Interactomics: Basics and Applications
Prof. Sanjeeva Srivastava
Prof. Santhosh Noronha
Department of Biosciences and Bioengineering
Indian Institute of Technology, Bombay

Lecture - 55
Basics of Statistical Analysis-II

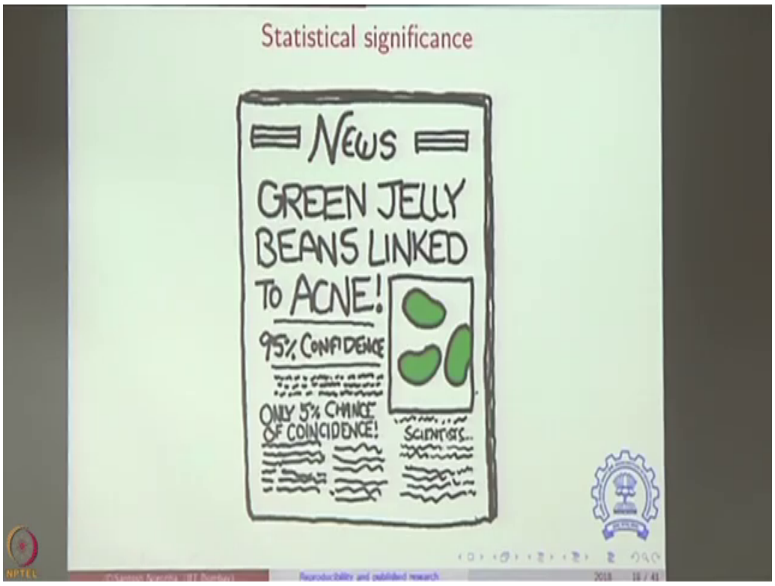
Today, Prof. Santhosh Noronha from IIT, Bombay will continue his lecture about considerations of data analysis especially for the omics data sets. Today's lecture is going to be about why basic understanding of data analysis is required. For example, 0.5 percent accepted error rate in significance used without basic understanding of data may result in false interpretations. Prof. Noronha will also talk about the importance of replicates and how one should choose controls which are usually one of the very important samples for the big data or the omics based experiments.

So, again thinking about a good experimental design what should be replicates, what should be your strategy for data analysis. Actually, determine the meaningful sense of your experiments despite all the advancements in these technologies and the pace at which we could generate the data. But, it is still getting meaningful data is not a straightforward, it is not easy.

So, I hope today's lecture and based on the previous lecture; these two lectures will illuminate your knowledge and give you the concepts about good experimental designing and what should be the considerations to look for to get the meaningful insights from your data. So, let us welcome Santosh Noronha.

The systematically tested so many possible candidates for significance and if there was a 5 percent error rate in your analysis. You would have randomly found a candidate and called it significant. And, we end up fixating all our energies on these one or two candidates that we get when it is sure randomness that has caused these two turn ups. So, unless you have an independent way of carrying out an analysis with these candidates and validating are they important to you it is kind of pointless proceeding further.

(Refer Slide Time: 02:29)



Statistical significance

NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

©Santosh Noronha (IIT Bombay) Reproducibility and published research 2018 18 / 41

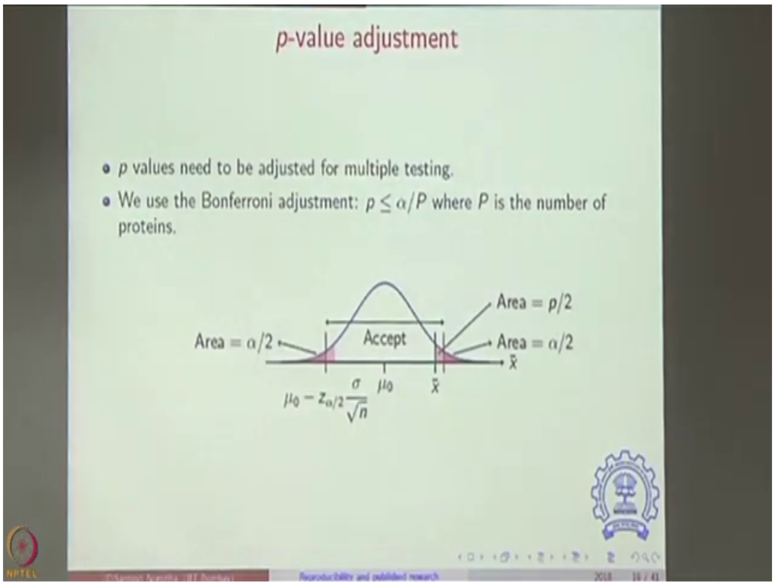
(Refer Slide Time: 02:34)

Statistical significance

WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).

Now, at this point it for if you are in the publishing game it is very important that you notice that publications, do not allow you to publish negative results. So, all these other things you cannot publish here. So, the only thing you can publish is this particular result. So, there is pressure on you to find that needle in a haystack as a positive result and publish it and that is the nature of those confirmation bias and pee hacking which pushes you into now focusing entirely your research on this particular candidate the green candidate as if it were the only relevant candidate.

(Refer Slide Time: 03:03)



p-value adjustment

- *p* values need to be adjusted for multiple testing.
- We use the Bonferroni adjustment: $p \leq \alpha/P$ where P is the number of proteins.

Area = $\alpha/2$
Accept
Area = $p/2$
Area = $\alpha/2$
$\bar{x}$
$\mu_0 - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$
μ_0
$\bar{x}$
NPTEL

©Santosh Noronha (IIT Bombay) Reproducibility and published research 2018 19 / 41

So, what is the way around this? So, again something that we typically do not do is an adjustment to this. So, the only way to if there is a 5 percent error rate in analysis and 5 percent error rate is a dangerous thing if you are doing a10000 studies I mean; I want you to appreciate this 5 percent at a different level.

You take any 100 papers published which are scientific hypothesis being tested and I can tell you without even reading those papers that 5 of them have got to be false. Because all of them have used a 95 percent confidence level for executing this analysis. And, if you are saying there is a 5 percent sheer bad luck error rate, then several of those researchers have suffered unfortunately with the randomness in the data they collected which means their results are going to be false.

It is not that they have set out to cheat. It is unfortunate given that they are unable to reproduce their own data and they are trying there is no rush to publish. So, the trick the way to control for others in some ways. So, how do I reduce my error rate? So, if my error rate is the red portion, the 5 percent. If I want to reduce my error rate I have to move my goalpost further out and that is only solution.

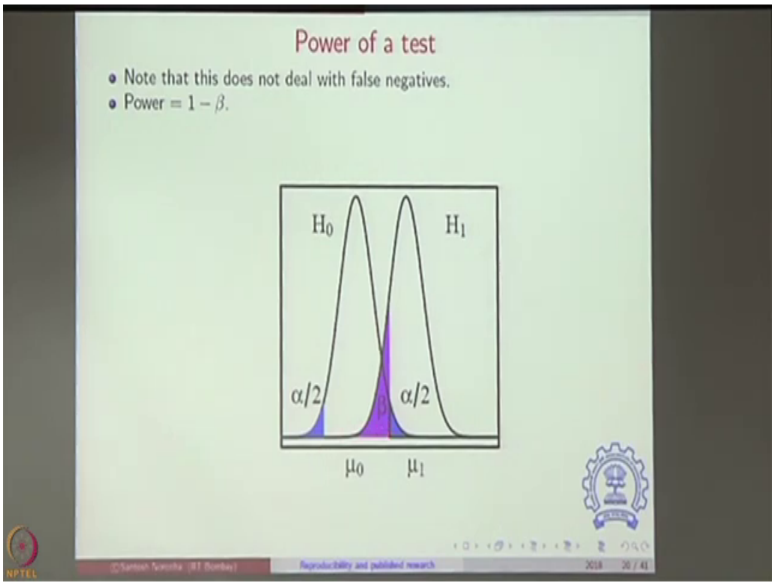
Of course, now it gets hard. The number of candidates you will get which passed this goalpost further out which are so extreme; what you are saying is your results must be so extreme that they are well outside these wider goalposts range. So, what do you say if you are going to do10000 tests. Each test should not have been done at a 5 percent level, instead each test should have been done at a 5 divided by the 1000 or10000 if I am doing10000 tests.

So, that 5 percent error rate should be spread across the10000 tests that you are doing. And if you were to attempt that this area now is 5 over 10000. So, it is become a tiny area, but I have effectively pushed my goal posts out. So, the odds of now passing my tests are much much lower and the odds of randomly passing my test have gone down. And, that is a core trick to the statistical analysis it is called a Bonferroni adjustment.

And good packages, software packages for omics testing will have this as a setting where you can correct for the number of tests that you are doing and try to refine this. And, it is a critical thing so in other words one of the things you ought not to be doing in omics framework where statistical tools are being provided to you by the manufacturer is user default setting in a workflow.

You got to ask the question, what are the settings ok which control for statistical significance and do these need to be tweaked to correct for the number of studies you propose to do on that software.

(Refer Slide Time: 05:56)



Power of a test

- Note that this does not deal with false negatives.
- Power $= 1 - \beta$.

H_0
H_1
$\alpha/2$
β
$\alpha/2$
μ_0
μ_1
©Santosh Noronha (IIT Bombay)
Reproducibility and published research
2018
30 / 41
NPTEL

There is this aspect of power of a test and what I want you to appreciate is while all the emphasis on asking is a genetic candidate is a gene candidate significant or not. All of this involves in only looking at this particular curve. So, if you look at this particular curve forget the other curve for a moment. If you are looking at this particular curve then your 95 percent confidence leaves out this blue area on either side that is what 5 percent, the blue area would have been 5 percent.

But for the sake of argument I am going to I pretend as if the reality was some of the hypothesis. Under which this would have been a mean value and this would have been a range of outcomes I would have seen if some other hypothesis were true and that is just for the sake of argument because now you will see something problematic happen.

If this hypothesis were true, then this blue area corresponds to that percentage of time you are going to get your hypothesis outcome wrong. So, that 5 percent of the time we are getting a hypothesis outcome wrong under the null hypothesis. What do these thresholds mean for you? Within these thresholds you say this hypothesis is ok. I am in agreement with that hypothesis. Outside that those thresholds you end up saying I do not believe in this hypothesis I will go with the other hypothesis in this case H 1.

If you now look at H 1, H 1 is allowed to be true only from this coordinate to the right beyond that region you believe in H 1 to the left you are you have already argued you prefer to go with H naught as a hypothesis. But do you now see under H 1, this area in purple corresponds to an error where H 1 could have been the true hypothesis, you have gone with h naught. Therefore, as a technicality you are committing a mistake by saying H naught is true when H 1 should have been true. So, there is a mistake there is a mistake it is just like false positives and false negatives.

In fact, it is related to the concept of false positives and false negatives. You will make one mistake or the other. If you were to create a diagnostic kit and you are going to change the threshold for detection of a particular measurement in trying to cut down the false positives, just think about this if I take this threshold and I move it to the right. If I take this threshold and I move it to the right under the H naught curve, what will happen to the blue area? The blue area goes down.

I commit less of a mistake with respect to my original hypothesis. But if I move this coordinate to the right, what happens to the purple area? The purple area grows. If you are trying to minimize false positives in your analysis, you run the risk of increasing false negatives and vice versa ok. So, that is the key issue. So, the headache comes about because if you look at what we have done in the previous thing. We only paid attention to one curve we did not ask the question what might the other hypothesis behave like which is the case here ok.

So, if you start paying attention to alternate hypothesis you suddenly realize that; yes I might have a diagnostic kit. For example, which is accurate 95 percent of the time that is what that confidence in told tells you. But what it and therefore, a 5 percent of the time I am making a mistake of a certain kind, let us say I am falsely calling somebody positive, so false positives.

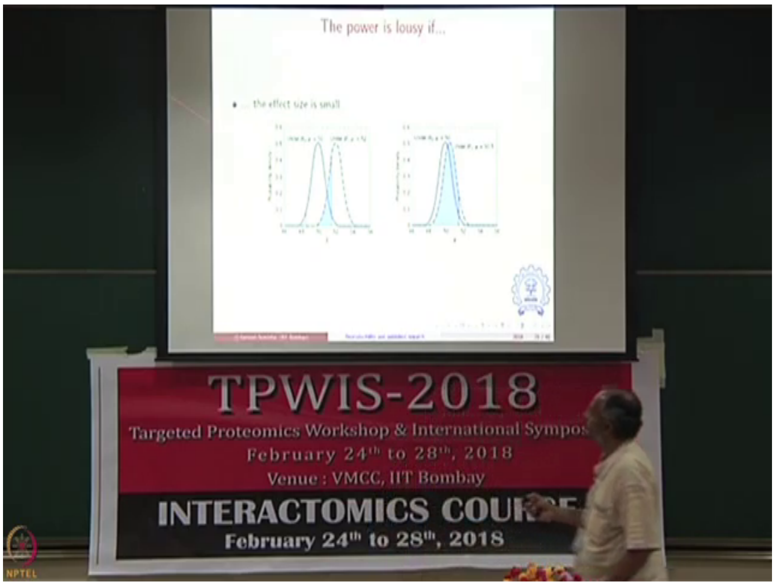
But what is not giving me information at all is, what is my false negative rate? And for the false negative rate you ought to be looking at the other curve and this beta. So, in other words you want beta to be small, you want alpha to be small, you want beta to be small. 1minus beta is called the power of a test and it is a good practice to ask whenever you claim that something was a significant candidate. This is significant target do not just tell me how significant was that result. Tell me how powerful that test was.

In other words tell me what is this value beta that I might therefore, actually have a lousy test. So, this power of a test is a concept which most again is there it is buried somewhere in software typically as an option for you to report. But it is not something researchers are in the habit of reporting. So, when somebody tells me I have found a candidate and. In fact, I found a short list of candidates which are all significant.

What they are not telling me is, how powerful was the analysis, what you are not telling me therefore, is what was the probability that they have got the analysis wrong, the other way around. While they are telling me that they are confident within 95 percent. What they are not telling me is whether this was so bad that this was 20 percent or 30 percent or 40 percent, the purple area.

If one of these values is greater than 10 percent your study your analysis is already in trouble. So, both alpha and beta, both these shaded areas cannot be large because they are both errors in your interpretation.

(Refer Slide Time: 10:39)



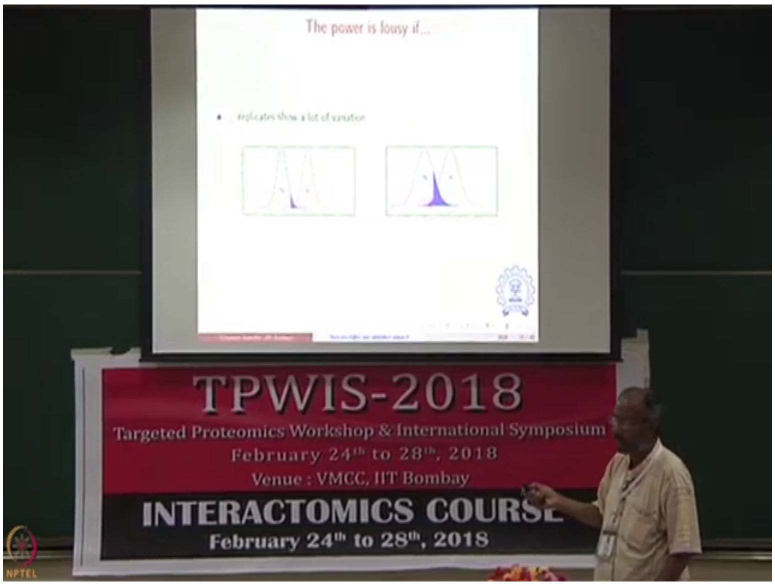
The power is lousy if...

- ... the effect size is small

TPWIS-2018
Targeted Proteomics Workshop & International Sympos
February 24th to 28th, 2018
Venue : VMCC, IIT Bombay
INTERACTOMICS COUR
February 24th to 28th, 2018
NPTEL

And, here are cartoons which quickly improve the point. So, if you are trying to distinguish two hypothesis and your two hypothesis are so similar to each other. Therefore, the effect size is small, you will have such an overlap in the to the predictions coming out of the two hypothesis that they are unable to discriminate and say which hypothesis is true. You will not be able to do that ok.

(Refer Slide Time: 11:01)



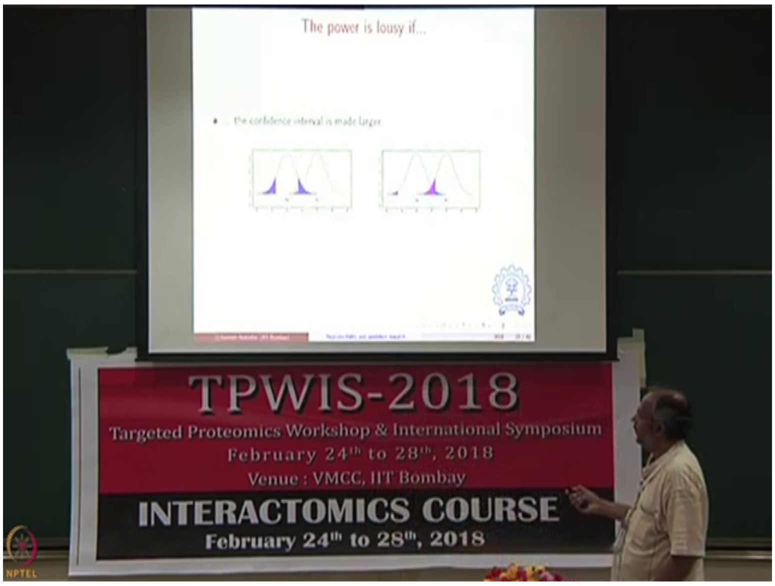
The power is lousy if...

- replicates show a lot of variation

TPWIS-2018
Targeted Proteomics Workshop & International Symposium
February 24th to 28th, 2018
Venue : VMCC, IIT Bombay
INTERACTOMICS COURSE
February 24th to 28th, 2018
NPTEL

If you have a bunch of replicates, if you have a bunch of replicates; you, I am not getting into the math of this, but your curves get thinner. If your curves are thinner, so the spread in here is thin. If your curves are thinner the overlap is reduced compared to here. So, you in a nutshell you want more replicates of any analysis that you do, otherwise your errors are going to be large ok.

(Refer Slide Time: 11:25)



The power is lousy if...

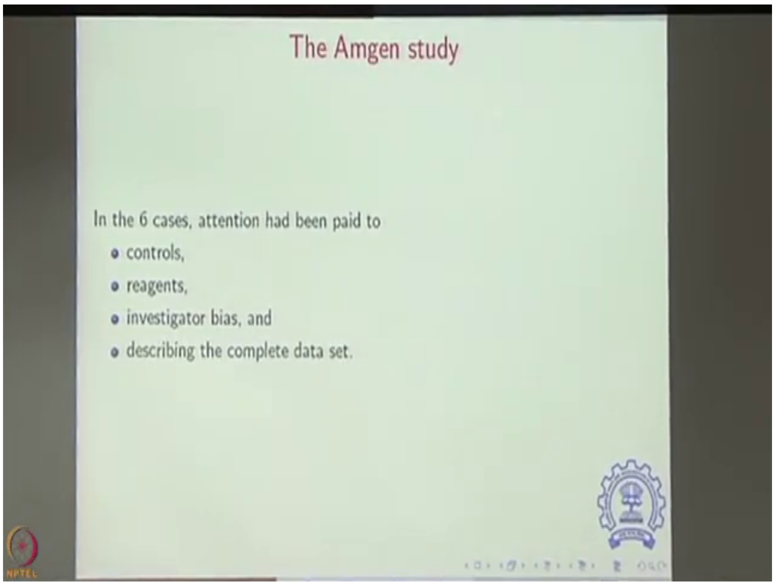
- ... the confidence interval is made larger.

TPWIS-2018
Targeted Proteomics Workshop & International Symposium
February 24th to 28th, 2018
Venue : VMCC, IIT Bombay
INTERACTOMICS COURSE
February 24th to 28th, 2018
NPTEL

If you are going to make a confidence large if somebody says that really not to be talking about 95 percent confidence, you need 99 percent confidence. Then the immediate outcome to you is the moment you move your error sorry your, but your thresholds further out this purple area which was small here has become larger. So, that that was since the case.

So, there is no clean solution here. The moment you try to improve once in situation something else in your analysis is going to worsen somewhere else and my point is to prove that everything is interlinked and you therefore, ought to be talking about significance of a result as well as whether there is a powerful test being done.

(Refer Slide Time: 12:00)



The Amgen study

In the 6 cases, attention had been paid to

- controls,
- reagents,
- investigator bias, and
- describing the complete data set.

This is a famous study and it is really worth looking this up on your own later on. I am jealous one of the two top bar tech companies in the world ok. They make their that is the dominant manufacturer bar pharmaceuticals, protein drugs for the most part and while initially the worked on things which are already discovered in research labs increasingly they have been doing their own research trying to find out what is the next generation of pharmaceuticals that have to be manufactured.

They obviously, keep track of literature. So, one of the things that it was they took 53 landmark papers published in the top journals in oncology and haematology. These are publications coming out of MIT, Caltech, Stanford, Berkeley, the top labs the top universities in the world. 53 and their logic was these are all published in the top journals.

Let us repeat these results in house and if as is published these candidates are good candidates let us get into the business of manufacturing these candidates that that is where they were going ok. So, out of 53, they could reproduce only 6 papers and this is MIT, Caltech, Stanford, you are not talking some small tiny college somewhere. So, what is going on?

It is not that people at MIT and Berkeley and Caltech were cheating, it is not that they were deliberately cheating. But there is a situation where the results coming out of even these top labs cannot be reproduce. So, why do you think they cannot be reproduced ok. In the 6 cases where the results could be reproduced when you look at carefully what happened ok.

Attention had been paid to doing the right controls in the experiment. You need the right controls. You do not make claims about results based on only a test case you do the controls. The reagents were reproducible and this you will realize most of you are doing experiments, my lab experiments reagents, especially in the immunology space are hard to obtain in a reproducible fashion antibodies in particular batch to batch variation exists and your enable to reproduce results.

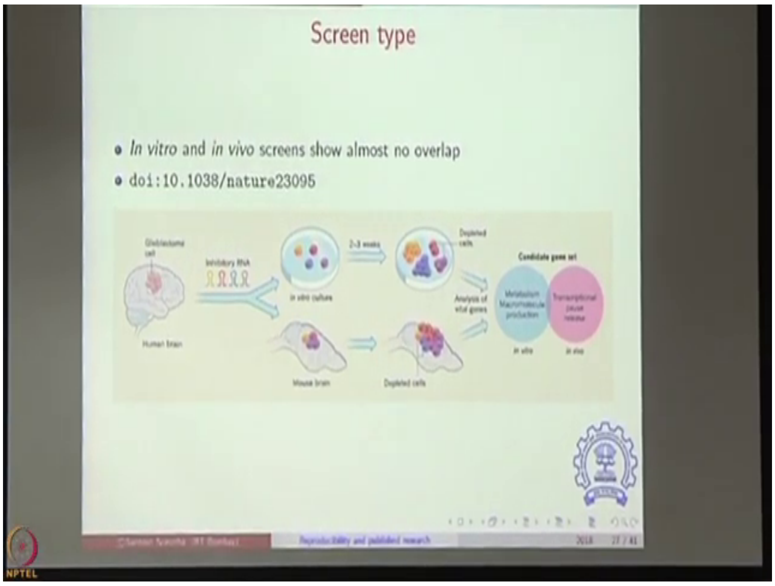
So, in the 6 cases, there was the ability to manufacture these reagents reproducibly and that made a huge difference ok. The investigators were not biased, they did not they were not trying to push for a particular insight or an outcome and importantly they were honest about reporting all the data.

So, you remember that straight line plot where I deleted the mid section of the data and then you claim a better result than it actually is. They were honest enough to claim all or at least report all their data which meant that is when somebody tried to reproduce it they also saw some bad data equally bad to what these guys had found.

It is a surprising result. It tells you to what extent there is pressure on people to publish positive insights even of the top labs. And the moment this study came out and when this pharma company published this insight ok. Many companies started paying suits.

So, buyer did a similar study. They looked at 67 targets published in the literature bias another big pharma major and out of 67, they could reproduce 14 results which tells you this is a serious problem.

(Refer Slide Time: 15:33)



Screen type

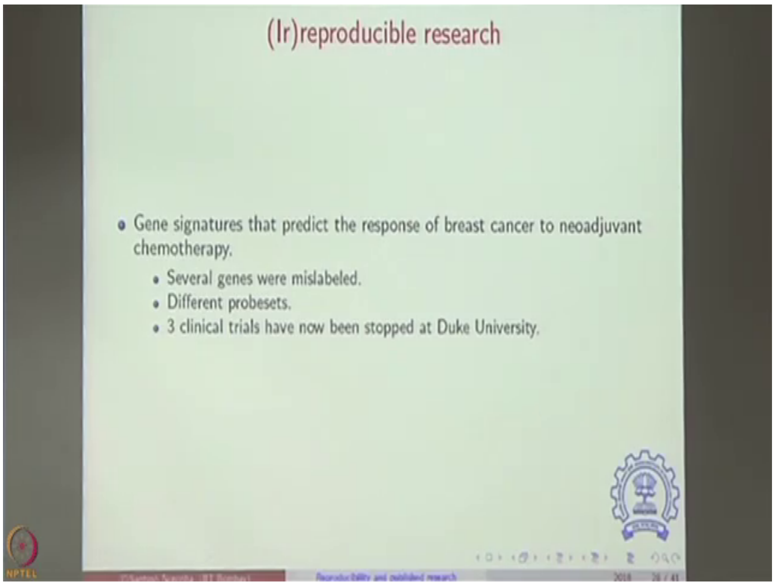
- *In vitro* and *in vivo* screens show almost no overlap
- doi:10.1038/nature23095

Now, this problem goes beyond just statistics alone. So, you can argue that a lot of that is bad luck with data not being reproducible because of the one time you do this experiment with that one material and inherently it is not as a reproducible experiment ok. But it also reflects other aspects of poor design.

So, here is one experiment of poor design where you are screening for certain drugs to do with epigenetic control in globba, in a globba blastoma system and the screen which was done was an in vitro screen. So, there is two ways finally, this got done and in vitro screened using you are basically relying on RNA interference kind of protocol to try to identify targets.

And the in vivo screen; where this is the in vivo screen where you are directly loading these cells onto the brain and then looking for changes in function. The in vivo screen and the in vitro screen have practically no overlap in terms of what is up regulated and what is down regulate ok. Which means, if you had just done the in vitro experiment and you are generated a bunch of targets and you are then proposed to now design drug candidates against these targets. You are a wasting huge amount of money ok.

(Refer Slide Time: 16:43)



(Ir)reproducible research

- Gene signatures that predict the response of breast cancer to neoadjuvant chemotherapy.
  - Several genes were mislabeled.
  - Different probesets.
  - 3 clinical trials have now been stopped at Duke University.

©Santosh Noronha (IIT Bombay) Reproducibility and published research 2018 28 / 41

There have been several such things you start going through literature you will see these things you know it is not something conventionally with that journals published. So, this is published elsewhere these some of these ironically are published in blogs they are not published (Refer Time: 16:53) for example, been studies on looking at gene signatures predicting the response of a breast cancer to chemotherapy ok.

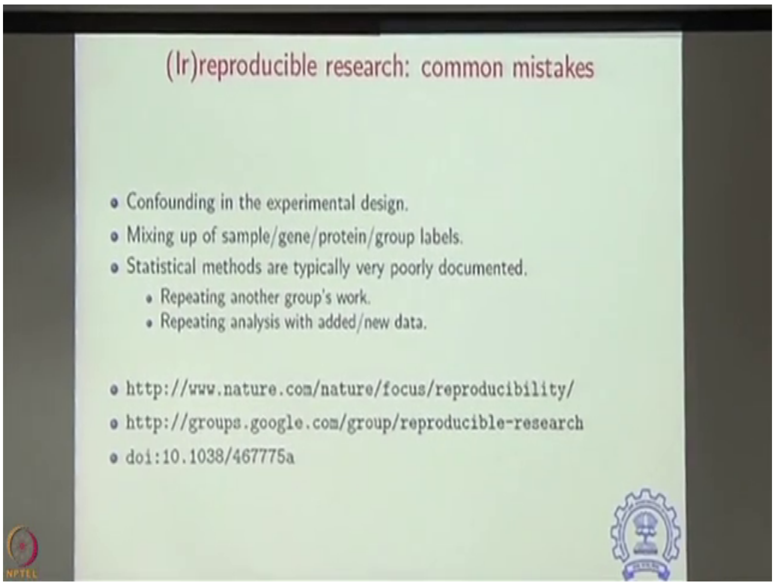
And in this case, the problems are even more ridiculous. In this case and this is another strategic problem with handling large datasets, one of the things that happen here is when the research and they finally, traced it to a student this was a duke university when this data set omics data set was finally, taken into a spreadsheet and subsequent analysis was done and this data set was sorted many columns got sorted one column did not get sorted.

And now every gene is being assigned or all these numbers are being assigned to the wrong gene labels gene ids. This was one mistake which happened very early on in a rush to carry outs this analysis. Nobody followed that up and it went through an entire analytics pipeline ok. Bio informatics drug candidates were created probe sets were created ok. Three clinical trials were started on human patients on this basis and a huge amount of money was spent by NIH and running to clinical trials you will appreciate a billion dollar experiment sometimes ok. And millions of dollars later in this case because this was an early clinical trial, early stage clinical trial, millions of dollars later.

When the duke researchers went back to NIH and said our results are not reproducible. These candidates despite the by informatics do not seem to work in reality and the hard question got asked why show your lab notebooks you go all the way back and you look at the printouts of the spreadsheets and suddenly realize one column has not been shuffled and sorted. And comes down to a simple it mistake we just wasted a lot of time and money these are clinical trials.

It could have been worse, if people had died as a consequence of being seriously hurt as a consequence of the trial because you are actually playing around with therapies or proposed therapies you could have been much much worse in terms of how this sort of hurt the university and the researchers.

(Refer Slide Time: 19:20)



(Ir)reproducible research: common mistakes

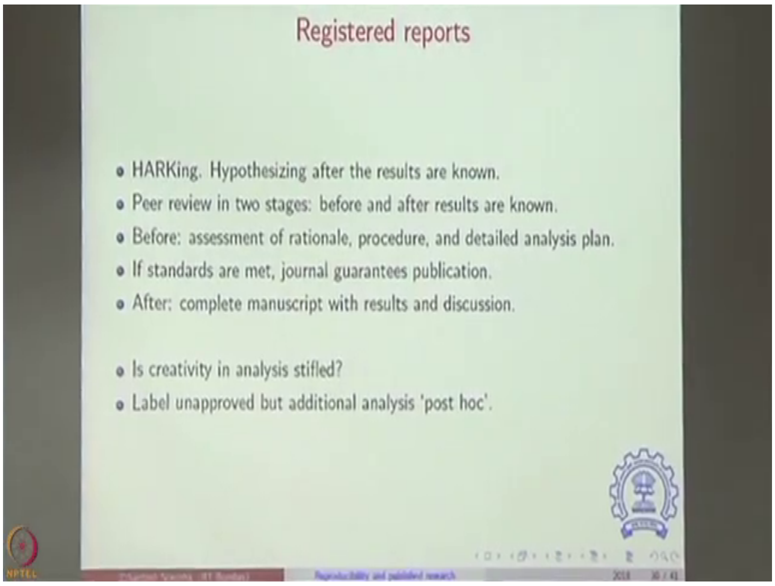
- Confounding in the experimental design.
- Mixing up of sample/gene/protein/group labels.
- Statistical methods are typically very poorly documented.
  - Repeating another group's work.
  - Repeating analysis with added/new data.

- http://www.nature.com/nature/focus/reproducibility/
- http://groups.google.com/group/reproducible-research
- doi:10.1038/467775a

So, I am giving you a bunch of links here really what I wanted to appreciate is leaving data analysis and statistics as an afterthought to a bio informatics pipeline is a dangerous business. You remember that phrase I came up with hypothesis after results are known. There is this philosophy that getting the data is a hard thing; therefore, all the effort goes into getting the data and once you get the data you say you will actually get down to doing the science.

But, really it should have been the other way around they should having a robust experimental method and then computational method identified before the experiment was even done and then you report whatever results you get as per that methodology. In fact, this is the whole publishing paradigm in the omics space is not going to change.

(Refer Slide Time: 20:02)



Registered reports

- HARKing. Hypothesizing after the results are known.
- Peer review in two stages: before and after results are known.
- Before: assessment of rationale, procedure, and detailed analysis plan.
- If standards are met, journal guarantees publication.
- After: complete manuscript with results and discussion.

- Is creativity in analysis stifled?
- Label unapproved but additional analysis 'post hoc'.

©Santosh Noronha (IIT Bombay)
Reproducibility and published research
2018
30 / 41
NPTEL

Now, and for example, germs like nature, already starting to follow an altered publication protocol, where they are So concerned about the fact that data was generated and hypothesis are created that they are saying look entire review process must now happen in two stages.

So, in stage number one; before you even do your experiments you actually try to publish a paper and this is a weird thing. You try to publish a paper and you submit to the editor a protocol that you wish to follow. So, you say that look I wish to work on such in such a system these are the experimental methods I am going to use. And this is the statistical analysis that I propose to do once I get this data. And you want a reviewer system a bunch of reviewers to look at this protocol and tell you in a advance whether this is correct or not ok.

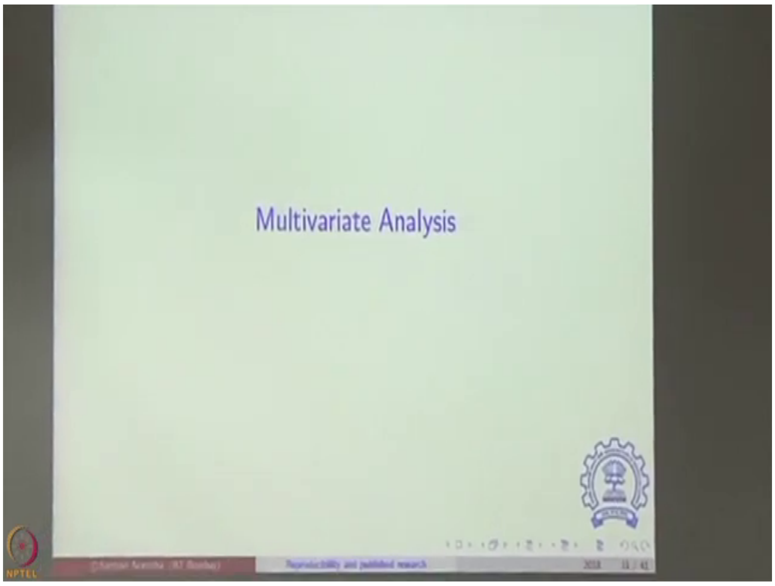
Why would you do this? Because we know we are under pressure to publish positive results and not negative results. So, how do you take away that pressure? So, one way to take away this pressure is to say let your methodology be accepted by the peer group the editors and the reviewers. And at this point regardless of whether your results are good or not if they publish the paper. So, you get guaranteed publication of this paper after a protocol is approved ok.

So, therefore, before you even publish you talk of how you are going to assess your datasets ok. What is the hypothesis in other words, what is the protocol that you going to follow both experimental and computational and what is your detailed analysis plan of how you are going to interpret with gene sets are important to you and which are not ok.

And at this point if they are of acceptable standards you are guaranteed publication. And afterwards you publish the data that you are actually generated both good data and bad data, because now there is no penalty if you publish bad data. One argument against this has been that if you are going to allow people to just publish a protocol and in fact, have to announce my analysis protocol in advance. Does that allow you to do more creative analysis later on because, you have force to you are locked into some kind of analysis already because that is what you going to prove.

But the reality is ok. As long as you label those extra analysis that you do. In fact, it is called a post hoc analysis as long as you flag it in your publication that this was done afterwards, it is still acceptable to the peer review committee. So, this is a game changer in the way the omics ok. Industry is going to potentially function down the road. So, the other moment it is small it is probably like 30 journals which are signed on to this kind of a paradigm for publication. But it is a community which is so concerned of that what they are doing is not a reproducible that they are willing to ok. Collectively go by this protocol of publication.

(Refer Slide Time: 22:44)



Multivariate Analysis

©Santosh Noronha (IIT Bombay) Reproducibility and published research 2018 11 / 41

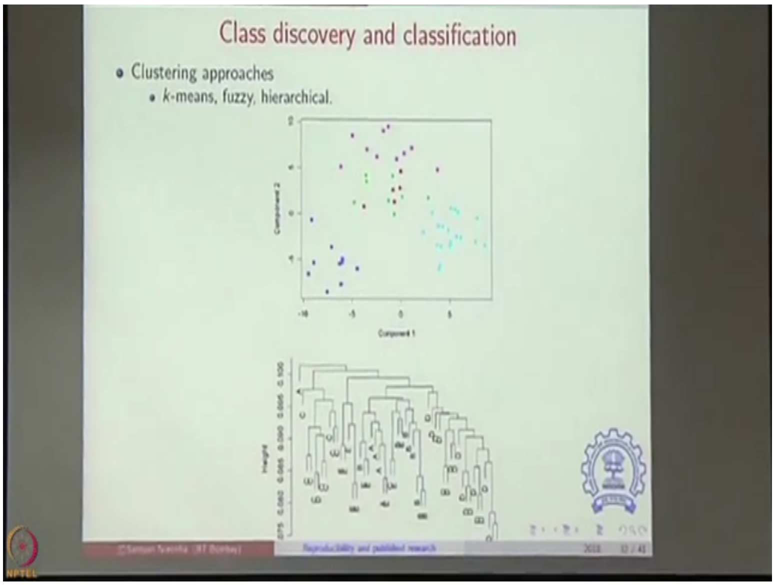
So, I want to spend a few last minutes on what might happen. So, if you are not if it is not a good idea to analyze one gene at a time and ask what is important and what is not important as a candidate what else can you do.

One of the things that is lost to you when you analyze one gene at a time, basically in any system when you analyze one component at a time, what you lose sight of is what are the linkages between the components? It is kind of like saying in a car, I have got each component and I will try to separately study each component. And if you were to study each component yes you know precisely how the brake works, how an accelerator works.

But what you do not know is how the car works given a brake and accelerator. You do not understand how the system works ok. And, clearly there is some interaction between the brake and the accelerator which finally governs all the system works and these interactions are lost to you if you study things in isolation. So, the only solution therefore, in a computational manner is if you take things into a multivariate mode you do not study things one at a time study the whole data set at one shot not one variable at a time.

It turns out there are many ways in which you can do this I am just doing a few buzz words out there some of you will be familiar if you have done courses in bio informatics you will be familiar with things like clustering ok.

(Refer Slide Time: 23:58)



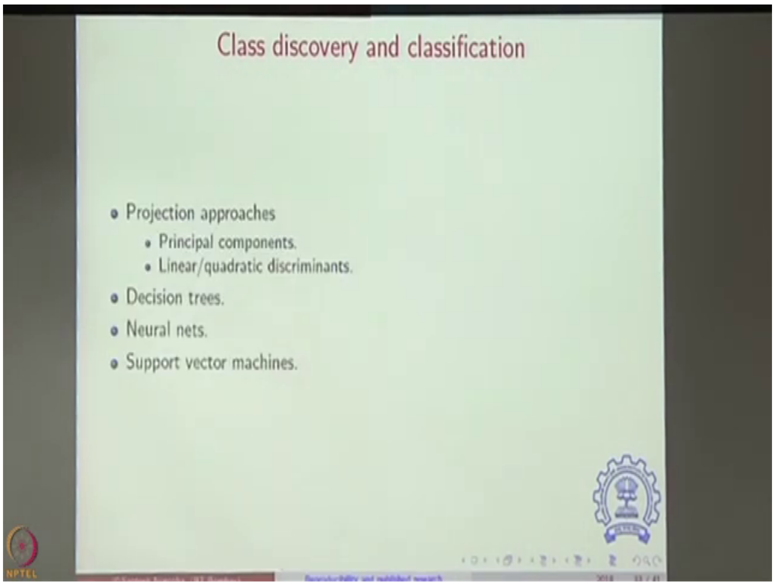
Class discovery and classification

- Clustering approaches
  - *k*-means, fuzzy, hierarchical.

©Sanjeeva Srivastava (IIT Bombay) Reproducibility and published research 2018 12 / 41

Hierarchical clustering is something that for example, phylogenetics a multiple sequence alignment methods would require ok. For example, if you are building some kind of a tree or species or how genes have evolved over time. So, these are all approaches where whole data sets get interpreted at one shot.

(Refer Slide Time: 24:20)

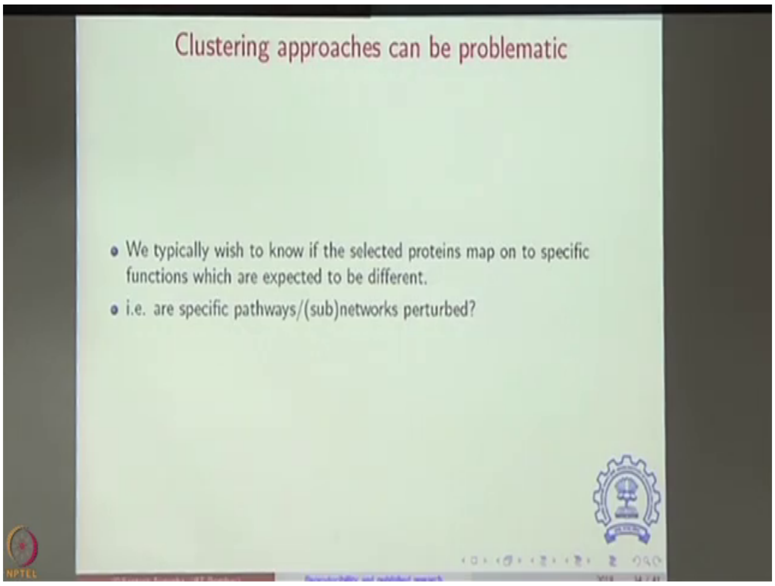


Class discovery and classification

- Projection approaches
  - Principal components.
  - Linear/quadratic discriminants.
- Decision trees.
- Neural nets.
- Support vector machines.

And, if you start looking through the pattern recognition literature and again I am throwing more buzzwords at you, you will realize there is a whole bunch of methods available to you out there. Some of these may get built into some omics tools, but they are more likely to be present in some statistics toolbox in which case you have to make the effort to go to that toolbox and try to figure out what is going on.

(Refer Slide Time: 24:42)



Clustering approaches can be problematic

- We typically wish to know if the selected proteins map on to specific functions which are expected to be different.
- i.e. are specific pathways/(sub)networks perturbed?

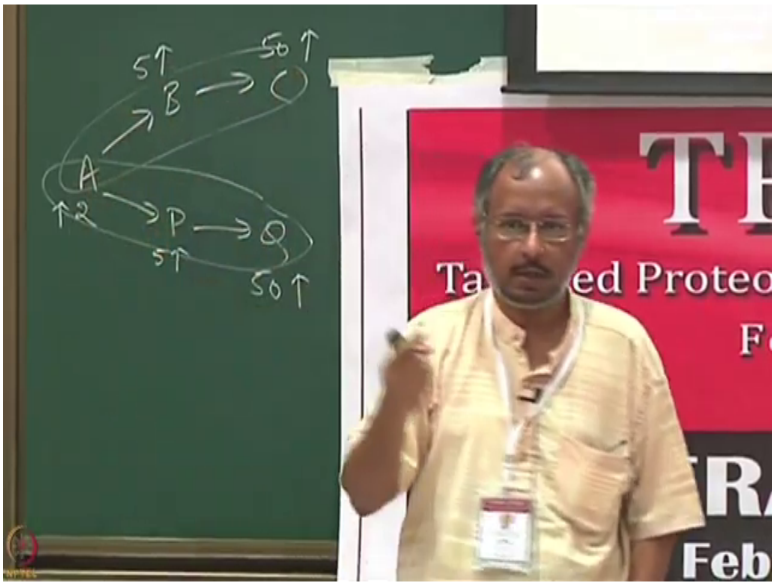
Um Now trying to find patterns and multivariate data sets can be problematic for many reasons. For example, go back to those omics question of you have carried out an experiment control versus test case and you are looking for fold change. You will ask the question what is been you would normally have ask a question to what extent does something mean up regulated or down regulated. If somethings up regulated on a log scale beyond let us say two fold that is got to be significant for you. So, you are going to make that kind of an argument.

Now, one of the problems is why is twofold an important cut off and not some other lower cut off. And I can give you for example, I can give you a simple kinetic argument for why a value of two is arbitrary think of two branched pathways A is going to B going to C, A going to B going to C and A is going to P go into Q. So, two branch pathways A, B, C, A, P, Q.

These let us say our metabolic pathways. There is some metabolism is going on there is branched metabolism at A something is going down one pathway to C something is going down to Q. If you were to look at fold changes. So, if I up regulate something at A that up regulation of an activity at a cascades into some change for B and some change in to C. It is starts impacting changes for B and C and similarly for P and Q.

Which guys would you expect to be the most up regulated as a function of fold change at A. If I have a fold change of a as two fold, what can I expect at B and P? B and P can go up 5 fold because I have typically a transcriptional regulator being toggled a little bit that effect starts impacting some effect a genes a bit more and that goes further down.

(Refer Slide Time: 26:34)



So, very quickly this is what I was saying. So, if I have gone up two fold here and you are trying to say this is significant then it is typically the case in a metabolic pathway that this goes up something like 5 fold and this goes up something like 50 fold. Because the end products start accumulating even more.

So, and why is nature being this way because it does not make sense to directly push this up 50 fold. Because then you lose control over you lose fine tuned control over how things propagate down different pathways and you want to control the expression levels of each intermediate to various pathways ok.

If I ask you to find out those species which are most up regulated. You would have told me C and Q are most up regulated because they have the highest fold changes ok. Therefore, if you were to cluster them, if you did not know better, if I did not draw this pathway structure and you simply told me this went up 50 fold from a spreadsheet and this went up 50 fold from a spreadsheet. One in one temptation at this point is to assume this is a relationship between C and Q.

This C and Q are both path part of let us say some operon and that is why they are whole operons gone up 50, 50 fold. When there is no connection between C and Q, but the connections are via this ok. So, if you wanted a cluster. If the question was being asked what is the cluster of effecter genes which are going up as a response to, what our intervention you did. The cluster was not C and Q as one cluster and B and P as another cluster.

Because they would be clustered on the base of all changes remember that is not a cluster what should have been a cluster this should have been a cluster and this should have been a cluster. Because there is a more obvious biological explanation as to how there is a cascade effect in terms of up or down regulation as you go down pathways.

And you can immediately see therefore, that any clustering approach which now clusters on under an assumption of fold change alone is problematic if you are going to start grouping together candidates or targets in the basis of expression levels alone that is a problem ok. So, you have got to be looking for relationship. So, what is the relationship?

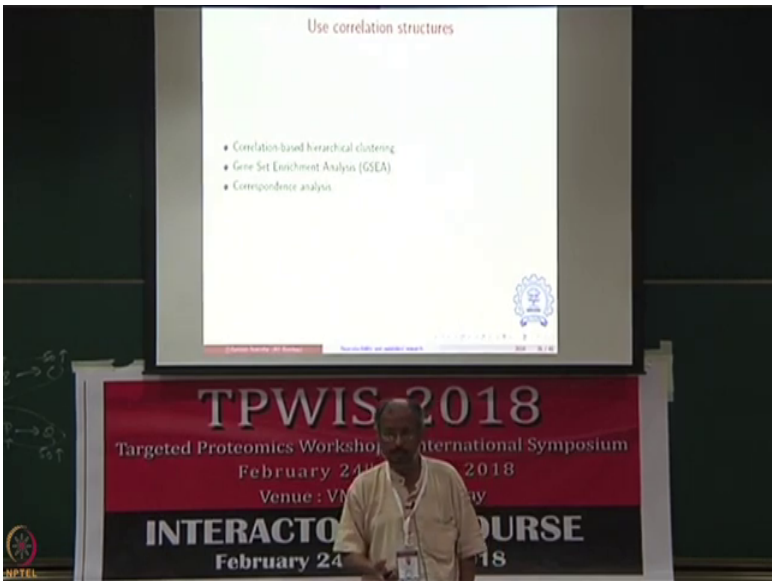
What is you want to start looking for is if I moved this up is something else going up, is something else going up, is something coming down, is something else is going up and what you want to see is in every patient across every patient, across every disease condition if these things are going up and down in coordinated fashion, then there is something going on between this bunch and that bunch deserves to be clustered. This is some genotypic relationship that you are now seeing across these species because they ultimately related by one physical process ok.

Now, that is the subtlety because I am now saying I am not. So, interested in the raw magnitudes of these up and down fold changes that is not important to me. What is more important to me is whether the level of this goes up when this goes up whether this goes up two fold or whether it goes up 1.5 fold, does this go up across all patients ok. And when this goes up, does this go up and those kinds of pair wise relationships is what I start looking for. But what is what do we call those pays why if I were plotting a line between x and y you would call that pair wise relationship or correlation coefficient.

That r square value that I showed you a while back. So, here is certainly an insight instead of simply saying let me look at fold changes and ask is a fold change is important and then trying to identify targets on that basis. Sometimes it is more intriguing to ask the question are correlations between pairs of candidates important and is that telling you something.

And now the reason I bring that up is if I were to somehow plot this data if you look at the previous slide, these are the things where clusters are based on magnitude. So, the 50 fold change guys are all together the 5 fold change guys are all together and so on, but if I wanted to look at correlations that is a different model ok.

(Refer Slide Time: 30:54)



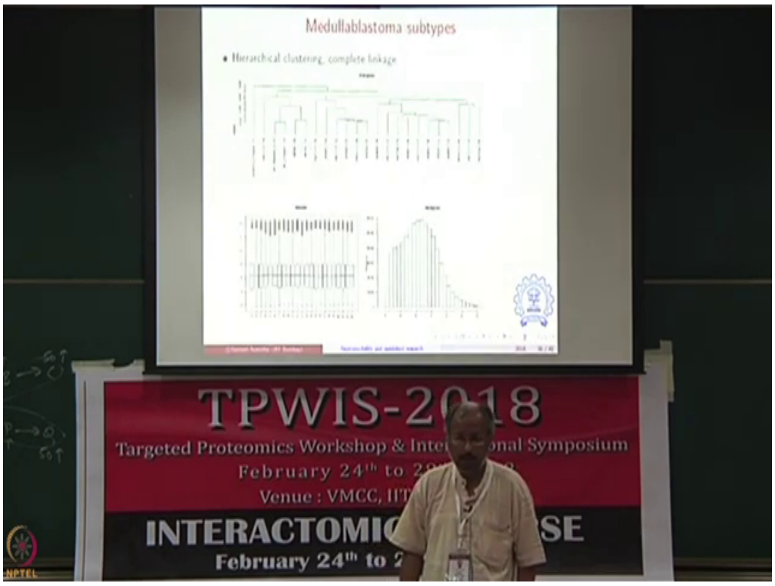
Use correlation structures

- Correlation-based hierarchical clustering
- Gene Set Enrichment Analysis (GSEA)
- Correspondence analysis

So, correlation structures are usually way more important in biology than simple fold changes. Because that fold change could have occurred which is your bad luck that 50 fold for example, could remember all our discussion of randomness 50 fold could have been because of bad luck.

So, instead you need correlation based analysis. If you are talking about hierarchy is that other called our species correlating amongst themselves in a hierarchical analysis. So, choose something based on the correlation analysis ok. There are methods out there for example, on gene set enrichment which say that we build clusters based on which genes are together. One statistical tool which is what I will end up with is something called correspondence analysis where you have looked at how things cluster together.

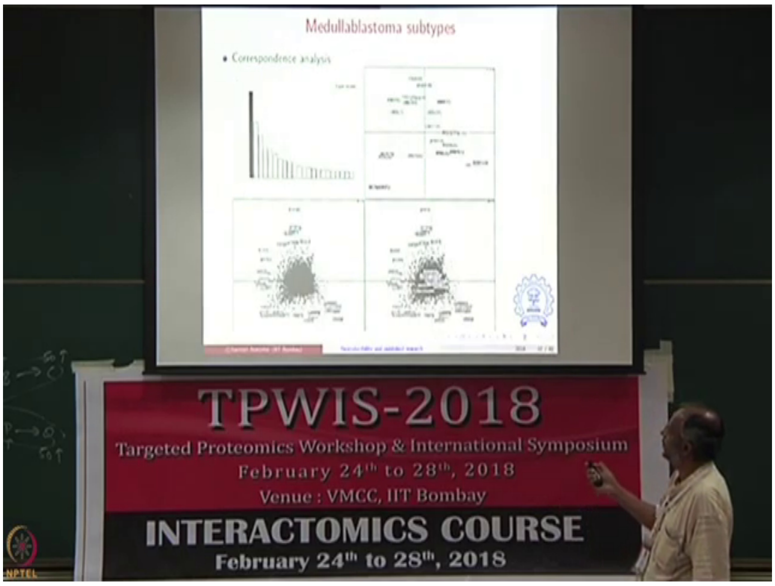
(Refer Slide Time: 31:33)



It is something we have done for a series of a data sets in this case for a medullablastoma k analysis across different types. So, what you are seeing are different patient types and you are looking different patients and samples then you are looking for what so relationship between them. These are all exploratory methods where somebody is saying we have got so many tissue samples across so many patients. Can you find out how many subsets of medulloblastomas you might find and where this is going is nobody knows the cause ok.

What, how many subgroups might exist with this particular disease condition? And then later on you ask a question what could be causing or what is the signature for that subgroup which genes are signatures for each subgroup. But the question as to how many subgroups exist in the first place is itself an open question ok.

(Refer Slide Time: 32:25)



Medullablastoma subtypes

- Correspondence analysis

TPWIS-2018
Targeted Proteomics Workshop & International Symposium
February 24th to 28th, 2018
Venue : VMCC, IIT Bombay
INTERACTOMICS COURSE
February 24th to 28th, 2018
NPTEL

So, if I were to do a hierarchical clustering I will find this kind of analysis. If I were to use something called correspondence analysis that same data is plotted it is literally the same spreadsheet, but it is being plotted different ways. And I wanted to appreciate that there is no one perfect way to do this which is why a better analysis try different ways. Now in this case an interpretation is slightly different. So, here most of you are familiar with how to interpret this two nodes in here are very closely related relative to something else over here.

Whereas here ok, you are essentially asking whether things are far away from the center, at the center you have got an average condition for a tissue sample. As you move further away these are all patients, these are all patients. As you move further away; you are more deviating from the normal ok. So, distance from the origin matters and if you are moving on a diagonal away from the origin. All these guys on a diagonal are related..

So, my notion of a cluster is no longer one nice cloud, spherical cloud. So, this group of patients is a cluster out here is some other cluster and there is another group of patients behaving differently. Which is not which kind of shows up here there is one cluster here, there is one cluster here, there is another cluster here of patients ok.

So, different ways of interpreting this through different insights out. What was very useful about this method was the fact that it allowed allows one to not just plot patients, but you can also on the same coordinate system plot genes. So, remember your dataset. You have got different patients or different sample conditions and for each sample condition based on your omics throughput you have so many gene expression levels or protein expression levels, same logic. And now I am plotting just the genes and asking these are all the normal gene housekeeping genes probably marginally changing the expression levels.

And asked the question which genes are sitting out of the extremes. Which genes radially or furthest away from the origin. Those genes are probably doing something interesting in terms of having their expressions always go up or down based on a correlation with other patients. What is being plotted is not raw magnitude, but correlation coefficients ok.

So, these genetic candidates are all related to each other somehow and one insight by the way is that when one goes looking in these gene candidates are all related to one particular signalling pathway and no surprise that they are all nicely correlated with each other. One guy went up so many other genes responded to that signal and went up and down.

So, they all show up as a cluster on this axis another bunch of genes are clustered around here and so on. And, what is very powerful about this analytical procedure is you can then superpose this on top of this and you then ask remember the clusters of patients we had, there is a cluster of patients here and another cluster of patients. Now what are genetic signatures. So, these genes over here are signatures case specific to this cluster of patients ok.

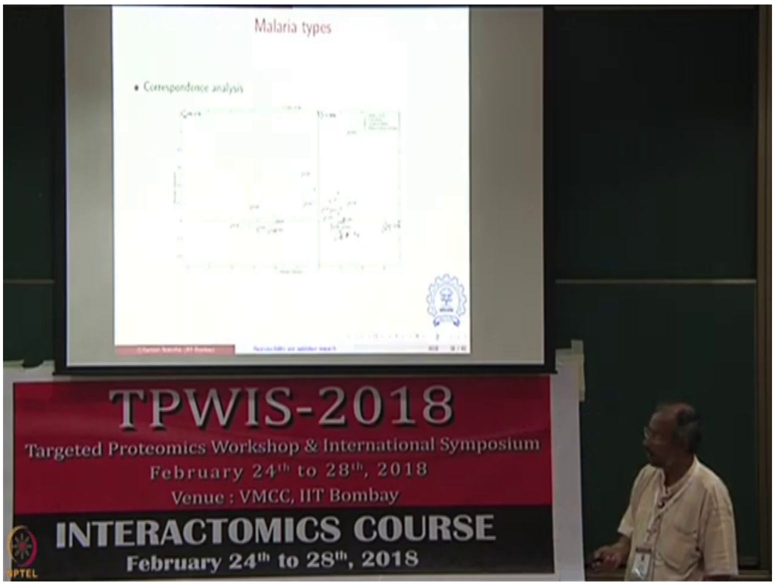
Now, what has happened here is rather than test one gene at a time and we know the problems now of testing one gene at a time by sheer bad luck 5 percent of the time you get things wrong. That can mount as an error rate if I am doing10000 analysis instead the intact lot of data, the entire matrix of data is being analyzed when you think about this these are columns, my patients are columns, my genes are rows in a data set.

So, I am looking at columns of patients I am looking at rows of genes and I am looking at the two things superimposed and I am looking at all my data somehow projected at one shot. And, what I learned from this methodology is that a subset of genes here is associated with these patients. A different subset of genes is associated with a different set of patients and so on. And I already found my clusters and my markers for those subtypes ok.

So, it turns out that in the statistics world, at least in the multivariate statistics world, the appropriate methodology for statistical analysis of this data set existed. It just was the case of being a little adventurous and going out there and trying to find out could was there a method which would more accurately ask answer this question of what were relevant targets. And, not simply trust the least complicated statistical procedure and the least complicated statistical procedure was just one gene at a time and that procedure is prone to a large number of mistakes ok.

Whereas a more robust approach which looks at all the data at one shot in a multivariate mode captured relationships very fast we go looking it is turned out there are nice insights about why these genetic, why these genes were part of a signalling pathway. And, how a defects in one particular gene could escalate into this condition that has led to better science.

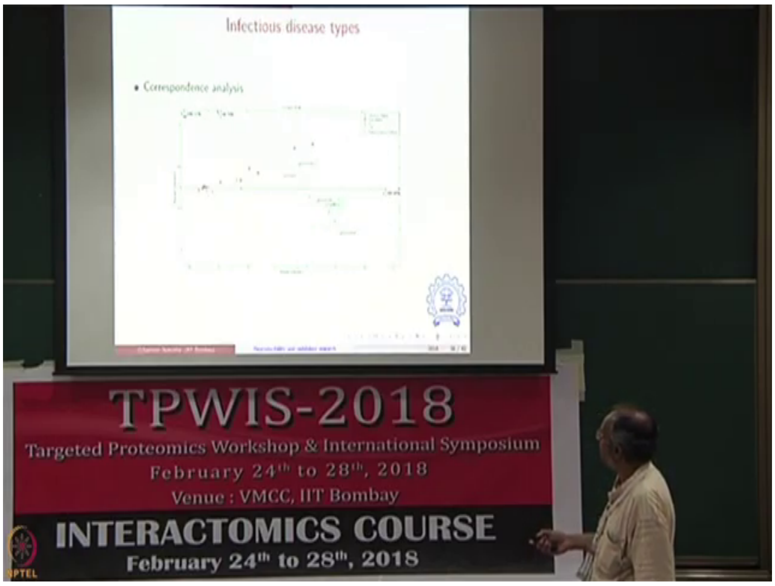
(Refer Slide Time: 37:14)


Malaria types
• Correspondence analysis
TPWIS-2018
Targeted Proteomics Workshop & International Symposium
February 24th to 28th, 2018
Venue : VMCC, IIT Bombay
INTERACTOMICS COURSE
February 24th to 28th, 2018

The same thing has been done later on ok. Again, for proteomics data for different for classifying different types of infections from blood. So, if you are looking so, I am not sure you can make this out other than the colours here, but these are healthy patients, blood from healthy patients and there are a nice cluster on their own ok.

We are looking at falciparum malaria you are looking at vivax malaria and you can clearly see there is a differentiation between vivax and falciparum that shows out when I just cluster this data at one shot.

(Refer Slide Time: 37:50)



Infectious disease types

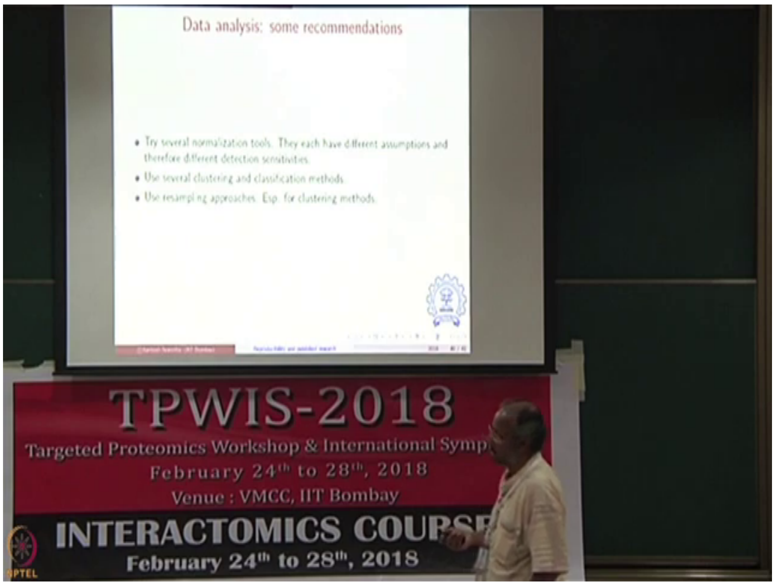
- Correspondence analysis

TPWIS-2018
Targeted Proteomics Workshop & International Symposium
February 24th to 28th, 2018
Venue : VMCC, IIT Bombay
INTERACTOMICS COURSE
February 24th to 28th, 2018

So, we are able to therefore, fundamentally differentiate falciparum from vivax as a malaria type. And in fact, we have gone further beyond those to ask is there a differentiation. For example: from leptospirosis which are all conditions that you would normally see as blood infections causing a high fever and so, if somebody in wants a rapid diagnosis here is an approach which does this. And not sharing all the data here, but you are seeing a subset of your gene candidates and clearly these gene candidates are capable of differentiating multiple clusters.

So, multivariate analysis; it is not a question of one of these genes being analyzed at a time. In fact, you go the other way around. You analyze all the data at one shot on one plot ask which gene subsets are important and then go and ask for each individual gene why did it turn out to be important. You do not flip it the other way around and ask each gene are you important or not and then try to make a story out of it instead the whole data set gets analyzed at one shot, a subset is chosen and each one is reconfirmed as being important one at a time ok.

(Refer Slide Time: 38:44)



Data analysis: some recommendations

- Try several normalization tools. They each have different assumptions and therefore different detection sensitivities.
- Use several clustering and classification methods.
- Use resampling approaches. Esp. for clustering methods.

TPWIS-2018
Targeted Proteomics Workshop & International Symp
February 24th to 28th, 2018
Venue : VMCC, IIT Bombay
INTERACTOMICS COURSE
February 24th to 28th, 2018

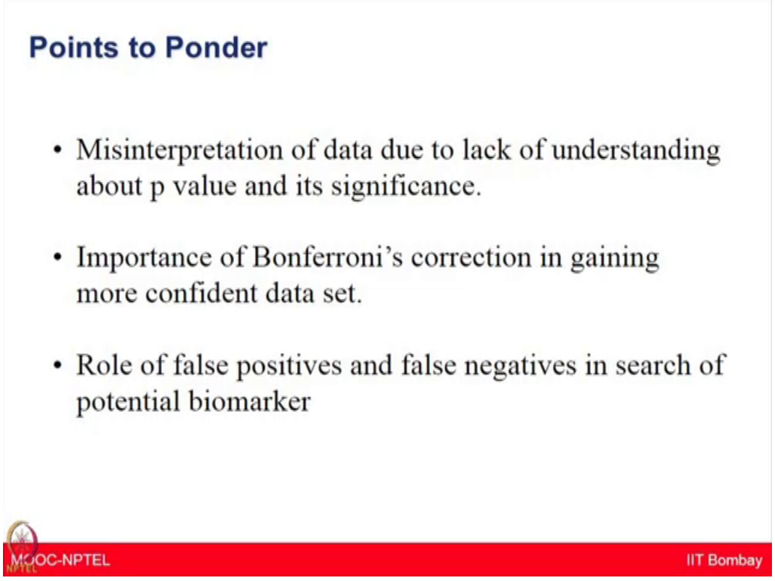
So, I am not expecting you guys to turn statisticians overnight, but this is more in terms of being aware that there are methods out there. And there is several other methods out there which improve the quality of your analysis. So, in a nutshell there are several approaches and it is a democratic philosophy which is do not trust one method, do not trust one voter.

You trust many people to vote for a given candidate and if there are independent statistical methods which are all seemingly voting for the same target, then you have probably found a target. If one method alone is talking about a target then it is probably bad luck and surely not a significant target ok. So, that is another insight to take from this.

So, I will stop there.

.

(Refer Slide Time: 39:27)



Points to Ponder

- Misinterpretation of data due to lack of understanding about p value and its significance.
- Importance of Bonferroni's correction in gaining more confident data set.
- Role of false positives and false negatives in search of potential biomarker

MOOC-NPTEL
IIT Bombay

So, today's lecture I hope you have learnt about the errors created due to the lack of knowledge and understanding about the p values. We also studied how the Bonferroni corrections can help in reduction of false positive and false negative candidates from the data sets. You also heard the role of false positive and false negatives in search of potential biomarkers with include sensitivity and its specificity I hope it also reminded you Dr. Joshua Labaer one of the previous lectures about a good biomarker and considerations for biomarker discovery programs.

So, again you can see that you know different experts have same opinion about the experimental design. How to really find the right candidates, the right targets could be potential biomarker or the discovery targets especially, sorting out based on the false positives and false negatives. So, I hope these two lectures have made you much more aware about the need for the experimental design and various crucial considerations in data analysis. But before I close, let me give you the overall summary of all the lectures which we have covered in this course.

So, we started this course from the basic microarray technologies, especially the nucleic acid programmable protein array and when the leading experts in the area Prof. Joshua Labaer gave you some very interesting lectures about the basics of this technology as well as different applications with more focus on by a worker discovery program in various diseases. We then learnt about how to use NAPPA technology for a screening of various auto antibodies in different disease conditions or use the same technology platform for drug discovery screening.

We also learnt about how to use these technology platforms for protein interactions and looking at various type of protein modifications. So, various these examples, these applications have brought in a horizon that these technologies could be used for identification of biomarkers the therapeutic targets and for the functional proteomics based screaming.

We also got a chance to look into applications of other type of array based platforms. Especially, the reverse phase protein arrays and also the considerations of making good arrays and making good slides by doing good type of printing. Then different type of applications of purified protein arrays using heuprot (Refer Time: 42:23) were shown to you directly with the demonstrational sessions from a researcher scholars in the laboratory where you learnt about some examples of malaria and the cancer research, how it could be beneficial by employing the protein microarray based technologies.

Next we learned about very briefly Immunoprecipitation and the use of the advanced mass spectrometry based technologies. Of course, we did not talk too much about mass spectrometry in this course because that was not the scope of this course. But this is one of the very promising technology which is helping now the entire field of interactomics or entire field of proteomics to say for various applications. So, of course, you should try to get more advanced training in this area, but at this one of the application we try to give you emphasis that IP followed by MS is a strong platform to identify the potential in tractors.

During these lectures, we also try to give you the idea there are different type of label free biosensors are very important. By label based technologies may have some bias for what the signal looks like is that a real signal is there an artefact you have to negate many of the false positives, many of these false fluorescence signal those possibilities in these experiments.

But, the label free sensors label free technologies have tried to overcome that and look for just the biomolecular interactions in its original state. So, trying to avoid many of the confounding factors which one may observe in routine microarray based technologies.

So, I hope technologies like Bio Layer Interferometry, BLI. Surface plasma on resonance with technology like SPR and micro scale. Thermophoresis technologies have really given you the broad idea that many of the label free biosensors could also be used for biomolecular interactions studies. Along with these technologies of microarrays and label free biosensors one of the latest advancement in the entire biomedical field is about next generation sequencing technologies.

And these sequencing technologies have immense applications for the entire genome sequencing to RNA sequencing to variety of applications and we try to give you at least some idea for what can be done using NGA s platforms. The two of the leading industry key players and the replication scientist from illumina and thermo fisher to talk to you about the latest advancement in this area as well as the possible applications which could be used on these technology platforms.

Then we also had interaction with another reading scientist and a clinician Dr. Sanjay Navani who talked to you about another mega project of human protein atlas and the very important role of India in doing the pathology atlas project and they associated challenges of the journey and the major outcomes of this project. So, all of these rapidly evolving technology platforms have immense applications in life sciences and translational biology.

They also provide a much comprehensive picture for better understanding of the crucial physiological processes in systems approach. So, I hope these lectures various discussion points heavily made you aware the pros and cons of designing these experiments and using the technologies choosing the right technology for your given experiment.

I hope these weekly assignments and live interactive sessions, they are helpful and you enjoyed attending this course as much as we made efforts to teach you this course and these advanced technologies.

Thank you very much.