Next Generation Sequencing Technologies: Data Analysis and Applications

Differential Gene Expression 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 few classes, we have covered the preliminary steps for transcriptomic data analysis. So, namely, we have talked about transcriptome assembly, we have talked about normalization and now we are ready for the main analysis, right? This is a very popular analysis that we do on transcriptomic data sets, and the analysis is called differential gene expression analysis. So, this is the agenda for today's class.

We will talk about differential gene expression analysis; we will talk about the concepts; and then we will introduce some of the tools that we use for doing this kind of analysis. So, let us get started. Here are the keywords for this class that will come across these terms: The first is negative binomial distribution, dispersion and hypothesis testing.

So, just to remind you about the steps of the RNA-Seq data processing pipeline, we have covered mapping, read mapping, and transcriptome. We have talked about the challenges, then we have talked about assembly that we can do and sometimes do. Then, of course, we have the quantification step, followed by normalization. So, in the last class, we talked about different types of normalization, which can address different technical issues that are present in the read data. Now we are ready for the analysis part, and this is what we are going to discuss today, ok? So, the analysis is a differential gene expression analysis.

This is a very popular analysis that we do, and the goal is to identify genes that show differences in expression between two sets of samples. So, these two sets of samples could be compared between the control and treatment groups, right? So, imagine you are doing this experiment on a control group. You want to test how a treatment affects cell behavior, cell physiology, etcetera. So, you have control cells, and then you have the cells that are treated with a drug or some other type of treatment, and then you look at the change in the

expression of genes. And you want to identify which genes are responding because that might give you some insights about the function of the treatment that you are using.

We can also do this kind of analysis for disease-versus-healthy samples, right? So, you can imagine we have let us say disease tissue, and the most common example that we use is tumor tissue, right? So, we have tumor samples and then we have healthy samples, right? So, we can compare the transcriptome of tumor cells versus normal cells or normal samples, and we can identify genes that show differences in expression in tumor cells. So, this is something that will tell us again about the biology of tumors, right?

So, how the tumor progresses and how all the stages are developed, right? We can also do this by simply studying gene expression patterns across two different conditions. So, you can think about cells that are in a normal, healthy condition, and then you kind of expose them to some sort of stress, ok? Maybe this could be something like oxidative stress or maybe salt stress—those kinds of environmental stresses that are commonly found. And then you can ask, like, which genes are overexpressed when we expose the cells to stress?

So, again, by identifying genes that show differences in expression in the stress environment, you can identify genes that are actually involved in the process of stress response. So, all this analysis, right, and all these examples that I gave you, the goal is to identify this class of genes so that we learn more about the biology of the system, right? So, what is actually happening inside the cell, how genes are responding, which genes are responsible for what kind of task, etcetera, those kinds of goals we have at the end, ok? By the way, this analysis can be done, of course, not just for comparison between two samples or two sets of samples; you can do this for multi-class multiple sets of data, right? So, in the case of disease versus health, you can think of, for example, whether you have a tumor, are healthy, or have let us say other types of samples, right? Other diseases, not just tumors, maybe some other disease, right?

So, that kind of analysis can be done. And similarly for condition analysis, you can have multiple conditions, right? It is not just condition 1 versus condition 2, we can have, let us

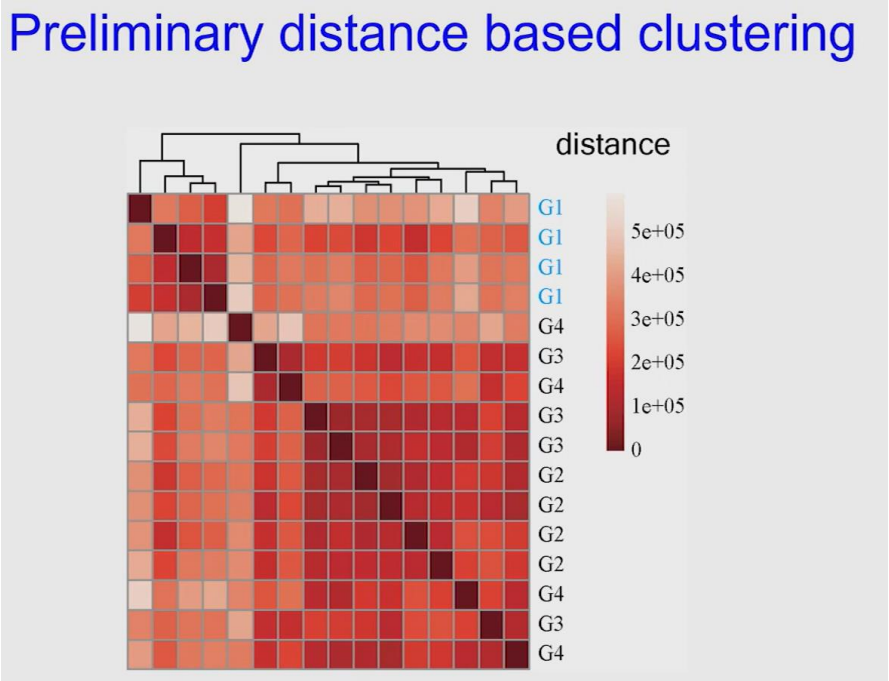
say, a normal condition, then you have, let us say, salt stress, you have oxidative stress, maybe DNA damage all those conditions and you can do all these comparisons as well, ok, alright. So, before we actually proceed to differential gene expression analysis, So, we sometimes prefer to perform something called preliminary analysis, right?

So, this preliminary analysis helps us identify which groups we should compare, ok, and which groups are similar in expression to each other, right? So, we have the normalized count data, which we can use to actually do this preliminary analysis, ok? So, I will talk about two different preliminary analyses that are commonly used, but you can, of course, think of other ways of looking at the data, and there are, of course, many innovative ways people are coming up with for this analysis. So, we can identify the most appropriate groups for comparison, right? So, let us say we identify where these groups are very different and will be most interesting to study, rather than groups that are actually very similar to each other.

And so, one of the first methods is distance- or correlation-based clustering. So, we can cluster the samples based on their normalized count data and see which samples cluster together and which samples are far apart in terms of their expression. So, again, if we want to do differential expression analysis, it would make sense to compare samples that are actually far apart, not the ones that are very similar to each other. And then we have another method, which is called principal component analysis, or PCA. Again, the idea is very similar, right?

So, we want to actually look into the samples that are actually similar to each other and the samples that are different from each other, ok? So, I will just give you some examples with some sample data, and we will show you what this will look like, ok? So, I have been talking about these 16 samples, right? So, we use this data for normalization. We explored different types of normalization with that data. So, I will use the same data to actually show you what this correlation-based clustering or distance-based clustering will look like, and if we do the principal component analysis on that data set, we will get, right?

So, that will again give us some insight into the data, ok, and the biology of the data, ok. So, let us look at the preliminary distance-based clustering, right? So, here what happens is that we calculate something like a distance between the samples, right, based on their normalized count data, and we do something called higher-accurate clustering, and what it achieves is that it will group similar samples in a cluster, right? So, they will occur together in the cluster, ok? So, again, we will see this now with the example, and you will probably better understand what we mean by this, ok?



So, here are the 16 samples, and they are in four groups, ok? So, we have group 1, group 2, group 3, and group 4. You can imagine those as four different conditions, right, and so we want to see, right, which conditions are giving similar transcriptomic profiles and which conditions are giving different transcriptomic profiles because, based on this analysis, we can decide, ok, maybe this is the most interesting comparison for differential expression analysis. So, what you see in this plot here is okay. So, here are the samples, right?

So, along the rows, you have the samples, and along the columns, you have the samples, right? So, this first one, this first matrix here, so this one, this cell actually gives you the value of distance between this G1 replicates, right? So, it is distance from itself, right? So, this will give you like 0 distance, ok? And on the right, you can see the color code here.

So, darker color means lower distance, ok, and lighter color means higher distance, ok. So, along the diagonal, you will see 0 because we are comparing that sample with itself, ok? So, there would not be any distance because there is the same sample. Now, with this one here, the next one is a comparison between these two ones, right? So, again, these are replicates, ok?

So, that is why I am saying 4 groups, and you might ask, Why do you see 4 G1s because these are technical replicates? These technical replicates have been done for four groups. So, there are 4 replicates for each group. So, in total, 16 samples are OK. And this one, so and so on, right?

So, as you can see, this is the replicate, or comparison of the replicate. It is actually not very similar. There is some distance. You can see a bit of a lighter shade. Then you have the comparison of this one with this one, right?

So, you have this one again. So, again, this is a comparison between this G1 and this G1, ok? For this one, this is the distance between these, right? So, you can now imagine, right, you have this matrix kind of form where you are comparing the distance between each pair, ok? So, you have replicates for each group, and then you are comparing the distance between each pair. Now, what you can probably notice if you look at the overall matrix, ok, what you probably will notice is that this G1 replicates that they are actually very close to each other, right?

So, again, this is now clustered, ok? So, which means samples that are close to each other in terms of distance will be clustered together, ok? So, what you see here is that these 4 G1 replicate their cluster together, right? So, they are present in this corner of this matrix, ok? So, which means they are actually very similar in expression to each other and you can probably notice the other cluster if you look carefully that there is another cluster here, ok, on this side.

You see that, right? So, what it means is that the rest of the groups are probably very similar to each other, ok? Again, because they share some transcriptome profiles, right? So, they have shared a transcriptome profile, and the genes are very similar in expression across this whole sample. So, what it means is that probably we have two clusters, one consisting of the G1 samples, right, G1 replicates, and the other group, right, the other cluster consists of all these G2, G3, and G4 samples. Interestingly, you have one G4 that actually stands out, right?

It is not actually close to G1; it is also not close to the other cluster because, as you can see, most of these distances are in a lighter shade, which means the distance is high. So, maybe there are some issues with this experiment with this sample, right, when RNA was isolated or processed. So, this might be an outlet that we might have to discard before we actually go for analysis, alright? So, you probably understand now that it gives us an insight, right? We have two clusters, and maybe it is important or it would be interesting to compare G1 versus the rest, ok? G2 versus G4 differential expression or G2 versus G3 will probably not give you some interesting results because they are probably very similar in terms of expression profile, ok?

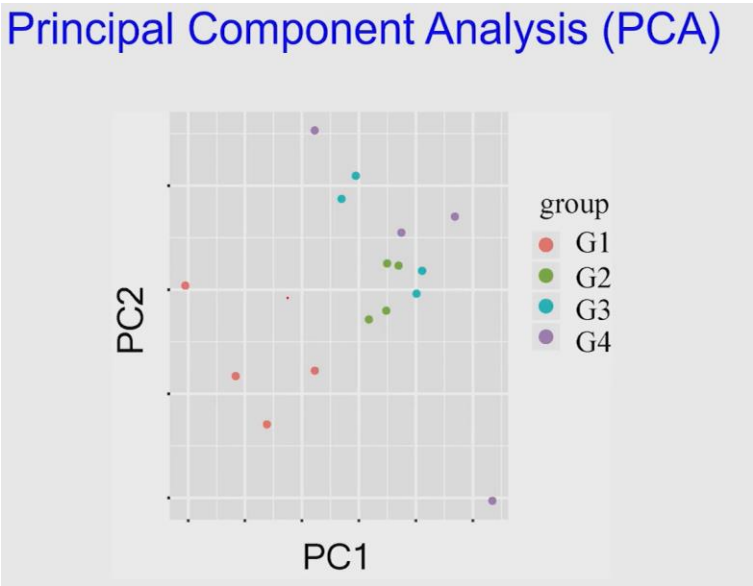
So, G1 versus rest would probably be the most interesting differential expression analysis that we can do, ok? So, this kind of insight we can get when you are using this kind of clustering approach So, coming to the principal component analysis, or PCA, So, this is again a very not very detailed discussion. I will just very briefly mention what we do in PCA, but you can find out a lot more here. There will be a lot more elaborate analysis that you can do with this PCA form. So, this is a dimension reduction technique that tries to preserve the information that is present in the data.

So, when I say dimension reduction, in our normalized count data, we have counts for maybe 10,000 genes, right, 7,000, 10,000, or 20,000 genes, ok? And working with that many dimensions is very, very difficult, ok? And maybe some dimensions are not so important, whereas some dimensions are important, ok? So, what principal component analysis tries to do is try to identify, right, a combination of these variables that kind of

explains the variation that is present in the data. And it will try to reduce these dimensions, right, from 10,000 to maybe 10 or 100 that can explain the variation that is present in the data, ok?

So, it tries to preserve the information as much as possible, ok? So, it derives its principal components. So, these new variables are called principal components, and these are combinations of the original variables, right? So, it can be a combination of 10 genes or 20 genes, right? So, it is looking at it will identify which gives which can explain the variation better, right.

And then, of course, once we have this, right, once we identify these principal components, we can see which samples are actually similar to each other and which samples are different. So, we can apply this principal component analysis to the same samples, right? Again, we have this group 1, group 2, group 3, and group 4, right? And we have four replicates for each group, ok? And again, you can see the colors, right, in red, green, blue, red, green, blue, and purple.



So, you can see these four colors, and you can probably notice them now. So, the x-axis is PC 1, the principal component 1, and the y-axis is the principal component 2. We can have principal component 3, and so on. But usually, principal component 1 and principal component 2 explain the most variance that is present in the data. So, what you probably

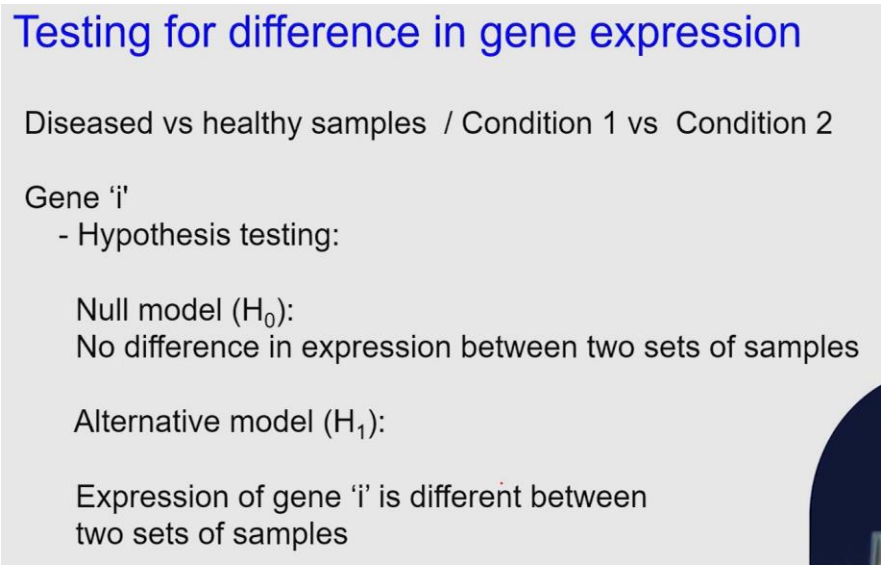
notice from this plot is that group 1 again stays apart from the rest of the sample, right?

This is clear, right? So, this is completely separated from the rest of the samples, whereas group 2, group 3, and group 4 are very close to each other, right? These replicates are very close to each other on the other side. Again, one point you probably notice is that there is one group 4 sample that is kind of far away from all these clusters, right? It does not cluster with group 1 or with group 2, group 3, or group 4 (right). So, it is probably an outlier that we may have to remove before we can actually do any differential expression analysis, right?

So, otherwise, this might affect the results, ok? And this could be an artifact of experimental errors, etcetera, ok? So, now that we have done the preliminary analysis, we have identified which comparison is the most informative, right? What would be the most informative comparison? We see it is probably G1 versus G4 or G4 versus 1, G1, right, or maybe G1 versus the rest of the samples, ok. Now, how do you test for differences in gene expression?

How do we identify these differentially expressed genes? What is the method? And let us take this example: right disease versus healthy, or condition 1 versus condition 2, that kind of will help us understand the process, right? So, we have let us say gene i, right? So, this could be of any value, right? So, again, depending on the organism, this will vary from, let's say, 1 to 10,000 or 1 to 20,000, depending on the organism.

And we do something called hypothesis testing for each gene, ok? So, this is the general approach, but we will then talk about specific tools and we will take one tool and specifically mention what kind of testing we are doing, what kind of statistical test we use, etcetera. So, for hypothesis testing, you might be aware that we need a null model, which we usually denote by H naught, and the null model is that there is no difference in expression between two sets of samples. So, if you are talking about disease versus healthy samples, the null model says, OK, there is no difference in expression. Now, since you have a null model, you also have an alternative model that says the expression of gene i is different between two sets of samples.



So, it is the opposite of the null model, ok? Now, in hypothesis testing, what we will have to do is then do a statistical test among replicates of this normalized count data that we have generated. So, in case we want to do a G 1 versus G 4 comparison, we will do a statistical test between the replicates of G 1 and G 4. So, this replicate number can vary between 3 and 5. Again, it will be better to have more replicates, but these experiments are expensive. Right, doing the RNA-seq is expensive. So, again, we kind of settle for 3 to 5 replicates per sample, right?

So, we cannot have more than that, usually. So, this is something we need to keep in mind because this will come to us later on, ok? So, because the sample number is low, we have to be a bit careful about the statistical test, ok? So, when you do the statistical test, I will tell you what kind of statistical test we can do a bit later, and from the statistical test, we get a level of significance or something like a p-value that will tell, ok, whether there is also significance, and we have a threshold.

This is usually 0.05 and sometimes 0.01. You can be more stringent. You can also set it to 0.

001 or so. Now, the usual threshold is 0.05. So, if p is less than 0.05, then we can reject the null hypothesis. If p is greater than 0.05, then we do not have enough evidence to reject the null hypothesis, right?

So, this is standard hypothesis testing. Now, when we reject the null hypothesis, what does it mean? It means that the gene is differentially expressed, ok? So, the null hypothesis was that there was no difference in the expression of gene i between these two sets of samples. When you reject the null hypothesis, it means that is actually not the case, right? So, we say the gene is differentially expressed between the two samples, right? So, we identify gene i as a differentially expressed gene between the two samples.

Now, one thing we will keep in mind is that since we are doing this hypothesis testing for all genes, which could be 1 to 10,000, we need something called a multiple hypothesis testing correction, ok? So, you will see this in some tools, right? This is already implemented. We will talk about multiple hypothesis testing in later classes, ok?

This is something we need to understand a bit better, ok? So, we do this hypothesis testing for all genes in the sample cell tissues, right? Again, I can vary from 1 to 5000, 1 to 10,000, right? So, we do this hypothesis testing for all genes, and then we can identify genes for which we can reject the null hypothesis, because these are the genes that are differentially expressed between these two sets of samples, ok? So, this can be, let's say, 500 genes or 100 genes, and this list will give us those genes that are differentially expressed between, let's say, disease and healthy conditions, ok? And as I mentioned, the goal is to understand the biological difference, right, between samples or control versus treatment or disease versus healthy.

So, these genes can help us understand the biological differences. Of course, we will talk about, right, how we actually better interpret the data once we have identified this set of genes. So, we will talk about that in later classes, right? There are ways to actually interpret this analysis. So, let us now talk about the statistical test that we do.

So, we can have something like parametric methods. So, I am not sure if you are aware of parametric methods and non-parametric methods. If not, I will just mention it, right? So, in parametric methods, there is an assumption about the underlying data, right? So, here is

the count data that we are working with, and there is an underlying assumption about the count data distribution. So, this is normalized count data that we are working with, and there is an assumption that the count data follows a certain distribution.

And of course, if your count data distribution differs from this assumption, that might lead to a bit of error in the results, right? So, these results will be sensitive to violations of assumptions. So, if your distribution is not following that assumption, let us say you assume, ok, your data follows normal distribution, but the distribution is not actually normal, then of course, the results will be a bit sensitive and it might give you some false positives. On the other hand, you have non-parametric methods, ok?

So, there is no assumption about the count data distribution, right? So, they are kind of not sensitive to these violations in assumptions, etcetera, because they are not assuming any count data distribution. So, they are usually more robust because they can take any form of shape or any form of distribution, and the results will be valid. So, here are some tools that I will just mention that use these parametric methods. Again, this is a list, and there are probably even more tools that you can use. Some of them I have just mentioned, for example, the cuffdiff2 mentioned earlier, then you have baysec, we have DEseq and DEseq2, we have edgeR, sleuth, and EBseq, ok.

So, out of these two, I will just discuss a couple of them in more detail. Then, on the non-parametric side, you have NOIseq and SAMseq, ok? Of course, we will not go into all these tools; we will not discuss them all. I will just mention some of them, ok? So, what parametric methods in general do is assume this distribution, right, for the count data, and you can assume a lot of different types of distribution, such as the Gaussian distribution. So, normal distribution, you can assume, or binomial distribution, Poisson distribution, or negative binomial distribution, right?

So, what researchers have seen is that a negative binomial distribution actually best fits the count data because it is usually overdistributed. So, the dispersion is quite high. So, the mean actually is less than the variance, ok? So, we usually use the negative binomial

distribution. So, all these tools that I have just mentioned use this negative binomial distribution assumption for the count data.

So, out of these tools, I will discuss DEseq2 in a lot of detail, right? We will talk about the steps, etcetera. So, you will understand how this analysis works, and I will also briefly mention edgeR. These are the two most popular tools for differential expression analysis, ok? So, for the rest, I will just give you the references so that you can look up and read.

So, here is the reference for basic. Again, the packages are available, and the tools are available as packages in Bioconductor in R. So, you can download them and install them. So, hands-on, we will talk about R and how to install packages. We will install some of them, for example, DEseq2, and we will analyze our data.

So, similarly for curve tape, you have the tool available. This is the reference, and you can look it up. So, again, for sleuth, the tool is available, and the reference is here. You can see and learn more about it. Similarly, for EBseq, you can have the reference and the link to the software.

For nonparametric methods, you have NOIseq. Again, there is a Bioconductor package for this tool available. You can install it in R, and you can use it. And finally, this is for the SAMseq. This is a tool, again implying a nonparametric approach for this differential expression analysis.

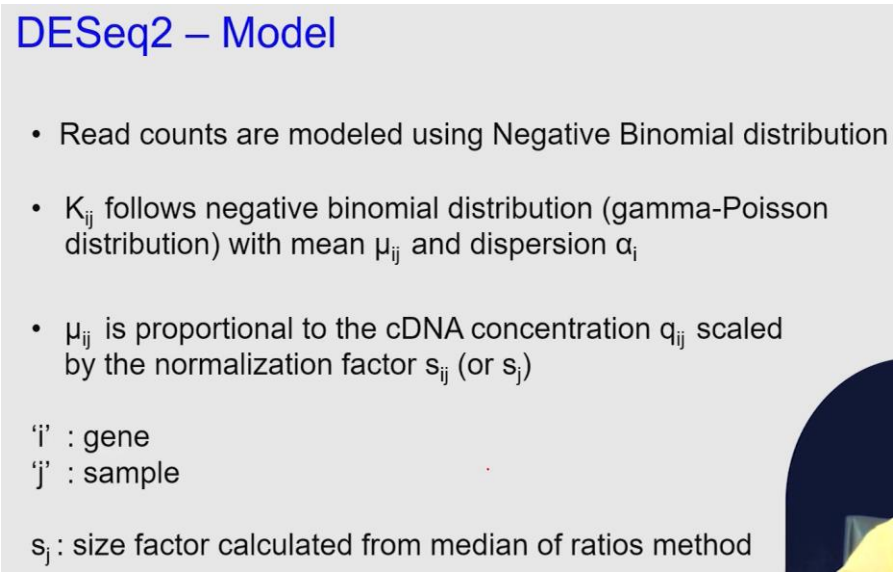
So, let us now go into the details of one of the tools. This is a very popular tool for doing differential expression analysis. It is called DEseq2. So, the DEseq was an earlier version; the current version is DEseq2, ok. So, how does this work? So, here is the reference for this tool, and then you have the R package. We will install this tool in our hands, and we will use it for differential expression analysis. So, how does it work, ok? So, this tool will start with a count matrix, and as we can see, the count will be k i j; this k i is the gene, right?

So, this count is for gene i in sample j, ok? So, let us say in our earlier example, we have let us say 16 samples and we have 7,000 genes, ok. So, we can see that I will run from 1 to 7,000 and J will run from 1 to 16, ok? And this tool starts with raw count data, ok? It will not take normalized count data; it will take raw count data, and it will consider only uniquely mapped reads. So, if you are giving it raw count data, you should give only the uniquely mapped reads, and this curve reads that map to multiple genes or multiple transcripts, and for paired-end data, the mate should map within the gene, ok?

So, if you are working with paired-end data, make sure that both mates map within the gene. If one mate maps outside the gene, you should discard that, ok? This is something that has been mentioned in the paper, and again, why is that the case? The reasoning is given there, ok? So, there is a normalization step. So, DEseq2 performs a normalization from the raw count data, and it uses this median of ratios method, which we discussed in the last class.

And we have seen that this method can correct for differences in sequence depth and, to some extent, for RNA composition bias. Now, let us move on to the model, right? This is the most important part of DESeq2. So, as I mentioned, this is a parametric method. So, the read counts are modeled using a negative binomial distribution, and what we assume then is that K i j follows a negative binomial distribution, sometimes called a gamma proportional distribution, with a mean mu i j and a dispersion alpha i.

So, two parameters are there and then this mu i j is proportional to the cDNA concentration in the library, right? So, which is denoted by q i j, and this is scaled by the normalization factor s i j or s j, right? So, we have talked about this normalization factor, or scaling factor, right? When you talk about the median of ratios method, we get a normalization factor for each sample, right?



So, here we can use that factor, right? For example, s j, but sometimes we want to calculate this normalization factor for each gene separately. So, in that case, this will be s.i.j., right. So, this normalization factor is unique for each gene, and then we can supply that as well, ok? So, just to remind you, i is for gene and j is for sample.

So, sj is a size factor that has been calculated from the median of ratios method. This is the default, right? This is what DESeq2 will do. So, just to put it in the notation, so k i j follows negative binomial with these two parameters s j q i j and alpha i, ok. And using the generalized linear model, it tries to fit this log 2 q i j with this x j r and beta i r, right? So, the summation of this beta is beta, the coefficient values, right, and x is the design matrix element, right.

So, for example, it just tells whether this sample belongs to a disease sample or a healthy sample, right. So, in our example, it will tell us whether this is group 1 or group 2, group 3 or group 4, and beta is the coefficient. So, we will take a very simple example that I think would make it even clearer. So, in simple experiments, this design part, right, this design matrix, right, is usually sample status, and you can say, OK, this is 0 for healthy and 1 for diseased.

DESeq2 – Model

$K_{ij} \sim NB(s_j q_{ij}, \alpha_i)$

Using Generalized Linear Model (GLM):

$\log_2 q_{ij} = \sum_r x_{jr} \beta_{ir}$

x_{jr} : design matrix element (e.g., diseased or healthy)

β_{ir} : coefficients

So, something like that. And you can also accommodate more complex designs, right? Because this is a generalized linear model, you can have much more complex designs and multiple factors, not just health and disease. You can also might add, let us say you have some treatment, right? For example, in some diseases, some patients have had treatments, so you are using that treatment data, and you can also think, OK, maybe the age of the patient is also important for this analysis, right? So, you can add age also, and you can see, right, whether this in which of these factors probably affects the expression, ok?

DESeq2 – Model

- Diseased vs healthy design
- $\log_2 q_{ij} = \beta_{ih} + x_j \beta_{id}$

  x_j=0 for healthy sample
  =1 for diseased sample
- Null model : β_{id}= 0, no difference in expression
- Alternative model: $\beta_{id} \neq 0$ with LFC given by β_{id}

And you can also have interaction terms, right? So, if you are familiar with linear models, right, you can see that you can use correction terms, and this being a linear model can also

do that, ok? So, let us take a very simple design if you are not very comfortable with these complex equations, right? So, we take a simple design and we will understand how this works, ok? So, we have disease versus healthy design, right, and you can simplify this model as in this way, right, log 2 q i j is beta i h plus x j beta i d, ok.

So, we can say x j; this is the design matrix part, right? So, x j is 0 for the healthy sample and x j is equal to 1 for the disease sample, ok? And the null model—right, this is where the statistical model comes in. So, the null model is beta i d equals 0, right? So, there is no difference in expression between the healthy and disease samples, right.

So, this second term here becomes 0, ok? But of course, then you have the alternative model. So, which says beta i d is not equal to 0, and that means there would be something called log fold change, or LFC, that is given by beta i d. So, just note that this is in the log, right? So, this is a log transform. So, beta id will give the log fold change, log2-fold change, or, in short, LFC.

We will discuss what this LFC is a bit later, ok? So, what this generalized linear model does, right, is give us estimates of these coefficients, right, of this beta that are present in the model and that best fit the data, ok? And so, in that process, what you get is the overall expression level for each gene as well as the log2 fold change value for each gene between a diseased and healthy sample, for example. So, you can imagine this, right? A similar concept will apply if you are talking about condition 1, condition 2 analysis, etcetera. So, now, at this model, right, you get this log fold change as well as the mean expression level for each gene, ok?

And there are now the next steps. So, the other step is the dispersion, because the statistical test that you are doing depends on this parameter, alpha i, right? So, we need to estimate this dispersion from the data itself, right? We need to do something called LFC shrinkage, and finally, we will also need to do hypothesis testing. I just mentioned, right, we do some hypothesis testing, right, but what kind of hypothesis testing have we not talked about? Here we have just set up the model; we have estimated the parameters, but we have not

identified the differentially expressed genes, which we will do in the next class.

So, here are the references for this class. To conclude, as we have seen, differential gene expression analysis allows us to identify genes that change expression patterns across different conditions or different samples. Here we took examples of disease versus LD. And we have also talked about some of the preliminary analysis that can help us direct towards more meaningful analysis. So, we can choose which groups we should compare when you are doing this differential expression analysis.

And we have also talked about parametric and non-parametric-based methods for model testing. We have seen that the negative binomial distribution best fits the data, and we have started our discussion on differential expression analysis with DESeq2. So, there are some steps remaining, which we will be discussing in the next class. Thank you.