VIDO BIOINFO
Bioinformatics
and Systems
Biology Lab

Canadian
Bioinformatics Hub

Canadian Bioinformatics Workshops

Introduction to R Programming for Bioinformatics

Day 2- Module 4A: Differential Expression Analysis and Mini Project

Mohamed Helmy, PhD
Principal Scientist and Adjunct Professor
Bioinformatics and Systems Biology Lab
VIDO, University of Saskatchewan

6-7 October 2025, VIDO, Saskatoon

VIDO

UNIVERSITY OF
SASKATCHEWAN

Learning Objectives

By the end of this module, we should have knowledge on:

- Perform a differential expression analysis using DESeq2.
- Interpret DGE results (log2 fold change, adjusted p-value).
- Use AI tools responsibly to assist coding and data analysis.
- Apply R and Bioconductor skills in a mini team project.

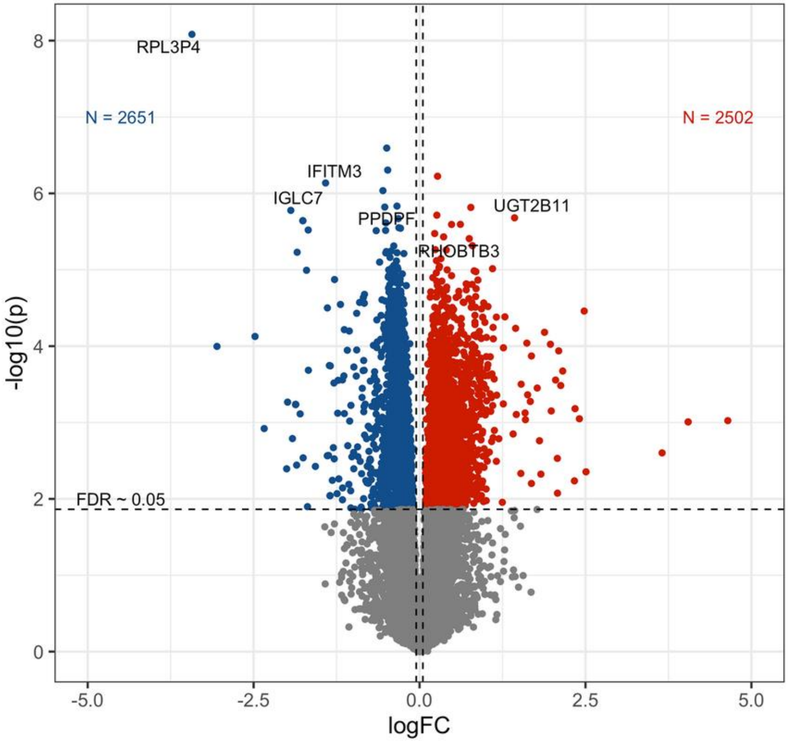
Differential Gene Expression (DGE) Analysis

What is Differential Expression?

- **Compares gene expression between two conditions (e.g., treated vs untreated).**

- **Identifies genes whose expression changes significantly.**

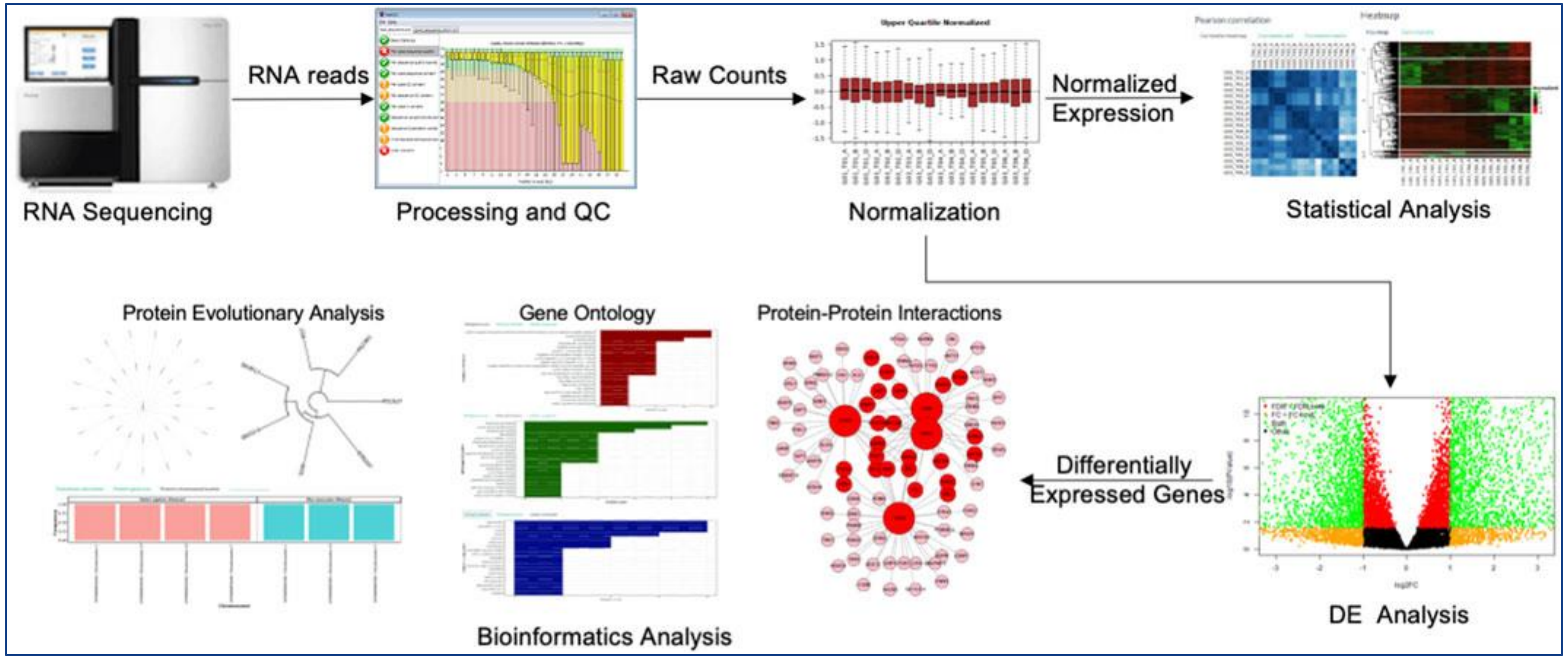
- **Key outputs:**
    - log2 Fold Change: magnitude and direction of change
    - *p*-value/adjusted *p*-value (FDR): significance of the change



Özgümüş, T., et al. *Sci Rep*,2021

DGE Workflow in DESeq2

- **Prepare data (DESeqDataSet)**
- **Run the DESeq() pipeline**
- **Extract results**
- **Visualize results**
  - Volcano plot
  - MA plot
  - Heatmap



Helmy M, *et al. Frontiers in Bioinformatics*, 2021

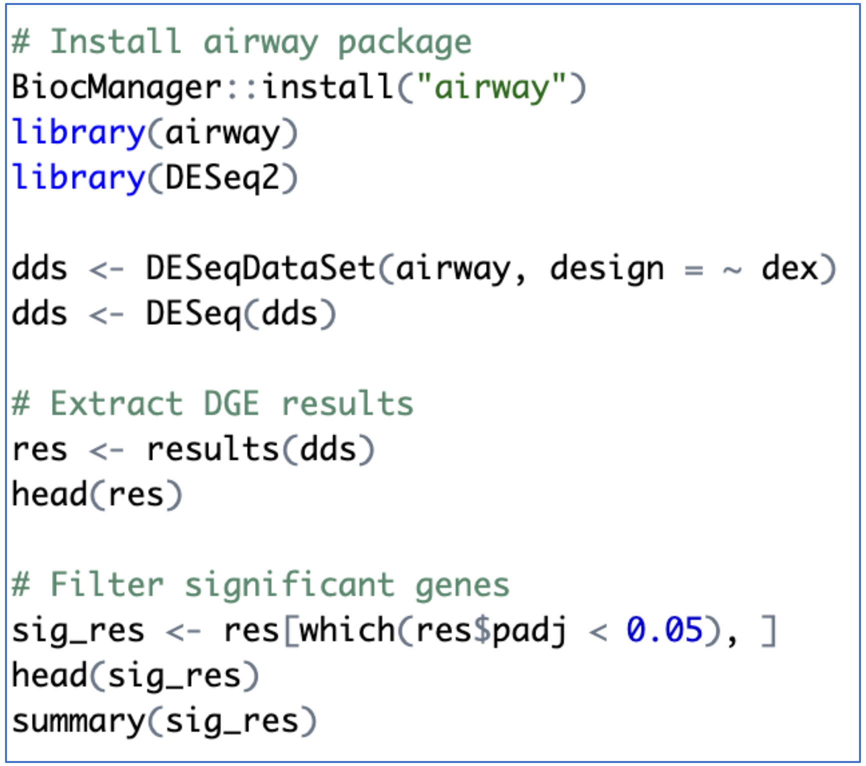
Demo: DGE Analysis Code

- **Extract DEGs using *airway* and *DESeq* packages**

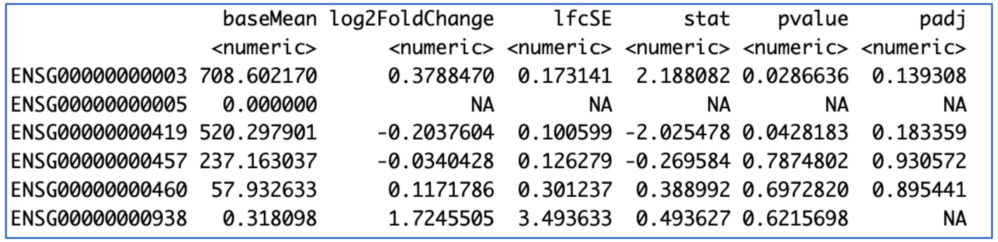
- **Create a table of genes with**
  - log2FoldChange
  - *p*-value
  - adjusted *p*-value

- **Extract DGE results**
  - Use *results()*

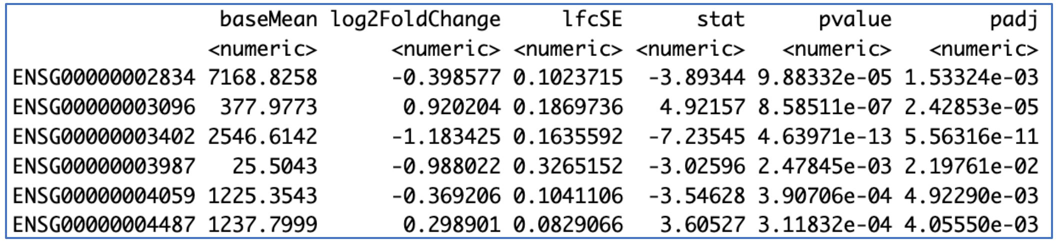
- **Interpret direction and magnitude**



```r
# Install airway package
BiocManager::install("airway")
library(airway)
library(DESeq2)

dds <- DESeqDataSet(airway, design = ~ dex)
dds <- DESeq(dds)

# Extract DGE results
res <- results(dds)
head(res)

# Filter significant genes
sig_res <- res[which(res$padj < 0.05), ]
head(sig_res)
summary(sig_res)
```

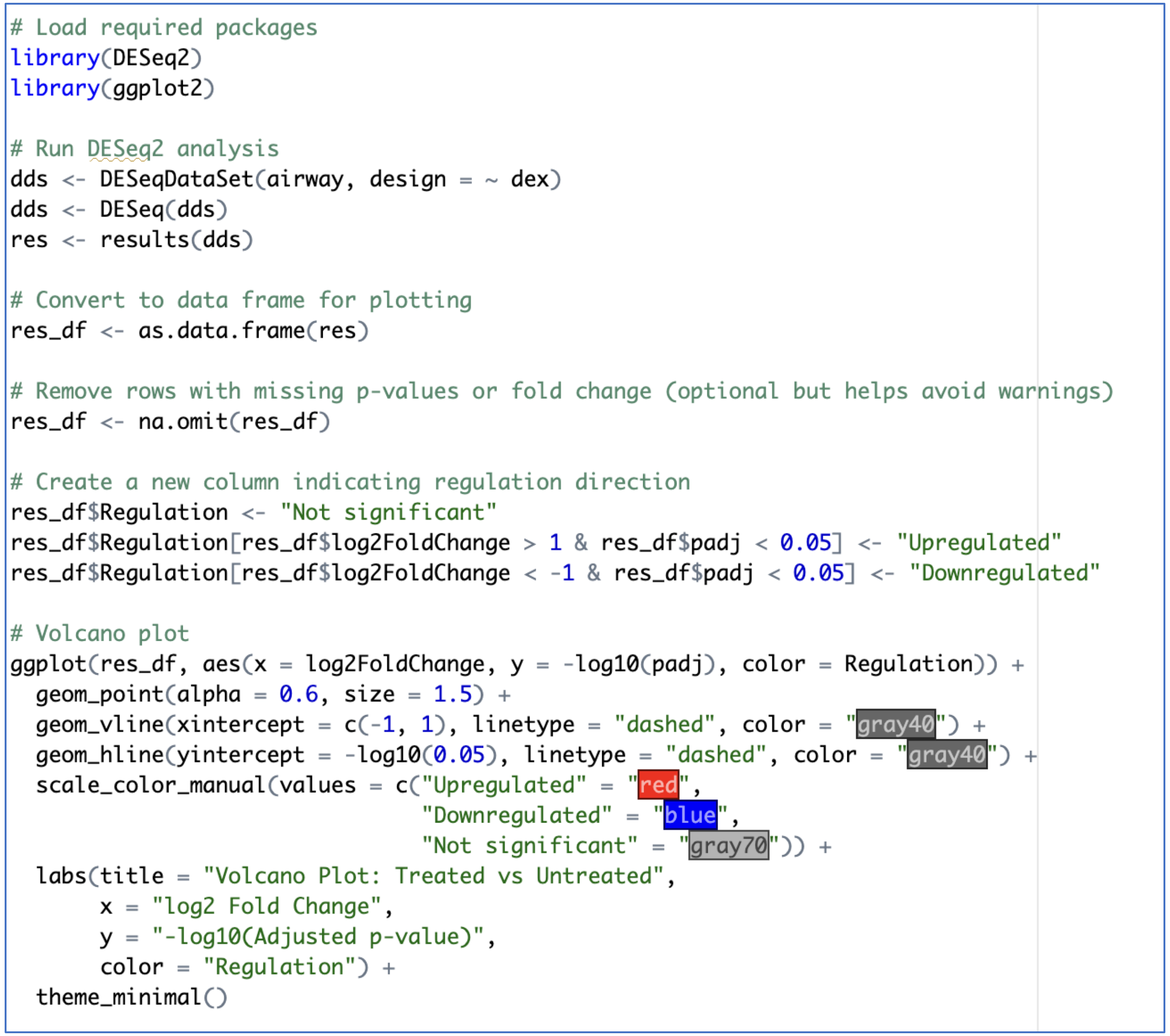


```
                 baseMean log2FoldChange     lfcSE      stat    pvalue      padj
                <numeric>      <numeric> <numeric> <numeric> <numeric> <numeric>
ENSG00000000003 708.602170      0.3788470  0.173141  2.188082 0.0286636  0.139308
ENSG00000000005   0.000000             NA        NA        NA        NA        NA
ENSG00000000419 520.297901     -0.2037604  0.100599 -2.025478 0.0428183  0.183359
ENSG00000000457 237.163037     -0.0340428  0.126279 -0.269584 0.7874802  0.930572
ENSG00000000460  57.932633      0.1171786  0.301237  0.388992 0.6972820  0.895441
ENSG00000000938   0.318098      1.7245505  3.493633  0.493627 0.6215698        NA
```

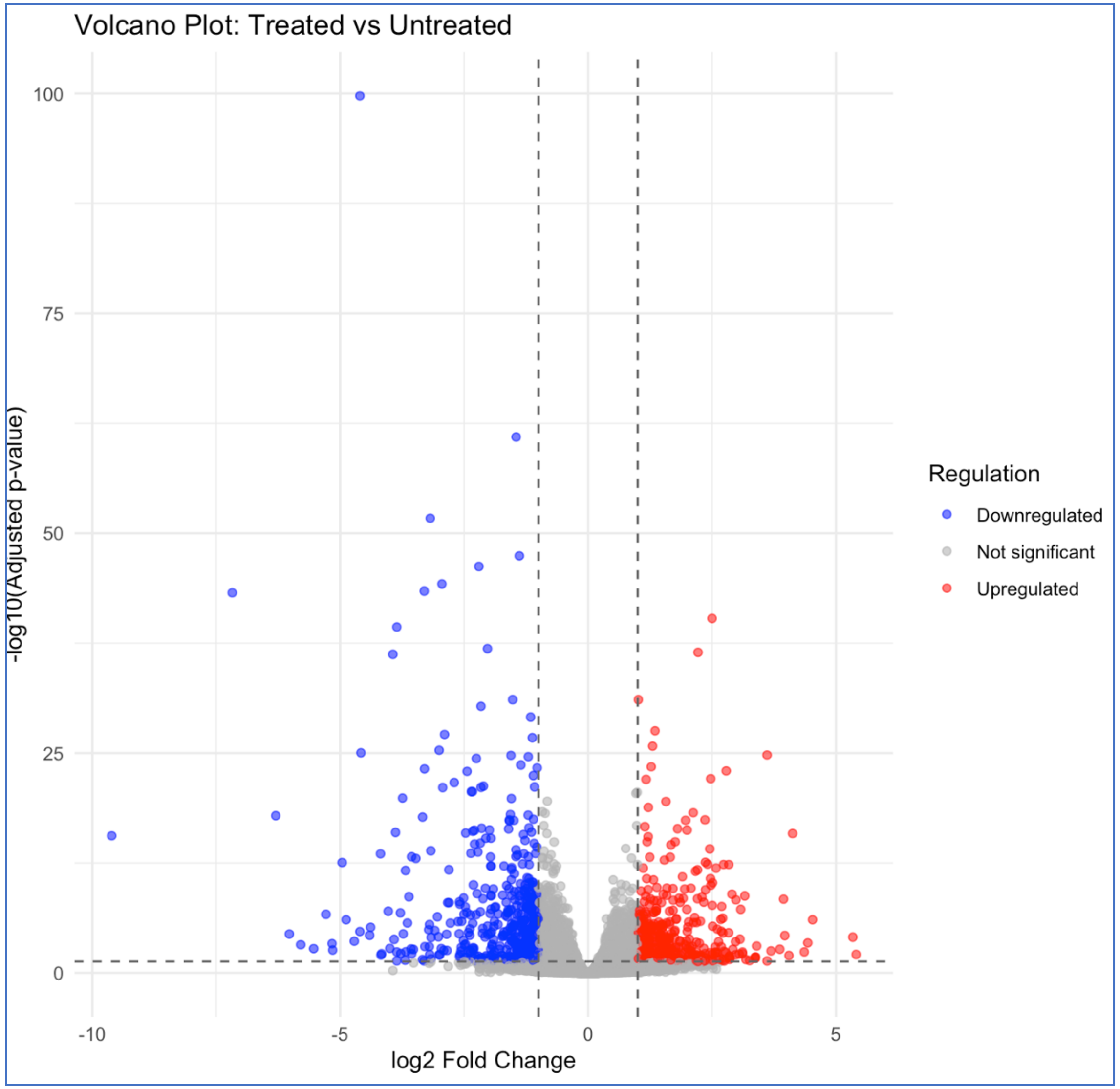


```
                 baseMean log2FoldChange     lfcSE      stat      pvalue        padj
                <numeric>      <numeric> <numeric> <numeric>   <numeric>   <numeric>
ENSG00000002834 7168.8258      -0.398577 0.1023715  -3.89344 9.88332e-05 1.53324e-03
ENSG00000003096  377.9773       0.920204 0.1869736   4.92157 8.58511e-07 2.42853e-05
ENSG00000003402 2546.6142      -1.183425 0.1635592  -7.23545 4.63971e-13 5.56316e-11
ENSG00000003987   25.5043      -0.988022 0.3265152  -3.02596 2.47845e-03 2.19761e-02
ENSG00000004059 1225.3543      -0.369206 0.1041106  -3.54628 3.90706e-04 4.92290e-03
ENSG00000004487 1237.7999       0.298901 0.0829066   3.60527 3.11832e-04 4.05550e-03
```

Visualization of DGE Results

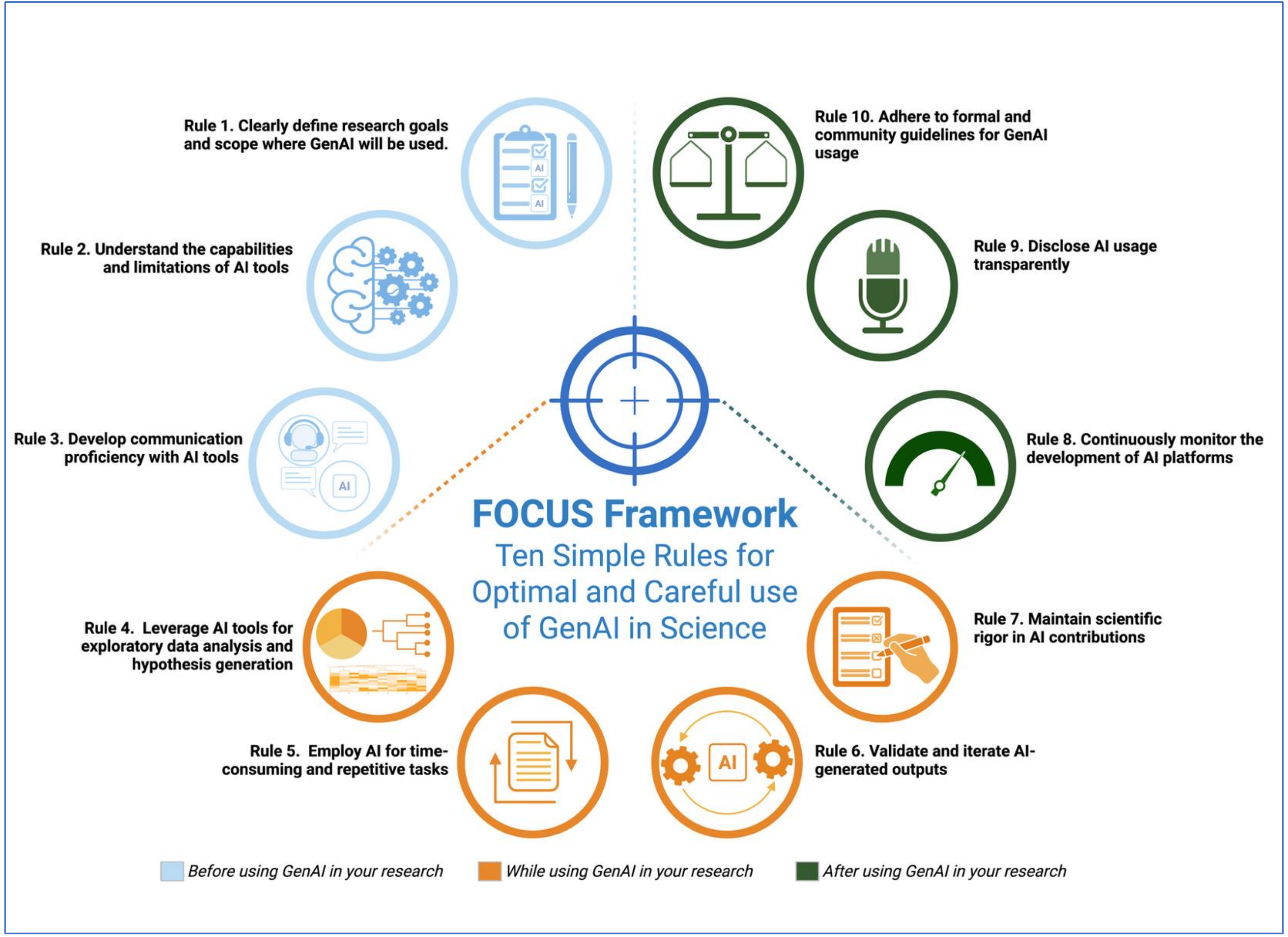


```
# Load required packages
library(DESeq2)
library(ggplot2)

# Run DESeq2 analysis
dds <- DESeqDataSet(airway, design = ~ dex)
dds <- DESeq(dds)
res <- results(dds)

# Convert to data frame for plotting
res_df <- as.data.frame(res)

# Remove rows with missing p-values or fold change (optional but helps avoid warnings)
res_df <- na.omit(res_df)

# Create a new column indicating regulation direction
res_df$Regulation <- "Not significant"
res_df$Regulation[res_df$log2FoldChange > 1 & res_df$padj < 0.05] <- "Upregulated"
res_df$Regulation[res_df$log2FoldChange < -1 & res_df$padj < 0.05] <- "Downregulated"

# Volcano plot
ggplot(res_df, aes(x = log2FoldChange, y = -log10(padj), color = Regulation)) +
  geom_point(alpha = 0.6, size = 1.5) +
  geom_vline(xintercept = c(-1, 1), linetype = "dashed", color = "gray40") +
  geom_hline(yintercept = -log10(0.05), linetype = "dashed", color = "gray40") +
  scale_color_manual(values = c("Upregulated" = "red",
                                "Downregulated" = "blue",
                                "Not significant" = "gray70")) +
  labs(title = "Volcano Plot: Treated vs Untreated",
       x = "log2 Fold Change",
       y = "-log10(Adjusted p-value)",
       color = "Regulation") +
  theme_minimal()
```



AI-Assisted Coding (ChatGPT, Gemini, DeepSeek)

- **The FOCUS Framework for Optimal and Carful Use of GenAI in Science**



Rule 1. Clearly define research goals and scope where GenAI will be used.

Rule 2. Understand the capabilities and limitations of AI tools

Rule 3. Develop communication proficiency with AI tools

Rule 4. Leverage AI tools for exploratory data analysis and hypothesis generation

Rule 5. Employ AI for time-consuming and repetitive tasks

Rule 6. Validate and iterate AI-generated outputs

Rule 7. Maintain scientific rigor in AI contributions

Rule 8. Continuously monitor the development of AI platforms

Rule 9. Disclose AI usage transparently

Rule 10. Adhere to formal and community guidelines for GenAI usage

FOCUS Framework

Ten Simple Rules for Optimal and Careful use of GenAI in Science

🟦 *Before using GenAI in your research* 🟧 *While using GenAI in your research* 🟩 *After using GenAI in your research*

Helmy M, *et al. PLOS Computational Biology*, 2025

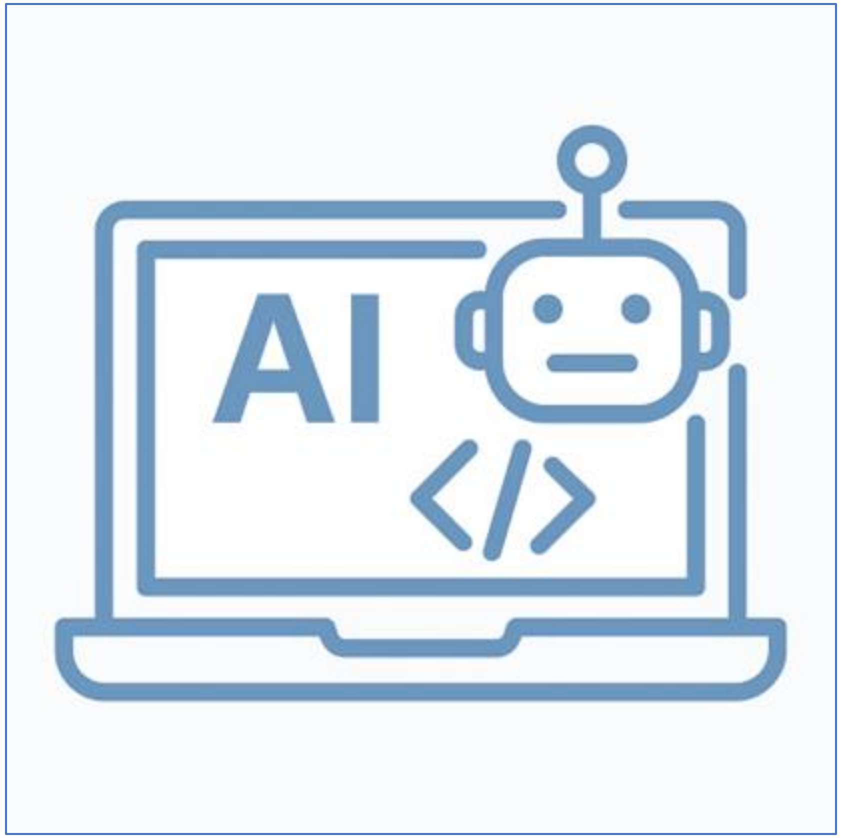
Introduction to AI-Assisted Coding

Why do we do sample clustering?

- **Tools like GitHub Copilot, ChatGPT, Codeium, or RStudio’s AI Assist can help:**
  - Write repetitive or boilerplate code.
  - Suggest functions or syntax corrections.
  - Generate documentation and comments.

- **Ideal for debugging and learning**

- **<u>A tool for understanding , nor a substitute for it</u>**



Demo: AI-Assisted Coding

- **Writing code**
  - Example: “*Write R code that filters DESeq2 results to significant genes and plots the top 10 with largest fold change*”

- **Debugging code**
  - Example: “When I run this code I got the following error

  dds <- DESeqDataSet(airway, design = ~ dex)

  Error: object 'airway' not found”

- **Explaining code**
  - Explain this R code for me “top10 <- head(order(geneVars, decreasing=TRUE), 10)”.

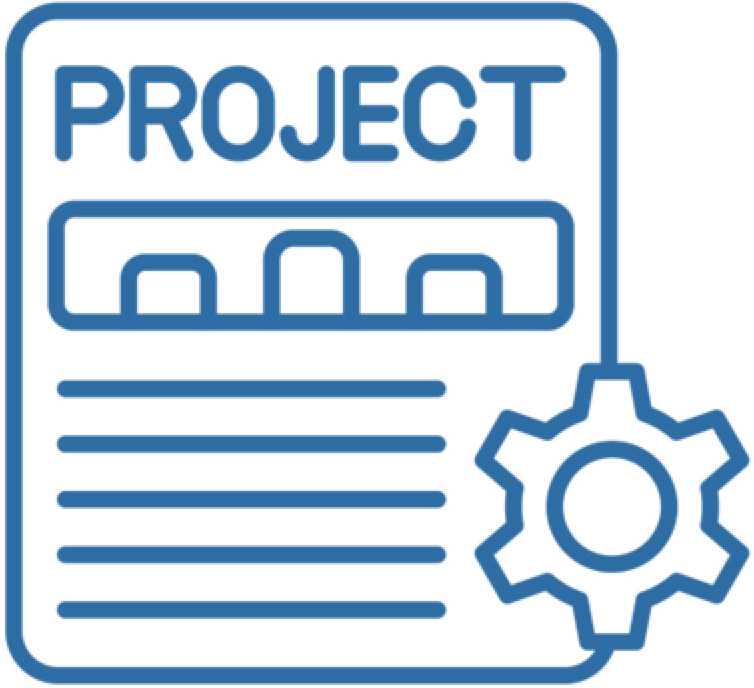
It is always useful to give your AI tools a context so that it gives you a better and more relevant code or explanation.

Capstone Mini Project

- **Goal:**
  - Work in teams to explore the airway dataset and generate insights from gene expression data.

- **Tasks:**
  - Identify top differentially expressed genes.
  - Visualize the results using:
    - PCA plot or clustering
    - Volcano plot or heatmap
  - Annotate selected significant genes with org.Hs.eg.db.
  - Summarize findings in a short presentation (2 slide).

- **Feel free to use AI assistance in the project**
  - Do not do the whole project using AI assistance
  - Keep it for debugging and explaining the errors



https://www.vecteezy.com/free-vector/project-icon

Tips and Suggested Workflow

1. **Load and preprocess data (*airway*, *DESeq2*, *vst*).**

2. **Perform DGE analysis.**

3. **Select and visualize top genes.**

4. **Annotate genes with biological names.**

5. **Interpret the biological relevance of results (treated vs untreated).**
   - How many genes were significantly differentially expressed?
   - Do treated and untreated samples separate clearly in PCA and heatmap?
   - What are some key upregulated genes and their biological roles?

Project Evaluation Criteria

Teams will be evaluated on:

- **Code execution and organization (30%)**
- **Quality and clarity of visualizations (30%)**
- **Interpretation of results (30%)**
- **Team collaboration and presentation (10%)**

Closing and Key Takeaways

- **You can now:**
  - Work confidently with R and Bioconductor
  - Normalize, visualize, and interpret RNA-seq data
  - Use AI responsibly to enhance coding productivity

- **Next steps:**
  - Explore DESeq2, edgeR, and limma in depth.
  - Apply these skills to your own datasets!

THANK YOU

VIDO

VACCINE AND INFECTIOUS DISEASE ORGANIZATION

VIDO.ORG

UNIVERSITY OF SASKATCHEWAN