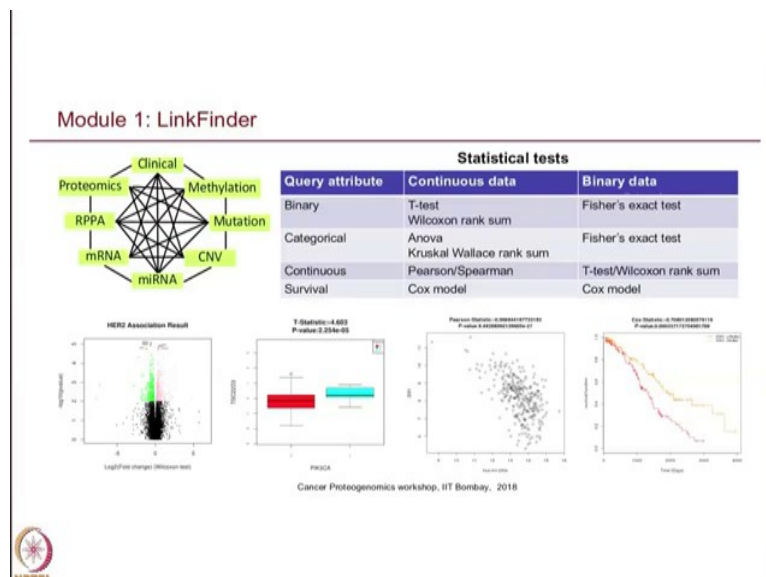
And of course, for all these analysis we end up with a list of genes or some statistical analysis results and then we also need to convert this to pathway and the network and understanding, and then we need to do some pathway that enrichment analysis. This type of turn to better understands the results.

(Refer Slide Time: 05:33)



So, together and we want to build a tool that can help users to use this data to discover compare and the interpret omics associations. So, you can start with any of this omics platforms or from the phenotype and then you can get, get connected to any other platforms, let us go, and then you can also compare your results across cancer types or across platforms.

(Refer Slide Time: 06:05)



So, in order to do that, we develop the 3 modules in the system to do this. First, is called the link finder. So, basically from any of the attribute you are interested in, like we said it could be survival or microRNA a expression or any gene expression or protein expression or

mutation and depending on the and then you want to compare it which is the other space meaning. For example, mutation against phospho phosphorylation or phosphoproteome or micro RNA against the proteome or a transcriptome right and then depending on the data type in your query attribute and also in your target space search space and you have to choose great statistical test in order to do the analysis.

I think Dr. Mani already talked about many of these tests and for what type of data you should we use which type of statistical test. But this provides a basic summary like if you have binary as your query dataset and then your search space is continuous then you use a T-test or Wilcoxon test one of them is parametric and data is non-parametric.


If both query and the target are binary then you use Fisher's exact test. So, but if you interest in survival as query and for continuous data you use cox model to do the analysis, but all these tests have already been implemented in a linked omics. So, you can just pick the right and or the system actually can help you to recommend the right test for you to do the analysis.

And after you do that without to do the analysis, you can get your overall result as a volcano plot. Here you have the effect size or for example, the t statistic on the x axis and the minus log p values and the y axis the volcano plot, show you the results and then for individual genes you have the differential results or correlation results or survival result showing us different types of plots.

(Refer Slide Time: 08:36)


Module 2: LinkCompare

- Compare two or more sets of LinkFinder results
 - Multi-omics analysis
 - Pan-cancer analysis



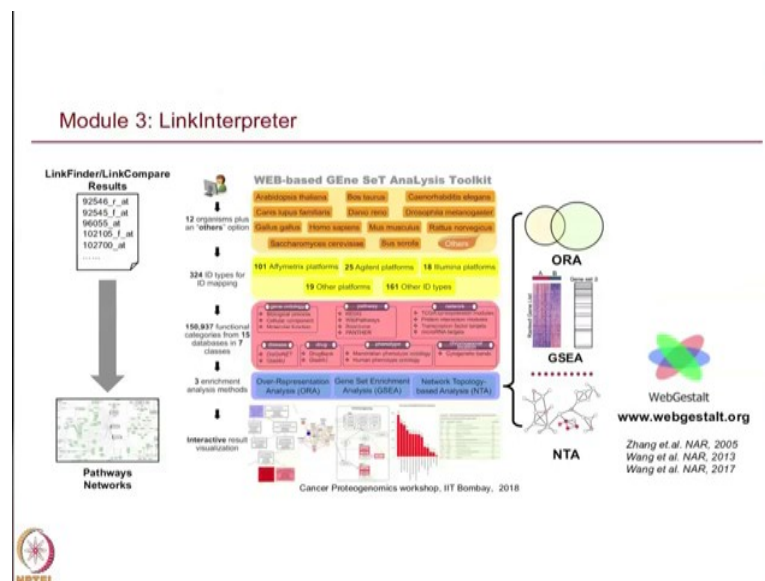
Scatter plot Venn diagram Meta analysis

Cancer Proteogenomics workshop, IIT Bombay, 2018



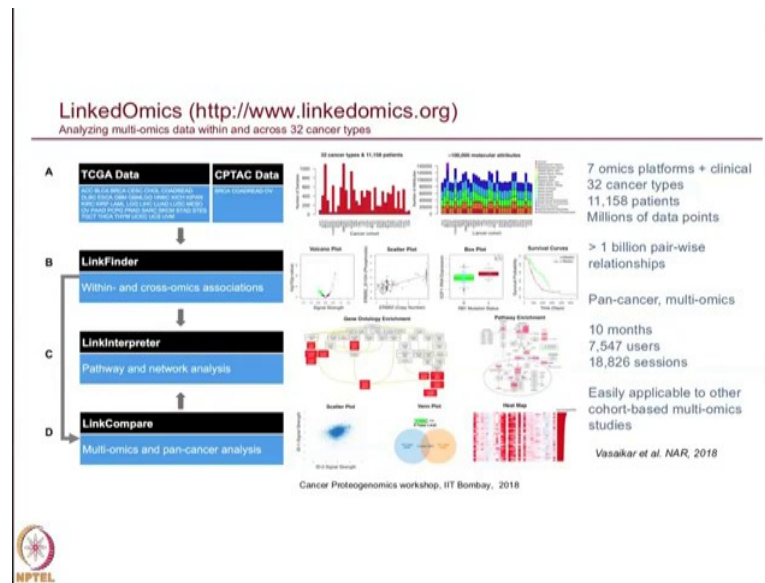
And then for link compare; basically, it can give you some visualisation to compare the results from multi-omics studies or from the pan cancer studies. For example, if you have mRNA or the transcriptome correlated to microRNA200C and also the proteome also correlated to the 200C, you can have a scatter plot to compare the result or you can have after you have some significant the genes you can use Venn diagram to compare or you can you if you have lets say survival results for many cancer types and then you can use a meta analysis to compare the results.

(Refer Slide Time: 09:26)



And, the link interpret the part is easy to understand now because basically values in the WebGestalt to the link interpreter.

(Refer Slide Time: 09:37)

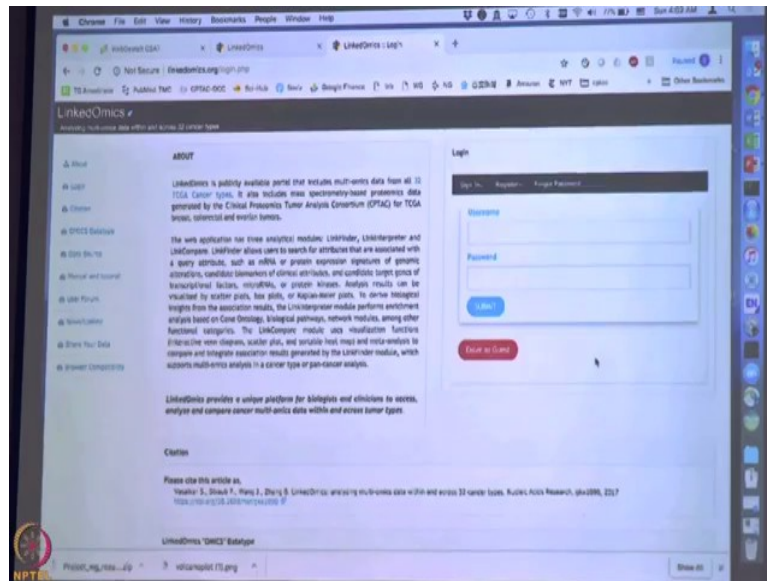


So, this is an overview of the system. It has the data from TCGA and CPTAC and then it used link finder to identify starting from one query attribute could be survival or microRNA expression or mutation or anything we are interested in and then you define a search space and then you get some results here, and the results can be visualised and then you can compare results across different the platforms or across different cancer types using this visualisation or data analysis. And, then the results from this tool can be used as input to an interpreter to generate the pathway level results.

We are going to use the ovarian cancer survival related genes as an example. So, the idea is that ovarian cancer has been studied by both TCGA and the CPTAC, right. And then from say TCGA we have copy number data, we have RNAseq data and the from CPTAC we have proteomics data and then we want to ask this question which genes are correlated with poor prognosis in ovarian cancer and based on all the copy number and mRNA and the protein it is an interesting question, but it is not that easy.

I mean, if you want to do it by yourself you first have to download the data from the TCGA and then you have to learn R to the associate survival analysis and then you have, so basically, there are not options you need to do in order to achieve this. But with the linked omics you can do this actually in probably just 20 minutes of course, without the traffic, internet traffic.

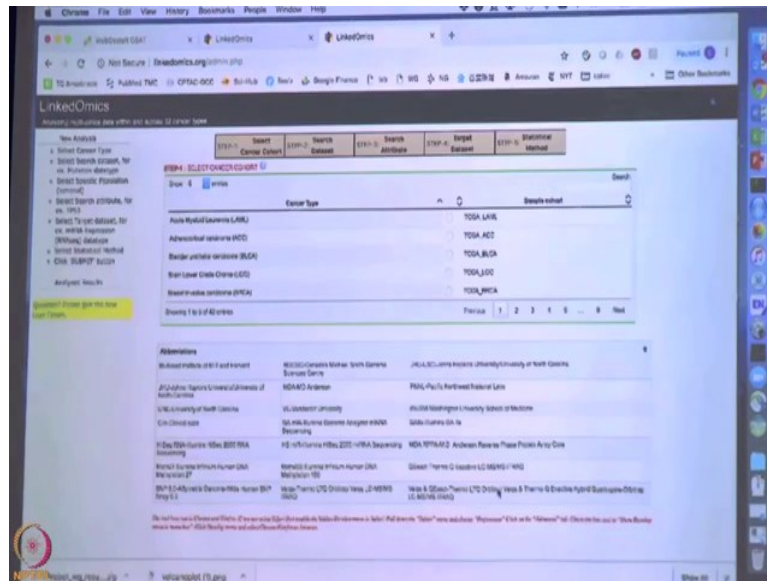
(Refer Slide Time: 11:47)



So, let us just go to the website. If you just Google linked omics and then you should be able to find this website. If you go to the website there are two options you can enter as a guest or you can register for an account. You can do any of those analysis without registration, it is free, the registration is also free, but the beauty of getting registered is that you can save your result in the database.

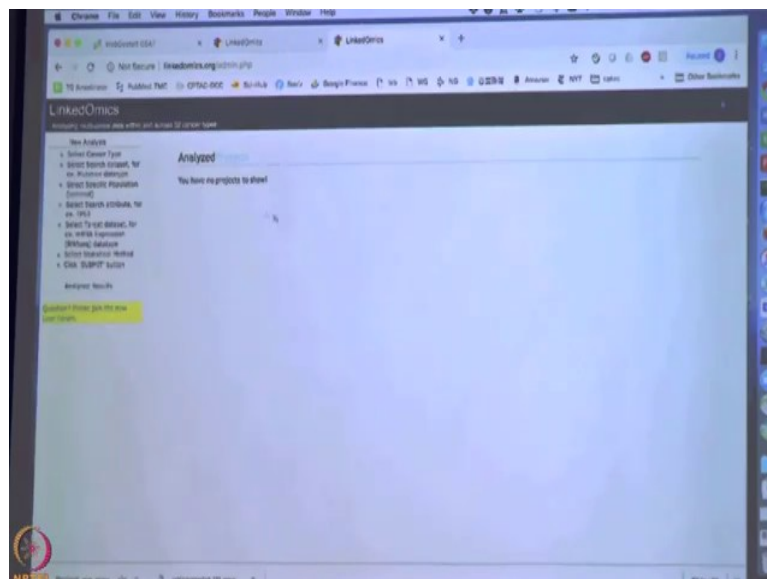
So, next time when you come back you do not have to repeat your analysis. For example, if you do analysis today you have account and the next time you are just logging and the all your results will be saved in the system. But if you just do as guest the results will only be good for today and the next time you come back you do have to repeat everything. So, it is up to you can either register now, or later if we want to make it easier today you can just enter as guest.

(Refer Slide Time: 12:58)



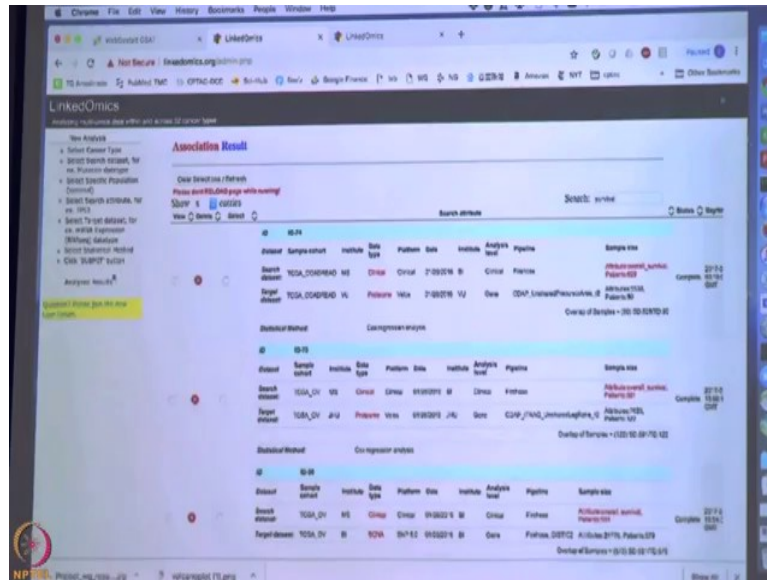
So, if you enter as guest or if you register and come here and at the back, so basically on the left it shows you and if you click on the new and I see a space create perform a new analysis and this shows you the multiple steps that you need to do in order to perform an analysis. And, if you can you can on the analyse the results, so, basically this will show you all the results you have generated so far right.

(Refer Slide Time: 13:24)



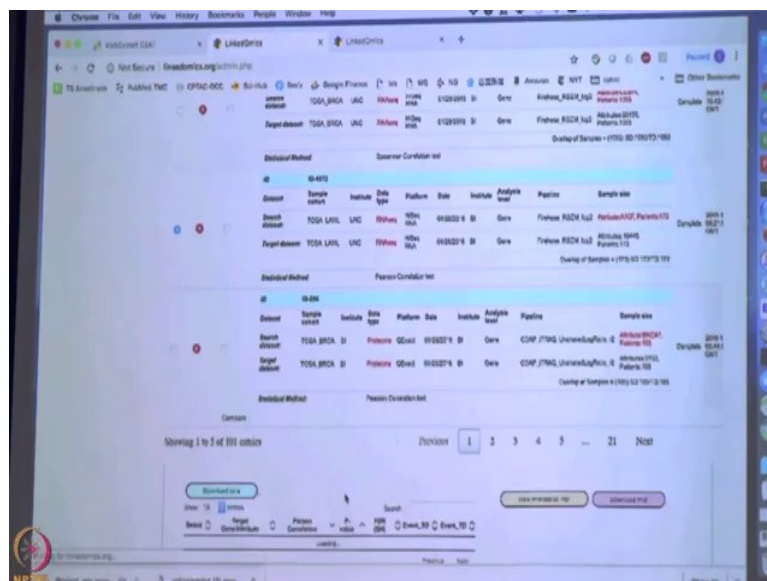
But of course, as a guest user or new user you do not have any results, but assuming next time if you would have a you are registered user you have some results. When you are log in next time you should be able to see some results here.

(Refer Slide Time: 13:45)



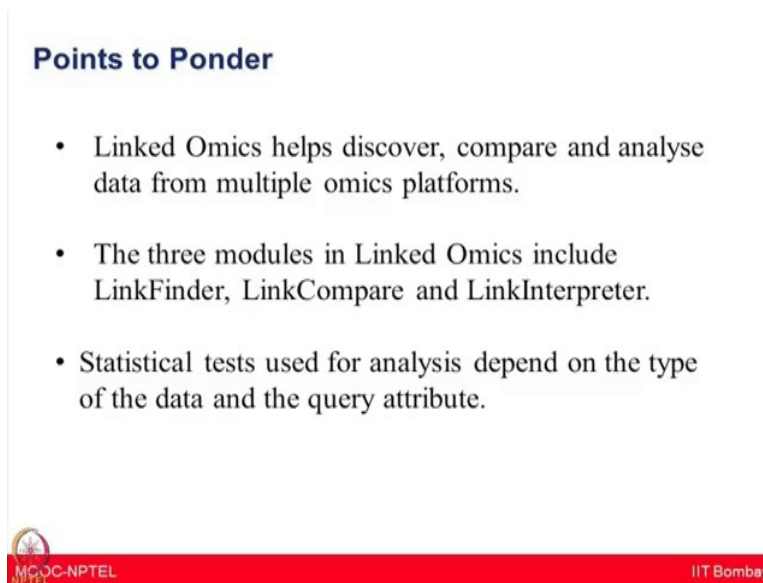
For example, this is my account and all my analysed results are saved here. So, I have a lot of analysis already performed.

(Refer Slide Time: 13:53)



So, I just the click on this and I can retrieve the result already; so, but for you guys you do not have the results.

(Refer Slide Time: 14:03)



Points to Ponder

- Linked Omics helps discover, compare and analyse data from multiple omics platforms.
- The three modules in Linked Omics include LinkFinder, LinkCompare and LinkInterpreter.
- Statistical tests used for analysis depend on the type of the data and the query attribute.

MOOC-NPTEL IIT Bombay

I hope today you have learnt that linked omics comprises of three modules, module one link finder which helps in comparing data from two attributes. For survival studies the cox model statistics can be used and it is very widely used in many publications. Module two consists of link compare.

This helps in comparing two or more data sets from the module link finder. Module 3 consist of link interpreter which makes use of web just start to interpret data from the modules 1 and module 2. In the next lecture, Dr. Bing Zhang we will continue the hands-on session on use of linked omics.

Thank you.