Next Generation Sequencing Technologies: Data Analysis and Applications

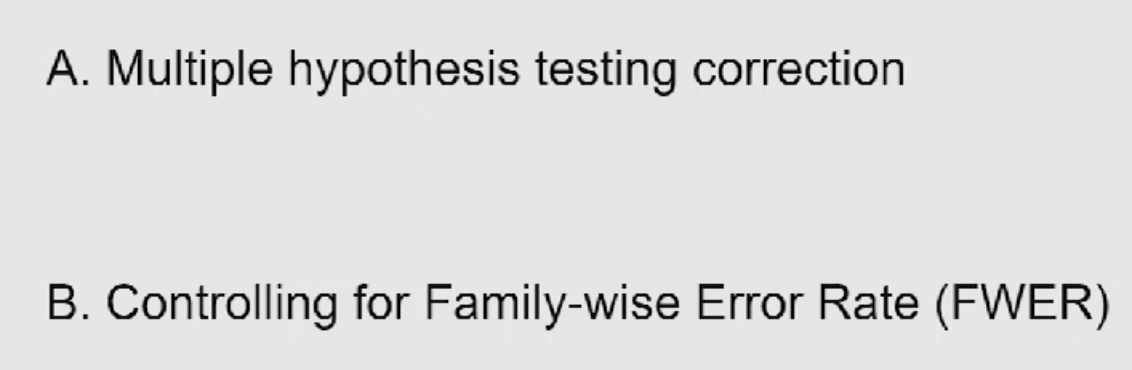
Multiple hypothesis testing correction

Dr. Riddhiman Dhar, **Department of Biotechnology**

Indian Institute of Technology, Kharagpur

Good day, everyone. Welcome to the course on Next Generation Sequencing Technologies, Data Analysis, and Applications. In the last couple of classes, we talked about differential gene expression analysis. We used DESeq2 to actually do the analysis, and at the end, we actually looked at some of the results that we got from that analysis. So, when we looked at the results, as you remember, we actually saw something called p-adjusted. So, we saw the normal p value.

Along with that, we saw something called p-adjustment. At that time, I mentioned that p-adjusted is something that we will discuss later. This p-adjustment is an adjustment based on something called multiple hypothesis testing. So, in today's class, we will be talking about this multiple hypothesis testing correction, and we will talk about the methods that we can use to actually generate or correct for this multiple hypothesis testing, and we will also generate the p adjusted.

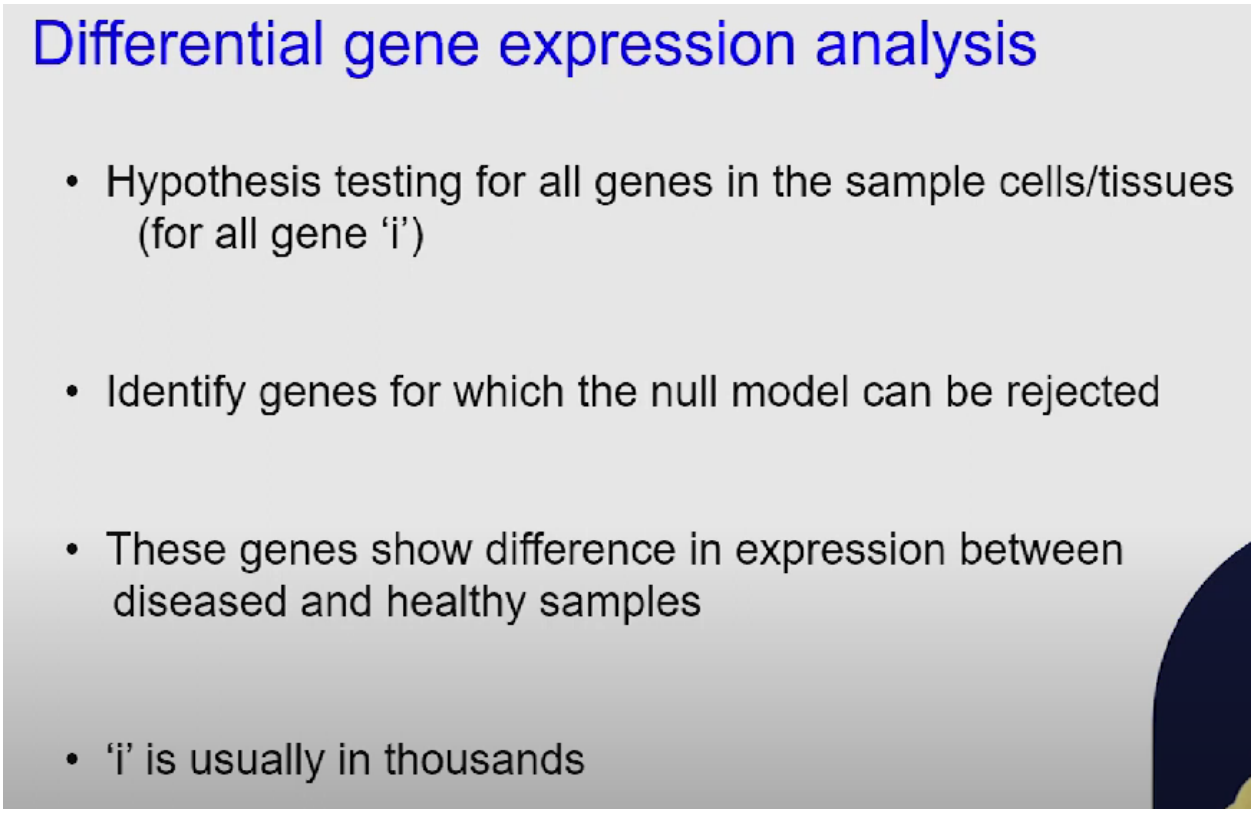


There are different methods, and at the end, I will also mention which method we use for getting this p-adjusted value in DESeq2. So, let us begin. So, here are the concepts that we will be talking about in this class. So, the first is multiple hypothesis testing correction, and the second is actually one of the methods that we can use for multiple hypothesis testing correction, called family-wise error rate correction. So, these are the two topics that we will be covering in this class. So, just to remind you of what we talked about in differential gene expression analysis, So, one of the integral parts of differential gene expression analysis is hypothesis testing. We talked about different methods for how we can do this hypothesis testing and especially for DESeq2, we use something

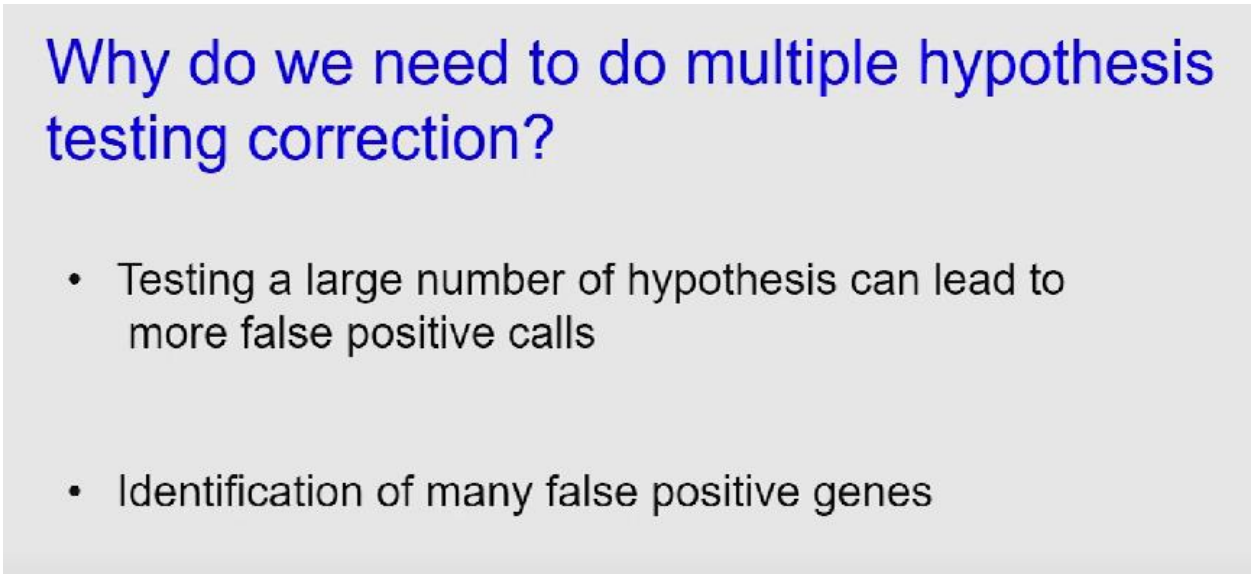
called the Wald test. So, this hypothesis is tested for all genes in the samples, cells, and tissues. So, for all genes i, if you remember the notation, this helps us identify genes for which the null model can be rejected.

So, when you do the hypothesis testing, we have a null model, and we also have an alternative hypothesis. So, the null hypothesis is that there is no difference in the expression of gene i between the samples. So, these samples could be disease versus healthy, control versus treatment, etcetera. So, that is the null hypothesis, and based on this hypothesis testing, we get this statistical significance, this p value, and based on that p value, we can decide for which genes we can reject the null hypothesis. So, this actually helps us identify the genes that show differences in expression.

So, again, this is the most important part. We want to identify the genes that show differences in expression between, for example, disease and healthy samples. So, these are the genes that we look for; this is the goal at the end. Now, this is usually in thousands or tens of thousands, depending on the organism. So, if you are talking about simple bacteria like E. coli, coli, we have 4000 genes; we are talking about yeast, which has 6000. If you are talking about human 20000 to 25000 genes, but if you talk about transcripts, if you are also considering the isoforms, then that number would increase even further. So, this is usually in thousands or tens of thousands. So, which means we are doing this hypothesis testing for tens of thousands of genes, right? So, this is a huge number of hypothesis tests, and we will see what actually happens when you test so many hypotheses at once.



Now, the question is, of course, why do we need to do multiple hypothesis testing and correction?



As the name suggests, when you are testing a large number of hypotheses, you should make this correction, and by the way, this is not just applicable for differential gene expression analysis. This is applicable not just to biological science or medical science; this is applicable everywhere. In

today's age, where you are actually probably testing a lot of hypotheses, you need to do this multiple hypothesis testing correction. Otherwise, you will get erroneous results and erroneous inferences. So, as you will see, you need to do this for all cases where you are testing multiple hypotheses, especially when you are testing a large number of hypotheses. So, the reason is very simple: testing a large number of hypotheses can lead to more false positives. So, I will mention what is false positive in a moment for those who are not familiar with all these concepts or statistics, which means in our case we will identify many false positive genes as differentially expressed. So, we will identify many genes, and by the statistical test, we will say that these genes are differentially expressed, but in truth, in the actual scenario, they are not. So, that means we are identifying them as differentially expressed when they are not, and maybe we will have some wrong inferences. When you want to interpret the data further, we will make the wrong inferences, ok? So, we want to avoid these false positive genes or false positive calls, and this is true for many situations where you are doing this kind of multiple hypothesis testing. So, the question is now: how does testing a large number of hypotheses lead to a higher number of false positives? I just mentioned that if you test a large number of hypotheses, you will get more false positives. So, the question is, why is this right, and how does this happen? So, I will show you two ways this can happen. So, one is the number of false positives, and the total number of false positives that you can expect will rise if you test a large number of hypotheses. In addition, the probability of getting a single false positive will also increase by many fold. So, these are the two points that I will just show you now, ok? So, let us start with the idea of false positives, etcetera, if you are not familiar with them. So, in a statistical test, you have two types of errors: the first is a type 1 error, and the other is a type 2 error. And if you consider this table right, then on one side on top you have the actual situation this is the actual truth, ok, and you have H naught is true, the null hypothesis is true, and then you have the null hypothesis is false, right? So, if you think about differential expression analysis, some genes are differentially expressed in actual situations, and some genes are not. So, you have H naught true, which means these are the genes that are not differentially expressed and then you have H naught false, where those genes are differentially expressed. So, our null hypothesis was that the gene I is not differentially expressed, so they are not showing any difference in expression between these two sets of samples.

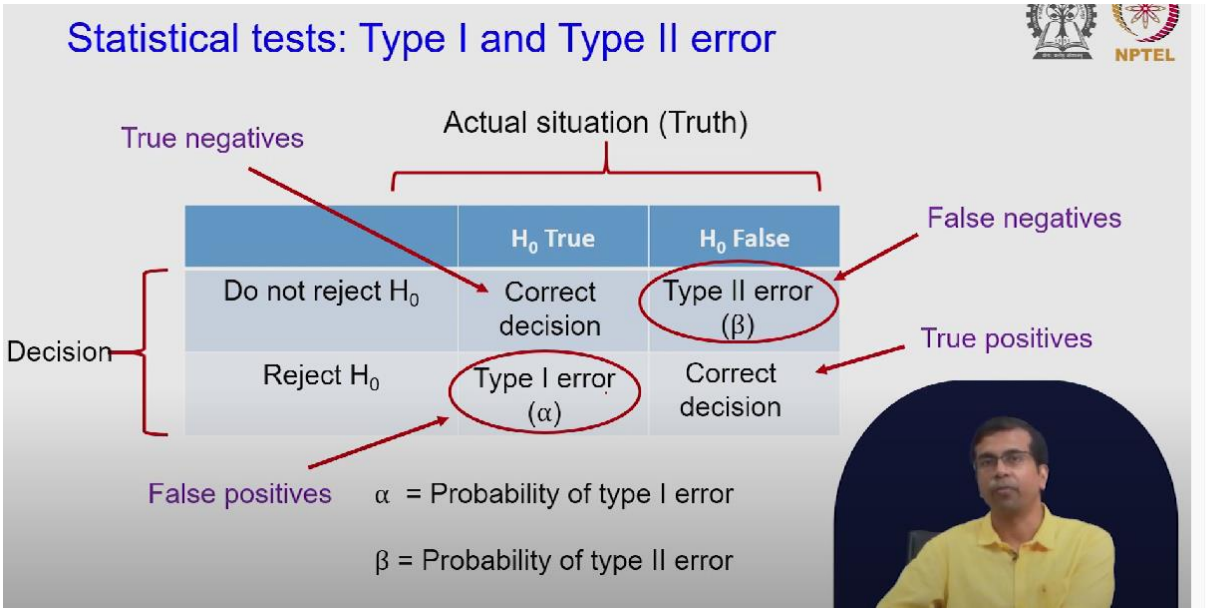
So, that is the actual situation, right? What is that what we call the truth or ground truth? On this

left side, we have something called a decision. So, what is this decision? So, this is the outcome of the statistical test, whatever test you are doing right, whether it is the old test or some other test, whatever is right. So, whatever test you are doing, the decision is ok based on the p value of the statistical test, and if the p value is below that certain threshold, let us say 0.05, we reject the null hypothesis, and if the p value is greater than or equal to 0.05, we do not reject the null hypothesis. So, that is the decision that you make based on the p value from the statistical test. Now, imagine this situation, right? So, you can have four scenarios. We have the ground truth and the decision from the statistical test. Now, when H naught is true and we are not rejecting H naught, that is a correct decision, right? The statistical test is making the correct decision.

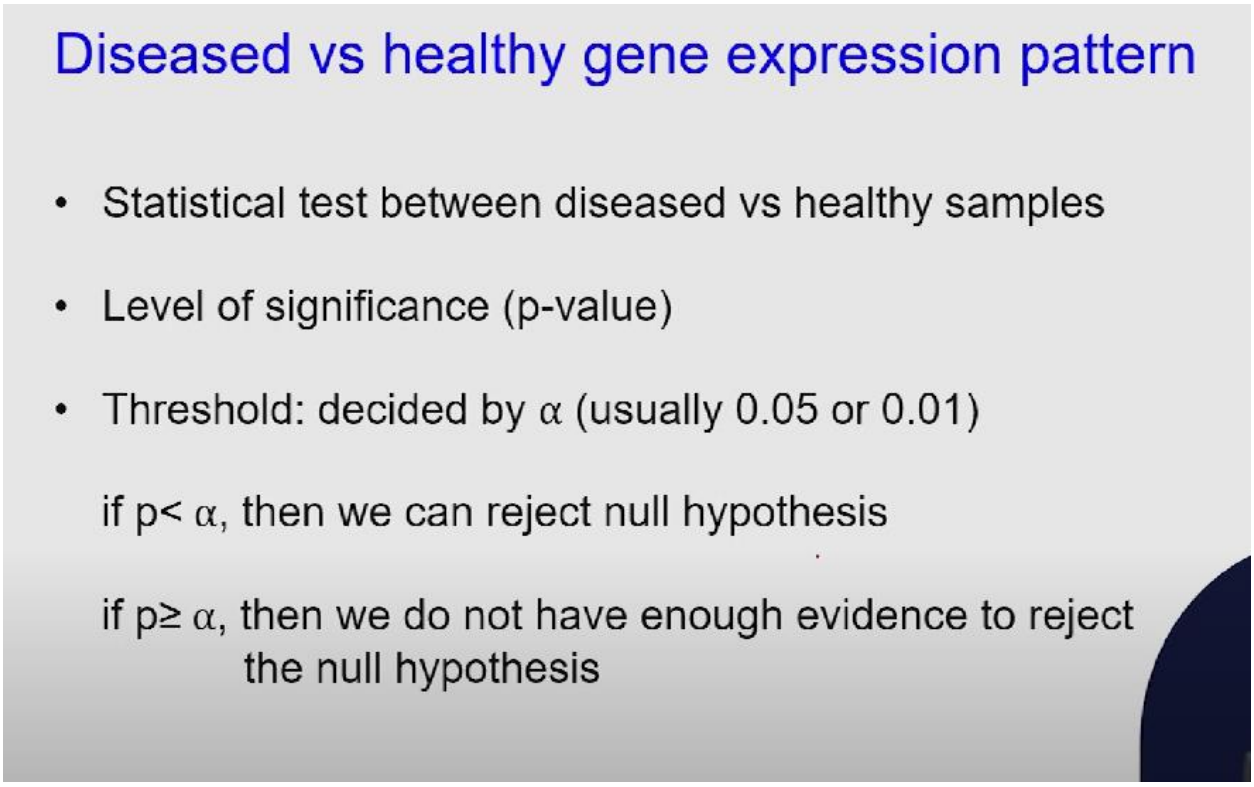
When H naught is false and we reject H naught, then that is also a correct decision, right? So, if a gene is actually differentially expressed and the statistical test says OK, this gene is differentially expressed, that is making the correct decision. But then you have the two other situations, right? These two empty cells, as you can see here, So, the first one here is H naught false; do not reject H naught right. So, even though H naught is not true, the statistical test is not rejecting H naught, and similarly, when H naught is true, it is falsely rejecting H naught. So, these two are errors, right? The statistical test is making some errors, okay? And this first one is called type 1 error, and the number alpha is given as the probability of type 1 error. And then you have the second type of error where H naught is false and we are not rejecting H naught, which is called the type 2 error or beta ok. So, again, beta is the probability of type 2 error, ok? So, if this is clear now, which are the false positives and which are the false negatives? So, these are the false positives, right? So, H naught is true, and we are rejecting H naught. So, the null hypothesis in our case is that the gene is not differentially expressed between two samples. But the statistical test is saying this gene is differentially expressed.

So, reject Haught. So, this gene is differentially expressed. So, this is a positive call, right? So, we are identifying this gene as a differentially expressed gene. So, this is a positive call. So, this is why it is falsely positive, right? So, this gives rise to false positives. On the other side, you have the false negative. So, where you have H naught is false, and do not reject H naught. So, we are not rejecting Haught, ok? So, that is false negatives. And then, of course, you have the true

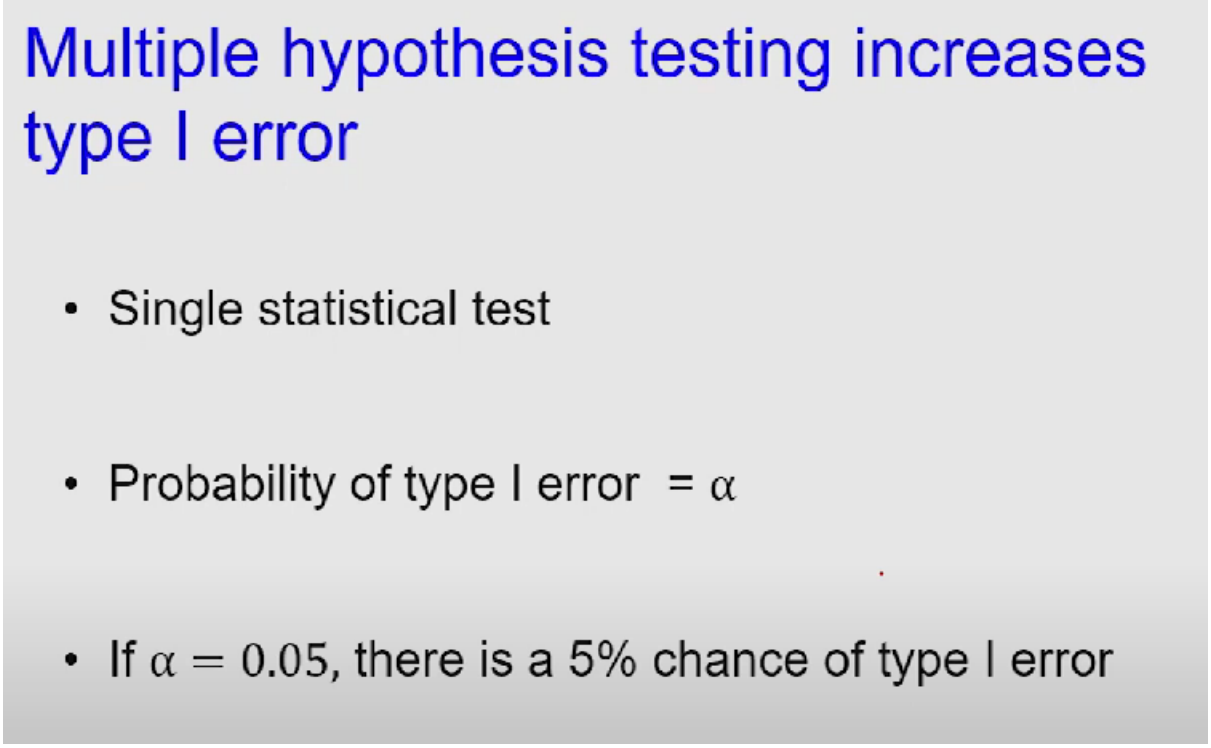
positives and the true negatives, which are the correct decisions. We are making the correct decisions. So, that is why they are true positives and true negatives, okay?



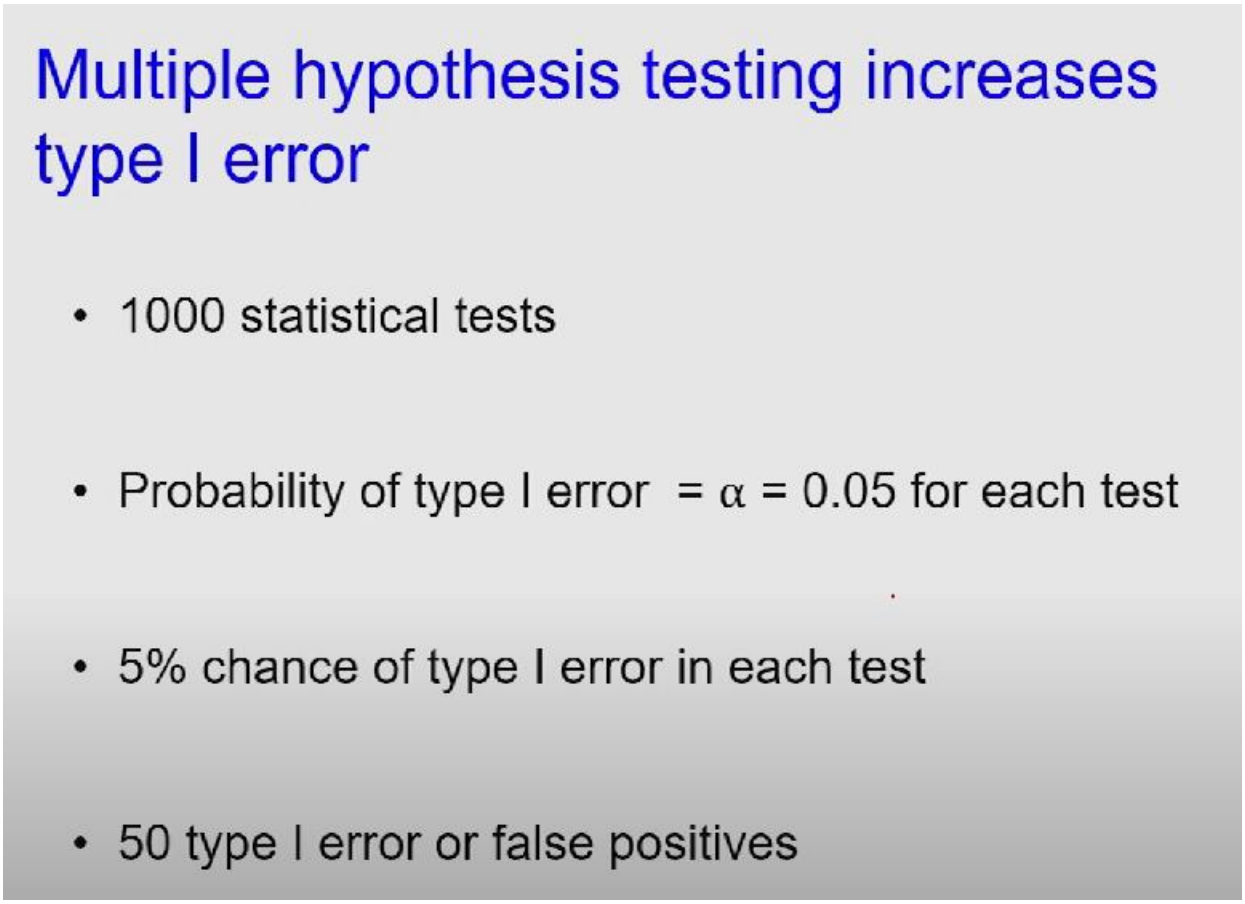
So, once this is clear, we will see right away how this type 1 error can increase when you actually do multiple hypothesis testing. So, this is a situation for individual statistical tests, right? So, you can imagine this scenario for each gene. We are doing hypothesis testing, and you have this table of truth. So, where do you have the actual situation versus the statistical decision? So, let us now imagine this disease versus a healthy gene expression pattern. So, what we do is a statistical test between this diseased and healthy samples. We determine this level of significance, and then we decide this threshold. This is actually decided by this alpha ok, which is usually 0.05 or 0.01 ok. This is related to the type 1 error. So, if p is less than alpha, then we can reject the null hypothesis. If p is greater than or equal to alpha, then we do not have enough evidence to reject the null hypothesis.



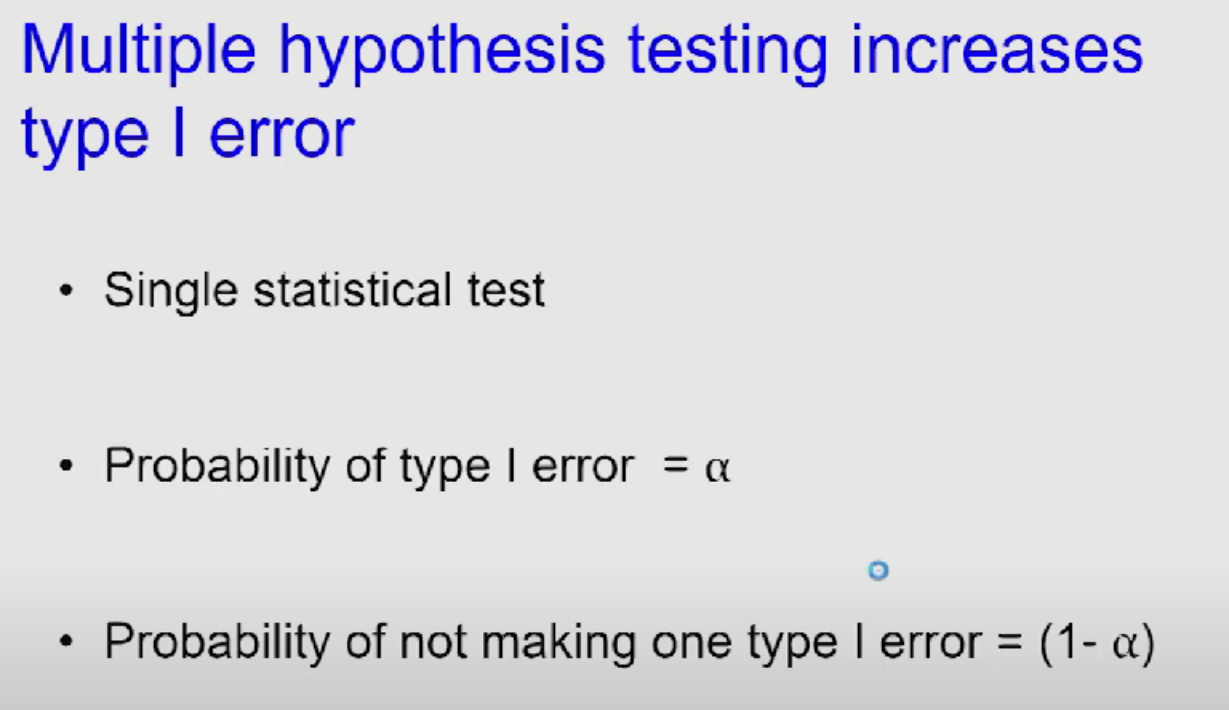
So, what it means is that when you set this alpha, we are expecting this; we are setting this probability of a false positive call. So, we want to avoid or minimize this positive call ok. So, imagine this single statistical test, and the probability of type 1 error is alpha ok, and if alpha is 0.05, which is usually the case, then there is a 5 percent chance of type 1 error ok. So, that is how we are setting this: 0.05. If you set this to 0.01, we have a 1 percent chance of a type 1 error.



Imagine this situation, which we are doing in the case of differential gene expression analysis. So, we have let us say 1000 statistical tests are okay. We are doing this individual test for the regime that is there, and the probability of type 1 error is still the same alpha. We keep this threshold of 0.05 for each test, and there is a 5 percent chance of type 1 error in each test. So, this has not changed; the alpha remains the same, but since we are doing 1000 statistical tests, given this 5 percent chance, we will expect about 50 type 1 errors or false positives.



So, you can see now that the threshold that we applied for a single statistical test will not be able to contain this when you do this 1000 statistical test. So, we will have a lot more false positive calls, if that is right. Now, the second situation is that it will also increase the probability of a single false positive call. So, we can look at it in another way, right? Again, the same situation, ok? Now, the probability of not making 1 type 1 error is 1 minus alpha, right? So, if the probability of a type 1 error is alpha, then the probability of not making a type 1 error is 1 minus alpha.



And let us say we are doing a statistical test, and the alpha remains the same for each test. So, since we are assuming they are independent, this statistical test is independent. So, to calculate the probability of not making 1 type 1 error, we can multiply 1 minus alpha into 1 minus alpha, etcetera. So, this will be 1 minus alpha to the power of n, ok. Now, what does it mean, right? So, if so, what is the probability of not making one type of error? So, if you want to calculate the probability of making 1 type 1 error, this will be given by 1 minus 1 minus alpha to the power of n.

Multiple hypothesis testing increases type I error

We are doing 'n' statistical tests

Probability of type I error = α for each test

Probability of not making one type I error

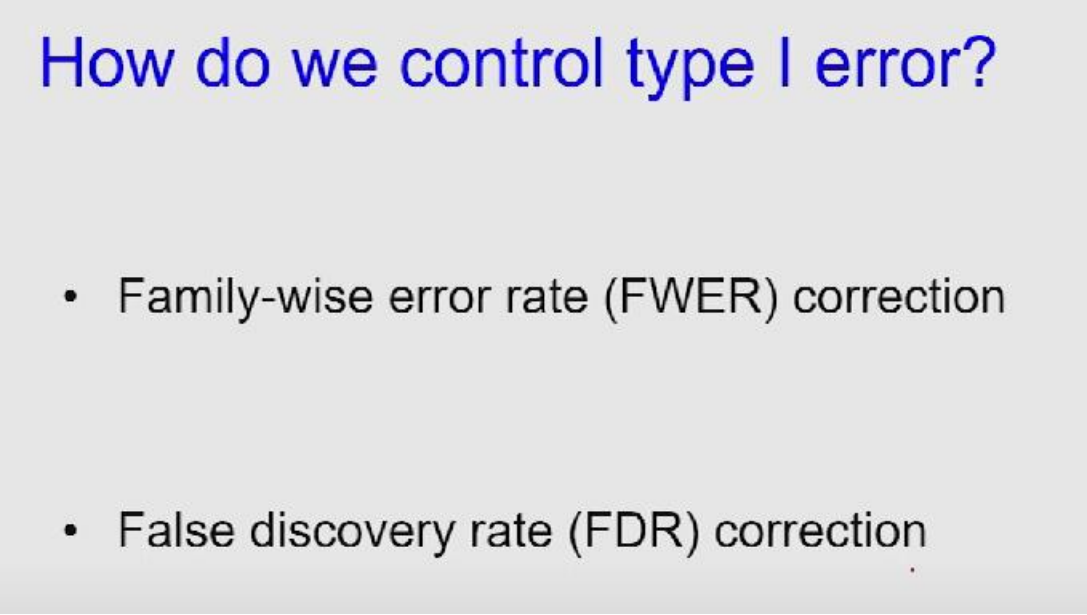
= $(1- \alpha)\times (1- \alpha)\times \ldots$
= $(1- \alpha)^n$

Multiple hypothesis testing increases type I error

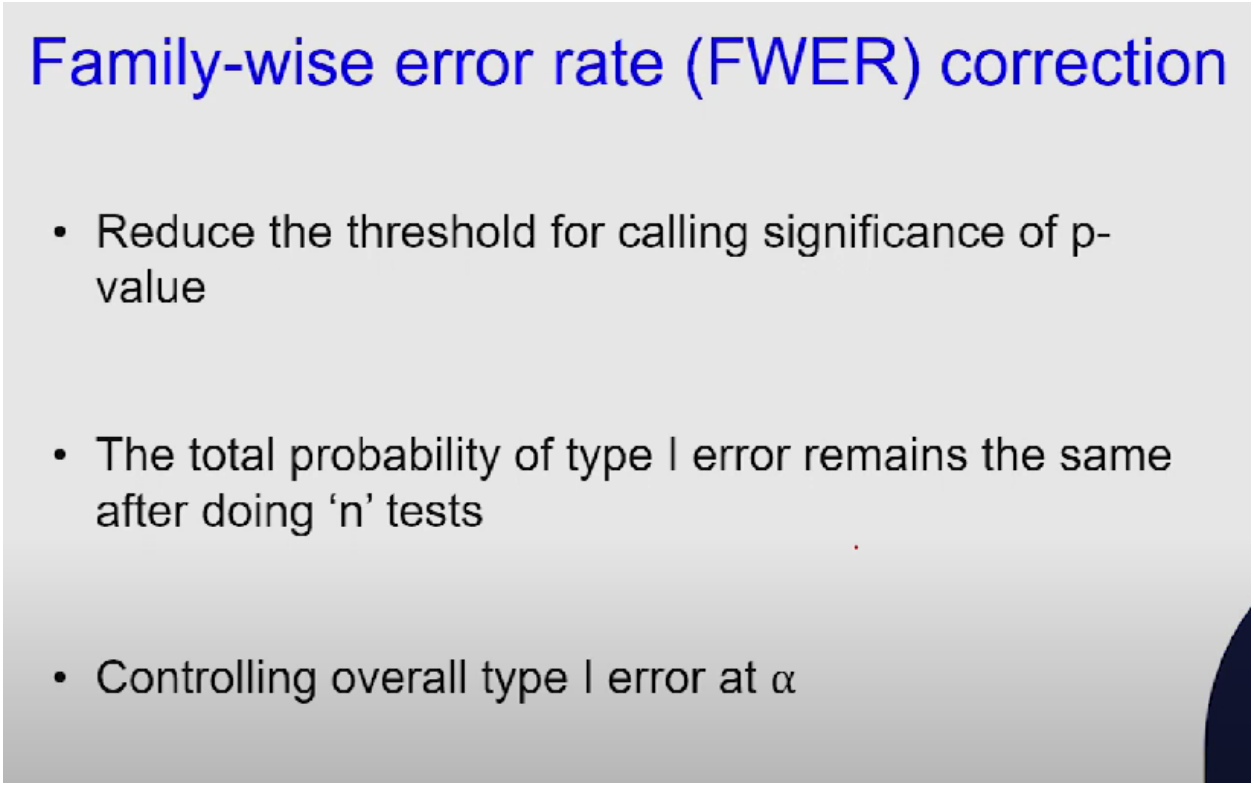
Probability of making one type I error = $1- (1- \alpha)^n$
(One false positive)

- As 'n' increases, the probability approaches the value of 1
- For n=100, this probability is 0.994

So, will the probability of getting 1 false positive be okay? So, earlier, it was 0.05, right? Now, as n increases, the probability will approach the value of 1. As you can see, alpha is a fraction. So, alpha is 0.05, and 1 minus alpha is 0.95. So, if you make it 0.95 to the power of n, then for n equaling 100, this probability becomes 0.994, right? So, what it means is that we are kind of sure now that if the probability is close to 1, that means we will definitely make 1 false positive call. So, that is the case here right? So, as we do multiple tests, the probability of making this false positive call also increases. So, now we understand, ok, the type 1 probability of type 1 error would increase right when we are doing this analysis. So, how do you control this type 1 error? We do not want to increase this type 1 error. We want to control this type 1 error, and there are two methods for this correction.



One is family-wise error rate correction, and the other is false discovery rate correction. So, you have FWER in short and FDR, ok? So, what we will do now is actually talk about this FWER in this class, and then in the next class we will talk about the FDR correction. These are different approaches for controlling type 1 errors. So, let us start with the family-wise error rate correction (FWER correction).

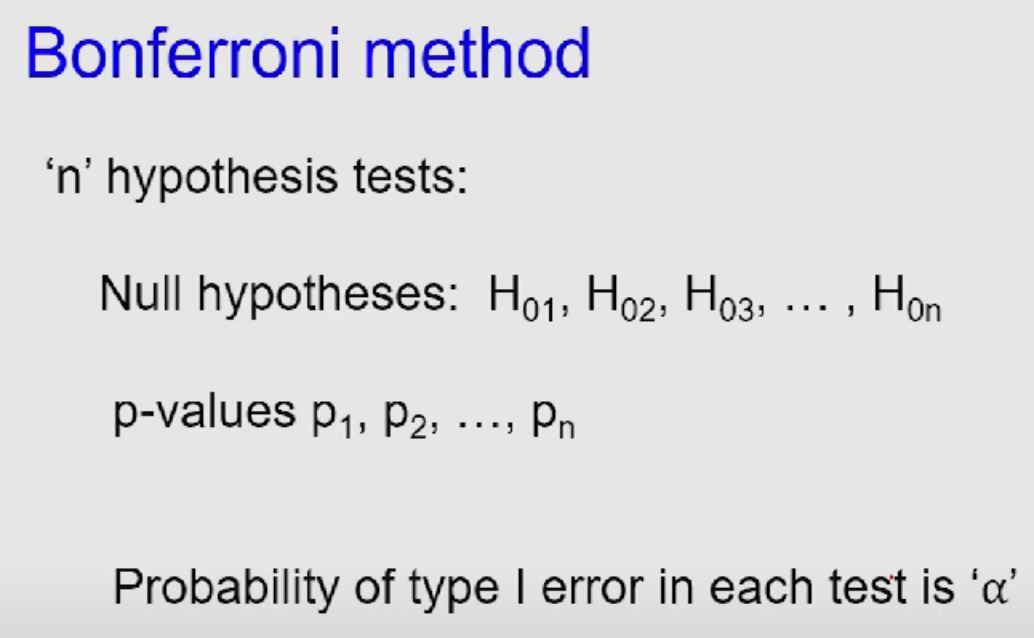


So, the idea is actually very simple here, right? Instead of setting the threshold to alpha for each statistical test, can we reduce the threshold for calling significance okay? Because we are doing 1000 statistical tests, if we keep the alpha at 0.05, we increase the number of false positives. But then, can we reduce this alpha threshold? Instead of keeping it at 0.05, can we make it lower right (0.01 0.001) or even lower right? And how do we actually do that? So, if you lower this threshold, that would mean we will again control the error rate. So, type 1 error rate is okay. So, what happens is that the total probability of type 1 error remains the same after doing the n test if you reduce this threshold. So, that is the idea that we want to keep this threshold low. So, we do not make this type 1 error okay. And what we are doing is controlling the overall type 1 error at alpha, which was 0.05 for a single test for 1000 tests. We also want to keep this overall type 1 error at 0.05. So, there are different methods for doing this. So, the first one is called Bonferroni correction, again named after the scientist who actually developed this method, and there is also Bonferroni correction. There are other methods as well.

- Bonferroni correction
- Holm-Bonferroni correction

So, I will not discuss all those methods, but I will talk about these two methods: Bonferroni and Holm-Bonferroni corrections. So, these are the corrections that will actually keep the probability of type 1 error at alpha by changing this threshold for significance. So, let us talk about the Bonferroni correction first.

So, imagine this situation: now we have n hypothesis tests, ok? So, then n could be 1000 or maybe 10000, depending on the number of genes that we are testing. And for each statistical test, we have the null hypothesis OK, which we number as H naught 1, H naught 2, H naught 3, and so on up to H naught n ok. So, as you have seen, for each gene, we do the statistical test, and we have a right null hypothesis for each of those tests. Now, also imagine that, corresponding to all these statistical tests, we get these p values: p 1, p 2, up to p n. So, once we have this right and also remember the right probability of type 1 error in each test, alpha is okay. So, the alpha is the probability of type 1 error in each statistical test. Now, the goal is to keep this alpha at the same level after doing these tests. We do not want this alpha to increase or the number of false positives to increase.



In the Bonferroni method, the idea is very simple. So, let us change the threshold for significance to lower right and how much lower we set this to alpha over n. So, n is the number of statistical tests that we are doing. So, let us divide alpha by n ok. So, it is a very simple idea, and then what you do is that we check all these p values that we generated, and if p i is less than alpha by n, then we reject the null hypothesis H naught i ok. So, this is a very stringent requirement, right? We have lowered this cutoff now instead of alpha. We have made this alpha by n ok, and this will again depend on the number of tests that we are doing. So, if you are doing 100 tests, this would be alpha by 100; if you are doing 1000, this is alpha by 1000, and if we can reject this null hypothesis H naught i, then we can call this gene i to be differentially expressed between diseased and healthy samples.

Bonferroni method

'n' hypothesis tests:

The threshold for significance is set to $\frac{\alpha}{n}$

If $p_i < \frac{\alpha}{n}$, then we reject the null hypothesis H_{0i}

- Gene 'i' shows difference in expression between diseased and healthy samples

For example, if you are doing this for disease versus healthy comparisons, So, again, this is just an example; you can have similar ideas. What are the drawbacks of the Bonferroni correction?

Bonferroni method

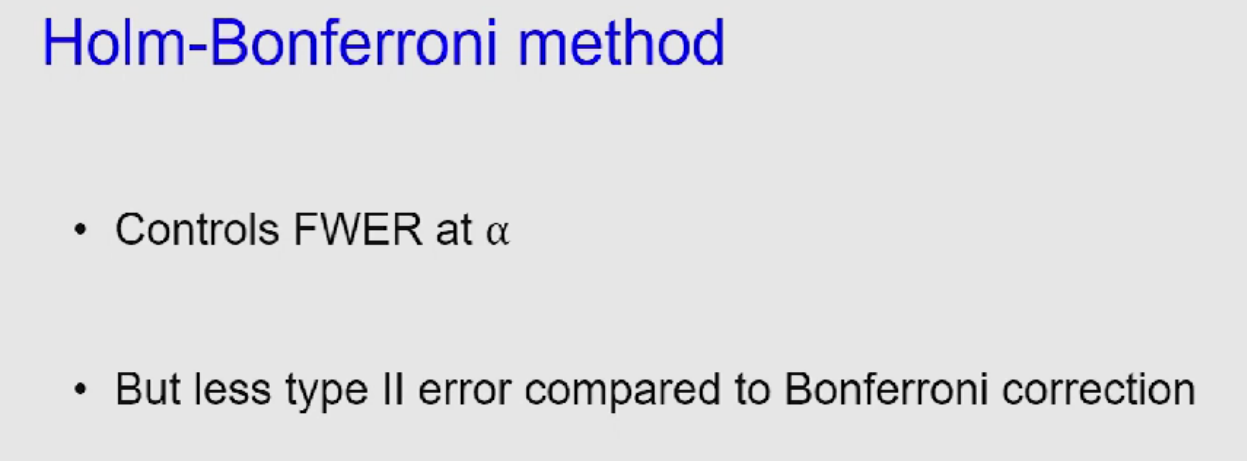
Drawbacks –

If n=10000, for example, the threshold is $0.05/10000 = 5 \times 10^{-6}$

If $p_i < 5 \times 10^{-6}$, only then we can reject null hypothesis

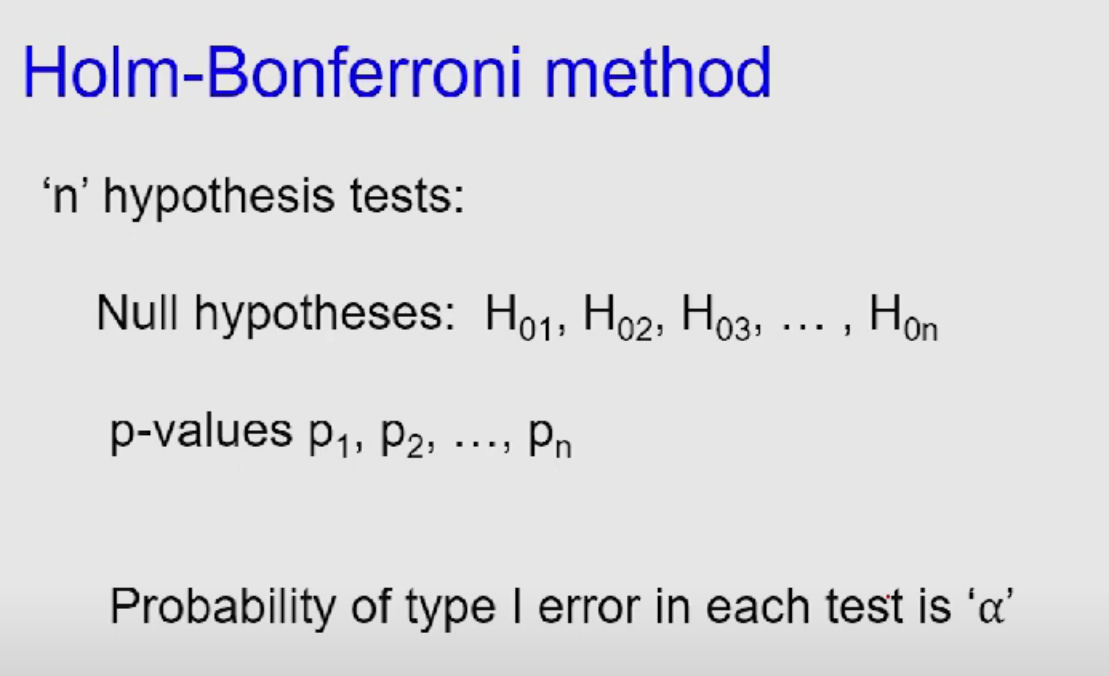
- Very conservative
- High type II error (False negatives)

So, one of the things you can probably imagine right So, if n equals 10000, for example, the threshold is 0.05 divided by 10000 and n is 10000. So, it becomes 5 into 10 to the power minus 6. If you are doing more statistical testing, this value will be reduced much further. So, you will get maybe 10 to the power minus 7 or 10 to the power minus 8; those will be the cutoffs, right? And you can now imagine that if p i is less than 5 into 10 to the power minus 6, then we can reject the null hypothesis. So, this is a very stringent cutoff, and very rarely will we get these p-i values so low. And this is especially true for our scenario right now, where we are doing this differential expression analysis with very few replicates. So, in the last class, we talked about the number of replicates in the experiments, and we said it was about 3 to 5 replicates for each sample. And if we have 3 replicates, it is very unlikely that you will get such p values, right from 5 to 10 to the power minus 6. That kind of p value you will very rarely get. Which means you will probably not detect differentially expressed genes with such a low number of replicates with this Bonferroni correction, ok? And this is the problem that researchers faced: this is a very conservative method. So, it is a very conservative means for calling this or rejecting the null hypothesis; it sets a very stringent condition. So, you need this p value to be very low, and only then will you get a significant call or a positive call, okay, where you reject the null hypothesis. On the other hand, what it means is that it commits a lot of high type 2 errors, ok? So, it will get you a lot of type 2 errors, which means you will get a lot of false negatives, ok? So, if you remember this table right, false negatives where H naught is false, but you do not reject H naught, ok? So, in our situation, what it means is that the gene is differentially expressed in the actual situation, which is the ground truth, but the statistical test here after the Bonferroni correction is saying this is not differentially expressed because the p value is very low, which is not lower than this cutoff. So, which means we are calling a lot of false negatives and we are not able to identify a lot of differentially expressed genes that are actually differentially expressed in the real situation. So, that means this sample has this method with low power, and we commit a lot of type 2 errors. So, what is the solution, okay? So, there is a solution that has been proposed that is called the Holm-Bonferroni method. So, this is a correction to the earlier method. So, the earlier method is very simple.

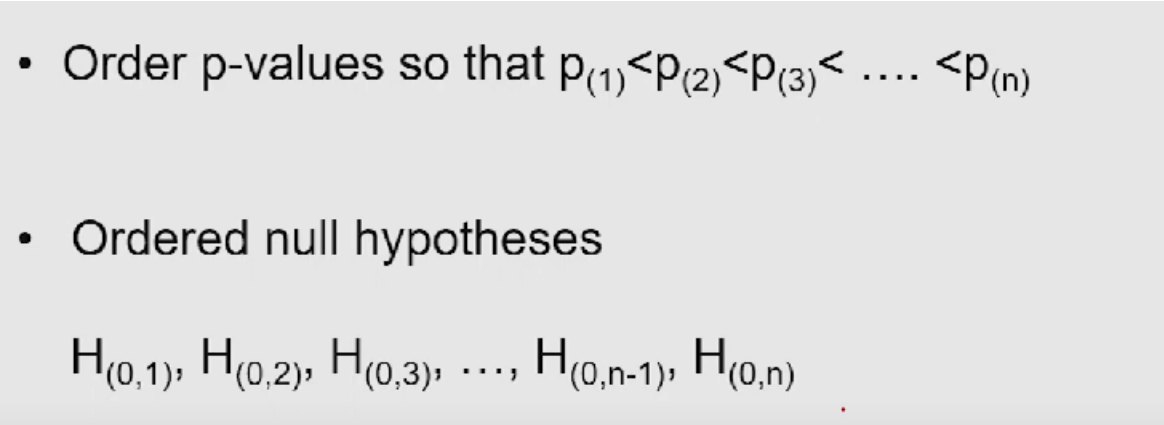


So, you can apply it very easily, but then you see you have a lot of problems, ok? So, this method also controls these family-wise error rates at alpha. So, you want this probability of type 1 error at alpha. You do not want to make false positive calls, right? And this method actually adopts a certain method that actually gives you less type 2 error compared to Bonferroni correction. So, we will see in a moment how we can actually generate or get more power and commit fewer type 2 errors compared to Bonferroni correction.

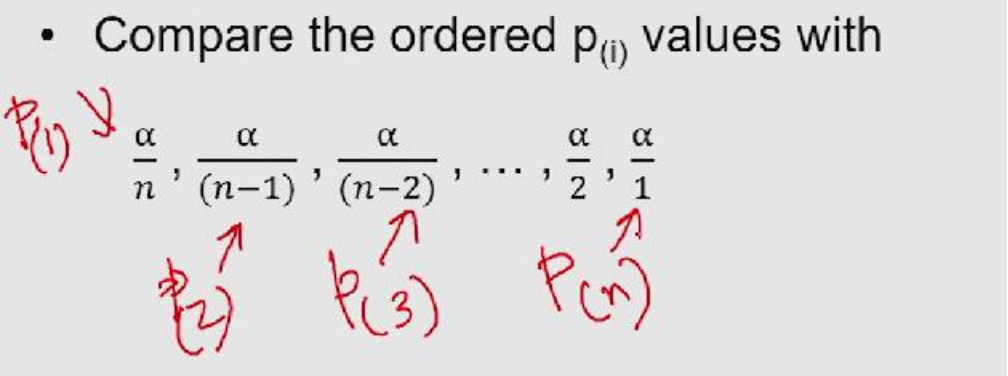
So, what does it do? So, again, coming back to the same situation where we have this N hypothesis test, We have the null hypothesis H naught 1, H naught 2, H naught 3, and so on. And you have p values corresponding to p values: for H naught 1, you get a p value of p 1 after the statistical test; for H naught 2, you get a p value of p 2 after the statistical test; and so on. And again, the probability of type 1 error in each test is set at alpha right.



So, that is what we are doing right. And we want to say, keep this alpha for all these N tests, right? So, after comparing these N genes for differential expression, we still want to keep the probability of type 1 error at alpha. So, we do not want to have this increase, right? We do not want type 1 error to increase, ok? So, in the Holm-Bonferroni method, the first step is to order the p values. So, as you can see now, I am writing this p 1 p 2 within brackets, which means these are sorted p values. So, p 1 is less than p 2, less than p 3, etcetera, right? So, p n is the largest p value that we have from the statistical test, right? So, you call you are doing this 10,000 tests, you get p values, and then you sort the p values, ok? And you have the lowest p value first and the highest p value at the end, ok? So, you can create this sorted list. And similarly, based on the sorting, you also have the order null hypothesis right, which we will call H naught 1 right within brackets, and then again you have H naught comma 2, H naught comma 3, and so on, ok.



So, what this method proposes is that instead of doing this alpha by N right, where N is the number of statistical tests, compare these ordered p values with alpha by N, then alpha by N minus 1, and so on, ok. So, what you do is then we start with, let us say, i equals 1 ok. So, for p 1, this is the smallest p value, which we compare with this alpha by N. So, alpha by N is the most stringent cutoff, and p 1 is the lowest p value. So, for the lowest p value, you apply the most stringent cutoff, ok? And then for p 2, this is the sorted value, which you compare with alpha divided by N minus 1. And then you continue this process for p 3. You compare against alpha by N minus 2, and finally, for p N, you compare with alpha by 1 or with alpha itself. So, in that case, alpha is 0.05, ok. So, what you are seeing here is that you are gradually shifting the cutoff, because in each test, the cutoff is different based on the original p value.



So, again, you sort these p values and set these cutoffs, and these cutoffs are different for each test. Now, you apply this condition. Right after you do these comparisons, you find the lowest k for which this p k is right. This is the ordered p value that is greater than or equal to alpha divided by N minus k plus 1. So, let us say if p you find this as k equals to 10 ok. So, alpha divided by N minus 9 is okay. So, you are comparing this, and you see p 10 is greater than alpha divided by N minus 9. So, what do you do then, ok? Once you find that, what you have to do is reject the null hypothesis H naught 1 to H naught k minus 1 ok. And then you keep the hypothesis or do not reject the hypothesis H naught k to H naught N ok. So, you kind of understand right because these p values are sorted by the first p value that actually crosses that threshold or is equal to that threshold. Then the rest of them, after this p value, will automatically be greater than their cutoffs, ok? So, that is why we actually keep this lowest k and find this lowest k for which this condition is satisfied, and then we reject the null hypothesis H naught 1 to H naught k minus 1, and we do not reject the null hypothesis H naught k to H naught N ok.

- Compare the ordered $p_{(i)}$ values with

$$\frac{\alpha}{n}, \frac{\alpha}{(n-1)}, \frac{\alpha}{(n-2)}, \cdots, \frac{\alpha}{2}, \frac{\alpha}{1}$$

- Find the lowest k for which $p_{(k)} \geq \frac{\alpha}{(n-k+1)}$

null hypotheses $H_{(0,1)}$ to $H_{(0,k-1)}$ are rejected

null hypotheses $H_{(0,k)}$ to $H_{(0,n)}$ are not rejected

So, again, this method is less conservative; as you can see, we are changing the significance cutoff. So, threshold for calling significance and we are kind of changing that from more stringent to less stringent right based on the p value ordering. So, what are the drawbacks of this family-wise error rate?

Drawbacks of FWER methods

- Stringent control for type I error
- Worried about even one false positive
- Higher type II error – reduced power

So, of course, this is very stringent control for type 1 error, right? We are even worried about committing a single type 1 error because we want to keep the overall alpha at 0.05. Even after doing these 10,000 tests, we want to keep this N at 0.05. So, we are worried about even one false positive out of this 10,000 test, and as a consequence the type 2 error is higher in the case of these methods, and we have reduced power. So, a higher type 2 error means we have more false negatives right where H naught is actually false, but we are not rejecting H naught right. So, this actually leads to a lot of false negatives, ok? So, these are the references for this lecture. So, what we have seen is that multiple hypothesis testing increases the number of false positive calls in the two approaches we have discussed. So, we have also talked about how we can do this correction right, and so, we have mentioned that this hypothesis testing correction is required when we are doing this large number of hypothesis tests. And I have mentioned that this is not just applicable for gene differential gene expression analysis. As you will see in other scenarios in the field, you will have to do this multiple hypothesis correction, and beyond the field, wherever you are testing a large number of hypotheses, you will have to do this correction. We have talked about the first type of method here, and we have mentioned two. There are two types of correction methods that exist: family-wise error rate correction and false discovery rate correction. And in this class, we have talked about family-wise error rate correction, and we have talked about two methods that actually help us with this family-wise error rate correction. The first is the Bonferroni correction, and the second is the Hombonferroni method, and we have talked about the drawbacks. So, one of the drawbacks is that we actually have very stringent control over type 1 errors, but as a consequence, we make a lot of type 2 errors. Now, the goal is to actually try to balance these two and not miss a lot of true positive genes. Thank you.