Next Generation Sequencing Technologies: Data Analysis and Applications

DGE analysis result and visualization

Dr. Riddhiman Dhar, Department of Biotechnology

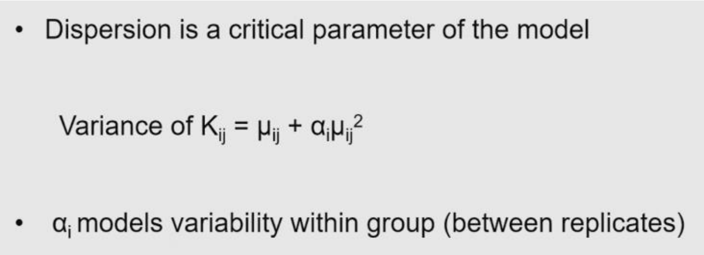
Indian Institute of Technology Kharagpur

Good day, everyone. Welcome to the course on Next Generation Sequencing Technologies, Data Analysis, and Applications. In the last class, we started discussing differential gene expression analysis, or DGA analysis in short. We have talked about the preliminary concepts, and we have also started a discussion on DEseq2. So, we have completed some parts about the steps that we need to do for DEseq2 analysis. So, we will continue that discussion in this class.

We will look at the results that are generated by this DEseq2 tool, and we will also visualize the results. So, once we generate the results, this is a huge result, and we need to use some visualizations to actually better understand what we have. So, let's set the agenda for today's class. We will look at the DGE analysis results.

We will generate those results, and we will look at some of the visualizations and some of the plots that are generated for visualizing the results. So, just to briefly summarize what we discussed in the last class regarding DEseq 2, So, DEseq2 starts with a raw count data matrix. It utilizes the normalization method with the median of ratios method, which can take care of some of the technical variations in the data and then it uses a model that is based on a negative binomial distribution. So, this is a parametric method.

It has some assumptions about the count data distribution, and it models them using this negative binomial distribution, and then it fits a generalized linear model, right? And it takes that approach to estimate the parameters, the coefficients that are fit in the model, ok? So, those coefficients give us the mean expression values of the genes as well as the differential expression, right? So, whether some genes are showing a difference in expression, So, in terms of differential expression, we get the log fold change value, ok, so or LFC, alright. So, again, I will remind you of the terms, right?



- Dispersion is a critical parameter of the model

Variance of $K_{ij} = \mu_{ij} + \alpha_i\mu_{ij}^2$

- α_i models variability within group (between replicates)

So, we start with a count matrix with count k i j for gene i in sample j, and DEseq2 assumes that k i j follows a negative binomial distribution with mean j and dispersion alpha i. The steps that are remaining, right? The first step is that we want to estimate the dispersion alpha i. So, once we have estimated the parameter beta from the model, we need to identify whether this beta is significantly different from 0. So, that will tell us whether a gene is differentially expressed or not. So, to do any statistical test, we also need to get an estimate for this dispersion, alpha i.

Now, this is quite challenging, right? We have to use the data itself to estimate this dispersion, ok? So, what is this dispersion? So, this is actually a very critical parameter of the model. As you now understand, for doing the statistical test, we have to use this parameter because we are assuming a negative binomial distribution, and this dispersion alpha i is a parameter of the distribution. This will determine the shape of the distribution. So, alpha is actually related to the variance of k i j using this term, right?

So, the variance of k i j is given by mu i j plus alpha i mu i j square, and this models the variability within the group and between replicates, ok? So, we can estimate this alpha i, right, from the replicates that are present in our data, but as I mentioned in the last class, the number of replicates is very small, right? So, it is usually 3 to 5, ok? So, it is a very small number of replicates because we cannot do these experiments with a large number of replicates. It is difficult; they are expensive, right?

So, we would have to settle for only 3 to 5 replicates, which means the estimates of this dispersion, or alpha i, could be noisy because we are estimating from only a small number of replicates, right? Because maybe one replicate is slightly like showing slightly different patterning expressions, and then, of course, the alpha i would be noisy, ok? So, how do we address that? How do we address that issue? So, it is actually addressed by something called dispersion shrinkage, ok? So, we want to reduce this noise in the dispersion data, and tie-daze is very simple, but there is an assumption. The assumption is that genes with similar mean expression should have a similar dispersion value, ok? So, this is an assumption that we make; this model DEseq2 makes, right, and based on that, it applies an empirical-based framework to shrink dispersion towards this expected dispersion value, right.

So, if you look at the original reference, you will see, right, that once we calculate this dispersion value for each gene, for some genes, the dispersion will be much higher than the group of genes that show similar mean expression. So, there is a relationship between dispersion and mean expression, ok? So, what this model does, right? It assumes, ok, genes that have similar mean expression also should have similar dispersion values, ok, because we are talking about within replicate measurements. So, there is no reason why the dispersion value should be different, ok? And then it uses this framework to actually shrink this dispersion towards this expected dispersion value, and this shrinkage is actually dependent on the number of replicates.

So, if there are a higher number of replicates, it means there will be less shrinkage, right? So, because if you have more and more replicates, you are more confident about the dispersion estimate, ok, but if you have a very small number of them, then maybe you have only three, and then in that case, you may not be that confident about this dispersion estimate. So, this means you need to have, you need to probably shrink the dispersion mode, ok? So, this is kind of the assumption behind this dispersion shrinkage. Now, once you have estimated this dispersion, right, we also calculate the low fold change, right, and there is another term that we also use that is called low fold change shrinkage, ok.

But what does it actually mean? This is also somewhat related. So, what we see in RNA-Seq data or also in microarray data is that genes that have a low read count or low mean expression actually show a lot of variability in the data. The technical replicates show a lot of variability, right? This is kind of expected, right? If you have a very small number of reads to choose from or a small number of fragments to choose from, there is a lot of variation.

So, if you are doing the same experiment, you can imagine that this is a pool of molecules that you have, and then you have for these genes only a very small number of molecules. Now, from this pool for each technical replicate, you can choose samples, right? So, you are sampling from that pool, and the genes that are underrepresented, right, or should have or have low expression are likely to have a higher variation between these replicates. So, when you are pulling when you are sampling this, right, for these genes, in some replicates you might get 10 reads, in some replicates you might get 5 reads, and in other replicates, you might get 2 reads, right? So, this is what we see: higher variability in genes with a low read count, ok?

And to account for that again, we want to shrink this lock-fold change value, and again, DEseq2 uses an empirical-based framework. And of course, you can use different frameworks to do this shrinkage, and this actually gives more robust results, as the researchers have seen. So, in the next part, we have estimated the dispersion, right? We have fit the model, and we have the parameter values estimated using a generalized linear model, right? So, we know the beta values; now it is time for hypothesis testing, right?

So, in the hypothesis testing in the simple example that I gave you in the last class, we want to have a null hypothesis, which says beta values are 0, which means there is no difference between disease and a healthy condition, ok? So, by hypothesis testing, we actually have this statistical basis for that, right? By saying whether we can reject that hypothesis or not, right? So, this is what DESeq2 does, right? So, precisely, we are testing whether coefficients are significantly different from 0, and coefficients that are significantly different from 0 will be identified as differentially expressed genes. So, to do that, DEseq2 does what DEseq2 does: actually, we will divide these log fold changes by the standard error, and it will test against a standard normal distribution using the Wald test.

So, you will get these Wald test statistics and the p values after the hypothesis testing for each gene, ok? And these p values will be their significance that will signify, right, whether we can reject the null hypothesis or not, ok? And I have also mentioned this because we are doing this for a large number of genes, right? So, sometimes with 10,000 or 20,000 genes, we have to do something called multiple hypothesis testing, because we are doing this hypothesis testing for each gene. So, individually, we are doing 10,000 hypothesis tests, or if you have 10,000 genes, 20,000 hypothesis tests.

So, we need to do something called multiple hypothesis testing, ok? We will talk about this multiple-hypothesis testing in the next class, right? So, this is something very important for getting accurate results, ok? So, I will very briefly mention the other tool, which is the edgeR tool. Here is the reference, and this is also a Bioconductor package. Again, this is a very popular tool for differential gene expression analysis.

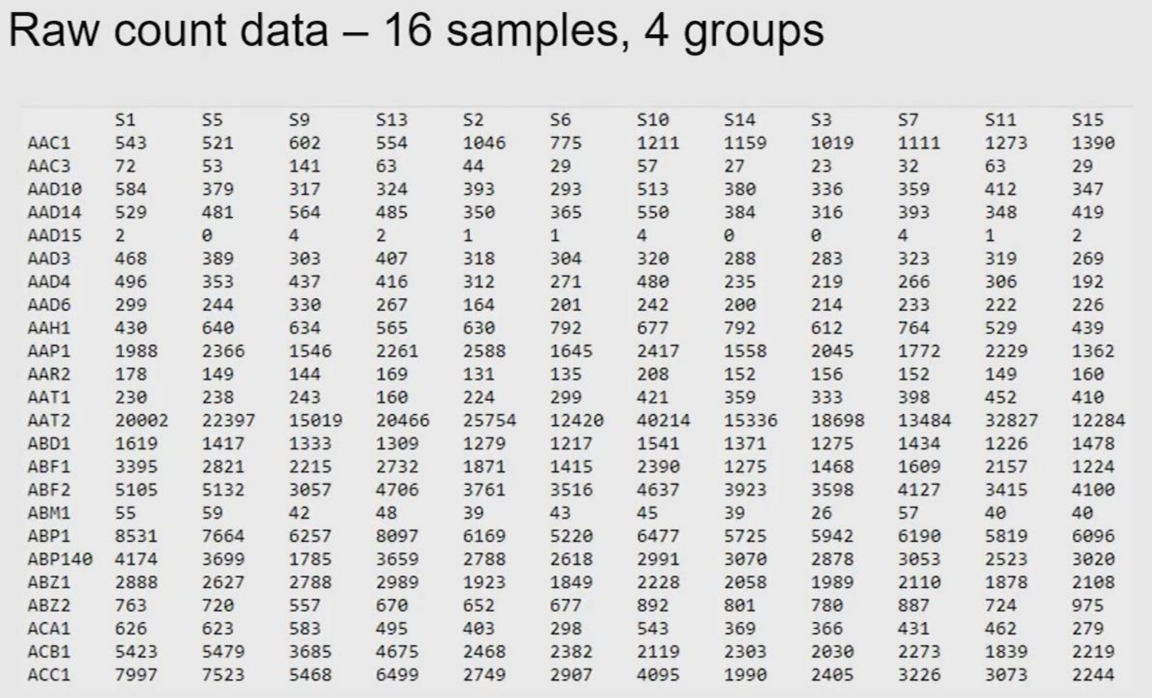
So, edgeR actually uses very similar concepts and similar steps, but there are some differences, ok? So, edgeR also starts with rock-on data, and it does the TMM normalization to generate the normalized count data. We have talked about TMM normalization, and again, we have seen that this method can account for some of the RNA composition bias that might be present. So, this is the normalization that is used by edgeR. So, edgeR is also a parametric model, right?

So, it uses the negative binomial distribution. So, it counts our model using this negative binomial distribution, and it also uses a conditional maximum likelihood approach to estimate gene-wise dispersion values. So, you see the similarity with the other tool DEseq2, right? But of course, there are some differences. If you go into the reference and see the details, you will see the dispersion methods are not exactly the same, right? So, this estimation process and the shrinkage of dispersion again use empirical-based methods, but there is some difference.

And one of the unique points about edgeR is that it can sometimes separate out this technical variation from biological variation. So, it can estimate the biological variation while calculating dispersion. Again, there is some underlying assumption. So, it assumes the count data within technical replicates follows a Poisson distribution, whereas between biological replicates you have a negative binomial distribution. So, with that assumption, it can actually calculate the biological variation while calculating dispersion.

And for statistical testing or hypothesis testing, right, we have seen that for DEseq2, it uses the wild test, but edgeR uses Fisher's exact test for differential expression. So, to identify this differential expression, it will use something called the Fisher's exact test. So, what is the advantage of Fisher's exact test? Because it can work with very small numbers, it is an exact test. So, it can work with a very small number of data points. So, even if you are working with only two replicates or three replicates, it would give you more robust results because this is an exact test that is used.

So, we have now talked about these two major tools—the two most popular tools for doing differential expression analysis. What I am going to do now is show you some results from the analysis using DEseq2. So, again, I will take the results from the 16 samples that I am using for all these illustrations. So, we have raw count data for 16 samples, divided into four groups. So, each group has four replicate measurements, right?



Raw count data – 16 samples, 4 groups

	S1	S5	S9	S13	S2	S6	S10	S14	S3	S7	S11	S15
AAC1	543	521	602	554	1046	775	1211	1159	1019	1111	1273	1390
AAC3	72	53	141	63	44	29	57	27	23	32	63	29
AAD10	584	379	317	324	393	293	513	380	336	359	412	347
AAD14	529	481	564	485	350	365	550	384	316	393	348	419
AAD15	2	0	4	2	1	1	4	0	0	4	1	2
AAD3	468	389	303	407	318	304	320	288	283	323	319	269
AAD4	496	353	437	416	312	271	480	235	219	266	306	192
AAD6	299	244	330	267	164	201	242	200	214	233	222	226
AAH1	430	640	634	565	630	792	677	792	612	764	529	439
AAP1	1988	2366	1546	2261	2588	1645	2417	1558	2045	1772	2229	1362
AAR2	178	149	144	169	131	135	208	152	156	152	149	160
AAT1	230	238	243	160	224	299	421	359	333	398	452	410
AAT2	20002	22397	15019	20466	25754	12420	40214	15336	18698	13484	32827	12284
ABD1	1619	1417	1333	1309	1279	1217	1541	1371	1275	1434	1226	1478
ABF1	3395	2821	2215	2732	1871	1415	2390	1275	1468	1609	2157	1224
ABF2	5105	5132	3057	4706	3761	3516	4637	3923	3598	4127	3415	4100
ABM1	55	59	42	48	39	43	45	39	26	57	40	40
ABP1	8531	7664	6257	8097	6169	5220	6477	5725	5942	6190	5819	6096
ABP140	4174	3699	1785	3659	2788	2618	2991	3070	2878	3053	2523	3020
ABZ1	2888	2627	2788	2989	1923	1849	2228	2058	1989	2110	1878	2108
ABZ2	763	720	557	670	652	677	892	801	780	887	724	975
ACA1	626	623	583	495	403	298	543	369	366	431	462	279
ACB1	5423	5479	3685	4675	2468	2382	2119	2303	2030	2273	1839	2219
ACC1	7997	7523	5468	6499	2749	2907	4095	1990	2405	3226	3073	2244

And this is how the data looks; this is the raw count data, right? So, what you can see here, right on the left column, right in the leftmost column, is that you have the gene names, right? So, these are the genes, right? So, these are the genes here, and here are the samples, right? So, you are starting from S 1, S 5, S 9, and up until there will be a total of 16.

Of course, I cannot show you the whole thing—all the columns here in this small space—but you have 16 samples, ok? And for each sample, these are the count data, right? So, you can see that gene AAC1 in sample 1 has 543 reads mapping to this chain, ok? So, these are ordered in such a way that these replicates are side by side. So, S 1, S 5, S 9, and S 13 are replicates of each other, ok?

So, this is how these are organized, ok? So, again, we have a list of 7000 genes, 6000, 7000 genes for which you have this count data, ok? And as you can probably see, the count data varies a lot, right? So, for some genes, you have very high-count data, and for some genes you have very low count data, right? So, you have this example here. You can see that this gene, AD 15, shows very low count data across all samples. Whereas, you have this gene here, A T 2, and you can see very high expression and a very high count, meaning this gene is probably more highly expressed.

Of course, this is raw count data, so we do have to be careful in connecting these two expression levels because we have to do some normalization before we can do that. So, I will show you, right? So, as I mentioned, segment 2 uses this median of ratios method, right? So, it actually calculates something called the scaling factor or normalization factor and this package is also called the size factors, right? So, in this Seg. 2 package, you see it will be called the size factor.

This is nothing but the normalization factors or scaling factors calculated in the median of ratios method, ok? And here are the steps: So, if you run D seg 2 on this sample, here are the five steps—the right six steps, actually. So, the first step is estimating the size factor. So, as you will see when we do the hands-on, it will actually give you all these steps one after another.

So, the first is estimating size factors. This is where the normalization is taking place, using the median of ratios method. In the next step, it is actually estimating dispersions. You understand now what dispersions are and how they are calculated. For each gene, you get this dispersion estimate. So, you have this gene-wise dispersion estimate, right?

So, you are getting these inverse values, and then the next step is, of course, the dispersion shrinkage, right? So, as we mentioned, you need to do the shrinkage on dispersion. To do that, it needs this information on the mean dispersion relationship, right? So, you need to generate something like a mean expression or mean count versus dispersion, and you will probably see something like this.

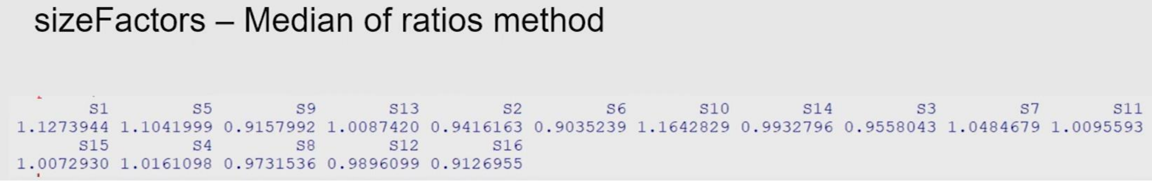
So, you will see these points, right? If you see the origin paper, you will see these points around, and this kind of relationship you will see, ok? And for some genes, right where you have a higher dispersion compared to these genes of similar mean expression, So, that is the assumption in this model, right? The genes with similar mean expression should show similar dispersion values, right?

So, imagine one gene here that has higher dispersion. So, for that gene, we need to perform the shrinkage, right? So, of course, the method will perform that shrinkage again depending on the number of replicas that are present in the data, right? So, it will bring it closer to the expected value, of course—not exactly to that expected value, but closer to that expected value. So, by doing that, it will generate these final dispersion estimates, right? So, this is what it means by final dispersion estimates because it is doing this dispersion shrinkage, and then the final step is the fitting model and testing, right?

So, the fitting model is this generalized linear model, and the testing part is hypothesis testing. So, this generalized linear model will give the estimates for these coefficients' beta values, and then you have hypothesis testing to identify genes that are differentially expressed. So, it will simply generate the test statistics, the wild test statistic, and the p value, and of course, it will also do a p value adjustment for multiple hypothesis testing. So, let us now look at the results, right? So, how do these results actually look and what you get?

So, before you actually see the results, these are the actual size factors that were estimated by this tool preset 2 on this data set, ok? So, here you see a number for each sample. So, for S 1, this number is 1.

127, and so on. For S 5, this is 1.10, right? So, these are the size factors that have been determined. So, we have to use this in the model, right? So, this will be used by set 2 in the model, right? You remember this count modeling, which is proportional to the fraction of molecules in the cDNA library and multiplied by the size factor, right?



sizeFactors – Median of ratios method

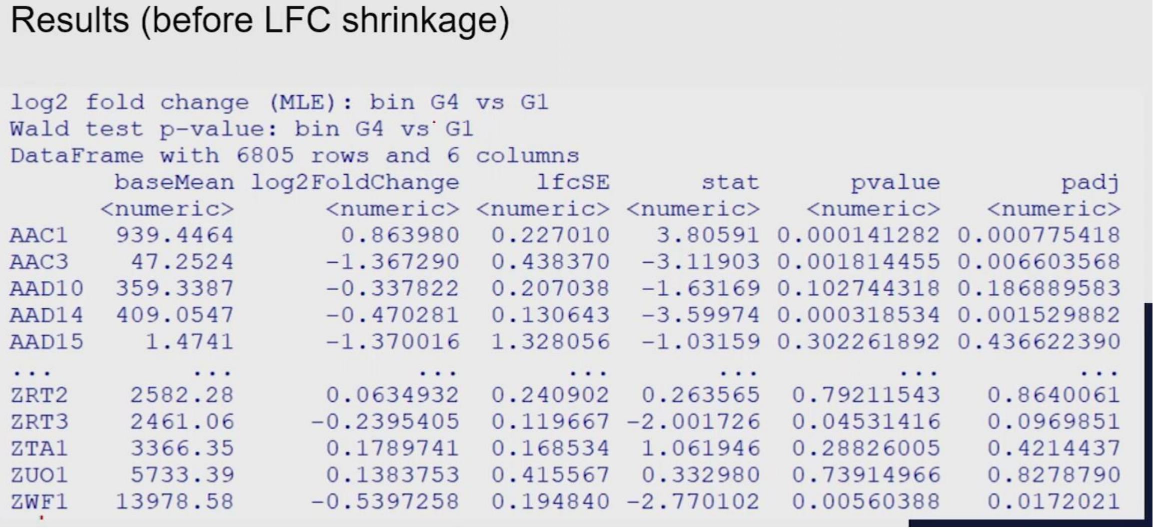
```
       S1        S5        S9       S13        S2        S6       S10       S14        S3        S7       S11 
1.1273944 1.1041999 0.9157992 1.0087420 0.9416163 0.9035239 1.1642829 0.9932796 0.9558043 1.0484679 1.0095593 
      S15        S4        S8       S12       S16 
1.0072930 1.0161098 0.9731536 0.9896099 0.9126955 
```

So, SJ, ok. So, these are the size factors for these samples. Now, in the results part, what we will get, we will see in a moment, we will get this mean expression of genes, we will get something called log2-fold change which is actually looking at the differential expression, right? How much is the difference in expression of a gene between two sets of samples? So, it is whether it is disease versus healthy or condition 1 versus condition 2.

It will also give us the standard error of the LFC value, ok? So, how much error is there in the estimation? It will give us the results of the statistical analysis. So, because this set 2 performs a world test, it will get a world test statistic, you will get a p value, and you will also get something called an adjusted p value. This is a p value corrected for multiple hypothesis testing. So, you will get all these values for every gene, ok? So, if you have 7000 genes, you will get these values for 7000 genes, ok?

And from that, you can then identify which genes are showing significant differences, right? So, there is a significant difference in expression, and you also get the log2-fold change, ok? So, here is the actual data, right? So, summary data that we will get once you run these sets 2, ok? So, it says log2-fold change, and we have done this as a G 4 versus G 1 analysis, right? Because this makes sense.

P 1 and G 4 are quite different, and it says there are about 6,805 rows, which means 6,805 genes, and 6 columns here in the results. Okay, because we have this base mean. So, this is the average expression of the gene. You have log2-fold change, you have the standard error of log2-fold change, you have a statistic that is the world test statistic, you have the p-value, and you have the adjusted p-value, or p edge, right? So, you get these 6 columns in the results, and then you have the first column, which is for genes, and for every gene, you can see these values, right? So, here is the mean expression for this gene, AAC1. You have a log2 fold change, right?



Results (before LFC shrinkage)

```
log2 fold change (MLE): bin G4 vs G1
Wald test p-value: bin G4 vs G1
DataFrame with 6805 rows and 6 columns
       baseMean log2FoldChange     lfcSE      stat      pvalue        padj
      <numeric>      <numeric> <numeric> <numeric>   <numeric>   <numeric>
AAC1   939.4464       0.863980  0.227010   3.80591 0.000141282 0.000775418
AAC3    47.2524      -1.367290  0.438370  -3.11903 0.001814455 0.006603568
AAD10  359.3387      -0.337822  0.207038  -1.63169 0.102744318 0.186889583
AAD14  409.0547      -0.470281  0.130643  -3.59974 0.000318534 0.001529882
AAD15    1.4741      -1.370016  1.328056  -1.03159 0.302261892 0.436622390
...         ...            ...       ...       ...         ...         ...
ZRT2    2582.28      0.0634932  0.240902  0.263565  0.79211543   0.8640061
ZRT3    2461.06     -0.2395405  0.119667 -2.001726  0.04531416   0.0969851
ZTA1    3366.35      0.1789741  0.168534  1.061946  0.28826005   0.4214437
ZUO1    5733.39      0.1383753  0.415567  0.332980  0.73914966   0.8278790
ZWF1   13978.58     -0.5397258  0.194840 -2.770102  0.00560388   0.0172021
```

You have the standard error, statistic, p value, etcetera. Now, you can see that this p value is quite significant, it seems, right? So, 0.0001, and you also get the p adjusted, ok? And once we understand how p-adjusted data is generated, we can probably better interpret the data. One of the things you probably notice is this log2-fold change, which can be positive or negative.

So, a positive log2-fold change means this gene is overexpressed in the G4 sample, ok? So, because this is a G 4 versus G 1 comparison, G 1 is used as the reference. So, this positive number means this gene, AAC1, is overexpressed in the G4 sample, ok? Whether this is significantly different, right, or significantly overexpressed that will come from the p-value, and we have to interpret the p adjusted later on.

A negative number, right, minus 1.36 for AAC3, would mean that this gene is poorly expressed. So, it shows lower expression in the G4 sample compared to G1, ok? And how is the difference in log fold change? It is about 2 to the power of 0.86 in the case of A C 1, and in the case of A C 3, it is 2 to the power of minus 1.

36, ok, because we have to do that because it is a log 2, right, log base 2. So, we have to take the 2 power, 2 to the power, ok? So, we have these results for all genes and this is before LFC shrinkage. So, I have talked about this LFC shrinkage, right? Why why we need to do that? And these are the results after LFC shrinkage, ok? So, here the statistic part is gone because we usually do not use that statistic much, right? We are usually more interested in these p values, p values, p adjusted, and the log fold change, right? So, you can see these numbers have changed slightly because of this LFC shrinkage.

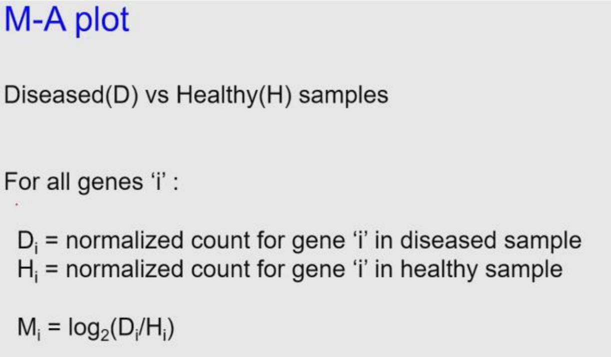
For example, for AAC1, it was 0.86 or something before this LFC shrinkage; here it is 0.78, ok. So, these numbers will change a little bit again depending on the number of replicates that are present in the data and its mean expression, ok? So, we have now seen the results, and we know what to expect after we have done the analysis and how they will appear. Now, we need to visualize, of course, right? Because we have 7000 genes, we cannot go to every individual gene and see its results, right?

So, we want to plot the results in such a way that we can interpret or identify the number of genes in the data set, ok? Of course, there are different ways to do this in the R package. In DEseq2, we can do that; in R, we can do that, but we also want to look at the results, right, and with some plots, because visually, we can then identify certain important things. And so, that is what we are going to do now, ok?

So, there are three different plots that we use very commonly. So, one is called the MA plot, the second is the volcano plot, and the third is the heat map. So, I will discuss these different types of plots, and we will actually plot the results that we have just generated using these plots, and you will see they will give us quite informative inferences, ok? So, the first one is the MA plot. So, what we look at here is the mean expression versus log2 fold change, ok?

So, how do you calculate this? In a moment, I will talk about this. So, we will calculate two parameters, M and A. So, what happens is that when you plot this mean expression versus log2-fold change, you can actually see the number of differentially expressed genes, ok, and their distribution of log2-fold change, ok. And also, the MA plot helps in proper normalization, right? So, we can check whether we have done proper normalization, and this was actually very useful for microarray data analysis, but here also in un-anesthetic data analysis, we can say, ok, whether the normalization has been done properly or whether there is a need for another round of normalization or a different normalization.

So, into the technical details, how do we actually generate this plot? So, imagine this scenario again diseased versus healthy samples, ok, and for all genes I, we can say DI is the normalized count for gene I in the diseased sample and HI is the normalized count for gene I in the healthy sample, ok. And we calculate these two parameters, M and A, for each gene, ok? So, MI is log 2 Di divided by Hi, and AI is log 2 Di plus log 2 Hi divided by 2. So, A is kind of like the average expression, ok, and Mi is the log fold change, right?



M-A plot

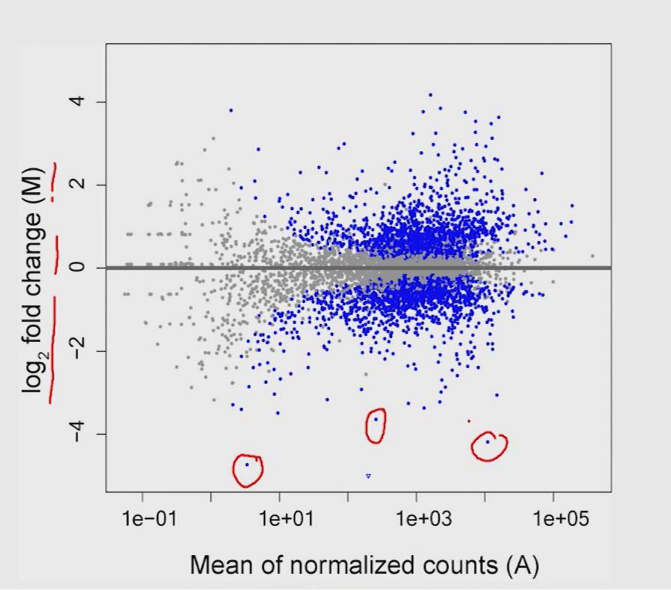
Diseased(D) vs Healthy(H) samples

For all genes ‘i’ :

D_i = normalized count for gene ‘i’ in diseased sample
H_i = normalized count for gene ‘i’ in healthy sample

$M_i = \log_2(D_i/H_i)$

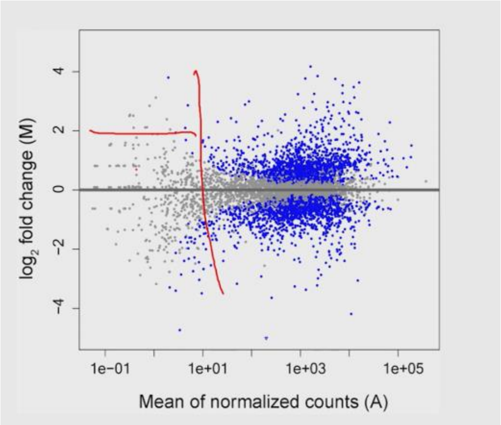
We are looking at differences in expression, ok? So, here is the plot—actually, this is how it would look. So, on the x axis, you have the mean of normalized counts, or the A values, right? This is the average value, and then along the y axis you have the log 2 fold change, or M, ok. And as you can see, the log2 fold change can be positive or negative and we have actually discussed this scenario, right? So, a positive change would mean the gene is overexpressed in the G4 sample; a negative change would mean this gene is under-expressed or shows lower expression in the G4 sample compared to the G1 sample.

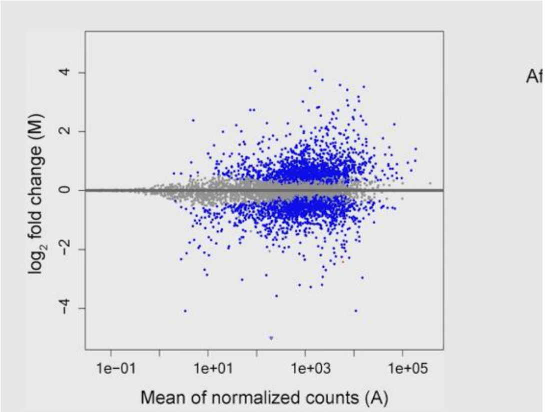


So, now what you see are these dots. So, each dot is a gene result. So, we calculate M and A for each gene, and then we plot this, right? And every point that you see is a gene, ok? And you can see these two different colors, right? So, you have points in two different colors; some points are in gray and some points are in blue, ok? So, what is the difference? So, the gray points are not statistically significant, ok?

And blue points are statistically significant, right? They show a statistically significant difference in expression, ok? So, this means these genes that are in blue and in this top part are the genes that are overexpressed or show higher expression in the G4 sample compared to G1. And this gene blue in blue here, right, shows a negative log2 fold change, and in blue, they are showing lower expression in the G4 sample compared to G1, ok? So, you can kind of get an idea, ok, these are the genes these are the number of genes that are showing this kind of thing, ok.

And you can also see the distribution of the log2 fold change. So, you can see if there are quite a few genes that are above these two values in log2 fold change, right? So, that means they are actually showing four-fold overexpression, right, 2 to the power of 2. So, that is a four-fold overexpression in the G4 sample compared to G1. And similarly, you can interpret the data for the genes that show very low log2 fold change or negative log2 fold change, right?





And you can understand they are showing significantly lower expression in the G4 sample, ok? So, this is before LFC shrinkage, right? So, what happens after LFC shrinkage is that you see, right, the genes that showed very low mean expression or mean count are they kind of disappear from this here, right? So, one of the points is probably obvious, right? So, we see most of these blue genes only in the high mean expression. For this gene that shows low mean expression, right, even though you see the log2-fold change is high here, right, they are not statistically significant, ok.

This is because there is a lot of variation in the data. And because of this variation, when you do the hypothesis testing, it cannot identify that the p value is not significant. It cannot be identified as a statistically significant difference, ok, because the statistical test depends on the variance of the data, right? So, for genes that have low counts or show low expression, they show very high variability, which is why they are not identified as significantly different. Even though the average log2-fold change you can see is quite high, right, they are statistically not significant. So, what LFC shrinkage does is kind of shrink these LFC values, because they are highly variable.

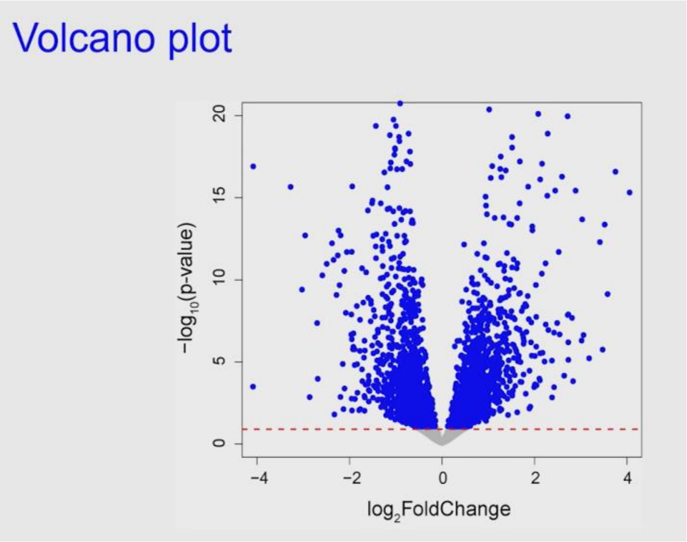
It actually uses this empirical base method to shrink these values. And what you see is that these values have been shrinking now, and you can see they are kind of close to 0, right? And this kind of makes it actually very robust for this method, because we are kind of taking this variable into account depending on the number of replicates, ok? So, I also mentioned that we can check the quality of normalization from the M-F plot, right? So, this is something we researchers have been doing for microarray data as well, and this can also be extended to anistic data. So, one of the assumptions in this kind of analysis is that most of the genes do not show any difference in expression across samples.

And this means that for most genes, their log2 fold change value should be 0, right? So, they should be lying on that 0 line, ok? And if you see a dependence of log2 fold change on the mean expression, that would mean it probably requires better normalization, or maybe you need to use some other normalization on top of the current method, ok? So, what I will mention right here is that what you see is that most of the genes are lying on these 0 lines. So, this seems pretty good, but sometimes what you might see is that if you have the 0 line, the samples are lying like this, ok?

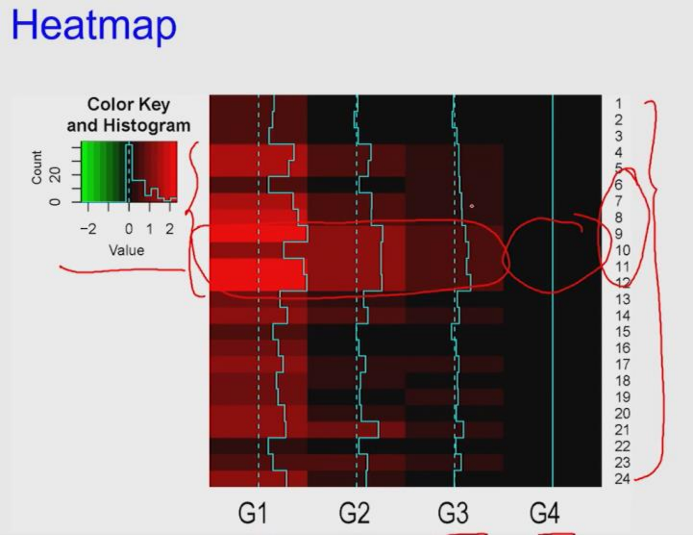
Most of the samples are below 0, or you see some sort of curve nature in this plot, ok? Instead of this kind of nature centered around 0, you see, maybe this plot is slightly curved, ok? Now, if this is the case, that would mean, right, this is the 0 line, right? If this is the case, this would mean you need better normalization methods, ok? And by inspecting this M-A plot, you can probably say, OK, the data is properly normalized, or whether you need some sort of normalization or different normalization methods on this data.

All right. So, the other plots are there. So, another is the volcano plot, and this actually looks at log2-fold change versus log 10 p value. It kind of gives you information about how log2-fold change values vary with significance level, right? So, this is what we get. So, the x axis is log2-fold change, right, the LFC, and the y axis is the minus log 10 p value, right.

So, a higher minus log 10 p value means more significance, right? Because you have the minus. So, if you had let us say 10 to the power minus 2 p value, your minus log 10 p value would be 2, right? If you have 10 to the power of 10, your minus log 10 p value would be 10, right? So, a higher value of minus log 10 p means more significant, right? And as you can see, we have set this threshold here, and based on this threshold, these genes are colored; these are significant; these are in blue. The genes that are not showing a significant difference, right? There is no statistical significance, right? They are shown in gray here and in this part.

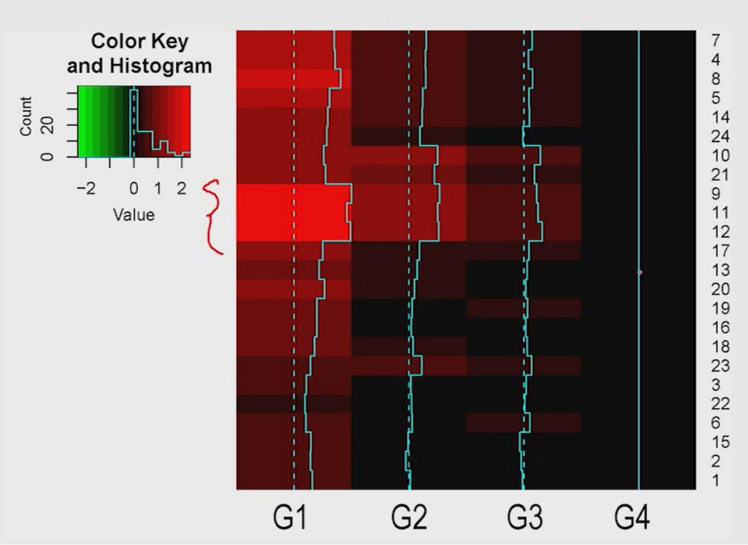


You can also see, for example, what fraction of genes are showing a higher log-fold change than 1 or genes that are showing a lower log-fold change than minus 2 or minus 1, right? All those things can be visualized using this Falcon plot. And finally, the heat map, right? So, we use a heat map for comparison of the expression levels of a set of genes across many samples and not just two samples, and here is a heat map for the same thing, right? So, you have g 1 g 2 g 3 g 4. So, we are looking at, let us say, one replicate of each, and we are looking at the different genes, which are numbered here along the rows, right.



So, you can see genes 1 and 2, etcetera. So, here you are seeing these expression levels; these are color-coded, and the colors are given here, ok? So, if you have a red color, this means this gene is overexpressed; ok, it shows higher expression or positive log fold change; ok. So, you can probably see this very clearly here, right? For example, this set of genes is more highly expressed in the group 1 samples, right, compared to group 4 or group 3.

In addition, what you can probably see is that there is kind of a gradient, right? So, g 1 has the highest expression for these genes, then you have g 2; they show some expression, g 3 is slightly lower, and g 4 shows almost no expression, right? So, this kind of pattern can be identified with this heat map analysis. You can also do hierarchical clustering on top of this. So, this heat map is without clustering; the genes are ordered, right, from gene 1 to gene 24, but you can do hierarchical clustering.



So, it will cluster genes that show similar expression patterns across the data, right? So, you can see it has clustered these groups of genes because they show a similar pattern across these groups (1, 2, 3 4, ok. So, this set of genes has the highest expression in group 1, then slightly lower expression in group 2, then lower expression in group 3, and lowest in group 4. So, we have talked about this difference visualization that helps us understand the results better, ok? Here are the references for this class.

So, we have completed the differential expression analysis with the package DESec 2. At least I have discussed the theoretical steps; of course, we will do it hands-on. And we have seen the results that we get and two values. Two results are very important for us. One is a log2 fold change, and the other is p-adjusted. We have visualized the results through the MA plot, volcano plot, and heat map. Again, we have seen some important insights that we get from this visualization. Thank you.