Introduction to Proteogenomics
Dr. Sanjeeva Srivastava
Dr. Karsten Krug
Department of Biosciences and Bioengineering
Broad Institute of MIT and Harvard
Indian Institute of Technology, Bombay

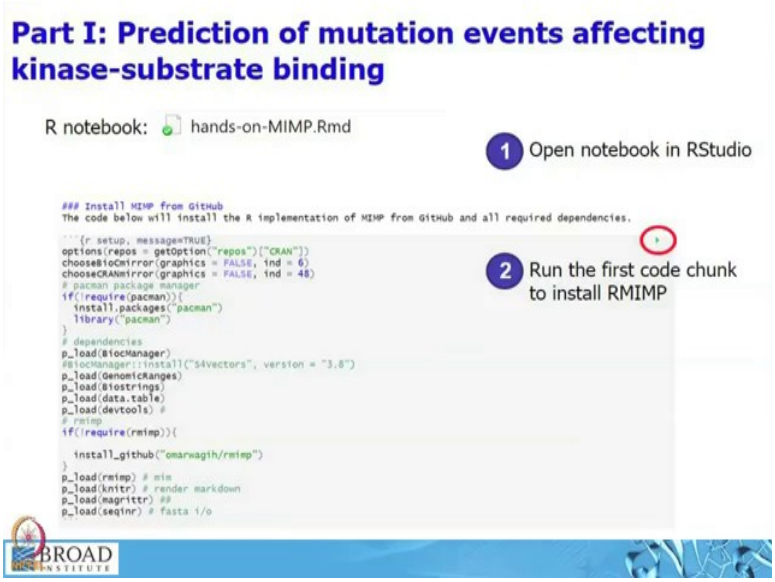
Lecture – 57
Sequence - GSEA

Welcome to MOOC course on Introduction to Proteogenomics. In last few lectures, we have heard about various ways to analyze pathways from the Dr. Karsten Krug. After understanding about how mutations effect phospho relation leading iterations in the signaling pathways.

Today, Dr. Karsten will talk about how one could use MIMP and GSEA in the hands on session. He will talk about how to use GitHub to obtain the basic codes and to use them without actually coding, but by manipulating codes as per your data an analysis requirement. The Dr. Krug will talk about use of two different formats for GCT files, GCT 1.2 and GCT 1.3 and conditions when one could make use of this format to provide better results.

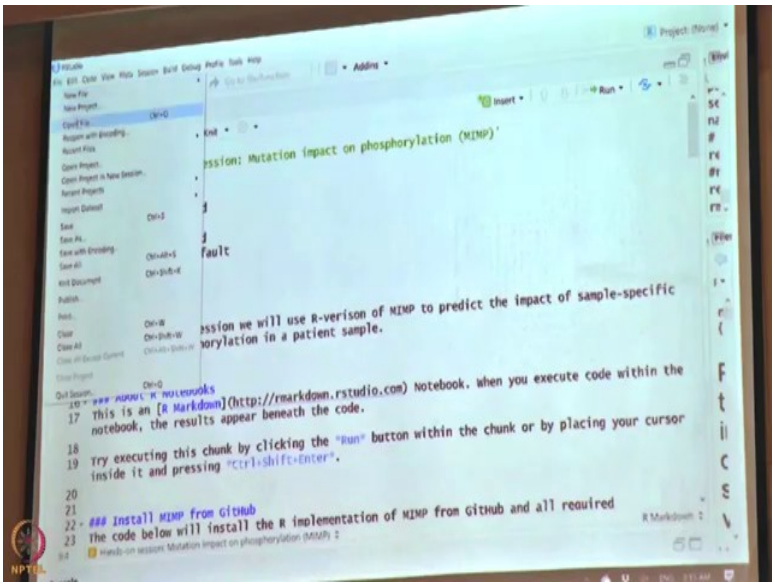
So, let us welcome Dr. Karsten Krug for his last lecture and learn more about usage of MIMP and GSEA.

(Refer Slide Time: 01:37)



So, there will be two parts. First part will that we will try to use to predict mutation events affecting kinase substrate binding. So, this we relates to the first part of my talk. And here we will try something very experimental. We will try to use R actually I already prepared on a notebook, far there will you know talk about more over that is now notebook and so on and so forth. But, you will find this kind of files here hands on MIMP dot Rmd which stands for R markdown in this zip file that you hopefully all downloaded.

(Refer Slide Time: 02:17)



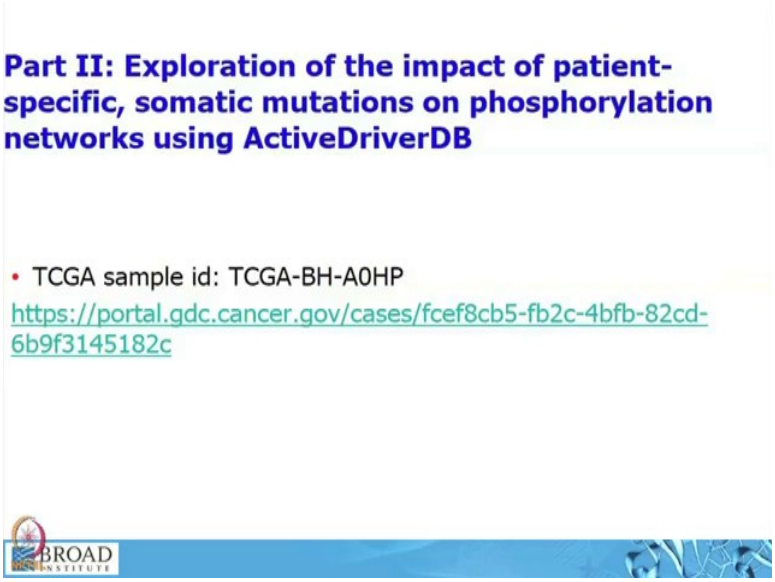
Just open R studio.

Student: yes sir we did it

Then you just load file go to file, open file.

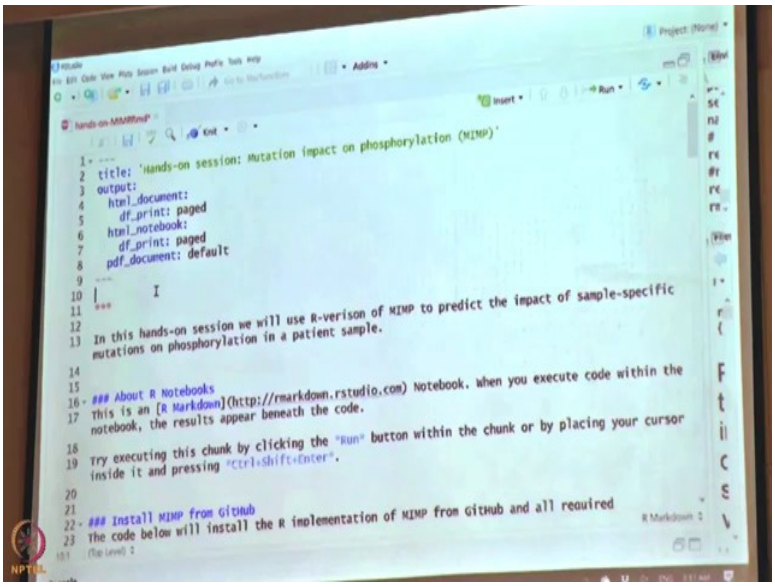
Then open R markdown, file which is called hands on slash MIMP dot Rmd.

(Refer Slide Time: 02:33)



So, you were looking at this R is so called notebook. It is R code inter mingled with you know just text. Who of you have heard about markdown in general? Ok. So, that is ok, that is great.

(Refer Slide Time: 02:42)

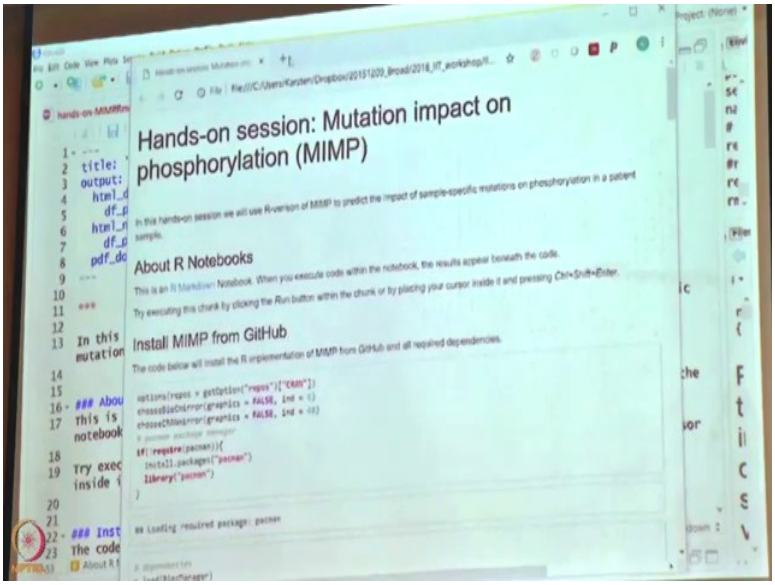


Student: It has gone up.

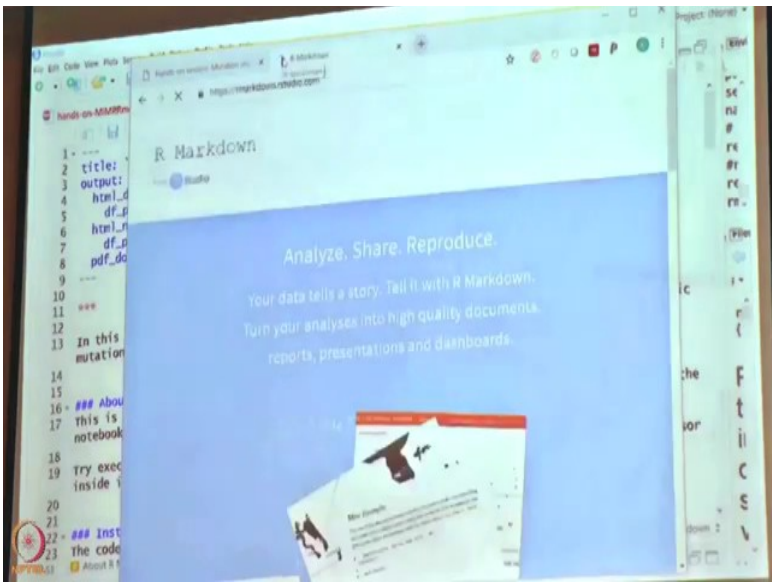
Markdown; it is a very simple text based format to create you know structured documents, so on like HTML page does not stuff like that you know. And there is extensions of that allow

you to execute code in these map, in these documents. Actually, there is a link here about our notebooks, and if you click on this link here you get some more information here.

(Refer Slide Time: 03:38)



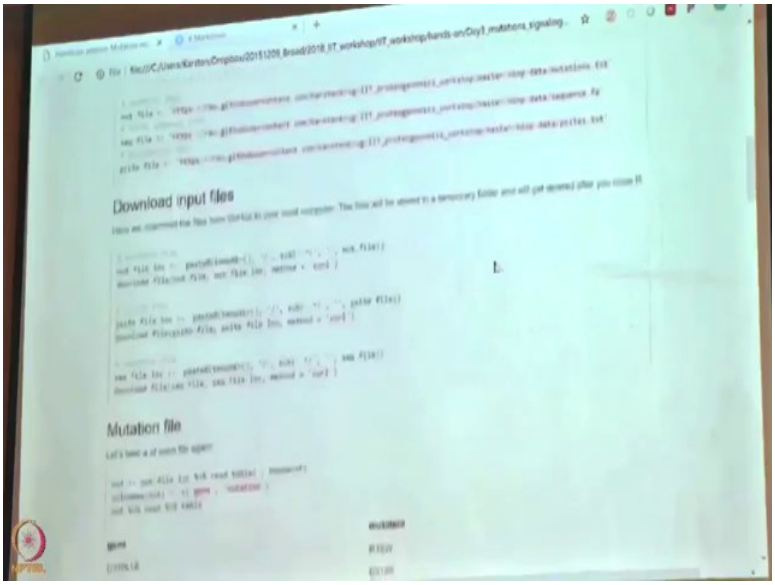
(Refer Slide Time: 03:42)



And it is a very convenient way to analyze, to document you analysis results and your code and also to share these with your collaborators. May be I quickly show you the result of this whole analysis. At the end of the day you will have an HTML document. It looks like that which you just open in your browser and there you have some documentation about like two

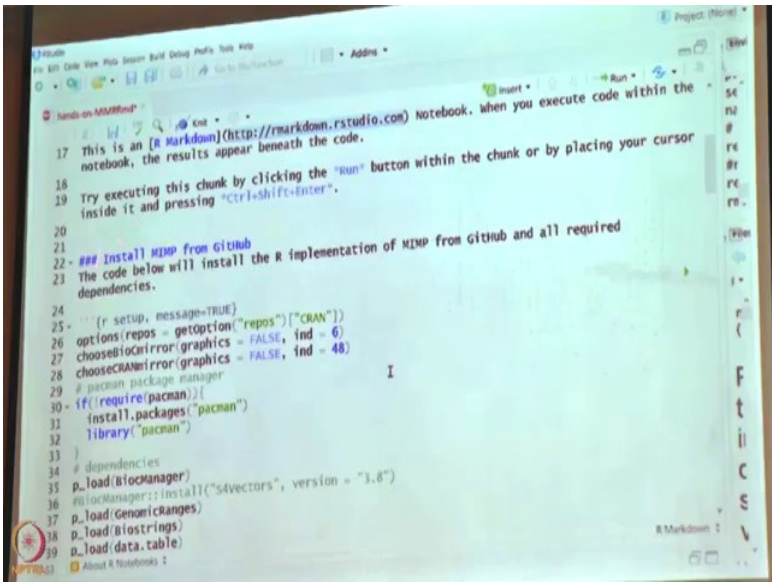
different steps. So, this is text that you entered, so you can describe about both kinds of goal of my analysis, what is their input and output.

(Refer Slide Time: 04:26)



But, you will also have all kind of you know R code that has been executed in order to get through that analysis results. This you can easily share with your collaborators and you can just re run everything in order to get through these results.

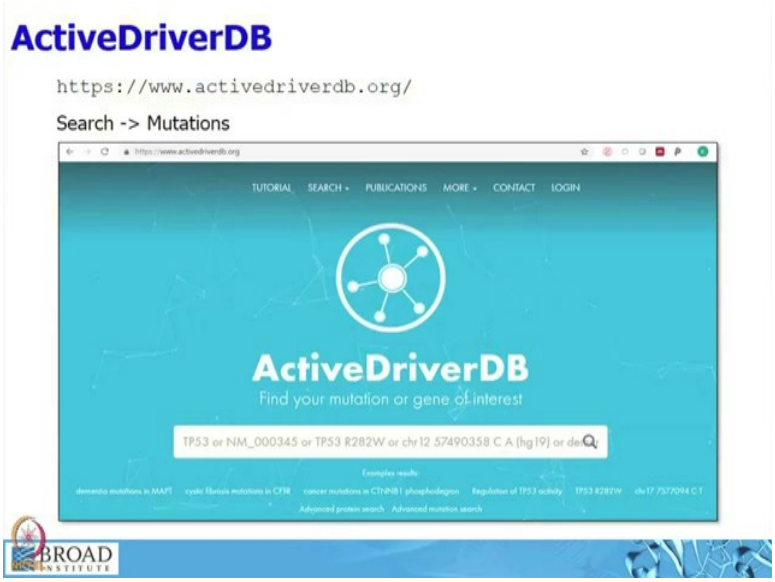
(Refer Slide Time: 04:50)



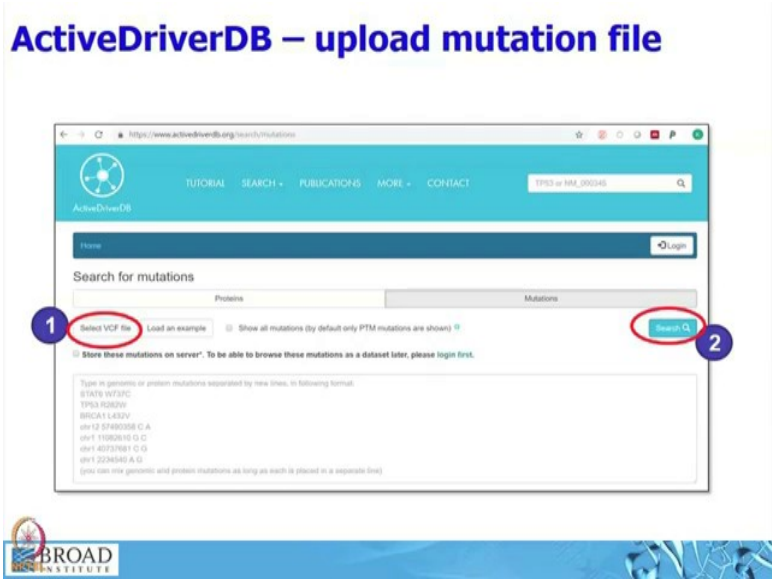
So, but, right now we just focus on that one. So, in as we see this code on a place to see you know this block here starts with R. So, this is actually R code here in this kind of block. And

you can execute this code by just clicking on this little triangle here. Please try to do that. So, this might take, depending on Wi-Fi connections might take some time because the first part actually sets up the entire analysis, so it downloads a couple of packages again and so on and so forth. So, if you could please try to click on this little triangle here that is a first, it is a you know in line 25 that is probably easier, so in line 25 if you click on the on the triangle.

(Refer Slide Time: 05:57)



(Refer Slide Time: 06:05)



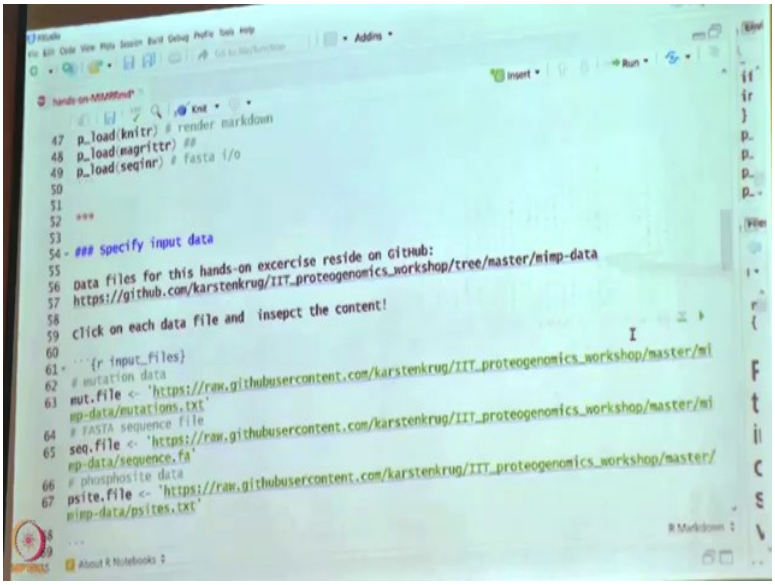
Student: 9 10, I just want to very quick demonstration of that second part as it might not work as website is down. So there we wanted to use main mutation for one particular TCGA sample

Then you can search Mutations, then you can upload your vcf files. Here you can upload and search your file

can you please explain how to make the VCF file that

I mean vcf file is called variant calling format. depending on your pipeline you have used to call your variants to look at vcf files , that is pretty standard data format for variant calling. So, it is not at all that you have to create a VCF file, its more like that you should have got it from somewhere. If you. If you ship your samples for sequencing post sequencing you get a BAM file back they may ask for genome sequence. Then typically you would have some pipeline one in your lab or in your collaborative lab have genomics pipeline. They will take this Bam file and do permutation on it the result you get is the vcf file for example. I mean it is very, it is not very complicated all that I have, right.

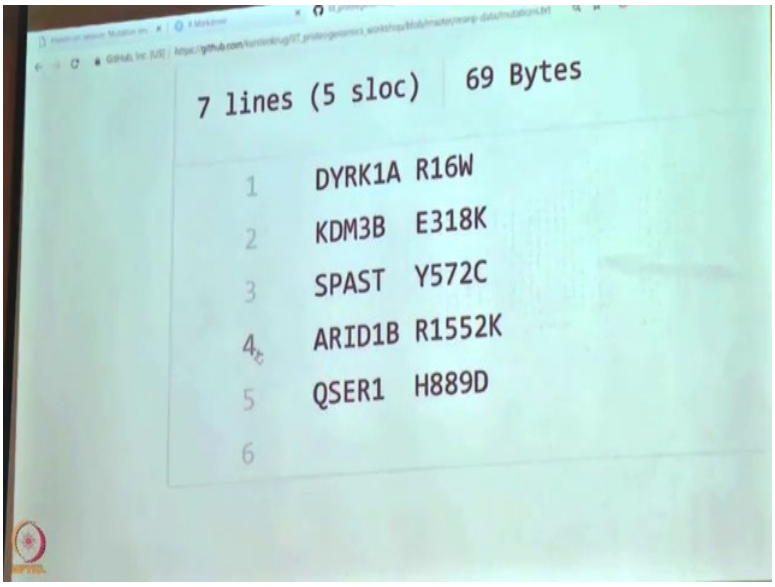
(Refer Slide Time: 07:13)



But to do here, so first we install the packages you need to specify the data, right here I have chosen the mutation file and a phosphosite format and its corresponding data base file format TCGA patients again, in order to make more convenient for you I Programmed everything on Github So, its already click any where to get your data from Github. So you don’t have to go

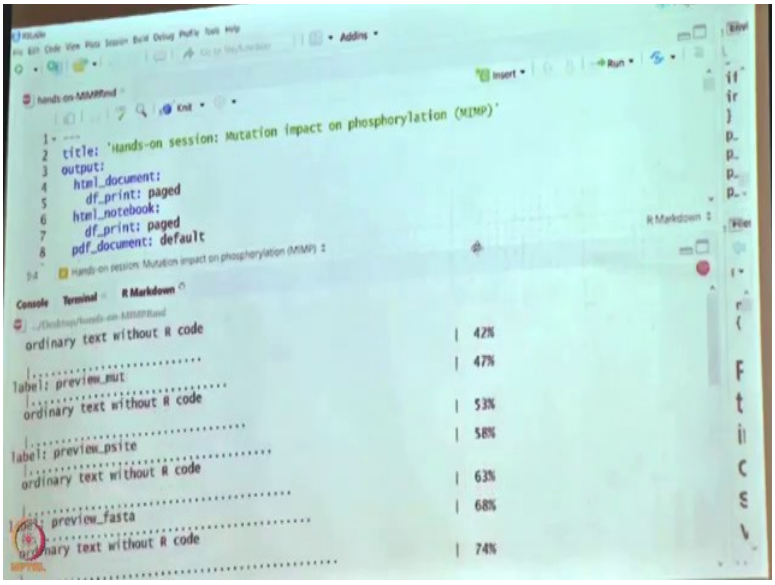
anywhere to download. So, you can also just you know copy this link and if you go there and you will see these three files. So one is called Mutations. Txt, the other is psite.txt and another is sequence FASTA file So again, a very simple format

(Refer Slide Time: 07:58)

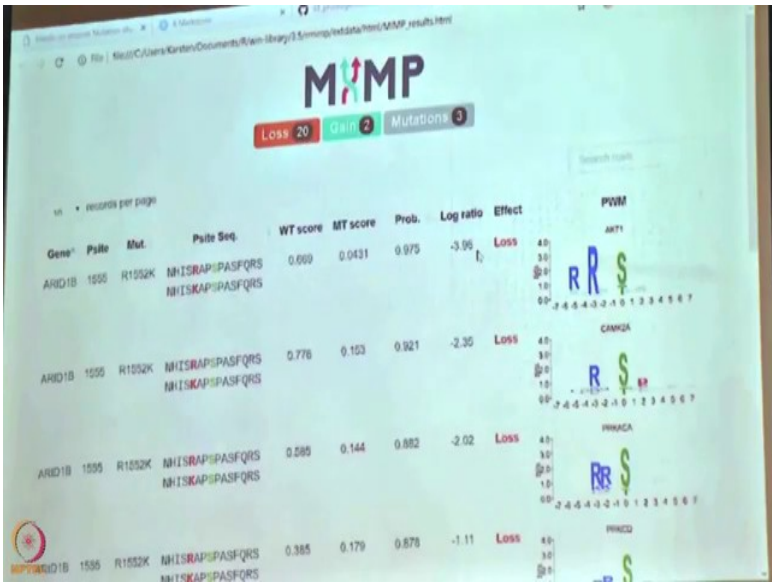


So everybody can see that? Probably not. So this is very specific to the file, right. So, this is now; you know the mutation file is you know is divided into 2 columns. the first column, is the gene name and the second column you can see amino acids and positions of the mutation caused.

(Refer Slide Time: 09:45)



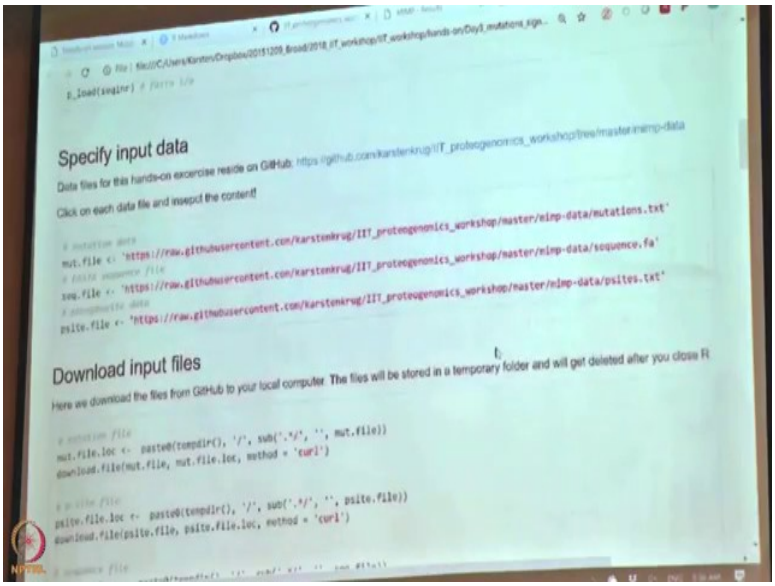
(Refer Slide Time: 09:52)



Let us see how these different sites have different results. So, actually, before we go on to these are this is results of the fragments. So you can press on the growth and look at the HTML file in it and if successful then this html document should be saved in the same folder. So in the same file you should have an HTML file to double check that.

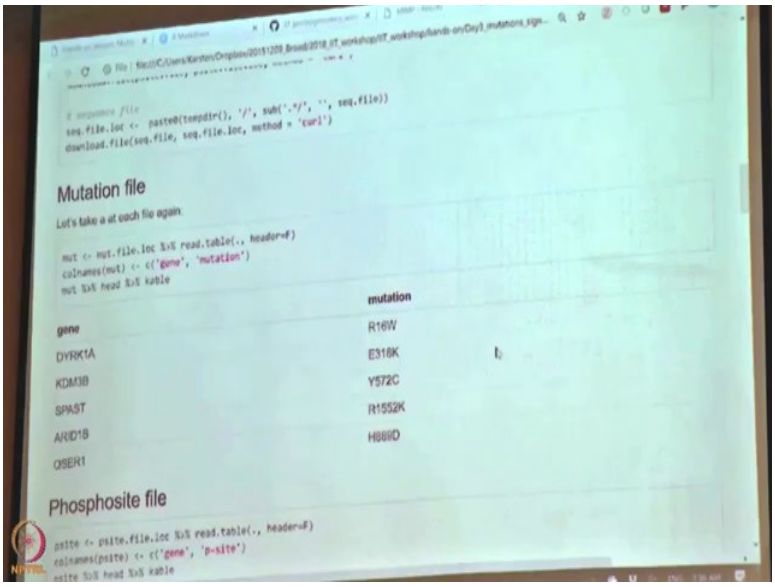
And here we have the entire report as HTML which includes you know some description about a project and also documents all different steps that we have done here. So, first we install all of the packages. We also see the output that was generated by the different code chunks here.

(Refer Slide Time: 10:50)



Here, now we have the direct link to the GitHub. If you just click on that you end up on my GitHub page and some of the files. So, meaning we specify the input data, what the script then does it downloads the input files, right.

(Refer Slide Time: 11:08)

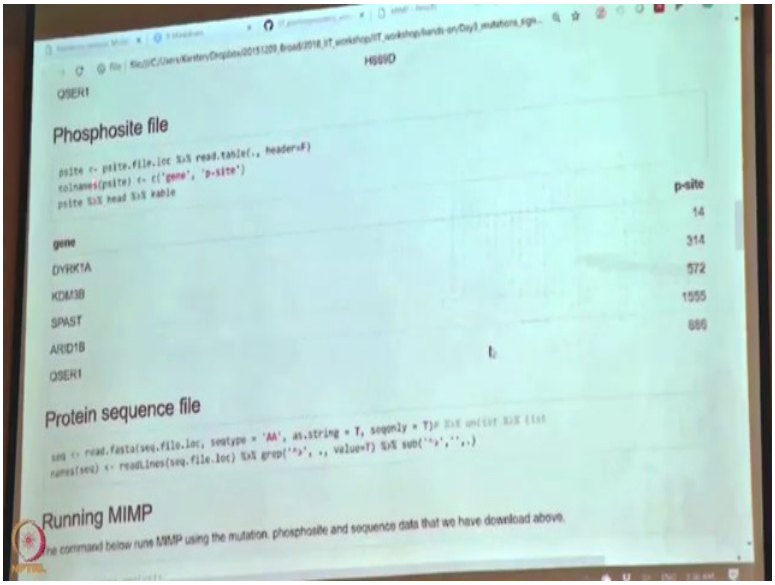


And then what the script does that, it imports all of these files and here just shows the various couple of entries, you all have that in your HTML file here. So, this is the output that creates that table. Does that make sense?

Student: Yeah.

So, again are we are looking at mutation file, we have two counts, gene mutation, gene name mutation.

(Refer Slide Time: 11:40)

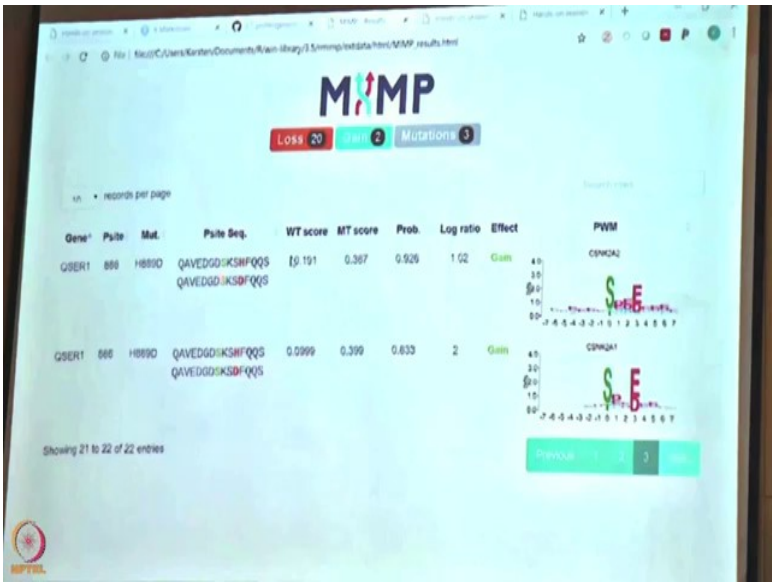


Phosphosite file is similar again. So, this is the code that generates this table. And, what happens then is it will open up this result page. So, this is what just happened when I ran MIMP. This is also one example that I was showing in a like earlier in this morning, whereas result page you see there is a 3-mutations that effect in total like 22 possible phosphorylation events.

We see that most of them are losses. So, meaning motif got lost, so the phospho like the phosphosite is most likely not phosphorylate involved by the kinase that use to be. But we also observed two gains. So, meaning now we will certainly see like an increase of our potential, predicted increase of that phosphsite to be occupied.

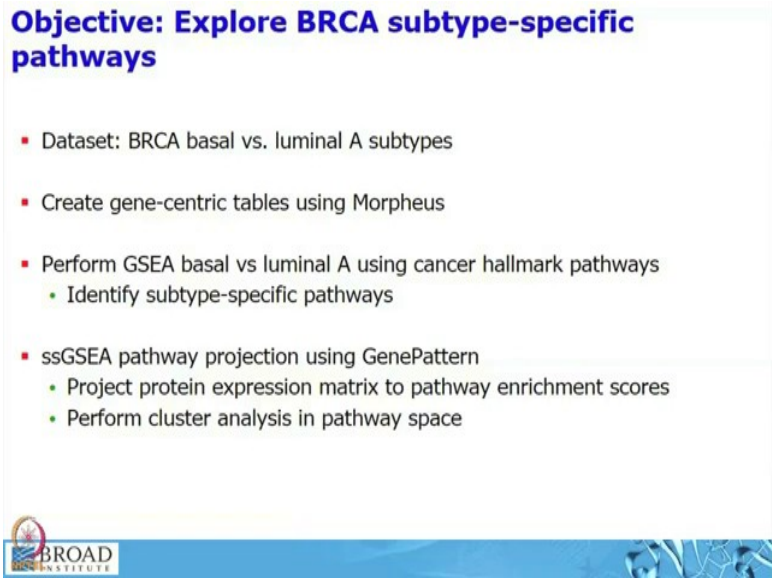
Student: Yes.

(Refer Slide Time: 12:55)



If I just go through the last page, so these are the two events where we protected the gain and here I just was showing or show you two examples of a phosphorylation gains. So, basically that is the wild type, text the mutated version. So, you see this aspartic acid here which is now we have recognized, why like a noble kinase, right. So, this motif fits to that particular kinase. So, we predict or we assume that this phosphosite is more like to be phosphorylated now. So, that is type of data, right that we get out of here.

(Refer Slide Time: 13:44)



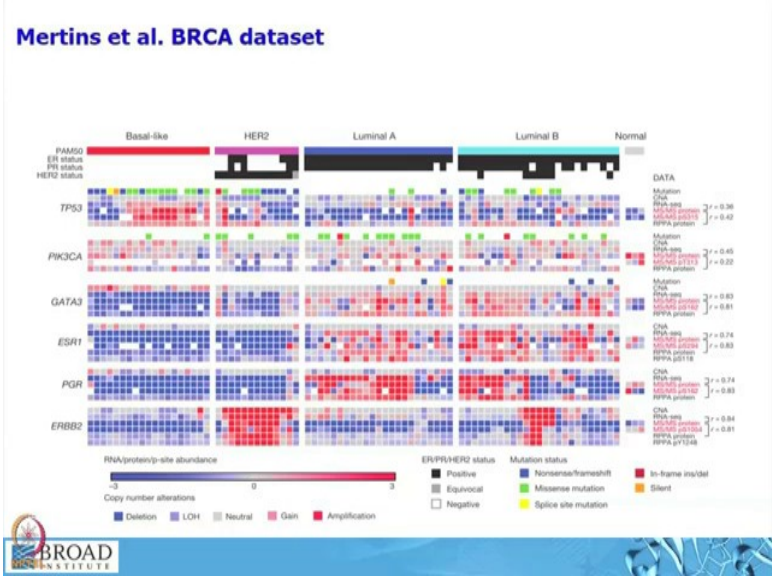
The objectives, what the goal of this hands on session is to explore first cancer subtype specific pathways. And we are going to use the first cancer data set that we have used, but only looking at two sub type just to make it you know that easier. I am just going to look at basal and luminal A and do the tables that I have already created only contain these two subtypes, right.

So, we are going to do. So, this first exercise is optional. So, this would involves Morpheus, I am not sure how well these works here again. So, we can skip that I can just demonstrate how we use Morpheus, so how you could use Morpheus to create or to convert your gene protein centric tables, so your protein centric tables into gene centric tables. So, this is always what you have to do, if you want to do pathway analysis. So, again this is optional. So, you have already the correct tables to move on.

Now, we will do two different ways of pathway of GSEA analysis, one is like the classical so to say GSEA using this Java application that I think everybody was able to download. Most of you; in the second approach we are going to use same data set and we go use single stranded GSEA to project our protein matrix into pathways.

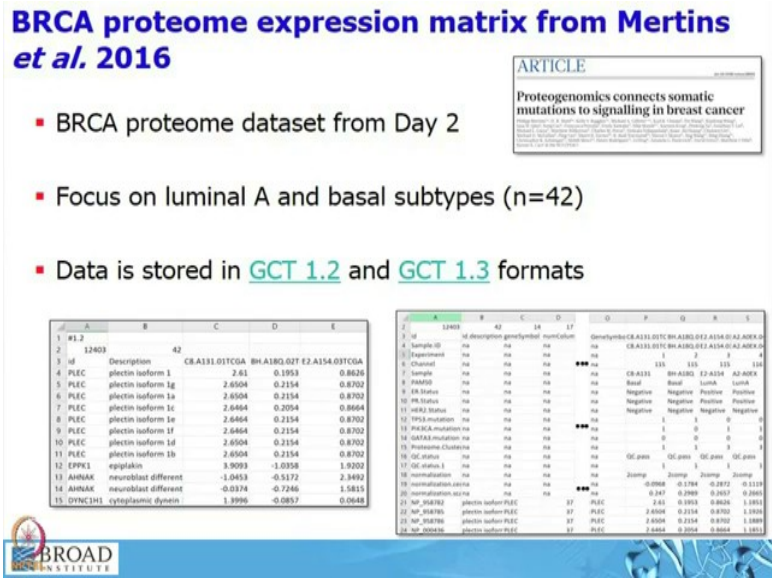
And then I actually plan to use Morpheus again to perform some cluster analysis on marker selection on the pathways. So, again I cannot guarantee that this is going to work with internet connectivity part. We will at least try to do that. And again, so I try to make the slides as compared to the pathways. So, you should be able to do theory now to just go home, you have a data, you go through the slides, you can repeat these exercises on your own. This is quick recap.

(Refer Slide Time: 15:57)



So, that is a dataset we you know heard about the dataset couple of times already, and now we only wanted to look at the basal versus luminal A. And just by eyeball in protein space, so like in proteogenomics space you clearly see differences. Now, we interested in, so what are these pathways that are differentiating basal and luminal. And in this case only we are looking at the cancer hallmark pathways. So, it is a very small compatible and annotated curated pathway database.

(Refer Slide Time: 16:34)

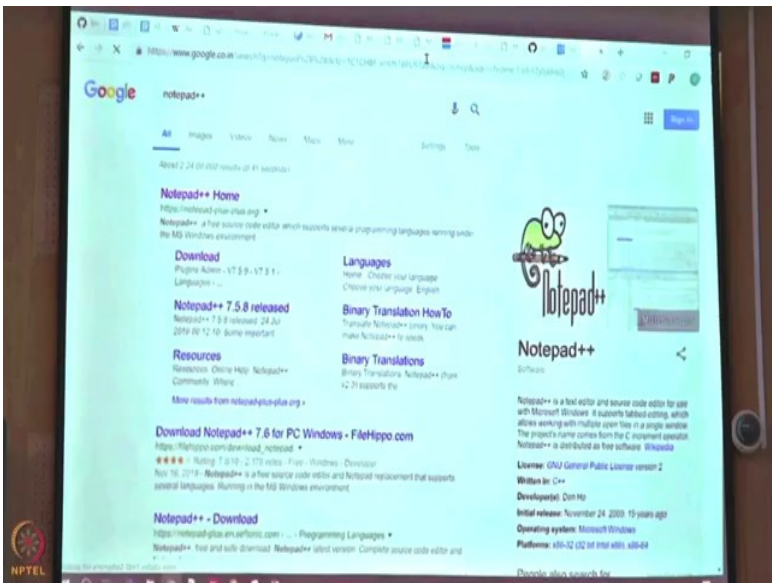


So, especially just for what I sat here. So, we have two different data formats and I got a lot of questions about GCT, and how to create GCT files and how to open GCT files and you know these files are simple text files. You can open them, these files in any text editor and I would highly recommend to install what you use a text editor in your on your PC.

Student: Which text editor.

I would highly recommend notepad plus plus.

(Refer Slide Time: 17:16)



This, is what I use on windows systems and you know if you just Google notebook, notepad plus plus, you know would little web page, you just download that. That is like a general recommendation from my side. So, in this hands on we are going to use two different versions of GCT, one is called 1.2 that is the like older version of GCT which you know has been long for 10 years now I suppose.

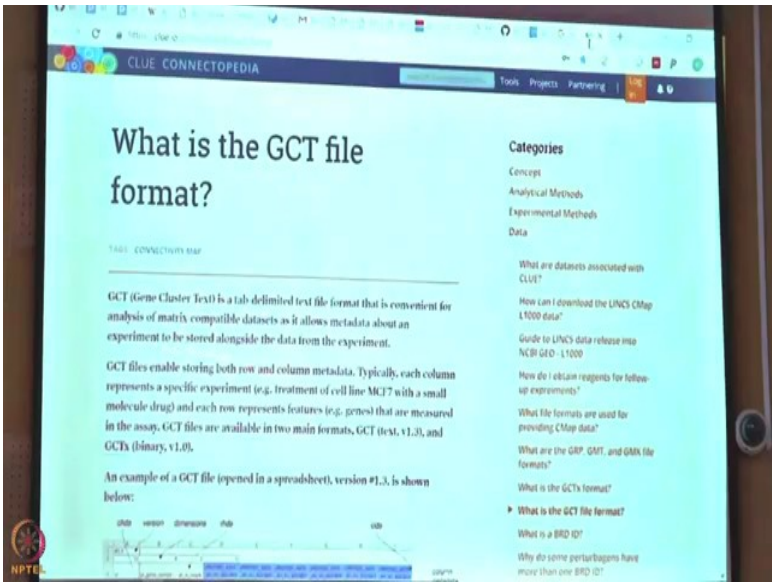
And since the couple of years you know the they we revised the format and came up with a new one which is called 1.3. And the only difference is that in 1.2, we only store the data, right, so we just store the data and you know we have two annotation columns of basically, two annotation columns to describe the data.

So, that is the somebody made that like you know hard code, so you cannot change that. You always have two annotation columns and then we have the samples and the data. GCT 1.3, you can store metadata which describes your experiment you know for example. So, for

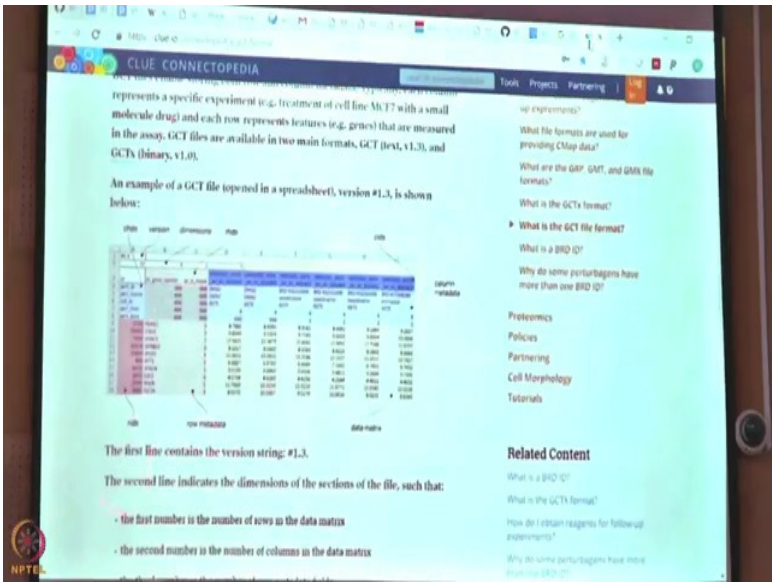
example, so this is like a snapshot of these database that we are going to use and here is the data in this corner and on top of the data we had all meta data this that describes the samples. So, this is one sample you have to GCT id and then we have all kinds of information so, which you know the details, you know has been used to quantify that which subtype it is you know go through ER 2, PR 2, PR status and so on and so forth.

So, the advantage of this format, although it might not be very intuitive in beginning, but if you are used to that its very convenient because you store all of your meta data together with your data, right. You do not have to look through your computer and you know find meta data that actually annotates your data matrix.

(Refer Slide Time: 19:27)

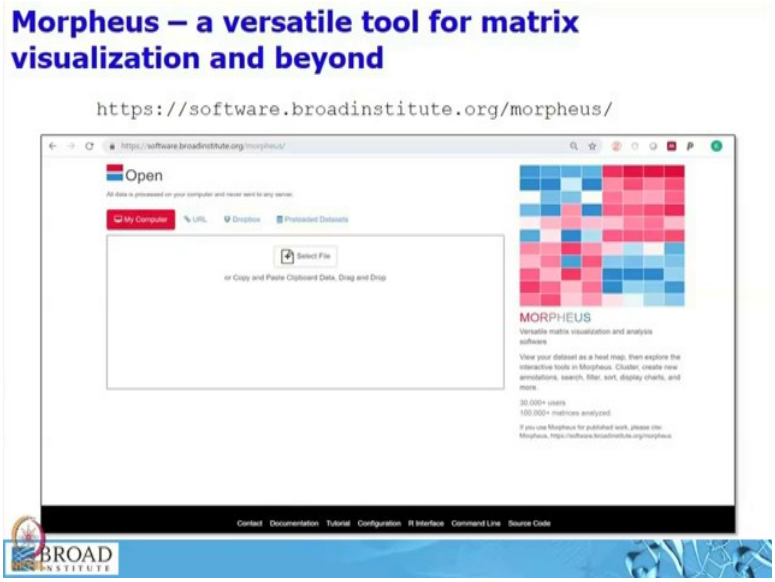


(Refer Slide Time: 19:33)



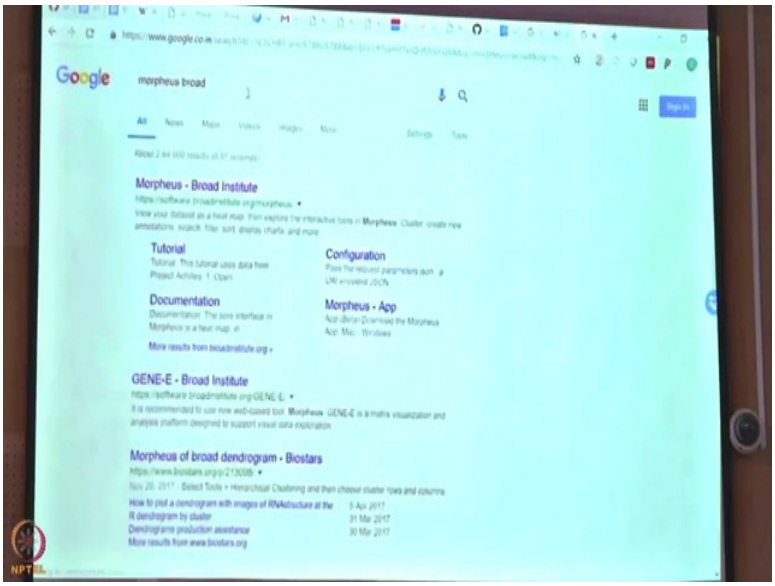
What is the GCT file format? That is what you know want to know, I can how this file format is organized here. I mean here if you spend you know 10 minute or so, you would better understand what GCT means and how to create one. We have to use both versions because the Java application does only support GCT 1.2, ok.

(Refer Slide Time: 19:48)



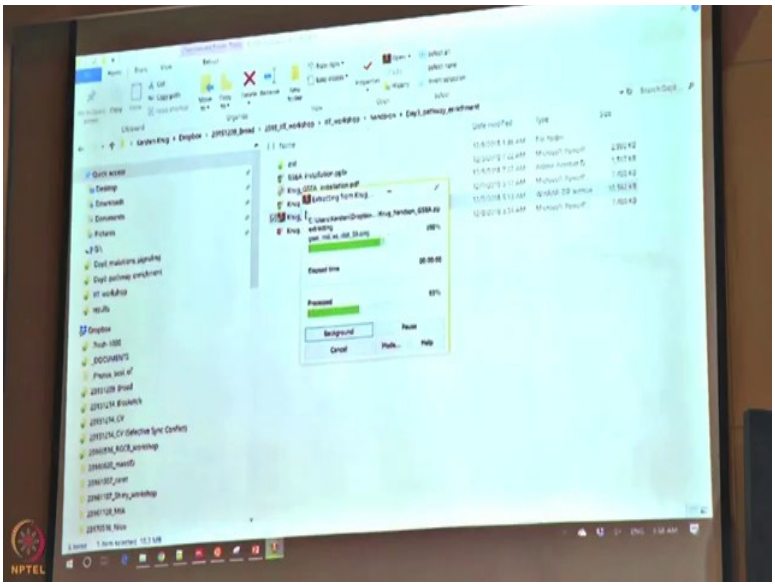
So, this type is now optional. Let us try to make it work; let us try to grow to Morpheus.

(Refer Slide Time: 19:53)



And here there are different ways to you know import the data, you can parse your computer, if you have it in your dropbox, so you can provide an URL or you can choose simply drag and drop it through this window here, right.

(Refer Slide Time: 20:21)



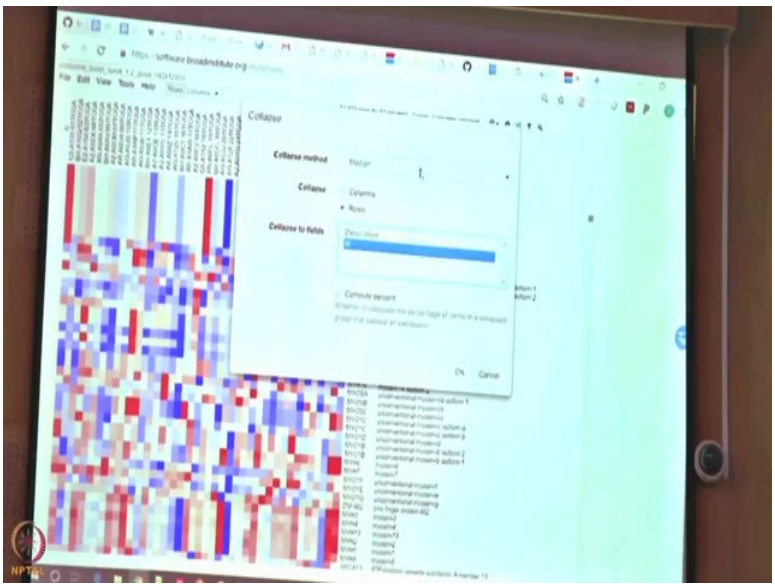
So, now what we going to do; we go to so, that is a zip file I am going to quickly extract that here. So, that is the one that you download it. So, now if you now go, if I now go into this file you see two folders GSEA and single stranded GSEA, you are going to focus on the GSEA 1, right. So, everybody with me.

Student: Yes sir.

And here you have two GCT files; so, one and proteome basal luminal A 1.2 and the other one size proteome genes. So, this is gene centric, this is protein centric. So, in case we are not able to use Morpheus we already have the gene centric matrix, that is what I am going to this what I am want to say it here. But right now we try to just as an exercise which is I just want to show here how would you do that and for that you just drag and drop these file the one without genes into Morpheus. So, drag and drop means you do this. Does that work for everyone?

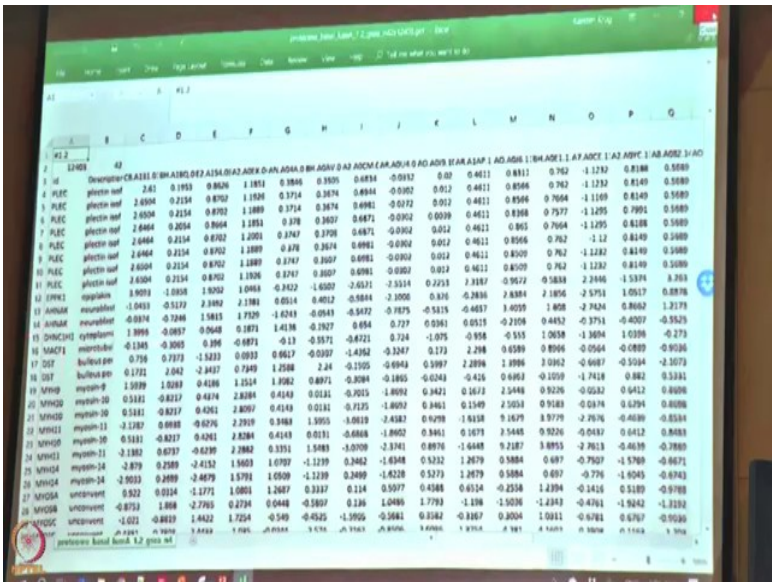
Student: Yes.

(Refer Slide Time: 21:36)



And here you already see you know that you have genes that appear multiple times here, right. So, these are the different isoforms and you know these types basically tell you that we cannot really resolve these isoforms where very similar expression patterns because they have many peptides that are share between those. And for pathway analysis we need to have the single row for each gene. Does that make sense?

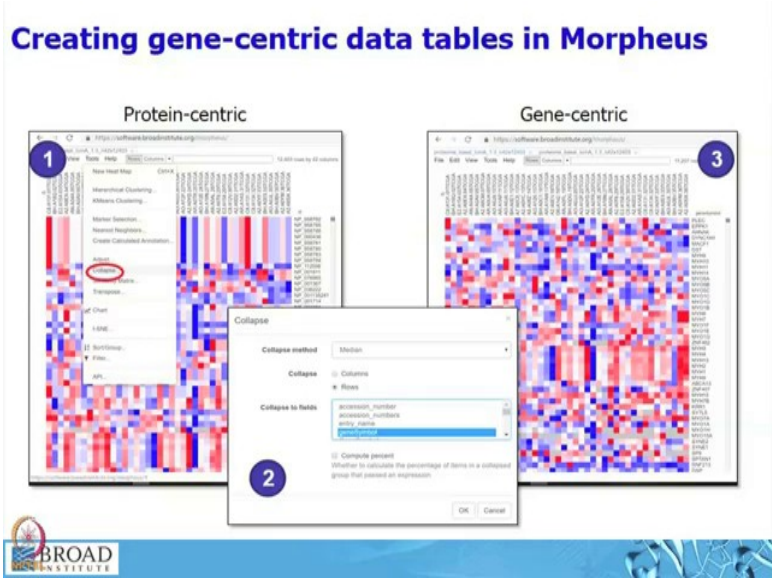
(Refer Slide Time: 22:08)



If you would open it in excel this is what you would get that, right. Just the excel it is a same file, you see the number of genes here or like proteomes and number sample columns like 42, then you have the gene ID, you have some description and then you have data matrix that is 1.2 format, easy as that. That is the same file just we look at this in Morpheus.

You can easily I mean you know that I am going to too much data, but you can easily I think at the later slides I show how to change their notations. For example, you could go and say, ok I also want to look at the description, right. So, if this is highly customize to go, and may be again you have to spend some like you know couple of minutes, just play along with your own data, but everything is possible here. So, that would we want to do, we want to do; we wanted to create gene centric tables.

(Refer Slide Time: 23:11)

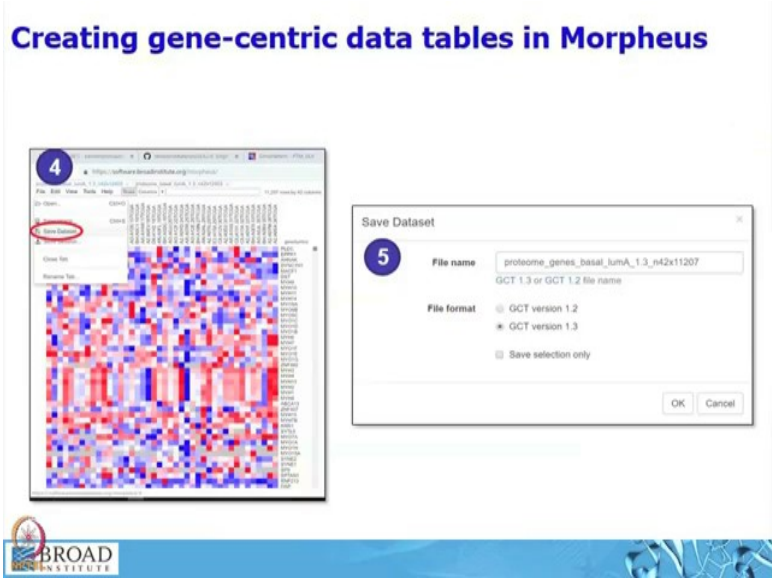


So, you click on tools that is the first step, then you click on collapse, then you should see this window popping up here. I going to do the same and parallel here. Tools collapse. So, then you have to pick the field that you want to use to collapse in your case its the ID columns. So, this is the first column here shown which contains the gene IDs, right.

And you can also specify whether you want to collapse rows or columns truly depends, I to, we want to combine or collapse different rows. And here you can choose how you want to collapse them, median or mean. Again, it is very that is no clear answer what you know what would be best. So, median is usually more robust against standing like outliers. So, we are going to do that.

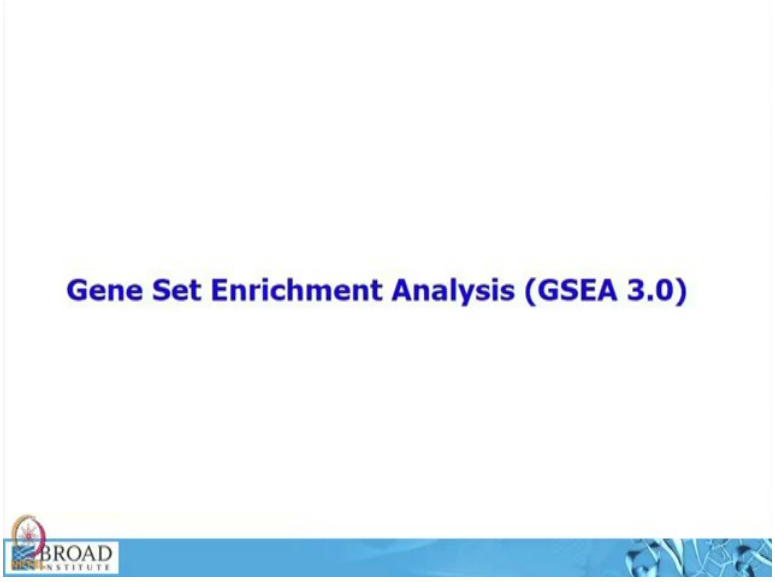
Then I just click, ok. Then you will see that does not, you will get a new data tab here, now you can also go back and forth, right. If I click here that is it protein centric matrix, if you click here that is a gene centric matrix. So, you see that each gene symbol now is listed only once. So, it is you know very convenient to create these tables. And if you want to download that without table, you can just do so, by clicking on file save dataset you can you know pick a GCT version, and you can give the name and so on and so forth and you can just click, ok. So, now here ready you do GSEA that is what we are going to do now, ok.

(Refer Slide Time: 25:13)

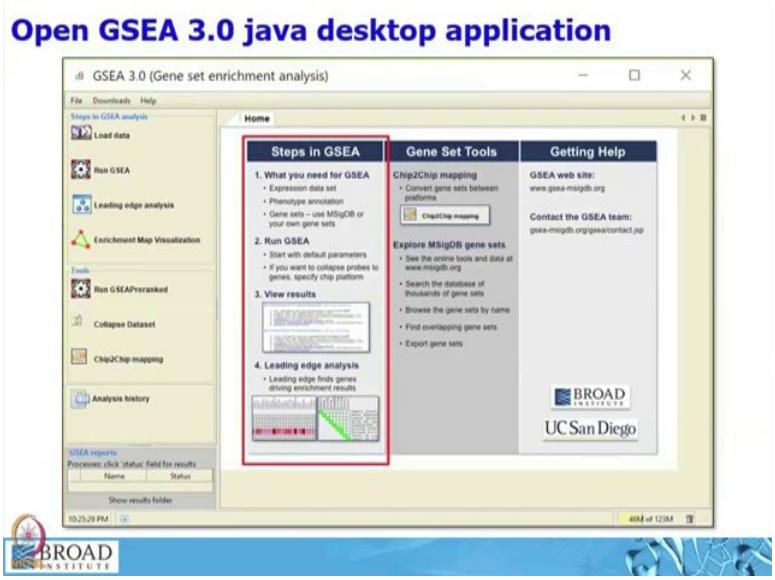


So, now, I want you and again. So, here step by step manual how to do all of these, right. So, you should be able to do that at home.

(Refer Slide Time: 25:22)

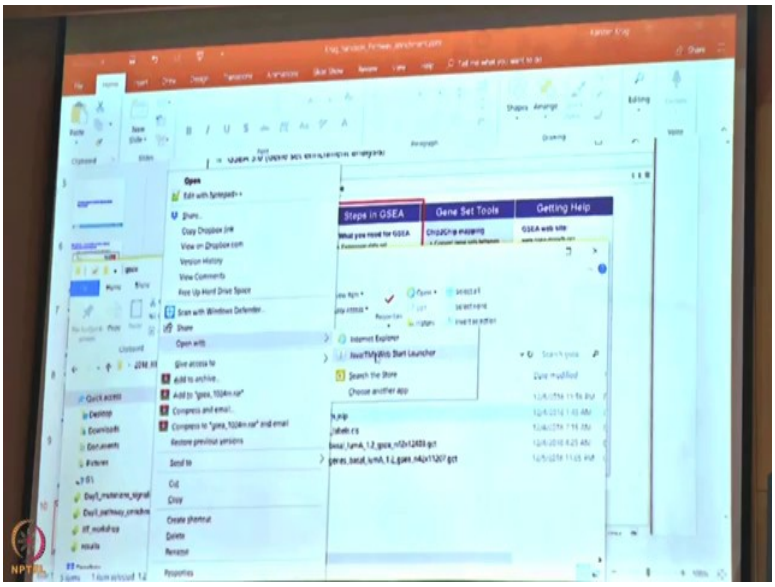


(Refer Slide Time: 25:27)



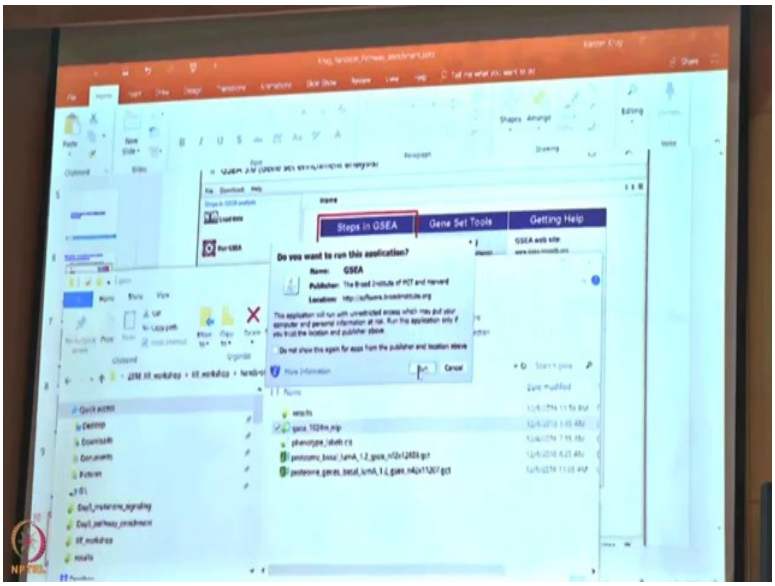
So, now its GSEA time. And I saw, a large fraction of you guys, you got it run. So, please try to open GSEA, the Java application. So, you should be able to see this kind of screen shot here, like this kind of window. Now, try to do same my PC, you see.

(Refer Slide Time: 25:47)



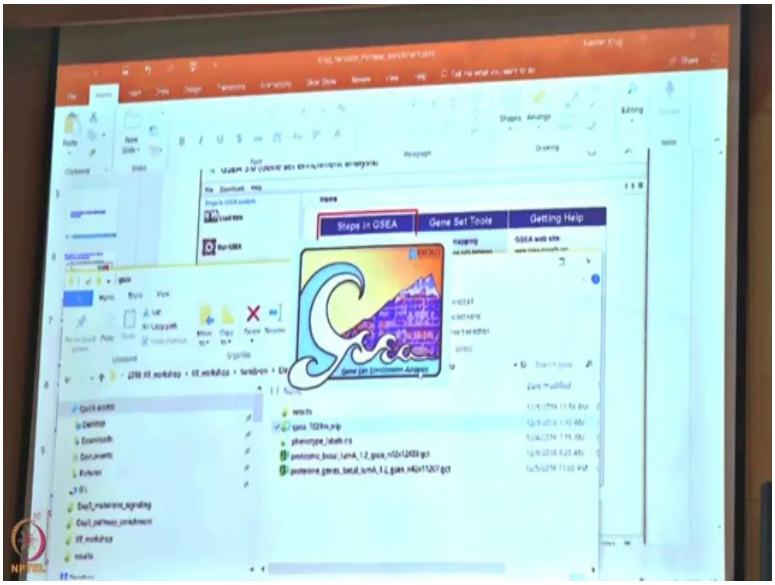
So, if this jnlp file is not automatically associated with Java like here on my PC you can just right click on it, then you should be able to see Java web start launcher, launcher. So, then you should be able to open the app. Now, once this is finished you should be able to see the GSEA window.

(Refer Slide Time: 26:28)

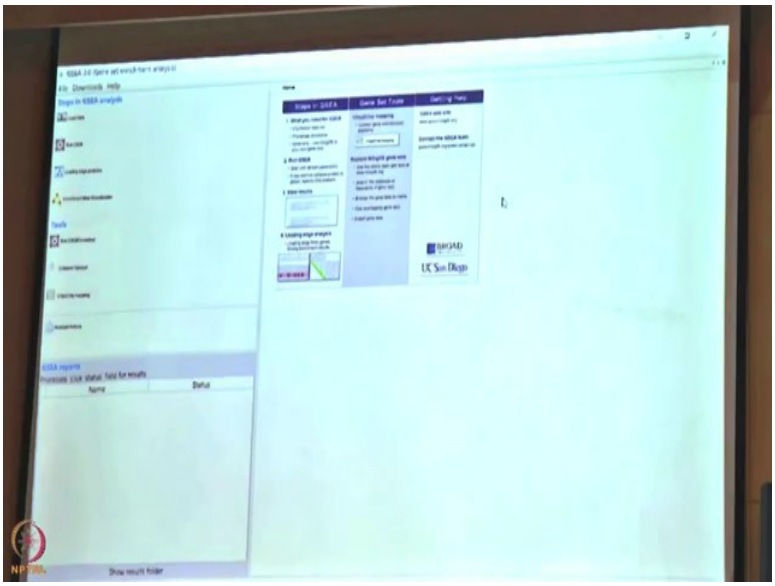


So, now it asks me to whether I want this application and I just say yes one, right.

(Refer Slide Time: 26:36)



(Refer Slide Time: 26:40)

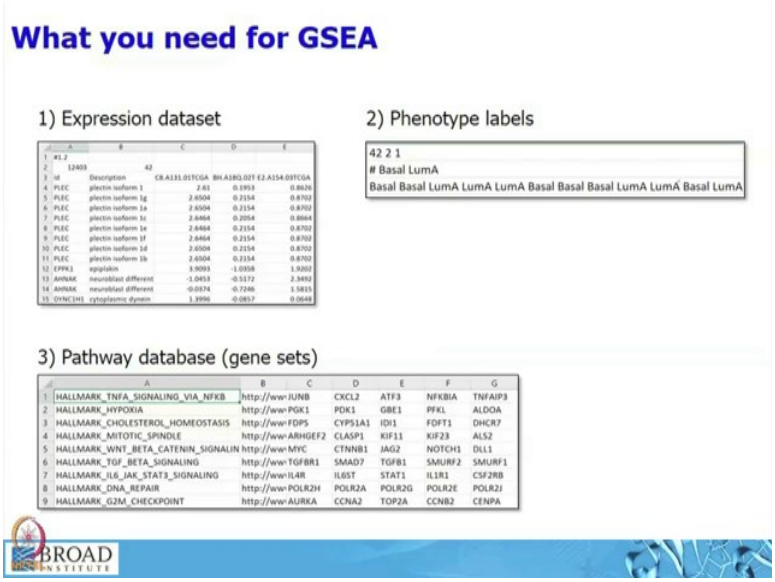


So, this little bit smaller my screen here. So, because of my resolution, but Java also or GSEA also comes with a very extensive documentation and also liked a entire user interface is know if you if you pay attention is very intuitive, because where here on this start page where your data actually you know describe all the different steps that you do, steps in GSEA. So, this is exactly what we are going to do now.

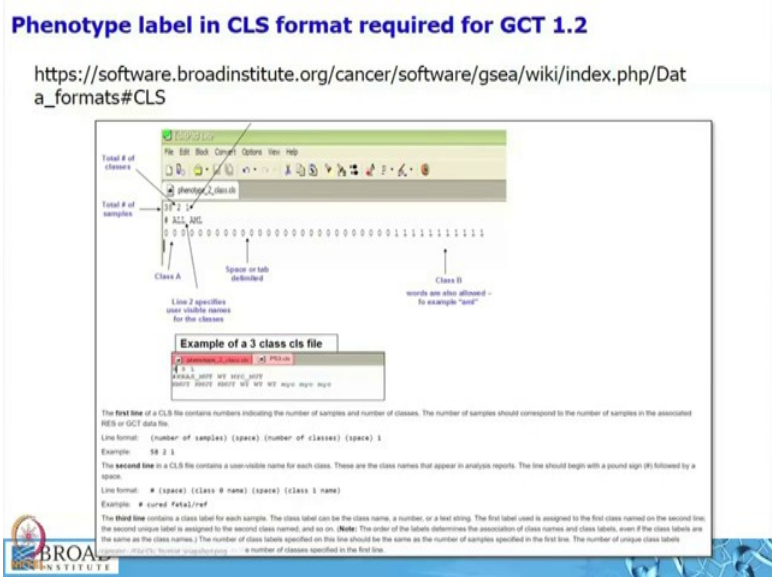
So what we need for GSEA? We need expression data. So, this is how a GCT 1.25 that we have just created. We need phenotype annotation, so because we do not have the metadata about our samples in our GCT file because it is a 1.2, we need an extra file and I have to tell the software what is luminous samples, what are basal samples, right.

And that is why GCT is so convenient because we you do not have to worry about any other files, right everything is in your file. But, if know to make this work we have to create a phenotype label file and I am going to show you how you do that. And we have to this take a gene set database, and again so here you can upload your own gene cells, you can download different databases which you can upload here where it also you know directly links to the MSigDB page. So, you are sure you always get an the latest version.

(Refer Slide Time: 28:09)



(Refer Slide Time: 28:15)

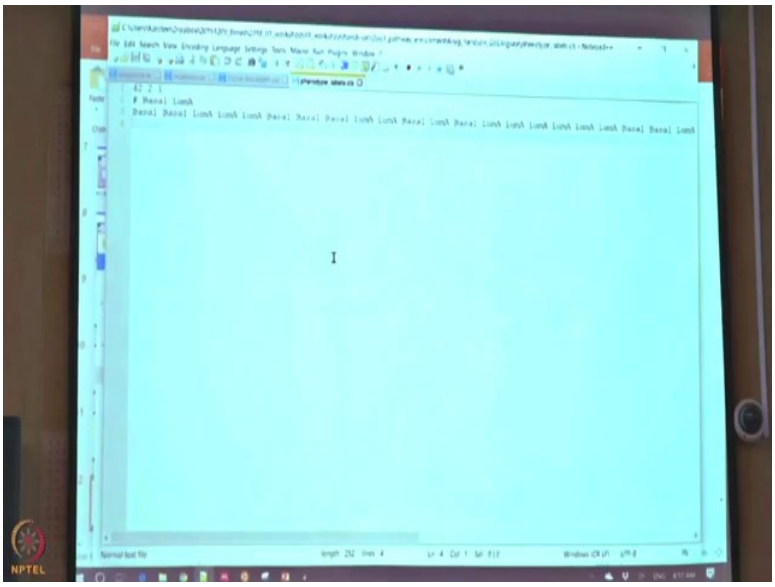


The phenotype labels are stored in so called CLS format. Here if you follow that link you will get again more information about that format. It is again something Broad specific that has been used for while, but now because we have GCT 1.3 you know that is not what we call anymore, but however, for this particular application still is.

And I mean, what is the word here, it is the third line. So, the third line contains the same number of you know columns here, then your GCT file has samples. So, in this case we have vowels here CLS file we will have 42, which is in the same order then the columns in your

GCT file, right. Then you can say, ok, first column is basal, second column is luminal and so on and so forth. Now, already prepared that file and you can find that in your GSEA folder, so which is called phenotype labels dot CLS.

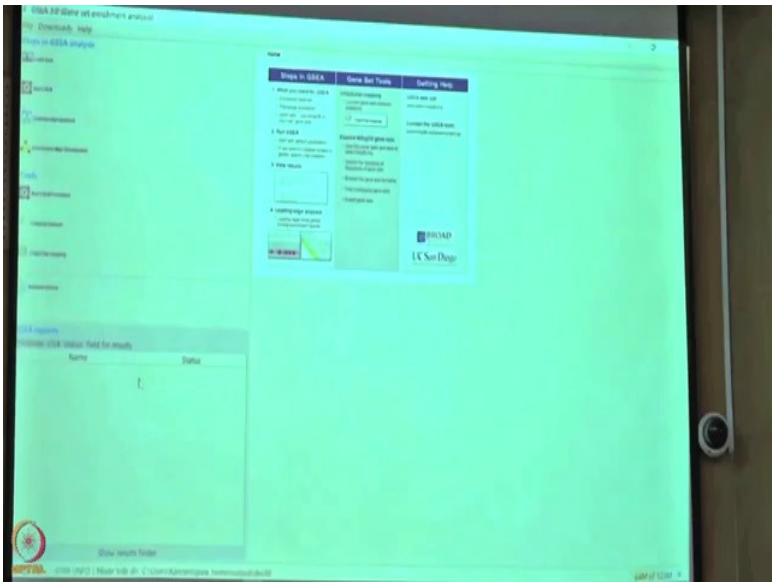
(Refer Slide Time: 29:29)



So, if I open that in word pad. So, the first line tells you how many samples do you have, 42. The second line specify, the second cell here specifies how many groups do you have in your files with two luminal basal. And the third one has to be always one. Do not ask me why, but this is what it says in the web page.

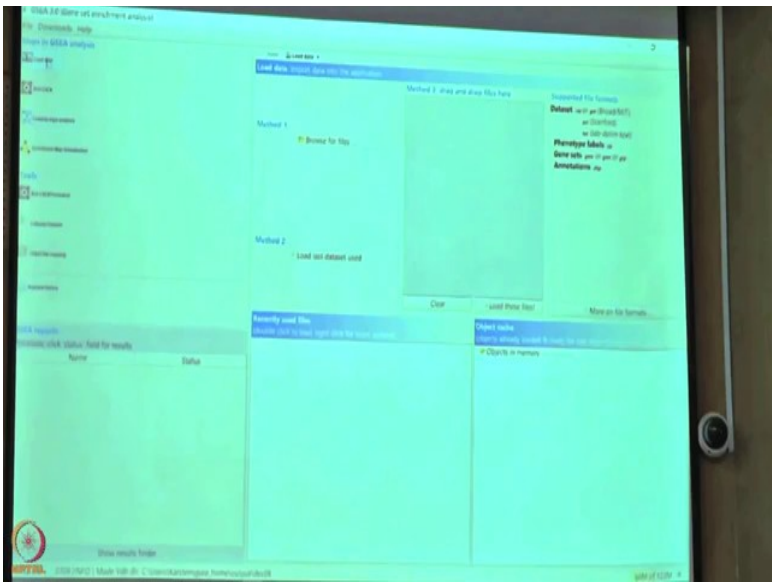
The second line always starts with the hash tag here and then it list both labels like a unique you know what mutation of your class labels. Now, the third line is important one. We just now define for each sample again this has to be in same order, then your GCT file is say, this is your first sample basal or second sample basal or third sample luminal A, ok.

(Refer Slide Time: 30:28)



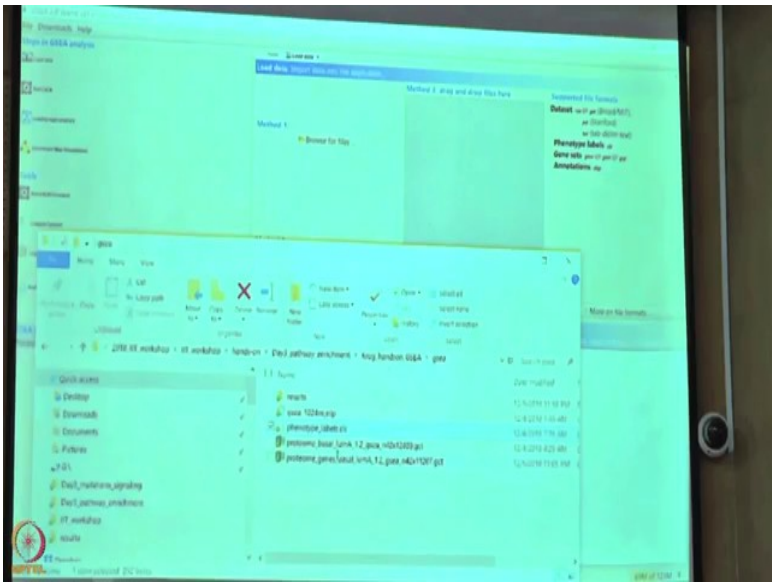
Now, let us try to input the data first. So, we are going to work through these steps and this is also the order which is shown here on the left. so the first step is to load data.

(Refer Slide Time: 30:40)



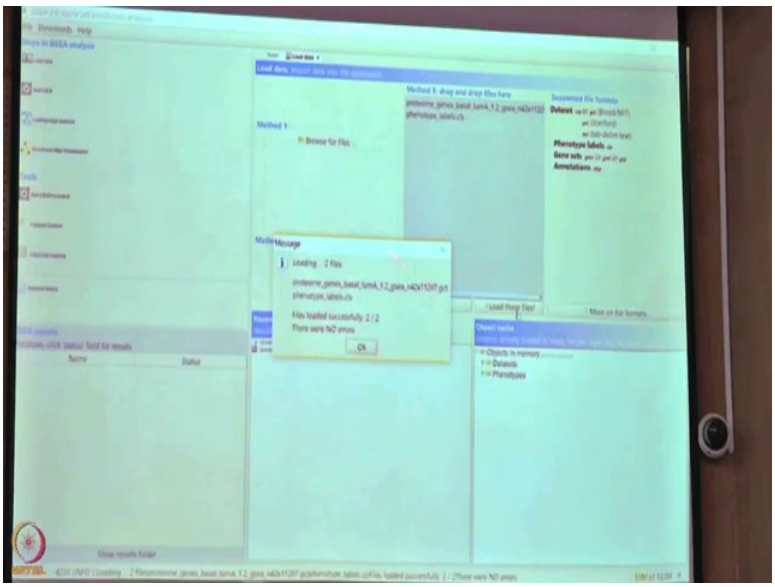
So, we will end up on this page here. So, again you have a different options how to import data.

(Refer Slide Time: 30:47)



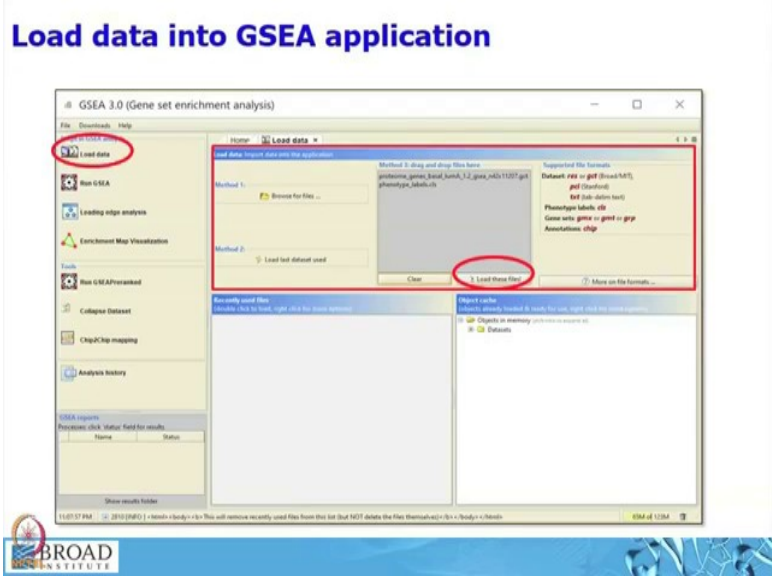
Go to a data and now just make sure that you will select the genes version this time, that is the gene centric version and I just drag and drop it here. So, now the file is here. Now, I do the same with the phenotype labels, alright. So, we have both files here and other direction below these two files you have to press this button here, below these files.

(Refer Slide Time: 31:30)



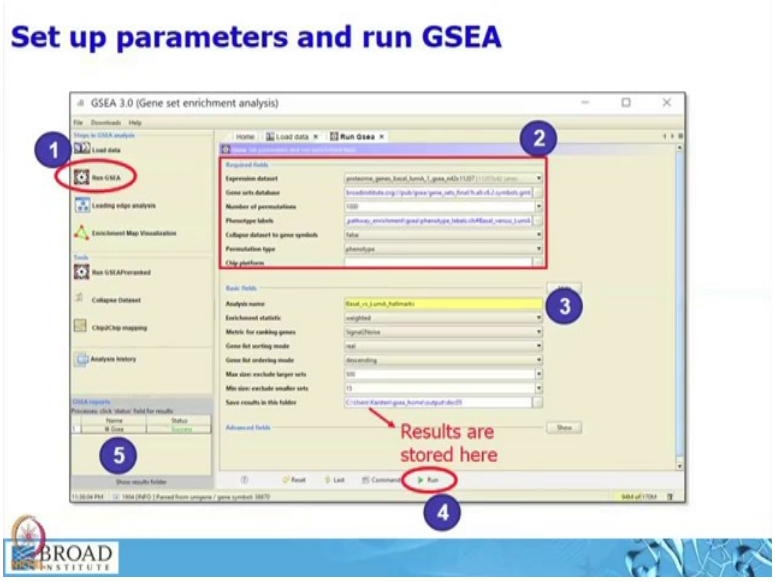
So, whatever be this pop up window which tells you, up uploaded two files is the names of the files and files loaded successfully two out of two which is promising and there were no errors. So, now just going to hide this window I just click, ok, exit.

(Refer Slide Time: 31:51)



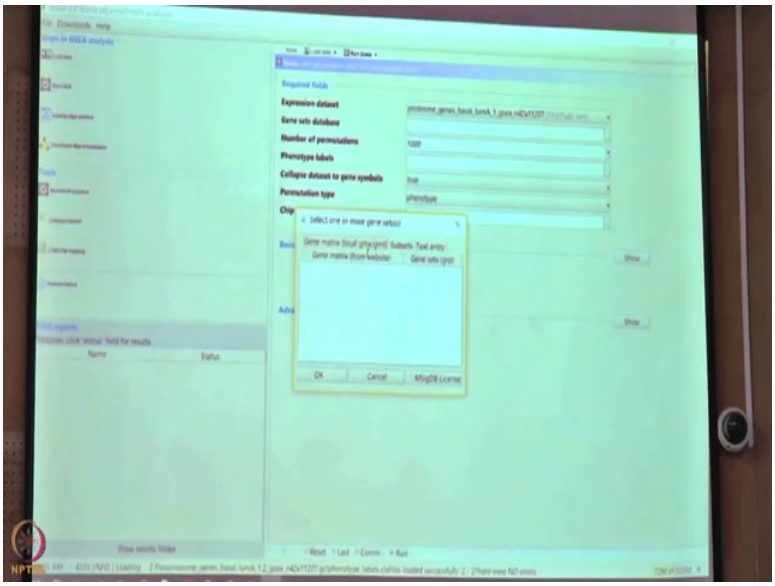
Again, if you go back to my PowerPoint presentation, you will see or these steps that we have just have done here, ok.

(Refer Slide Time: 32:00)



Now, we are going to next page which is called one GSEA. So, here we are going to define parameters that we are going to use during our pathway analysis. So, please click on one GSEA. I am going to do the same on my PC here, one GSEA.

(Refer Slide Time: 32:19)



So, here on the first set of parameters, so these are required fields. So, we have to define those. So, which is you know this is the expression dataset like, gene set database, number of permutations. So, why do we need to do permutations? Say again.

Student: Out of 1000 patients, it will select the best cases.

Almost; so, we are doing permutations. I mean which is calculated once we would get a enrichments code, which would tell us much because we do not know, you know what does it going to tell me. So, we do permutations where permuting the class labels or samples repeating this entire analysis 1000 times in this case.

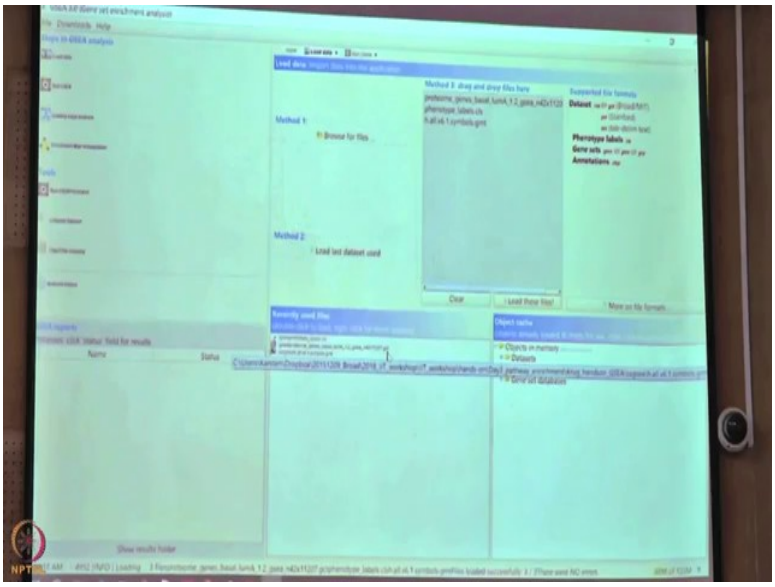
So, we will get this distribution of enrichment scores. Then we can go back and actually calculating the probability that our actual enrichments code that we got, it is in the tail of the distribution or not, which tells us ok. This one is significant was not we can use that to calculate p values, right. So, that is the main purpose.

So, we are generating a back on distribution know of false positives enrichments means course because we randomly shuffle this class labels, so it should make any sense. Like mega random distribution. Then we look where this distribution does actual our enrichment code fall into, ok.

These are the number of permutations and then we have to specify phenotype labels. So, let us how about we just do it, we just start as a top here and if you click on that there will be

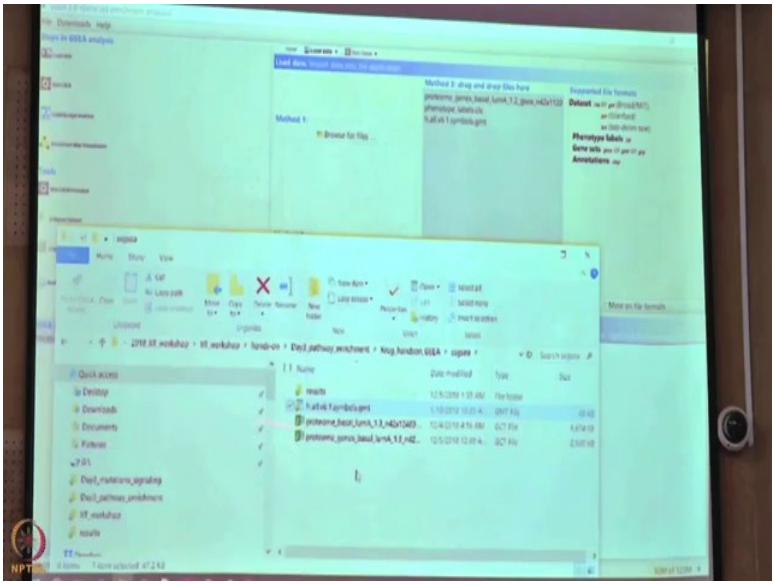
only one data set loaded, right. So, you specify that one, gene set database. So, this might take a second or two because now it is connecting to the Broad servers, ok. Here we go. So, now, I am going to click on here. So, it is as gene matrix, gene matrix local GMT, ok.

(Refer Slide Time: 34:54)



So, you have to import this database first like we have, like we imported the GCT file and the class label file. So, please go back to data and then please go to the single sample GSEA folder. So, right.

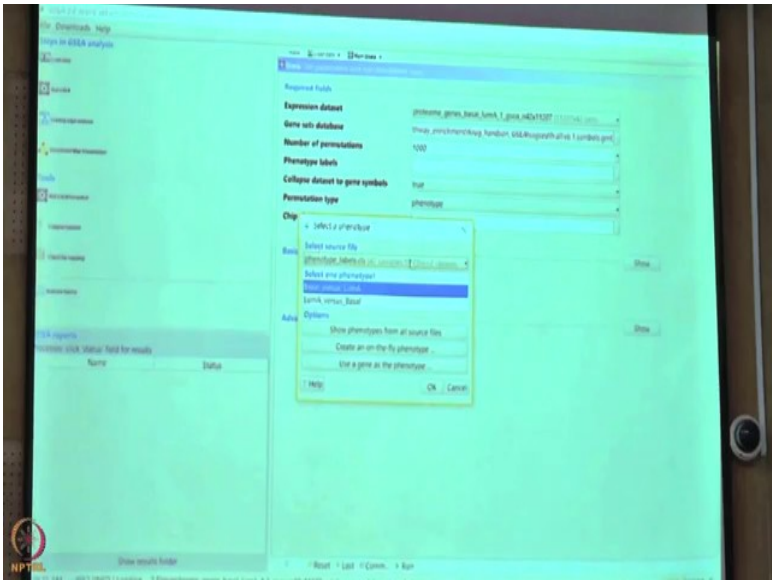
(Refer Slide Time: 35:08)



Now, we were here in the GSEA. Now, I go one full up and there would be another one called in the single sample GSEA, and in that further you will find this file h dot version 1, version 6.1. So, that is the hallmark database. You can just again simply drag and drop into GSEA.

So, now, if you go back to one GSEA, we will we should be able to see that database, once we get the message again, ok. Now, I am able again now see this file here, right, ok. So, number permutations we just discussed now, so now, we load phenotype labels. So, here you can say that in the first panel here you can select the source file, so if there is only one, right.

(Refer Slide Time: 36:04)



So, you could also have multiple phenotype labels and you can you know play along with different ones, but here we only have one. But here now it actually let us you select what kind of comparisons do you wanted to do, do you want do basal versus luminal or do you wanted to do luminal versus basal. I think you just you know pick one and I just leave it at basal versus luminal and I click, ok. Now, we are almost done so.

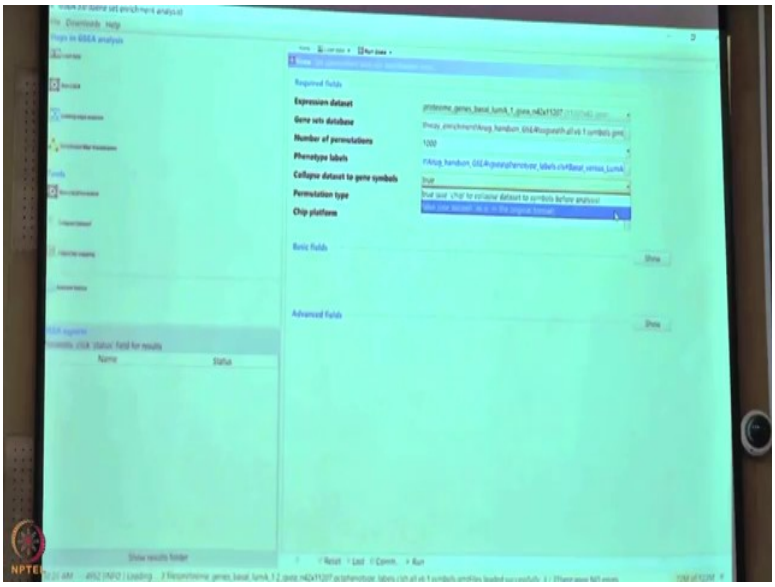
So, the next option here is actually very important.

So, collapse dataset to gene symbols, so that is what we have done already in Morpheus. So, just keep in a mind that this software has been developed in 2005, I mean not this particular software itself, but the principle of gene set enrichment analysis. So, there was no proteomics

and no RNA seq, no, I mean you know there was no RNA seq and no proteomics pretty extended we know now.

So, this has been developed for microarrays, right. And a software comes with the option to collapse microarray probes to genes. So, this is what is option is for. So, we can we are not able to use that here. So, that is why we have to reselect that and just say, false. Use dataset as is, so that is what you want to do.

(Refer Slide Time: 37:40)



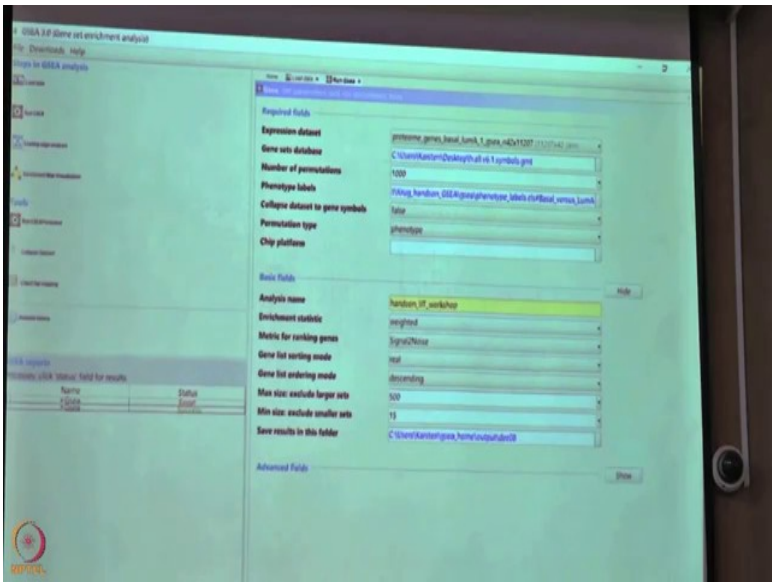
So, permutation type you can either select whether you want to do permutations on your phenotypes, meaning on your sample columns or you can do a permutations in your gene sets. So, why would you do? Why would you have to choose or change this option? So, can you think of a scenario where you cannot do your permutations on your phenotype labels on your sample columns? Over that be. So, one is based on the phenotype and the other is based on gene set. So, in option 1, we would permute the sample columns and a phenotype labels, and in option 2, we would permute the gene sets we would then only generate gene sets, nonsense gene set to create our background distribution.

Student: Actually, when we have lot of sample about 500 500. So, you would like to perform permutation to get to get best results.

Exactly. If you have a sufficient number samples and I would already consider forty or you know sufficient; you can kind of permit, you can do a permutation, you got a samples. If you

just a through a biological replicates you cannot do permutations on two replicates, right. So, then you would choose gene set, but in this case we have like 40 samples in total it is totally find to do out permutations there, ok.

(Refer Slide Time: 39:21)



So, now we actually filled out all required fields. So, now, we can we expand the basic fields let me just going to do some small reductions here. So, first of all we can just give it a in an analysis name, Hands-on IIT workshop. What is? Oh, the most important option here.

I mean you know principally do not have to worry about these kind of parameters, but you would have to worry about probably if some scenarios is how you do your ranking. If you remember, so GSEA works on a ranking of your genes. So, it would rank in our case luminal versus a basal. So, it has to do some sort of marker selection or some sort of ranking that differentiates luminal from basal.

So, the 4th option is signal to noise, which is basically if I am correct it is basically that the average between luminal and basal, the difference in average is divided by the product of the standard deviations of both groups that gives you measure. Specifically, the fold change divided the fold change between luminal and basal scaled that is standard deviation, right.

If you have a higher fold change but higher standard deviational as well you would end up with a lower fold change, whereas, if you have a higher difference the higher fold change

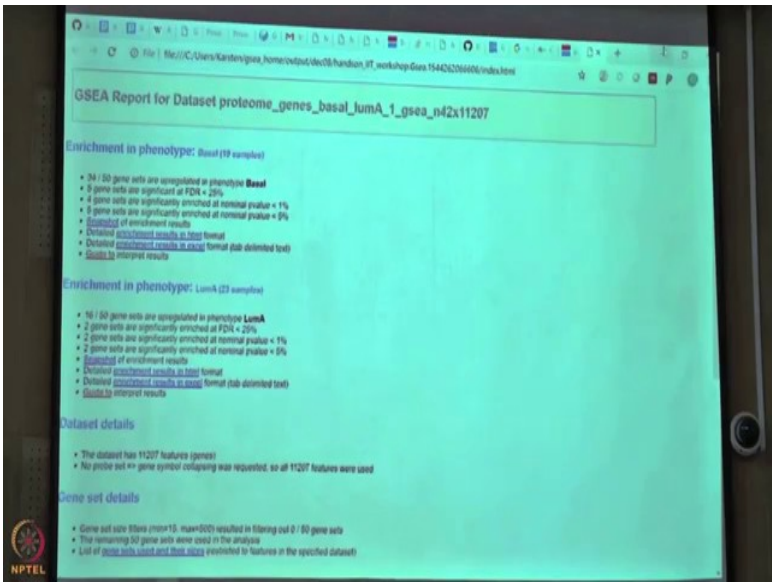
with a lower standard deviation your denominator would be very small and you would still get a higher fold change.

You could also do, I think this requires at least they say on a web page you should have at least 3 or so samples in each phenotype, if you want to do that. You could also do for example, t-test you know that is probably the second data would recommend or other matrix like you know Euclidean distance between luminal and basal or correlation and things like that, right. That is different matrix how to rank a genes.

In general, I would recommend just leave it to signal to noise as long as you have sufficient number of samples. So, we are going to leave that here. And you know, so this might be interesting for you because that is the folder where you can find the results afterwards. So, this is where GSEA is towards its results and you can also change that folder, but this is like the d folder, you will find the results here.

And these other filters here. You can exclude gene sets that have fewer than 15 members and more than 500 samples, I mean these are pretty good fold parameters; you do not have to worry about them too much, ok. I think this should be everything that we need to actually perform the GSEA analysis and in order to run that just you have to click this little one button here, ok. Now, it shows success too me so.

(Refer Slide Time: 43:21)



Let us take a look at a results and we can just click on success and then there should be page should pop up like an HTML report which summarizes your results.

Student: What is the significance of FDR.

Student: That is false detection rate, no.

FDR is a false discovery rate.

Student: False discovery rate.

Exactly.

Student: That has to be less.

Yes.

This is basically the fraction of let us say we have like 100 pathways, right and if its axis if you have an FDR 5 percent, this tells you that 5 pathways actually falls positive, similar.

Student: Ok.

Yes.

List of significant files

Student: Let 95 files are actually genuine.

Yes, here the fold parameter in GSEA is 25 which is very loose right

Student: Yeah.

But you can you can also adjust the parameter. So, we are looking at FDR, so false discovery rate smaller than 25 percent. So, 25 is actually pretty high, right. So, that is the default setting here and you know as protein nothing you would put in a paper, right, but it helps you do you know get a first claims on your data. You will have all of these results in an excel sheet as well, where you have the FDRs and then you can just filter or look at a pathways that have a certain FDR.

So, this is basically, I think this parameter has been used here which man study like in 2005, and still made it until where this version here. So, it is basically a summary of you know very high level summary of results. So, you have at least two blocks, the first block tells you this is a enrichment of a phenotype basal. So, everybody with me?

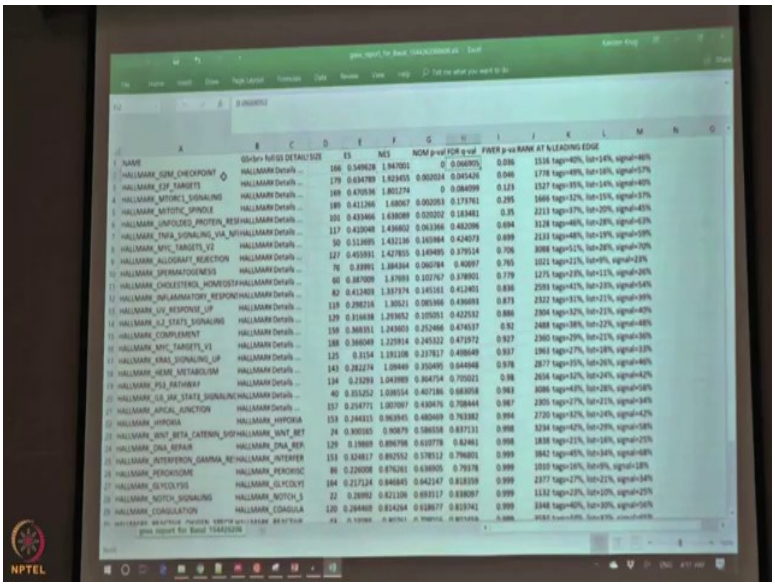
Student: Yes.

So, we see we have 19 samples in basal. The second block is a enrichment in phenotype luminal A where we have 23 samples, then you can get some you know high levels summary here. So, for example, 34. So, I told you that the hallmark database has 50 gene sets, so that is why here is a 34 out of 50 gene sets are up-regulated. So, they have positive enrichment score. So, they are ha more in which in basal compared to luminal. Luminal we only have 16, right and they should adopt to 50. This does not mean that they are significant or whatsoever, this tells you the direction.

So, here a FDR 25 percent and again this is very high, I know, that again that's a summary and it tells you it is 5 gene types that are below 25 percent for basal, and there is two gene sets that are below 25 percent of luminal A. And all of these, you know this, this entire page is again these are different hyperlinks that work forward you to the actual results. So, here you have a summary about the datasets. So, you were looking at the 11,000 genes and so on and so forth and here is the summary about your gene set database.

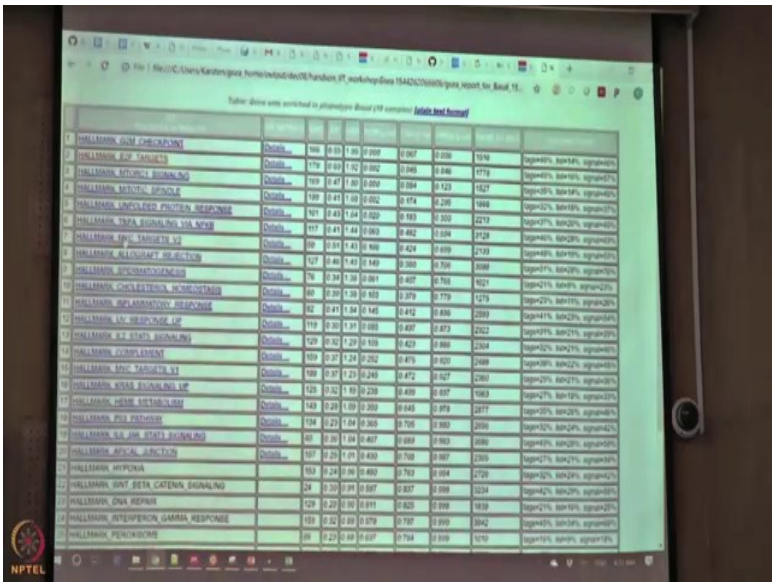
And also, you have very detailed and very extensive documentation about GSEA because it is such an old software, old approach, very well developed, very well maintained and curated level of documentation and tutorials on line. And here already you find the direct link how to interpret the results. You can just click on that and you will find all the information you need in order to make sense out of these result page. So, what I am going to show you and here yourself direct link to the excel sheet as you can see, right. If you click on that should be excel sheet.

(Refer Slide Time: 47:27)



So, that is now the gene that the pathways for basal, right and here you also have the FD'R val, the FDR value, where you can just focus on the first tool which are below 5 percent and then below 10 percent, right, ok. Let us look at an example here. So, what I did now, I clicked on enrichment in phenotype basal, I clicked on detailed enrichment results in the HTML format.

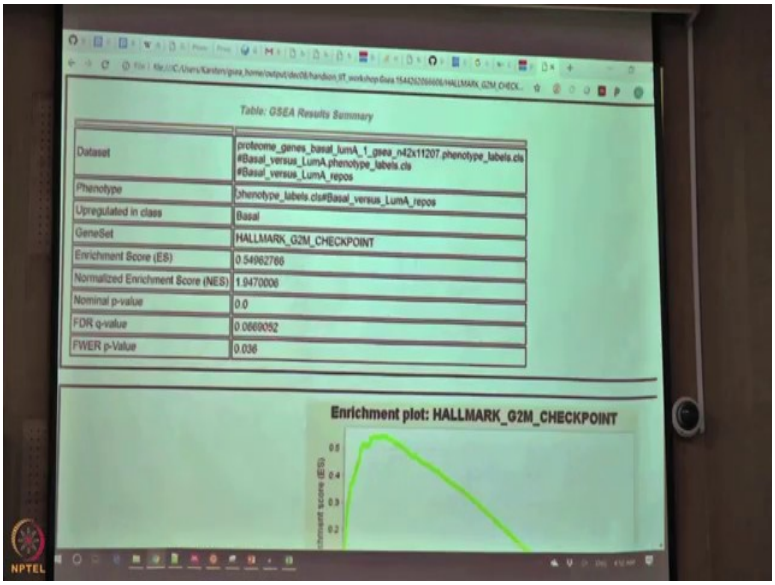
(Refer Slide Time: 48:14)



So, this is kind of a similar table that you got in the excel format, but now is HTML and you have this different hyperlinks here. So, the most significant, most differential pathways

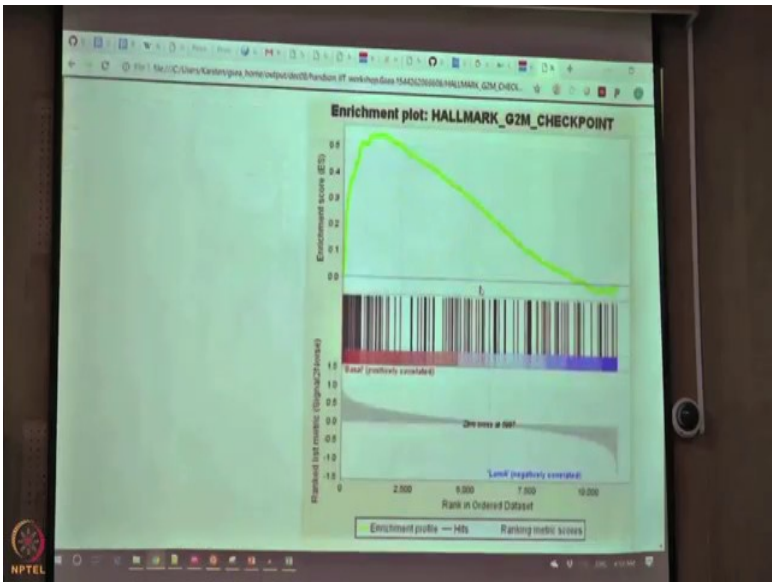
family is G2M checkpoints, so some cell cycle. So, now, for each of these gene sets where pathways you can click on GSEA details.

(Refer Slide Time: 48:38)



And you will actually get these enrichment plots, right. So, here we are looking at G2M checkpoint signature, you have p-value and the FDR value which is associated through with this pathway. And we also have this enrichment plot here, right.

(Refer Slide Time: 49:00)



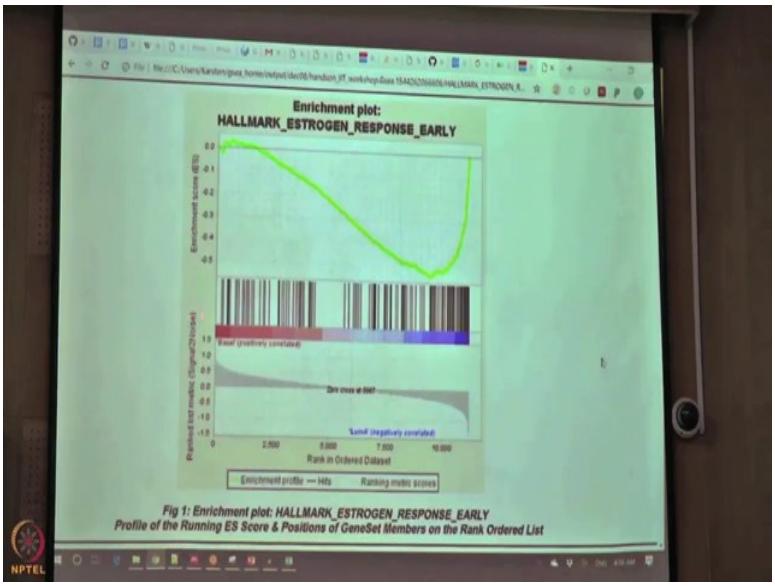
So, in the x axis we have the genes which are ranked ordered according to their differential expression between basal which is shown on your left.

And luminal A which is shown on right. So, these are all genes. So, these are all the genes that are more abundant and basal, here are the genes that are more abundant and luminal A. And all of these Were dig a bar in this case are members of that particular pathway. And again, just by eye balling in CA, in your this cluster of members here, right which are which basically do cluster about genes that are very very abundant in basal subtitle, right; and if you calculate the enrichment score.

(Refer Slide Time: 50:23)

So, if you now look at another example for luminal A, we can also look at the snapshot of an enrichment results. This should be like summary about all of the pathways. So, what are the; so, this is the most significant ones, this is the second most significant one, and we see both of them are estrogen related, estrogen response early, estrogen response late. Does that make sense? So, we are comparing luminal A versus basal. Now, we are looking at luminal specific pathways and many of these luminal tumors are positive. This is actually the hallmark cancer pathways, pathway that we are seeing here in this set.

(Refer Slide Time: 51:10)

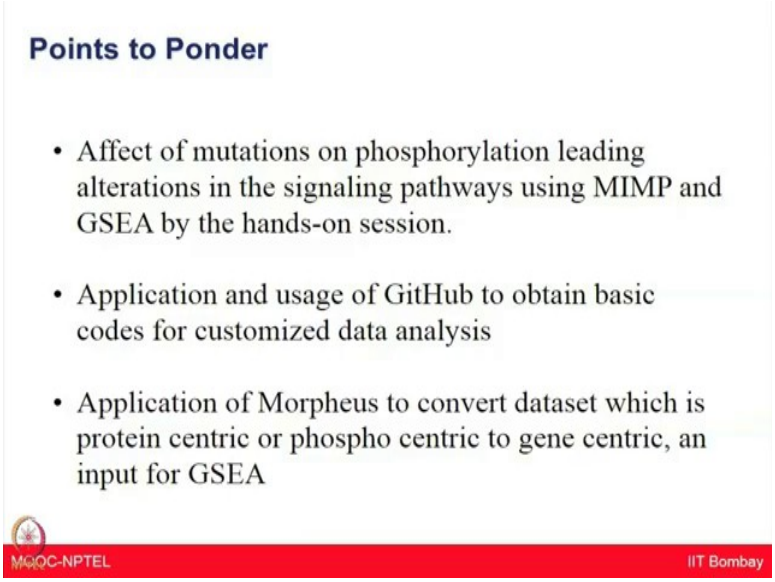


So, what shown here is the signature noise that we got correlated. It also says that here and here according to that this is the ranking of my genes in my entire data set. So, we had like 11,207 genes in this dataset, in this ranking, according to the signal noise signal to noise characteristics comparing basal and luminal A sub types.

So, again, so what these genes here are more abundant and basal sub type. These genes are no; on that side of the ranking, are more abundant in luminal A subtype, right. And again, so here we see a clear enrichment. So, this is very good bias against our members of this particular gene set, right, it has to generate response early we see clear cluster of these members here in genes that are more abundant and luminal A subtype.

So, I think the most difficult part is to get the data in to the right format and I gave you some hints to use Morpheus so on and so forth.

(Refer Slide Time: 52:21)



In today's lecture, I hope you learnt how to use R scripts and incorporate data in MIMP and GSEA tools for understanding and visualization of your data. Use of Morpheus to convert to your dataset, which is protein centric or phospho-centric to gene centric which can be used as an input for GSEA.

The next session is going to be again hands-on session in which the Dr. Bing Zhang will talk about how one could use linkedomic tools.

Thank you.