An Introduction to Proteogenomics

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

Lecture – 19
Data Analysis: Batch Correction and Missing Values

Welcome to MOOC course on Introduction to Proteogenomics. This is Dr. D. R. Mani's second lecture. In the first lecture, he had discussed about various strategy employed for data normalization. In today's lecture, we are going to continue the discussion about data analysis and specially the batch correction and missing value imputation. These two things are very important for any type of omic data analysis. Just because you can generate large amount of data set in a very, very short time using mass spectrometry or NGS platform that does not mean that the data quality is very high, you have to be very cautious, very clear that what data you are analyzing, and make sure that you do the proper ways of normalization, batch correction as well as the missing value imputation before you start further data processing.

In this way, the batch correction removes the technical differences in the data, whereas missing values need to be imputed to get the better outcome. Sometime, you do not take a call if there are many missing values in your data set, then probably that is not a good data, you need to trash the data, and it starts all over. Or let us say you know if you have seen only very few places the data points are missing, but rest of the data is there then software can utilize some resources, some you know ways of averaging and imputes the values to fill out those missing values.

Again there is lot of statistics and considerations required, what should be the way to do the missing value imputation, but that is what Dr. Mani is going to talk to you today in his lecture. He will also explain and talk to you about batch collection methods like limma and ComBat. On the other hand, he will discuss about the missing value imputation which is one of the very important considerations in the big data analysis. So, let us welcome Dr. Mani again for his lecture on batch correction and missing values imputation.

So, the next topic in this set of slide is a batch correction. So, we just talked about normalization, where we are trying to make all the samples similar. So, I do not know how many of you have heard of batch correction, just a few. So, batch correction is something you apply when you think you have different batches of experiments you did, and you think there might be a difference between those batches. So, you have 15 TMT experiments you need to do. You did 5 in January, 5 in May and the remaining 5 in winter, and you this is all one project, and you had lot of vacation left over, so you were not there to finish all of them.

So, now, you want to put all of them together, and do an analysis. So, it is likely that the data you get from each set of 5 are going to be very different or significantly different, and that is because of technical variation, not because of biological differences in the sets of samples you used. So, correcting for that is called batch correction.

So, let us say you have a scenario where you are looking at breast cancer, you got you have ER positive, ER negative and HER2 positive samples. So, you run all your ER positive samples in January, you run all your HER-2 positive samples in May, and then you run all your PR positive samples in December. And now you have three batches again which have differences. And there is big differences between what you did in January, and what you did in May, is it because of the difference between ER and PR, or is it because of the batch? Answer?

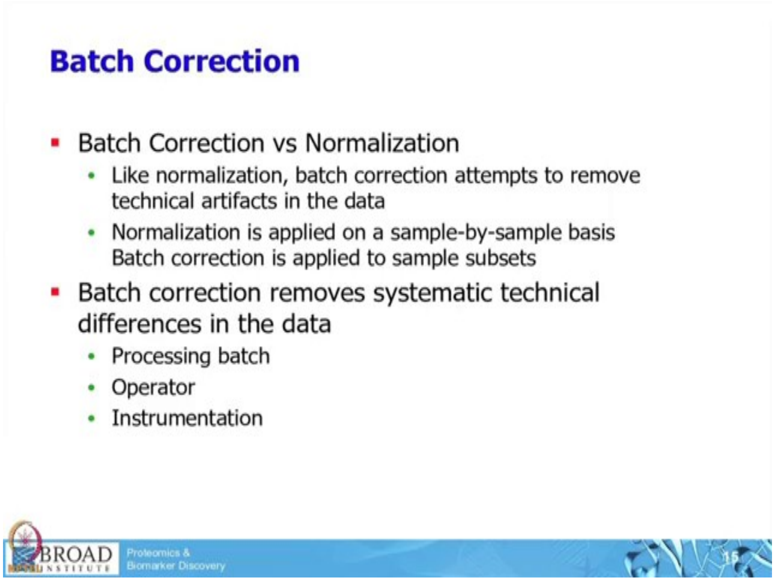
Student: Sir unknown, where we cannot.

Yes, you cannot tell because the way you designed your experiment, we cannot deconvolve, whether it was because of the batch or whether it was because of biological differences. So, that is why statisticians say you always have to run a randomized subset when you are splitting things into batches. So, if you had ER, PR and HER2 positive samples in January you pick a random subset of ER, PR and HER2 positive samples, you put them in your 5, 10plexes and run it. You do the same in May and you do the same in December.

Now, if there is a big difference between January and May, you know it is because of your batch because there are ER, PR and HER2 positive samples in both sets and it is not a biological difference. So, that is why when you run experiments you generally tend to design them, so that you lay them out and you randomize the samples to the degree possible. So, if three people are working on a project, you do not want one person to be working on ER positive samples, another to be working on PR positive samples and so forth.

You want each one to work on a random subset of all the samples. So, any distinction that is of importance to you in the biology that has to be randomized. If you do not, if you have batch effects, you cannot separate the batch effects from the biology. And if you did your experiment correctly and there is a batch effect, then how do you deal with it and that is kind of what I am going to talk about in the next few slides.

(Refer Slide Time 05:51)



Batch Correction

- Batch Correction vs Normalization
  - Like normalization, batch correction attempts to remove technical artifacts in the data
  - Normalization is applied on a sample-by-sample basis Batch correction is applied to sample subsets
- Batch correction removes systematic technical differences in the data
  - Processing batch
  - Operator
  - Instrumentation

So, what I say like normalization batch correction attempts to remove technical artifacts in the data, but normalization is usually on a sample by sample basis, whereas batch correction is on a sample subset basis. So, you have a group of samples that you ran at some point that needs that resulted in some technical difference between some other group.

So, batch correction can be used to remove systematic technical differences, so different operators, different time of running it, different machines, different column you use for your LC things like that. But as long as you have normalise you have randomised your data, you are in good shape. So, you decided to run all your ER positive samples in the beginning of your study. You are doing everything in 1 week, but you decided to run all your ER samples on the first 3 days. And then you are doing your PR samples on the next 3 days. And on day two and a half, your column got clogged and you had to throw it away.

So, now all your ER samples were run with the old column and most of your PR samples were run with the second column. Is there is a difference? Is that the column or is it biology? So, you cannot tell. But if you are alternating ER and PR samples, and there is a big difference you know it is your column, so that is the importance of like randomizing your analysis when you are doing sample processing and experimental analysis.

So, the thing is if you have an internal, if you have a reference that you ran and you try to use that to minimize the difference, it would be for a pilot project but if you try to publish it, the reviewers will not accept it. So, you have to have normal randomized study, and then do some sort of batch correction. You can help the if you have at least same thing you ran across your study, you can maybe try to use it to do better batch correction, but it is not ideal. It is better than nothing but it is not the ideal approach.

If the question is whether the protein was different between type A and type B, so like cancer versus normal, a protein is different you do it 3 times and you see the same difference. So, in that case you probably do not need batch correction, because all you are doing is ratios of cancer to normal in the same batch. But if you are now look if you have 3 replicates your statistics will be much more robust. So, you can combine all the three replicates to find out which proteins are different.

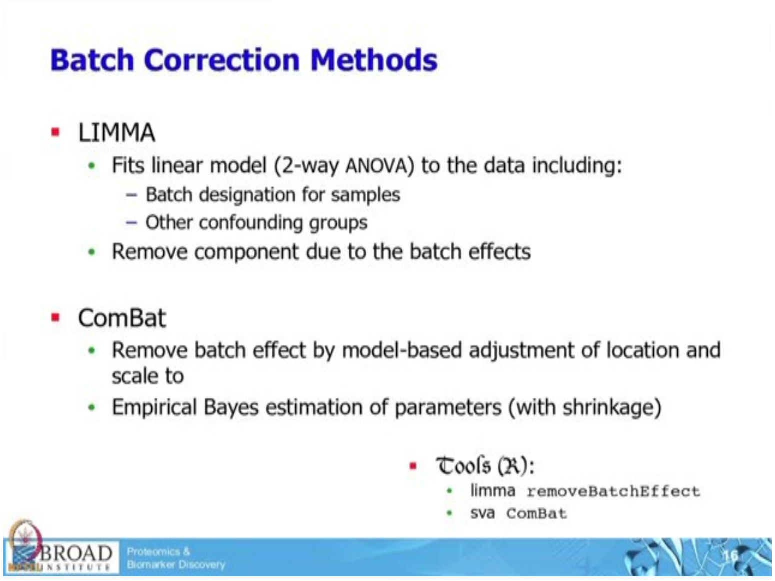
In that case, if there is a big difference between the same protein measurement in your different runs, then your statistics will fail, because there is way too much variation, your p values should be very high. So, in that case, it will help to have batch correction. So, it depends on how you analyze the data. If you are looking at each replicate separately drawing some conclusion and checking that conclusion across, you do not need batch correction. But if you are combining all your replicates and doing a unified statistical analysis then hopefully you would have randomized your replicates. But if you did not, you would do batch correction.

You have only 10 samples that you can put on a TMT 10plex and then you submitted it, and the reviewer came back and said run it 2 more times. So, then you can run it 2 more times, you do not need a reference. But if you are doing a study which requires 15 TMT

10plexes just to do one replicate of the study, then you would have a reference, because you have to link the different 10plexes.

So, if you are I think the rule of thumb base, if your experiment will fit in 1 TMT reaction or 1 TMT experiment, you do not need a reference. But if it is going to span more than 1, you should think about having a reference. In some cases you if there are like 2 or 3 10plexes depending on the conditions and the experiment, maybe you do not need it. But if it is a big discovery study where you have a 100 samples and you need to do like biomarker discovery or proteomic analysis of all those samples then you would have a reference. So, what are methods we have for batch correction? I have a couple of methods listed, I would not go into too much detail here.

(Refer Slide Time 10:09)



Batch Correction Methods

- LIMMA
  - Fits linear model (2-way ANOVA) to the data including:
    - Batch designation for samples
    - Other confounding groups
  - Remove component due to the batch effects
- ComBat
  - Remove batch effect by model-based adjustment of location and scale to
  - Empirical Bayes estimation of parameters (with shrinkage)
- Tools (R):
  - limma `removeBatchEffect`
  - sva `ComBat`

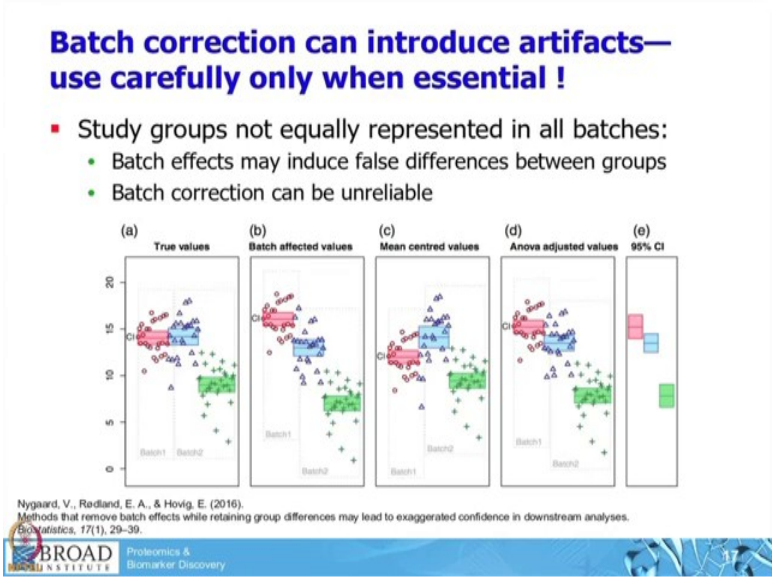
BROAD INSTITUTE Proteomics & Biomarker Discovery 16

But two common ones used are LIMMA is a package that is used for a lot of analysis. You can also use it for like differential marker analysis and so forth. It builds a linear like a 2-way ANOVA model and then uses that to do the batch correction. And there is a R package called LIMMA, and it has a function called remove batch effect. I would not go into the theory or details here, but you can explore that and I think they have some examples and stuff you want to take a look at it. Another option is called ComBat. So, this is a empirical base estimation of how to do the batch correction.

So, Bayesian analysis usually tends to be more robust than traditional frequentist analysis. So, this kind of uses a little more robust method, but this also has mechanisms that make it robust. It uses what is called moderation which is essentially another way of doing empirical Bayesian analysis. So, both these tools are relatively useful, and you can you try both. They have differences that might make it more appropriate for one project or not.

So, suppose you have two batches, but one batch is more sacred than the other one. So, in your batch correction, you want to take the second batch and make it similar to the first one instead of just putting the two together and correcting however. So, if you want to do that, I think ComBat will let you do that, but LIMMA will not let you do that. So, there are differences like that that might dictate which you want to use, but both are reasonable to take a look at.

(Refer Slide Time 11:47)



Batch correction can introduce artifacts—use carefully only when essential !

- Study groups not equally represented in all batches:
  - Batch effects may induce false differences between groups
  - Batch correction can be unreliable

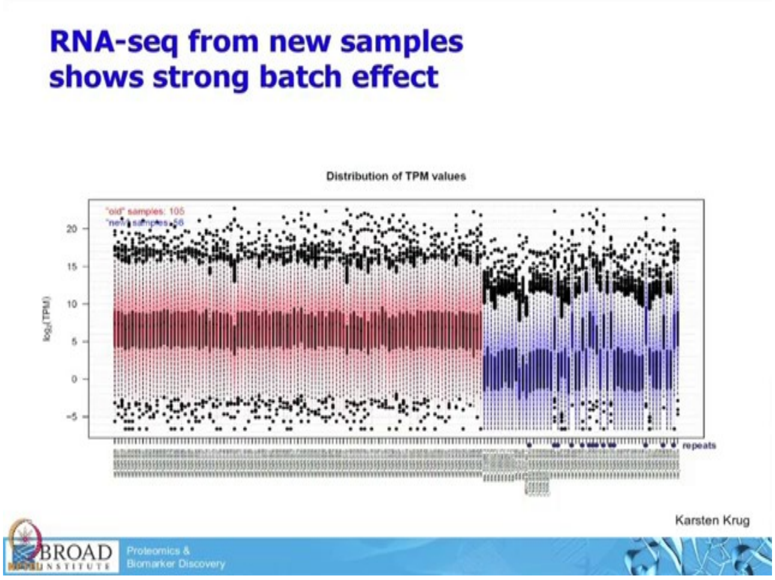
Nygaard, V., Rødland, E. A., & Hovig, E. (2016).
Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses.
Biostatistics, 17(1), 29–39.

BROAD INSTITUTE Proteomics & Biomarker Discovery 17

So, I think I have made this point multiple times. So, batch correction can obviously introduce artifacts. So, because of that I would not do batch correction unless there is proof that batch correction is necessary. So, you look at something and you say this set is very different from that set. If you can show that then you would go and do batch correction; otherwise batch correction can introduce artifacts that might look like signal, that might get rid of signal. So, you can get a lot of false positives and false negatives if you indiscriminately use batch correction.

And one example showing here is what I was saying, the experiment was not properly randomized, and so one batch has more of one type of things than the other batch. And when you do batch correction, it looks like there is a differential signal, because batch correction says, this batch was one this batch was the second batch and I need to go correct for both. And when you do that, there it the differences introduced could look like biological signal.

(Refer Slide Time 12:49)



RNA-seq from new samples
shows strong batch effect

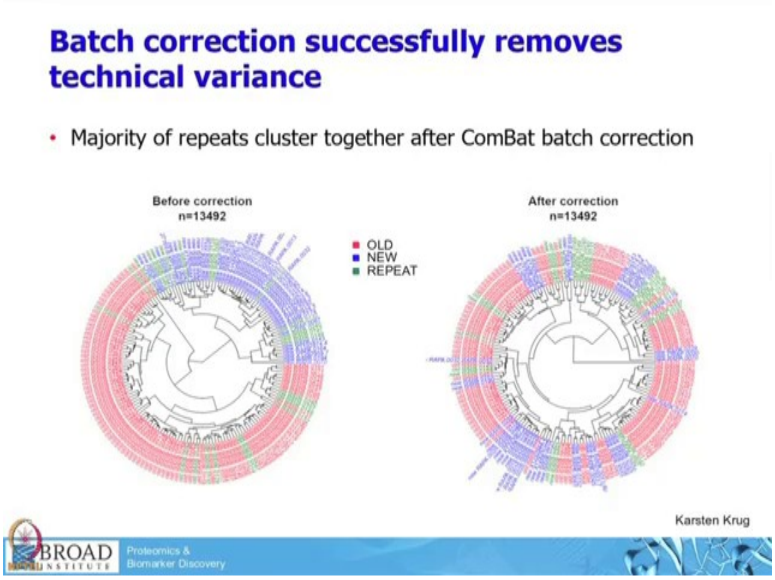
Karsten Krug

So, here is an example that Karsten worked on in our group. So, we had a experiment where there was RNA-seq done from about 105 samples, so those are the red samples. So, you can see this is plotting the expression value for all the genes measured using RNA-seq. This is the box plot. So, you can see where the first batch was all fine. There were one sample that was slightly low, but most of them were fine.

If you look at the second batch, this one was new sets of samples, but it also had replicates from this set. So, things marked with a star underneath are replicates from this batch. So, you take the same sample from this batch and you run it like a year later, you can see how different they are. And so the question is now I want to put these two together and do an analysis, because I got new more samples. When you have more samples, you get more statistical power.

And so how do I do it now? And so for that this is a case where you show there is a batch effect and without correcting the batch effect, you cannot do the analysis. So, we do go and correct it. And once you correct they all look similar. And so here is a example of how that looks similar. So, there the left side shows hierarchical clustering, hierarchical clustering groups similar things together.

(Refer Slide Time 14:13)



Batch correction successfully removes technical variance

- Majority of repeats cluster together after ComBat batch correction

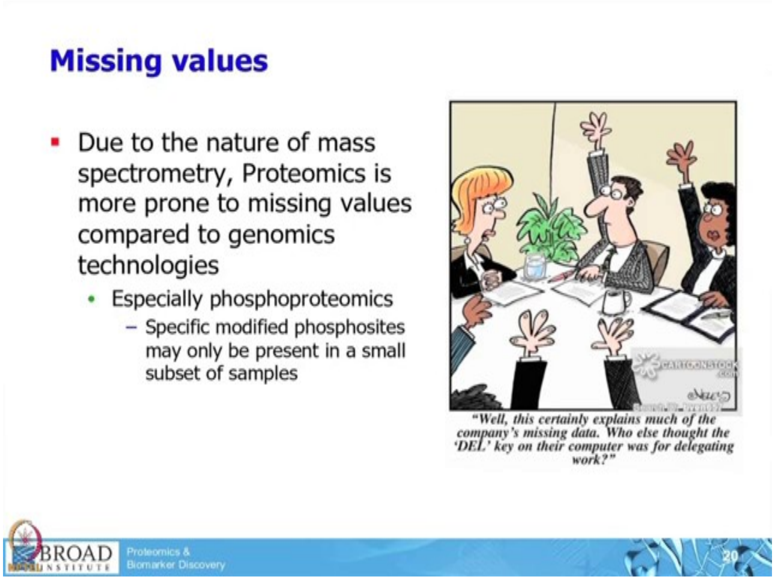
Karsten Krug

BROAD INSTITUTE Proteomics & Biomarker Discovery

So, when you say here are all my samples, Shomi thinks that are similar and put them together. You can see all the blue ones which are the new batch grouped together, all the red ones which is the old batch grouped together. And the green things in the middle are the replicates. So, even the replicates do not go together, they go with the batch. So, this is a very strong batch effect that overrides any biological signal you might have. And so when you correct it you can see now that the reds and the blues are all mixed up and the greens actually line up the replicates line up with each other.

So, now after batch correction a sample and its replicate are the most similar, which is the correct place to be not with things in the same batch. So, we say that the after batch correction the batch effect has been removed, and now you can do the analysis. So, that is kind of an example of where we used it. The next topic is missing values, this one I have a lot of slides, but I think many of them are technical and probably unnecessary. So, I will just zip through it quickly.

(Refer Slide Time 15:23)



Missing values

- Due to the nature of mass spectrometry, Proteomics is more prone to missing values compared to genomics technologies
  - Especially phosphoproteomics
    - Specific modified phosphosites may only be present in a small subset of samples

"Well, this certainly explains much of the company's missing data. Who else thought the 'DEL' key on their computer was for delegating work?"

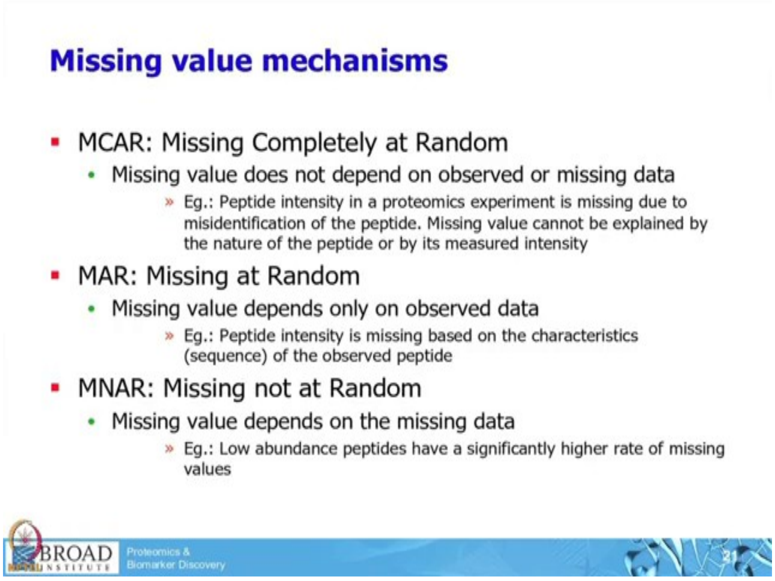
BROAD INSTITUTE Proteomics & Biomarker Discovery 20

So, mass spectrometry is prone to more missing values than RNA-seq or most of the genomics methodologies. So, in like 10-15 years ago, if you used an affymetrix microarray to measure a sample, you would get a measurement for every single gene there would basically be no missing values. But in proteomics that there is it is not possible, because proteomics is stochastic. You are measuring things that fly in a mass spectrometer and what flies ones may not fly again, what flies in one sample may not fly for another sample. And so you are you tend to see a lot more missing values in proteomics.

To make this worse when you are doing phosphoproteomics which is really the interesting part of proteomics because you can look at signalling and kinases and how they work. The phosphosites may be present in one sample and completely absent in another. So, maybe some pathway is activated in a subset of your cancer samples, but is not activated in all your other samples. So, the phosphosites that represent the activation of that pathway will be present in like some small subset of your samples, but not present in all your other samples and so you get a lot more missing values in phosphoproteomics.

When you try to do statistical and machine learning analysis many of the tools need all the data to work. If you have missing values, the algorithm cannot be applied. And so the question is how do you deal with missing values. So, in statistical theory, there are three types of missing values in increasing order of I guess a worry index.

(Refer Slide Time 17:05)



Missing value mechanisms

- MCAR: Missing Completely at Random
  - Missing value does not depend on observed or missing data
    - Eg.: Peptide intensity in a proteomics experiment is missing due to misidentification of the peptide. Missing value cannot be explained by the nature of the peptide or by its measured intensity
- MAR: Missing at Random
  - Missing value depends only on observed data
    - Eg.: Peptide intensity is missing based on the characteristics (sequence) of the observed peptide
- MNAR: Missing not at Random
  - Missing value depends on the missing data
    - Eg.: Low abundance peptides have a significantly higher rate of missing values

if you are in the first one you do not need to worry about it, you can throw away the data and you will be fine, so that is called missing completely at random. So, randomness is very important. If your data is missing in such a way that it the things that are missing are completely random then they are not going to affect your biology. And so you can say fine I do not care, I am going to throw it away, I have 200 proteins that are completely missing at random and I can just ignore it.

The second part missing at random is that the missing value depends on some part of the observed data. So, for an example here is peptide intensity is missing based on characteristics of the peptide. So, if the peptide has a specific amino acid it is missing, so here so if the if the peptide has a specific amino acid at random times it is missing, it is not always missing. So, there you know the characteristic and once you know the characteristic, then the missingness is random, because which sample its missing is random.

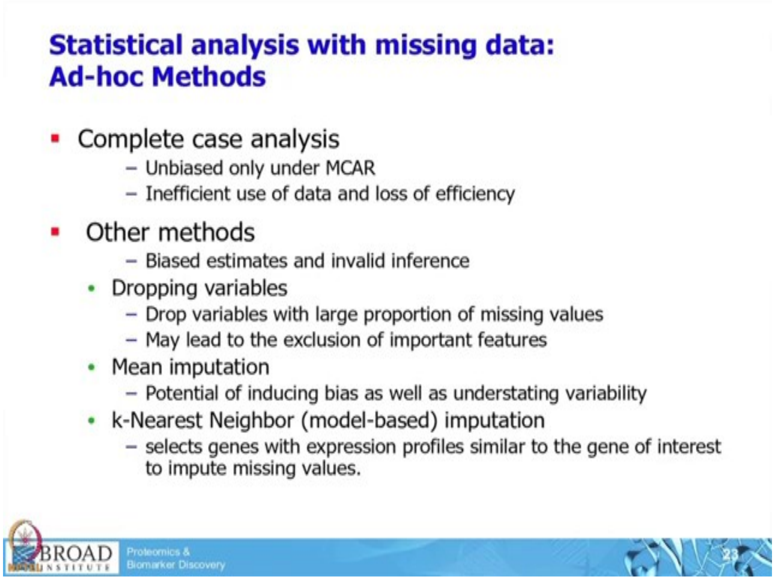
But if the sample has a specific amino acid in that peptide, then it is likely that it might be missing, but whether it is actually missing or not is random. So, in that case it is called missing at random. So, these things you have to be a little more careful you cannot just like throw it away, but the most worrisome part is missing not at random and unfortunately all proteomics data, most proteomics data is missing not at random.

Missing not at random means the missing value depends on the missing data; in other words it is missing when the value is some specific thing. If your intensity is less than some count, the data is missing so that is missing not at random. And so that is the most worrisome type of data and there you have to be very careful on how you deal with it, you cannot just like casually throw it away, but it is also the hardest to deal with even in with statistical theory. So, I will just go through like a pragmatic approach to how to deal with this, I will not go into all the details of what we do, we do not need that.

So, here are some approaches that people use to deal with missing data. So, one is called complete case analysis. So, this is you throw away any data that is missing and then you do your analysis. So, if a protein was observed in 95 percent of your samples, but missing in the 5 percent you throw it away and then you do your analysis, so this loses a lot of data. And in phosphoproteomics this may not be a good thing, because you will lose the loss a lot of your signalling peptides, you may lose a lot of your biomarkers too in even in the proteomics data.

So, you want some other method. So, one way is throwing away things that have a large proportion of missing values, so that may be reasonable. So, in other words if you have something that is missing in more than 80 percent of your data, you can say that is way too much I do not want to deal with it. So, you if a protein is missing in more than 80 percent of your data, you do not look at it that makes sense, because statistically there is not much you can derive from looking at a small set and also you do not know how it is different in the others.

(Refer Slide Time 20:29)



Statistical analysis with missing data: Ad-hoc Methods

- Complete case analysis
  - Unbiased only under MCAR
  - Inefficient use of data and loss of efficiency
- Other methods
  - Biased estimates and invalid inference
  - Dropping variables
    - Drop variables with large proportion of missing values
    - May lead to the exclusion of important features
  - Mean imputation
    - Potential of inducing bias as well as understating variability
  - k-Nearest Neighbor (model-based) imputation
    - selects genes with expression profiles similar to the gene of interest to impute missing values.

BROAD INSTITUTE Proteomics & Biomarker Discovery 23

And so it may be to drop variables that are missing. So, in the analysis we do we have a relatively high threshold, I think if the data is missing in more than 70 percentage of the samples, then we throw that away. So, we take a protein if that protein has not been observed in 70 percent or more of the samples, we throw it away. So, if we start with 12000 proteins and we apply that filter about a few 100 the maximum 1000 proteins will get thrown away, so it is about like 5 to 10 percent of proteins generally fall into that category. If you look at phosphoproteomics the number is larger, it is more like 25, 30 percent, but still it is not that big number and it is a reasonable thing to do in most situations.

So, the other thing to do is impute the missing data. So, you can say I do not want to throw away things that have missing data my analysis method needs the data, so what can I do. So, one thing you can do is you can say, I am going to carefully figure out what the value could have been or should have been and then fill it in, so that is called imputation. So, you can impute missing data and then complete your data set for that there are several methods the best ones use some sort of a machine learning model. So, they look at all your data, they look at things have been observed for the thing that is missing and then they kind of predict the missing value should have been this by looking at the entire data set. So, they builds a machine learning model by looking at the entire data set and then comes up with imputed value for data that is missing.

So, we have found that that there is this method called k-nearest neighbour imputation that works reasonably well there are a couple of other machine learning based methods that also work well. So, if you really have to fill in, so the way we do our analysis we will we start with the full table, we apply the threshold if it is missing in too many samples. We just throw it away and then for the remaining we use at k-nearest neighbour imputation to fill it in to do the analysis.

There are some analysis methods that can actually deal with missing data in those cases we do not impute, but if you are using a method that cannot deal with missing data we impute using KNN or some machine learning based method. So, before you do a marker analysis you should not use the information of the classes for making any decision, because if you do that then you are biasing your analysis to the groups that you know and you should not do that. So you would look at the entire data set.

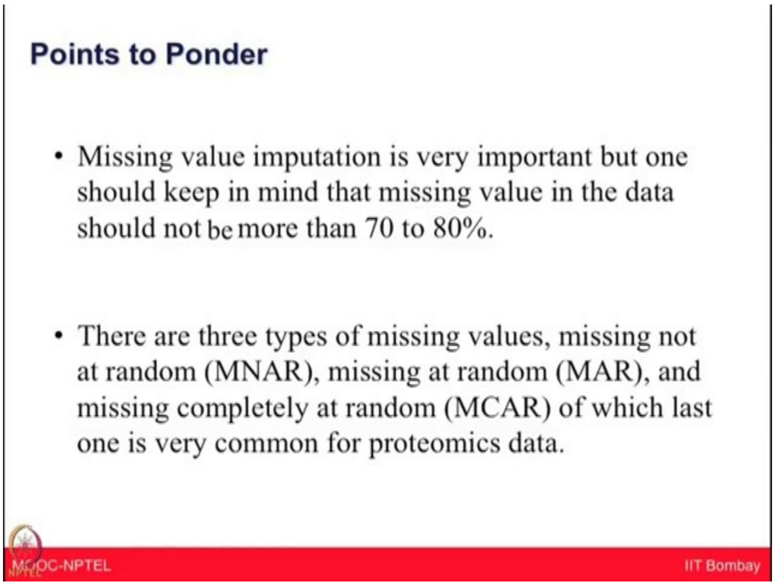
Let us say your perfect biomarker is present in your cancer and not in the controls. And so your you have 50 percent cancer, 50 percent controls; if you set this threshold to be 40 percent, you will throw away all your biomarkers. So, you would set the number to be high enough, so let us say you set it to be 70 or 80 percent. So, then you retain things that are present in only a few of the cancers, but are missing in all the normal's and maybe even a few cancers, then you will keep those markers.

When you do k-nearest neighbour implementation remember it is called k-nearest neighbour, so what it does is it is going to say ok, I need to fill in this protein for this sample, I am going to find samples that are similar and find what values they have and then fill it in. So, it is going to look at other control samples and then fill it in, but if there are if it was missing in all the control samples, then the imputed value will not be that that clear. But if some had some small values because of noise or something like that then it will just average the noise and create a value that will fall into the noise, but if it was missing in the tumors, then when you look for similar samples you to find more of tumors and then it will fill in with the value that is specific to the tumors.

But so this algorithm will actually correctly take group into account without knowing about the group, it does not have to be actual technical replicates, but you need biological replicates for.

If you had two samples and it was missing in one, there is not much you can do to fill it in the other, but if you had 100 samples and it was missing in 5, you can use the other 95 to fill it in.

(Refer Slide Time 24:49)

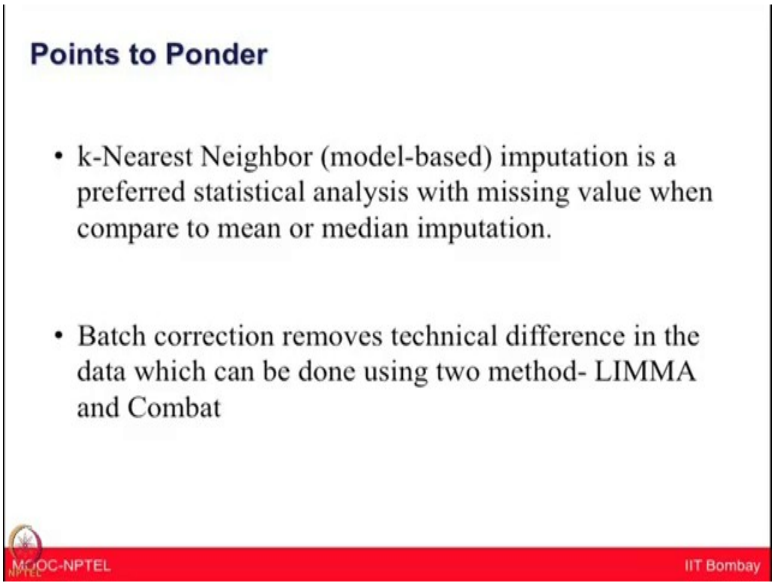


Points to Ponder

- Missing value imputation is very important but one should keep in mind that missing value in the data should not be more than 70 to 80%.
- There are three types of missing values, missing not at random (MNAR), missing at random (MAR), and missing completely at random (MCAR) of which last one is very common for proteomics data.

MOOC-NPTEL IIT Bombay

(Refer Slide Time 25:01)



Points to Ponder

- k-Nearest Neighbor (model-based) imputation is a preferred statistical analysis with missing value when compare to mean or median imputation.
- Batch correction removes technical difference in the data which can be done using two method- LIMMA and Combat

MOOC-NPTEL IIT Bombay

So, I hope today you have learnt about what is batch correction and how to perform this kind of analysis, you also learned the important strategies of batch correction which is LIMMA. The design matrix is used to describe comparisons between the samples. For example, the treatment effects we should not be removed, the function in effect fits a linear model to the data which includes both batches and regular treatments, then removes the component due to the batch effects.

Another method for batch correction is combat which is robust, as it can do a model-based adjustment to remove the artifacts. The batch correction also need to be done correctly, otherwise a wrong strategy may lead you to the artifacts. So, this important thing is the missing value imputation which you have heard, this is you know one of the common facts; people see in the various experiments, especially the mass spectrometry based data generation whereas, some time you know some values you cannot see for every single protein and what should be the considerations to impute these missing values is very important.

Again as I mentioned in the beginning that you need to ensure that missing values are not too much in your data set, especially not more than 70 to 80 percent, otherwise the data is not real you should try to are not use the data set at all. If there are only very few missing data points, then you can utilize the missing value imputation strategy to try to recover that information. The different type of missing values like Missing Not At Random – MNAR or Missing At Random – MAR as well as Missing Completely At Random or MCAR of which the last one MCAR is commonly used for the proteomic data set analysis.

You also heard the k nearest neighbour model-based imputation which is a preferred statistical analysis with missing values, when compared to the mean or median population. We will continue our discussion about different strategies employed for data analysis and the lecture will be continued by Dr. D.R. Mani in the next lecture.

Thank you.