Lecture – 39
Applications of Interactomics using Genomics and Proteomics Technologies

Welcome to the MOOC course, applications of interactomics using genomics and proteomics technology, today we have a guest lecture by Professor Santhosh Noronha, he is a faculty at IIT Bombay, who works in the areas of big data analysis and various device development, today lecture is about data analysis in evaluating the various hypothesis, like searching for a potential drug delivery target or a genetic target in a disease context. When we are trying to search for a biomarker candidate or a potential drug candidate, it becomes very crucial to know how to make a good experimental design in which manner our data should be reproducible, if you think about a test which is has to be given to the patient the test has to be reliable, reproducible as well as it has to serve the large community. So, therefore your number of patients to be included for the study has too much larger, your study design has to think about various flaws, various mistakes, various errors which might be happening, if you don't carefully consider your end goals, your actual questions to address. So, in this slide, professor Noronha is going to illuminate your understanding

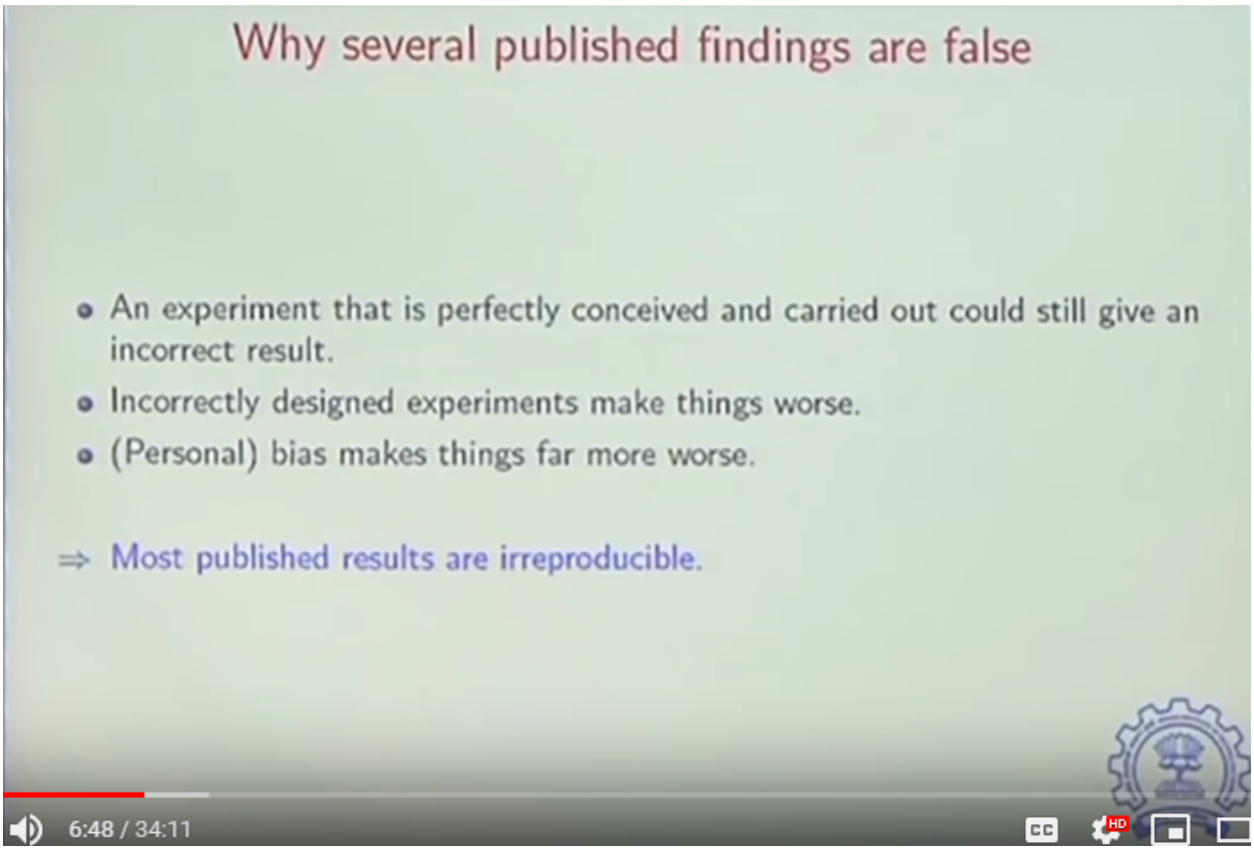
about, how one should think about the good experimental design, before actually planning and executing the omics experiments, by now you are familiar that data generation using protein microarrays or using various type of NGS platforms or various type of proteomic technologies are quite straight forward as long as your sample preparation is good, as long as you know what exactly biological question you want to ask for, however to really get a meaningful data, meaningful information it's not so straightforward and that's where you need people who are good knowledge for a statistics, who can work with you in designing good experimental plan. So, if the aim is to look for biomarkers or to choose the drug target your experimental design, your study plan, your number of patients replicates all of these things becomes very crucial. So, let us welcome person Arun, to give you this lecture about how to really carefully plan a good experimental design for omics-based studies.

Refer Slide Time: (03:10)

From what I see is a background of most of the participants here students, that to from a life sciences background most of you, there's some researchers here as well I understand. So, I thought it would be very useful and provocative both, to bring up a discussion of data analysis as it applies to multivariate experiments such as the ones you are doing in proteomics. So, a big issue with statistics is how confident are you about any insight you get out of any analysis, that you do and while clearly you're spending three to four days learning about experimental protocols, the question is, what are you going to do with this data, that is generated and what hypothesis does it drive and for that matter, how valuable is this data that is generating in the first place. So, I'm not actually going to take a case study and the reason I'm not taking any case study is because I'm not in this domain as a core domain, I'm a biochemical engineer, I'm

interested in production of pharmaceuticals, I usually do not ask questions about what's the role of this gene, in that particular disease context and so on, but one of the things as an engineer you're forced to do is to deal with large amounts of data especially in a process analysis context and this data analytics kind of need actually, we find is very useful in evaluating hypothesis which is what most scientists are concerned with. So, these are a collection of thoughts that have had over, while which I think apply mostly to scientists, less so to engineers but they have to do with the fact that we're generating data at very large scales, we generating data in very large amounts and the question therefore is given this rate at which we are generating data and the assumption by the ways that experiments we are doing, our well designed experiments which are therefore generating good data, but given this data the question is what can you infer and what are potential errors that one can make in trying to arrive at conclusions. So, that's why I'm titling this particular talk, as one on reproducibility in particular and one flipside to this is as we see a lot of publications come out especially no mixed domains, the question comes up which aspects of these are reproducible and if they are not reproducible, what kinds of mistakes are people making which prevents good insight from, Okay. Potentially being leveraged into some kind of future drug development pipeline. So, remember most of the reasons for working in this space of omics is that you're trying to find targets typically for drug delivery, Okay. So, at every level here you can be asking the question how I statistically found a good target, second given a target how I statistically found a good drug candidate, given that there's a library of a zillion candidates, have you found a good candidate could deploy, Okay? As a potential drug and if you think about what's actually occurred out there, the success rates are very low and success rates invariably are low because of issues to do, with the reproducibility and that's what I want to deal with, if you look broadly at why most published findings are false, you can break it down into different possibilities, which is that you've actually thought of a good research design for your experiment

Refer Slide Time: (06:39)



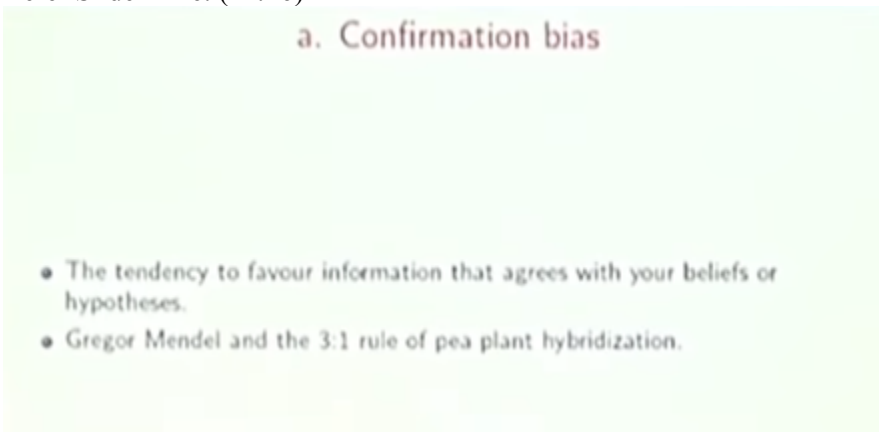
but given the fact that for example, if you are talking about disease conditions given the fact, that you might not be able to access that particular tissue type in more than one patient, therefore you are not in a position to do replicates of this as a study systematically over a larger population, you might therefore end up with your bad luck and with an incorrect result, it's kind of like saying you'll toss a coin hundred times and with your bad luck the one time you toss the coin hundred times you might have ended up, with thirty heads out of hundred, which is theoretically possible and 30 out of100 sounds, as if you have a biased coin not a fair coin, Okay. Nothing wrong with getting 30 out of hundred, it's just that it's a random event, it has to do with the fact that coin tosses are random events and with your sheer bad luck, the one time you did it you could have ended up with 30. Now, your problem is you are trying to ask the question, why is this coin, Okay. Behaving in this particular fashion, you have expected a fair coin, therefore you would have expected 50 heads out of 100 and instead when you see a 30 you're actually pushed into this question as to, whether what you're seeing is an example of an unlock event with a, with a fair coin where you should have expected 50 heads but instead you saw 30 or have you instead seen a biased coin. So, which way is this? So, in other words are you looking for a particular hypothesis and you're saying that the hypothesis is not true, are you saying that there's an alternate hypothesis is true, which way do we go? So, this is a situation in which instead of that fair coin now and biased coin, think of whether you're looking at a genetic target for manipulation in a disease context. So, is that target something that

you pulled out as a random event or is it so significant to you, that you feel that it should now go to follow-up experiments, the context of this is therefore that if you are trying to find targets and there's an investment by the way, in both the experiment that you're perform to find targets and an investment subsequently in trying to validate these targets, Okay? How confident are you in carrying out these experiments and of course the assumption knowledge is that the experiment is designed correctly, that you've done it with appropriate controls, that you have randomized, that you have got appropriate, Okay. Controlled experiments that you're doing and if you have not done that then, things only going to get worse, in terms of the quality of data that you have and even more critically, and this is a sad fact, if you as an investigator or already biased towards one cluster of genes as being important to you, then personal bias makes things even worse and the brute truth of this is that most published results are therefore irreproducible so much. So, if you go to the journal nature, is an entire subdomain on that website which talks about irreproducible research and their concern is for all those papers trying to get published in nature, how do you guarantee that whatever insights or whatever results that somebody is trying to package in the paper, how confident are you that those are reducible enough that they're worthy of being published and being published not just in any journal, but in nature. So, they are so concerned that there's a sub site that they have created on reproducibility and the lack of it, there are several causes for why there's this bias, ultimately. And, I'm just pleased, I'm going to list a few as we go along and not all of these have to do with omics, but I want you to appreciate the broad idea and finally towards the end, we'll talk how to control for some of these.

Refer Slide Time: (10:02)

But, the most important reason for why results are not of producible is something called a confirmation bias, which is you already have a hypothesis going in your mind, about what do you think is the underlying cause for this disease and now when you're generating data, you're only paying attention to those datapoints. So, that subset of data, which you think confirms your hypothesis, Okay? And in doing so because you're only looking at a subset of data point, you're probably missing out something more important, which might have told you something else about the disease condition, Okay. The other reason is very common, which is that people manipulate that analysis outright and a third situation is that the results are not independently tested, you may have already had some discussion about this, in terms of your omics pipelines but a typical headache for example, a tissue sample which you're processing for an omics study is not available to a different researcher. So, there's no way of validating whether your actual experimental workflow was executed correctly or not, therefore whether your targets are therefore relevant or not. So, let's look at these one by one. So, the oldest example of confirmation bias is actually somebody you'd be surprised with Gregor Mendel, to give you an idea of why, you know, if Gregor Mendel were to publish in this day and age, in fact he would not publish he would not be published, because you would say that he cheated, he plagiarized this data, why? You sit and do that pea plant experiment, which is famous for,

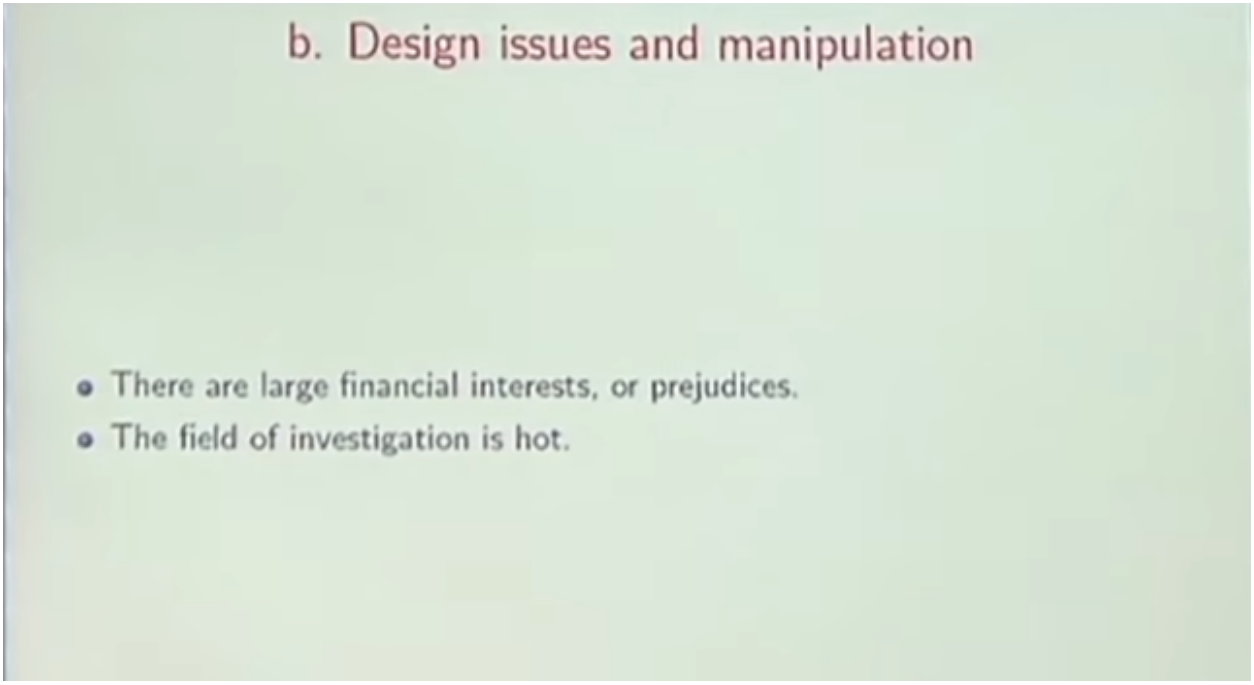
Refer Slide Time: (11:26)



various expected to get a proportion of three is to one daughter's with a certain phenotype, you do that experiment starting with pea plants, your proportion of plants of daughter plants which have, which are carrying a particular phenotype will not turn out to be 25 percent, 25 percent to give Gregor Mendel credit for something which he realized intuitively was some number to round off to, at the three is to one proportion, is probably a nice rounding off of numbers, but if you sat in with the raw experiment yourself, when you collected a whole bunch of plants and you categorize them by the length of the leaves and so on, you would not get 25 percent of them having shot leaves versus long leaves, for example, what does that tell you, it tells you that he was already biased towards reporting a result of three is to one, he wanted

a number three is to one of course there was no statistician there at that time to challenge him, of course these days there are, when you send out an article for a review process. So, when he says three is to one, whole bunch of people accept three is to one, as if it's the truth but here's the thing that experiment could never have been reproduced, Okay? And he goes into this with a specific bias, and the irony is the law of inheritance as he found it is an accurate description of how inheritance happens, but inherently you have a problem that this is not inheritive, the second reason why most results are not of producible is that there, is manipulation of those data.

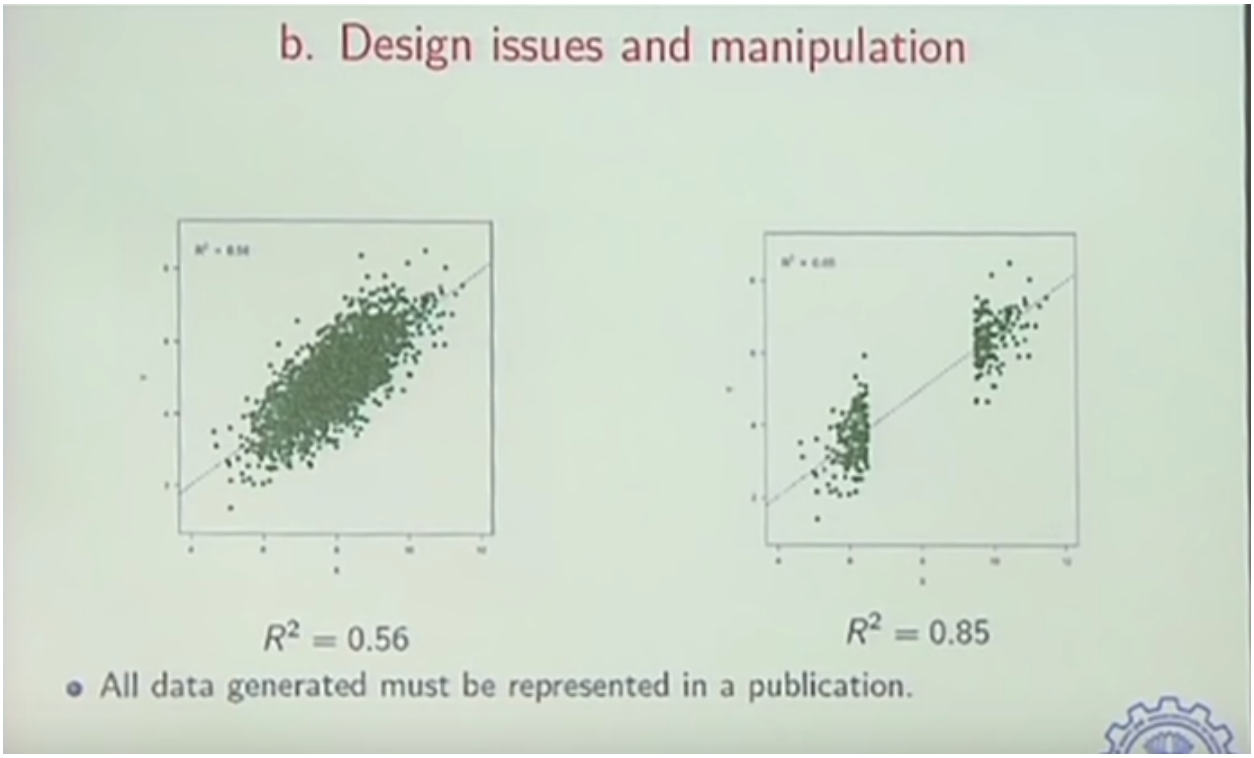
Refer Slide Time: (12:50)



And especially in large departments in the west, you're starting to see a paradigm where there is sponsorship of the research in the department by some large entity, a pharmaceutical company and that starts influencing, because the whole process of which types of hypothesis will you look at, what data will you generate, what kinds of experiments we do is actually perform and report, and you're not free therefore to actually carry out certain types of experiments, because you're using the pharma companies money, you're not free to do, what you want to do and therefore you're not truly reporting, what you think is relevant inside, Okay. So, this is an example of a manipulation happens and this is a spin-off of this is, at any given point in time, one particular type of investigation suddenly becomes hot and everyone tries to do that and that's accepted as a standard protocol for data generation, but there's nobody challenging, Okay. Because there's nobody capable of challenging given that it's not a kind of facility that's availability

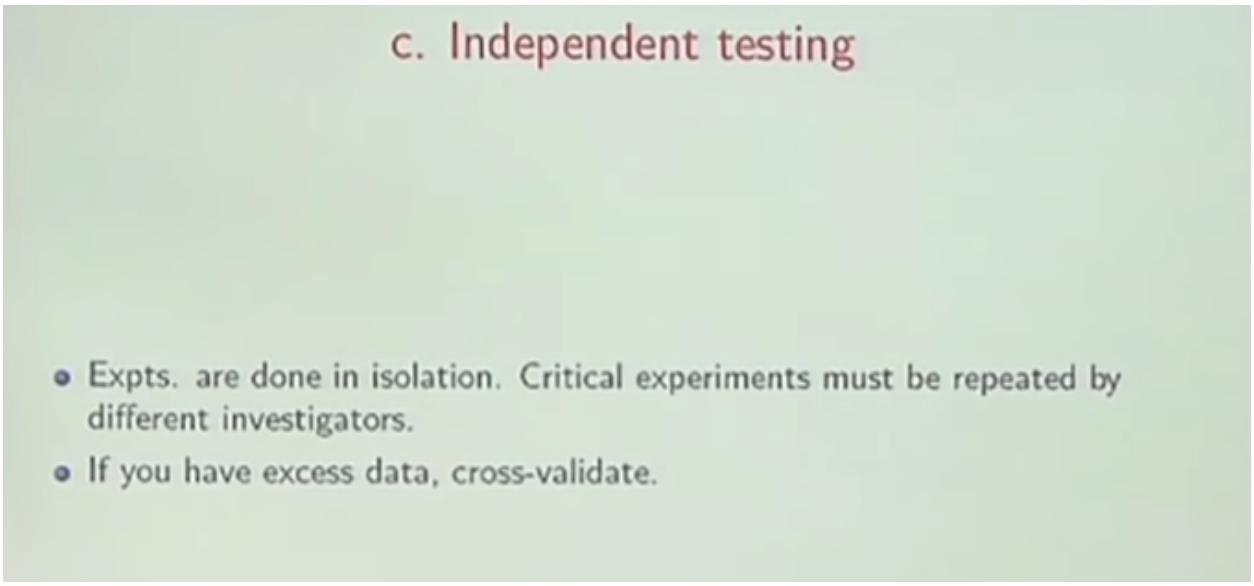
everywhere, Okay. What the truth is of that particular data set or whether there's some prejudice involved in analysis.

Refer Slide Time: (13:50)



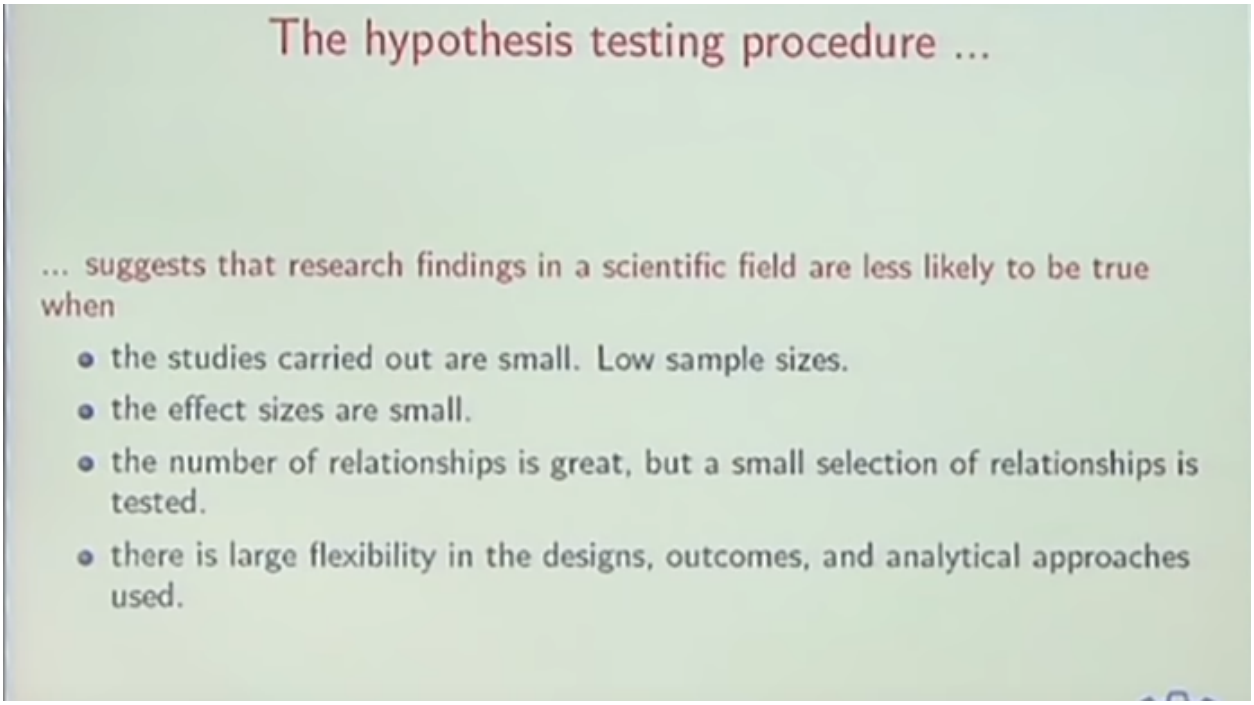
And, this is a simple example to point out, what happens if you manipulate data outright by misreporting on, in this case not reporting data. So, if you look at the plot on the right and you look at the reported as per value, it's a simple straight line fit which I'm reporting, you'll swear that it's a great line that as per value, if you just go by a published as per value, you think it's a good model, but you see how you have arrived at a good model by just omitting a few data points and the risk associated with selective, Okay? reporting of data in a publication.

Refer Slide Time: (14:22)



And a final point on this line is that when you are talking about testing an independent testing, the fact remains that we are all working in isolation and therefore, there is no culture of materials being shared and being retested elsewhere, if you're saying that something's a critical hypothesis, a critical gene for example or a critical protein, in some kind of a disease condition, the first thing you have to do, is for this protocol to be repeated by somebody else by different investigator, who ideally doesn't have access to your raw materials and therefore is independently validating that your idea can be reproduced elsewhere, Okay.

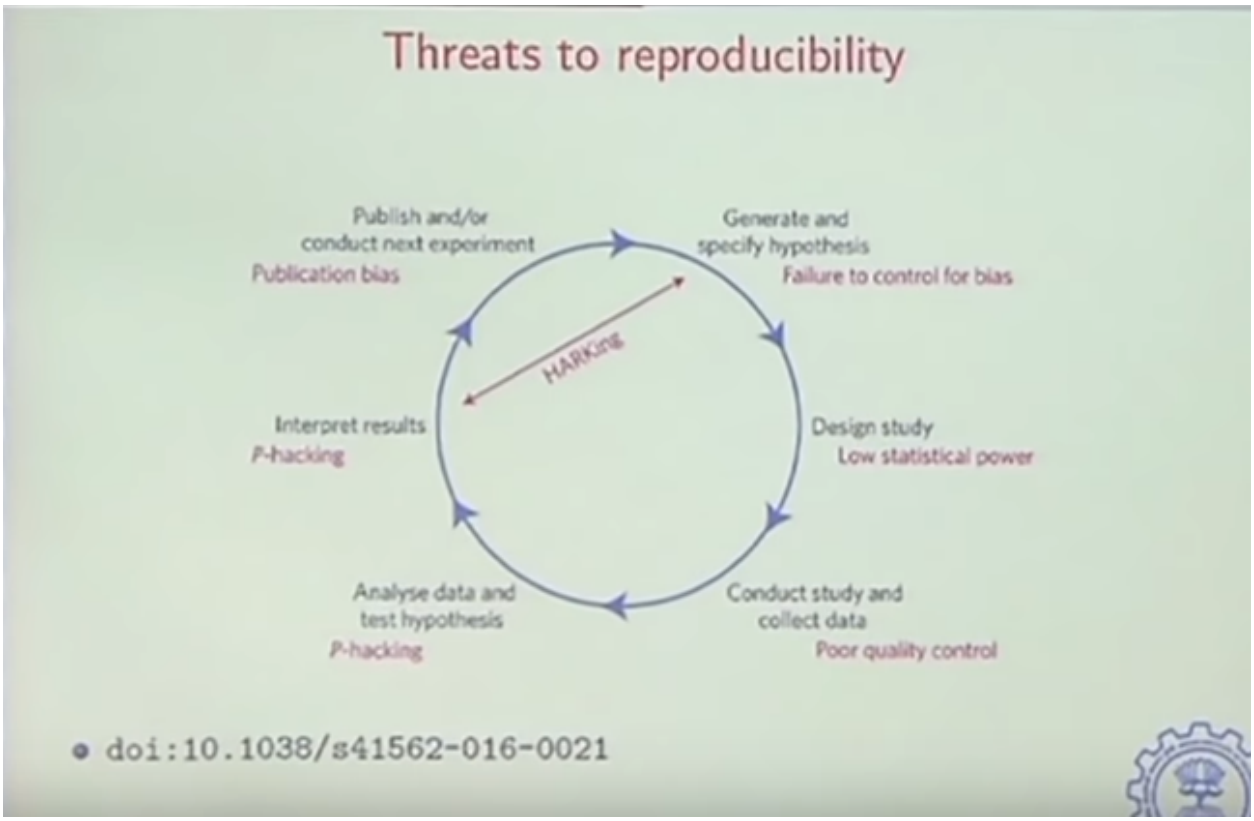
Refer Slide Time: (15:02)



All of this comes down to how you set up for the most part in especially an omics type of approach, a hypothesis, the entire statistical framework of how we evaluate a data hinges around the fact, that you ask whether something's an important gene or not as a hypothesis and then you try to shoot down that process. So, the easy way for you to remember this is actually, how the legal system works. So, how does the legal system work, whether it is in India or in the West, what what is it one tries to do somebody is what, somebody's innocent until proven guilty and there's an additional clause there, what is that clause? You're innocent until proven guilty beyond the shadow of a doubt, that's critical, beyond a shadow of a doubt, you're innocent until proven guilty. So, the, the effort is on somebody to try to prove you guilty not to try to prove you innocent. So, somebody has to try and prove you guilty, and that too that guilty has to be shown beyond a shadow of a doubt, that shadow for doubt is something that you would have heard of, Okay? as a confidence level. So, I trust a result, within a certain confidence interval and if I see a measurement beyond that particular range of confidence, I say that this result that I've gotten is not consistent with the original hypothesis, I now go with an alternate hypothesis, one simple example here is, if I toss that coin hundred times and I get 15 out of hundred heads, 15 out of hundred as heads, instead of calling the coin of fair coin, I say my outcome was so extreme, fifteen hundred that I would rather go with the hypothesis, that it is a biased coin and not go with the hypothesis that it's a fair coin. I got an extreme result, beyond a confidence interval of sorts. So, therefore the entire hypothesis testing procedure involves on taking one particular statement asking whether our observations are extreme relative to a range within which, that statement can be true and then therefore deciding on whether we are on the inside, in which case that statement is, Okay. Or, whether we are on the outside in which cases, what we have seen is

something extreme. Now, if you think about how you identify targets, you look at gene one, it's fold change an expression, is that fold change extreme, is that fold change is extreme, then you say as you shortlisted and say that how extreme it was could not have been an even by chance, therefore you believe that it's are actually a good clinical target and for all the thousands of these sets that you look at, you ask, is that been some random variation in their expression levels within what range could that random variation have been, and you therefore assign some confidence interval and when you see a measurement well beyond that, Okay. In terms of either fold upregulation of fold down regulation at this point, you say I saw something so extreme it could not have been by randomness, could not have been by chance, therefore is something going on there as to why this particular candidate is over or under expressed, Okay? However, what we know is that these findings, these kinds of studies in terms of statistics, they are limited for several reasons and and in principle the reason that these findings are less likely to be true, is if whatever you're looking for is a small effect, then odds are that you know you are going out on a limb by saying something's a significant gene, when all you were out to do in the first place was to create a small effect, Okay? Odds are you're found a wrong candidate, also if your effect sizes are small, if your sample size is small and you have really really no business saying, that something is an extreme result and if you are looking at many relationships, if for example in the omics case if you look at many genes and trying to make the argument, that this chain is important, that gene is important and you've got a whole bunch of genes to analyze one after another, the larger that pool of things being tested, odds are you making mistakes and come back to that in a minute, but take these points, Okay. In the context of what threatens reproducibility,

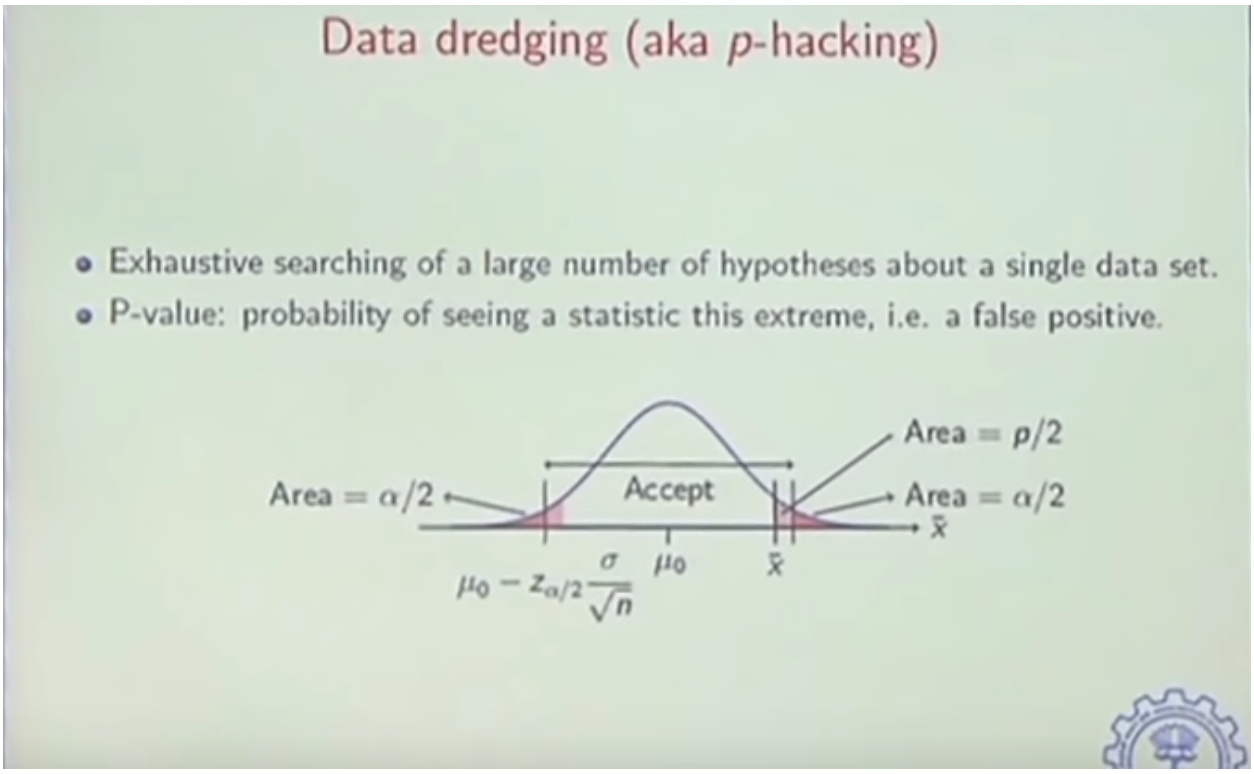
Refer Slide Time: (18:58)



I'm not sure you can see this very clearly, some interesting buzz words, interesting terms which later on I really recommend all of you go look up, Okay. So, if you look at, if you look at the cycle and in fact your cycle because it's a publication, publication cycle, it's how one does research with publications in mind. So, if you look at the way you're supposed to operate, you're supposed to operate at the top right there will generate, very specific hypothesis, you generate the hypothesis, you are supposed to not design experiments, collect samples, Okay? Designers study and when your design is studied by the way you are supposed to design, what is called a powerful study. I'll come to that power a little in a little bit, but your design a study which is a protocol which is going to help you discriminate between a control and a test case, that's that's what you mean by designing, a study well and you then actually conduct the study and collect data. So, that's at the bottom, again notice that in red I'm putting down things which can go wrong at each level. So, when you're generating and specifying the hypothesis, you might already be biased, you're already looking out for a particular result, Okay. And you might already be biased in how you're, therefore trying to now carry out the study, if you think about how tobacco company is operated for a while in carrying out trials as to whether nicotine is addictive and this paper after papers saying from a tobacco company saying nicotine is not addictive, there was a bias. Because they cannot be in a situation, where the report that nicotine is addictive and harmful, because obviously they will kill their own business. So, you can see they're going to treat the way their results are, and the study is done, Okay. To

end up with the result which is obviously good for them. So, you generate and specify hypothesis, your design and study and then you're worried about, how capable you are of differentiating your control result from a test result, you conduct the study and then collect data and usually a problem here is, that you don't have good quality control in terms of your methodology and then finally you get down to what's called data analysis, and some of you, if not all of you have probably heard of something called a p value, how many of you have not heard of a p value, Okay? We all heard of a p value. So, let's discuss that p value, because a lot of omics critically hinges on the interpretation of a p value and what you can do with it, Okay. There's in fact, unfortunately a phenomenon called p hacking and again you really want to look up p hacking, you have not seen this before which comes about where people use the p value concept without having a true understanding of it, and try to label some targets as being important and others as unimportant. So, their interpretation of this data via hypothesis testing approach leads people, usually once you get this data then you're going to do some follow up experiment, which means you're in the next publication cycle, you set up some follow up activity and now maybe you're testing these targets or you're testing different tissue samples using this popular, this particular protocol, you would normally get locked into some kind of a cycle, where if you have already published group 2 papers on a particular cell type and with a particular subset of genes, you are forced into publishing more just to offset the costing of their search, that you're doing and this is where a bias can creep in, because you're not in a position to say that what you have been doing all along this poor, you'll notice that there's this double line as an arc cutting across, Okay. The circle and never written harking there, harking hark refers to hypothesis, after the results are known, which is unfortunately what happens to most of us, which is we generate the data and then we start asking, Okay. What what exactly are we trying to as an insight, because at this point you've put in a lot of time and effort into your study, you better come up with an insight. So, therefore you ask the hypothesis after the results are known which fundamentally is cheating, because you know you can decide what you're going to call significant, what you want to call insignificant and this goal post business, if you go, if you we're ever going to have a goal post to decide what's a good target, what's a bad target, that should have been done before, you even looked at data, otherwise you have the potential to cheat.

Refer Slide Time: (22:57)

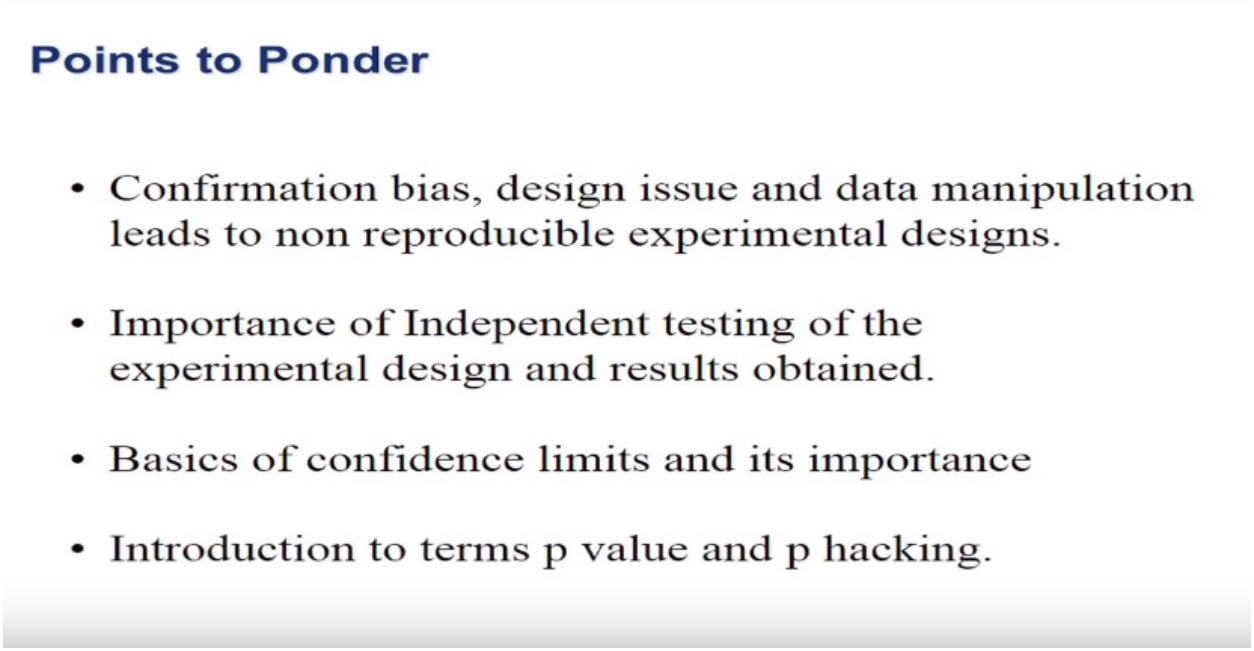


So, there's P hacking business is the crux of most issues to do with reproducibility. So, let me, Okay. Quickly go through that concept Okay. P hacking also has a nice phrase called data dredging, basically involves you searching through a large number of hypothesis, typically with a single data set. So, the data set for example, could be that you have one genomic data set or one omics, proteomics data set and you're asking the question which subset of candidates are important as targets, how does a hypothesis testing procedure work? So, if you look at this, what what what does this mean, we expecting the fact that in an experimental condition, this should have been the average value, I should have seen, if replicates wherever done. Now, there's already a practical problem which is in an omics world, you're probably not going to have the luxury of doing many replicates, because it's a very expensive protocol ready executing and you may not even have tissue samples to play around with, Okay? If you had the luxury of doing many replicates, you would see a range of values for the same gene, that range of values would follow some kind of a distribution like this, Okay. And, with your luck you may end up with the result anywhere under this distribution just like, when you toss a coin hundred times it is possible for you to get 50 heads out of hundred, but it's also possible for you to get 30 out of hundred and for that matter 70 out of hundred, these are all possible outcomes, given that you are not in the business of doing the experiment again and again and again, notice that statistics ideally says, do things again and again and again and average them out, because then you're more likely to find the true value intuitively, you know that that if you work with for example, if ask you what's the average height of a person, what's the average height of

a human being, either I can work with your one, I can take one individual and use your height as a representative of the average, but you know that's not smart, because I might end up with a short person or a tall person instead the safer thing to do is what take an arithmetic mean of everybody out here and then use that as an estimate of the average, you know that you know that's a better representation of an average, Okay? So, the whole notion, that results can occur under some distribution, Okay. But, if this is what you expect, then we know that between these two boundaries here, between these two boundaries, we're in a situation where for example, if this were 50 heads out of hundred, this could be 30, this could be 70, then we are in a situation where any outcome, we see the one time, I do my experiment between 30 and 70, any time I do the experiment once and I see an outcome between 30 and 70, what do I say, then what do I infer as the hypothesis result, I say that my result is close enough to 50 and I probably saw a result associated with this is a fair point, it is only if I see an extreme measurement sitting out here or sitting out here, that I said look what I just saw is so extreme, I saw 15 out of 100 or I saw 85 out of hundred that is so extreme it could not have been because of a fair coin, therefore I should be in considering this coin to be a biased coin, Okay. That's how we interpret typical hypothesis, now if you look at this, these two barriers which by the way of confidence and rules and for us typically, what is a magnitude of this conference interval, what do we choose? Anybody has an idea? What what magnitudes do you choose? You would have heard of a 95% confidence interval as a typical setting for confidence intervals. So, there's 95% Okay. Of confidence, what you're saying is 95% of the area under this curve probability, which is probability is between this threshold on that which means the 5% sitting outside. Now, the headache is how do you interpret this 5% sitting outside. So, what does therefore this dark red shaded area mean to you, that dark red shaded area is a situation, where you could have had a fair coin, because the whole thing is associated with being a fair coin, these are the range of outcomes you would see with a fair coin, you could have had a fair coin, but it's sheer bad luck, you see an extreme result like 15 out of 100 or 85 out of 100, and at this point, you go with the other hypothesis, that it's a biased coin, but it could have been a rare outcome with a fair coin, in which case as a technicality you've made a mistake in your hypothesis, why because you're gone with the other suggestion, are the other hypothesis than with the original hypothesis you have gone with the assumption, that the coin is biased, then you should have gone with the assumption of the coin is fair and what you see is a rare event with a fair coin. So, the five percent, which is something we tend to take for granted is actually critical in a no omics test, because it represents an error, it represents an error in interpretation. So, this five percent is an error, where we should have gone with the null hypothesis as this is called, we should have gone with that original hypothesis, but instead we are preferring to go with an alternate belief, an alternate hypothesis. So, five percent of the time, we commit an error in an analysis. So, we commit mistakes five percent other time. So that's the take home message of that but how many times did you carry out to study, how many

of, how many hypotheses did you test in an omics study, how many genes are you testing? If you are testing 10,000 genes, you know study one after another and I just told you, you're committing a mistake 5% of the time then, have you committed mistakes in your analysis of 10,000 genes, you would have, somewhere in there, you would have ended up with the wrong conclusion just by sheer bad luck, because your error rate is 5% the error rate doesn't matter much, if you are testing something once ,what are you tester something 10,000 times. Well! you didn't test one thing 10,000 times, you carried out 10,000 different tests one per gene and if you tested 10,000 things one after another you've made a mistake by sheer bad luck at least 5% of the time, which amounts to a large number of candidates potentially being wrong, Okay? Therefore, if you are looking at your null hypothesis, being this gene is not important and you're trying to shoot down the null hypothesis. So, if my hypothesis is this gene is not important because it's expression fold change is one and I'm looking at fold up or down regulation close to one, then any range around one corresponds to my null hypothesis being right and what is the null hypothesis that gene is not important and it is only when I see an extreme measurement would, I now say this gene is important but 5% of the time for a given gene, I would make a mistake and I just did this analysis 10,000 times one after another, which means most of the times, I'm actually calling something important as a gene candidate which is up or down regulated, it is not even statistically important to you it could have occurred by sheer chance, sheer chance why, because we did the experiment few times and with sheer bad luck we are seeing extreme results, and we are being fooled into thinking that our insights are important, when they are not, Okay. So, the p value, which is now, so let's say our outcome is over here, if my outcome is over here, you then ask the question, how far away from this 95% threshold was I and how close am I to changing my mind, Okay. And this p value, if this by the way, if the five percent of the area outside this threshold on the right, there's slightly more than five percent of the area outside this threshold that's intuitive, because I just move the goalposts inside. So, there's more than five percent of an error that you, Okay. Committing the question is how close you to this threshold, what are the odds of this error being extreme, Okay.

Refer Slide Time: (31:19)



I hope you are convinced that there are many minors, but very crucial considerations about a good experimental design. I hope you have learned, how to analyze your data and to make best sense out of it, while thinking some very minute aspects of experimental design and data analysis, he also studied how conformational biases may lead to missing out some very important information from your data sets, also there are various reasons like confirmation bias, design issues, data manipulation along with the lack of independent testing, which could lead to the non-reproducible experimental design and if you want to publish in very good scientific journals, you need to ensure that all of these considerations have been met, I hope you have also learned about the importance of confidence interval in the selection of a potential reliable candidate, we have also learnt about how p value could affect the interpretation of results on your data. So, you need to be very careful about knowing some of the terminologies using statistics, how these tests could be performed and how one should really make a meaningful interpretation from the this data set, which is available to you which is usually very large data set from the omics experiment but what is the most significant leads out of that needs, lot of consideration needs, a lot of considerable thought process and your understanding from a experimental design to the various type of tests being performed and then only finally you can come up with a reliable list of the proteins, biomarkers or candidate drug targets which could be meaningful for the future experiments. So, the next lecture will also be continued by professor Noronha, and he will further talk to you about various factors involved in good data analysis and experimental designs,
Thank you.