Differential Expression of RNA-Seq Data

UND Genomics Core

Differential gene expression in R

- DESeq2
  https://bioconductor.org/packages/release/bioc/html/DESeq2.html

- edgeR
  https://bioconductor.org/packages/release/bioc/html/edgeR.html

- To get help
  https://support.bioconductor.org/

Steps in Differential Expression Analysis

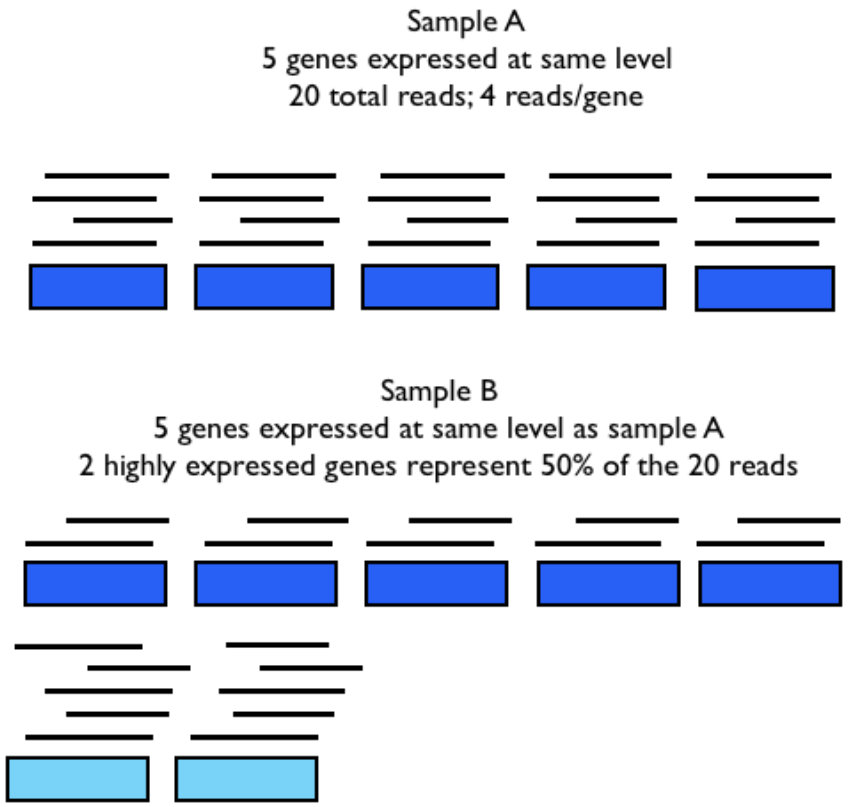
1. Normalize read counts
2. Calculate Dispersion
3. QC
4. Statistical testing

Normalization

- Both DESeq2 and edgeR only account for factors that influence read counts between samples
  - Sequencing depth
  - RNA composition
- RNA composition bias occurs when few transcripts represent a large portion of the reads resulting in other transcripts being underestimated

RNA composition bias

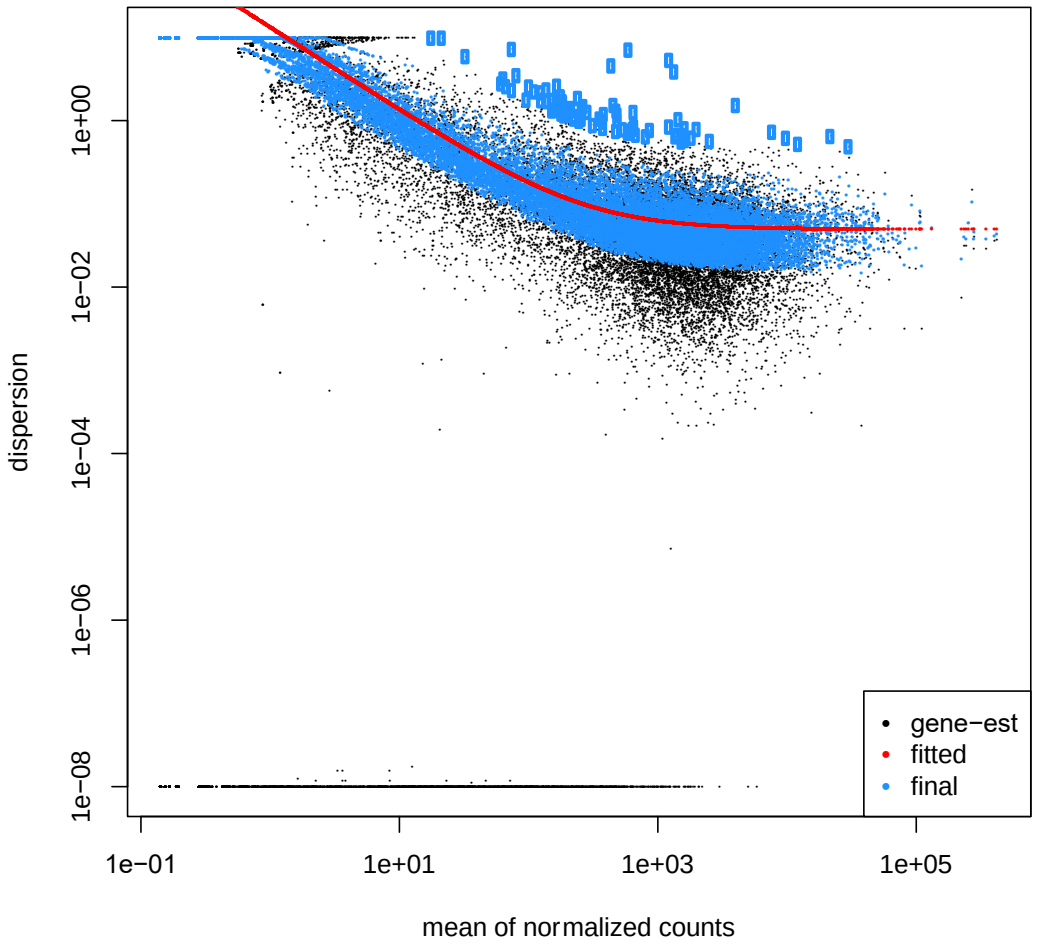
Both samples are sequenced at same depth. Sample A has 5 genes expressed at the same level. Sample B has the same genes as sample A plus two highly expressed genes that take up 50% of the reads



Dispersion

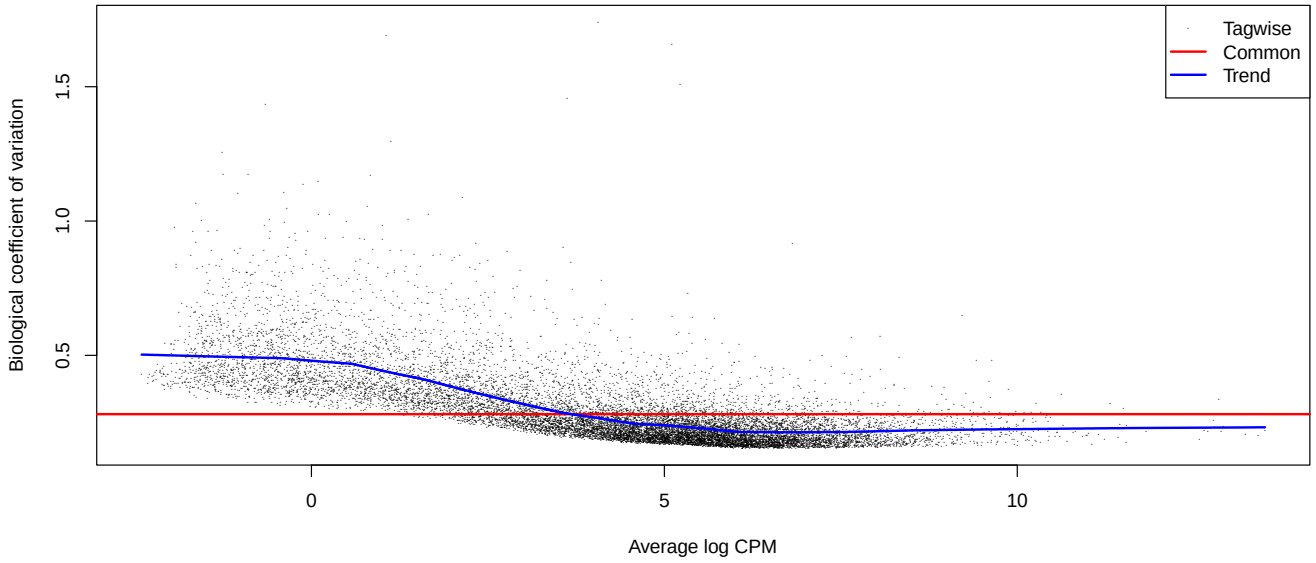
- Dispersion is a measure of variability
- Two sources of variability in RNA-Seq experiments:
  - Technical variation – This is the variation that would happen if you were to re-sequence a sample
  - Biological variation- differences between different samples
- Both DESEq2 and edgeR share information between genes to get better dispersion estimates

Dispersion DESeq2

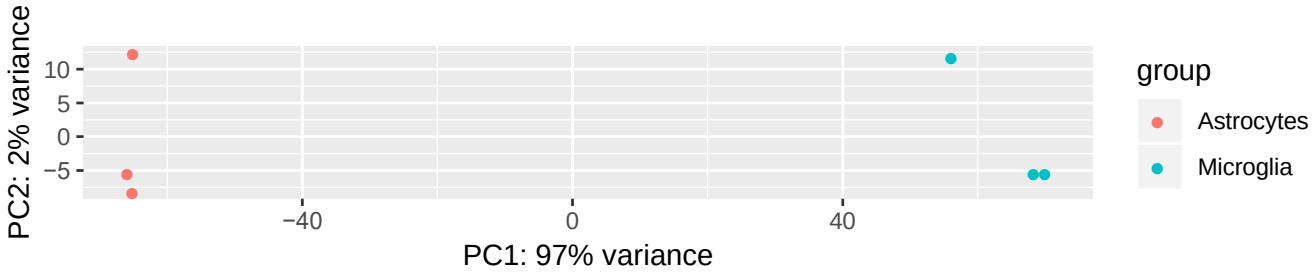


Dispersion EdgeR

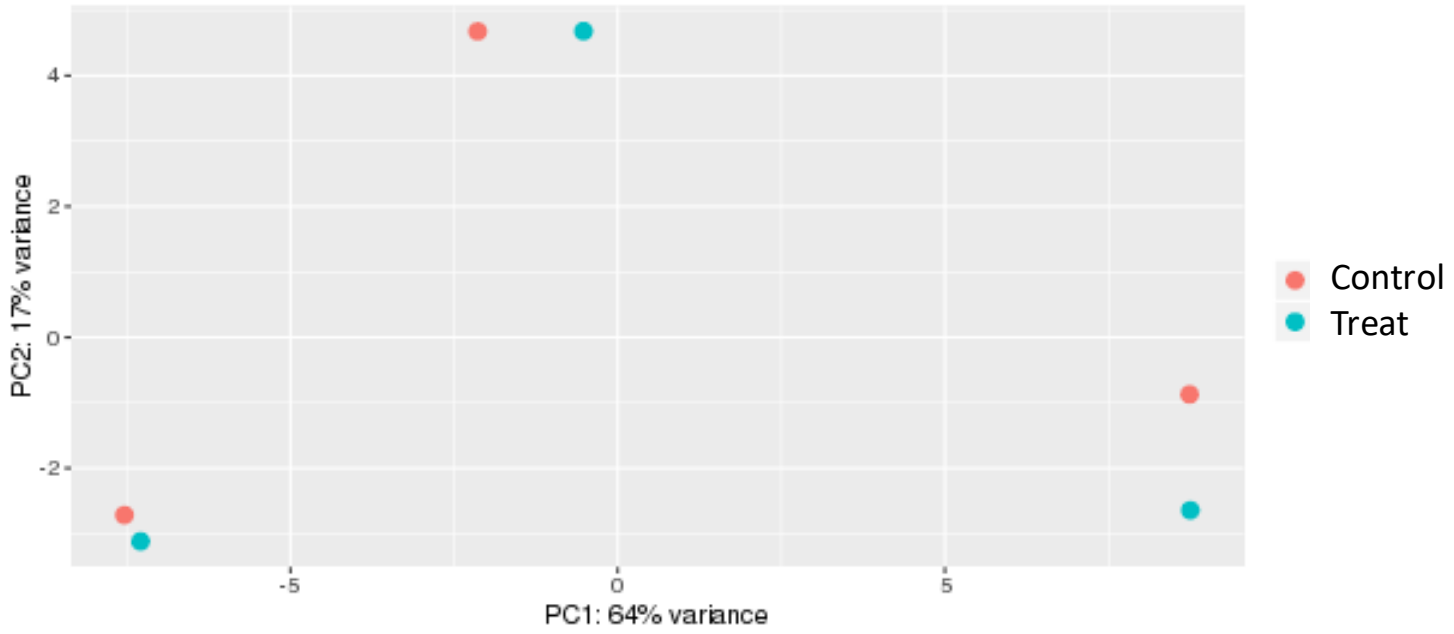
Biological CV
% variation of read counts around replicates
Expected Values:
~ 10% for experiments with using cell-lines or genetically identical organisms
~ 40% for experiments using human samples.



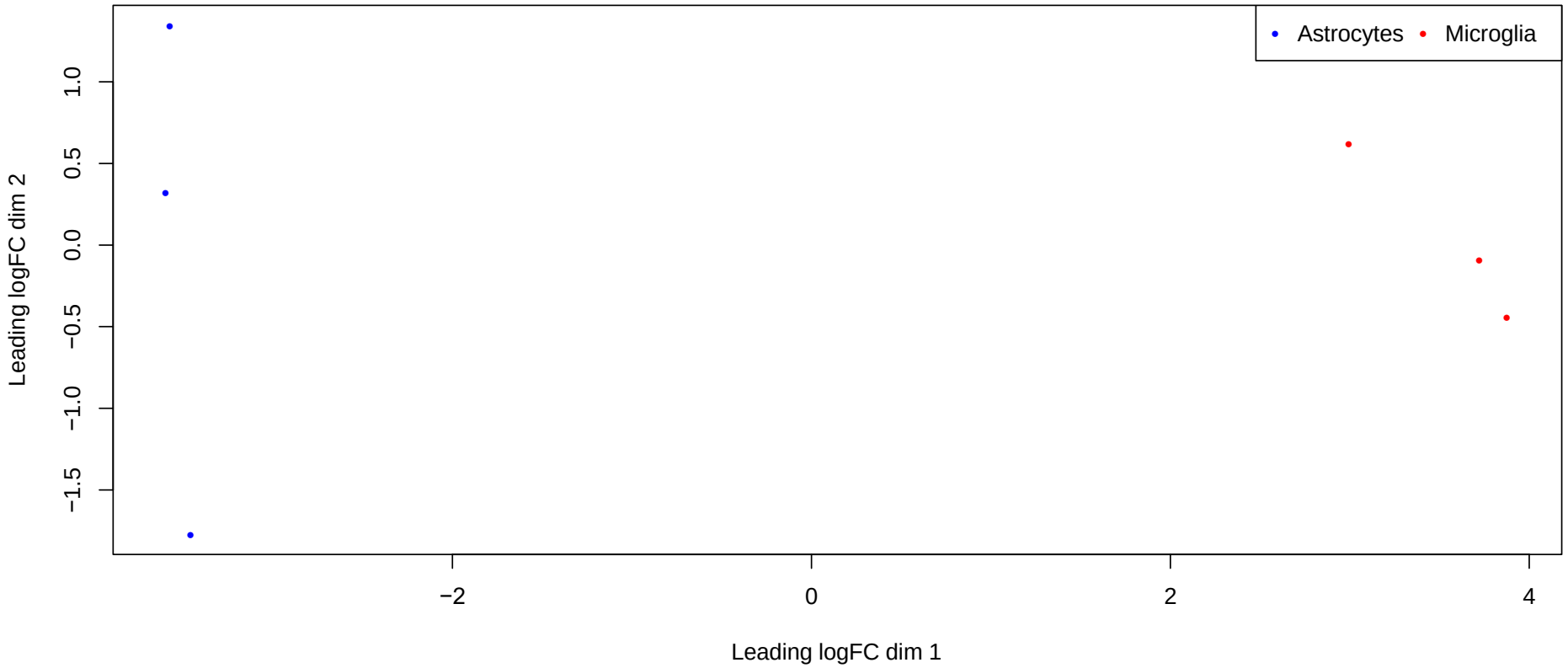
QC - PCA


PC2: 2% variance
10
5
0
−5
−40
0
40
PC1: 97% variance
group
Astrocytes
Microglia

Batch effects



QC -Multidimensional Scaling



DE testing

For ***each*** gene, test if average gene expression in Condition A is significantly different than the average gene expression in Condition B

Gene ID	A1	A2	B1	B2
0610005C13Rik	5	4	2	0
0610007P14Rik	117	119	82	83
0610009L18Rik	39	40	30	22
0610009O20Rik	347	303	164	126

DESeq2 results

```
log2 fold change (MAP): condition Astrocytes vs Microglia
Wald test p-value: condition Astrocytes vs Microglia
DataFrame with 6 rows and 6 columns
             baseMean   log2FoldChange             lfcSE             stat                pvalue                 padj
            <numeric>        <numeric>         <numeric>        <numeric>             <numeric>            <numeric>
Id4   5020.40734581943 8.15034901988938   0.2516571197447 32.3867213777131  4.2216228432672e-230 7.29158697489111e-226
Scg3  4105.51732360569 7.38929345960283 0.231789414675741 31.8793395718264 5.16341880601948e-223 4.45912848087843e-219
Nrbp2 6421.29629667546 7.77342794344814 0.258311906797815 30.0931847850612 5.94979618032589e-199 3.42549598755296e-195
Fhl1  10539.1551633426 7.85196639706473 0.264085540591698 29.7326630586134 2.90582168237698e-194 1.25473380245038e-190
Ddah1 11998.1635305717 7.34888538678968  0.25676747217624 28.6207802121663 3.70485936311317e-180 1.27980661839381e-176
Cnn3  29081.6656279462 8.08581987380511 0.284596918238846 28.4114807842689 1.45881081765198e-177  4.1994300737475e-174
```

baseMean: The average of normalized counts across all samples.

log2FoldChange: The fold change for specific comparison in log2 scale.

lfcSE: The standard error of the log2 Fold Change estimate.

stat: The test statistic. Equals log2FoldChange/lfcSE

pvalue: P-value for that gene.

padj: P-value adjusted for multiple comparisons.

edgeR results table

```
Coefficient:  1*Astrocytes -1*Microglia
                 logFC   logCPM        F       PValue          FDR
Scg3          7.315899 6.736840 936.7542 4.363507e-10 9.340098e-07
Arhgef25      6.558039 5.508114 926.8187 4.566708e-10 9.340098e-07
Klhdc8a       7.812319 4.785456 925.9325 4.585392e-10 9.340098e-07
Zcchc18       7.183448 5.620235 911.7363 4.897970e-10 9.340098e-07
Id4           8.082817 7.031316 887.7936 5.486999e-10 9.340098e-07
Pipox         8.416771 5.667708 857.9814 6.347762e-10 9.340098e-07
Gdpd2         7.460712 5.500371 831.1791 7.267847e-10 9.340098e-07
```

LogFC Fold Change between Condition A and Condition B in $\log_2$ scale

LogCPM Average $\log_2$ Counts per Million across Condition A and B.

PValue unadjusted p-value for each gene

FDR p-value adjusted for multiple comparisons. Default is 0.05

Why adjust for multiple testing?

- P-value is the probability that the observed difference is due to chance.
- For each test, a significance level of 0.05 = 5% error that the test result is significant by chance
- Testing 30,000 genes, 30000 x 0.05 = 1500
- 1500 genes could be significantly DE due to chance alone
- Control these false discoveries with the False Discovery Rate (FDR)

You try!
Follow along with an Differential expression analysis using DESeq2 described in the file DESeq2_handout.docx