Next Generation Sequencing Technologies: Data Analysis and Applications

FDR correction and interpretation of DGE analysis results

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 class, we started talking about multiple hypothesis testing and correction. We talked about why this type of correction is required. We have seen that when you are testing a large number of hypotheses, the number of false positives can increase, and we have started a discussion on correction methods. So, how can we correct or control for these false positives or the type 1 error?

So, in the last class, we mentioned the family-wise error rate correction method, and in this class, we will be talking about the FDR correction method. So, once we have discussed this FDR correction method, we will move on to something called the interpretation of the results that we obtain from DGE analysis. So, this is the agenda for today's class. So, first we will discuss the FDR correction method, and then we will talk about the interpretation of results from DGE analysis.

So, we generated this table from DGE analysis and how we actually proceeded with that result, ok? So, let us start with the FDR correction today. So, the first keyword that we will see is the adjusted p value, ok, and you have come across this term before, and the second keyword we will see is gene ontology, ok. So, briefly, we have two types of methods for controlling type 1 error when we are doing multiple hypothesis testing. So, the first type is the family-wise error rate correction, which we discussed in the last class, and then the second type of method is a false discovery rate correction, or FDR, which we will be discussing in this class.

So, let us start with this FDR correction method, and what you have seen in FWER correction methods is that we are worried about making even a single type 1 error, right? So, that is why you want to control alpha at the same level even after 10,000 tests, ok? But in many real-life scenarios, what you will see is that we are not really worried about making even a single type 1 error, and

you can sometimes accept a certain number of false positives. It is not that we want a lot of false positives, but we can tolerate a certain number of false positives in real-life situations, right? So, in our example, let us say that when you are doing this differential expression analysis, we identify 100 genes as differentially expressed, and imagine if, let us say, 3, 4 genes are false positives.

So, if you remember this table of truth that we talked about, we have the actual truth and the statistical decision, right? So, if only 3 or 4 genes are false positives, then that is not going to bias our analysis that much, ok? So, in that case, we can tolerate a small number of false positives; they will not substantially change our results, ok? So, that is what we state here, right? So, we can probably control the type 1 error at a certain value, rather than trying to control it at a very stringent cutoff.

So, we can perhaps tolerate 5 to 10 percent of false positives again, depending on the test or the inference that you are making. We can tolerate this 5 to 10 percent of false positives in our data, ok? So, this is actually the approach that these false discovery rate correction methods take. And there are different methods out there. The first one and the most popular one is the Benjamini-Hochberg method. You also have other methods, for example, the Benjamini-Yekutieli method and then you have the Storey-Tibshirani method. This is a Q-value method called the Q-value method.

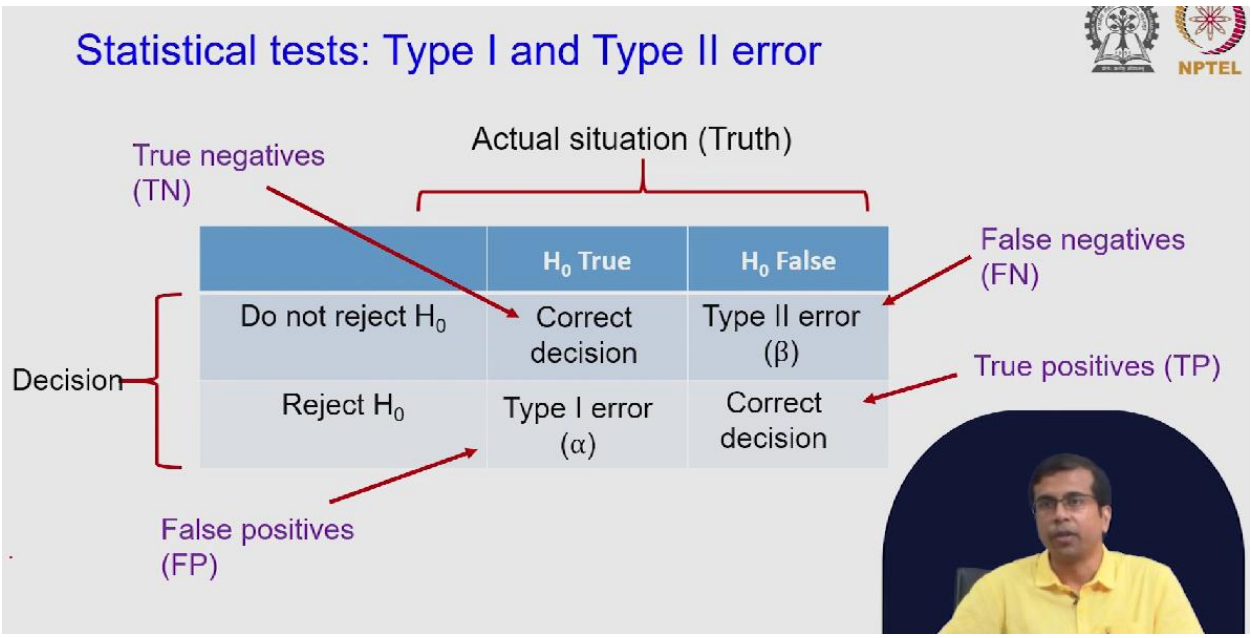
Methods for FDR correction

- Benjamini-Hochberg method
- Benjamini-Yekutieli method
- Storey-Tibshirani method

…

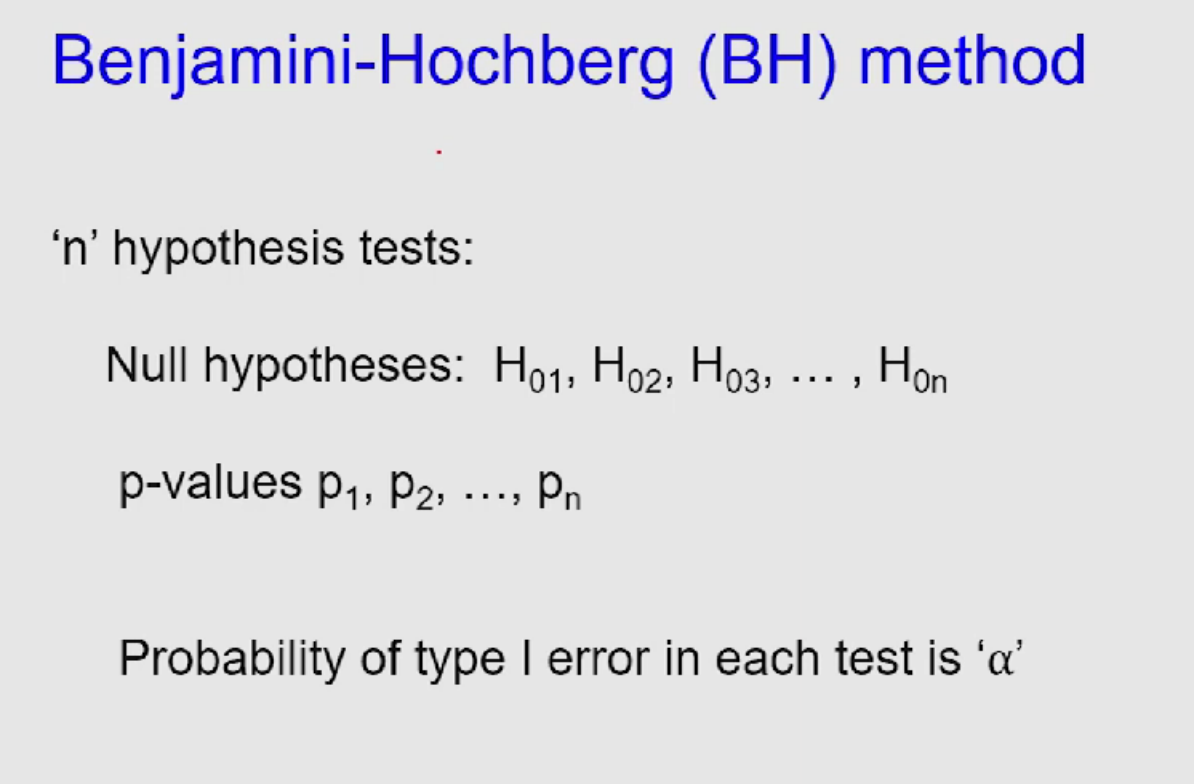
So, we will talk about the first method, the Benjamini-Hochberg method; this is the most popular one, and for the rest of them, I will just give you the references, and of course, you can have a look.

Statistical tests: Type I and Type II error

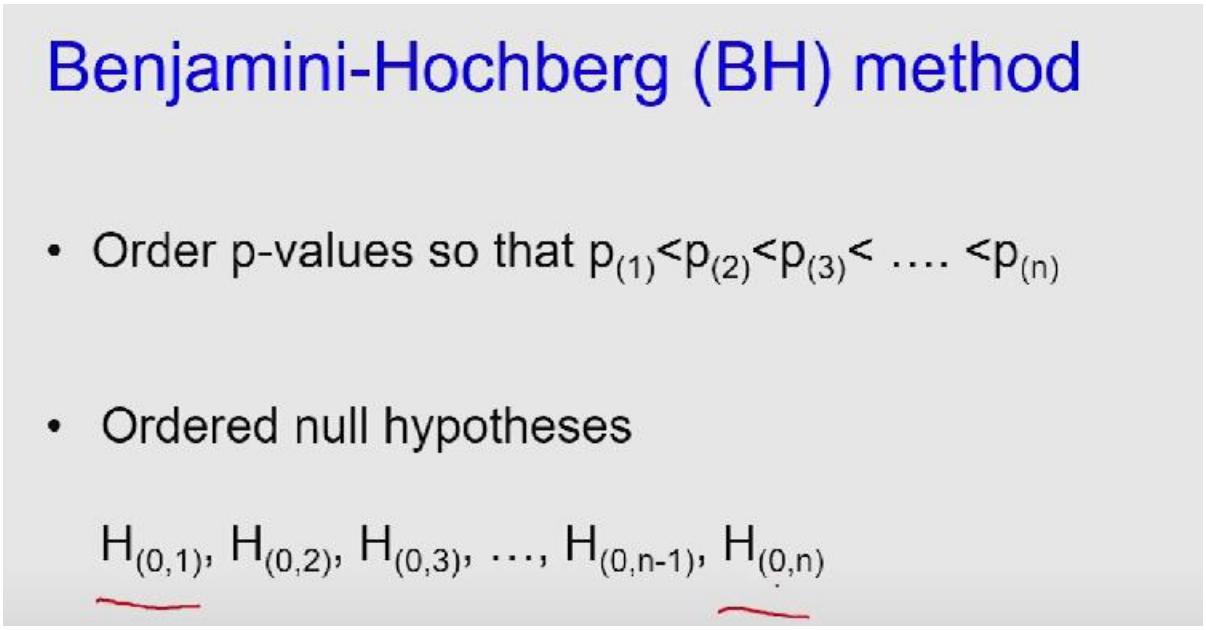


$$FDR = \frac{FP}{TP+FP}$$

So, coming back to the statistical test where we talked about this type 1 and type 2 error just to remind you, right, we have these situations; we have this actual situation or the truth, and then you have the statistical decision on the other side, and then you have also defined the false positives, which are the type 1 error, then we have the false negatives and of course, we have the true positives and the true negatives, right? So, the false discovery rate actually relates to these false positives and true positives, right and is defined as the false positive number divided by the true positive plus the false positive. So, let us now go into the Benjamini-Hochberg method, or sometimes you will see in many tools that it is written as the BH method, ok, and again, we will take the same example that we took earlier, ok. So, we have N hypothesis tests; right, we have this null hypothesis H naught 1, H naught 2, and so on up to H naught N, and we have corresponding p-values p 1, p 2, and so on up to p N, ok, and the probability of type 1 error that was that in the original situation, right in each test is alpha.



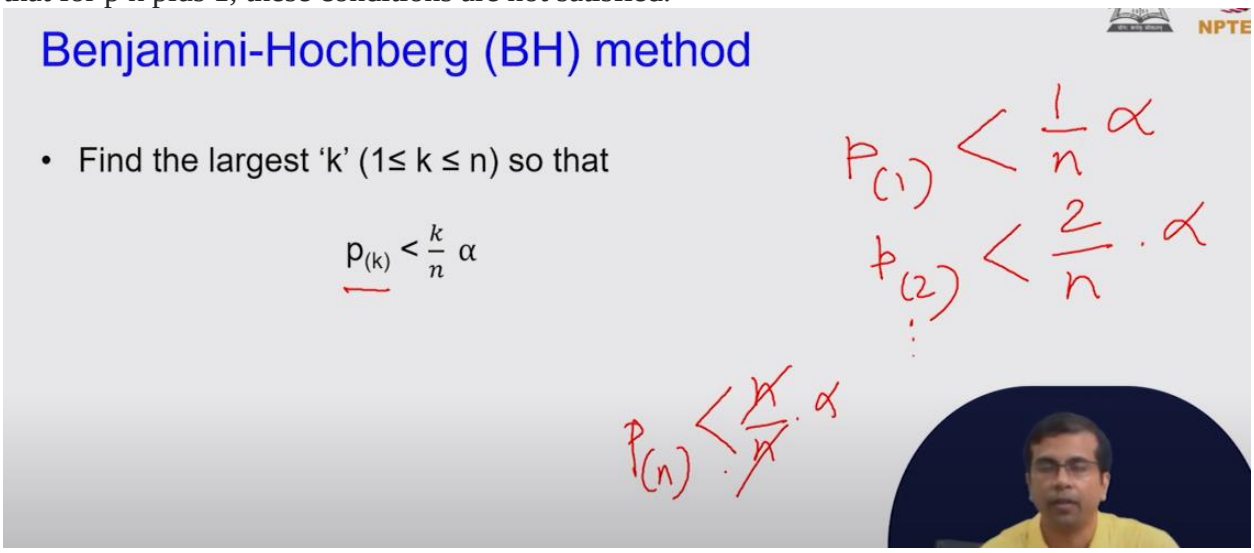
Now, what we do in this method is that we order the p-values first. Right again, we have seen this kind of situation in the Holm-Bonferroni parent method, where we actually order these p-values according to our magnitude. So, the smallest p-value will come first. This is p 1, right within the bracket. Right again, this is denoting that these are sorted p-values. Right, this is not the original p-values. So, the smallest p-value comes first and the largest p-value comes last, right at the end, and we also have the order null hypothesis according to the sorting, which we call H 0 comma 1. So, this is the first null hypothesis according to the p-value sorting, and similarly, we have up to H naught N, ok? These are again the null hypotheses that are ordered against the p-values.



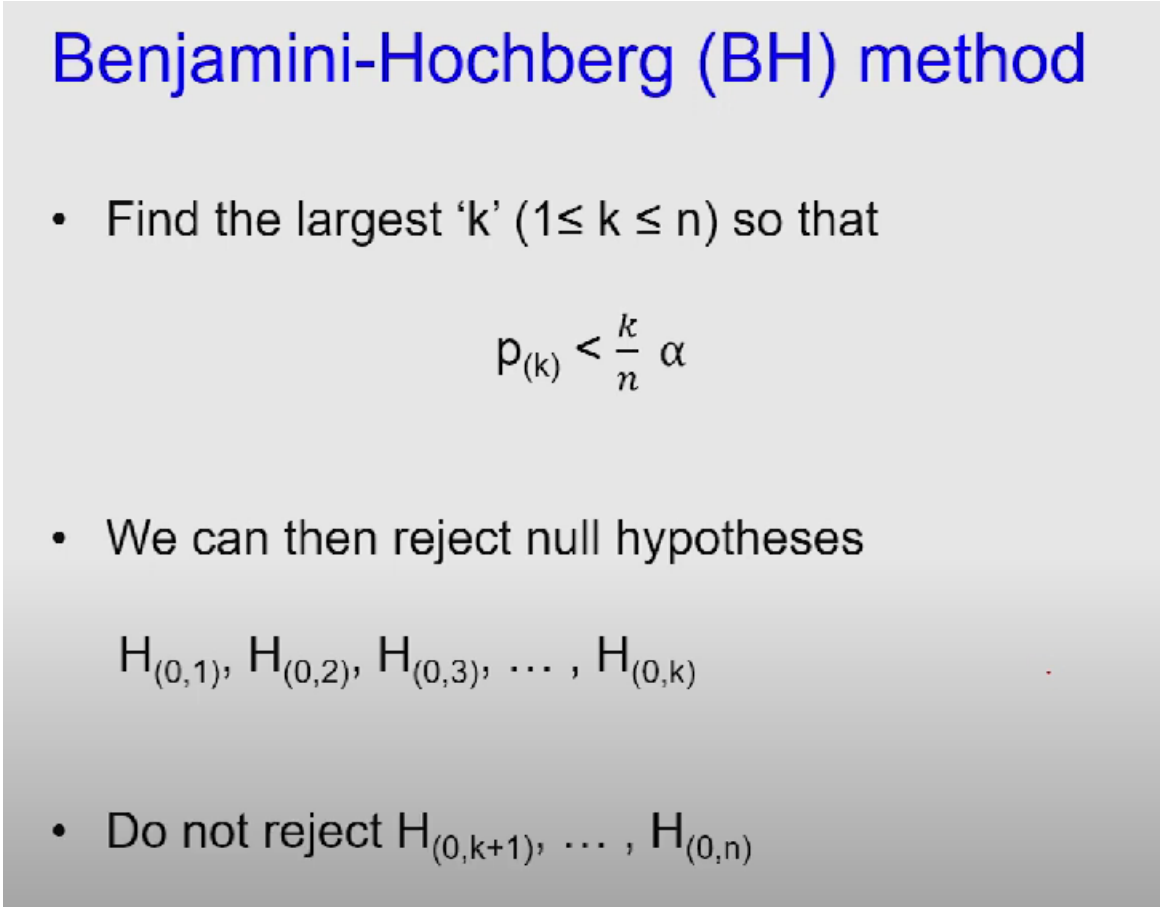
So, what we do in the case of the Benjamini-Hochberg method is actually try to find the largest K. K is between 1 and n. So n is the number of statistical tests that we are doing. So that the pk, that is, the ordered p value, is smaller than k divided by n time alpha. Alpha, of course, will be 0.

05-1. So what we are doing is comparing these order p values now. So, if you remember, we have this p 1 p 2 right. So, since this is p. 1,. So, this is the smallest p value.

So, k is 1. So, what we do is compare this with 1 divided by n times alpha. For p 2, we are doing this by comparing it with the right 2 divided by n times alpha, and so on. We keep on doing this, and for p n right, this is order 1; we have n divided by n times alpha, which means we are simply comparing this with alpha. So, when we are doing this comparison, we want to find the largest k for which we have this condition satisfied, ok? So, p k is less than k by n times alpha, which means that for p k plus 1, these conditions are not satisfied.



And so, what you can do now is then reject this null hypothesis H naught 1 up to H naught k, which satisfy this condition. Where they are less than the cutoff that is set for them, ok. Again, the cutoff is adaptive, and as you can see, we are changing the cutoff based on the p values. And we do not reject this null hypothesis: H naught k plus 1 to H naught n ok. So, what this method will do right is control the error.



So, we will control this number of false positives at a certain value, ok? It is usually 5 percent or 10 percent again, depending on the alpha that we choose. And if you have seen this right, once we apply this method, we calculate something called adjusted p values. And sometimes they are also called q values, but this could be misleading again depending on the method that you are using. You should always look at which method has been applied for this FDR correction. So, this is something that is important.

So, we either call them adjusted p values or sometimes also q values. So, these adjusted p values are calculated from the original p values of the statistical test by taking this FDR rate and the false discovery rate into consideration. And I will give you the formula for how we actually calculate these adjusted p values. So, again, going back, we have this order of p values from the n test, and again, p values from p 1 to p n. And what I do is we start from p n, right? This is the last p value. This is the largest p value.

Adjusted p-values

- Adjusting p-values for taking false discovery rate into consideration
- Ordered p-values from 'n' tests : $p_{(1)} < p_{(2)} < p_{(3)} < \ldots < p_{(n)}$

And we calculate this adjusted p, right? This is the adjusted p value. So, adjusted p i, this is the minimum of p i into n divided by i, or compare with p i plus 1 ok. So, if i is the rank of that p value, we multiply that with n, which is the number of statistical tests divided by i. And then we calculate, compare with p i plus 1 right, and take the minimum of these two. And so, n is the total number of tests, and i is the rank of the p value.

Adjusted p-values

- Start from $p_{(n)}$
- $Adj.p_{(i)} = \min \{ p_{(i)} \times \frac{n}{i} , Adj.p_{(i+1)} \}$

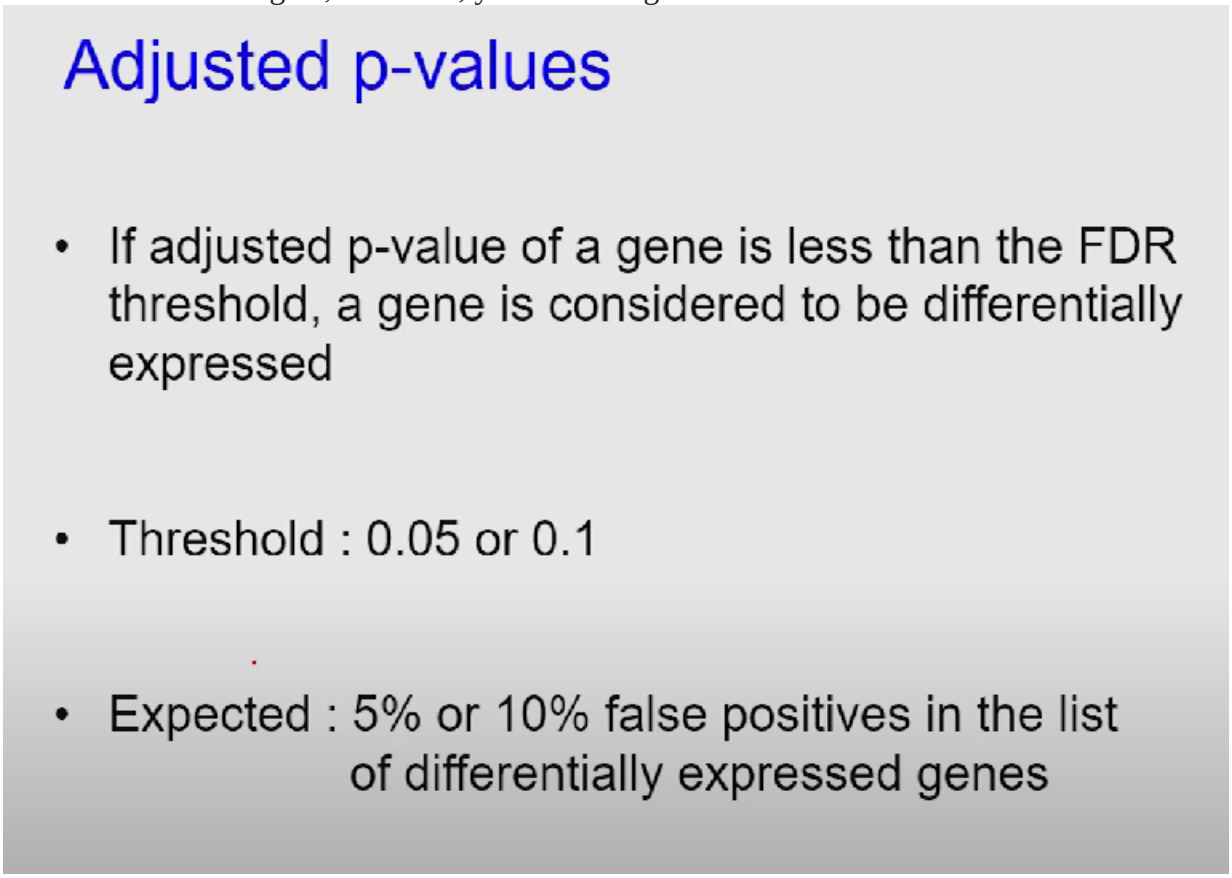
  n = total no. of tests
  i = rank of the p-value

- If $Adj.p_{(i)} > 1$, then set $Adj.p_{(i)}=1$

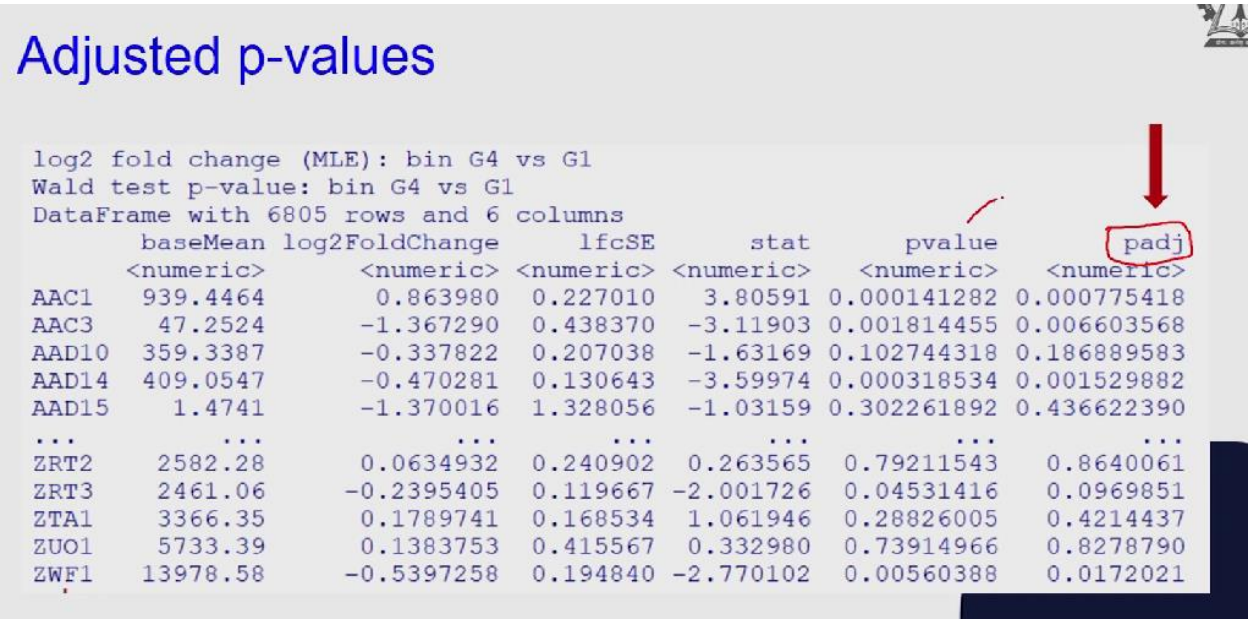
And in any case, if p i is greater than 1, we then set p i equal to 1, right? So, this might apply to the largest ones, right? So, the adjusted p might go above 1, right? So, these are actually adjusted p values, okay? So, this is adjusted p i when you are calculating this value, ok?

So, this is how we calculate this adjusted p value, and what do we now do after we have adjusted p values? If an adjusted p value is less than the FDR threshold, then we consider a gene to be differentially expressed. Now, what is this threshold? It is usually set at 0.05 or 0.1, and it corresponds to a 5 percent or 10 percent false positive rate. So, if we set this threshold to 0,.

1 right We can expect about 10 percent of false positives in the list of differentially expressed genes. So, finally, we identify, let us say, 100 genes that are differentially expressed; about 10 percent of them are false positives, ok, because we have set the threshold at 0.1. Now, if you want to make it more stringent, of course, you can change it. You can make it 0.



05, 0.02, and so on. So, going back to this table, this is the results table that we looked at, and we have this column p adjusted, which is actually adjusted according to this Benjaini-Hochberg method. And these p-adjusted values are calculated from these p-values. What you will now also notice is that this p adjusted is bigger than the p value right, and that is because of this adjustment. If you remember we multiply by the number of tests and divide by the rank of that p value. So, you will notice that this p-adjusted value is always larger than the p-value that you get from the statistical test.



One thing you should not confuse is that this list is actually ordered according to the gene names; it is not ordered or ranked according to the p value. So, once you have generated this list, and this is done inside DESeq2, once you have generated this list, you have to sort the p values, rank the p values right, calculate all these adjusted p values, and put them back here, ok? So, we have now talked about this FTR correction method, and you see, this is less conservative than the FWER methods, and they have more power in the detection of true positives. And this is why this is the method of choice now in the case of differential expression analysis. Now, one question you might have is: which method should we choose—FWER or FDR? Right again, it depends on the question that you are answering.

So, in situations where you can tolerate certain false positives or are doing a huge number of tests, you probably want to apply FDR. And in situations where you have a smaller number of tests, maybe you are doing 10, 20, and you do not want to commit or identify any false positive cases. In that case, you will use this FWER method. So, now we want to move into the results part, right? So, we have generated these results, and we have understood all the components of the results. So, we have talked about log tuple change; we have the p values; we have the p adjusted values; and we now understand how these p values are generated. p-adjusted values are generated, and we have generated a list of differentially expressed genes by applying an FDR cutoff of 10 percent.

Now, what do you want to do after you have got the list right? So, let us say you have 100 genes on the list, okay? And this is what we are going to talk about now, right? How do you actually interpret these results that you have gotten after this differential gene expression analysis? So, how do you interpret this data? So, we have this list of differentially expressed genes, and the question that we are mostly interested in is: what do these genes do in the cell or tissue?

So, we want to know the functions of these genes. Why do you want to know these functions? Because they will give us insights into biological processes. So, you can imagine this disease versus healthy situation, right? We are comparing this disease versus healthy samples, and we have done the differential gene expression analysis. We have identified a list of genes, and let us say we

have identified 100 overexpressed genes in the disease sample compared to the healthy sample. And the question we probably want to understand is: What do these genes do right? 100 genes: what are they doing? Are they in any way associated with the disease? So, whether they are causing the disease right, whether they are helping in disease progression, or whether they are actually doing certain functions right, that kind of makes the situation worse. So, the question that you want to answer often is: do these genes belong to specific functional classes?

And this is what we do with something called functional enrichment analysis; sometimes these are also called gene set enrichment analysis, although there is a tool that is actually called gene set enrichment analysis, or sometimes we call this pathway enrichment analysis. These are kind of equivalent, but not exactly the same. So, let us now go into the functional enrichment analysis, and these are the questions that we ask here ok? Are there specific functional classes that are overrepresented or underrepresented in the list of differentially expressed genes? So, we are probably asking this question because we want to understand the biology of what is actually happening inside the cell in the case of disease samples or in the case of treatment versus control analysis.

So, the genes that are overexpressed or the functional classes that are overexpressed are overrepresented, right? They are probably very important for the disease progression, and so on. And the question is: are we analyzing disease versus healthy samples? So, this is probably one of the questions that will come to our minds, right? Do these pathways play any role in disease progression or response to a treatment, a drug, or a therapy? So, maybe because of these functional classes, they are kind of generating some sort of resistance to therapy, or if we are comparing, let us say, condition 1 versus condition 2, we are giving some stress to the cells. Are these pathways associated with stress responses helping in the adaptation of the cells or the organism to that stress condition? So, this is again giving us molecular insights, right?

So, we understand how the disease is probably progressing, or we understand how the cells are responding to certain stress and how they are adapting to that kind of stress. So, we are getting through this analysis; we are now getting into the biology, right? So, we are getting more biological insight, and we are trying to understand the molecular basis of disease, stress response, and so on. So, depending on the question that you are asking, you are going to get molecular insights. So, the first step in this process when you are going to do this functional enrichment analysis is to identify the gene sets.

So, how do you define these gene sets, and what do we mean by these gene sets? So, before we can do this functional enrichment analysis, we need to identify or associate each gene with a function in the cell. So, you have a list of 100 genes that are differentially expressed, and for each of them, you need to have some association with certain functions inside the cell. So, this is the first requirement we need to have this association. This information is required, ok? And this one gene set consists of genes that are associated with a specific cellular function.

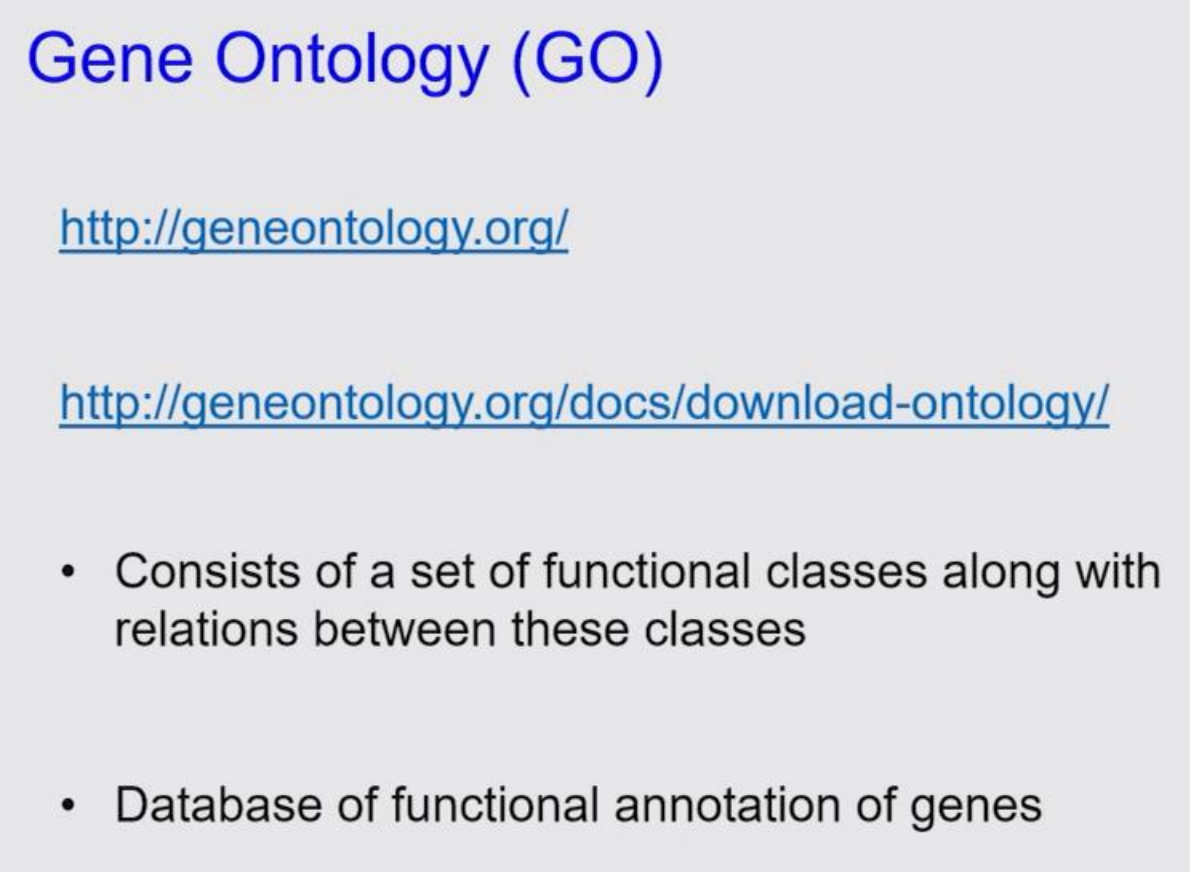
So, let us say we have glycolysis genes that are involved in glycolysis. You have ATP synthesis genes, right? You have TCA cycle genes, and so on. So, you have these gene sets that are associated with specific cellular functions, okay? And the other point that you should also keep in mind is that a single gene can carry out multiple functions. So, genes are pleiotropic; they often

have multiple functions, and thus they can be part of multiple gene sets. So, it is not necessary that all gene sets are unique; there can be overlap between gene sets.

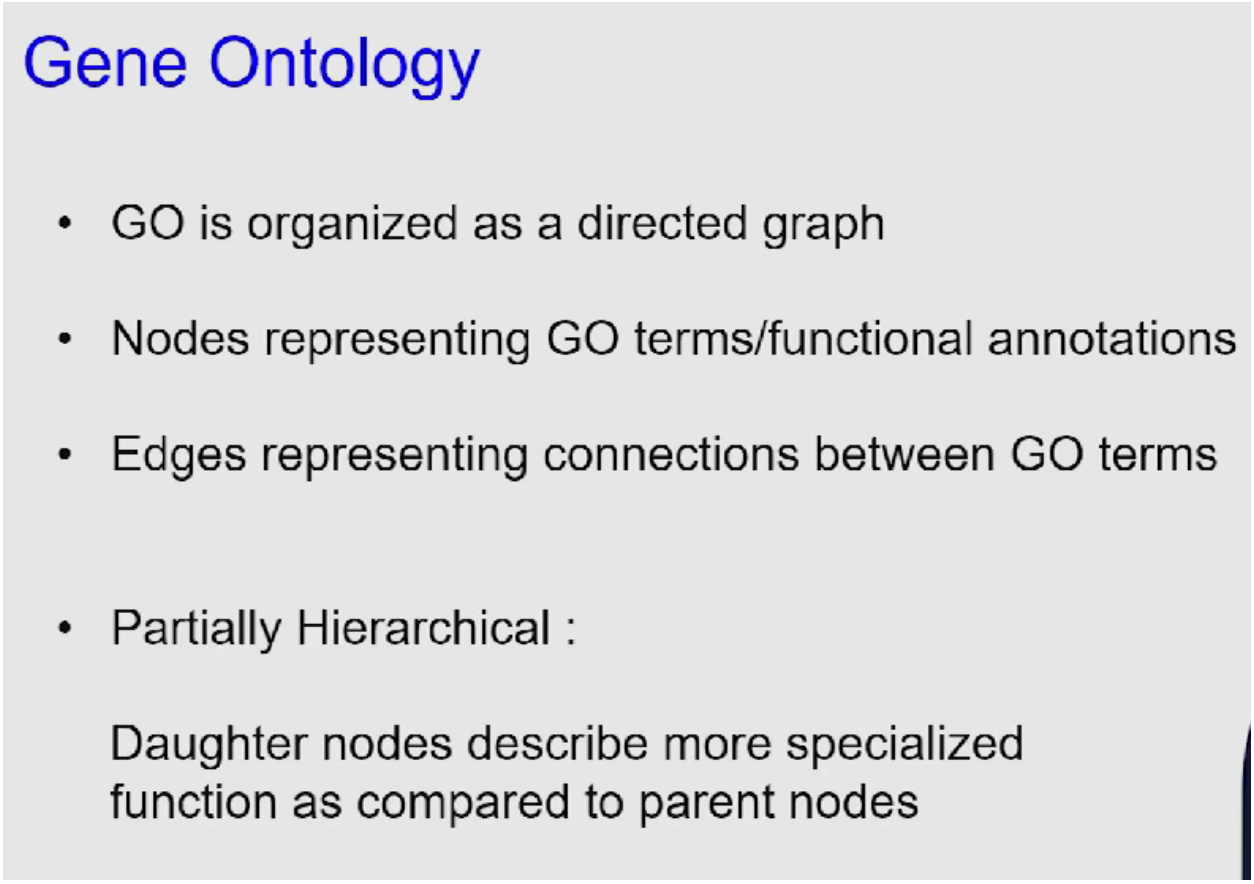
So, a gene that is involved in, let us say, the salt stress response to high salt concentration or high osmotic stress can also respond to oxidative stress. So, many genes can have this kind of multiple function, and they can be present in different gene sets. So, one of the major tools for this gene set enrichment analysis is the gene ontology. This is a resource where this kind of association information is present. Here is the link where you can actually find this information, and these are the references, ok?

So, I will briefly mention what this gene ontology is, and of course, I will encourage you to explore more by going through the resources that are present on the website. And you can also download ontologies for any organism that you are working with, and there are many organism specific ontologies that are available on the website. So, very briefly, gene ontology consists of a set of functional classes along with relations between these classes. So, we have many functional classes, and they are kind of related to each other, right? So, one is probably, let us say, glycolysis, which is something associated with carbon metabolism, right?

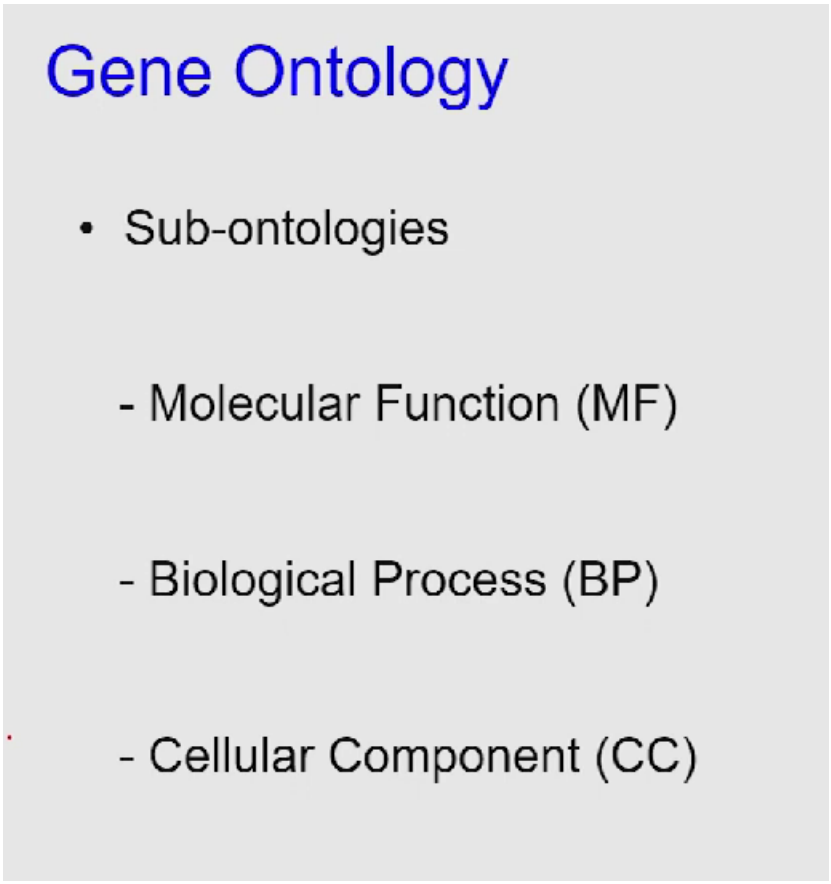
So, these are functional classes that are related to each other. So, gene ontology will also preserve this information, ok? So, it has this association of single genes to certain molecular functions, but it will also have this information about relationships between functional classes. And this is a database of functional annotations for genes. So, you can get this information: which gene is involved in which pathway or which cellular process?



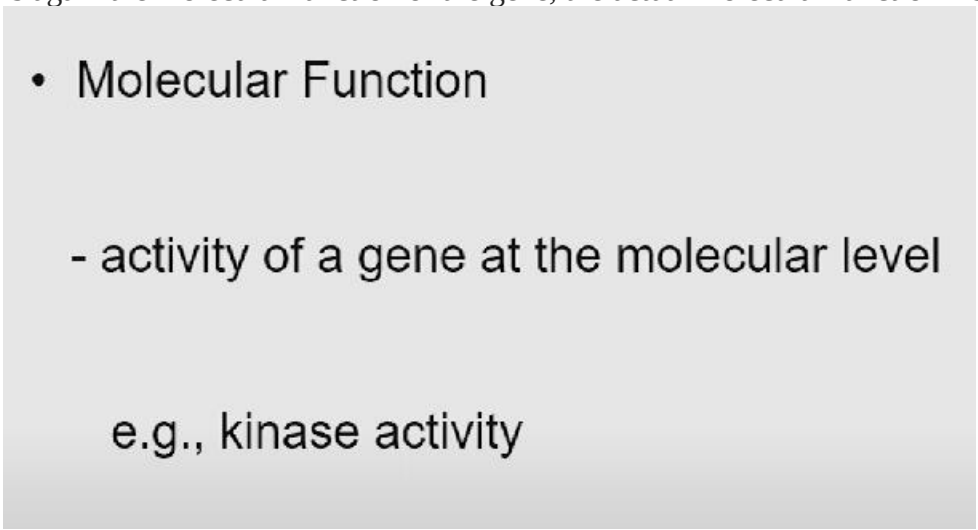
So, gene ontology is organized as a directed graph, as you can see once you go into the website. And each node represents a geo-term or a functional annotation, and then the edge represents the connection between geo-terms. I just mentioned one example, right? So, you can have these connections between differential terms, ok? And it is also partially hierarchical. If you go and see the networks like this ontology and the directed graphs, you will see that there is some sort of hierarchy a little bit right where daughter nodes describe more specialized functions as compared to parent nodes.



So, again, if you take an example, let us say we have tryptophan biosynthesis. So, that is, that will be a daughter node, and maybe a parent node could be amino acid biosynthesis, right? So, you have these relationships or terms that are related to each other, okay? So, in gene ontology for each in also you will find sub ontologies ok. So, what are the sub-ontologies? So, there are three sub-ontologies that are present.

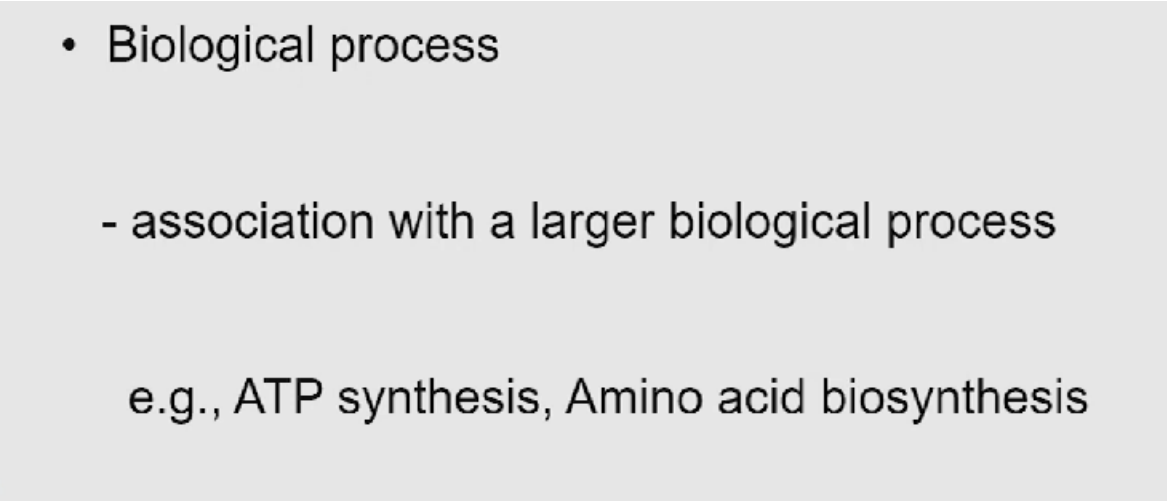


So, the first is molecular function, or, in short, MF, then you have biological process, or BP, and cellular component, or CC. So, you see these three terms in any ontology that you look at, okay? And what do they mean? So, the molecular function means it actually specifies the activity of a gene at the molecular level, which is what this gene actually does inside the cell. So, for example, this gene has kinase activity, or maybe this gene is involved in isoleucine biosynthesis ok. So, this is again the molecular function of the gene, the actual molecular function inside the cell.

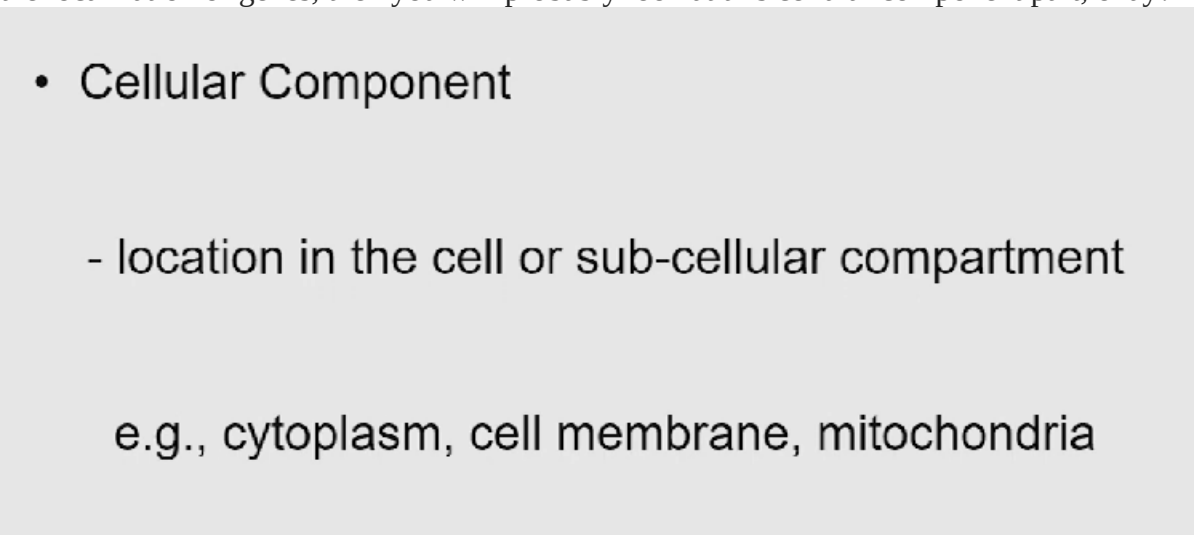


So, this would be specified here. The second part is the biological process, okay? So, here, it

actually will associate the gene with a larger biological process, ok? So, for example, you can say the gene is involved in ATP synthesis. It does not specify the actual function, right? The actual molecular function, which part of the ATP synthesis process is it involved in? That is not specified, but it is associated with ATP synthesis or maybe amino acid biosynthesis and so on. So, you see, these are actually associated genes with a larger biological process.

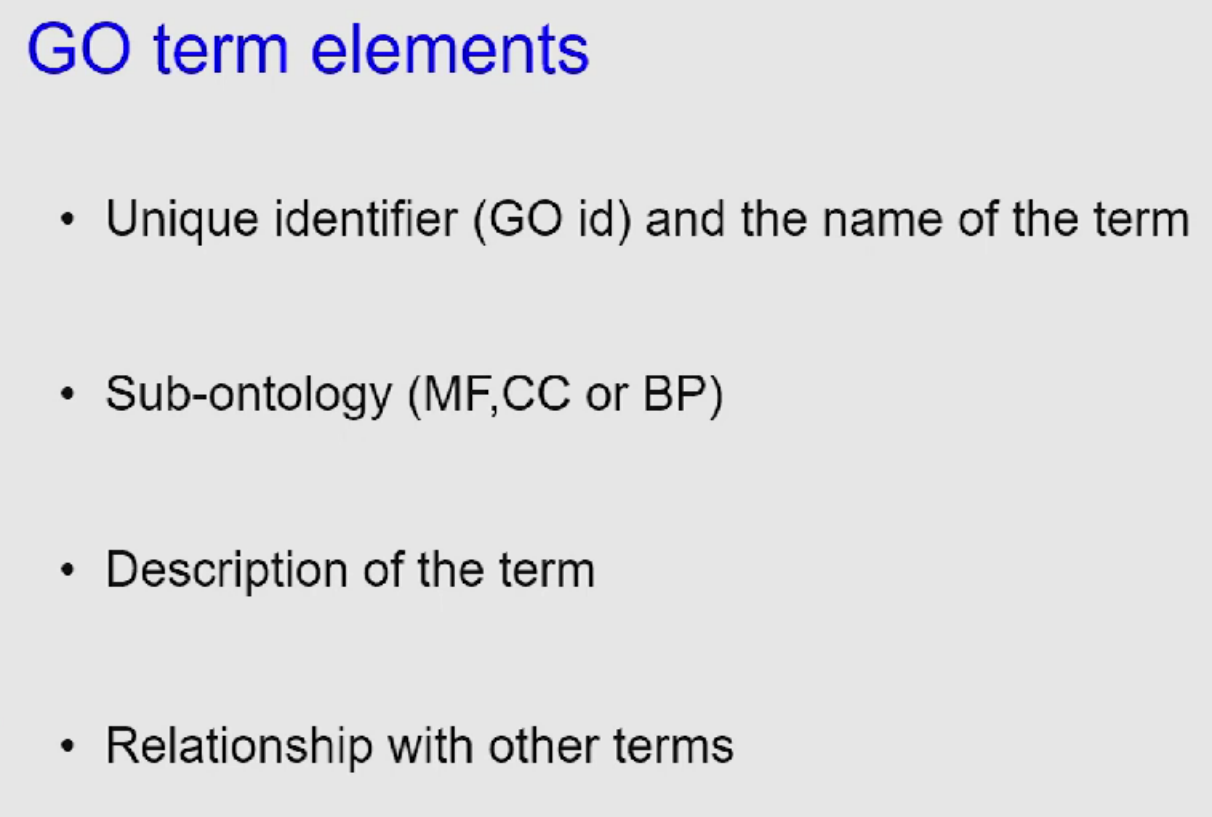


And again, depending on your requirement, depending on your analysis, what you are looking at is whether you are interested in the actual molecular function or you are looking at the larger biological picture. Depending on that you will choose the type of sub-ontology. And the last one is the cellular component, right? So, this is actually associating a gene with its location in the cell or sub-sample cellular compartment. For example, if a gene is located in the cytoplasm, cell membrane, or mitochondria, this will be mentioned in this sub-ontology. So, if you are looking at the localization of genes, then you will probably look at this cellular component part, okay?



So, there are some GO-term elements that will be present in ontology. So, the first one is the identifier, or sometimes we refer to it as the GO ID, and the name of the term. It will also mention the sub-ontology MF, CC, or BP, and it will have a description of the term and what this means, as well it will also describe the relationships with other terms that are present in the ontology. So,

for the annotation part right, there is an annotation, and this is organism-specific. Again, you can look at the links right, and this is regularly updated, and you can see that this is the most recent version that I have mentioned. You can go and see these annotations. These are organism-specific annotations because here the actual association happens.

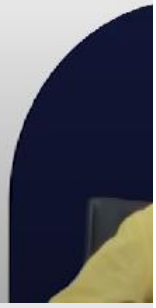


So, each gene in an organism is connected to a specific function. So, these are the components of this annotation file you will find. So, if you have genes or any protein names present, then you have an association with a GO ID or GO term that will have a very brief description of the function, and then you have some evidence, ok? So, you can now see how you can combine ontology with this annotation. So, ontology gives us this GO ID, the term, and the explanation to the description of the term right here. Here in the annotation, you are associating the genes with the GO ID or GO term, and hence associating them with specific functions.

GO Annotation

http://current.geneontology.org/products/pages/downloads.html
Release date: 2023-06-11

- Organism-specific annotation
- Gene/RNA/Protein
- Association with GO id/ GO term
- Brief description of function
- Evidence



Evidence

- Experimental (EXP)
- Inferred from expression pattern (IEP)

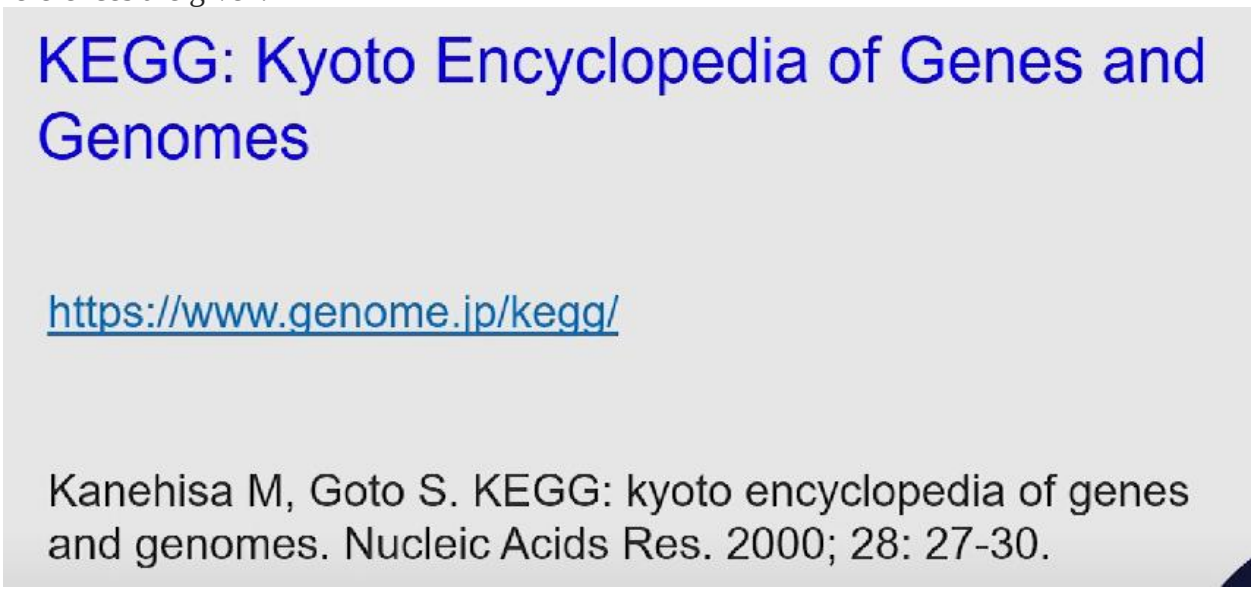
…..

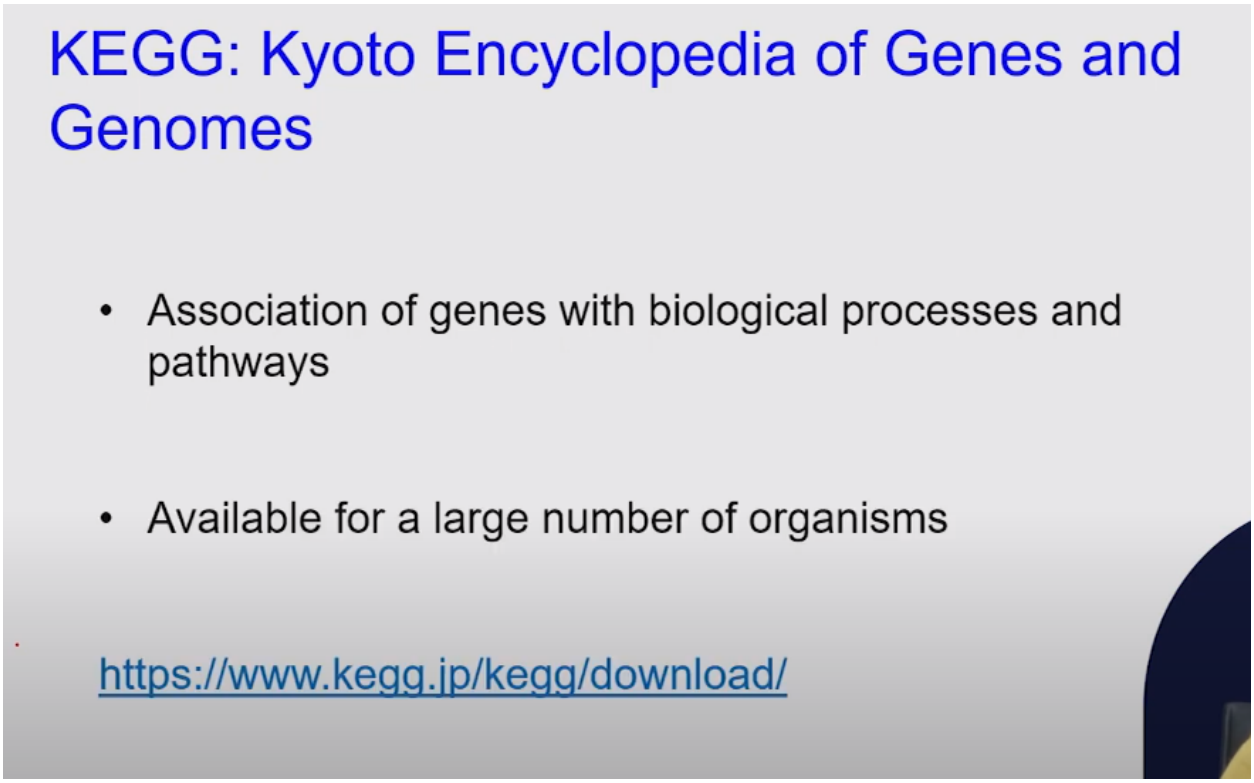
More details:
http://geneontology.org/docs/guide-go-evidence-codes/

For the evidence part, you will see in the data that there are a lot of details about how these evidences are collected and how they find which gene is doing which function in the cell. It could be experimental evidence, right? This has been experimentally determined in the organism right? But sometimes this is not possible. So, it is also maybe inferred from the expression pattern or

from computational analysis. So, all these things are given in the annotation part, right?

So, how strong is the actual evidence for that functional association? So, again, you can go to the link and see the details. There is another database where you can find this kind of association of genes to biological processes or pathways; it is called the Kyoto Encyclopedia of Genes and Genomes. Again, here is the link, and here is the reference paper. You can again look at the link and find out about these associations. So, you can associate genes with biological processes and biological pathways here, and this is also available for a large number of organisms. Similarly, you have another database called MSigDB molecular signature database. Again, the link and the references are given.





And here you have organism-specific gene sets, which again look at molecular signatures. So, identifying which genes are involved in which specific processes is right. So, again, you have something called hallmark gene sets. For example, you can have oncogenic signature or cell type-specific signature data sets. So, some cell types will have the expression of a certain gene set, and so on.

MSigDB: Molecular signatures database

- Organism-specific gene sets
- Hallmark gene sets
- Oncogenic signature, cell-type signature data sets ….

So, these kinds of data sets are present in this database, right? So, you can again explore and see. And similarly, there is another database called Reactome. Again, this means you have similar information right where you have associations of genes with biological processes and pathways. So, these are the references for this class—quite a few, actually.

Reactome

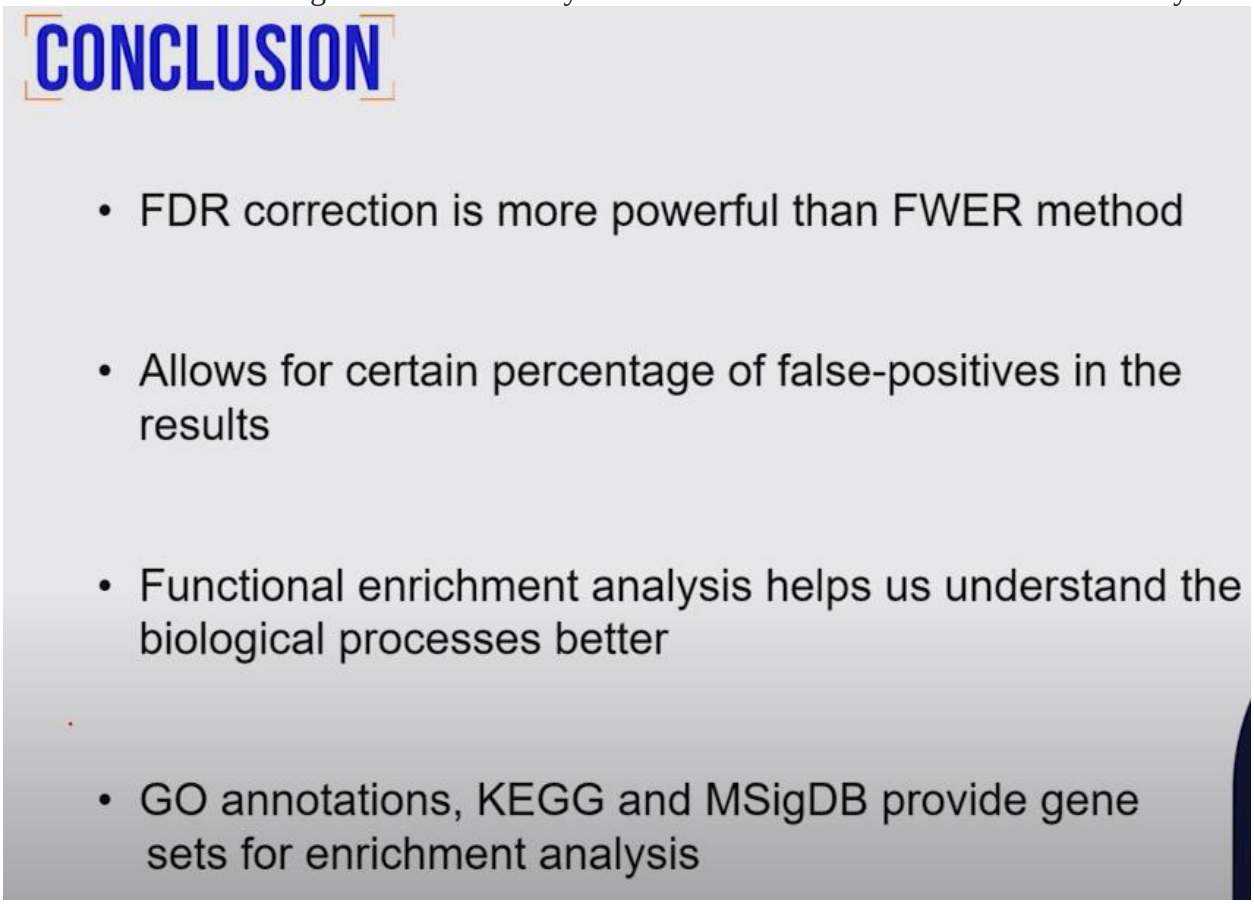
- https://reactome.org/
- Jassal *et al*., The reactome pathway knowledgebase. Nucleic Acids Res. 2020; 48(D1):D498-D503.
- Association of genes with biological processes/pathways

So, again, look at all the databases here. So, in the first part of the class, we talked about the FDR correction method, and compared to the earlier method, like the FWER method, what we have

seen is that FDR correction is more powerful. So, the power actually comes from this type 2 error. So, FWER commits a lot of type 2 errors, whereas FDR reduces those type 2 errors. So, what you see is that FDR can tolerate a certain amount of type 1 error, but it also increases the type 2 error.

So, these two type 1 errors and type 2 errors are balanced now. So, FDR can detect true positives more efficiently than FWER. So, FDR allows for a certain percentage of false positives in the results as we have seen, and this is acceptable in many cases, including differential gene expression analysis. So, if you are identifying a large number of genes, if 5 or 10 percent of them are false positives, they are not going to substantially bias our analysis or the downstream inferences. So, this is why we can tolerate this kind of false positive in many situations. Again, I have also talked about where you will apply the FDR method and where you will apply the FWER method.

Then we have moved on to the interpretation of the DGA analysis results, and one of the first steps in this process is the functional enrichment analysis. And as we have seen, we want to understand the functions of these genes that are differentially expressed. So, we want to associate these genes with certain molecular pathways and we want to see whether certain pathways or certain processes are up- or down-regulated. And this will actually help us gain more biological insights into the disease process, for example, stress response, drug treatment, etcetera. And we have talked about some of the databases right now that actually allow us to do this functional enrichment analysis.



For example, we have talked about geo annotations, we have talked about KEGG MSigDB, and we have seen right, we have very briefly mentioned right how they actually associate these genes

to their specific functions. So, again, there are different types of evidence that come into the picture. What is remaining is that we have not talked about the statistical test or the enrichment analysis itself. We have just talked about where we can get these gene sets and these functional associations of genes, but what you want to do ultimately is see whether any functional class is overrepresented or underrepresented in the list of differentially expressed genes.

And this is something that we will cover in the next class. We will actually look at the statistical techniques that will allow us to do this enrichment analysis and, finally, answer that question. Thank you.