


Lecture-40

Experimental design and Statistical analysis II

Welcome to the MOOC course applications of interactomics using genomics and proteomics technologies. To the professor Santosh Naronha 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 zero point five percent accepted error rate in significance used without basic understanding of data, may written false interpretations. Professor Naronha 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 they're still getting meaningful data is not as straightforward it's not easy. So, I hope today's lecture and based on the previous lectures 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.

Refer Slide Time (2:17)



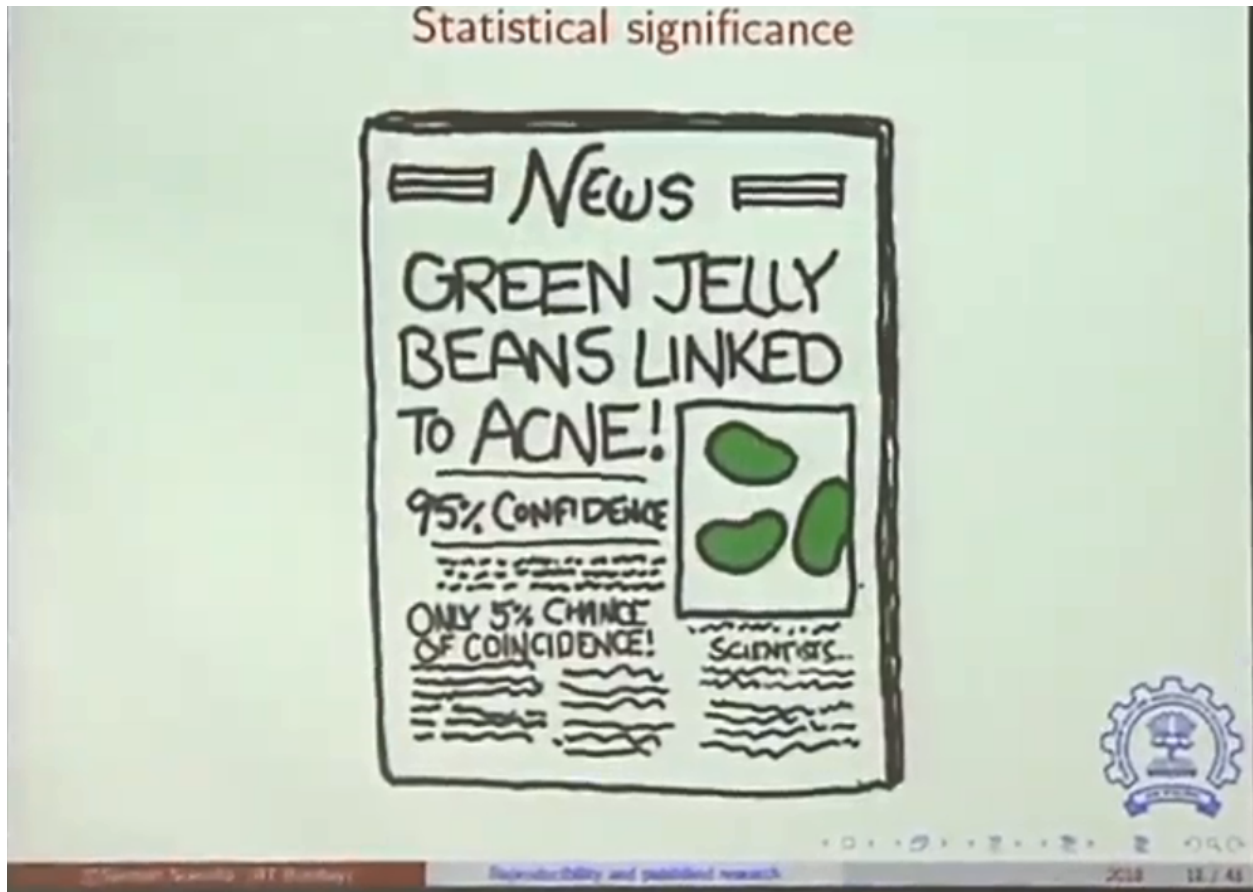
Prof. Santosh Naronha
**Department of Chemical Engineering,
IIT Bombay**

let us welcome / Santosh
[Music]

So let us welcome Mr. Santosh Naronha, They systematically tested so many possible candidates for significance, and if there was a five 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 well it's sure randomness that has caused these to turn up. So, unless you have an independent

way of carrying out an analysis with these candidates and validating that the important to you it's kind of pointless proceeding further.

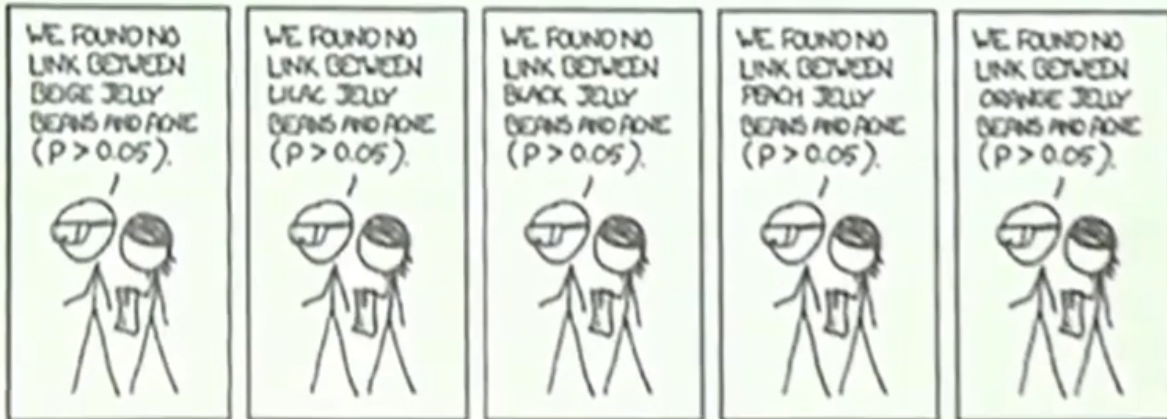
Refer Slide Time (2:49)



Now at this point it for if you are in the publishing game it is very important that you notice that publications

Refer Slide Time (2:54)

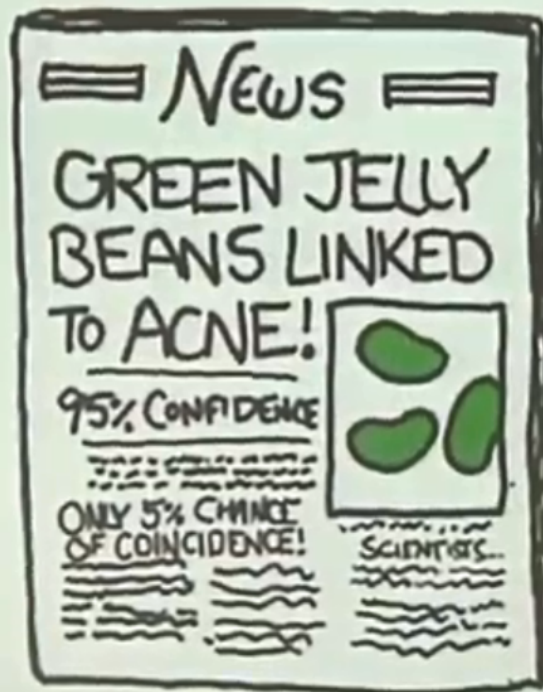
Statistical significance



don't allow you to publish negative results. So, all these other things you cannot publish it.

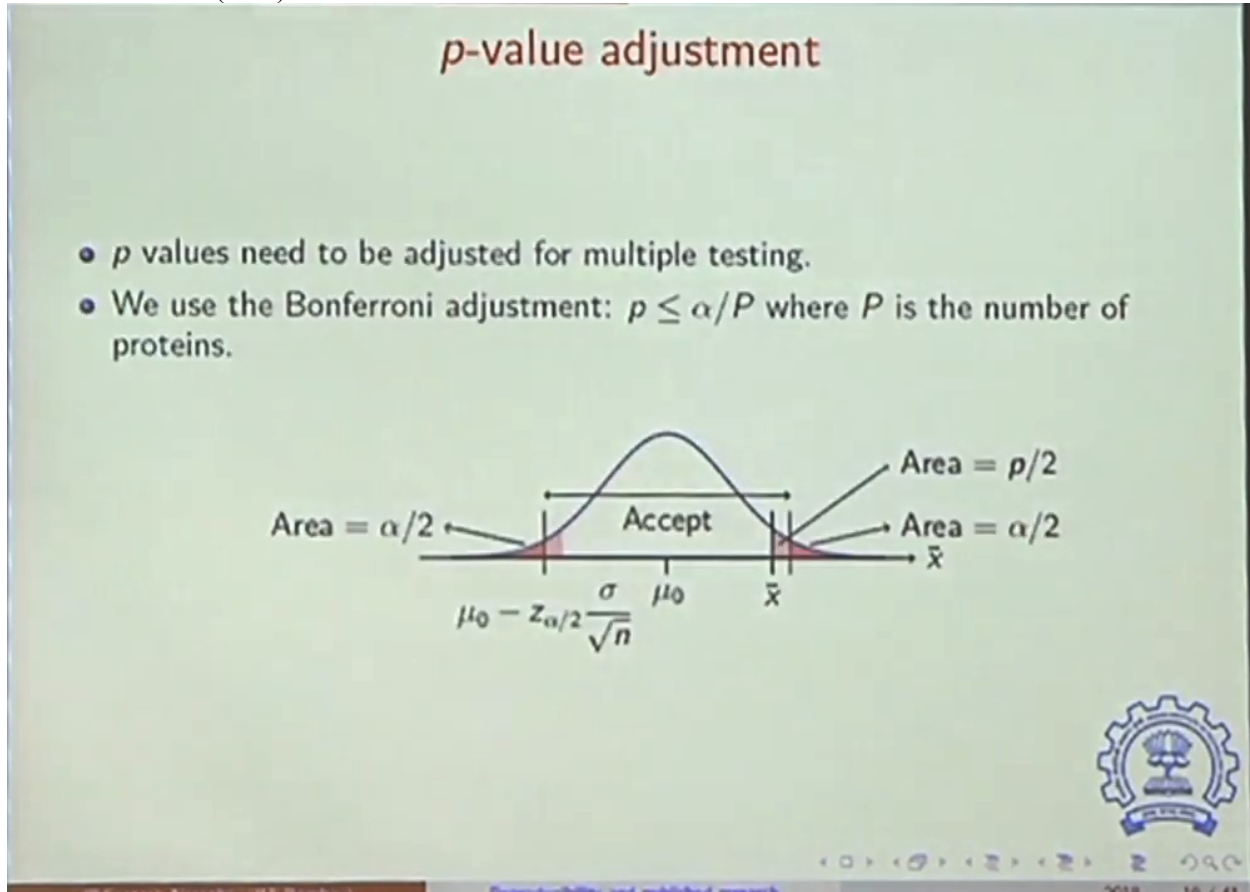
Refer Slide Time (3:02)

Statistical significance



So, the only thing you can publish is this particular result. So there's pressure on you to find that needle in a haystack as a positive result, and publish it, and that's the nature of this confirmation bias, and P hacking Okay? Which pushes you into now focusing entirely or research on this particular candidate the green candidate as if it we're the only relevant candidate.

Refer Slide Time (3:23)



So, what's a way around this, so again something that we typically do not do is is an adjustment to this, so the only way to if, there is a five percent error rate in analysis, and five percent error rate is a dangerous thing if you are doing ten thousand studies. I mean, I want you to, appreciate this five percent at a different level you take any hundred papers published, which are scientific hypothesis being tested and I can tell you without even reading those papers that five of them have got to be false, because all of them have used a ninety five percent confidence level for executing this analysis, and if you are saying there's a five percent sheer bad luck error rate, then several of those researchers have suffered unfortunately with randomness in the data they collected which means your results are going to be false, it's not that they have set out to cheat it's unfortunate given that they are unable to reproduce their own data, and they're trying there is no rush to publish. So, the trick the way to control for this in some ways so, how do I

reduce my error rate so if my error rate is this red portion the five percent, if I want to reduce my error rate I have to move my goalpost further out, and that's only solution of course? Now it gets hard number, of candidates you'll get which passed this goalpost further out which are so extreme what you're saying is your results must be so extreme that they're well outside this wider goalpost range.

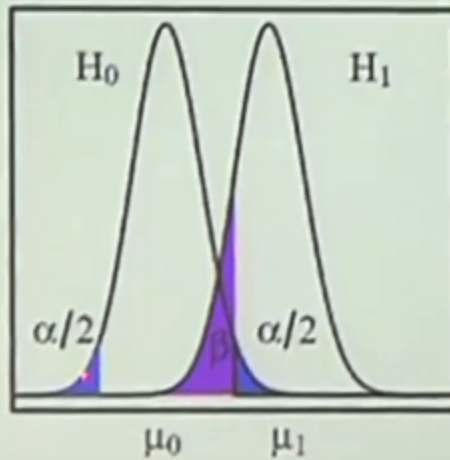
So, what what do you say if you are going to do ten thousand tests each just should not have been done at a five percent level, instead each test should have been done at a five percent divided by the thousand, or ten thousand if I'm doing ten thousand tests so that five percent error rate should be spread across there ten thousand tests that you are doing, and if you were to attempt that this area now is five over ten thousand so it's become a tiny area but I have effectively push my goal posts out so the odds of now passing my test are much much lower and the odds of randomly passing my test have gone down, and that's a core trick to the statistical analysis it's 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's a critical thing so in other words one of the things you ought not to be doing in an omics framework, with statistical tools or being provided to you by the manufacturer is user default setting in a workflow you what to ask the question what are the settings Okay? 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.

There's 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.

Refer Slide Time (6:16)

Power of a test

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

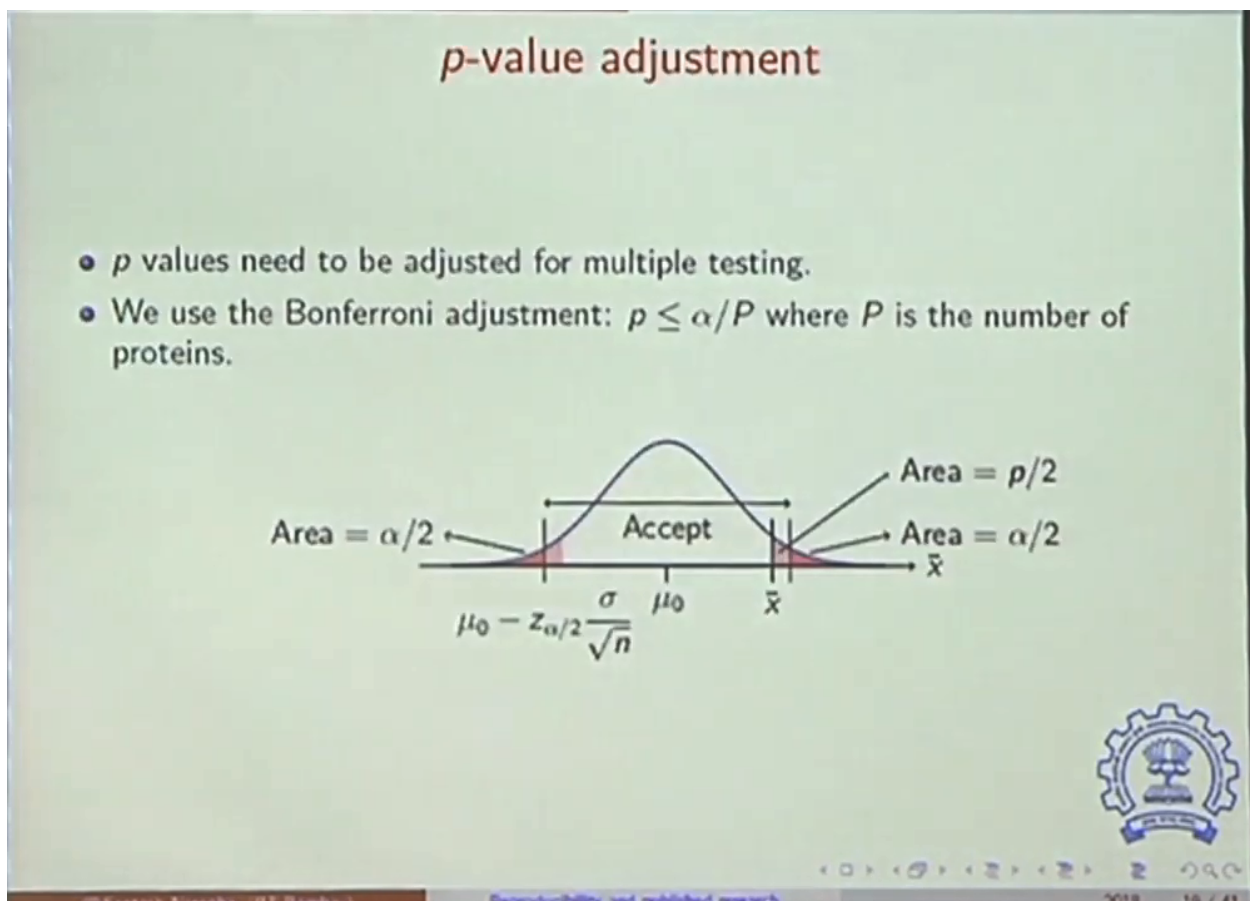


All of this involves only looking at this particular curve so, if you look at this particular curve forget the other curve for a moment, if you're looking at this particular curve then your ninety-five percent confidence leaves out this blue area on either side that's what five percent, the blue area would have been five percent. But for the sake of argument I'm going to pretend as if the reality was some of the hypotheses, 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 is better, and that's just for the sake of argument because now you'll see something problematic happen. If this hypothesis were true, then this blue area corresponds to that percentage of time you're going to get your hypothesis outcome wrong, so that five 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 okay I'm in agreement with that hypothesis, outside that those thresholds you end up saying I don't 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're you you have already argued you prefer to go with H_0 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_0 therefore as a technicality you're committing a mistake by saying H_0 is true, when H_1 should have been true. So, there's a mistake there's a mistake it's

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're 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_0 curve what'll 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 Okay?

Refer Slide Time (9:06)

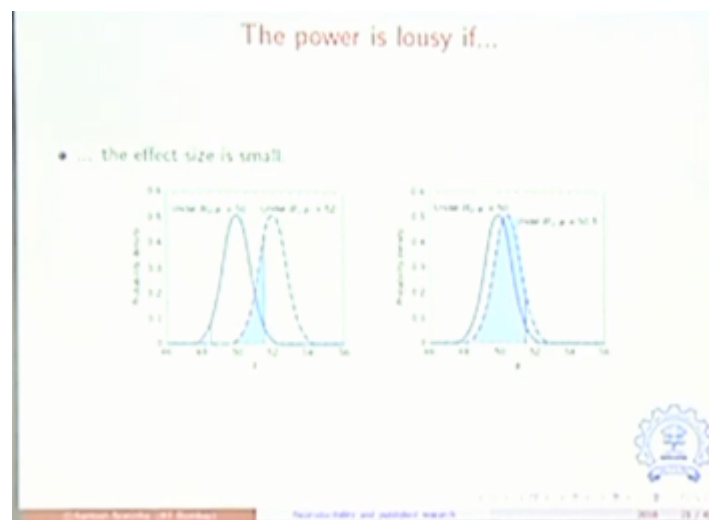


So that's key issue, so what the the headache comes about because if you look at what we have done in the previous thing we only paid attention to one curve we didn't ask the question what might the other hypothesis behave like which is the case here Okay? So, if you start paying attention to alternate hypothesis you suddenly realize that yes I might have a diagnostic kit for example, which is accurate ninety-five percent of the time that's what that confidence in told tells you but what and therefore a five percent of the time, I'm making a mistake of a certain kind let's say I'm falsely calling somebody positive,

so false positives but what is not giving me information at all is what's 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 one minus 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 don't 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 I will lousy test.

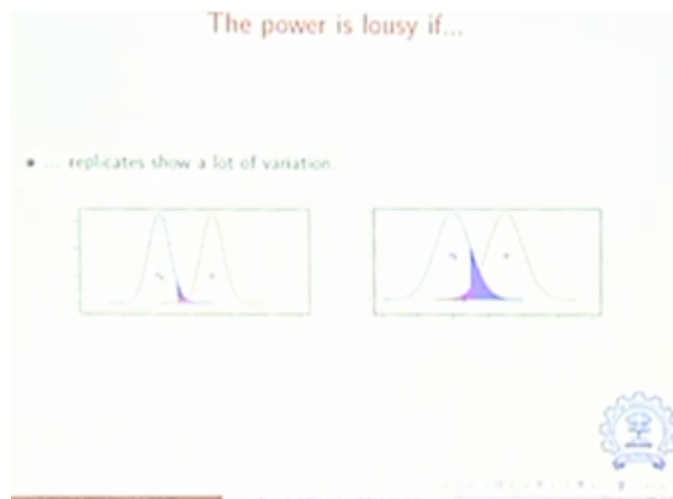
So, this power of four test is a concept which most again is there it's buried somewhere in software typically as an option for you to report, but it's not something researchers are in the habit of reporting so when somebody tells me I've found a candidate and in fact I've found a short list of candidates which are all significant what they're not telling me is how powerful was the analysis, what you're not telling me therefore is what was the probability that they've got the analysis wrong the other way around well they're telling me that they're confident within ninety percent what they're not telling me, is whether this was so bad that this was twenty percent or thirty percent or forty percent the papillary. If one of these values is greater than ten percent your study, your analysis is already in trouble. So, both alpha and beta both these shaded areas cannot be large because they're both errors in your interpretation.

Refer Slide Time (10:59)



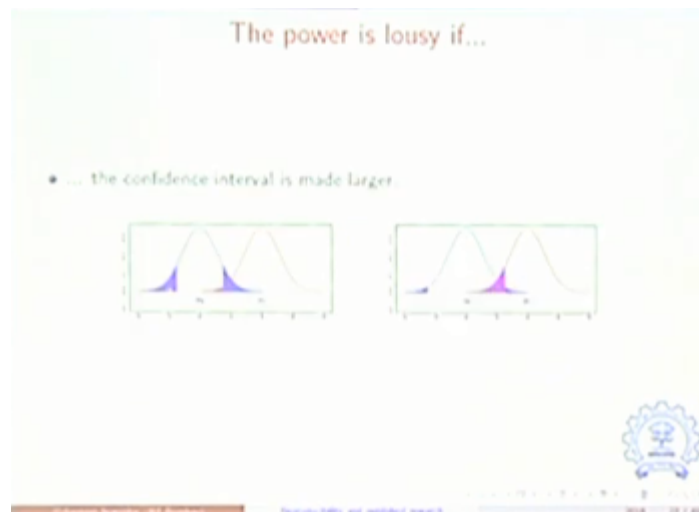
And here are cartoons which quickly improve the points, so if you are trying to distinguish two hypotheses and your two hypotheses are so similar to each other therefore the effect size is small you will have such an overlap between the two the predictions coming out of the two hypothesis, that they are unable to discriminate and say which hypothesis is true, you will not be able to do that Okay?

Refer Slide Time (11:20)



If you have a bunch of replicates if you have a bunch of replicates you I'm 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 in a nutshell you want more applicates of any analysis that you do otherwise you are going to be large.

Refer Slide Time (11:45)



If you're going to make your confidence into a large if somebody says that really ought not to be talking about ninety five percent confidence you need ninety nine percent confidence, then they made it out come 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 here so that that was since the case. So there's no clean solution here, the moment you try to improve one cent situation something else in your analysis is going to worsen some errors, 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 Okay? Whether this is a powerful test being done.

Refer Slide Time (12:22)

The Amgen study

- Amgen is a Biotechnology company producing a large number of biopharmaceuticals.
- They tried to confirm published findings from 53 'landmark' papers, in oncology and hematology.

© 2018 Amgen, Inc. All rights reserved. | Research and published results | 2018 | 18 / 42

This is a famous study and it's really worth looking this up on your own later on. I'm jealous one of the two top bio tech companies in the world Okay? They make they're the dominant manufacturer bar pharmaceuticals protein drugs for the most part, and while initially they worked on things which are already discovered in research labs increasingly they've 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 fifty-three landmark papers published in the top journals in oncology and hematology. These are publications coming out of MIT Caltech Stanford Berkeley their top labs, the top universities in the world fifty-three 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's get into the business of manufacturing these candidates that's where they were going Okay?


So, out of fifty-three they could reproduce only six papers, and this is MIT Caltech Stanford you're not talking of some small tiny college somewhere, so what's going on. It's not that people at MIT and Berkeley and Caltech were cheating it's not that they were deliberately cheating, but there's a situation where the results coming out of the even these top labs cannot be reproduce so why do you think they cannot be reproduced Okay?

Refer Slide Time (13:58)

The Amgen study

In the 6 cases, attention had been paid to

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



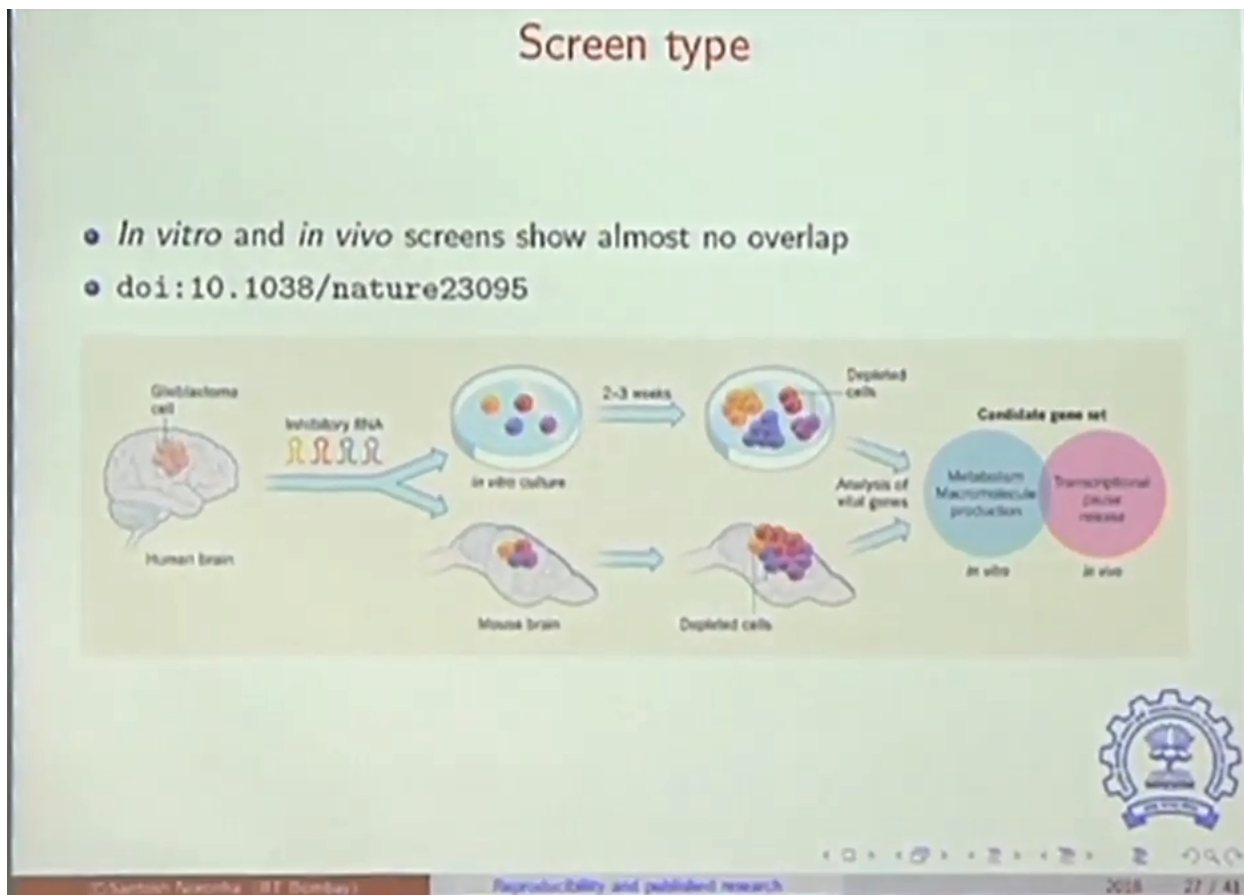
2011 25/11

In the six cases where the results could be reproduced when you look at carefully what happened Okay? Attention had been paid to doing the right controls in the experiment you need the right controls you don't make claims about results based on only a test case you do the controls the reagents were a producible, and this you will realize most of you are doing experiments met lab experiments reagents especially in the immunology space I had to obtain in reproducible fact antibodies in particular Okay? a batch to batch variation exists, and you are unable to reproduce results.

So, in the six cases there was the ability to manufacture these reagents reproducibly and that made a huge difference Okay? The investigators were not biased they didn't they were not trying to push for a particular insight or an outcome, and importantly they were honest about reporting all their data, so you remember that straight line plot where I deleted the mid 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's when somebody tried to reproduce it they also saw some bad data equal equally back to what these guys had found. It's a surprising result it tells you to what extent there is pressure on people to publish positive insights Okay? Even other top plugs. And the moment this study came out, and when this

pharma company published this inside many companies started paying suit so buyer did a similar study they looked at sixty seven targets published in the literature bias another big pharma major and out of sixty seven they could reproduce fourteen results, which tells you this is a serious serious problem Okay?

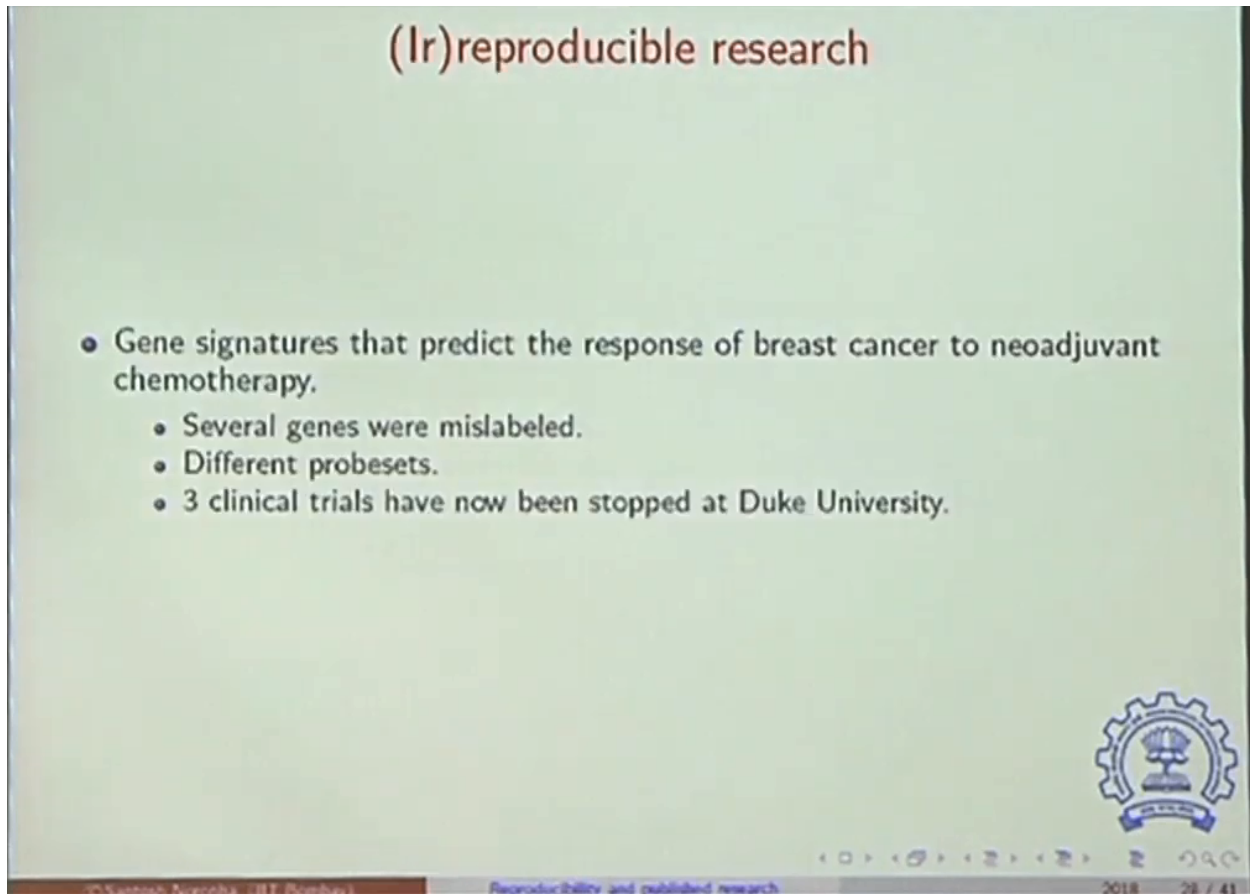
Refer Slide Time (15:53)



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's not a producible experiment. Okay? But it also reflects other aspects of poor design so here's one experiment of poor design, where you're screening for certain drugs to do with epigenetic control in glioblastoma system, and the screen which was done was an *in vitro* screen. So there's two ways finally this got done and *in vitro* screened using basically relying on RNA interference kind of protocol to try to identify targets, and the *in vivo* screen 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 in the *in vitro* screen have practically no overlap in terms of what's up regulated, and what's down regulated Okay? Which means if you had just done the *in vitro* experiment and you generated a bunch of targets

and you're then proposed to now Okay? design drug candidates against these targets, you have a huge amount of money.

Refer Slide Time (16:59)



(Ir)reproducible research

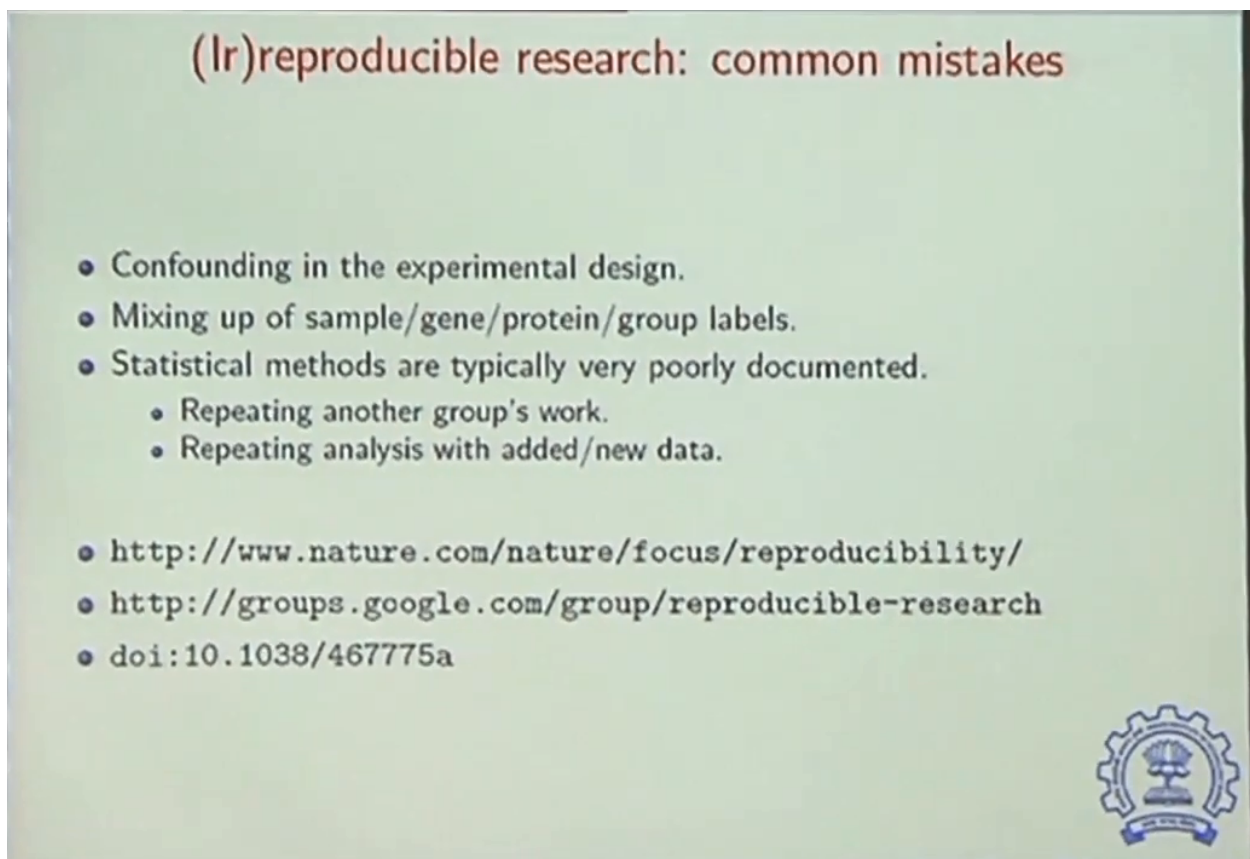
- Gene signatures that predict the response of breast cancer to neoadjuvant chemotherapy.
 - Several genes were mislabeled.
 - Different probesets.
 - 3 clinical trials have now been stopped at Duke University.

© Kavitha Nandha - IIT Bombay | Reproducibility and published research | 2018 | 28 / 41

They have been several such things you start going through the literature you will see these things you know it's not something conventionally with that journals published so this is published elsewhere some of these ironically are published in blogs they're not published Nestle engines they they have. For example mean Studies on looking at gene signatures predicting the response of breast cancer to chemotherapy Okay? 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 out this analysis nobody followed that up and it went through an entire analytics pipeline Okay? By informatics drug candidates were created, probe sets were created Okay?

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'll appreciate a billion-dollar experiment sometimes Okay? 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 a producible these candidates despite the by informatics don't seem to work in reality, and the hard question what asked why? show your lab notebooks you go all the way back and you look at the printouts of these 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 Okay? Could have been worse if people had died as a consequence of being seriously hurt as a consequence of the trial because you're actually playing around with therapies or proposed therapies you could have been much much worse in terms of how this sort of what the university, and the researchers.

Refer Slide Time (19:37)




(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>

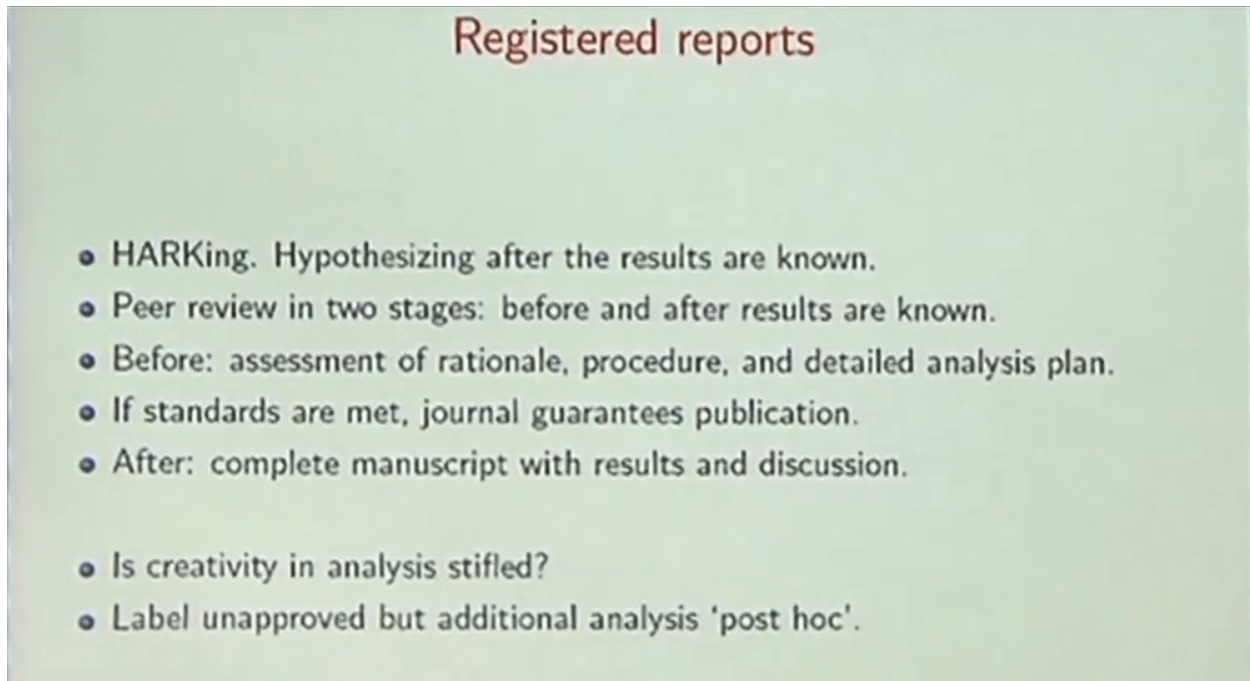
● doi:10.1038/467775a



So, I'm giving you a bunch of links here really what what I wanted appreciate is leaving data analysis and statistics as an afterthought to a by informatics pipeline is a dangerous business, 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'll actually

get down to doing the science, but really it should be the other way around they should have been 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.

Refer Slide Time (20:24)



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.

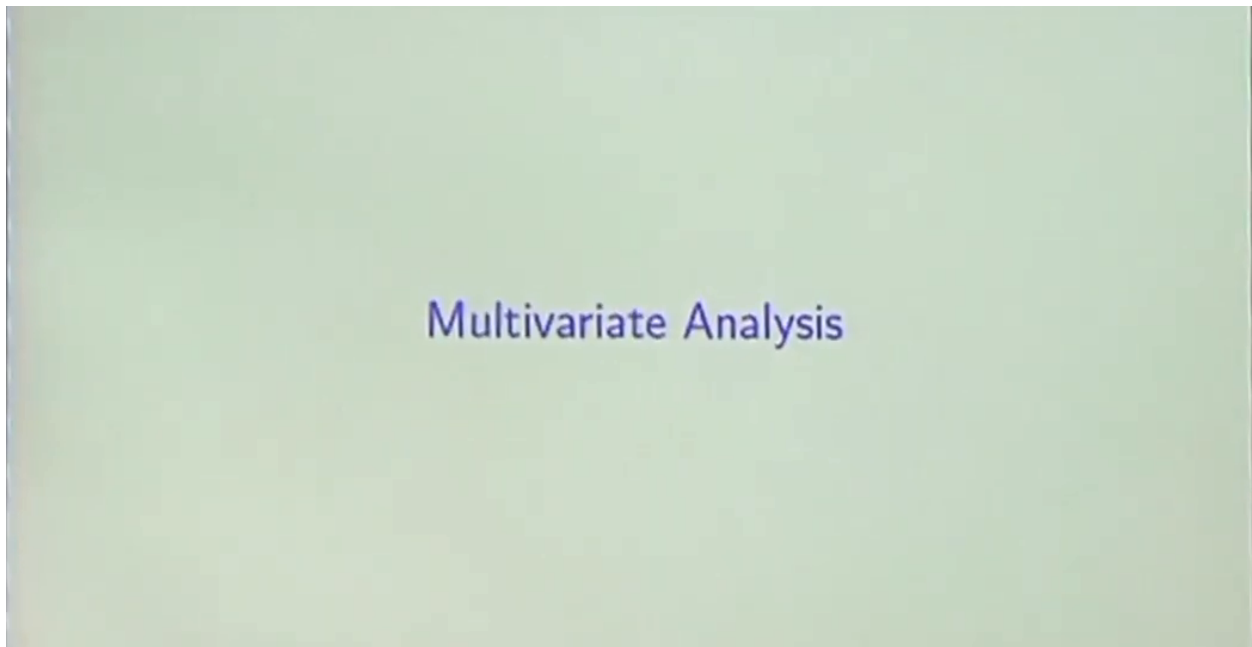
- Is creativity in analysis stifled?
- Label unapproved but additional analysis 'post hoc'.

In fact the whole publishing paradigm in the omics space isn't going to change now and for example germs like nature already starting to follow an altered publication protocol, where they're so concerned about the fact that data was generated and hypothesis are created that is 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'm 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 advance whether this is correct or not Okay? Why why would you do this because 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 that peer group the editors, and reviewers and at this point regardless of whether your results are good or not we they publish the paper. So, you get guaranteed publication of this paper, after a protocol is approved. Okay? So, therefore before you even publish you talk of how you're going to assess your datasets Okay? What is the hypothesis in other words what's the protocol that you want to follow both experimental, and

computational, and what's your detailed analysis plan of how you're going to interpret with gene sets are important to you and which you're not?

And at this point if they are acceptable standards, your guaranteed publication and afterwards you publish the data that you 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're 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 forced you are locked into some kind of analysis already because that's what you gotta prove. But the reality is Okay? As long as you label those extra analysis that you do in fact it's called a post hoc analysis as long as you flag it in your publication that this was done afterwards it's still acceptable to the peer review committee. So, this is a game changer in the omics Okay? industry is going to potentially function down the road. So the other moment it's small it's probably like thirty journals which are signed on to this kind of a paradigm for publication, but it's a community which is so concerned of that what they are doing is not a participant that they willing to Okay? Collectively go by this protocol of publication.

Refer Slide Time (23:03)

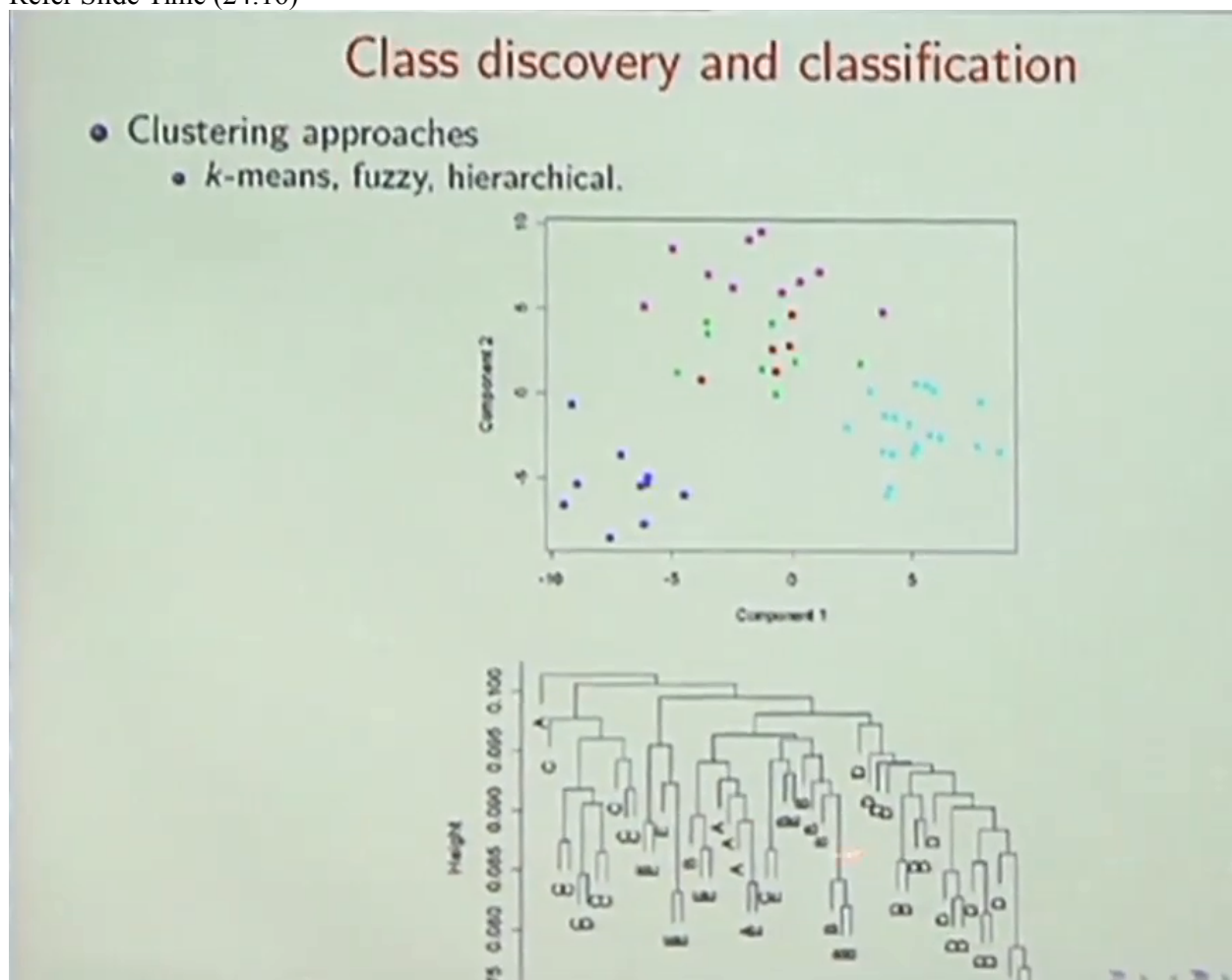


So, I want to spend a few last minutes on what might happen so if you're not if it's not a good idea to analyze one gene at a time, and ask what's important to watch not important as a candidate what else can you do? one of the things that's lost to you when you analyze one gene at a time basically in any system when you analyze one component at a time, what you lose sight of is what are the linkages between the components. It's kind of like saying, in a car I've got each component, and I'll try to separately study each

component and if you were to study each component yes you know precisely brake works, how an accelerator works but what you don't know is how the car works, given a brake and accelerator you don't understand how the system works Okay? and clearly there's 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 don't study things one at a time, study the whole data set at one shot Okay? not one variable at a time. It turns out there are many ways in which you can do this.

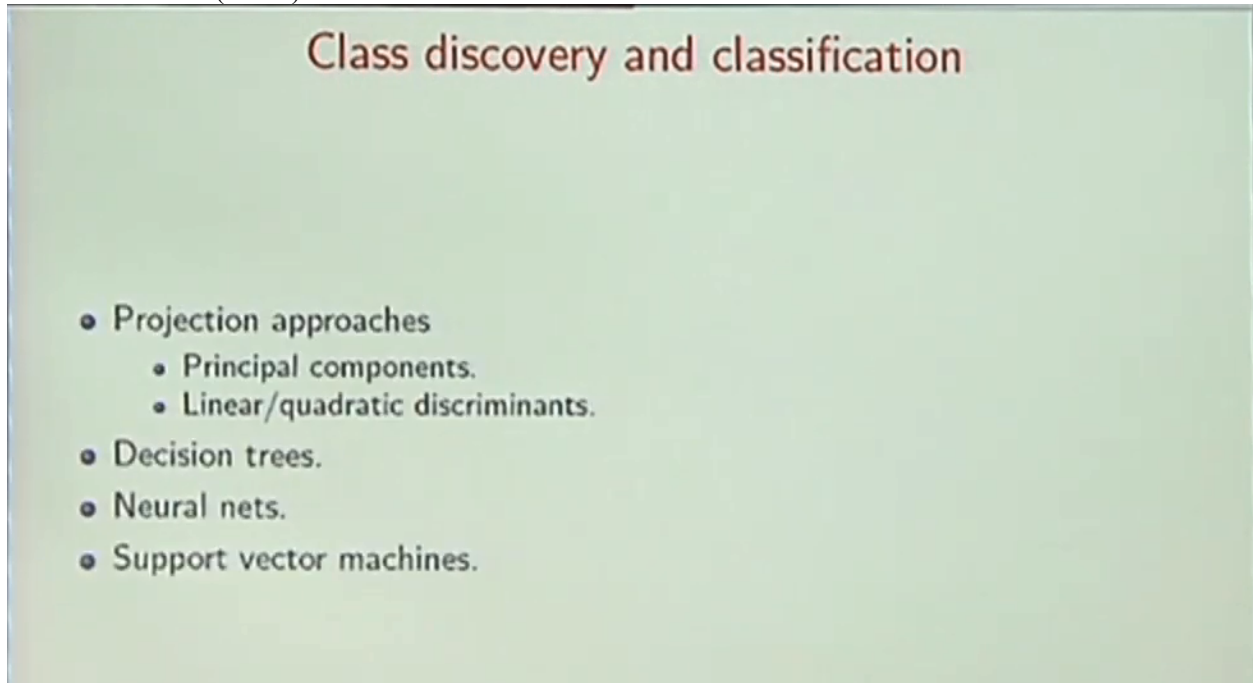
Refer Slide Time (24:16)



I'm just doing a few buzz words out there okay some of you will be familiar if you have done courses in by informatics, you'll be familiar with things like clustering Okay? hierarchical clustering is something that for example phylogenetics a multiple sequence alignment methods would require Okay? For

example, if you're building some kind of tree, or species Okay? or how genes have evolved over time. So, these are all approaches where whole data sets get interpreted at one shot.

Refer Slide Time (24:42)



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'm throwing more buzzwords at you you'll realize there's 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's going on.

Refer Slide Time (25:02)

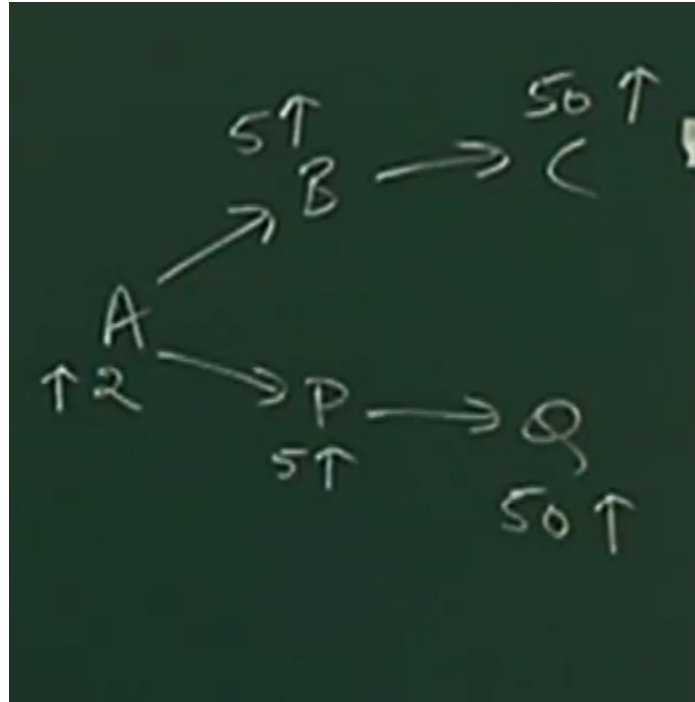
Clustering approaches can be problematic

- 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?

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 was so stressed case, and you're looking for fold change, you'll ask the question what's? been you you'd normally have ask a question to what extent does something mean up regulated, or down regulated if something's up regulated you know on a log scale beyond let's say two fold that's got to be significant for you so you're going to make that kind of an argument. Now one of the problems is why is twofold an important cutoff and not some other lower cutoff, 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 ABC, APQ. These let's say are metabolic pathways this sub metabolism going on this branched metabolism at a something's going down one pathway to C something's going down to Q.

If you were to look at fold changes so if I up regulate something at A that upregulation of an activity at a cascades into some change for B and some change in to C it starts impacting changes for B and C, and similarly for P and Q, which guys would you expect to be the most appreciated as a function of fold change at A, if I have a whole change of a as two fold what can I expect at B and P, B and P can go up fivefold because I have typically a transcriptional regulator being toggled a little bit that effect starts impacting some effector genes a bit more, and that goes further down so very quickly,

Refer Slide Time (26:55)



so this is what I was saying so if I've gone up twofold here and you're trying to say this is significant then it's typically the case in a metabolic pathway that this goes up something like fivefold and this goes up something like fifty fold because the end products start accumulating even more. So and why is nature being this way because it doesn't make sense to directly push this up fifty fold because then you lose control over you lose fine-tuned control or how things propagate down different pathways, and you want to control the expression levels of each intermediate through various pathways. Okay? If I ask you to find out those those species which are most up regulated, you would have told me C and Q are most up regulated because they are the highest fold changes Okay? Therefore, if you write to cluster them if you didn't know better if I didn't draw this pathway structure and you simply told me this went up fiftyfold from a spreadsheet and this went up fiftyfold from a spreadsheet. One in one one temptation at this point is to assume this is a relationship between C and Q at C and Q are both paths part of so let's say some operon and that's why their whole operands gone up fiftyfold.

When there is no connection between C and Q but the connections are wireless this Okay? So if you wanted a cluster if the question was being asked what's the cluster of effector genes which are going up as a response to whatever 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 basis of all changes remember that's not a cluster, what should have been a cluster this should have been a cluster, and this should have in a cluster, because there is a more obvious biological explanation as to how there's 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

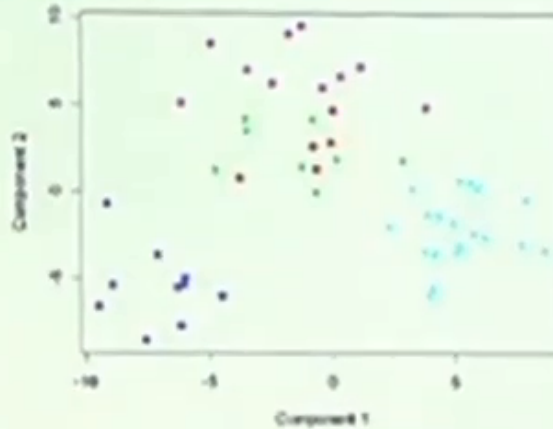
which now clusters on under an assumption of fold change alone is problematic, if you're going to start grouping together candidates or targets the basis of expression levels alone that's a problem. Okay? So, you have what to be looking for relationship, so what's the relationship what is one to start looking for is if I move this up is something else going up, is something else going up, is something coming down, is something else 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's 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. Okay?

Now that's a subtlety because I'm now saying I'm not so interested in the raw magnitudes of these up and down fold changes, that's not important to me, what is more important to me is whether whether the level of this goes up, when this goes up, whether this goes up two fold or whether it goes up one point fivefold does this go up across all patients Okay? and when this goes up, does this go up, and those kinds of pairwise relationships is what I start looking for but what is what what do we call those pairwise why if I were plotting a line between x and y you would call that pairwise relationship or correlation coefficient that r-squared value that I showed you a while back. So, here's certainly an insight instead of simply saying let me look at fold changes and ask is a fold change important, and then trying to identify targets on that basis. Sometimes it's 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. Okay?

Refer Slide Time (30:57)

Class discovery and classification

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



If you look at the previous slide these are things where clusters are based on magnitude, so the fifty-fold change guys are all together the fivefold changes are all together and so on.

Refer Slide Time (31:09)

Class discovery and classification

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

But if I wanted to look at correlations that's a different model Okay?

Refer Slide Time (31:13)

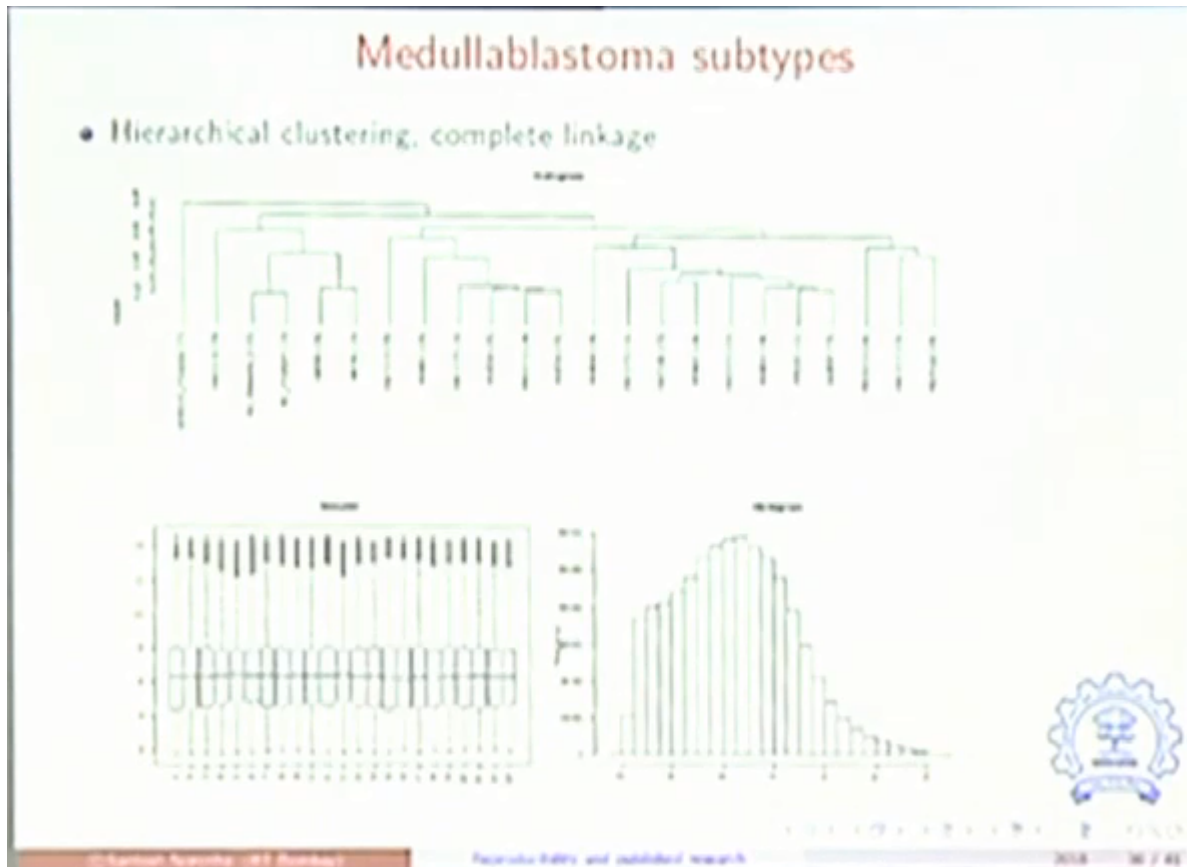
Use correlation structures

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

So, correlation structures are usually very more important in biology than simple fold changes because that fold change could have occurred with your bad luck that fifty-fold for example could remember all our discussion of randomness fifty-fold could have been because of bad luck. So instead you need to correlation-based analysis you should talk about hierarchy is that other called our species correlating amongst themselves in a hierarchical analysis. So, choose something based on the correlation analysis

Okay? 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'll end up with is something called correspondence analysis where you have looked at how things cluster together.

Refer Slide Time (31:53)

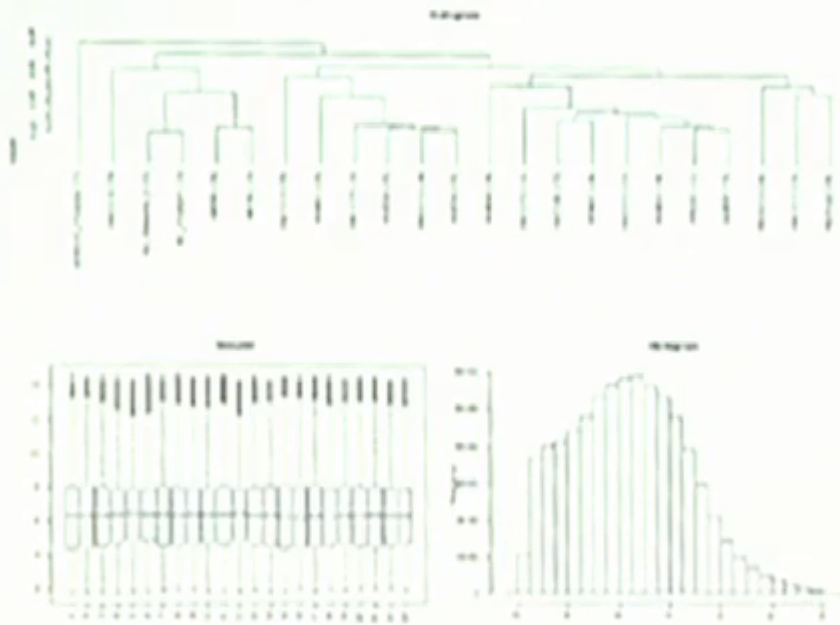


It's something we have done Okay? For a series of data sets in this case for a medulloblastomas analysis across different types so what you're seeing are different patient types and you're looking different patients and samples and you're looking for what's the relationship between them. These are all exploratory methods where somebody's 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 Okay? What how many subgroups might exist with this particular disease condition, and then later on you ask a question what what could be causing or what 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. Okay?

Refer Slide Time (32:44)

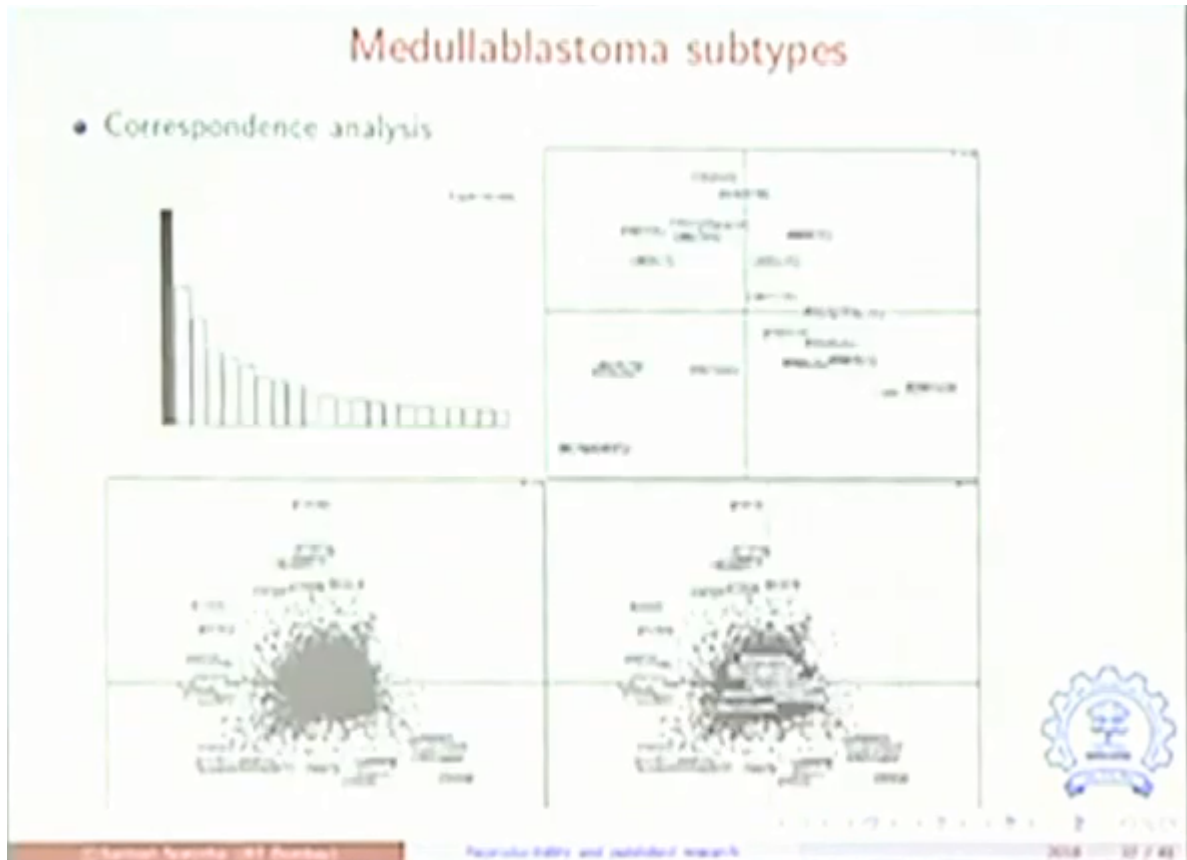
Medullablastoma subtypes

- Hierarchical clustering, complete linkage



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.

Refer Slide Time (33:07)

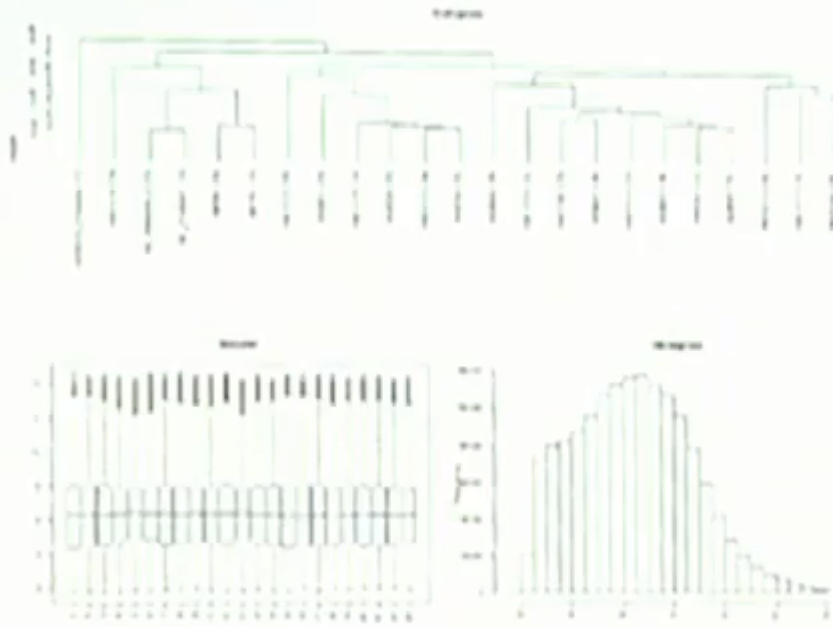


But as here Okay? you're essentially asking whether things are far away from the center at the center you've got an average condition for a tissue sample as you move further away these are all patients, these are all patients as you move further away you're more deviating from the normal Okay? So, distance from the origin matters and if you're 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 a nice cloud spherical cloud, so this group of patients is a cluster out here is some other cluster, and there's another group of patients behaving differently.

Refer Slide Time (33:45)

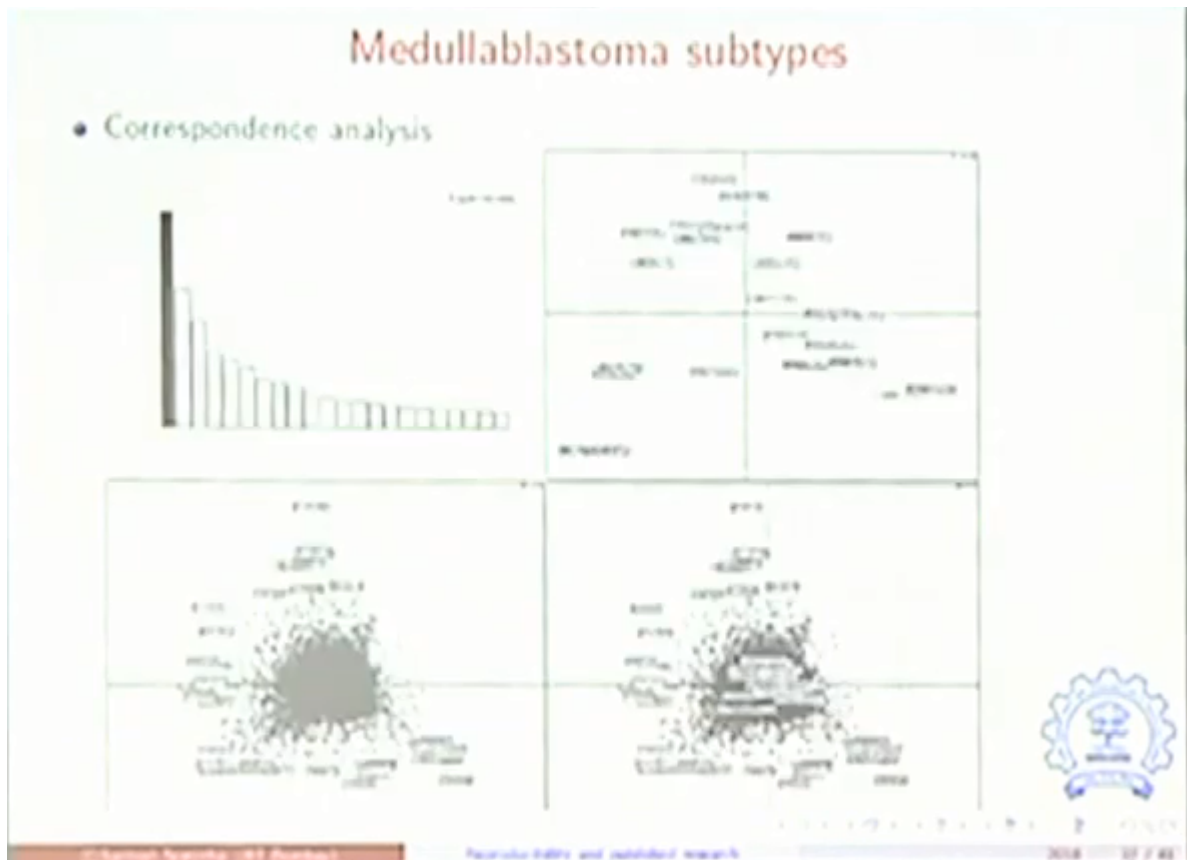
Medullablastoma subtypes

- Hierarchical clustering, complete linkage



Which is not which kind of shows up here there's one cluster here there's one cluster here there's another cluster here of patients Okay?

Refer Slide Time (33:52)



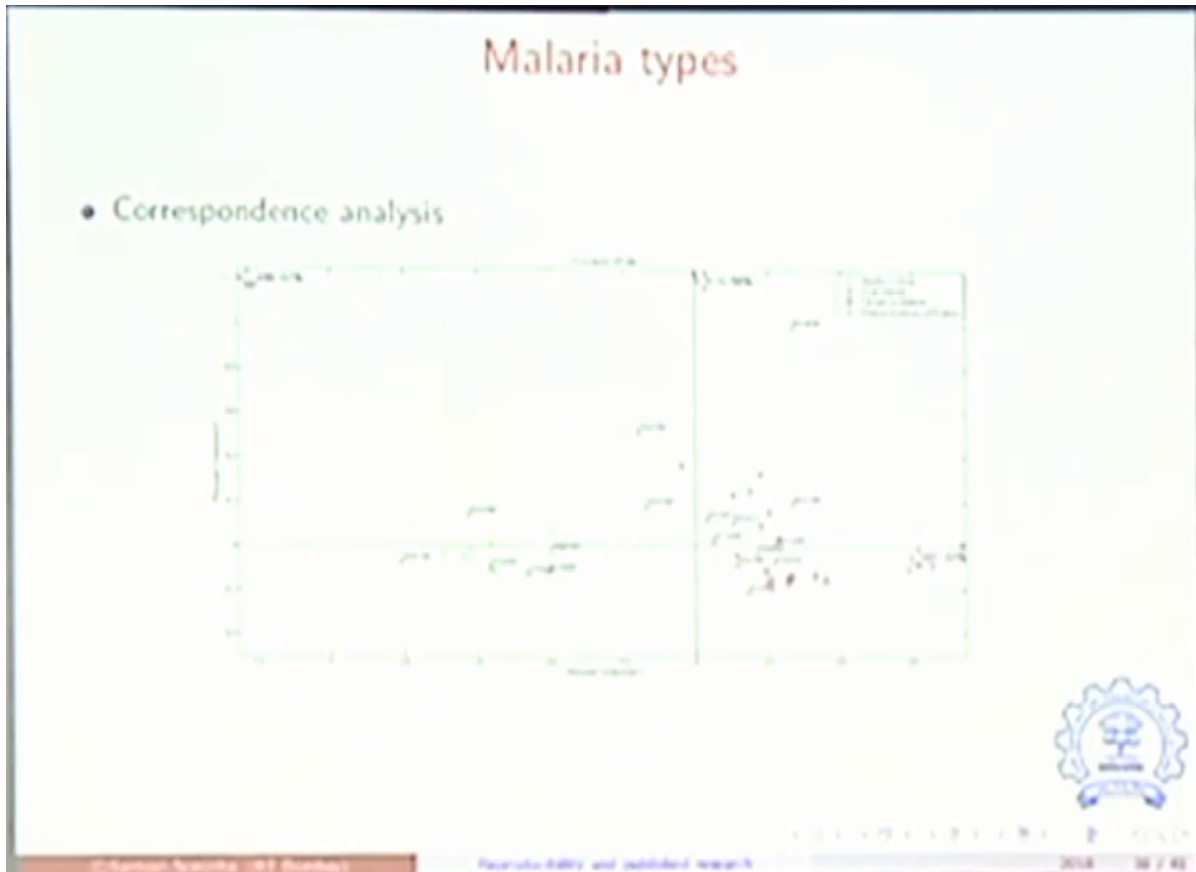
So different ways of interpreting this through different insights out what was very useful about this method was the fact that it allowed allows one to not just plot patients, but you can also on the same coordinate system plot genes. So, remember your dataset you've got different patients or different sample conditions and for each sample condition based on omics Okay? throughput you have so many gene expression levels for protein expression levels same logic. And now I'm plotting just the genes and asking these are all the normal genes housekeeping genes probably okay marginally changing the expression levels, and asked the question which genes are sitting out of the extremes with genes read 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 Okay? a correlation with other patients. What's being plotted is not raw magnitude but correlation coefficients Okay?

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 signaling 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's 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's a cluster of patients here and another cluster of patients now what are genetic signatures so these genes over here are signatures Okay? Specific to this cluster of patients Okay? 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 five percent of the time to get things wrong that can mount as an error rate if I'm doing ten thousand analysis instead the intact cloud 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'm looking at columns of patients I'm looking at rows of genes, and I'm looking at the two things superimposed and I'm 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. Okay?

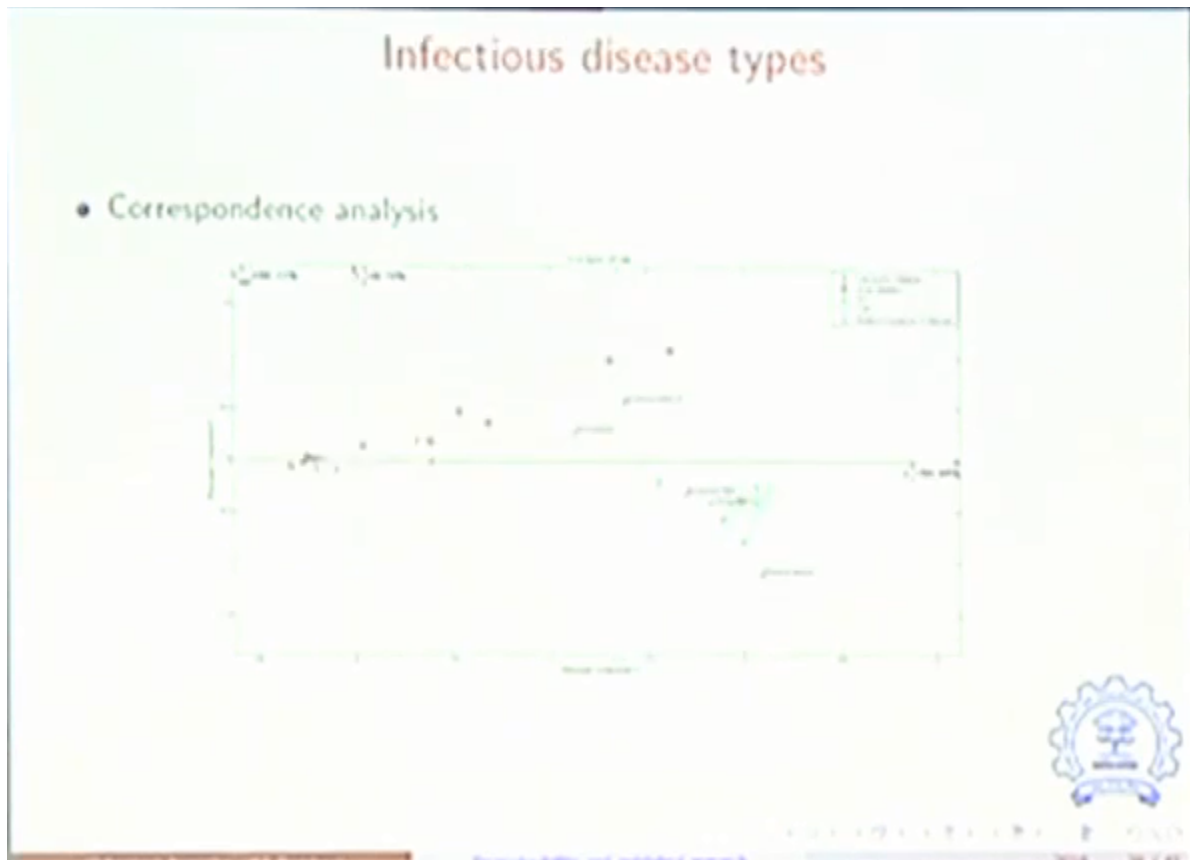
So, it turns out that in the statistics world, at least in the multivariate statistics world, the appropriate methodology for statistical analysis of this data set exists it just was the case of being a little adventurous and going out there and trying to find out could was there a method which would more accurately ask answer this question of what we are 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 Okay? 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's turned out there are nice insights about why these genetic why these genes were part of a signaling pathway and how a deflection one particular gene could escalate into this condition that has led to better science.

Refer Slide Time (37:32)



The same thing has been done later on Okay? Again, for proteomics data for different for classifying different types of infections from blood. So if you're looking so I'm not sure you can make this out other than the colors here but these are healthy patients, blood from healthy patients and they're a nice cluster on their own okay you're looking at falciparum malaria you're looking at vivax malaria and you can clearly see there's a differentiation between vivax, and falciparum that shows out when I just cluster this data at one shot. So, we're able to therefore fundamentally differentiate falciparum from vivax malaria type.

Refer Slide Time (38:12)



And in fact we've 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's an approach which does this, and not sharing all the data here but you're seeing a subset of your gene candidates and clearly these gene candidates are capable of differentiating multiple clusters.

So multivariate analysis it's not a question of one of these genes being analyzed at a time in fact you go the other way, 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 don't 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 Okay?

Refer Slide Time (39:06)

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.



So, I'm 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's several other methods out there which improve the quality of your analysis. So, in a nutshell there are several approaches and it's it's a democratic philosophy which is don't trust one method don't trust one voter, you trust many people to vote for a given candidate and these are independent statistical methods which are all seemingly voting for the same target then you've probably found a target. If one method alone is talking about a target then it's probably bad luck and surely not a significant target Okay? So that's another insight to take from this. Thank you. So, I'll stop there.

Refer Slide Time (39:47)

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



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 specificity. I hope it also reminded you dr. Jeorg rebates one of the previous lectures about a good bio worker and considerations for biomolecular discovery programs. So again, you can see that you know different experts have same opinion about the Simental design how to really find the right candidates, the right targets could be potential biomarker or drug 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 this mental 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 Jeorg well-aware gave you some very interesting lectures about the basics of these technology, as well as different applications but more focus on bio worker discovery program in various diseases. We then learnt about how to use the upper 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 screening. 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 array than making good slides by doing good type of printing then different type of applications of purified protein arrays using few prod ships they are shown to you directly with the demonstration 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 amino precipitation 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 it was not the scope of this course but this is one of the very promising technology which is helping now entire field of interact omics 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 interactors. During these lectures we also try to give you the idea there 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 an artifact 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 bilayer interferometry PLI, surface plasmon 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 NGS 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 leading scientist and a clinician dr. Sanjana Vani 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're 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.