Next Generation Sequencing Technologies: Data Analysis and Applications

Functional Enrichment Analysis

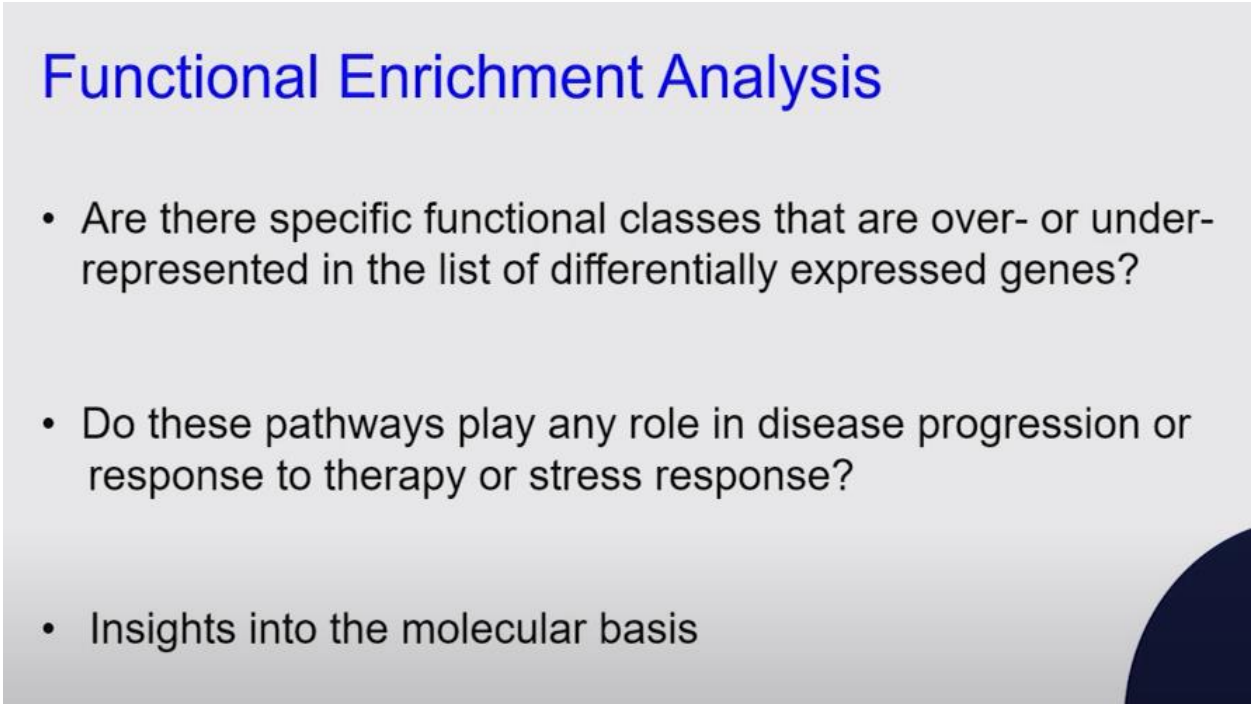
Dr. Riddhiman Dhar, **Department of Biotechnology**

Indian Institute of Technology, Kharagpur

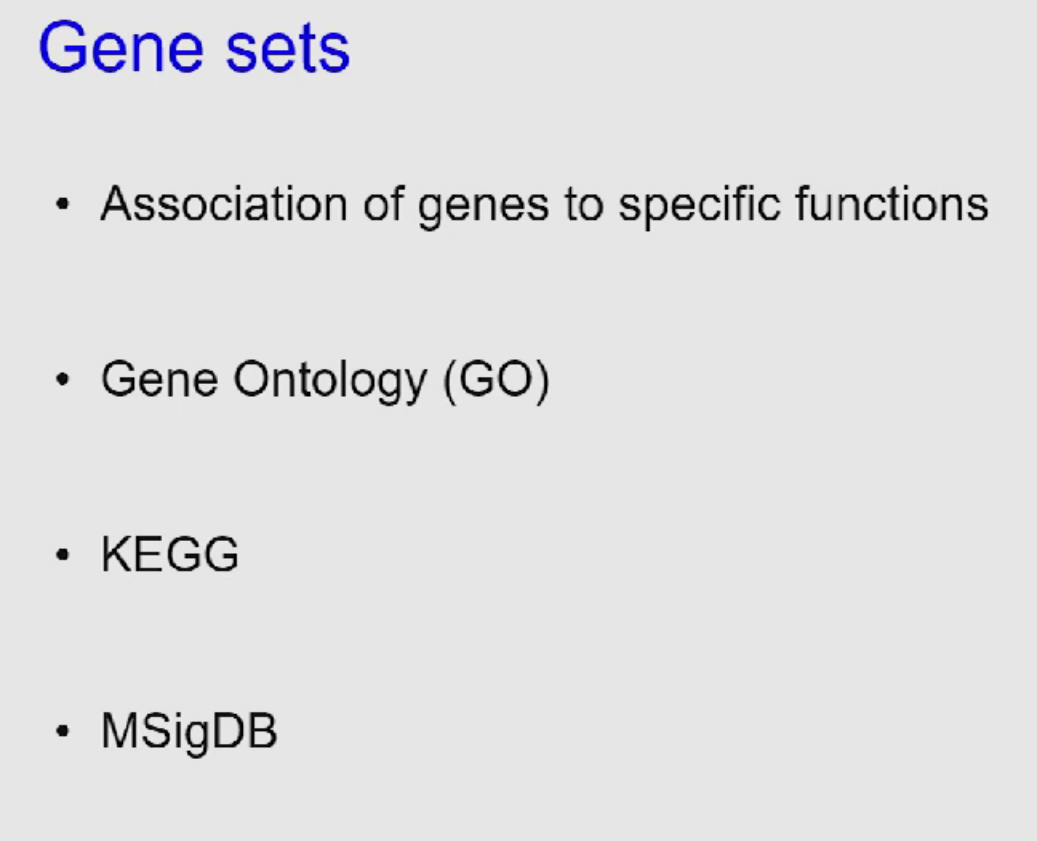
Good day, everyone. Welcome to the course on Next Generation Sequencing Technologies, Data Analysis, and Applications. In the last class, we started discussing the interpretation of DGE analysis results, and as part of that, we talked about functional enrichment analysis. This is a very common analysis that we do to understand the functions of the differentially expressed genes. And what you want to know is whether there are specific functional classes that are up-regulated or down-regulated in certain samples, and this will give us molecular insights and insights into the biological processes that are happening inside these cells. So, this is what we are going to discuss in this class, right?

So, we will be talking about this functional enrichment analysis, the actual enrichment analysis part, and this would be the agenda for this class. These are the keywords that you will come across: one is hypergeometric testing, and the other is GSEA. So, we will see that this is the gene set enrichment analysis. This is a specific tool that actually performs this enrichment analysis. So, just to remind you, we do something called functional enrichment analysis, which is also sometimes called gene set enrichment analysis or pathway enrichment analysis. They are not exactly the same, but they are also used interchangeably.

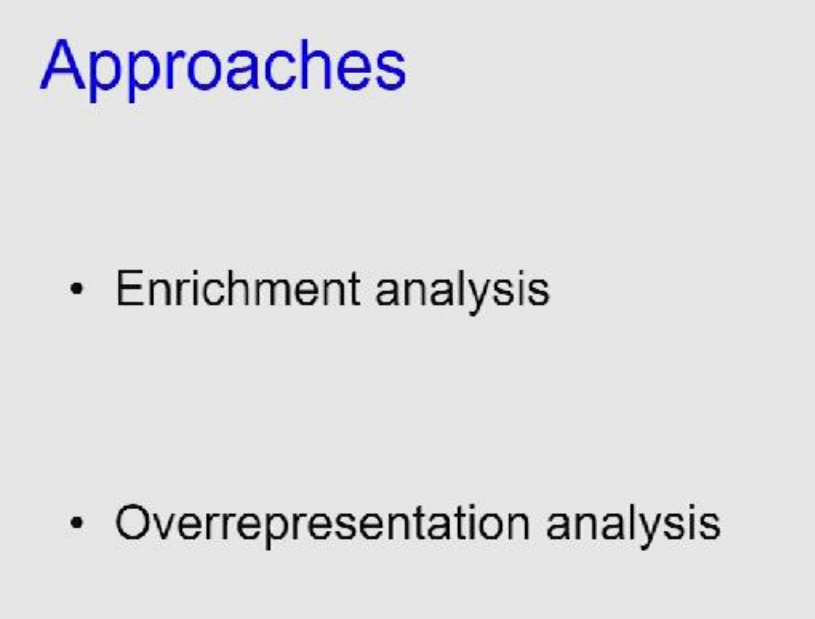
So, what functional enrichment analysis tries to do is answer these questions. So, whether there are specific functional classes that are overrepresented or underrepresented in the list of differentially expressed genes, And as I mentioned, the ultimate goal is to understand whether these pathways play any role in disease progression if we are doing this disease versus healthy sample comparison or whether they have any role in response to therapy. For many diseases, some therapy does not work. And in that case, we want to know which pathways are actually responsible for this therapy resistance and so on. Let us say we are analyzing stress response studies, where we expose cells to different conditions, such as condition 1 versus condition 2, and so on.



We want to know which pathways actually help in stress adaptation, whether in cells or in organisms. So, these are the ultimate goals of the analysis, and this actually gives us more insights into the biological process and helps us understand the molecular basis. So, for doing this functional enrichment analysis, the first step is to define the gene sets. And this actually associates genes with specific functions, and in the last class, we talked about different resources where we can find this information. So, we have introduced gene ontology, or GO, where you have the ontology part and the annotation part.

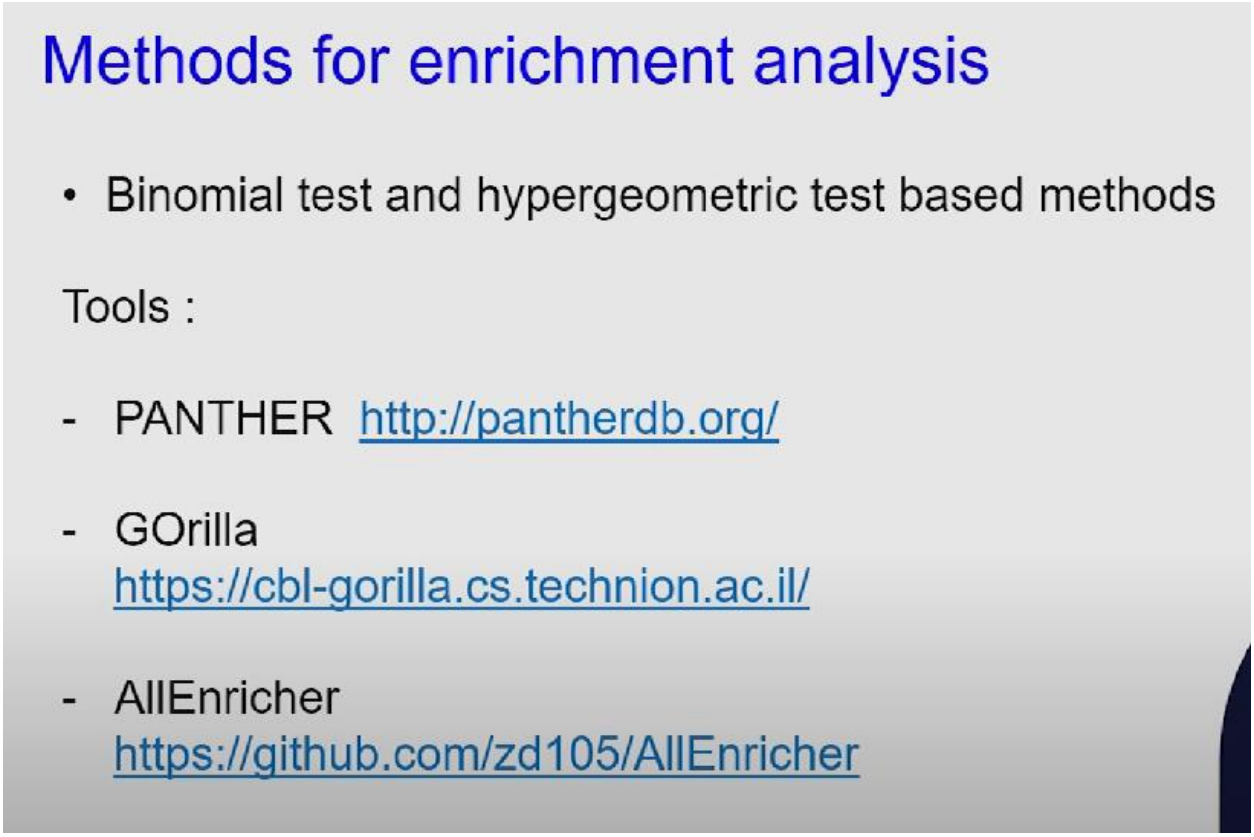


So, the annotation part is organism-specific, where you are associating genes in organisms to specific geo-terms and, hence, to specific functions. Then you have talked about the KEGG database very briefly. So, this is again associating genes to molecular pathways or biological pathways and then you have the MSigDB database, which connects genes to some gene sets. So, it organizes genes into gene sets, which could be signature gene sets or hallmark gene sets. So, once you have these gene sets, you can now do the enrichment analysis, ok?



So, there are two different approaches statistically for doing this enrichment analysis. The first is actually the enrichment part, and the other is overrepresentation analysis, and we will discuss these two methods. How we actually go about this enrichment versus overrepresentation is okay. They are statistically different. So, we will go into details about these statistics, ok?

So, the methods for enrichment analysis they actually test for enrichment of functional classes within a list of genes. So, in our scenario, where we have done the differential expression analysis, this list of genes will be the genes that are differentially expressed. So, these methods actually test whether certain functional classes are enriched within that list. And they usually compare this based on expectations from a global gene list, right? So, if you are looking at an organismal level right?

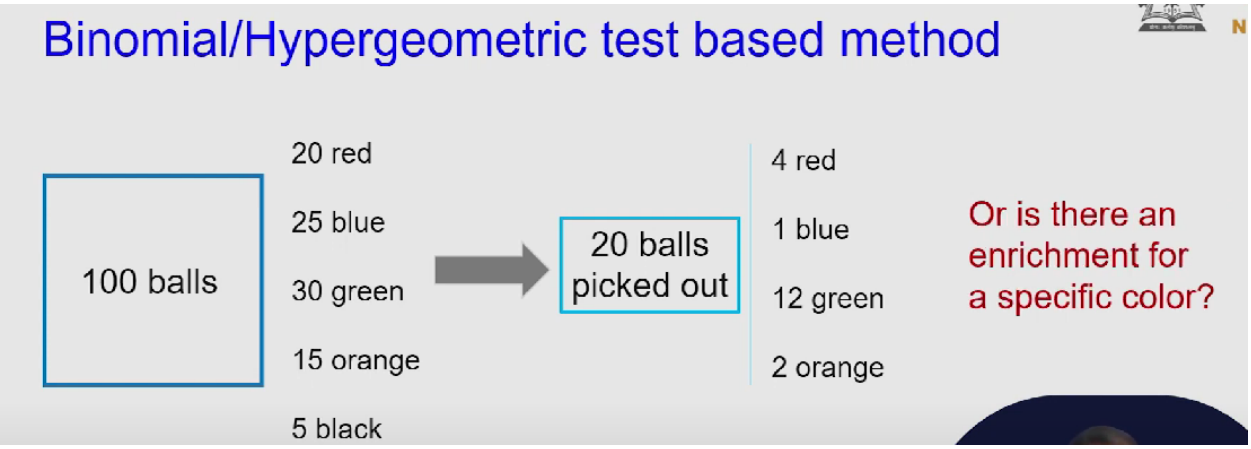


So, this is this with a global gene list, or maybe the gene list that you started your experiment with, right? So, maybe you are studying only a part of the genome, right? So, that will be your global gene list. Now, these methods rely on specific statistical tests. So, some of these methods are implemented in binomial tests, and some of them are implemented in hypergeometric tests.

Again, we will talk about these differences when you go into the actual statistical test. So, here are some tools, of course; there are many more out there, and the links are right. So, some of these tools are available online. So, you can simply paste your list, you can give the reference list or the global list, and you can compare whether there is any enrichment. So, you can find the results through these online tools.

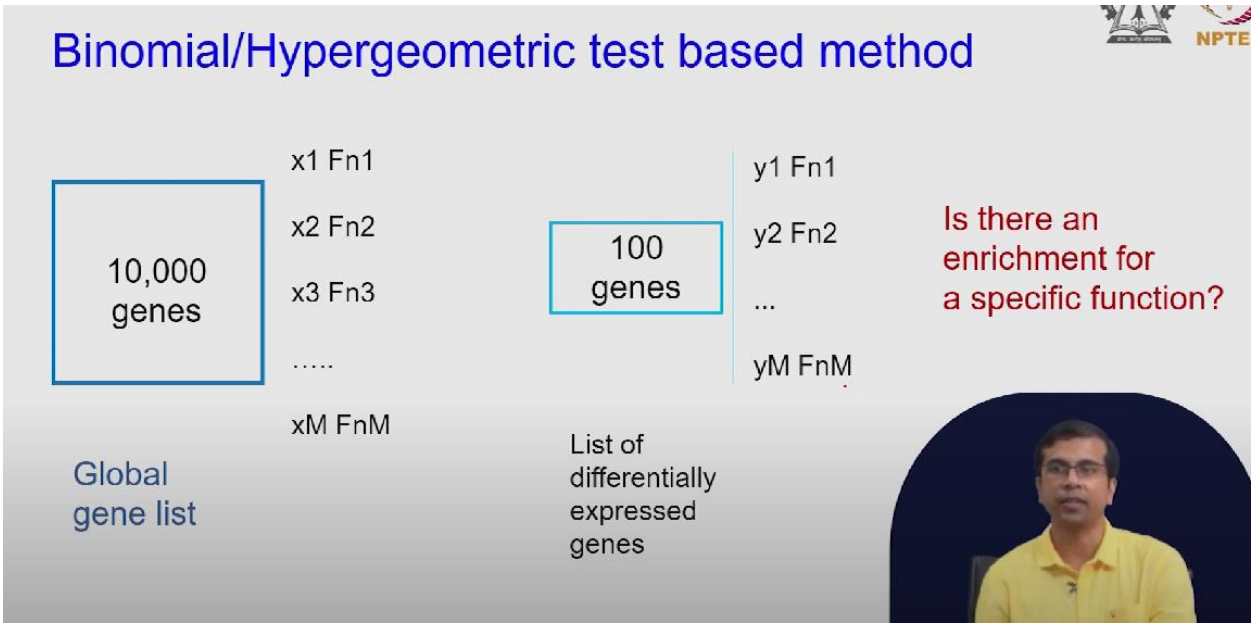
So, we will talk about this binomial hypergeometric test-based method and how it works. So, what is the idea behind this hypergeometric test or binomial test-based method? So, we will take an analogy first, and then we will come back to the gene analysis. So, let us imagine that we have a

box and 100 balls in it, ok? And out of them, you have 20 red, 25 blue, 30 green, 15 orange and 5 black.

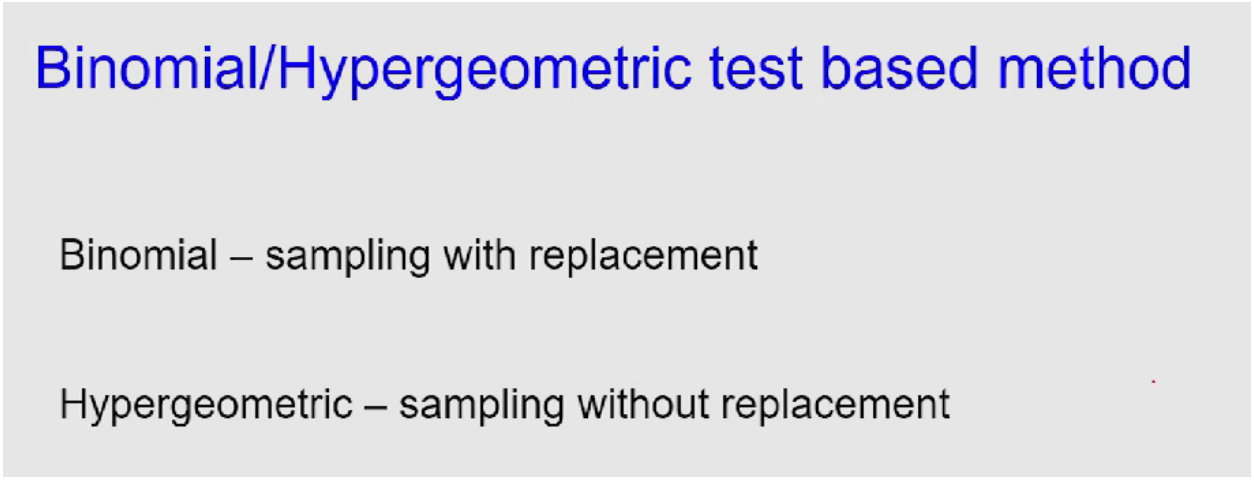


So, what do we do? Let us say that out of this 100, we pick out 20. So, we pick out 20 balls in another box, and what we see is that out of these 20, 4 are red, 1 is blue, 12 are green, and 2 are orange. Now the question that we ask, for example, is: is this distribution of colors right in this sample expected by chance alone? So, if you randomly pick 100 balls, would you get this distribution? Is this possible? Or is there an enrichment for a specific color, or do you see this is not a random sampling? Maybe you have picked out specific colors preferentially.

So, once you think of this example, it is actually very similar to what you want to do in the case of a hypergeometric test-based method or enrichment. So, this is the global box, and in our case, this would be the global gene list, ok? So, where we have about 10,000 genes, let us say that instead of the colors here we now have these functions. So, we have these x 1 functions x 2; these are the numbers, right? So, we have x1 genes that are in function class 1, x2 genes in function class 2, and xm genes in function class m.



So, this is the distribution of the 10,000 genes. Now, let us say we have this list of differentially expressed genes, and in our case, let us say we have 100 genes right on the list. And out of these 100 genes, we see y1 number of genes that belong to function class 1, y2 number of genes that belong to function class 2, and ym genes that belong to function class m. Now, once we have seen this right, the question is: is there an enrichment for a specific function, or is it just a random sample that we will get right? So, the question is, would we get the same distribution right here (y 1, y 2, y m) if we just picked a random sample from this global gene list?

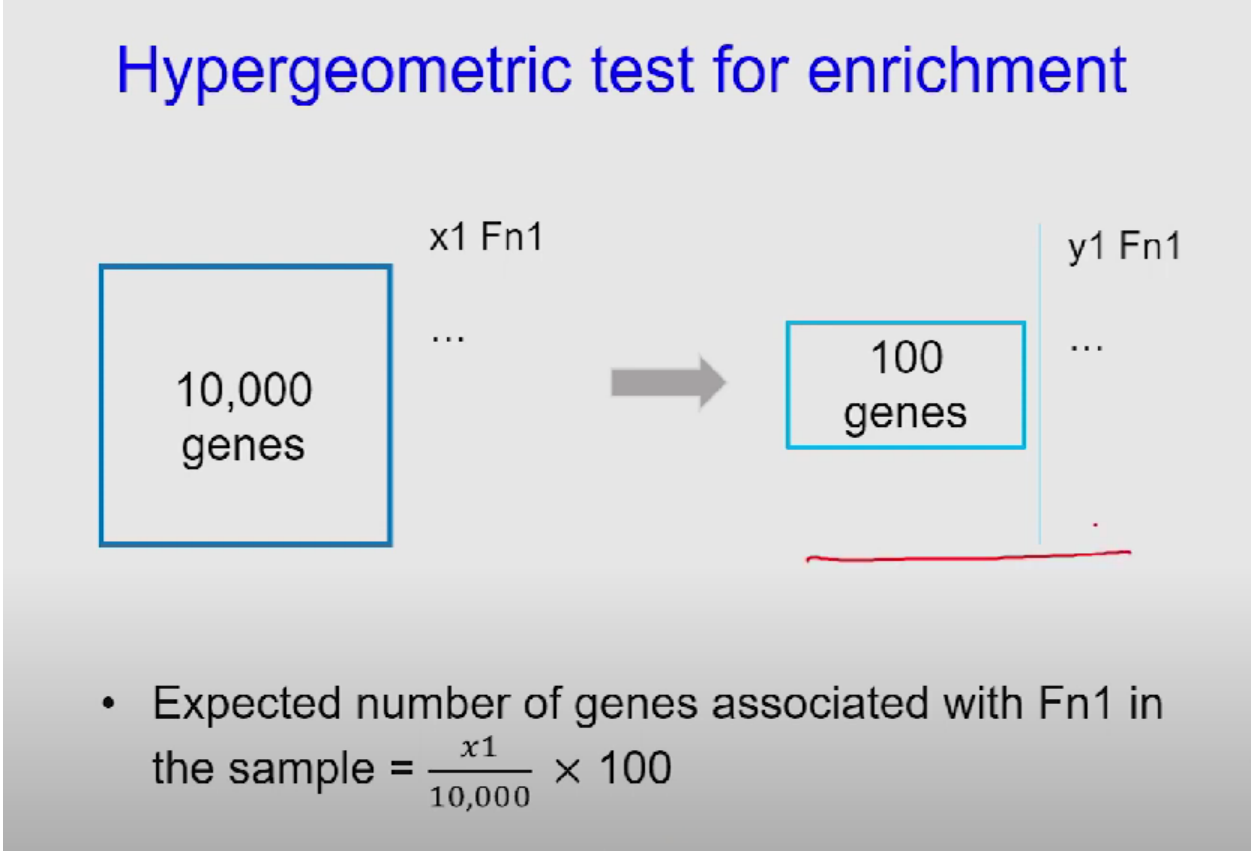


So, this is actually what we are going to answer, right? So, this will tell us whether there is an enrichment in the list. Now, the decision is whether we want to use the binomial test or the

hypergeometric test. So, binomial is sampling with replacement if you go back to your statistics classes, and hypergeometric is sampling without replacement, right? So, as you see in our scenario, this is mostly sampling without replacement, right?

So, if we pick a gene from the full list, that gene is not replaced, and we cannot pick that gene again. So, this is a sample without replacement in our case. So, we should use the hypergeometric test, and this is the most appropriate out of these two, ok? But some of the tests and some of the tools still use the binomial test, and this is an approximation of the hypergeometric scenario, ok? So, what do we do in this case? So, the question is whether there is any enrichment for specific functions in the list of differentially expressed genes.

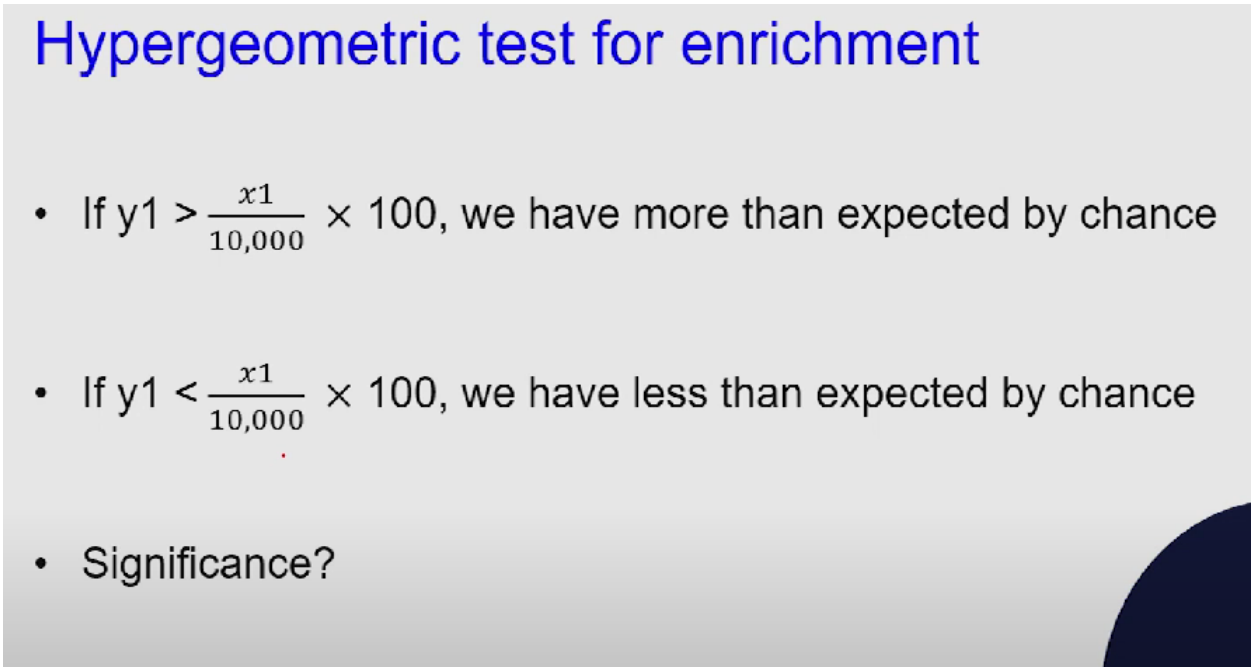
So, on the right-hand side, we have the list of differentially expressed genes. Now, to answer this question, what we do is perform a hypergeometric test, which is done on all functional classes, one by one. So, we start with functional class 1, then go to functional class 2, and so on, and we continue this up to functional class 3. And we do hyper-geometric tests for all these functional classes, ok? So, we start with functional class 1, and we will discuss how we do this hypergeometric test.



Now, for functional class 1, we have x1 genes right that belong to this functional class 1 out of 10,000. So, what is the proportion of these genes with genes that are in functional class 1 in the global list? This is x 1 out of 10,000. Now, given this proportion, if you pick 100 genes randomly from this global list, how many genes would you expect in the sample to belong to functional class 1? So, the expected number would be: if it is a random sample this will be function x 1 divided by 10,000 into 100. This means that many genes will be expected in this sample to belong to functional class 1. Now, based on this expectation, what you want to do is compare this y 1 with this expectation, and we want to ask how different or extreme y 1 is compared to this expected number x 1 divided by 1000 into 100.

So, this is the statistical test that we want to do, and for that, as you have said, the hypergeometric test, or hypergeometric distribution, is the best method. And we can transform this question into this question right when we are going to a distribution-based test: how often can we observe a more extreme value compared to y 1? So, how common is y 1 or how extreme is y 1 based on the

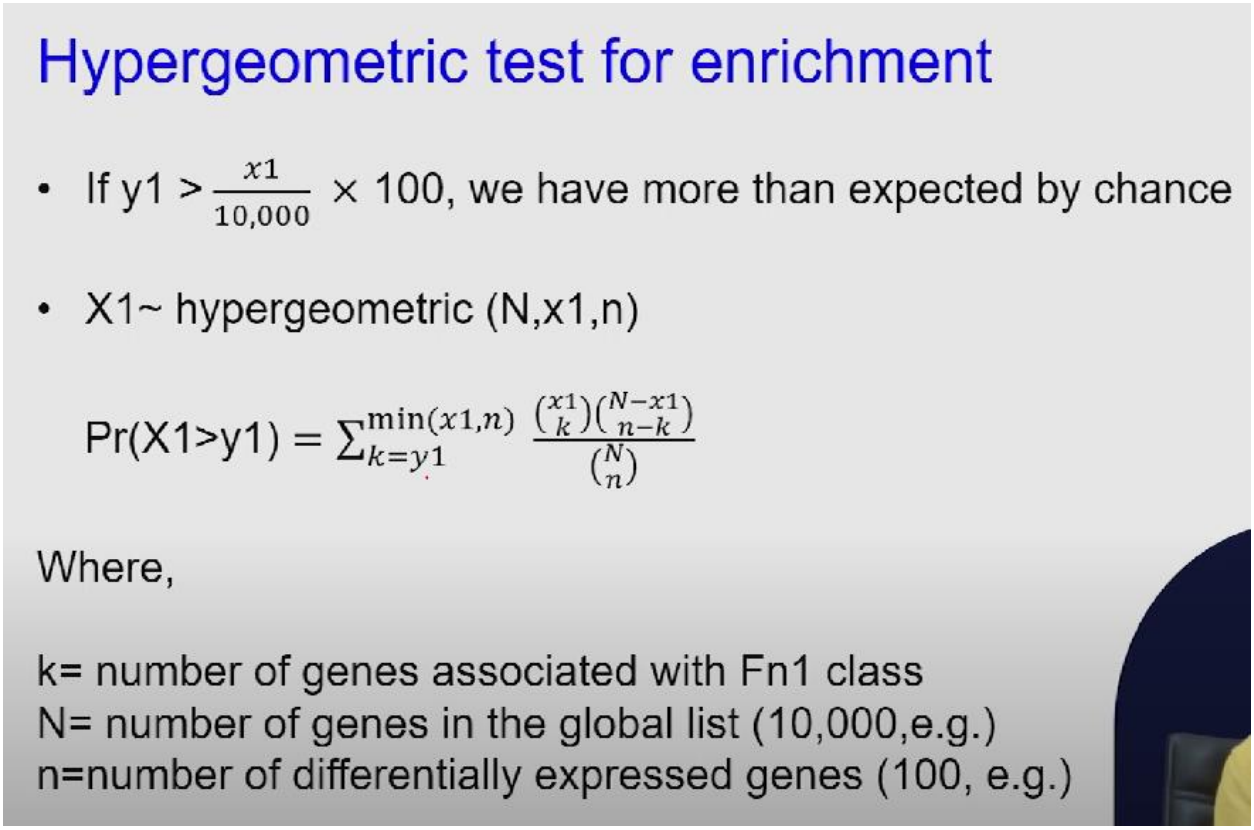
expectation that we have said? Now, if y 1 is greater than right, this x 1 divided by 10,000 into 100 is the expected number; we have more than expected by chance. So, if we pick a random sample right, this would be the number that would be expected right?



So, this is what is expected by chance. If y1 is greater than that, we will say we have more than expected by chance. On the other hand, if y1 is less than this, this is expected by chance. So, we will say we have less than expected, right? So, this is what is expected again, and if we have less than that, we say we have less than expected by chance. But now the question is whether these differences are significant, right? We have not talked about anything about statistical significance here, right?

We just said this is more than expected by chance or less than expected by chance, but we have not talked about significance. We need to calculate the statistical significance; without that, we cannot make any decisions. So, let us look at the statistical test that we can do, and maybe we can calculate a p value, which will help us make this decision. So, as we have mentioned, right x 1 follows hypergeometric. So, x 1 is the number of cases, or the number of cases or genes that belong to this function class 1 right in the sample. So, this is the variable that defines that, okay?

And this follows hypergeometry with these three parameters. capital N. This is the number of genes in the global list, right? So, if you are doing this genome-wide comparison, you will have all the genes of the organism in this list. So, you just simply count that number x 1 is the number of genes that belong to function class 1 in the global list, and small n is the number of samples or number of genes in the sample right. So, the number of genes is 100, for example, in our scenario, and this small n would be 100.



What we calculate is right. So, what is the probability that this x 1 is greater than y 1 or the chance of observing more extreme values compared to y 1? So, this is the probability that gives this probability value, right? So, this is what we calculated, okay? And this is given by the hypergeometric distribution. So, if you look at the hypergeometric distribution formula, this comes from this hypergeometric distribution.

And we are taking the sum right because we are calculating this probability of greater than y 1 x 1 greater than y 1 ok. And this is calculated by k equal to y 1. So, the sum over k equals y 1 to the

minimum of x 1, and small n is okay. So, this will give us the probability of observing more extreme values compared to Y1. So, you can look at the hypergeometric distribution formula, and you will see we can calculate this in this way.

Now, once we calculate this, So, just to mention k is the number of genes associated with function 1. This is what we have discussed. So, this can vary from y 1 to x 1 or n. And then you have capital N, which is the number of genes in the global list. So, in our example, this is 10000, and N is the number of differentially expressed genes, which is 100 in our example. So, this probability value will give us the p value right.

So, if you have let us say this comes to 0.01, right? So, very rarely, we will see more extreme values than y1, right? So, that will give us the p value. Similarly, if we can think about the other situation, So, if y1 is less than the expected value, we have less than expected by chance, but we need the statistical test or the p value from this.

Again, x 1 follows this hypergeometric distribution with these three parameters, and here we calculate the probability of observing lower p values than y 1 right. So, again, we are looking for extremes, right? So, how extreme is y1 in the whole distribution? Again, the formula is that this part is the same, but we are taking the sum differently, right?

So, this starts from 0 to 1, right? So, we can see, let us say, 0 observations, or we can have up to 1 observation. So, we are again looking at how many extreme values there are or how likely it is to observe more extreme values compared to y 1 and the k N capital N and small n; they are the same in this case. So, once we have this right, So, once we calculate this probability value, this sum is right. So, for each of these equations, we know all these values.

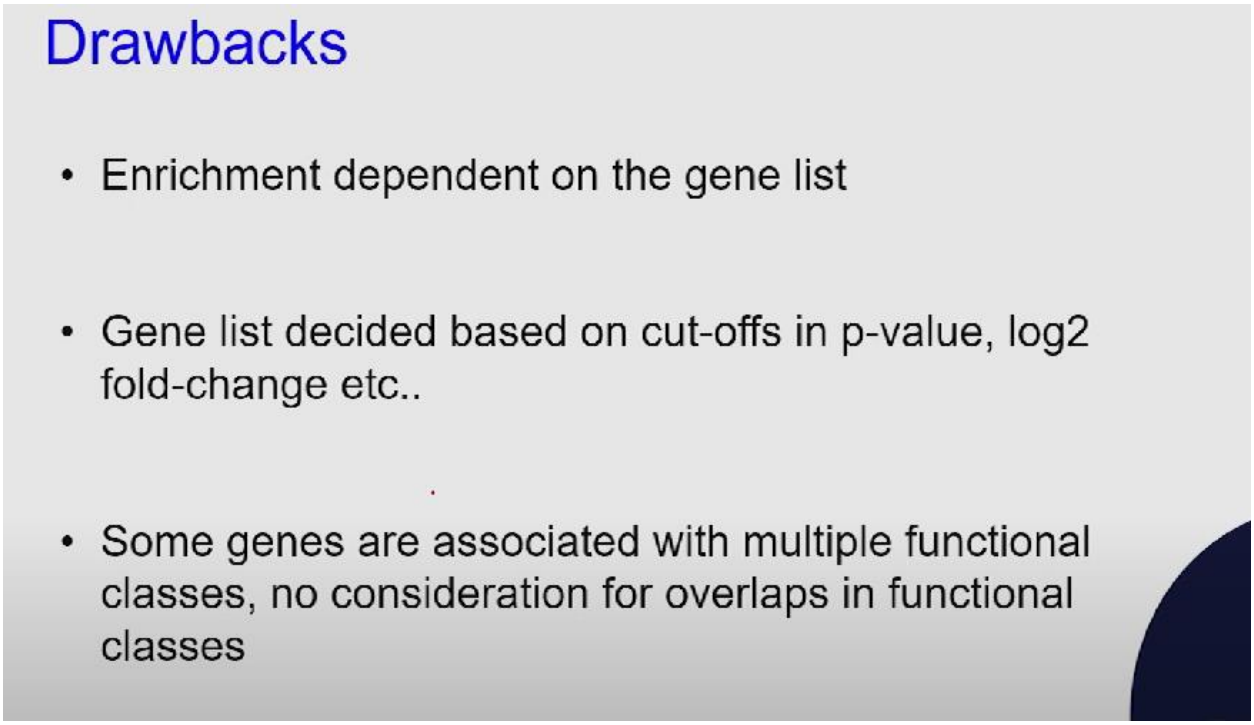
So, we can calculate the probability, okay? We know x 1; we know k; we know small n; we know capital N; depending on the situation, all these values we know, and then we can calculate these probabilities. So, these probability values provide p values for this test, and after we have these p values, we have to do the multiple hypothesis testing correction because we repeat this test for all functional classes. So, or gene sets, right? So, if we have, let us say 8. So, maybe for 100 functional

classes, we are doing this hypothesis test or hypergeometric test for all functional classes, right?

So, we are doing 100 tests. So, we will have to do the multiple hypothesis testing correction we just talked about, and usually we use the FDR correction. So, again, we can tolerate certain false positives, maybe 5–10 percent. So, it is right that we use this FDR correction.

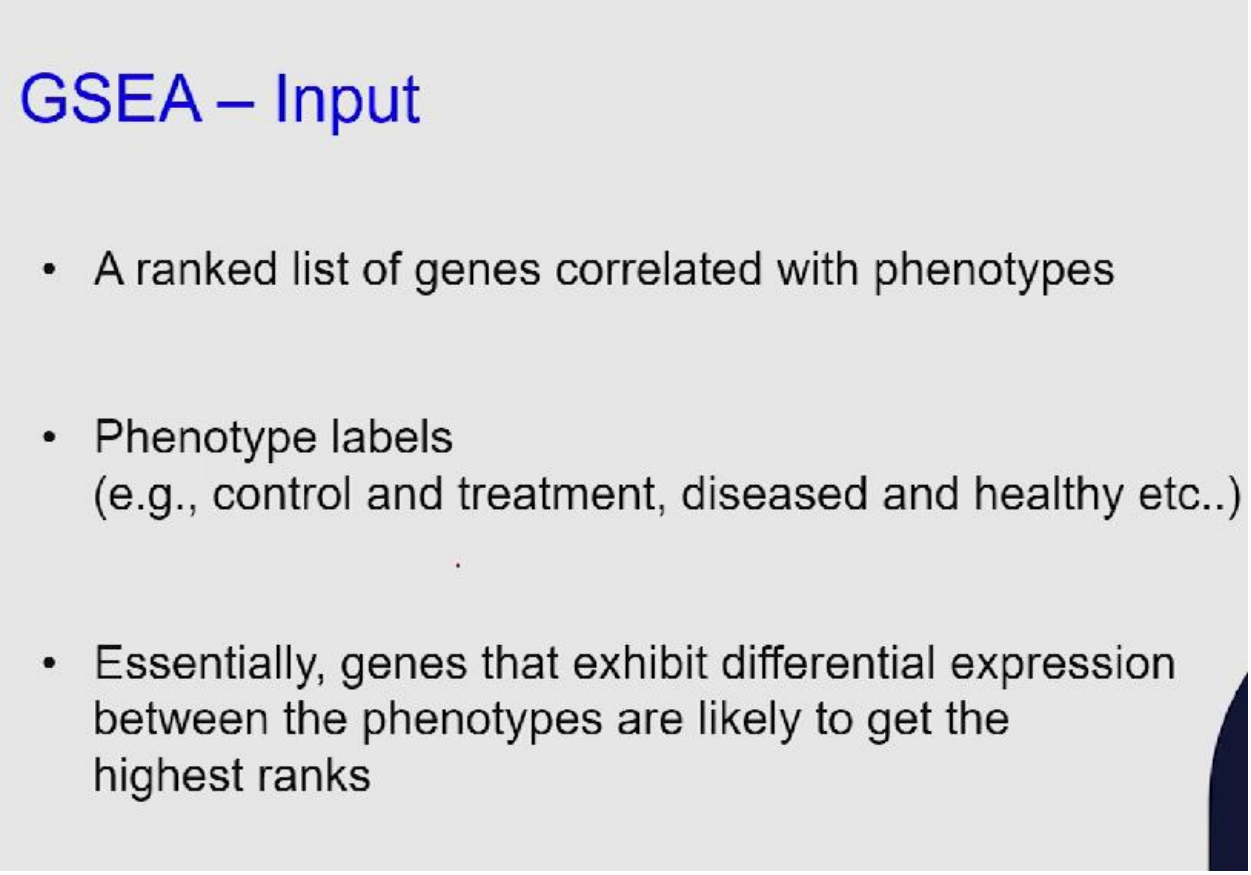
So, again, we said this FDR was less than 0.05 or 0.1 for identifying significantly enriched or underrepresented functional classes in the list of differential expressions. So, what are the drawbacks of this method? This method is quite simple: we are taking the list and looking for enrichment compared to global expectations. So, global expectation means if we look at all genes in the organism, So, what is the drawback, or what are the drawbacks of this method? So, this enrichment that you will get is dependent on the gene list that you are providing. So, if we have, say, 100 genes we will get some sort of enrichment; if we have, say, 120 or 150, we will get some other enriched functional classes.

Now, the reason I am saying this is because this gene list that we decide is often based on the cutoffs. So, this is something that we choose: the p-value cutoff, the FDR cutoff, or if we want to impose a log2 fold change cutoff, etcetera. So, these are hard cutoffs that are decided by us most of the time. So, in that case, this gene list is somewhat arbitrary. So, this can change depending on the person who is deciding on these p-value cutoffs or log2 foldchange cutoffs.



So, what it means is that the results will be variable, and we are also ignoring this fact when we are doing this analysis because some genes are associated with multiple functional classes and we are not considering overlaps in functional classes. So, there will be overlaps in functional classes because one gene can be part of two different functional classes or even more, and that information is not considered at all. So, we consider every gene independent, and we just do the statistical test. Now, the other approach is the overrepresented overrepresented analysis based on GSEA. So, the tool is a gene set enrichment analysis tool. It is a hugely popular tool, and here is the reference and the link for looking at the tool, and there is a very nice detailed discussion or explanation of the tool.

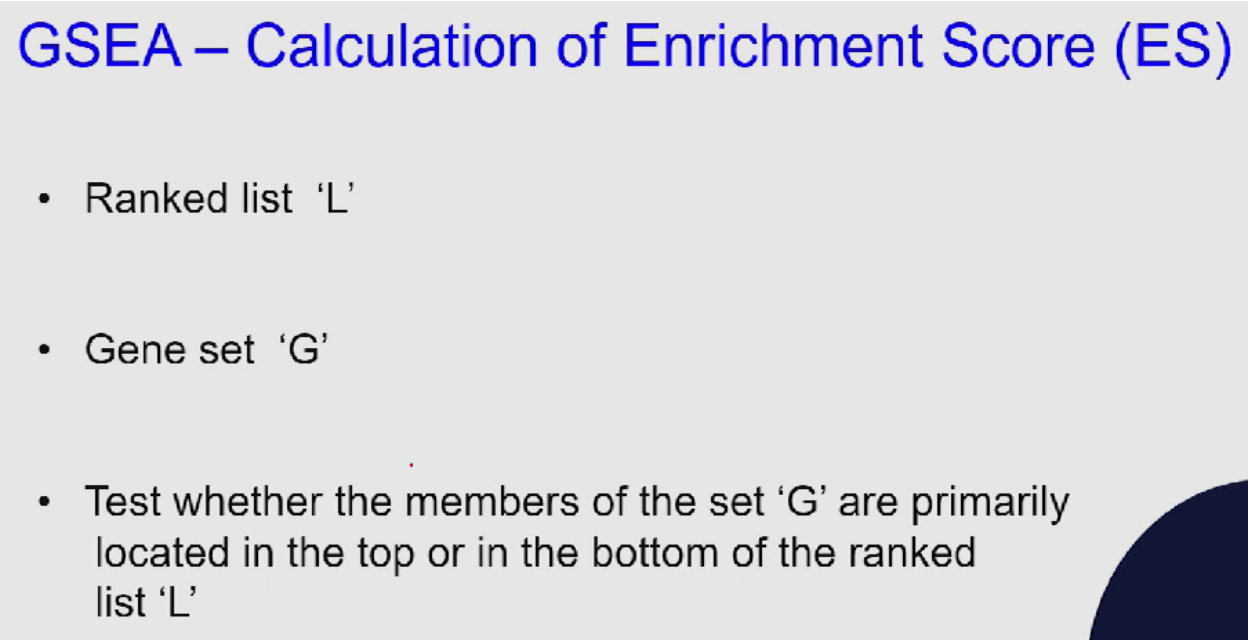
So, for GSEA, we will go into how it differs from the earlier methods, such as the geometric test or the binomial test. For GSEA, the input is a rank list of genes correlated with phenotypes. So, it is not the list of differentially expressed genes anymore, ok? It is the rank list of all genes that are present in your assay, right? So, when you are doing this ionistic let us say you have identified, you have your sequence of 6,000 transcripts, you have data for 6,000 transcripts, you just rank list, you provide the rank list of these genes or the transcripts and you correlate that with the phenotypes. So, it means you have to sort this list according to their correlation with phenotypes.



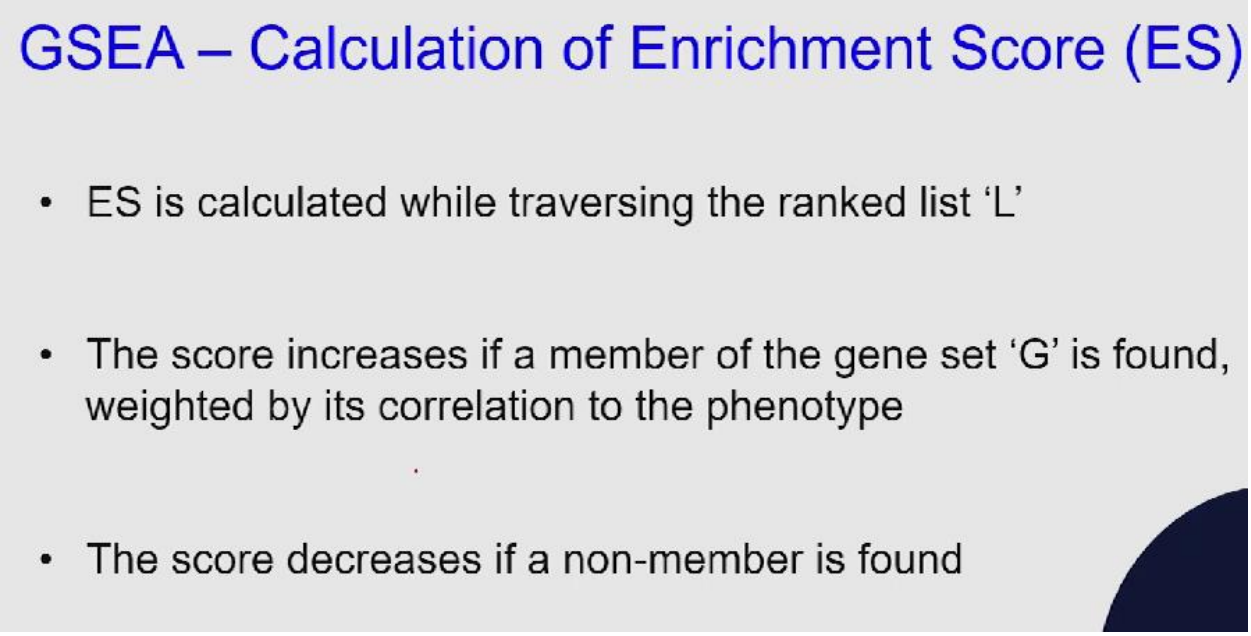
And you also have phenotype levels for example, you have control versus treatment levels, disease versus healthy levels, etcetera, and it actually essentially boils down to the genes that exhibit differential expression. Okay, but here you are not providing this list of differentially expressed genes only. What you see is that the differentially expressed genes between these two levels will come at the top; they will get the highest ranks because they are able to differentiate between these different phenotypes. And by the way, it is not necessary to have just two levels; you can have multiple levels in many cases, right? So, for example, if you are doing this expression analysis for different conditions, you can have 10 conditions, 10 treatments, or 100 treatments for that kind of situation.

So, you can have multiple levels, ok? So, as part of the input, you also need the gene set data. As I mentioned, you can get this gene set data from MSigDB or other databases where you have genes that are part of specific gene sets or hallmark sets that are defined. And once you have this data,

this GSEA calculates something called an enrichment score, or ES. So, what you will do here is that it has a rank list L that has been provided by you. So, you are providing this rank list based on your experimental data. You have a gene set G right that has been obtained from MSigDB. So, what this method will do now is test whether some of these gene sets are overrepresented in the rank list.



So, or in terms of whether these gene sets are overrepresented in the top part of the rank list or the bottom part of the rank list, So, why the top part? Because again, this is a rank list. So, the top genes are very interesting because they show probably the most differential expression. So, if these gene sets are overrepresented at the top, this means they are probably associated with differential expression and have some important biological consequences. So, this method, as I mentioned, is to test whether members of set G are primarily located at the top or bottom of the rank list, and that is what will give us more insights into the biology of the system.



So, how does this method calculate this ES score? So, ES is calculated through the traversal of the rank list L. So, this method will start from the top of the rank list, right, and it will traverse; it will take a gene, right, and you remember it has this gene set. So, what it will do is see what the gene is in the rank list and compare it with the gene set. And if there is a match, if a gene in the rank list is present in the gene set, it will increase the ES value. And there is a weight, of course; it is weighted by its correlation to the phenotype.

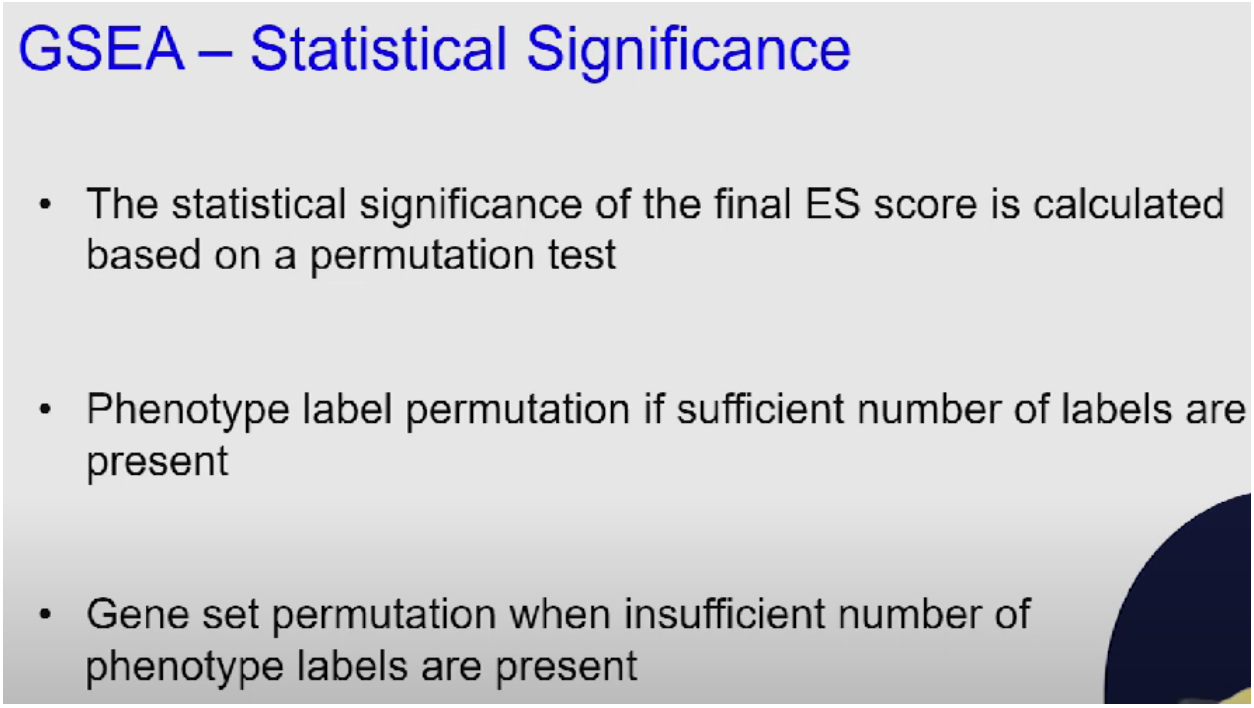
So, if it is at the top right, So, it will get this correlation, and that will be weighted by that correlation, ok? So, this increase is weighted by the correlation to the phenotype. And in this comparison, right? So, this comparison is going on if the score is wrong and the math is not right. So, if a gene is not found right on the list, the score will decrease.

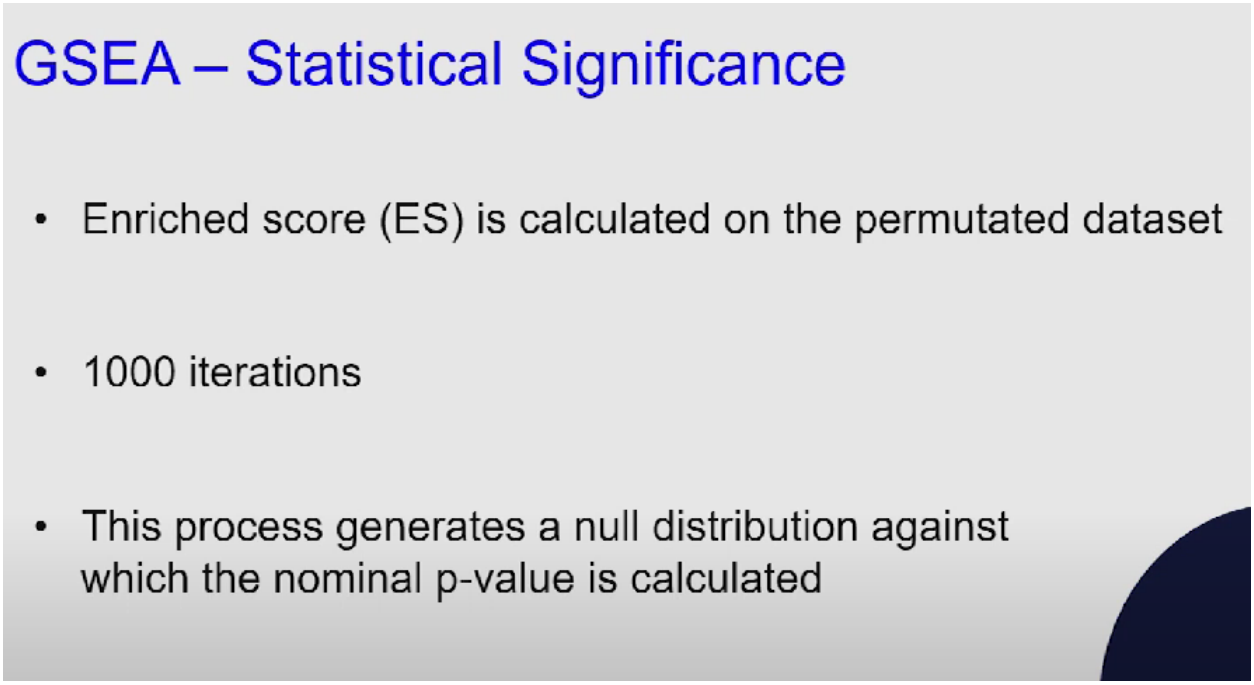
So, you can have these two counts going on, right? So, you are looking at the rank list, you are considering these genes from the gene set, and then you are counting right; you are seeing whether there is a match; if there is a match, you have an increase in score; if there is no match, then there is a decrease in score. So, how do you calculate the final score? This final ES is the maximum deviation from 0 while traversing the list L. So, if you see some of these genes in the gene set are enriched at the top right, you will see this score going up, and then this will go down. And the

maximum deviation from 0 is the final enrichment score. Now, once you have calculated this score, how do you actually generate the statistical significance?

So, this is a very important part, right? You can generate any score, but you need to say whether this is statistically significant or not. So, in the case of GSEA, the statistical significance is calculated based on a permutation test. So, how does it actually work? So, you can imagine, right? You have this rank list, right? You can draw here, right? So, you have this rank list of genes right rank 1, rank 2, etcetera. You also have the phenotype levels right, for example, P1, P2, P3, and P4. So, you have these data sets right, and once you have these phenotype levels right,.

So, what you do is permutate these phenotype levels. So, you change these phenotype levels and see whether the gene set is still enriched or not. So, you calculate this enrichment score again, okay? So, you perform this phenotype level permutation right if you have sufficient number of levels present right? Unless you have let us say 7, 8, or so that is recommended by GSEA, you cannot do this process, ok? So, it does not make any sense to do this permutation if you have only let us say 2 or 3 levels. So, in that case, what you do is do something called gene set permutation. When you do not have enough phenotype levels, you permute these gene sets, which means that you are working with keeping the size of the gene set the same. You do not change the size of the gene set.

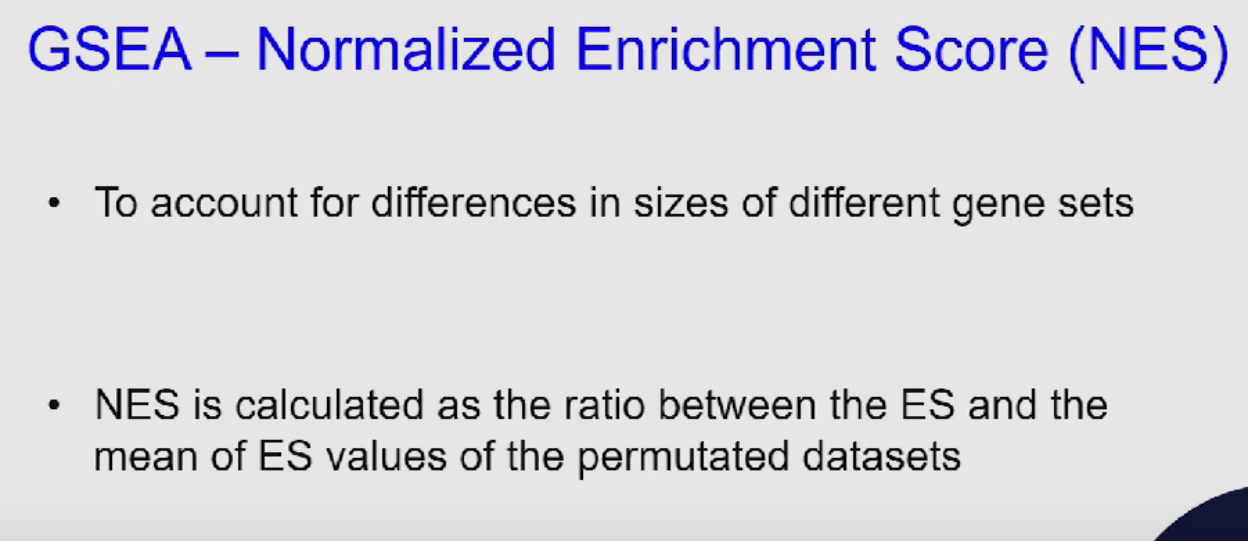




So, let us say that in a gene set, you have, say, 15 genes. You keep the size the same, but you permute the genes and then again calculate the enrichment score. So, once you have done this permutation, it actually generates something called a null distribution, and you perform this permutation test or permutation process usually 100 to 10 times, and 1000 iterations is

recommended by GSEA. And this will generate the null distribution against which you can calculate the p value. So, you can compare this enrichment score with the enrichment score from the original data set, and then you compare it with the null distribution and calculate the nominal p value again, looking at how extreme this enrichment score is compared to the null distribution that has been generated by the permutation. So, GSEA, we are again doing this multiple hypothesis testing because you are testing each gene set separately for the data set.

So, you again have to do multiple hypothesis testing and correction. So, this is required, and this is done through the FDR correction and the Benza or Benjamini-Hochberg method. In GSEA, there is something called a normalized enrichment score that is also calculated and is called NES. So, this is actually done to account for differences in the sizes of different gene sets. So, when you are working with these gene sets in MSigDB, you will see that not all gene sets are of the same size. So, a gene set might have 10 genes, some might have 15 genes, and some might have 50 genes.



So, again, to account for these differences, the number of genes will affect the enrichment score because you will find you are likely to find more matches, etcetera. So, to account for differences in size between these different gene sets, this NES is calculated, and this is calculated as a ratio between the actual enrichment score and the mean enrichment score values of the permutated data sets. So, once you have done this permutation, let us say 1000 times, you will get this enrichment

score for all those data sets, and you calculate the mean and take the ratio of the actual enrichment score versus this mean of the enrichment score that is generated from the permutated data sets. So, here are the references for this class. So, what we have seen is that functional enrichment analysis can give us insights into the actual biological processes.

So, we can ask questions that are actually biologically relevant. For example, you can say whether a certain gene is associated with a certain disease, whether it is helping in the progression of a disease, whether it is involved in the response to a therapy, or whether it is generating resistance to a therapy. We can also see whether a gene or pathway is associated with stress response. So, stress adaptation, and so on. We have talked about two different approaches. The first one is hypergeometric test-based enrichment analysis, where we assume a hypergeometric distribution.

So, we are assuming a sampling without replacement scenario. So, the idea is very simple: we talked about a list of differentially expressed genes. So, we get this list of genes, and we test whether certain functional classes are overrepresented or enriched in that list. So, that is one approach, and we do that through the hypergeometric test. We discuss the method, etcetera, and how we generate the p value and do the multiple testing correction. This will give us the functional classes that are enriched in that list or underrepresented in that list.

Now, the drawback of this method is that it relies on that list. So, this list is generated by these p-value cutoffs or log foldchange cutoffs and is a bit subjective. So, depending on the person who is doing this analysis, So, we might get different results for enrichment because the list might differ. In addition, this hypergeometric test does not take the overlap in gene sets into account. So, we have a single gene that could be part of multiple functional classes, and this overlap is not taken into account when we are doing this hypergeometric test.

We assumed all these sets were independent, and we just went on with the analysis. So, there is another approach that we have talked about, which is overrepresentation analysis, and this method is GSEA. So, we have talked about this, and the goal of this method is slightly different. So, it starts with the rank list, okay? So, rank the list of genes that have a correlation with specific phenotypes, and we also have the phenotype labels. And what GSEA does is look for gene sets

being overrepresented or underrepresented at the top or bottom of this rank list, because this will give us the most relevant biological information.

And it calculates something called an enrichment score from this process, and through the traversal of this list, it calculates this enrichment score, which is weighted by their correlation. And we have seen that the statistical test is done by a permutation test, where the method actually shuffles the phenotype labels when enough phenotype labels are available, or it shuffles the gene sets and generates a nominal p value, and finally, we also get a normalized enrichment score calculated. Again, for each of these methods, we need to do multiple hypothesis testing corrections because we are doing statistical tests for multiple functional classes. So, once we have done this, we identify a list of functional classes that are overrepresented, and then we can perhaps probe these classes further and see whether some of them are responsible for disease progression, for example, or maybe stress response and adaptation, and so on. So, these methods actually give us valuable biological insights into the questions that we started with.

So, when we start this differential expression analysis with those questions, So, what is the difference between, for example, disease versus healthy samples? We want to identify at the end the molecular basis or the biological process that is actually probably driving the disease. Thank you.