Differential Expression

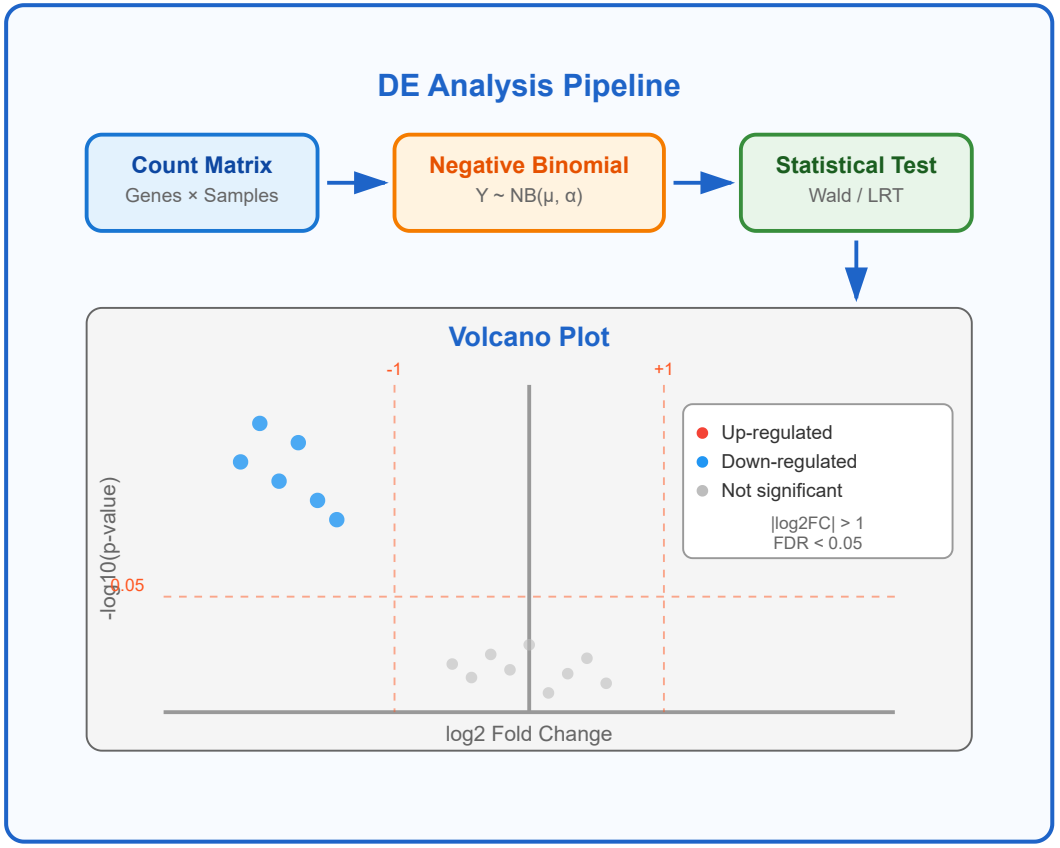
Statistical Models
Account for biological and technical variability

Negative Binomial
Models count data with overdispersion

Fold Change Thresholds
Typically |log2FC| > 1 for biological significance

FDR Control
False Discovery Rate < 0.05 for multiple testing

Volcano Plots
Visualize FC vs statistical significance

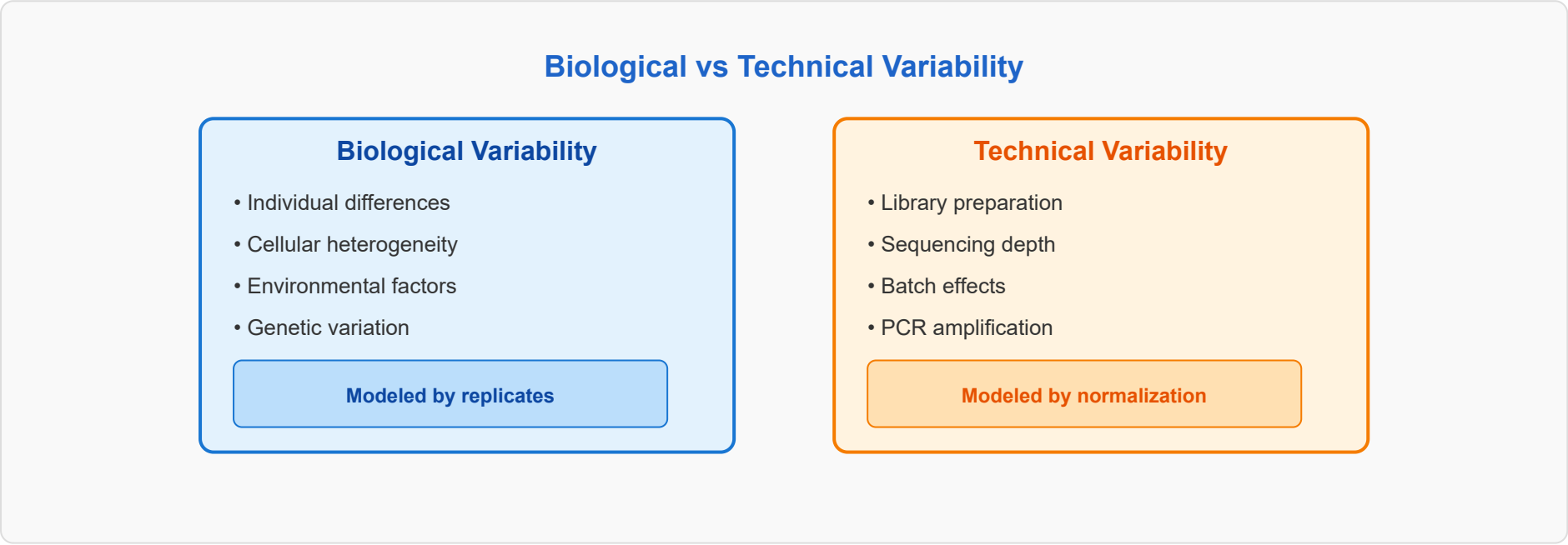


💡 Balance statistical significance with biological relevance

1. Statistical Models for Differential Expression

Statistical models form the foundation of differential expression analysis by providing a mathematical framework to distinguish true biological differences from random variation. In RNA-seq data, we encounter two major sources of variability that must be accounted for.

Sources of Variation



Why Standard Tests Fail

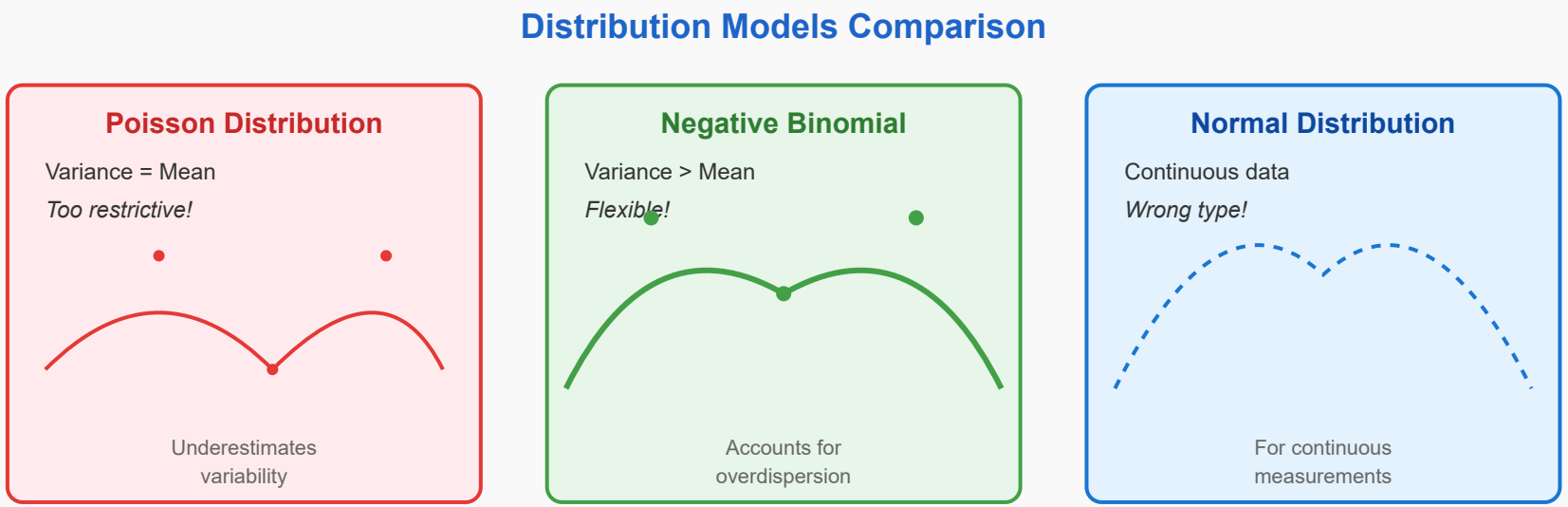
Simple t-tests or ANOVA are inadequate for RNA-seq data because they assume normally distributed data with constant variance. RNA-seq count data violates these assumptions in two critical ways:

Characteristics of RNA-seq count data:

- **Discrete counts** not continuous measurements
- **Mean-variance relationship** variance increases with expression level

- **Overdispersion** variance exceeds what Poisson distribution predicts
- **Zero inflation** many genes have zero counts in some samples

Model Comparison



Modern differential expression tools like DESeq2 and edgeR use **generalized linear models (GLMs)** with negative binomial distributions to properly model RNA-seq count data while accounting for both biological and technical variation.

2. Negative Binomial Distribution

The negative binomial distribution is the cornerstone of modern RNA-seq differential expression analysis. It extends the Poisson distribution by adding a dispersion parameter, allowing the variance to exceed the mean, which is characteristic of biological count data.

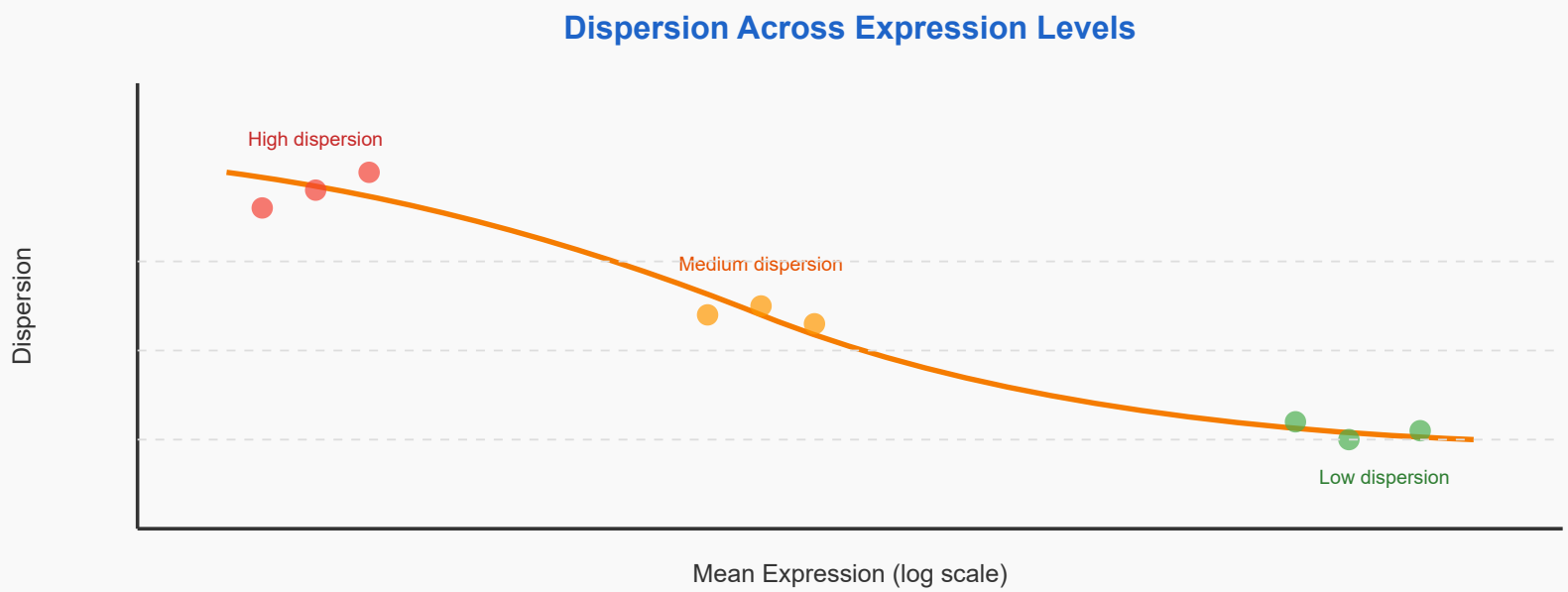
Negative Binomial Formula:

$$Y \sim NB(\mu, \alpha)$$

$$\text{Variance} = \mu + \alpha \times \mu^2$$

where μ = mean expression, α = dispersion parameter

Understanding Dispersion

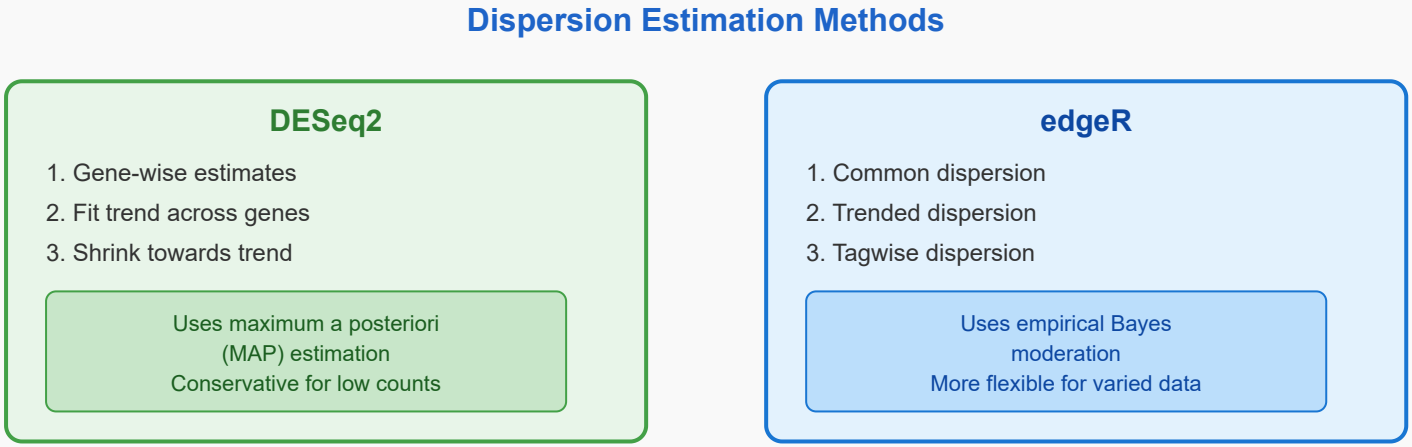


Key properties of dispersion:

- **Gene-specific** each gene has its own dispersion estimate

- **Expression-dependent** typically decreases with higher mean expression
- **Shrinkage estimation** improves accuracy by borrowing information across genes
- **Biological coefficient of variation** BCV = √α represents biological variability

DESeq2 vs edgeR Approaches



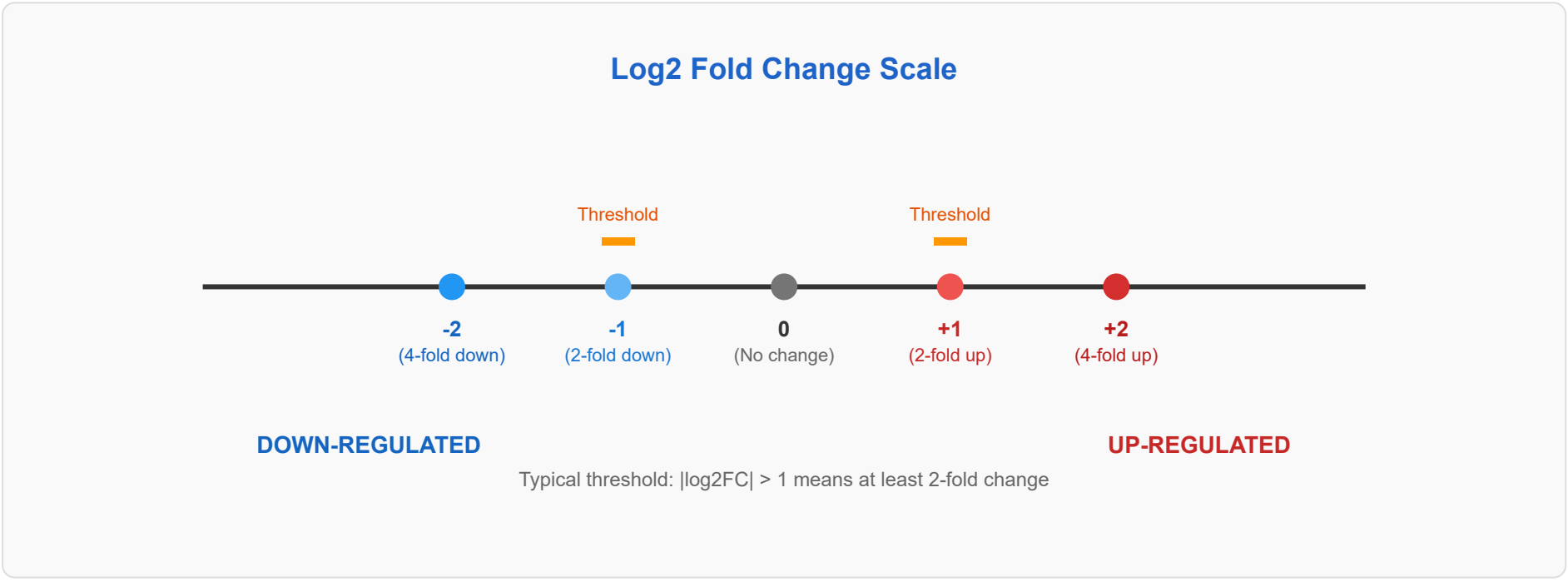

```
# Example R code for dispersion estimation in DESeq2 library(DESeq2) # Create DESeq2 object dds <-
DESeqDataSetFromMatrix( countData = counts, colData = metadata, design = ~ condition ) # Estimate size factors
(normalization) dds <- estimateSizeFactors(dds) # Estimate dispersions dds <- estimateDispersions(dds) # Plot
dispersion estimates plotDispEsts(dds)
```

Both methods share information across genes to improve dispersion estimates, especially for genes with low counts or few replicates. This **shrinkage approach** balances gene-specific estimates with overall trends, reducing false positives while maintaining statistical power.

3. Fold Change Thresholds

While statistical significance tells us that a difference is unlikely to be due to chance, fold change tells us whether that difference is **biologically meaningful**. A gene might be statistically significantly different but only change by 5%, which may not be biologically relevant.

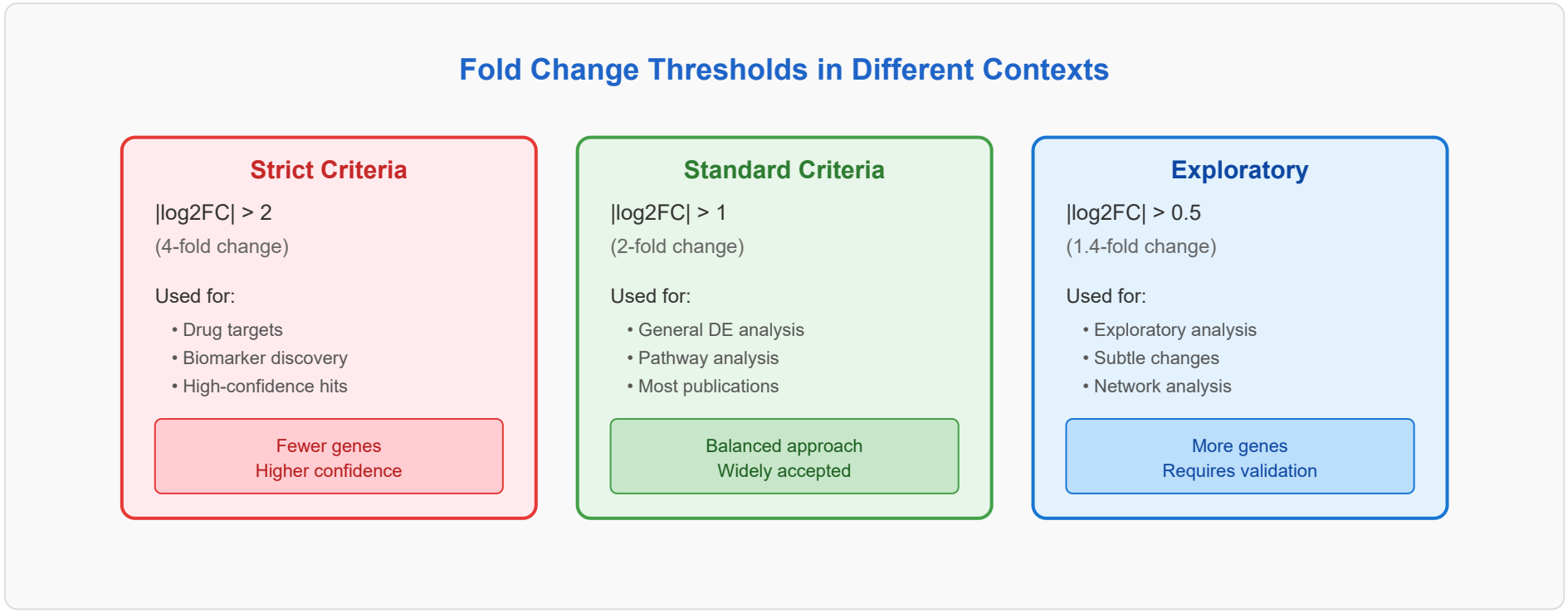
Understanding Log2 Fold Change



Why use log2 fold change?

- **Symmetry** up and down-regulation are symmetric (2-fold up = +1, 2-fold down = -1)
- **Interpretability** each unit represents a doubling or halving
- **Statistical properties** more normally distributed than raw fold changes
- **Easy comparison** magnitude directly comparable across genes

Common Thresholds by Context



Fold Change Calculations:

$$\text{Fold Change (FC)} = \text{Expression}_{\text{Treated}} / \text{Expression}_{\text{Control}}$$

$$\text{Log2 Fold Change} = \log_2(\text{Expression}_{\text{Treated}}) - \log_2(\text{Expression}_{\text{Control}})$$

Example: Gene with 100 counts in control, 400 in treated

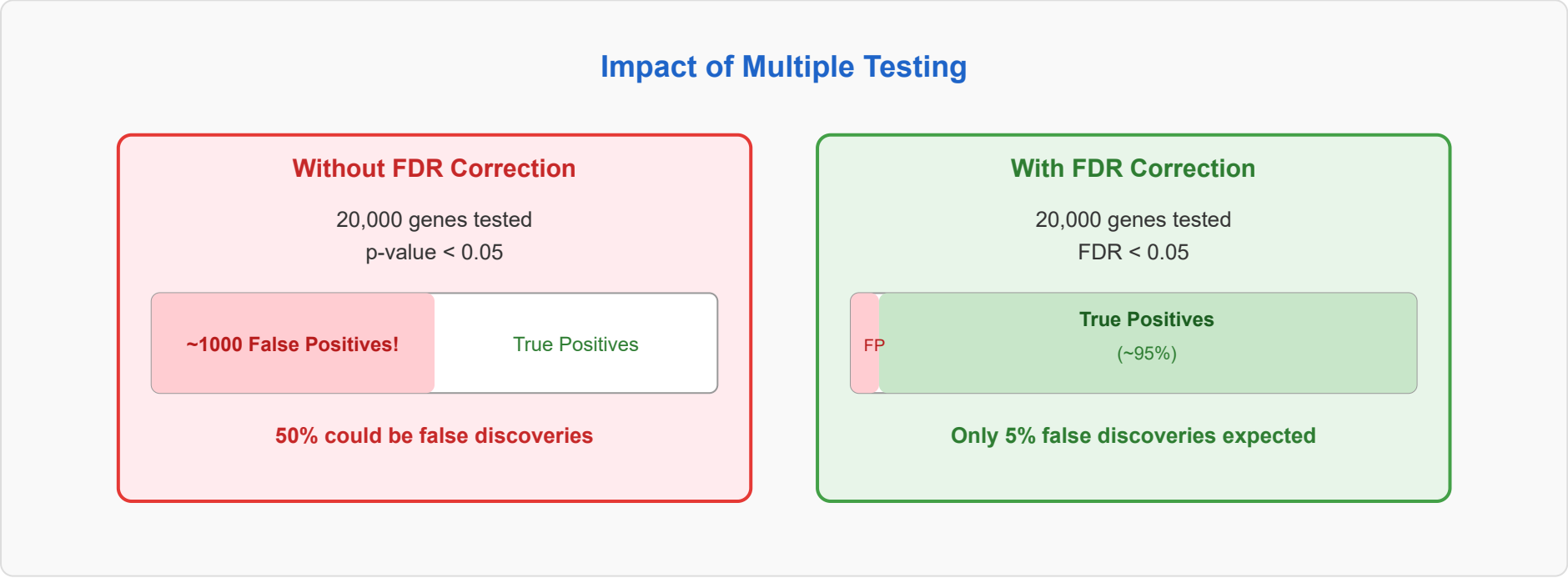
$$FC = 400/100 = 4 \rightarrow \text{log2FC} = \log_2(4) = 2$$

The choice of fold change threshold depends on your experimental context, validation capabilities, and downstream applications. Always consider both statistical significance and biological magnitude when interpreting results.

4. False Discovery Rate (FDR) Control

In RNA-seq experiments, we test thousands of genes simultaneously. This **multiple testing problem** means that even with a p-value threshold of 0.05, we would expect 5% of genes to appear significant by chance alone. For 20,000 genes, that's 1,000 false positives!

The Multiple Testing Problem



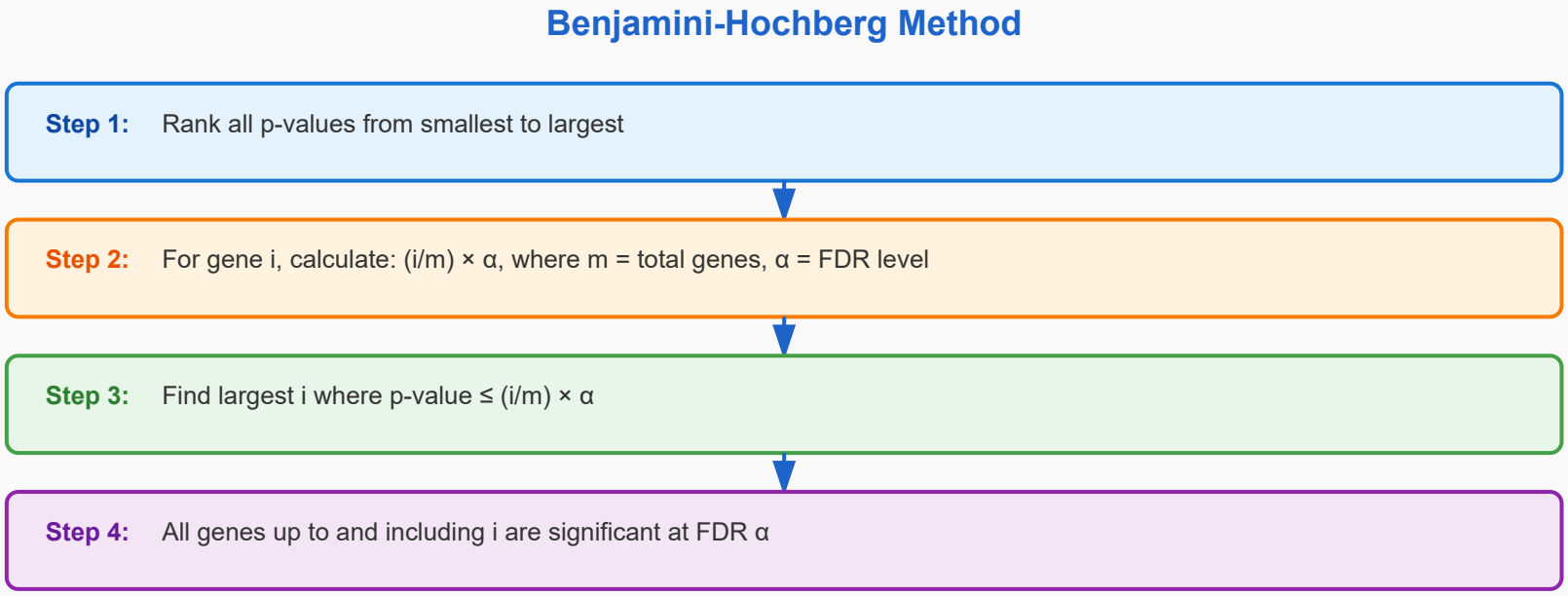
FDR vs P-value

Understanding the difference:

- **P-value** probability of observing data if null hypothesis is true (for a single test)
- **FDR (q-value)** expected proportion of false discoveries among all discoveries
- **Adjusted p-value** corrected p-value accounting for multiple comparisons

- **Benjamini-Hochberg** most common FDR control method in RNA-seq

Benjamini-Hochberg Procedure



Example: With 20,000 genes and α=0.05, the 100th gene needs p < (100/20000)×0.05 = 0.00025

```
# Example R code for FDR correction library(DESeq2) # Run differential expression analysis dds <- DESeq(dds) results
<- results(dds) # Results include both p-value and adjusted p-value (FDR) # padj is the Benjamini-Hochberg adjusted
p-value # Filter for significant genes significant_genes <- results[results$padj < 0.05 &
abs(results$log2FoldChange) > 1, ] # Count significant genes nrow(significant_genes)
```

FDR Interpretation:

FDR = 0.05 means:
"Among all genes called significant,
we expect about 5% to be false discoveries"

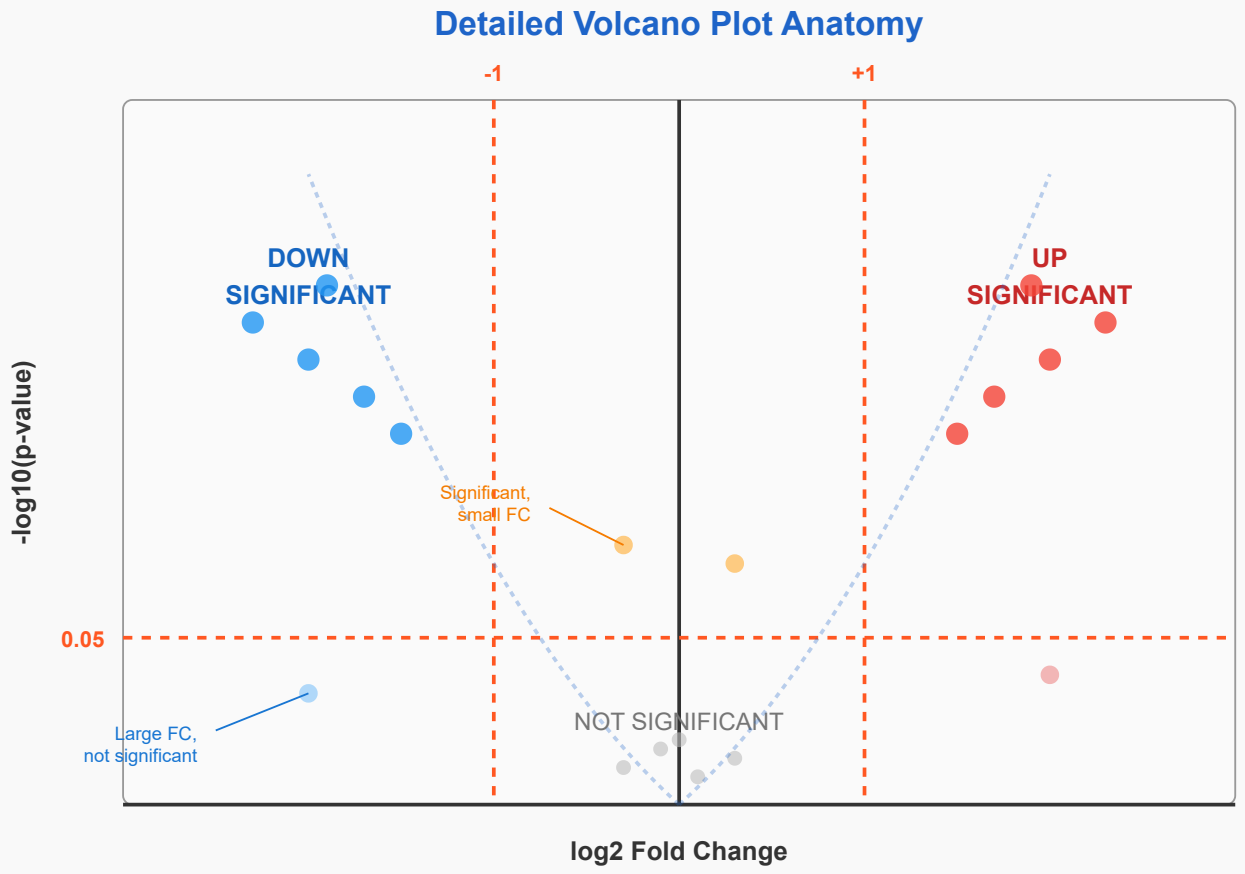
If 1000 genes are significant at FDR < 0.05,
expect ~50 false positives and ~950 true positives

FDR control is essential for reliable differential expression analysis. It provides a more interpretable measure than individual p-values when testing thousands of genes simultaneously, ensuring that your gene list has a predictable proportion of true discoveries.

5. Volcano Plots

Volcano plots are the most widely used visualization for differential expression results. They simultaneously display the **magnitude of change** (fold change) and **statistical confidence** (p-value) for all genes, making it easy to identify the most interesting candidates.

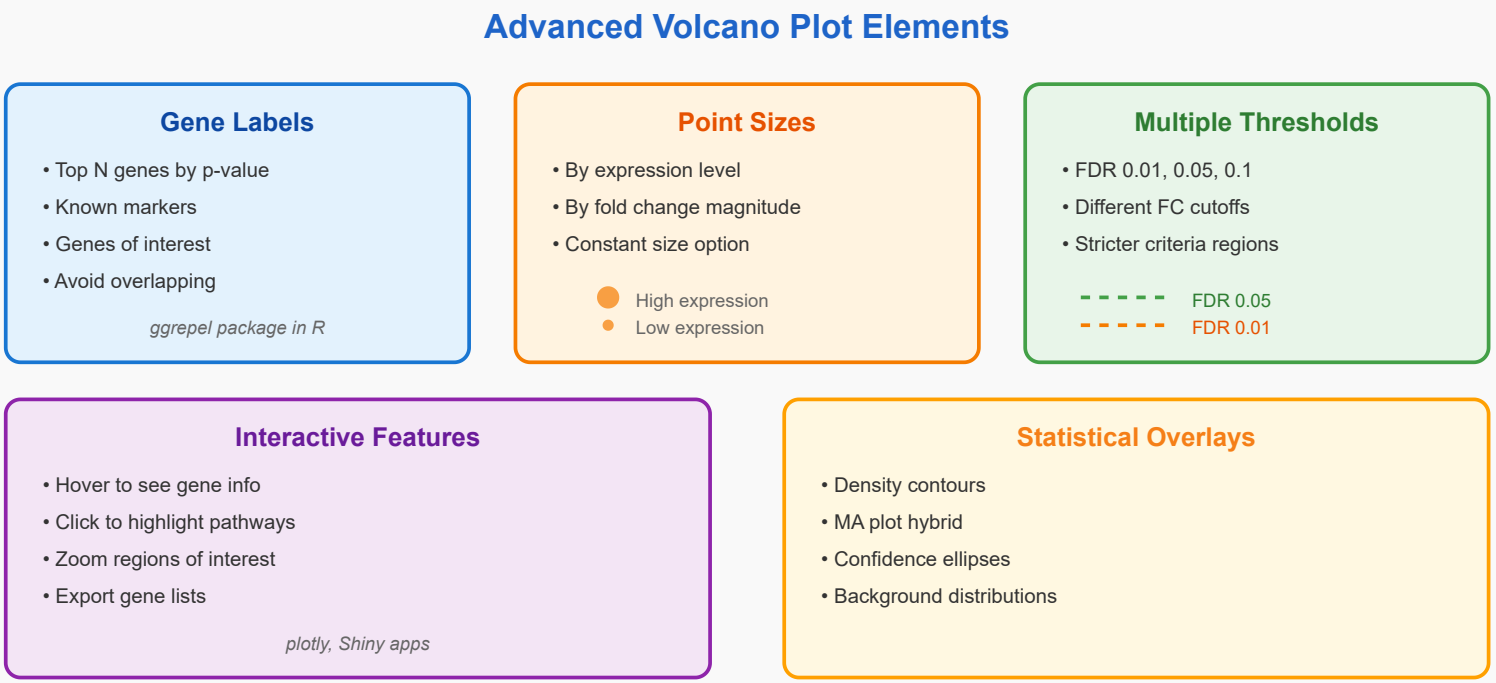
Anatomy of a Volcano Plot



Reading a volcano plot:

- **X-axis** log2 fold change (biological magnitude)
- **Y-axis** -log10(p-value) (statistical significance)
- **Top corners** most interesting genes (large change + significant)
- **Color coding** typically red/up, blue/down, grey/not significant

Enhanced Volcano Plot Features




```
# Creating an enhanced volcano plot in R library(ggplot2) library(ggrepel) # Prepare data volcano_data <- data.frame( gene = rownames(results), log2FC = results$log2FoldChange, pvalue = results$pvalue, padj = results$padj ) # Add significance category volcano_data$significance <- "Not Sig" volcano_data$significance[volcano_data$padj < 0.05 & volcano_data$log2FC > 1] <- "Up" volcano_data$significance[volcano_data$padj < 0.05 & volcano_data$log2FC < -1] <- "Down" # Create plot ggplot(volcano_data, aes(x = log2FC, y = -log10(pvalue))) + geom_point(aes(color = significance), alpha = 0.6, size = 2) + scale_color_manual(values = c("Up" = "#F44336", "Down" = "#2196F3", "Not Sig" = "#BDBDBD")) + geom_vline(xintercept = c(-1, 1), linetype = "dashed", color = "#FF5722") + geom_hline(yintercept = -log10(0.05), linetype = "dashed", color = "#FF5722") + labs(title = "Volcano Plot", x = "log2 Fold Change", y = "-log10(p-value)") + theme_minimal() + geom_text_repel(data = subset(volcano_data, padj < 0.001), aes(label = gene), size = 3, max.overlaps = 10)
```

Common Volcano Plot Patterns

Different experimental scenarios produce characteristic volcano plot shapes. A balanced experiment shows genes distributed symmetrically, while strong biological effects create distinct up and down-regulated clusters in the top corners. Technical issues or batch effects often manifest as asymmetric patterns or unexpected clustering.

Interpreting patterns:

- **Symmetric volcano** balanced regulation, good experiment
- **Skewed distribution** may indicate batch effects or normalization issues
- **Few significant genes** weak biological effect or insufficient power
- **Many significant genes** strong biological effect or technical artifact

Volcano plots are invaluable for quality control and hypothesis generation. They allow you to quickly assess the overall experimental results, identify candidate genes for follow-up studies, and communicate findings effectively in presentations and publications.

💡 **Complete Analysis: Combine all five components for robust differential expression analysis**