bioinformatics.ca

bioinformaticsdotca.github.io

CC BY-SA 4.0 DEED

Attribution-ShareAlike 4.0 International

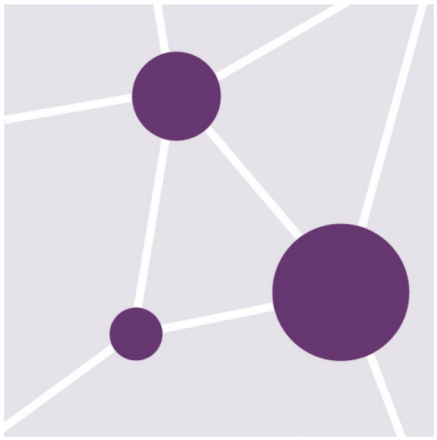
Canadian Bioinformatics Workshops

Module 2a: Differential Expression

Constance Li

Pathway and Network Analysis

May 12, 2026



UNIVERSITY OF CALGARY

Arnie Charbonneau Cancer Institute

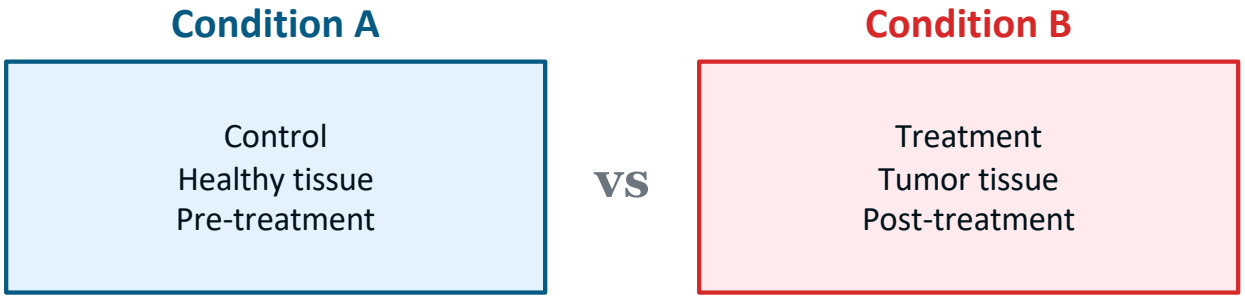
Learning Objectives

1. Understand the fundamentals of differential expression analysis
2. Interpret its statistical outputs
3. Create and interpret common differential expression visualizations
4. Create defined gene lists and ranked gene lists for downstream anlaysis

What is differential (gene) expression?

If we have multiple conditions, we expect

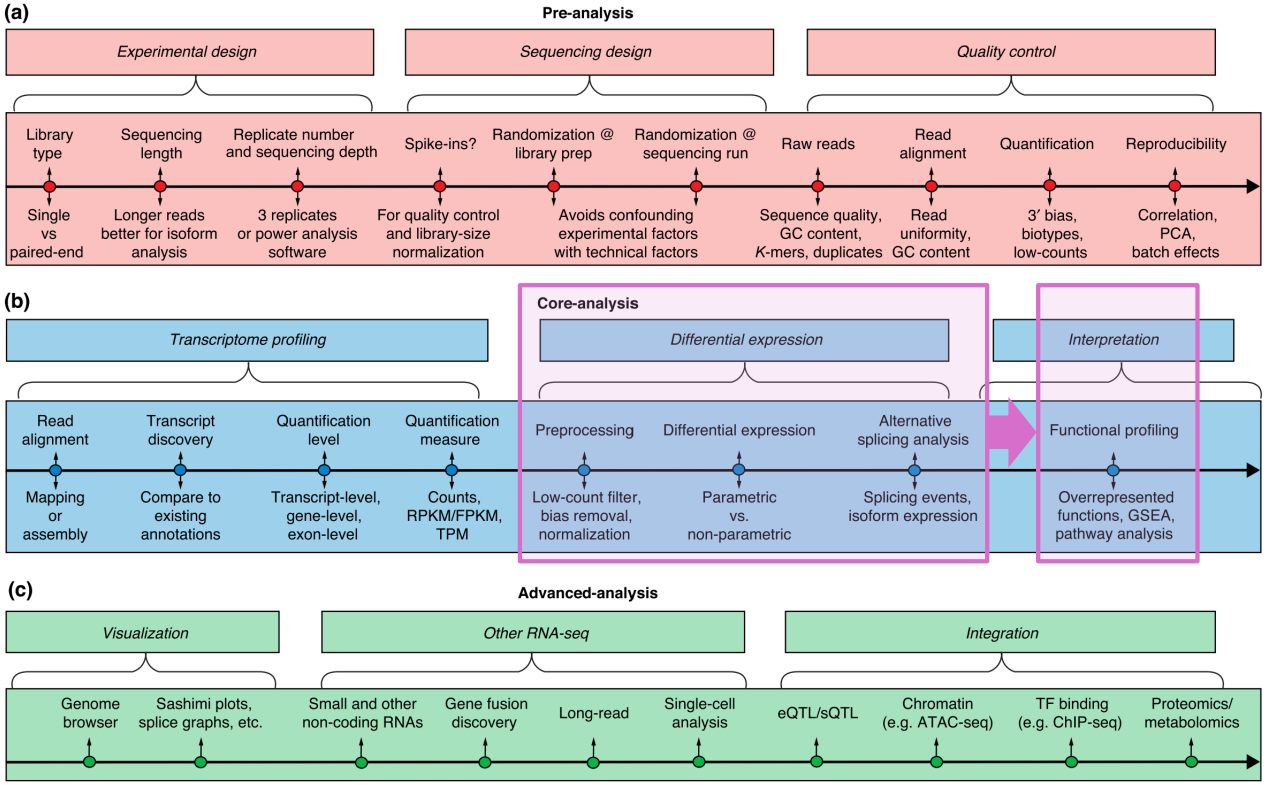
- Observations within the same condition to be similar
- Observations between different conditions to be different



Differential expression (DE) asks how the conditions differ

- Often, we interrogate the transcriptome

Where does DE sit?



https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0881-8

Understanding variability

Biological Variability

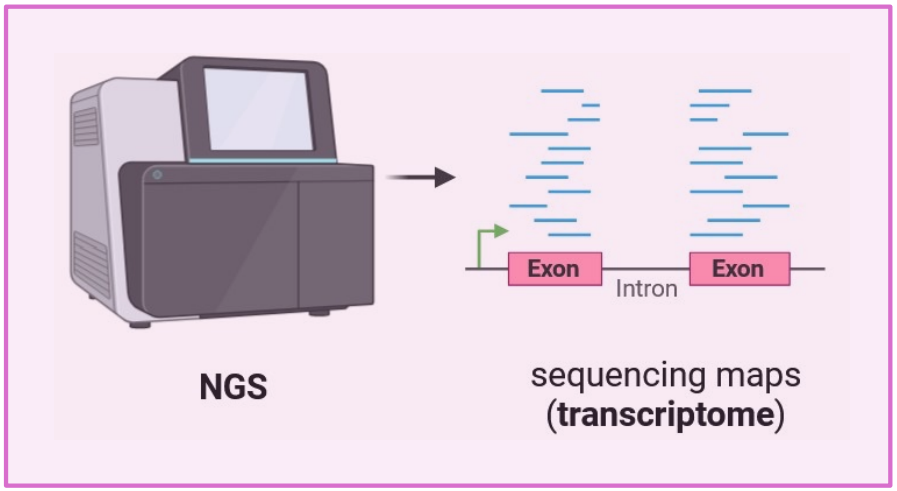
- Natural differences between individuals
- Differences in genetic background
- Cellular heterogeneity
- **This is the signal we want to capture**

Technical Variability

- Library preparation differences
- Sequencing depth variation
- Batch effects
- **This is noise we want to minimize**

- Our measurements include both types of variability
- We must include biological replicates to distinguish true differential expression from noise
  - We often see 3 or more biological replicates
  - Ideally, perform a power analysis

How do we profile gene expression?



- Microarray and RNA-seq are high-throughput gene expression profiling technologies
- Appropriate DE method depends on how data was generated

https://en.wikipedia.org/wiki/DNA_microarray
https://en.wikipedia.org/wiki/RNA-Seq

RNA-seq produces counts data

Samples

Genes

	TCGA.2J.AAB1	TCGA.2J.AAB4	TCGA.2J.AAB6	TCGA.2J.AAB8	TCGA.2J.AAB9	TCGA.2J.AABA
A1BG|1	167.92	134.85	141.16	60.56	98.70	158.63
A1CF|29974	52.00	127.00	14.00	36.00	20.00	30.00
A2BP1|54715	1.00	5.00	0.00	0.00	1.00	0.00
A2LD1|87769	370.02	263.92	278.94	197.66	155.54	158.52
A2ML1|144568	176.00	0.00	3105.00	18.00	599.00	1395.00
A2M|2	40392.80	37630.67	14564.83	15332.80	22682.89	16449.73
A4GALT|53947	3160.00	2744.00	1917.00	418.00	1050.00	1011.00
A4GNT|51146	893.00	113.00	2.00	10.00	47.00	10.00
AAA1|404744	4.00	1.00	2.00	1.00	0.00	30.00
AAAS|8086	1402.00	1268.00	1427.00	853.00	805.00	996.00
AACSL|729522	1.00	2.00	0.00	1.00	0.00	0.00
AACS|65985	2445.00	2915.00	994.00	726.00	600.00	1476.00
AADACL2|344752	0.00	0.00	0.00	0.00	0.00	0.00
AADACL3|126767	0.00	0.00	0.00	0.00	0.00	0.00
AADACL4|343066	1.00	0.00	0.00	0.00	0.00	0.00
AADAC|13	1222.00	77.00	12.00	216.00	118.00	89.00
AADAT|51166	56.00	190.00	103.00	89.00	83.00	100.00
AAGAB|79719	2393.00	3236.00	1628.00	1556.00	1031.00	2050.00
AAK1|22848	1600.00	2090.00	1560.00	646.00	852.00	1294.00
AAMP|14	7098.00	6413.00	3765.00	3340.00	2257.00	4814.00

Counts data represent the number of reads aligned to each gene

Note that raw counts do not account for

- Gene length
- Sequencing depth
- Library prep and other technical factors

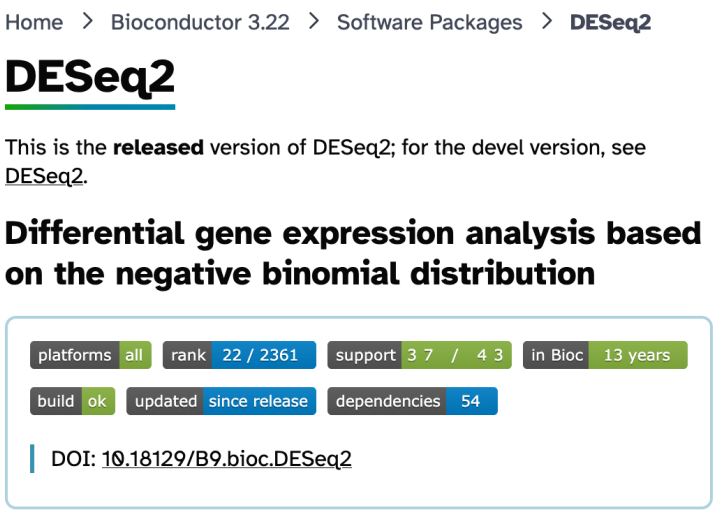
Count data are also

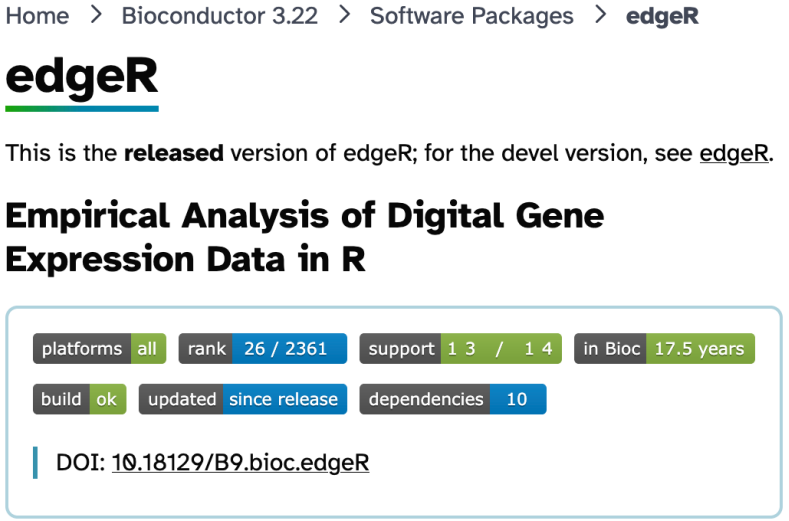
- Non-negative integers
- Right-skewed
- Over-dispersed

⚠ This means count data violates assumptions of standard statistical tests (t-test, ANOVA) which assume normally distributed data with constant variance.

Tools for RNA-seq DE analysis

- We must model RNA-seq count data appropriately





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

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

Tools for RNA-seq DE analysis

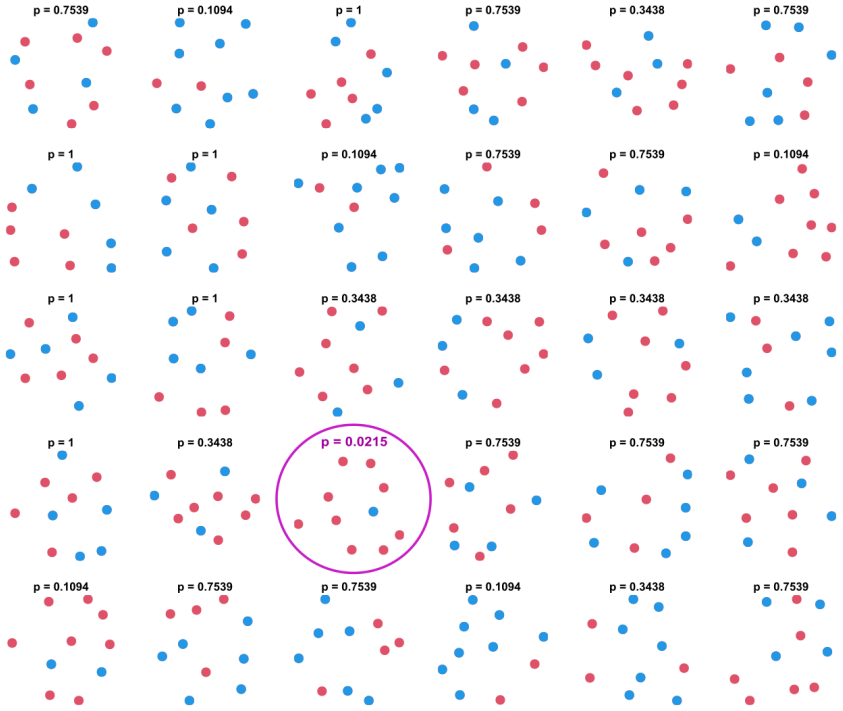
- We must model RNA-seq count data appropriately
- DESeq2 and EdgeR
  - Use negative binomial distribution models
  - Account for dispersion
  - Include tools for statistical analysis and visualization
  - Are both frequently used, well-documented, and well-supported
- We'll use DESeq2 in Lab 2a

Key statistical outputs

Log2 Fold Change log2FC	P-value p	Adjusted P-value FDR
Direction and magnitude of change	Probability of seeing this result by chance	P-value corrected for multiple testing
log2FC = 1 means 2× higher log2FC = -1 means 2× lower	Tests if expression differs from null hypothesis	Controls false discovery rate across all genes tested

Why correct for multiple testing?

The multiple testing problem



Suppose have a population with equal numbers of blue and pink dots

- We randomly sample 10 dots thirty times
- We test the null hypothesis for each sample
- We obtain one statistically significant result
  - Do we believe it?

Adapted from https://en.wikipedia.org/wiki/Multiple_comparisons_problem

The multiple testing problem

In bioinformatics, we often run thousands of tests and use a significance threshold (α) of 0.05

Problem

Testing 20,000 genes at $\alpha = 0.05 \rightarrow$ expect 1,000 false positives by chance alone!

Solution: Multiple testing correction

Two approaches:

- Family-Wise Error Rate correction is more stringent (eg. Bonferroni)
- False Discovery Rate correction is less stringent (eg. Benjamini-Hochberg)
    - Differential expression tools report FDR-adjusted p-values

Interpreting DE results

1. Understand the comparison

3. Consider statistical significance

```
log2 fold change (MLE): tumour.moffitt.cluster classical vs basal
Wald test p-value: tumour.moffitt.cluster classical vs basal
DataFrame with 20502 rows and 6 columns
                   baseMean log2FoldChange     lfcSE      stat      pvalue       padj
                  <numeric>      <numeric> <numeric> <numeric>   <numeric>  <numeric>
A1BG|1            184.80519     -0.0540375 0.1352394 -0.399569 6.89474e-01 8.34356e-01
A1CF|29974         97.43289      1.4938818 0.2305062  6.480874 9.11928e-11 7.08327e-09
A2BP1|54715         2.60974      0.3142817 0.2911807  1.079336 2.80438e-01 4.90682e-01
A2LD1|87769       258.36990      0.3710265 0.0960856  3.861418 1.12731e-04 1.30400e-03
A2ML1|144568      644.65539     -3.2550244 0.4799032 -6.782669 1.17976e-11 1.21562e-09
...                     ...            ...       ...       ...         ...        ...
ZYX|7791        1.20129e+04     -0.1275698 0.0757673 -1.683705  0.09223867  0.233321
ZZEF1|23140     2.23027e+03      0.2105631 0.0754541  2.790611  0.00526087  0.028755
ZZZ3|26009      1.18263e+03     -0.0234501 0.0610241 -0.384276  0.70077384  0.842145
psiTPTE22|387590 8.17595e+01      0.2341760 0.1591425  1.471487  0.14115950  0.313411
tAKR|389932      3.37914e-01      0.1712012 0.6382664  0.268228  0.78852350  0.893372
```

2. Examine effect size in the context of the comparison

$$\log_2 FC = \log_2\left(\frac{classical}{basal}\right)$$

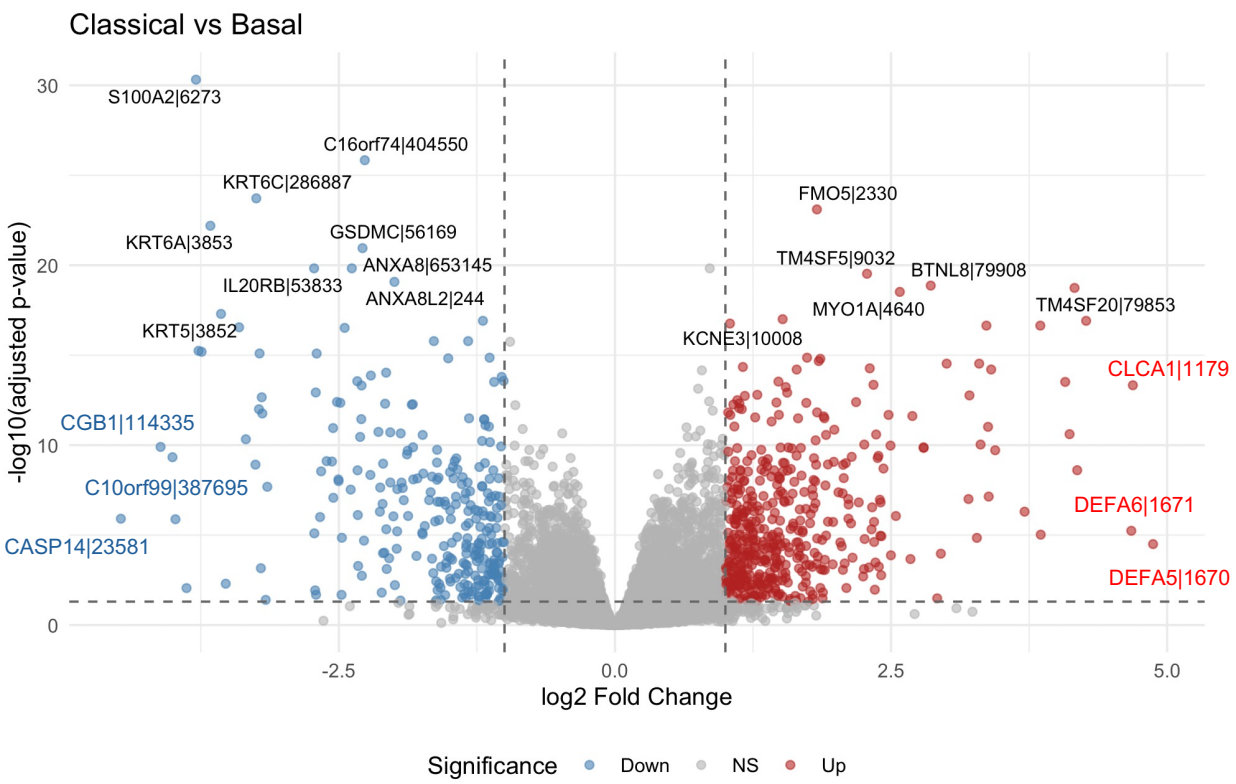
↑ ***A1CF*: Significantly upregulated in classical subtype**

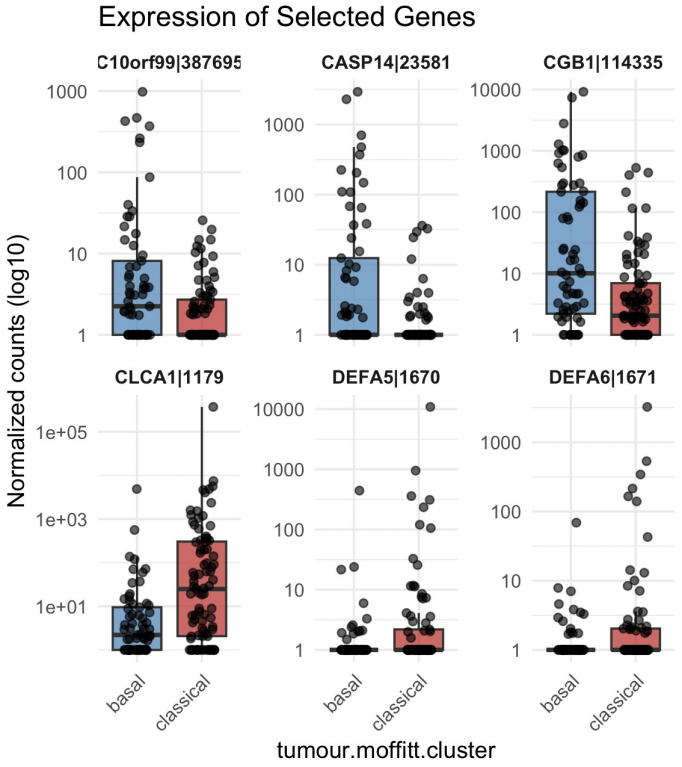
log2FC > 0, FDR < 0.05

↓ ***A2ML1*: Significantly downregulated in classical subtype**

log2FC < 0, FDR < 0.05

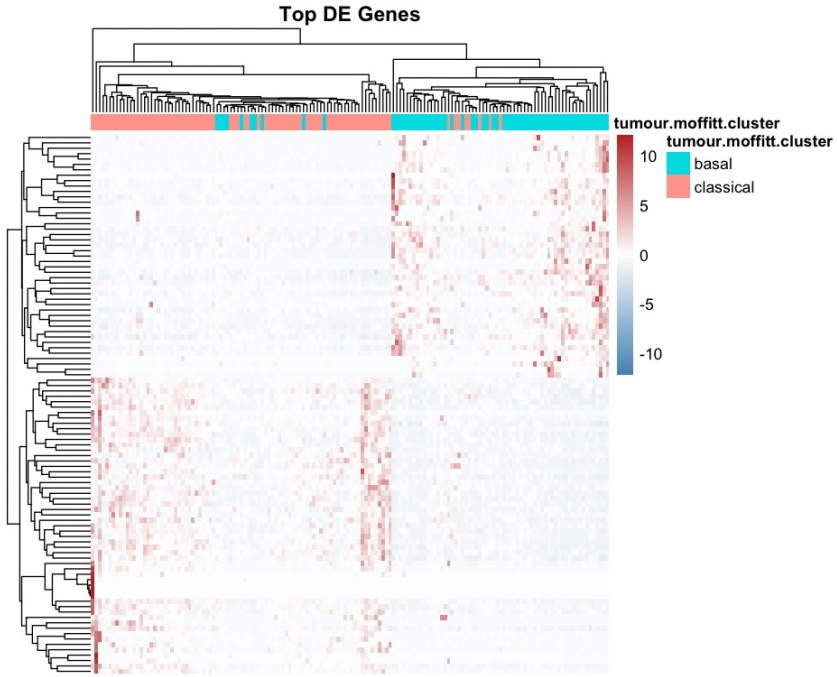
Visualize DE results to aid interpretation





Visualizing DE results

- Global view: volcano plots, heatmaps
  - Great results summaries
- Focused view: boxplots, violin plots
  - Gene-level data
  - Great for sanity-checking your understanding



DE-derived gene lists

Pathway analysis tools require different inputs depending on the method

Defined Gene List

For Over-Representation Analysis

- A simple list of significant gene names
- Filtered by thresholds (FDR, FC)
- Unordered
- Often separated: up vs down
- **Tool: g:Profiler**

```
TP53
BRCA1
MYC
...
```

Ranked Gene List

For Gene Set Enrichment Analysis

- ALL genes with a numeric score
- No arbitrary threshold cutoff
- Ranked from most up to most down
- **Tool: GSEA**

```
TP53    14.2
BRCA1   -9.8
MYC      0.5
...
```

Creating defined gene lists

Common filtering criteria:

- By statistical significance: Adjusted p-value < 0.05
- By biological significance: (2-fold change): |log2 Fold Change| > 1
  - Consider relaxing FC threshold if few genes pass (|log2FC| > 0.58 = 1.5-fold)
- Ideally, consider both

Upregulated Genes

FDR < 0.05 AND log2FC > 1

Higher in condition B vs A

Downregulated Genes

FDR < 0.05 AND log2FC < -1

Lower in condition B vs A

💡 Rough rule of thumb: Aim for 100-2000 genes per list. Too few = low power; Too many = loss of specificity

Creating ranked gene lists

- GSEA expects a ranking value for every gene

$$\text{ranking value} = -\log_{10}(\text{p-value}) * \text{sign}(\text{log2FC})$$

  - Combines statistical significance with direction of change
- You must include all genes tested (even not-significant genes)
  - GSEA looks at the distribution of pathway genes across the entire ranked list

The `.rnk` file format

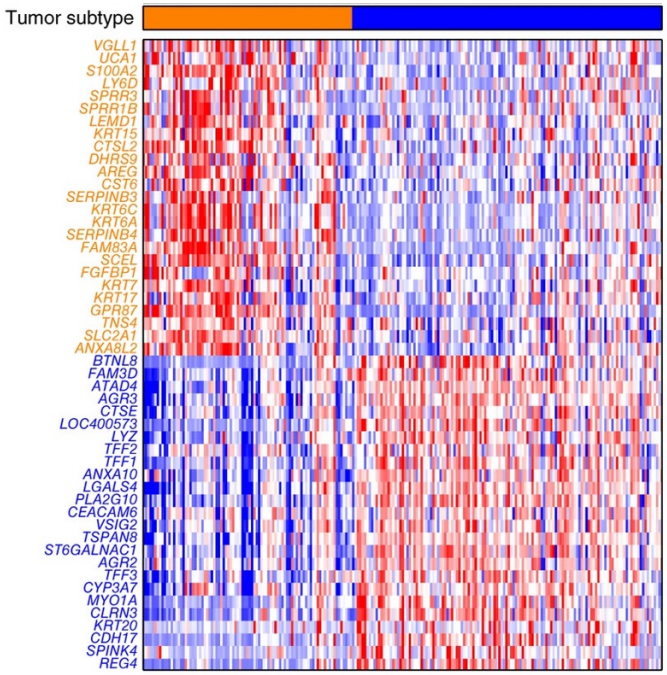
- GSEA requires `.rnk` files
  - Plain text file (`.rnk` extension)
  - Tab-delimited (two columns)
    - Column 1: Gene symbol
    - Column 2: Ranking score
  - No header row
  - Sorted by score (optional but helpful)

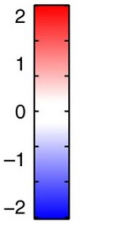
Example .rnk file

```
TP53        14.23
EGFR        12.87
KRAS        11.45
...
GAPDH       0.02
...
BRCA2       -8.91
BRCA1       -9.76
RB1         -11.34
```

⚠ **Common Pitfalls:**
Duplicate gene names (keep highest |score|) • NA/Inf values (remove or replace) • Wrong gene ID type (use symbols matching your GMT file)

DE Lab: TCGA pancreatic cancer RNA-seq





Module 2a Lab

DESeq2

Defined Gene List → Module 2b Lab

Ranked Gene List → Module 2c Lab

Beyond differential expression

- The same pathway analysis workflow applies to gene lists from many sources

ChIP-seq Peaks → nearby genes	**CRISPR screens** Hits from dropout/enrichment	**GWAS** SNP-associated genes
Co-expression Module members (WGCNA)	**Proteomics** Differentially abundant proteins	**Literature** Curated gene sets

- Any gene list can be analyzed for pathway enrichment

Summary

- DE analysis identifies genes with significant expression changes between conditions
- Use appropriate tools (DESeq2, edgeR) that model count data correctly
  - Other tools exist for other data (*eg*. limma)
- Always use FDR-corrected p-values
- Use visualizations to strengthen your interpretation of results
- Create defined gene lists (filtered) for tools like g:Profiler
- Create ranked gene lists (all genes) for tools like GSEA

Coffee Break & Networking Session

Workshop Sponsors