An Introduction to Proteogenomics

Dr. Sanjeeva Srivastava
Department of Biosciences and Bioengineering
Indian Institute of Technology, Bombay

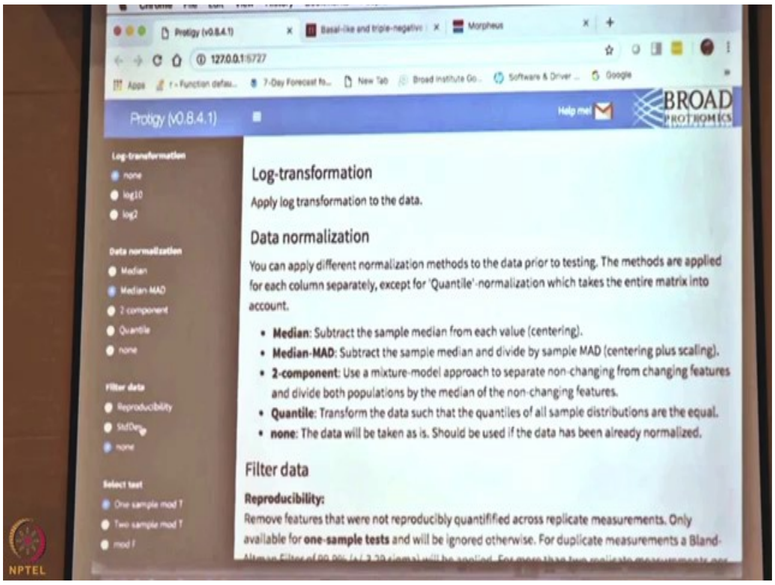
Lecture – 24
Hands-on: Protigy II

Welcome to MOOC course on Introduction to Proteogenomics. In the last lecture Dr. Mani provided you an overview of using Protigy software. Several features of Protigy includes data normalization and filtering, data quality control checks, marker selection interactive, visualization of results, integration of protein-protein interaction databases, as well as how to save your analysis sessions and then you can share with your colleague researchers and collaborators and finally, how to export your results into some output files like excel or PDF formats.

In today’s lecture continuing from the previous session Dr. Mani will talk about how to implement log 2 transformation, normalization, data filtering and selection of test using prodigy. He will also discuss about the output summary and data visualization in prodigy. He will finally, show you how protigy can represent your results in multiple way; like PCA plots, scatter plot, volcano plots, etc. So, let us welcome Dr. Mani for today’s session.

I want to show you what it can do so that you can explore things either with the given data or if you are adventurous enough you can try with your own data. So, let us just go through. So, pick PAM50 and then you will get the list of groups and how many samples are in each group and then click, then you get to the main page of protigy where you can decide what you want to do to your data.

(Refer Slide Time: 02:17)



Log-transformation

● none
○ log10
○ log2

Data normalization

○ Median
● Median-MAD
○ 2 component
○ Quantile
○ none

Filter data

○ Reproducibility
○ StdDev
● none

Select test

● One sample mod T
○ Two sample mod T
○ mod F

Log-transformation

Apply log transformation to the data.

Data normalization

You can apply different normalization methods to the data prior to testing. The methods are applied for each column separately, except for 'Quantile'-normalization which takes the entire matrix into account.

- **Median**: Subtract the sample median from each value (centering).
- **Median-MAD**: Subtract the sample median and divide by sample MAD (centering plus scaling).
- **2-component**: Use a mixture-model approach to separate non-changing from changing features and divide both populations by the median of the non-changing features.
- **Quantile**: Transform the data such that the quantiles of all sample distributions are the equal.
- **none**: The data will be taken as is. Should be used if the data has been already normalized.

Filter data

Reproducibility:

Remove features that were not reproducibly quantified across replicate measurements. Only available for **one-sample tests** and will be ignored otherwise. For duplicate measurements a Bland-

So, the first column is log transformation. So, suppose you had only ratios and you did not log transform then you can say I want to transform the data using log base 2 or log base 10. So, if you clicked on the appropriate dot it would transform your data. So, right now the data we have is already log 2 transform. So, we need to leave it at none you do not want to log transform, log transform data.

So, what is the log of a negative number, it is not defined; so, half of your data will get thrown away if you try to log transform again. Because the once you have log transformed the up regulated ones are positive, the down regulated ones are negative. If you try to log transform again, all the negative data will become missing, because the log of a negative number is not defined. So, you do not want to do that.

So, then we discussed many different ways of normalizing. So, some of those are implemented here. So, if you pick median, it is just medium centering it will not scaled; if you pick median mad it will median center and mad scale; then if you pick 2 component it will do the 2 component normalization that we went through. And if you want Quantile normalization really, you can also do it, but if you have data that has already been normalized then you would pick none.

So, in this case the data set I provided is intentionally not normalized. So, you would want to normalize it; you could do 2 component if you want to see, but 2 component normalization takes a while, each sample takes like a minute or so, and we have like a hundred samples. So, it would take like half an hour or so, to do it.

So, let us not try that, but median mad is a reasonable alternative. So, for this exercise we will use median mad, you can try 2 component later maybe in your in the evening or something like that. So, then this one has, the data set has about 15000 proteins that came out of spectrum Mill with basically very minimal filtering. So, if a protein was missing in 90 percent of the samples, it is still here; if a protein did not change in any of the samples it is still here. So, you might want to consider implementing some kind of a filter.

Ideally what we do in our analysis is to how like a missing value filter, where things that are missing in too many samples are excluded; here we do not have that filter. But a reasonable filter that I am going to use is called the standard deviation filter; what it does is it takes each protein, looks at the protein measurement across all the samples and calculates the standard deviation and says what is the value of the standard deviation. So, it ranks all the samples, all the proteins by standard deviation and keeps the ones that are which do you want to keep; the ones that are most varying or least varying.

Student: Least varying.

I hear least, anyone votes for most.

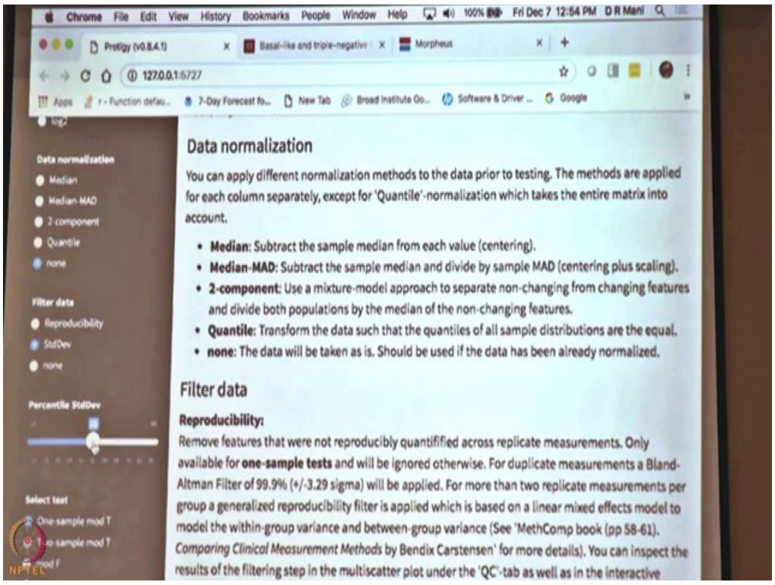
Student: Least.

Least; everybody thinks we should keep the least varying proteins. So, you have a protein that never changes in any of your subtypes, would you want to keep that.

Student: No

No, you want to keep the proteins that are most varied because that is hopefully your marker. Your ideal marker is not present in one and it is like high level in another one, another group. So, that has a high standard deviation. So, you want to keep things that are high that vary more high standard deviation

(Refer Slide Time: 05:36)



- ○ log2

Data normalization

- ○ Median
- ○ Median-MAD
- ○ 2-component
- ○ Quantile
- ● none

Filter data

- ○ Reproducibility
- ● StdDev
- ○ none

Percentile StdDev

Select test

- ○ One-sample mod T
- ● Two-sample mod T
- ○ mod F

Data normalization

You can apply different normalization methods to the data prior to testing. The methods are applied for each column separately, except for 'Quantile'-normalization which takes the entire matrix into account.

- **Median**: Subtract the sample median from each value (centering).
- **Median-MAD**: Subtract the sample median and divide by sample MAD (centering plus scaling).
- **2-component**: Use a mixture-model approach to separate non-changing from changing features and divide both populations by the median of the non-changing features.
- **Quantile**: Transform the data such that the quantiles of all sample distributions are the equal.
- **none**: The data will be taken as is. Should be used if the data has been already normalized.

Filter data

Reproducibility:

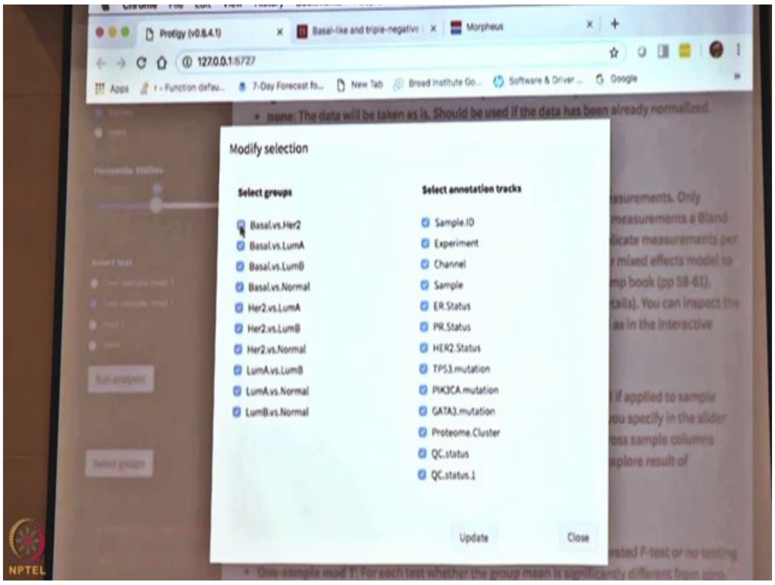
Remove features that were not reproducibly quantifified across replicate measurements. Only available for **one-sample tests** and will be ignored otherwise. For duplicate measurements a Bland-Altman Filter of 99.9% (+/-3.29 sigma) will be applied. For more than two replicate measurements per group a generalized reproducibility filter is applied which is based on a linear mixed effects model to model the within-group variance and between-group variance (See 'MethComp book (pp 58-61). *Comparing Clinical Measurement Methods* by Bendix Carstensen' for more details). You can inspect the results of the filtering step in the multiscatter plot under the 'QC'-tab as well as in the interactive

And here you can pick what percentile of things you want to keep. So, what fraction of things you want to keep, just to make things fast I will pick 50.

So, usually 50 is a more aggressive number, you would pick like throw away like the top that the bottom 10 percentile, but I am throwing away the bottom 50 percentile. So, this is just to make have fewer proteins in the analysis, but you see, you will still have about 7000 proteins in your analysis.

Then the question is, what kind of a test do you want to do? Do you want to do a 1 sample test, 2 sample test or a moderated or a F test. So, here I am going to pick 2 sample and then I will tell you why. So, if you pick 2 sample we pick PAM50 as the annotation and you want to compare some 1 or 2 of your PAM50 classes let say; then you would pick 2 sample and then to decide which two you want to compare, you click on select groups.

(Refer Slide Time: 06:42)



Modify selection

Select groups

- □ Basal.vs.Her2
- ☑ Basal.vs.LumA
- ☑ Basal.vs.LumB
- ☑ Basal.vs.Normal
- ☑ Her2.vs.LumA
- ☑ Her2.vs.LumB
- ☑ Her2.vs.Normal
- ☑ LumA.vs.LumB
- ☑ LumA.vs.Normal
- ☑ LumB.vs.Normal

Select annotation tracks

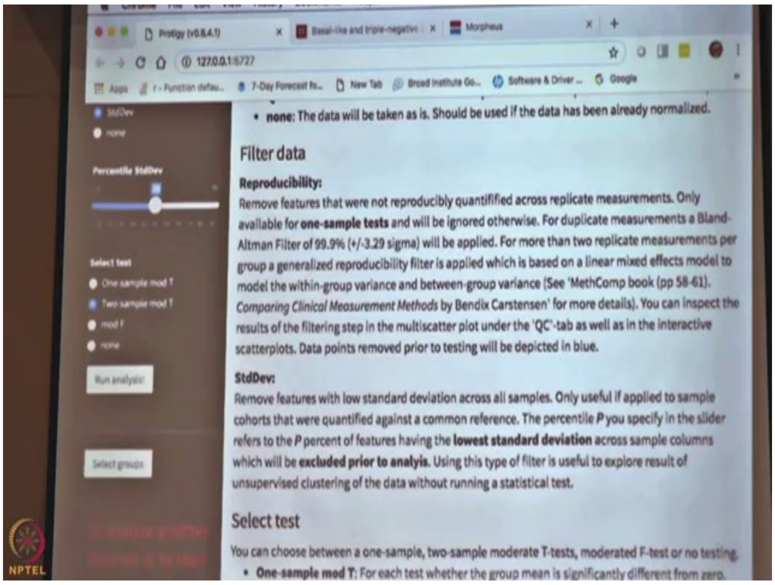
- ☑ Sample.ID
- ☑ Experiment
- ☑ Channel
- ☑ Sample
- ☑ ER.Status
- ☑ PR.Status
- ☑ HER2.Status
- ☑ TP53.mutation
- ☑ PIK3CA.mutation
- ☑ GATA3.mutation
- ☑ Proteome.Cluster
- ☑ QC.status
- ☑ QC.status.1

So, when you click on select groups it will tell you all the comparisons that are possible, given the annotation that you picked. So, it says you can do Basal versus Her 2, Basal versus Luminal A, Basal versus Luminal B and so on. So, I am just going to say I want to do Basal versus Luminal A and I will deselect all the others, you can do any number of these or all of these if you want.

But the results will be more confusing and harder to interpret, but you can pick which one you are looking at any point, I will just look at one. And when you do a display of the data as a heat map, it can show you annotation tracts, showing say you want to see which one was ER positive versus ER negative and so forth, you can keep that information and show it in plots if you want.

So, that is the annotation track selection. I am going to remove things that are basically likes a vary across every sample like sample ID, experiment, TMT channel, again sample name and QC status I will remove; and just keep the ER PR HER 2 status and mutation status for the three genes that are there, so we can look at that. Then click update to make sure that selection is registered and then close this.

(Refer Slide Time: 07:55)



- **none:** The data will be taken as is. Should be used if the data has been already normalized.

Filter data

Reproducibility:
Remove features that were not reproducibly quantified across replicate measurements. Only available for **one-sample tests** and will be ignored otherwise. For duplicate measurements a Bland-Altman Filter of 99.9% (+/-3.29 sigma) will be applied. For more than two replicate measurements per group a generalized reproducibility filter is applied which is based on a linear mixed effects model to model the within-group variance and between-group variance (See 'MethComp book (pp 58-61). *Comparing Clinical Measurement Methods* by Bendix Carstensen' for more details). You can inspect the results of the filtering step in the multiscatter plot under the 'QC'-tab as well as in the interactive scatterplots. Data points removed prior to testing will be depicted in blue.

StdDev:
Remove features with low standard deviation across all samples. Only useful if applied to sample cohorts that were quantified against a common reference. The percentile *P* you specify in the slider refers to the *P* percent of features having the **lowest standard deviation** across sample columns which will be **excluded prior to analyis.** Using this type of filter is useful to explore result of unsupervised clustering of the data without running a statistical test.

Select test

You can choose between a one-sample, two-sample moderate T-tests, moderated F-test or no testing.

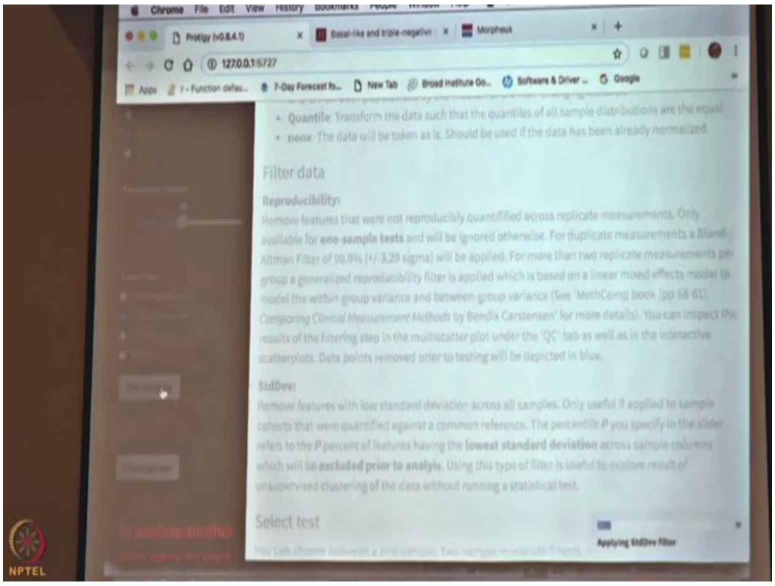
- **One-sample mod T:** For each test whether the group mean is significantly different from zero.

So, now you are ready to run your analysis.

Student: So, what is the data normalization was used?

We pick Median MAD now. So, we can pick median MAD again and that was not supposed to happen, but not sure. So, let us just make sure the select groups is still there yeah that is there.

(Refer Slide Time: 08:18)



- **Quantile**: Transform the data such that the quantiles of all sample distributions are the equal
- **none**: The data will be taken as is. Should be used if the data has been already normalized.

Filter data

Reproducibility:
Remove features that were not reproducibly quantified across replicate measurements. Only available for **one-sample tests** and will be ignored otherwise. For duplicate measurements a Bland-Altman Filter of 99.9% (+/-3.29 sigma) will be applied. For more than two replicate measurements per group a generalized reproducibility filter is applied which is based on a linear mixed effects model to model the within-group variance and between-group variance (See 'MethComp book (pp 58-61). *Comparing Clinical Measurement Methods* by Bendix Carstensen' for more details). You can inspect the results of the filtering step in the multiscatter plot under the 'QC'-tab as well as in the interactive scatterplots. Data points removed prior to testing will be depicted in blue.

StdDev:
Remove features with low standard deviation across all samples. Only useful if applied to sample cohorts that were quantified against a common reference. The percentile *P* you specify in the slider refers to the *P* percent of features having the **lowest standard deviation** across sample columns which will be **excluded prior to analyis**. Using this type of filter is useful to explore result of unsupervised clustering of the data without running a statistical test.

Select test

You can choose between a one-sample, two-sample moderate T-tests, ...

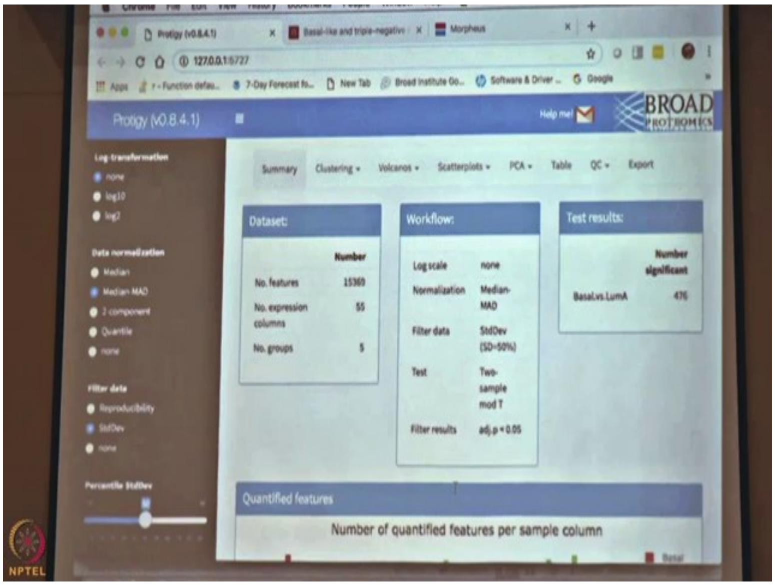
Applying StdDev filter

So, all the rest are fine. So, just double check and then click run analysis, it will take like half a minute.

It will tell you on the bottom what it is doing. So, it is applying standard deviation filter now, after normalization and then it will do the 2 sample T test and then you will get a page that shows results. So, while we are waiting I will just keep talking on what you will get at the end of this. So, at the end of this you will get a screen with multiple. So, actually it is going through, it is running the 2 sample test now.

I think it is done. So, here is the results now.

(Refer Slide Time: 08:49)



Log transformation
- ● none
- ○ log10
- ○ log2

Data normalization
- ○ Median
- ● Median-MAD
- ○ 2-component
- ○ Quantile
- ○ none

Filter data
- ○ Reproducibility
- ● StdDev
- ○ none

Percentile StdDev

Dataset:

	Number
No. features	15369
No. expression columns	55
No. groups	5

Workflow:

Log scale	none
Normalization	Median-MAD
Filter data	StdDev (SD<50%)
Test	Two-sample mod T
Filter results	adj.p < 0.05

Test results:

	Number significant
Basal.vs.LumA	476

Quantified features

Number of quantified features per sample column

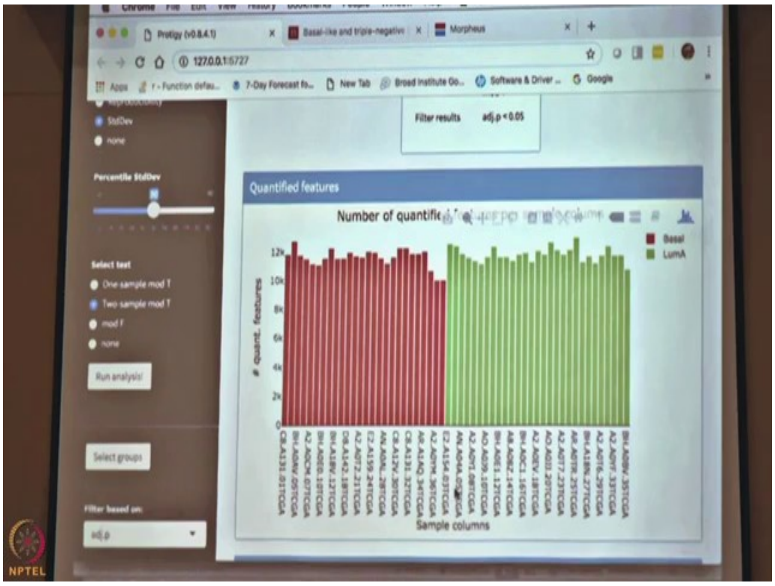
Basal

So, you will get a page like this, the top of the page gives you a summary of the data. So, you had 15369 proteins and you had 5 groups. So, number of expression columns so 50; so remember we had about 100 or 105 samples I think, but we did a Basal versus Luminal A comparison. So, the total number of basal and luminal samples together is 55.

Student: Yeah.

Then the workflow shows what log scaling, what normalization, what filtering you used and what test you ran. And so, we filter the results to look at only things that are statistically significant after adjusting the p-value and the p-value adjustment is Benjamini Hochberg FDR correction that is the only one that we have and that is always applied. So, the results are that there are 476 markers of Basal versus Luminal A; in the filtered data set that are statistically significant with an adjusted p-value of 0.05.

(Refer Slide Time: 10:11)



Filter results adj.p < 0.05

Percentile StdDev

Select test

○ One sample mod T

● Two sample mod T

○ mod F

○ none

Filter based on: adj.p

Student: Sir why we are getting in different results what is when significant going to 111

Maybe you are changed your p-values or you used a different normalization.

Student: Sir may be in the I mean, I have the Basal versus Her 2 may be you see the last number of significance.

Student: I have them use different kind of.

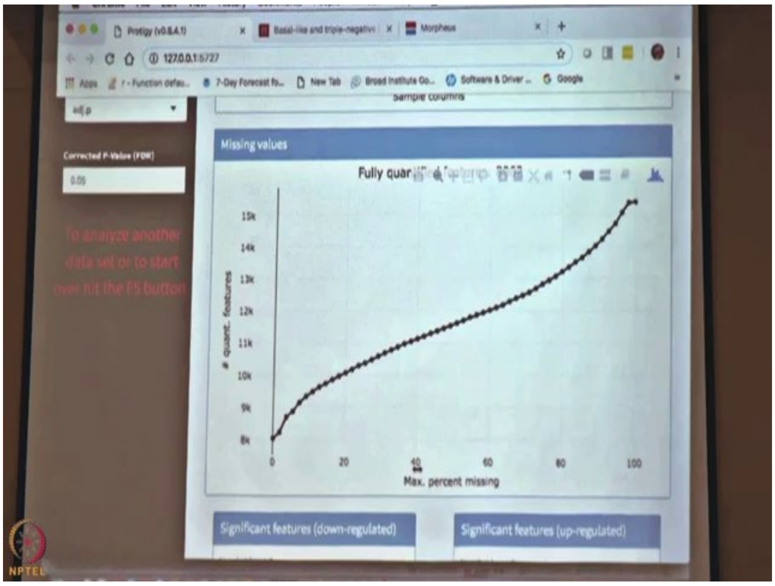
Yeah you could have used a different test or different groups.

Student: Ok basal versus it is a different thing

Yeah; obviously.

So, the bar plots here, show how many proteins were observed in each of your samples. So, this is like the number of proteins in this sample was about 12000; the second sample had a little more and so on. And the red bars are for the Basal samples and the green bars are for Luminal samples. So, suppose you looked at this and all the red bars were there and all the green bars were half the size; then you would be very worried because there is some batch effect or some effect somewhere that consistently is observing fewer number of proteins in the luminal samples only. So, this is basically like a QC check to kind of make sure that nothing is grossly wrong with your data.

(Refer Slide Time: 11:34)



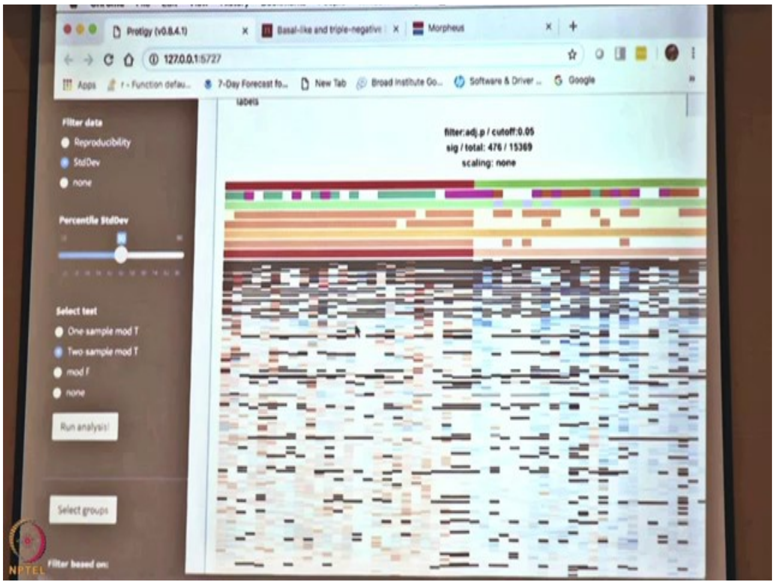
So, this one shows, how many missing values are there in each of the sample.

Student: (Refer Time: 11:41).

There are about 8000 proteins that are seen in every sample, all samples have these 8000 proteins observed; about 50 percent missing. So, when you take into account proteins, that have about 50 percent missing; so, they are not observed in 50 percent of the samples, then there are totally about 12000 proteins that fall into that category less than or equal to 50 percent missing.

So, this basically is to show the rate of missing values, how many do you have on the in the average sample and how many are observed in all the samples. So, you just to make sure that there are in too many missing values that you have.

(Refer Slide Time: 12:31)



Filter data
○ Reproducibility
● StdDev
○ none

Percentile StdDev

Select test
○ One sample mod T
● Two sample mod T
○ mod F
○ none

filter:adj.p / cutoff:0.05
sig / total: 476 / 15369
scaling: none

The other things you look at are there is a clustering tab, which can generate a heat map.

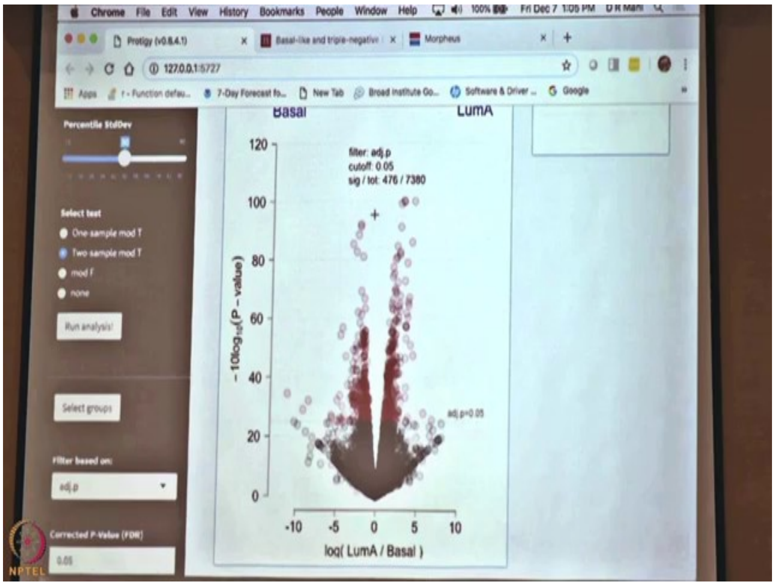
So, this is a heat map. So, you can see the annotation bars on top. So, red is Basal, green is Luminal A; and then these are other annotations that you have on the side, I think my screen is too small to see those maybe you can see it. But I think these tracks are ER PR Her 2 status; and you can see that ER PR and HER2 are basically all negative in basal samples, because generally triple negative samples fall into the basal category.

The other thing you notice here is that the black marks are missing values; remember we included missing values and the test that we do, can actually handle missing values, so we did not fill in missing values. And so, this display showing which values are missing for each of the proteins and these are all statistically significant proteins, that are different between basal and luminal. You can see at the top that basically this line is almost a complete black line with 1 or 2 things here and there.

So, that is saying that that protein was present in only a couple of samples. And because they were reasonably different between the basal and luminal the statistics is saying that was significant; would you really believe it, I would not. So, this is a strong indication that you really need to filter to remove proteins that are missing in too many samples. So, if you did that, this would kind of get chopped off somewhere here and after that it is fine. You have missing values here and there it is ok; but if your conclusion is made on two probe samples in one group and one sample in another group with all the rest missing that is not what you want.

So, this you can basically by visually exploring your data, you can get a feel for what is happening, whether your analysis is reasonable or not and you can constantly keep making sanity and quality checks to make sure that your analysis is a reasonable and the results you are getting is reliable.

(Refer Slide Time: 14:55)



Percentile StdDev

Select test
○ One sample mod T
● Two sample mod T
○ mod F
○ none

Filter based on:
adj.p

Corrected P-Value (FDR)
0.05

filter: adj.p
cutoff: 0.05
sig / tot: 476 / 7380

So, the other thing you can play around with lot of settings and things to make it look like you want I would not do all those, but the other thing you want to look at are volcano plots.

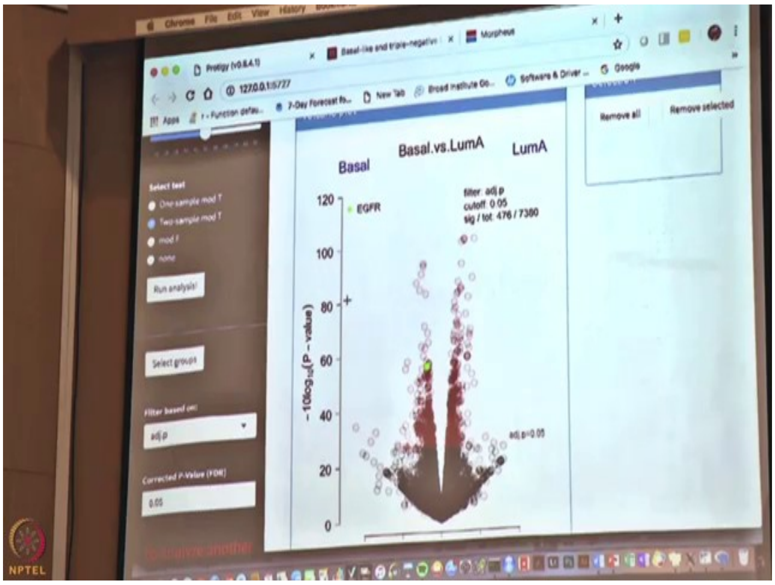
So, are people you are familiar with volcano plots.

Student: No.

No it is couple of them are. Now volcano plot is a plot of fold change on the x axis versus statistical significance or p-value on the y axis. So, on the x axis we have log fold change. So, if it is negative it is non-regulated; if it is positive it is up regulated, but this is comparing basal versus luminal. So, basically if it is on the left side it is up in basal; if it is on the right side it is up in luminal and the farther away it is from the x axis the more statistically significant it is.

So, p-values decrease when they become more significant, but we have multiply it by a negative sign, so it goes up. So, it is visually kind of impactful, right. So, if it is far out on the top then it is statistically significant protein. So, anything that is the beyond the threshold of 0.05 adjusted p-value is marked in red. So, remember it said there were 400 something markers that were significant, you are seeing all those markers and you can see which ones are up in basal and which ones are up in luminal. And in this one if you go and click on it; it will tell you what marker it is. So, if you click there.

(Refer Slide Time: 16:21)



Select test
○ One-sample mod T
● Two-sample mod T
○ mod F
○ none

Filter based on: adj.p

Corrected P-Value (FDR): 0.05

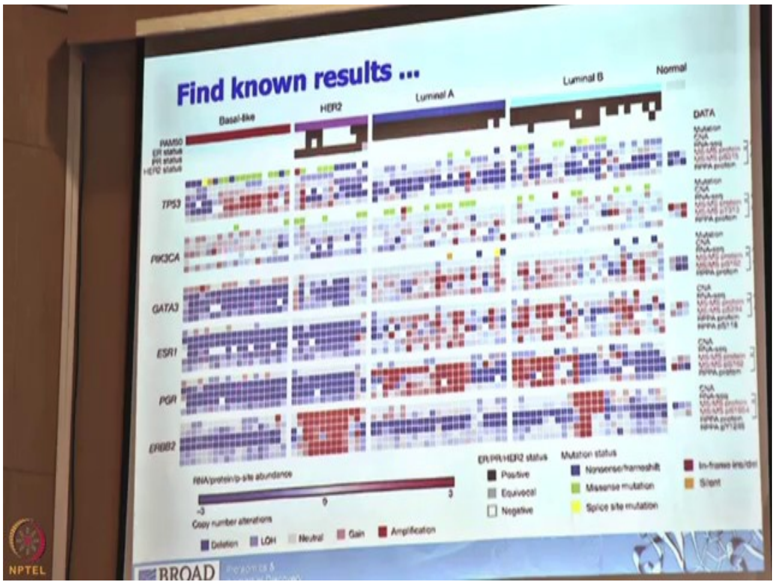
filter: adj.p
cutoff: 0.05
sig / tot: 476 / 7380

It will hopefully tell you.

Student: Just one below.

Oh that is true, sorry there. So, that protein is a that gene is AGR 2 and the protein REFSEQ ID is shown over there.

(Refer Slide Time: 16:46)

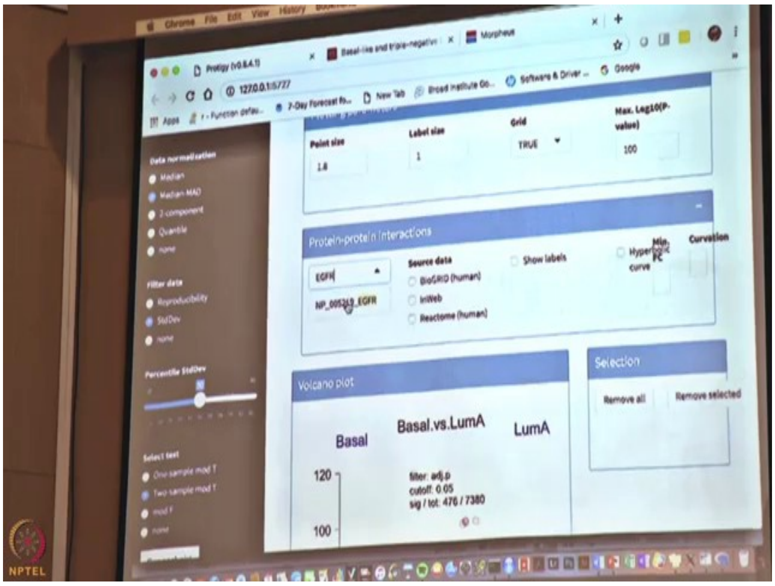


Find known results ...

So that, if you look at the breast cancer paper, so we said P 53 there are some up regulation in Basals and down regulation in Luminal A, if you look at ERGB 2 there are kind of similar in both Basal and Luminal A. So, we are comparing Basal-like and Luminal A in this comparison.

So, if you looked at P53 then it should be up in Basal and kind of down in Luminal. Similarly with PIK3CA and I think another protein that is known to be up in basals is EGFR. So, we can take a look at all those things in here, by doing the following.

(Refer Slide Time: 17:25)



Point size: 1.8
Label size: 1
Grid: TRUE
Max. Log10(P-value): 100

Protein-protein interactions

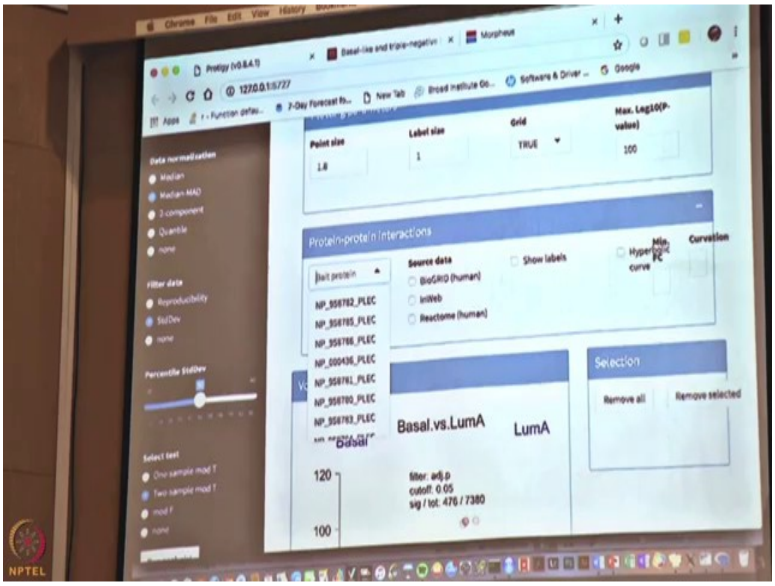
EGFR

NP_005219_EGFR

filter: adj.p
cutoff: 0.05
sig / tot: 476 / 7380

So, you go to protein-protein interactions, click on the plus sign. Deselect everything under source data and then typing the name of your protein.

(Refer Slide Time: 17:33)



Point size: 1.8

Label size: 1

Grid: TRUE

Max. Log10(P-value): 100

Data normalization
- ○ Median
- ● Median-MAD
- ○ 2-component
- ○ Quantile
- ○ none

Filter data
- ○ Reproducibility
- ● StdDev
- ○ none

Select test
- ○ One-sample mod T
- ● Two-sample mod T
- ○ mod F
- ○ none

Protein-protein interactions

Bait protein
- NP_958782_PLEC
- NP_958785_PLEC
- NP_958786_PLEC
- NP_000436_PLEC
- NP_958781_PLEC
- NP_958780_PLEC
- NP_958783_PLEC

Source data
- ○ BioGRID (human)
- ○ InWeb
- ○ Reactome (human)

□ Show labels

□ Hyperbolic curve

Basal.vs.LumA

filter: adj.p
cutoff: 0.05
sig / tot: 476 / 7380

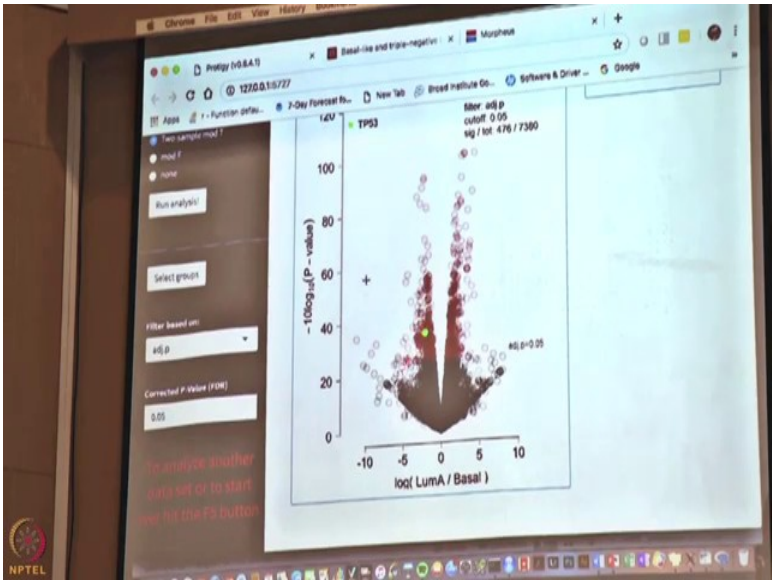
So, you want EGFR. So, it gives you the protein with the gene name you click on it, if you go here you can see EGFR is way over there it is statistically significant and it is up in basal, like we expected.

So, you see you can also try can you do multiple.

Student: No.

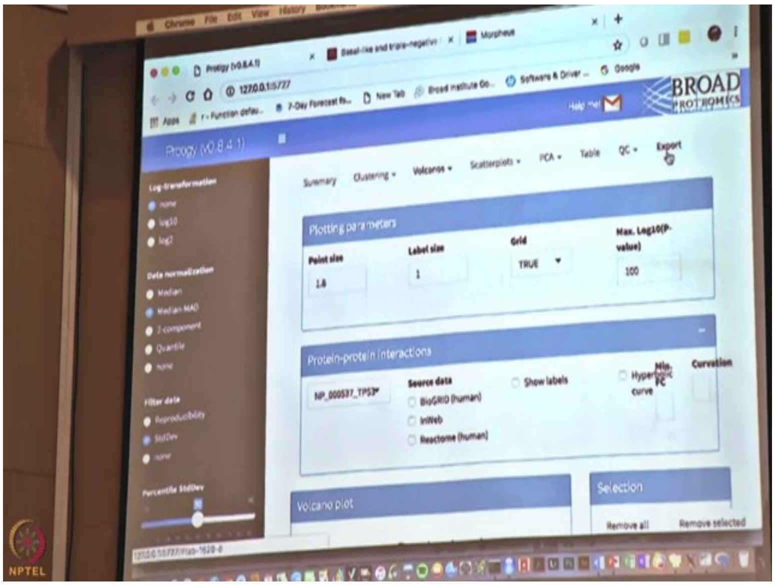
No. So, the other thing you can do is TP53. So, pick that.

(Refer Slide Time: 18:09)



You can see, this is also up in basal, but is not as just statistically significant as EGFR. So, you can explore your proteins, you can kind of export the list of statistically significant markers and look at it you can do all kinds of things.

(Refer Slide Time: 18:24)



Point size: 1.8

Label size: 1

Grid: TRUE

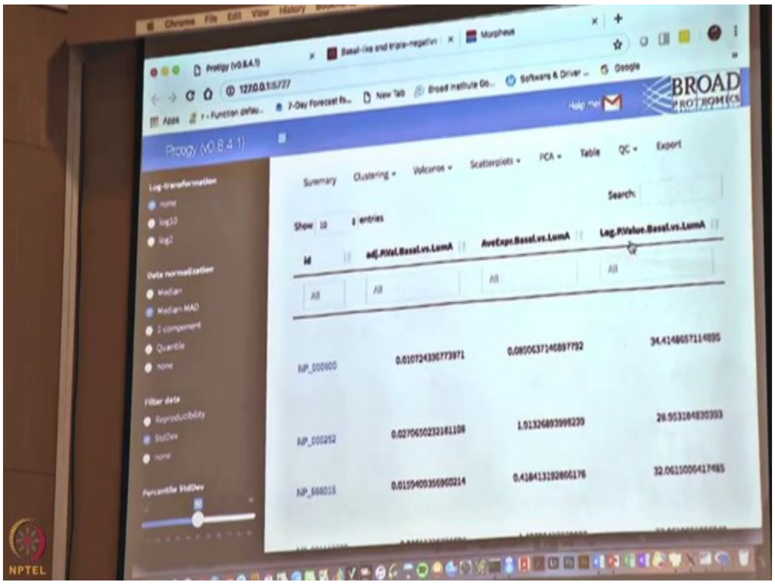
Max. Log10(P-value): 100

NP_000537_TP53*

Student: From there we can export from table

 There is export. So, you can use that.

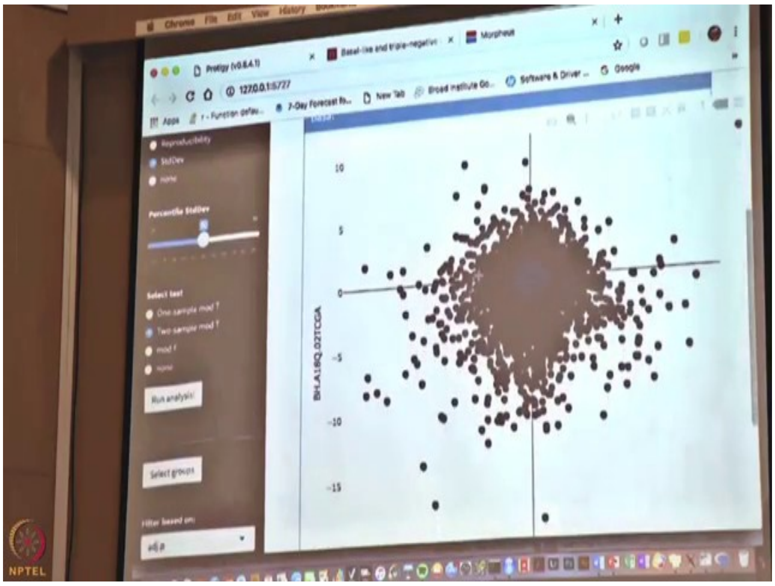
(Refer Slide Time: 18:31)



id	adj.P.Val.Basal.vs.LumA	AveExpr.Basal.vs.LumA	Log.P.Value.Basal.vs.LumA
All	All	All	All
NP_000600	0.0107243307773971	0.0850637146897792	34.4148657114895
NP_000252	0.0270650232181108	1.91326893998239	28.9531848303933
NP_666016	0.0155940356960214	0.418413192866176	32.0615006417485

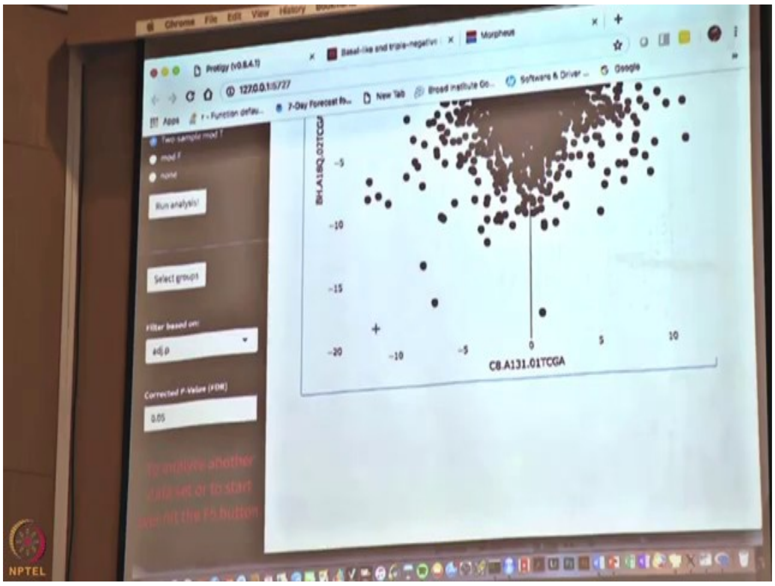
Now, if you look at the table, you can see the actual values. So, these are the adjusted P-values, the average expression, the log fold change; all the information that you might want to include in your paper or you want to like including important some other software you want to look at are all here and you can export this as a table to look work on it.

(Refer Slide Time: 18:55)



If you want to see how one sample plot; if you want to see it is a given protein, you know how it measures in one sample versus another you can do like a scatter plot. So, this is this sample versus that sample.

(Refer Slide Time: 19:07)

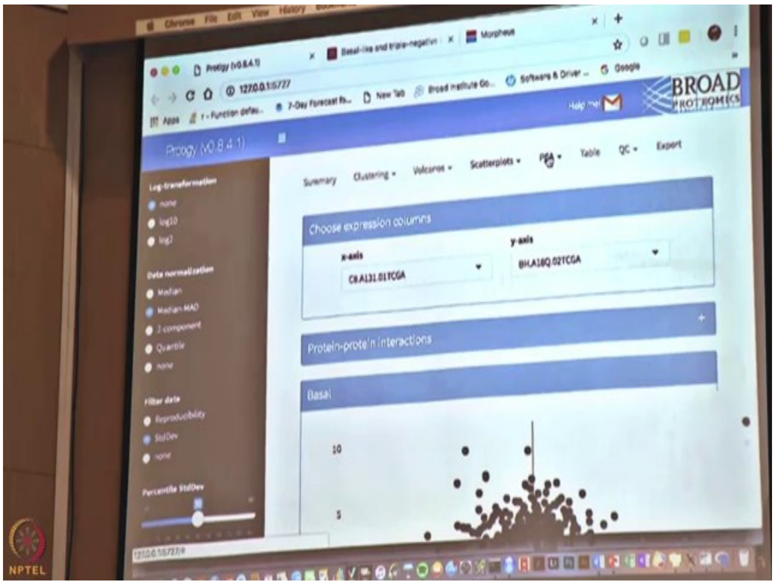


And it is showing all the proteins in that measured in those two samples.

So, you can see most of them are similar.

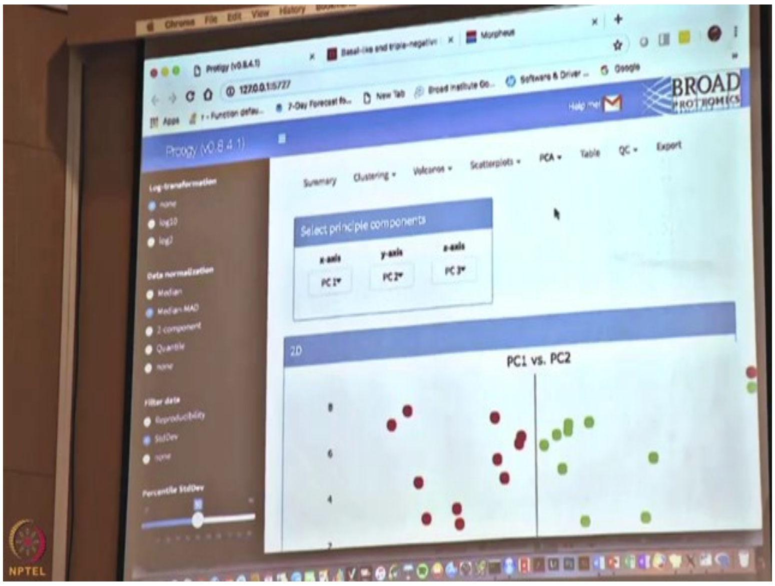
Some of them are more extreme.

(Refer Slide Time: 19:19)



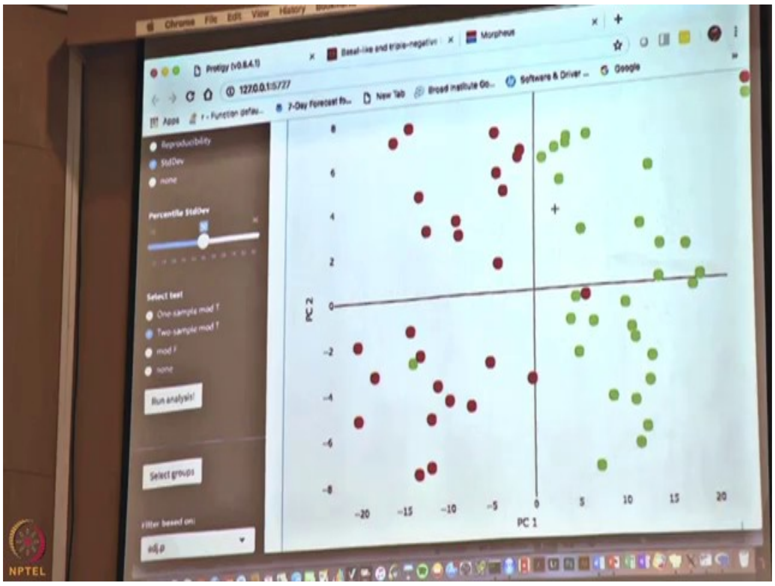
So, these are all like a basically things to look at the data and kind of get a feel for what is happening.

(Refer Slide Time: 19:27)



Now, other thing I want to show is PCA.

(Refer Slide Time: 19:30)



So, when you do PCA, you can see that is the plot you get and it is colored by basal versus luminal. So, you can see if you draw a vertical line it essentially separates Basals from Luminals; that is saying that this is the most dominant signature in the data, but you can also see.

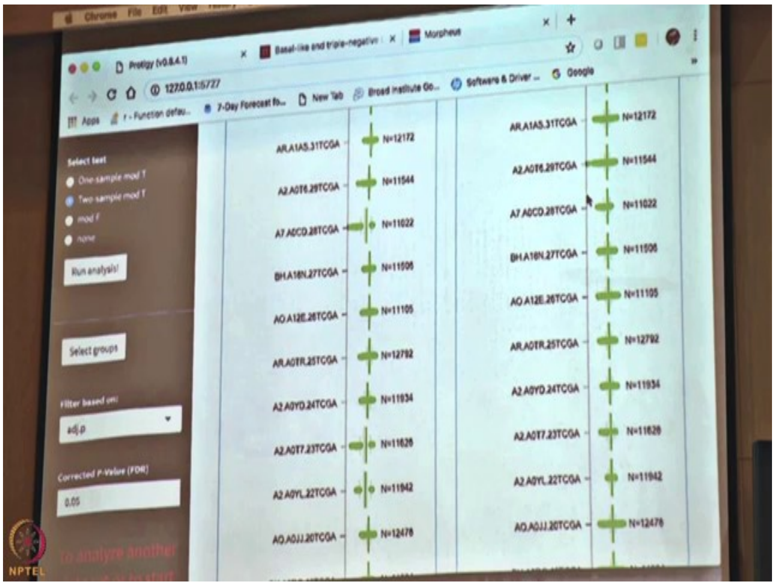
That there is ones red dot in the middle of all the greens and one green dot in middle of all the reds; so there are two samples that kind of behave like the other group. So, what are these you might want to go and explore?

So, in clustering also, if you do a look at the heat map you may be able to see this, but here it is much more striking. And now you can say ok, what happened with those samples? Are they mislabeled; what is the reason they behaving like the other group. So, you might want to go and explore those.

So, these are all tools to kind of generate hypothesis, so you can explore it more biologically and kind of build a biological story. It is not like this is going to build the story for you and write the paper, but this will give you tools to look at it from a biological perspective. And the tools are set up in such a way that people who do not do programming and who are primarily biologists or experimentalist can look at it. So, that is kind of the whole point of Protigy.

So, people who do not do R programming, can actually use the results of other peoples R programming. I think, I will stop there are QC section has many plot, actually let us just look at box plots.

(Refer Slide Time: 20:58)



Select test
○ One-sample mod T
● Two-sample mod T
○ mod F
○ none

Run analysis!

Select groups

Filter based on: adj.p

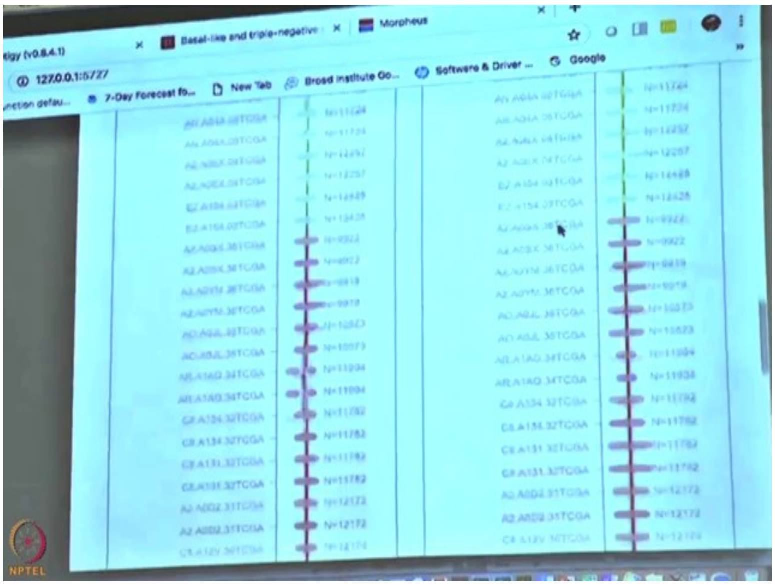
Corrected P-Value (FDR): 0.05

So, this will show you what happens after normalization. So, on the left side is the box plot before normalization, on the right side or box plots after normalization; I think the screen is so small it is all squish. But you can see this sample was actually adjusted quite a bit to get to normalization, to agree with all the others, but the other samples were adjusted to a lesser degree.

So, you can get a feel for are there samples that were that you had to use extreme normalization factors to get them to agree with all the others. In that case you might want to see whether those samples had any issues or if there was a less material for that sample for whatever reason or if the samples did not just work out; they failed for some reason. If you can show that they are failed for whatever reason, you can throw that out and do your analysis it will be more robust, if there are less offending samples.

But you should not through away sample simply to get a better result; but if there is a experimental reason why some sample failed, you can remove that sample and redo the analysis. The thing that is about it, I will stop there if there are any quick questions I will answer.

(Refer Slide Time: 22:13)



Student: Sub samples are margin rate, what is that color code?

So, here the red and green were Basal versus Luminal A.

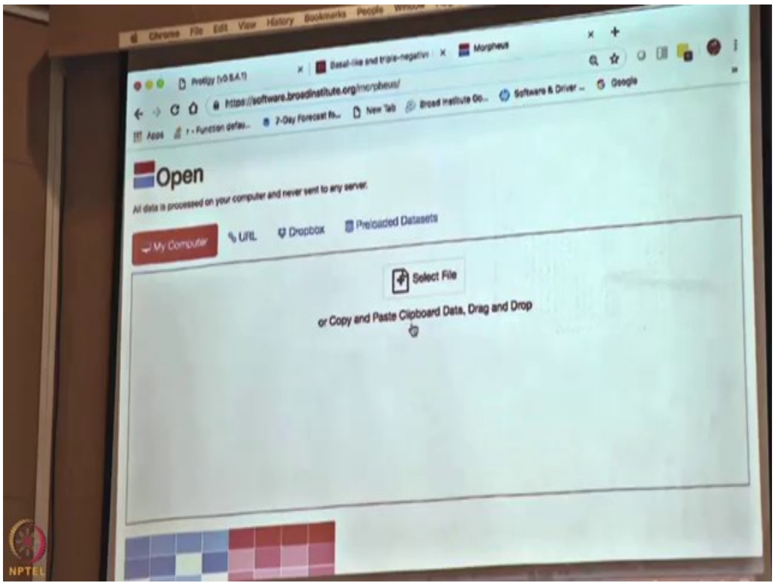
Student: CSV or excel format

Yeah.

Student: how to make gct file?

Yes. So, if you have a excel file or a CSV file, you can go to Morpheus.

(Refer Slide Time: 22:32)



All data is processed on your computer and never sent to any server.

or Copy and Paste Clipboard Data, Drag and Drop

So, like I, it is actually come up here. So, if you go to Google and search for Morpheus broad you will get the website and you can go to the software which is called Morpheus. And see there is select file or drag and drop file, you take your excel file or CSV file and drop it in there, it will open it and show a heat map and then you can save it as a GCT 1.3. You can add annotations using different files if you want and then you can save it as a GCT file and then use it in a protigy. You can also use CSV files directly in protigy; but when you load it, it is going to ask you to annotate the samples.

So, you have it will create a template and then you have to fill the template with your annotation and then load in the template. So for a hands on, it was little more complex. So, we did not do it, but it is also possible.

Student: How we drag CSV files in Morpheus and we go with the plots.

So, you go to once you do that you will get a top bar which as like a menu.

Student: Yeah.

And then there will be a way to export the table.

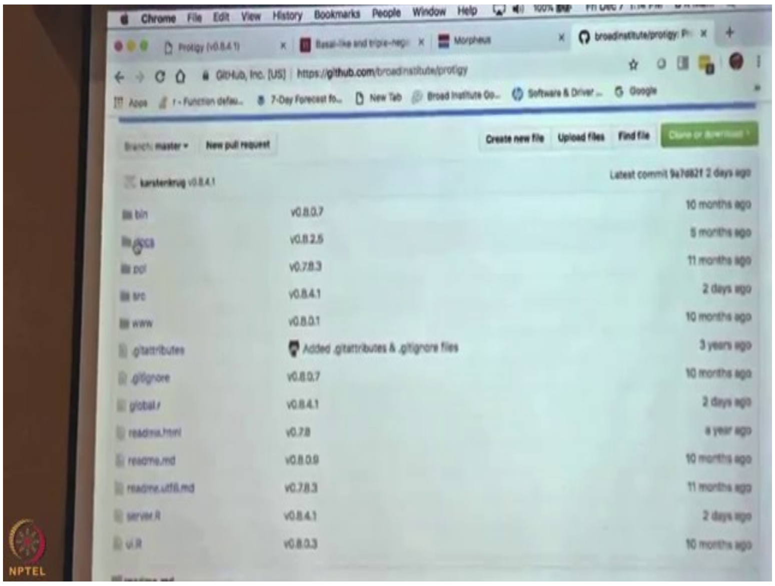
Student: Ok, we have to export the table.

Yeah you have to export the table and it will ask you what format and you pick GCT 1.3.

Student: Do we have a manual for protigy.

For protigy? So, go to github and search for prodigy, you will get broad institute protigy, jump to.

(Refer Slide Time: 24:00)

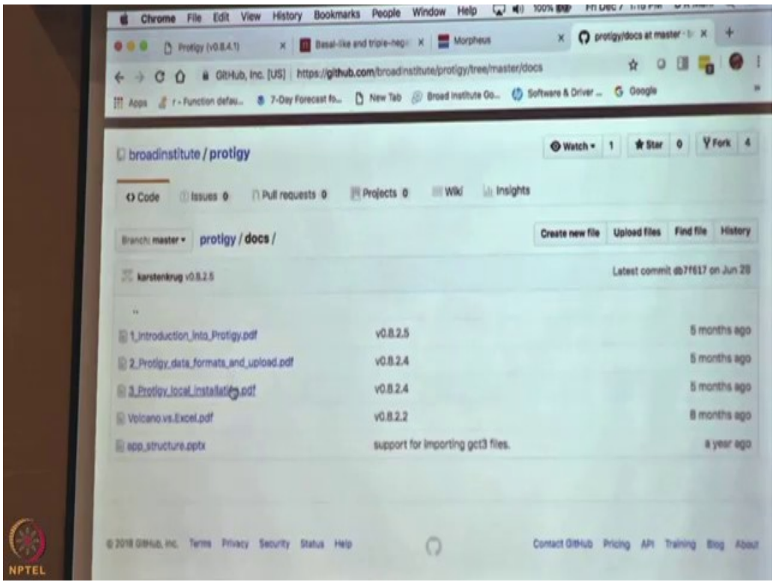


karstenkrug v0.8.4.1 Latest commit 9a7d821 2 days ago

bin	v0.8.0.7	10 months ago
docs	v0.8.2.5	5 months ago
pol	v0.7.8.3	11 months ago
src	v0.8.4.1	2 days ago
www	v0.8.0.1	10 months ago
.gitattributes	Added .gitattributes & .gitignore files	3 years ago
.gitignore	v0.8.0.7	10 months ago
global.r	v0.8.4.1	2 days ago
readme.html	v0.7.8	a year ago
readme.md	v0.8.0.9	10 months ago
readme.utf8.md	v0.7.8.3	11 months ago
server.R	v0.8.4.1	2 days ago
ui.R	v0.8.0.3	10 months ago

So, there is a directory here called docs.

(Refer Slide Time: 24:01)



broadinstitute / protigy

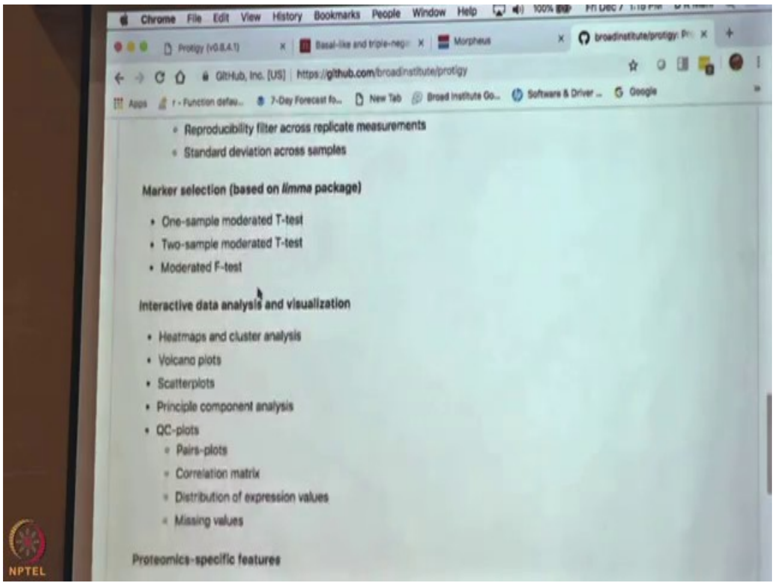
protigy / docs /

karstenkrug v0.8.2.5 Latest commit db7f617 on Jun 28

..		
1_Introduction_into_Protigy.pdf	v0.8.2.5	5 months ago
2_Protigy_data_formats_and_upload.pdf	v0.8.2.4	5 months ago
3_Protigy_local_installation.pdf	v0.8.2.4	5 months ago
Volcano.vs.Excel.pdf	v0.8.2.2	8 months ago
app_structure.pptx	support for importing gct3 files.	a year ago

So, there is an introduction, there is data formats, installation and some documentation there.

(Refer Slide Time: 24:12)


Chrome File Edit View History Bookmarks People Window Help
Protigy (v0.8.4.1)
Basal-like and triple-neg
Morpheus
broadinstitute/protigy: Pr
GitHub, Inc. [US] https://github.com/broadinstitute/protigy
Apps
r - Function defau...
7-Day Forecast fo...
New Tab
Broad Institute Go...
Software & Driver ...
Google
NPTEL

- Reproducibility filter across replicate measurements
- Standard deviation across samples

Marker selection (based on *limma* package)

- One-sample moderated T-test
- Two-sample moderated T-test
- Moderated F-test

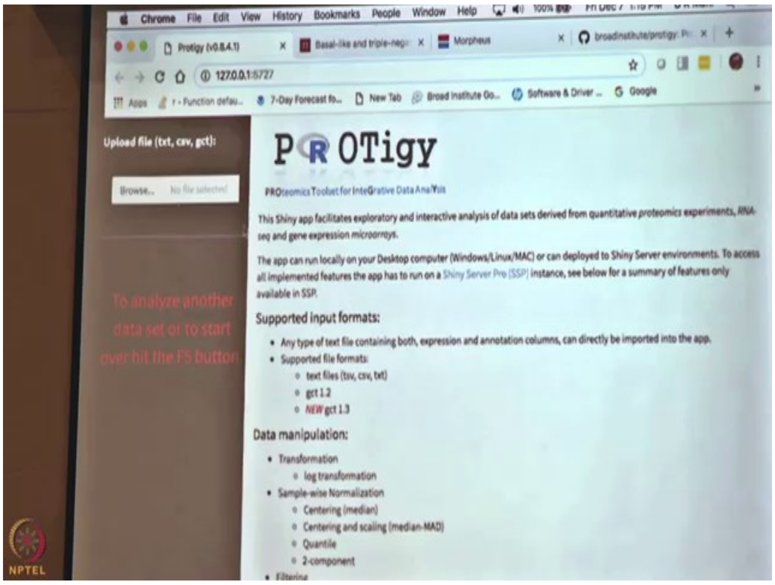
Interactive data analysis and visualization

- Heatmaps and cluster analysis
- Volcano plots
- Scatterplots
- Principle component analysis
- QC-plots
  - Pairs-plots
  - Correlation matrix
  - Distribution of expression values
  - Missing values

Proteomics-specific features

And I think on the main page also a there is some documentation on what all it supports and in protigy itself when you, I am going to refresh.

(Refer Slide Time: 24:22)



Upload file (txt, csv, gct):

To analyze another data set or to start over hit the F5 button

PROTigy

PROteomics Toolset for InteGrative Data AnalYsis

This Shiny app facilitates exploratory and interactive analysis of data sets derived from quantitative *proteomics* experiments, RNA-seq and gene expression *microarrays*.

The app can run locally on your Desktop computer (Windows/Linux/MAC) or can deployed to Shiny Server environments. To access all implemented features the app has to run on a Shiny Server Pro (SSP) instance, see below for a summary of features only available in SSP.

Supported input formats:

- Any type of text file containing both, expression and annotation columns, can directly be imported into the app.
- Supported file formats:
  - text files (tsv, csv, txt)
  - gct 1.2
  - *NEW* gct 1.3

Data manipulation:

- Transformation
  - log transformation
- Sample-wise Normalization
  - Centering (median)
  - Centering and scaling (median-MAD)
  - Quantile
  - 2-component
- Filtering

On the first page it has the same information, but when you load a data set it also tells you for each operation what it does and things like that.

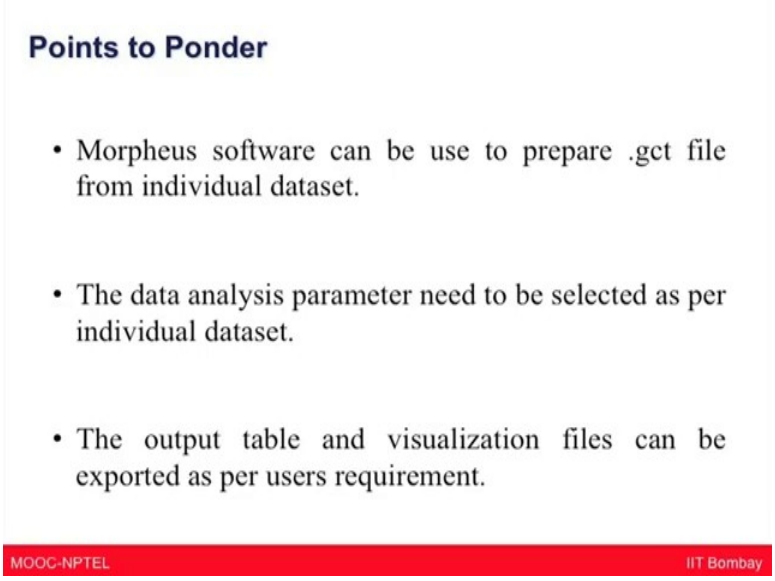
So, there is some documentation, but it is not like thermo documentation because we are not paid like thermo.

Student: Sir sign in needs in github.

Yes you need to sign in, you need an account to log into github it is free, but you have to sign in.

I am sure, you have a lot of, lot more questions and things where little unclear, but the more you explore on your own the more you will remember what you discovered; I could give you step by step for everything, but it will stick less.

(Refer Slide Time: 25:00)



Points to Ponder

- Morpheus software can be use to prepare .gct file from individual dataset.
- The data analysis parameter need to be selected as per individual dataset.
- The output table and visualization files can be exported as per users requirement.

MOOC-NPTEL IIT Bombay

I hope the last two lectures, especially todays last session was very helpful for you to get a glimpse about how to use protigy software for analysis as well as visualization of a data. We also learnt that Morpheus software can be used to prepare file with annotation, which could be used in the protigy analysis. Apart from the annotation, you can also visualize your data using heat maps or even explore the interactive tools like Morpheus.

Varieties of tools are available in protigy to really give you the visual glimpse of what is lying in your big Omics data set. I hope these sessions are giving you not only the basic understanding of various concepts involved in looking into data, but also providing you the open access tools, where you can start implementing them right away from any data set available from databases or your own data set you can start analyzing them now. In the next lecture we will have a guest speaker Dr. Debashish Das who will talk about Proteomics Data Analysis.

Thank you.