Introduction to DGE

View on GitHub

Approximate time: 60 minutes

Learning Objectives

- Understanding the different steps in a differential expression analysis in the context of DESeq2
- Exploring the importance of dispersion during differential expression analysis, and using the plots of the dispersion values to explore assumptions of the NB model

DESeq2 differential gene expression analysis workflow

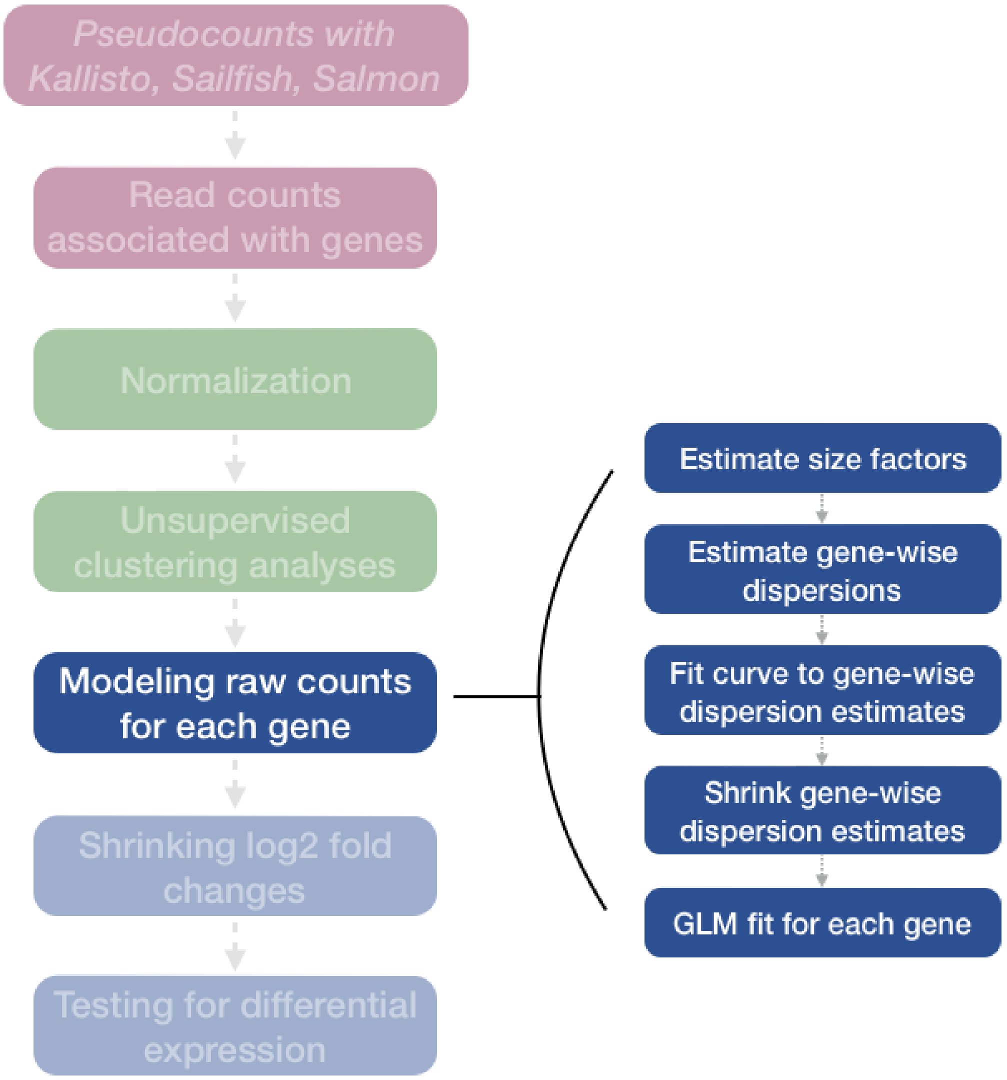
Previously, we created the DESeq2 object using the appropriate design formula and running DESeq2 using the two lines of code:

```
# DO NOT RUN

## Create DESeq2Dataset object
dds <- DESeqDataSetFromTximport(txi, colData = meta, design = ~ sampletype)

## Run analysis
dds <- DESeq(dds)
```

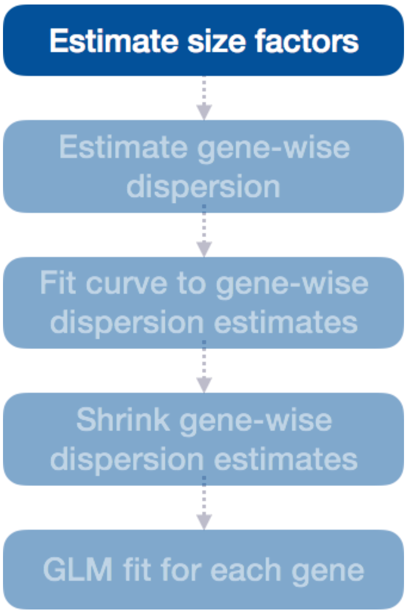
We completed the entire workflow for the differential gene expression analysis with DESeq2. The steps in the analysis are output below:


Pseudocounts with Kallisto, Sailfish, Salmon
Read counts associated with genes
Normalization
Unsupervised clustering analyses
Modeling raw counts for each gene
Shrinking log2 fold changes
Testing for differential expression
Estimate size factors
Estimate gene-wise dispersions
Fit curve to gene-wise dispersion estimates
Shrink gene-wise dispersion estimates
GLM fit for each gene

We will be taking a detailed look at each of these steps to better understand how DESeq2 is performing the statistical analysis and what metrics we should examine to explore the quality of our analysis.

Step 1: Estimate size factors

The first step in the differential expression analysis is to estimate the size factors, which is exactly what we already did to normalize the raw counts.


Estimate size factors
Estimate gene-wise dispersion
Fit curve to gene-wise dispersion estimates
Shrink gene-wise dispersion estimates
GLM fit for each gene

DESeq2 will automatically estimate the size factors when performing the differential expression analysis. However, if you have already generated the size factors using `estimateSizeFactors()` , as we did earlier, then DESeq2 will use these values.

To normalize the count data, DESeq2 calculates size factors for each sample using the *median of ratios method* discussed previously in the 'Count normalization' lesson.

MOV10 DE analysis: examining the size factors

Let's take a quick look at size factor values we have for each sample:

```
## Check the size factors
sizeFactors(dds)

Irrel_kd_1 Irrel_kd_2 Irrel_kd_3 Mov10_kd_2 Mov10_kd_3 Mov10_oe_1 Mov10_oe_2
 1.1149694  0.9606733  0.7492240  1.5633640  0.9359695  1.2262649  1.1405026
Mov10_oe_3
 0.6542030
```

These numbers should be identical to those we generated initially when we had run the function `estimateSizeFactors(dds)` . Take a look at the total number of reads for each sample:

```
## Total number of raw counts per sample
colSums(counts(dds))
```

How do the numbers correlate with the size factor?

We see that the larger size factors correspond to the samples with higher sequencing depth, which makes sense, because to generate our normalized counts we need to divide the counts by the size factors. This accounts for the differences in sequencing depth between samples.

Now take a look at the total depth after normalization using:

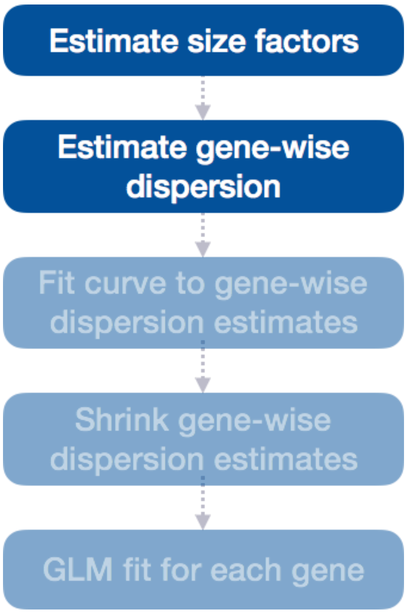
```
## Total number of normalized counts per sample
colSums(counts(dds, normalized=T))
```

How do the values across samples compare with the total counts taken for each sample?

You might have expected the counts to be the exact same across the samples after normalization. However, DESeq2 also accounts for RNA composition during the normalization procedure. By using the median ratio value for the size factor, DESeq2 should not be biased to a large number of counts sucked up by a few DE genes; however, this may lead to the size factors being quite different than what would be anticipated just based on sequencing depth.

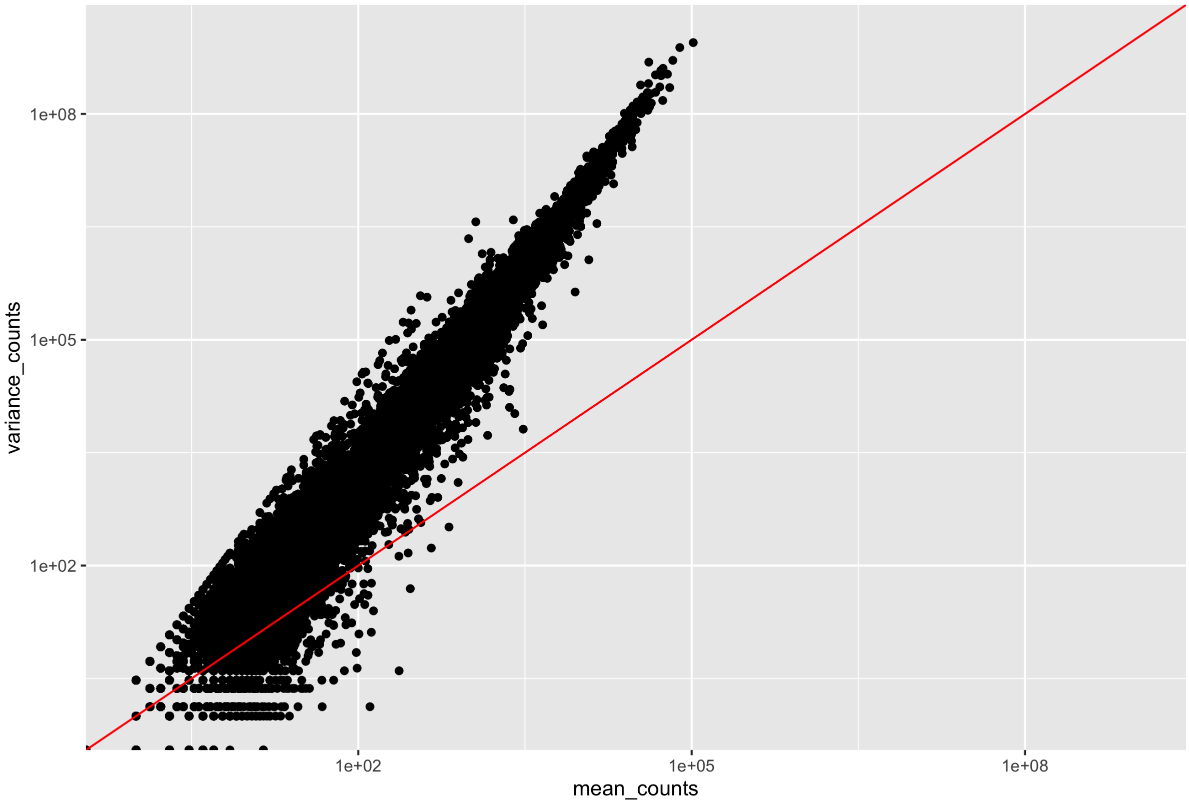
Step 2: Estimate gene-wise dispersion

The next step in the differential expression analysis is the estimation of gene-wise dispersions. Before we get into the details, we should have a good idea about what dispersion is referring to in DESeq2.


Estimate size factors
Estimate gene-wise dispersion
Fit curve to gene-wise dispersion estimates
Shrink gene-wise dispersion estimates
GLM fit for each gene

In RNA-seq count data, we know:

1. To determine differentially expressed genes, we need to identify genes that have significantly different mean expression between groups **given the variation within the groups** (between replicates).
2. The variation within group (between replicates) needs to account for the fact that variance increases with the mean expression, as shown in the plot below (each black dot is a gene).

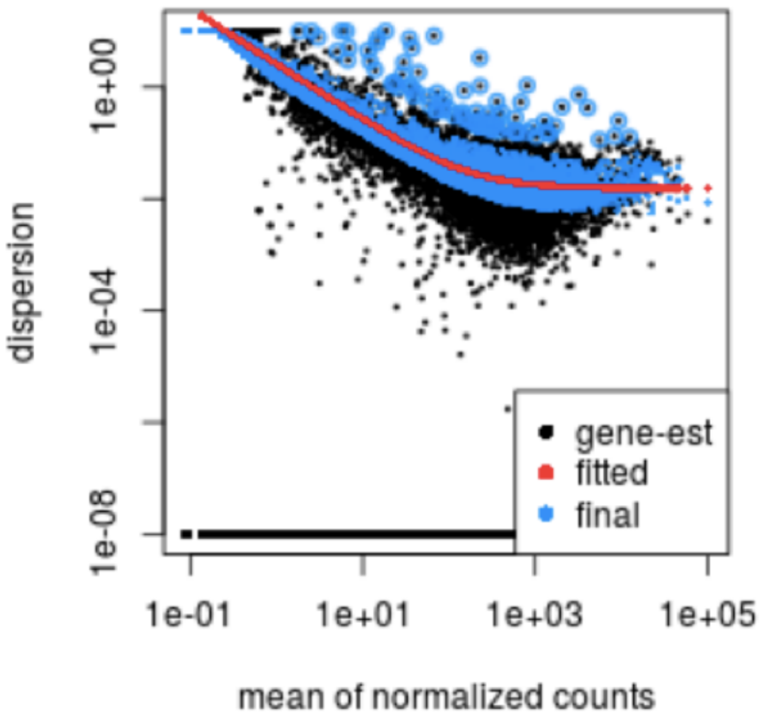


To accurately identify DE genes, DESeq2 needs to account for the relationship between the variance and mean. We don't want all of our DE genes to be genes with low counts because the variance is lower for lowly expressed genes.

Instead of using variance as the measure of variation in the data (*since variance correlates with gene expression level*), DESeq2 uses a measure of variation called **dispersion, which accounts for a gene's variance and mean expression level**. Dispersion is calculated by `Var = μ + α*μ^2`, where `α` = dispersion, `Var` = variance, and `μ` = mean, giving the relationship:

	Effect on dispersion
Variance increases	Dispersion increases
Mean expression increases	Dispersion decreases

For genes with moderate to high count values, the square root of dispersion will be equal to the coefficient of variation. So 0.01 dispersion means 10% variation around the mean expected across biological replicates. The dispersion estimates for genes with the same mean will differ only based on their variance. **Therefore, the dispersion estimates reflect the variance in gene expression for a given mean value.** In the plot below, each black dot is a gene, and the dispersion is plotted against the mean expression (across within-group replicates) for each gene.



How does the dispersion relate to our model?

To accurately model sequencing counts, we need to generate accurate estimates of within-group variation (variation between replicates of the same sample group) for each gene. With only a few (3-6) replicates per group, the **estimates of variation for each gene are often unreliable**.

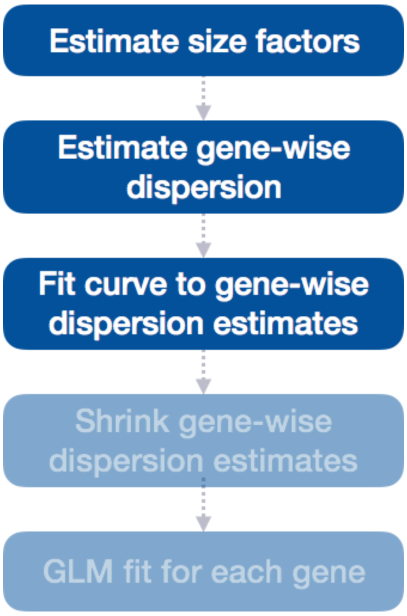
To address this problem, DESeq2 **shares information across genes** to generate more accurate estimates of variation based on the mean expression level of the gene using a method called 'shrinkage'. **DESeq2 assumes that genes with similar expression levels should have similar dispersion.**

Estimating the dispersion for each gene separately:

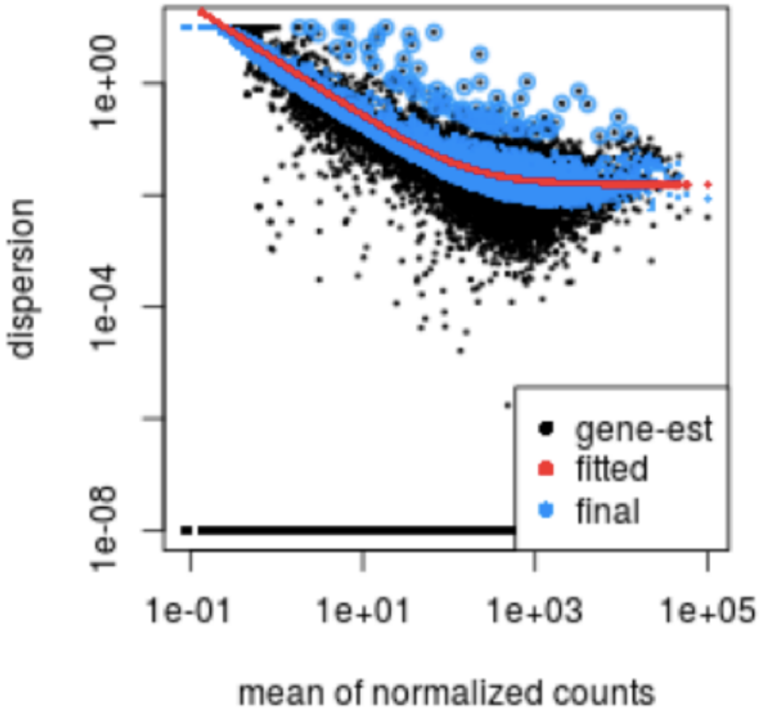
DESeq2 estimates the dispersion for each gene based on the gene's expression level (mean counts of within-group replicates) and variance.

Step 3: Fit curve to gene-wise dispersion estimates

The next step in the workflow is to fit a curve to the gene-wise dispersion estimates. The idea behind fitting a curve to the data is that different genes will have different scales of biological variability, but, across all genes, there will be a distribution of reasonable estimates of dispersion.

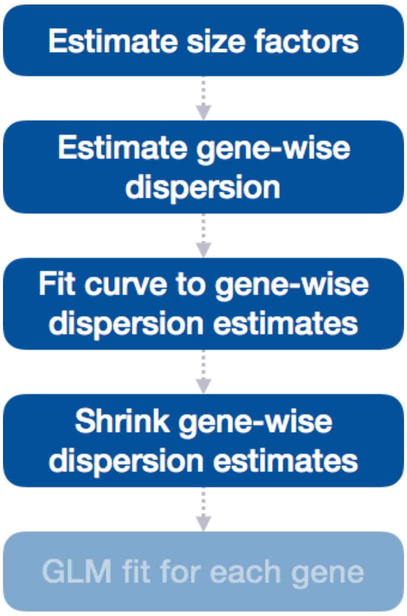

Estimate size factors
Estimate gene-wise dispersion
Fit curve to gene-wise dispersion estimates
Shrink gene-wise dispersion estimates
GLM fit for each gene

This curve is displayed as a red line in the figure below, which plots the estimate for the **expected dispersion value for genes of a given expression strength**. Each black dot is a gene with an associated mean expression level and maximum likelihood estimation (MLE) of the dispersion (Step 1).



Step 4: Shrink gene-wise dispersion estimates toward the values predicted by the curve

The next step in the workflow is to shrink the gene-wise dispersion estimates toward the expected dispersion values.

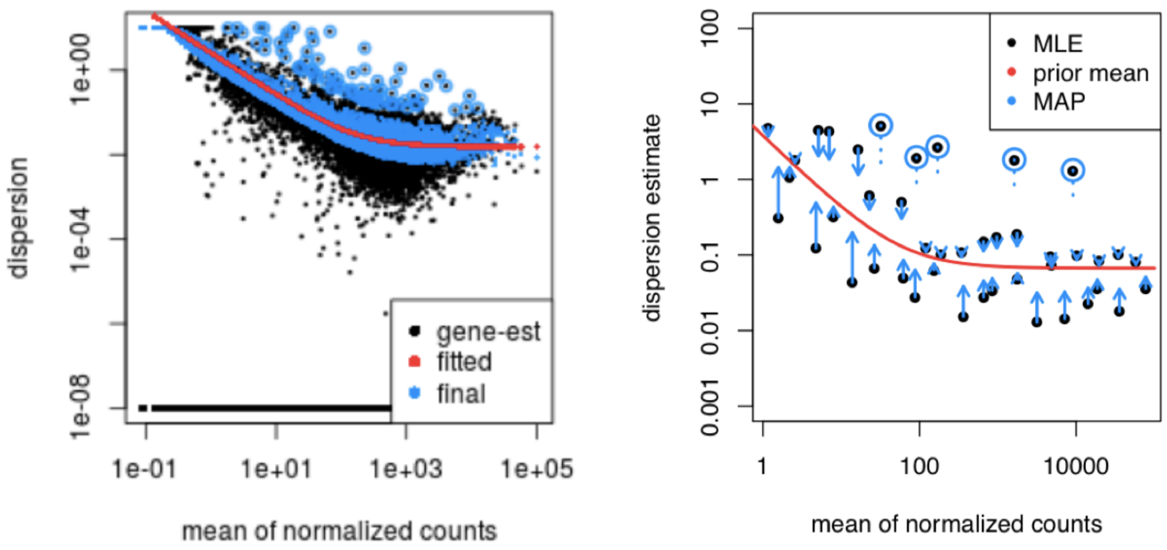

Estimate size factors
Estimate gene-wise dispersion
Fit curve to gene-wise dispersion estimates
Shrink gene-wise dispersion estimates
GLM fit for each gene

The curve allows for more accurate identification of differentially expressed genes when sample sizes are small, and the strength of the shrinkage for each gene depends on :

- how close gene dispersions are from the curve
- sample size (more samples = less shrinkage)

This shrinkage method is particularly important to reduce false positives in the differential expression analysis. Genes with low dispersion estimates are shrunken towards the curve, and the more accurate, higher shrunken values are output for fitting of the model and differential expression testing. These shrunken estimates represent the within-group variation that is needed to determine whether the gene expression across groups is significantly different.

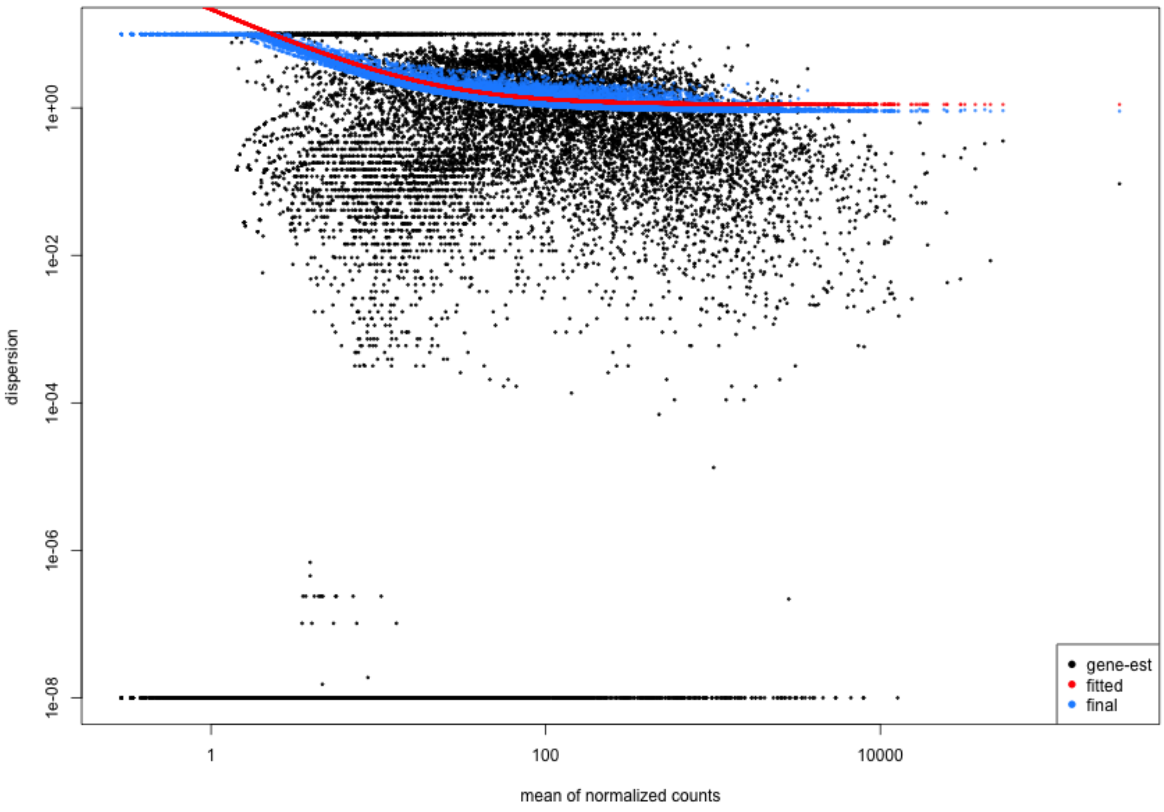
Dispersion estimates that are slightly above the curve are also shrunk toward the curve for better dispersion estimation; however, genes with **extremely high dispersion values are not**. This is due to the likelihood that the gene does not follow the modeling assumptions and has higher variability than others for biological or technical reasons [1]. Shrinking the values toward the curve could result in false positives, so these values are not shrunken. These genes are shown surrounded by blue circles below.



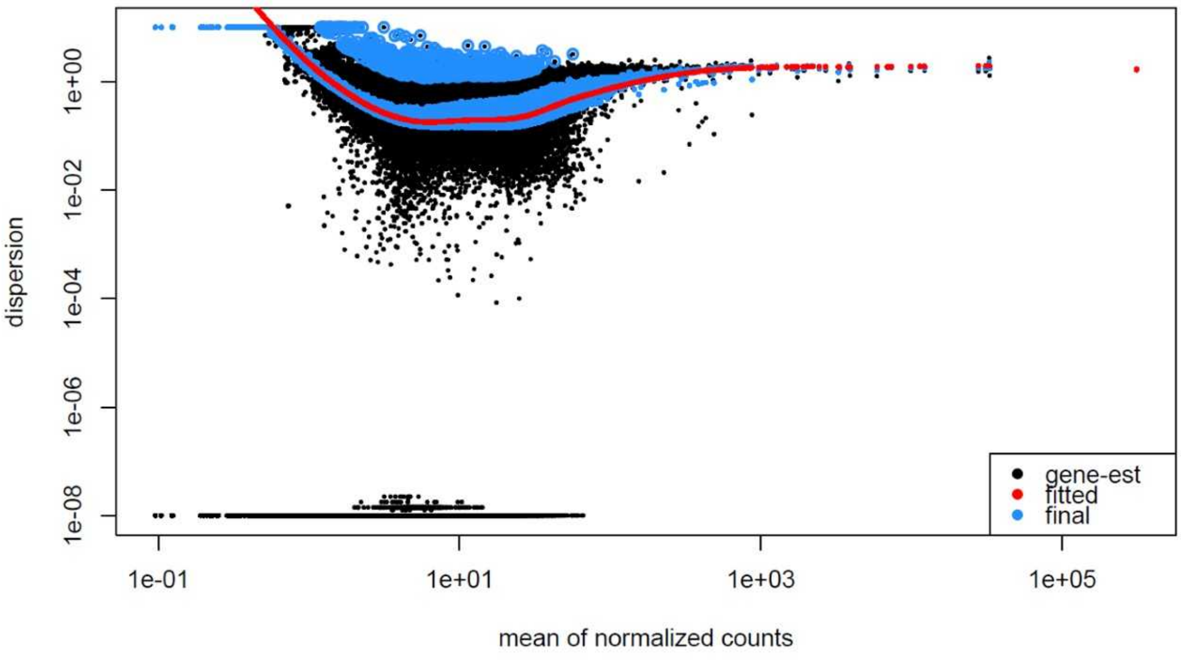
This is a good plot to examine to ensure your data is a good fit for the DESeq2 model. You expect your data to generally scatter around the curve, with the dispersion decreasing with increasing mean expression levels. If you see a cloud or different shapes, then you might want to explore your data more to see if you have contamination (mitochondrial, etc.) or outlier samples. Note how much shrinkage you get across the whole range of means in the `plotDispEsts()` plot for any experiment with low degrees of freedom.

Examples of **worrisome dispersion plots** are shown below:

The plot below shows a cloud of dispersion values, which do not generally follow the curve. This would be worrisome and suggests a bad fit of the data to the model.



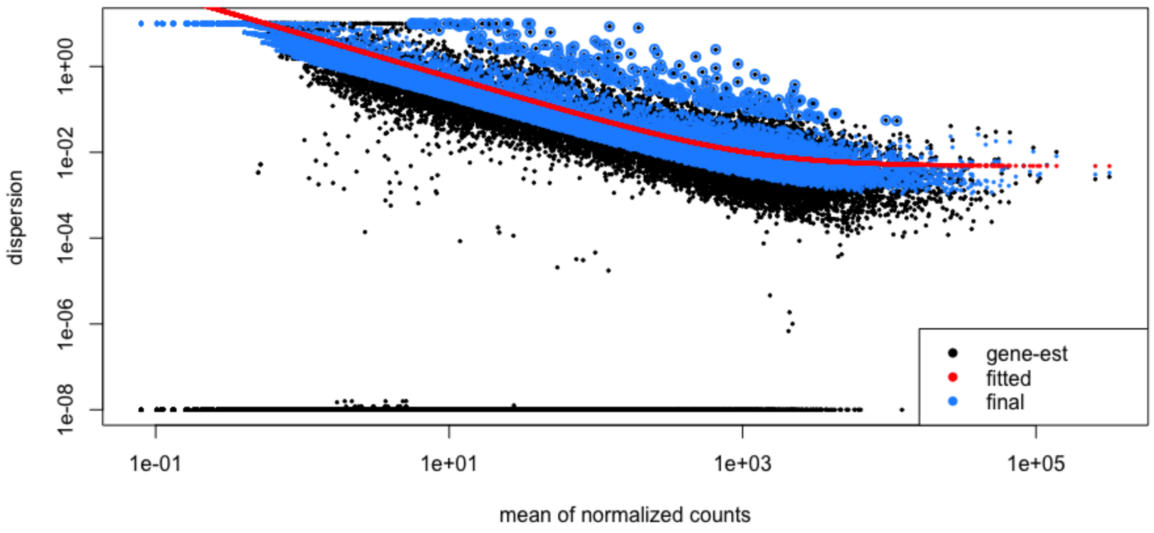
The next plot shows the dispersion values initially decreasing, then increasing with larger expression values. The larger mean expression values should not have larger dispersions based on our expectations - we expect decreasing dispersions with increasing mean. This indicates that there is less variation for more highly expressed genes than expected. This also indicates that there could be an outlier sample or contamination present in our analysis.



MOV10 DE analysis: exploring the dispersion estimates and assessing model fit

Let’s take a look at the dispersion estimates for our MOV10 data:

```
## Plot dispersion estimates
plotDispEsts(dds)
```



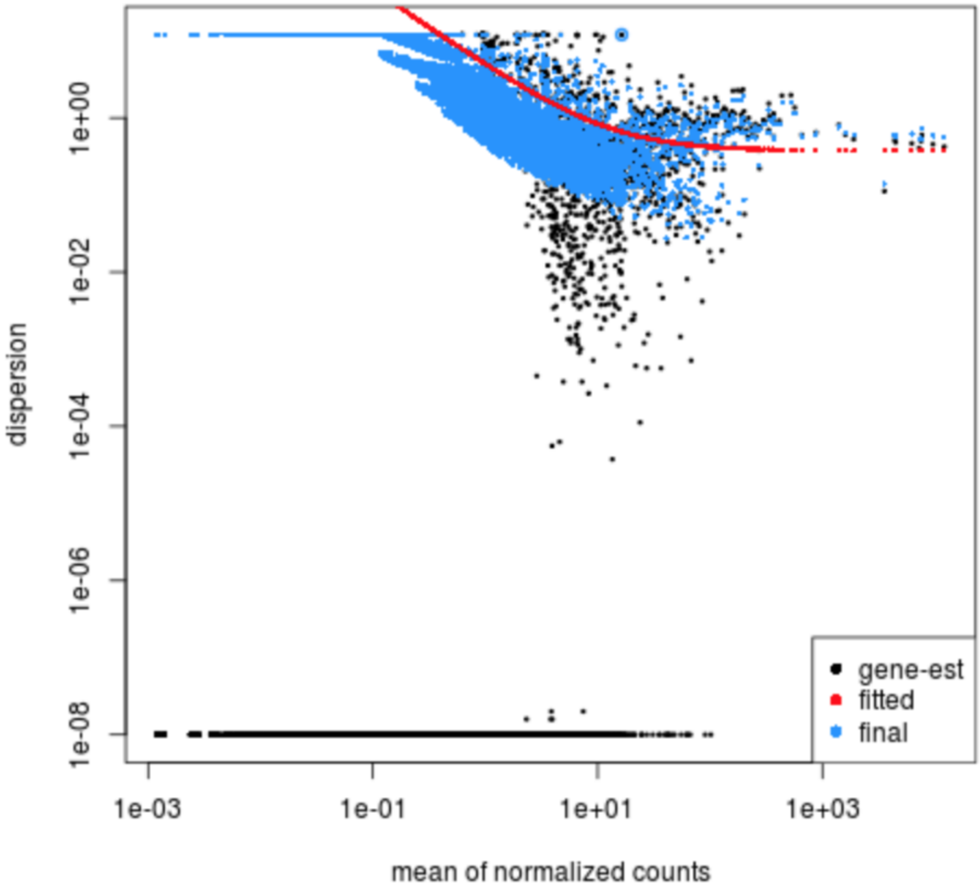
Since we have a small sample size, for many genes we see quite a bit of shrinkage. Do you think our data are a good fit for the model?

We see a nice decrease in dispersion with increasing mean expression, which is good. We also see the dispersion estimates generally surround the curve, which is also expected. Overall, this plot looks good. We do see strong shrinkage, which is likely due to the fact that we have only two replicates for one of our sample groups. The more replicates we have, the less shrinkage is applied to the dispersion estimates, and the more DE genes are able to be identified. We would generally recommend having at least 4 biological replicates per condition for better estimation of variation.

Exercise

Given the dispersion plot below, would you have any concerns regarding the fit of your data to the model?

- If not, what aspects of the plot makes you feel confident about your data?
- If so, what are your concerns? What would you do to address them?



This lesson has been developed by members of the teaching team at the Harvard Chan Bioinformatics Core (HBC). These are open access materials distributed under the terms of the Creative Commons Attribution license (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Some materials and hands-on activities were adapted from RNA-seq workflow on the Bioconductor website

DGE_workshop_salmon_online is maintained by hbctraining.
This page was generated by GitHub Pages.