# Module 6: Differential Gene Analysis (Part 1)

Key Concepts

- Differential gene analysis

- Read count normalization

- Read count transformation

- Gene filtering strategies

Differential gene analysis

- The aim of a differential gene analysis is to identify genes with difference in expression patterns under different experimental conditions (gene(s) versus condition(s) and vice versa) for a set of samples.
- By asking "which genes change and by how much?" and then mapping those changes to **pathways, processes, and regulatory mechanisms**, we can generate hypotheses about the underlying biological mechanisms.
  - **Disease mechanisms** (e.g., identifying driver genes in cancer)
  - **Therapeutic targets** or potential drug mechanisms
  - **Biomarker discovery** for diagnosis and prognosis
  - **Fundamental biology** of development, stress responses, or adaptation.

a
HCC tissues *vs* normal tissues

b
III-IV HCC tissues *vs* I-II HCC tissues

Zeng *et al.* 2022

# Which genes are up- or down-regulated in a given condition?
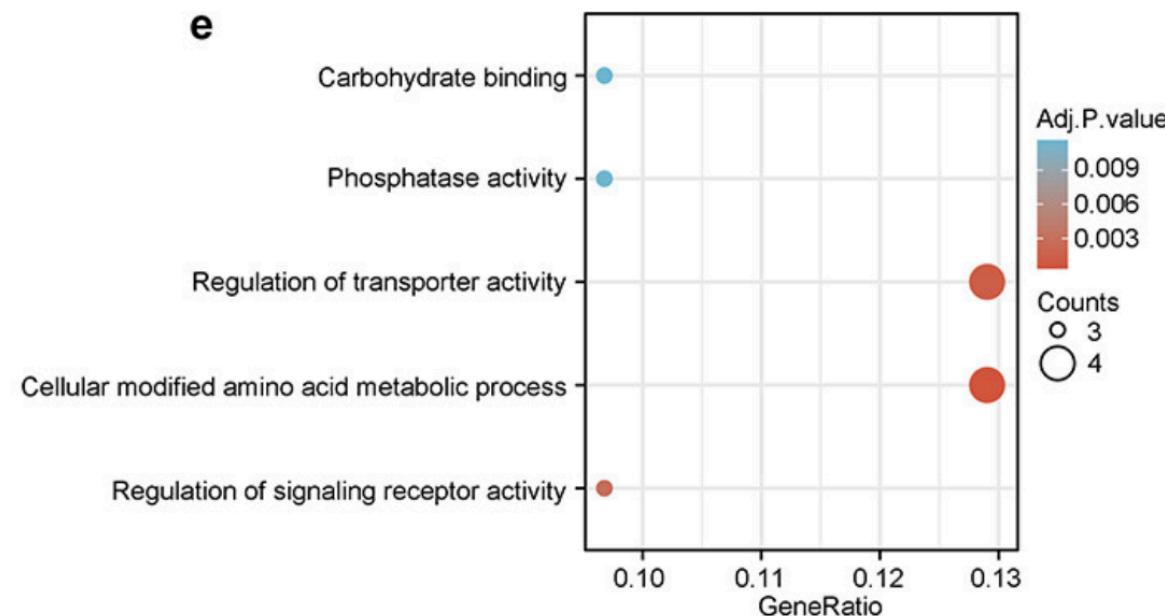
Question: Under condition A vs. condition B (e.g., treated vs. untreated), which genes show significant changes in expression levels?
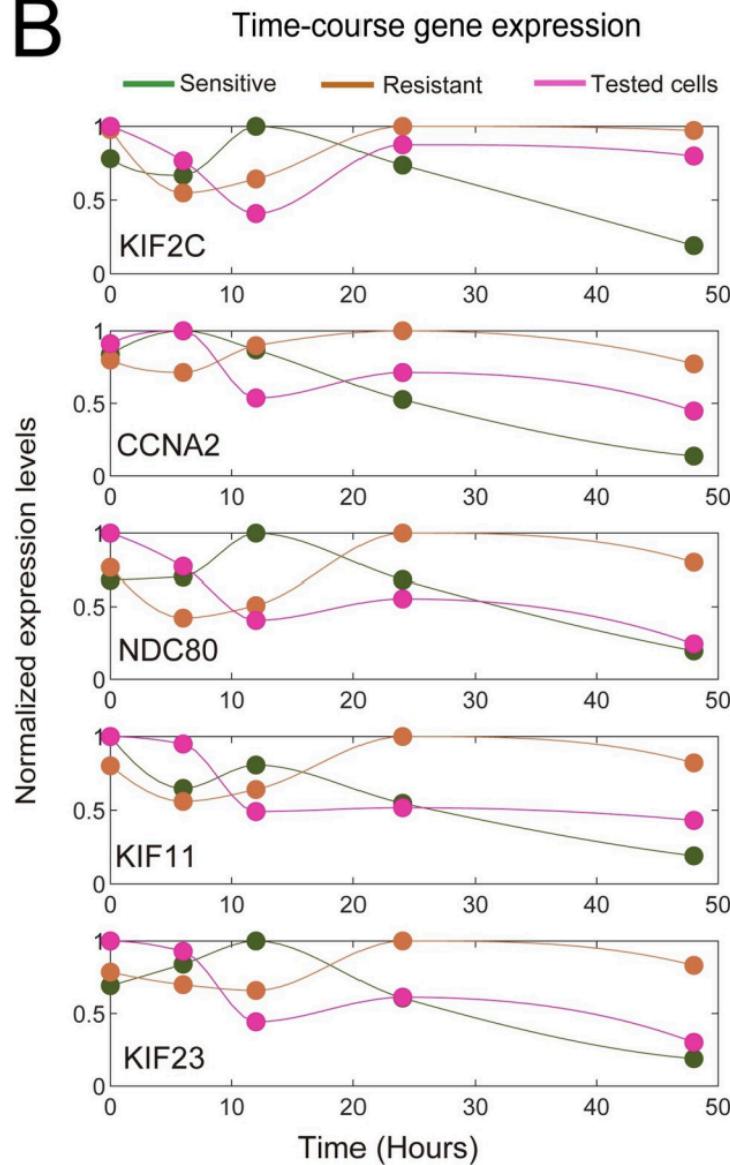
- The treatment can be drug exposure, knock-out mutation, or any external stimulus
- Identifying up-regulated genes can point toward processes or pathways that are activated under specific conditions, while down-regulated genes might indicate suppression or reduced activity of certain pathways.

What pathways or biological processes are affected?

Question: Are the significantly differentially expressed genes enriched in specific functional categories, pathways, or Gene Ontology terms?

- DGE results often feed into **gene set enrichment analysis (GSEA)** or pathway analysis (gProfiler) to reveal broader biological programs at play—like inflammation, cell cycle regulation, or metabolic shifts.

B

Time-course gene expression

Sensitive — Resistant — Tested cells

KIF2C

CCNA2

NDC80

KIF11

KIF23

Normalized expression levels

Time (Hours)

Zhang *et al.* 2019
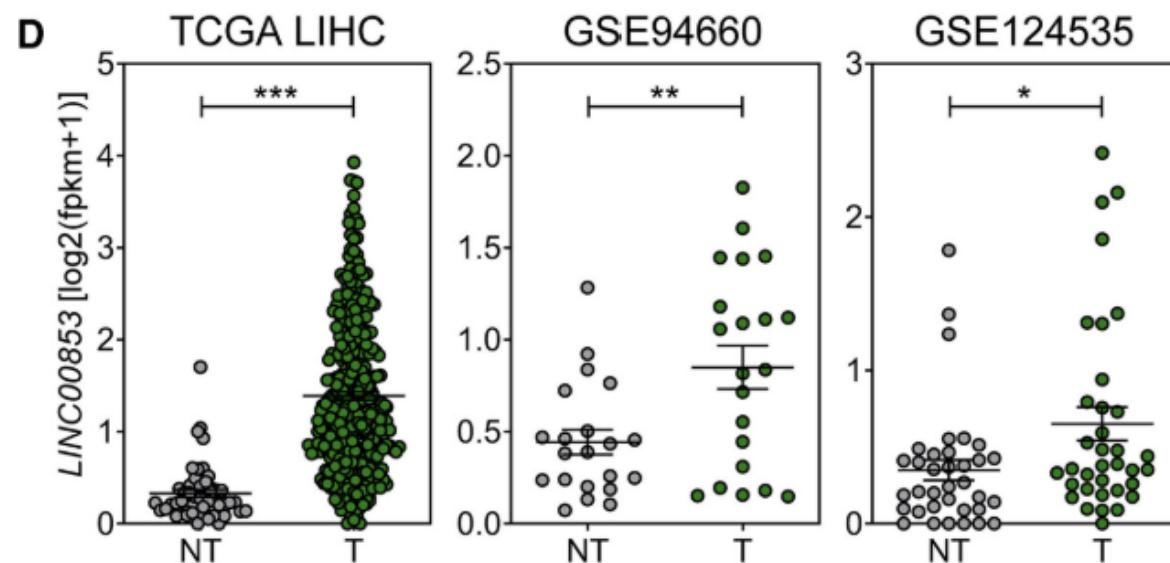
How does gene expression change over time (time-course experiments)?

- **Question**: What genes are differentially expressed at early vs. late stages of a response, disease progression, or development?
  - Time-course DGE studies can reveal **dynamic gene expression trends** that can help identify early-response genes, late-stage markers, or transitions in biological processes.

# Are there biomarker candidates for disease or treatment response?
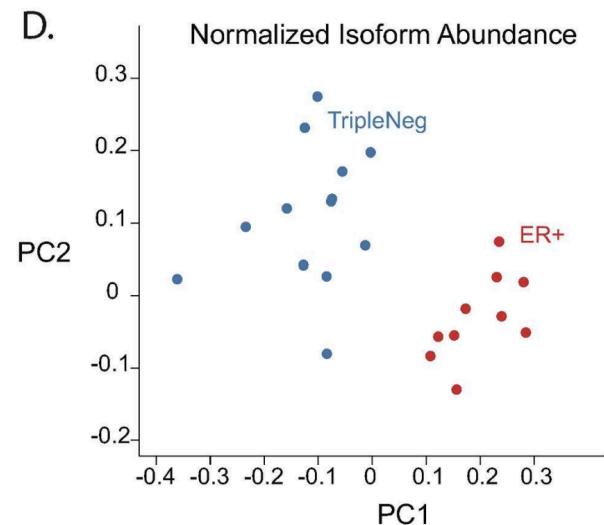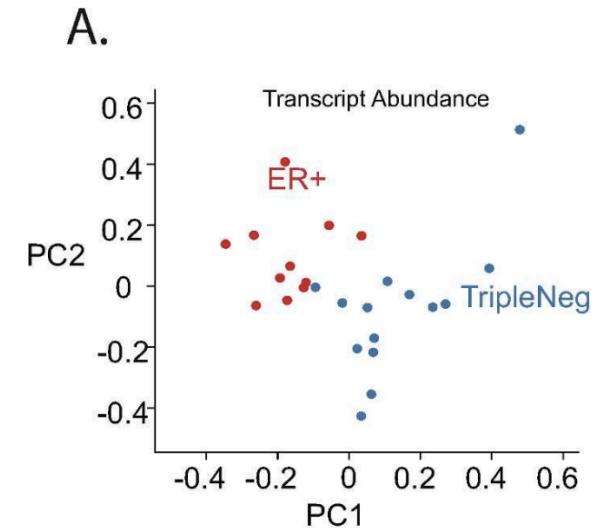
- **Question**: Which genes consistently show differential expression in a disease state (e.g., tumor vs. normal tissue) or in response to a therapy?
  - Such genes may serve as **potential biomarkers** for diagnostic, prognostic, or therapeutic monitoring purposes (e.g., predicting drug resistance or disease recurrence).

LINC00853 expression in the nontumor and the HCC cohorts in three HCC RNA-Seq datasets (TCGA_LIHC, GSE94660, and GSE124535).

Can we classify or subtype biological samples based on gene expression profiles?

- **Question**: In a cohort of samples (e.g., different cancer subtypes), which sets of genes distinguish one subtype from another?

  - Identifying **signature genes** for subtypes can guide more personalized treatment strategies

- **Example**: Different subtypes of breast cancer have very different expression profiles. These expression profiles are associated with different patient outcomes and with different treatment protocols. Different subtypes of breast cancer can be distinguished by using differential transcript isoform expression (Stricker *et al.* 2017).

A. Transcript Abundance

D. Normalized Isoform Abundance

DGE Tools

| Tool	Statistical Model	Key Strengths	When to Use
DESeq2	- Negative binomial with empirical Bayesian shrinkage of dispersion and fold changes - Based on the hypothesis that most genes are not DE	- Better handling of overdispersion - Good documentation and built-in plotting (MA plots, PCA) - Shrinkage of fold changes helps stabilize LFC for low-count genes	- Moderate sample sizes (≥3 replicates/condition) - Typical RNA-seq designs (two-group or multi-factor) - Ideal if you want an easy-to-follow pipeline with built-in diagnostics
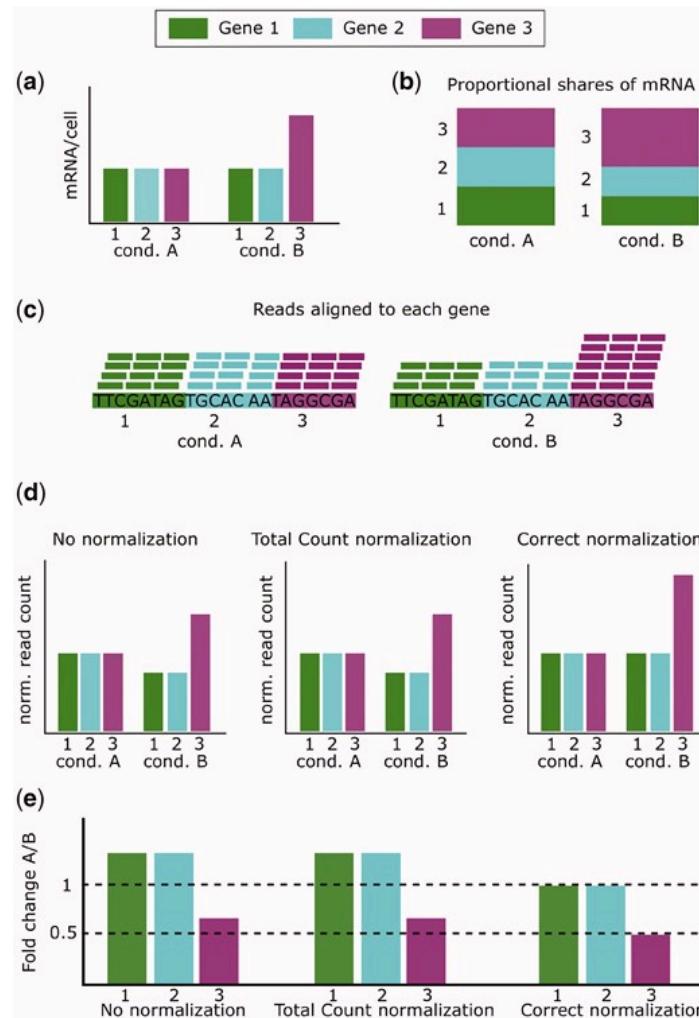
DGE Tools

| Tool	Statistical Model	Key Strengths	When to Use
edgeR	- Negative binomial with empirical Bayes estimation of dispersion - Based on the hypothesis that most genes are not DE	- Highly flexible for complex or unbalanced designs - Known to perform robustly even with smaller sample sizes - Deep configurability for normalization, dispersion, and model setup	- Complex experimental designs (time series, multi-factor) - Smaller datasets (though more replicates is still better) - Prefer a "light-weight" approach with fine-grained control

DGE Tools
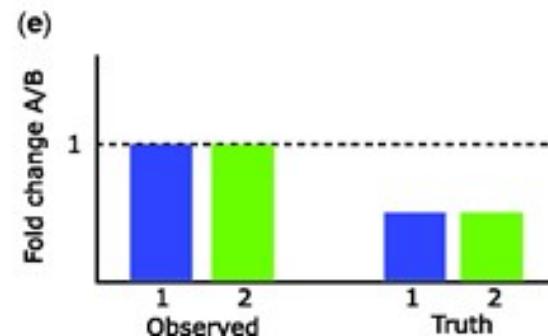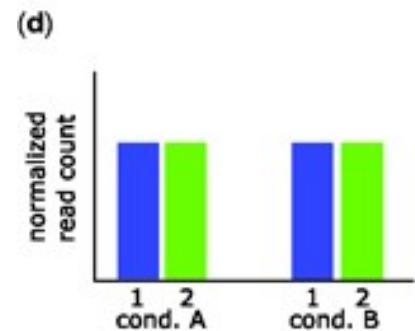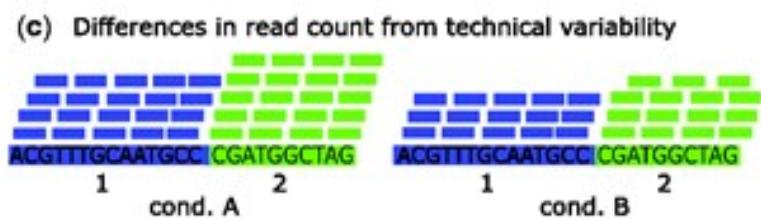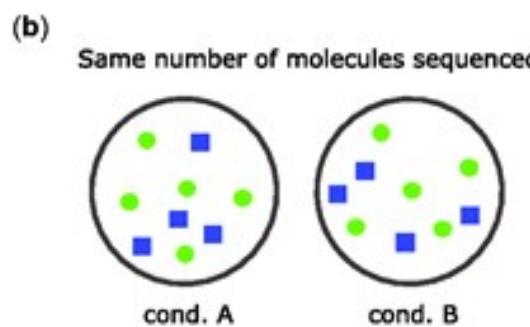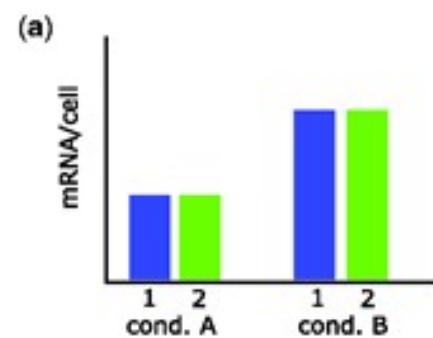
Tool	Statistical Model	Key Strengths	When to Use
limma (with **voom**)	Linear models with empirical Bayes shrinkage of variance estimates (after voom transformation)	- Stable release (works with microarray data) - Good for large datasets and complex designs - Fast and memory-efficient; integrates well with other limma functions (e.g., contrasts, gene set testing)	- Already familiar with limma or microarray workflows - Large sample sizes or multi-factor, repeated-measures designs - Looking for flexible linear modeling framework and advanced downstream analyses (e.g., GSEA, contrasts)

Read Count Normalization: Why do we need to do this?[2]

- The correct normalization method to use depends on which assumptions are valid for the biological experiment.

- Incorrect normalization leads to problems in downstream analysis, such as inflated false positives, that mean results cannot be trusted.

- **Example**: A small number of highly expressed genes creates the appearance that non-DE are DE, but the false DE calls may be corrected by normalizing read counts so that the expression levels of non-DE genes are equivalent.
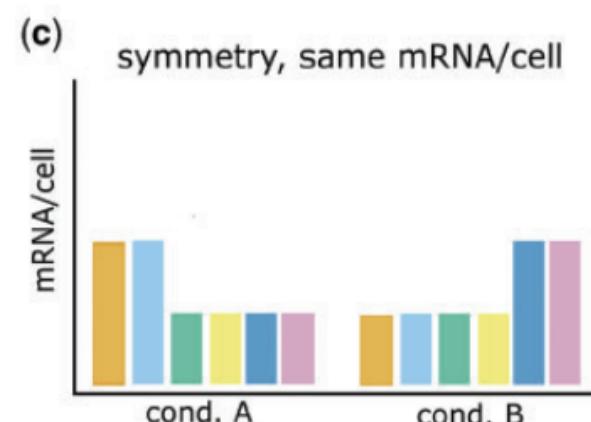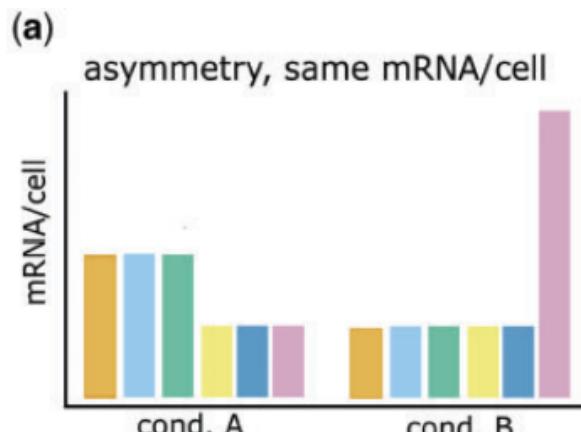
Read Count Normalization: Why do we need to do this?[2]

- In the case of a global shift in expression, it may appear that DE genes are non-DE or that up-regulated genes are down-regulated.

- It is necessary to take into account the differences in overall expression between conditions.

Read Count Normalization: Normalization by Library Size[2]

|	Details
Use Case	When the total mRNA remains equal or within the same range across different treatments/conditions, even if there is some asymmetry.
Assumptions	- Total mRNA/cell is consistent across conditions. - Differences in sequencing depth are the primary technical variability.
Limitations	- Cannot handle global shifts in expression (large asymmetry). - Assumes minimal variability in the total mRNA/cell across samples.
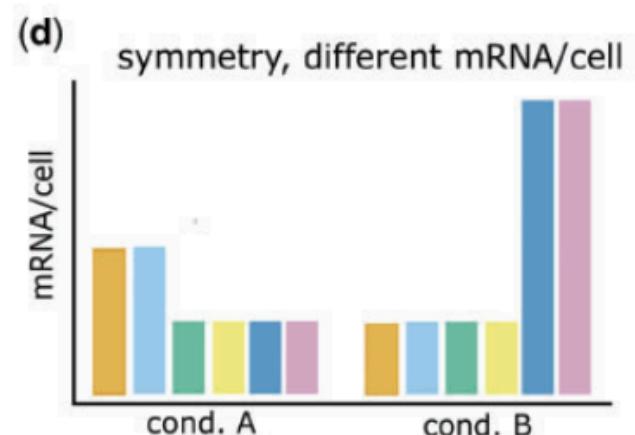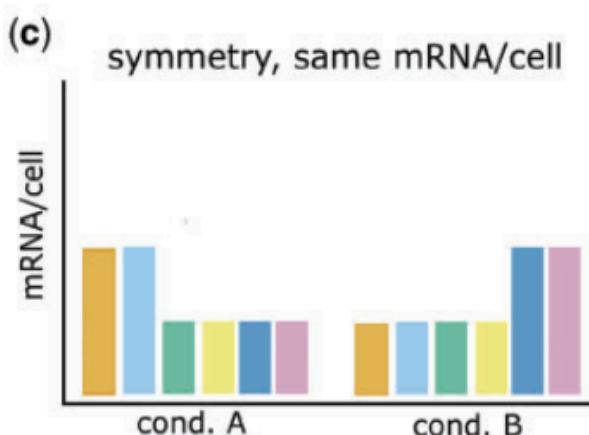
Read Count Normalization: Normalization by Library Size

Metric	Name	Calculation	Purpose
RPKM	Reads Per Kilobase (kb) per Million	(Read count × 10^6) / (Total reads × Gene length in kb)	Normalizes for sequencing depth and gene length (often for single-end data).
FPKM	Fragments Per Kilobase per Million	Same as RPKM but for paired-end reads	Accounts for paired-end sequencing data.
TPM	Transcripts Per Million	(Read count / Gene length) ÷ (Sum of all RPK values) × 10^6	Normalizes within a sample to allow better comparability of expression levels across genes.
CPM	Counts Per Million	(Read count × 10^6) / (Total mapped reads)	Normalizes only for sequencing depth (not gene length).

Read Count Normalization: Normalization by Distribution[2]

|	Details
Use Case	When there is symmetry in sample sizes, even if total mRNA levels differ across conditions.
Assumptions	- DE & non-DE genes behave similarly (technical effects are the same). - Balanced expression: approximately equal numbers of up- and down-regulated genes across conditions.
Advantages	Can handle small asymmetry (e.g., a few highly expressed genes) or differences in total mRNA/cell.
Limitations	Fails to correct global shifts in expression (large asymmetry).
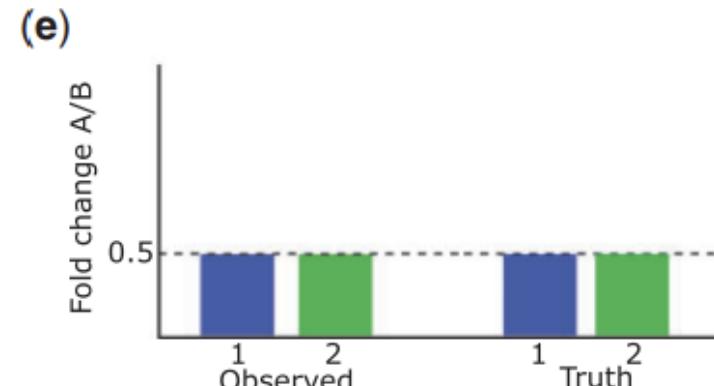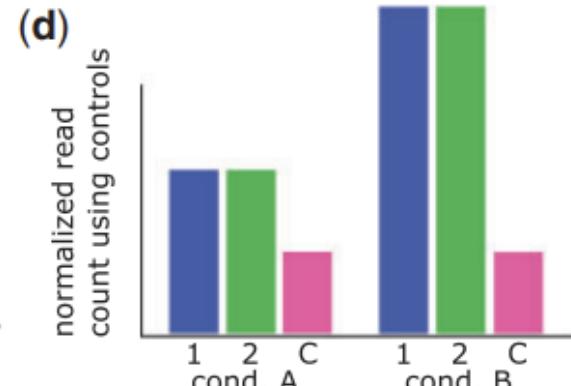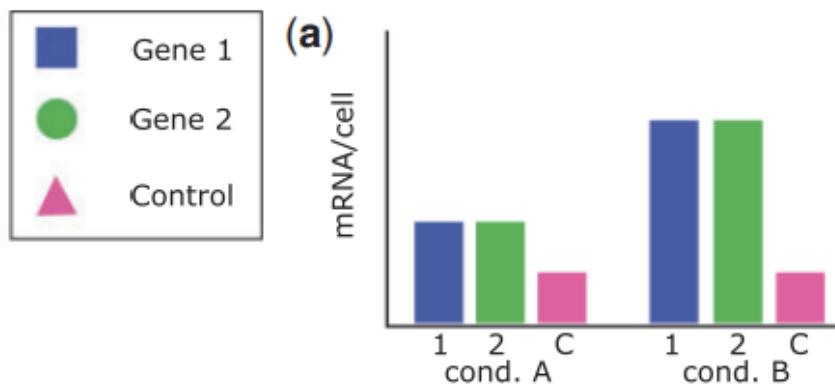
Read Count Normalization: Normalization by Distribution[2]

Method	Key Idea
Quantile Normalization	Forces the distribution of gene expression values (e.g., read counts or log-transformed counts) to match across samples.
Median/Upper Quartile	Normalizes counts based on a specified quantile (e.g., 50th or 75th percentile) rather than the total number of reads.
TMM (Trimmed Mean of M-values)	Trims extreme values (highly expressed genes) and calculates scale factors using average log-fold changes (M-values).
PoissonSeq / DEGES	Iteratively identifies non-differentially expressed (non-DE) genes and estimates scaling factors using those genes only.

Read Count Normalization: Normalization by Controls[2]

	Details
Use Case	When suitable controls exist that remain unaffected by biological conditions and behave consistently across samples.
Assumptions	- Controls exist and are non-DE in the context of the experiment. - Controls reflect technical effects for all genes.
Advantages	Corrects for variations in sequencing depth and technical effects using robust controls.
Limitations	- Requires reliable and representative controls. - Cannot correct global shifts in expression if controls do not reflect all genes.

Read Count Normalization: Normalization by Controls[2]

Method	Key Idea
Using Control Genes	Normalizes data based on a set of genes known (or assumed) to have **constant** expression across samples (non-DE).
Spike-in Controls	Uses known amounts of external RNA spikes (added to each sample) to determine and correct for technical biases.

Read Count Transformation: Why do we need to do this?

- **RNA-seq count data is highly skewed** (due to differences in sequencing depth, expression levels, and variance).

- **Standard log transformation (log2(count + 1)) is not sufficient** due to heteroscedasticity (high variance for highly expressed genes), so we use **VST** and **rlog** to help stabilize variance

Feature	Variance Stabilizing Transformation (VST)	Regularized Log Transformation (rlog)
Purpose	Stabilizes variance for high-count genes	Shrinks differences between low-count genes while preserving higher count relationships
Assumes Variance-Mean Relationship?	Yes, variance increases with mean expression	Yes, but applies stronger shrinkage for low counts

Feature	Variance Stabilizing Transformation (VST)	Regularized Log Transformation (rlog)
Dataset size	Optimized for large samples	Computationally expensive for large datasets
Computational Speed	Fast	Slower due to Bayesian shrinkage
Handles Low-Count Genes Well?	Partially, but less aggressive	Yes, applies stronger shrinkage
Normalization	Assumes normalized counts (DESeq2 size factors)	Assumes normalized counts (DESeq2 size factors)
Use Case	Large-scale RNA-seq datasets, differential expression analysis	Small RNA-seq datasets, noise reduction in low-count genes

Gene Filtering Strategies

- Expression Threshold (Minimum Counts or CPM/TPM)

- Sample Proportion

- Variance or Coefficient of Variation (CV) Filtering

- Filtering by Group-Specific Expression

Gene Filtering Strategies

Expression Threshold (Minimum Counts or CPM/TPM)

- Remove genes that do not meet a minimum read count or transcripts per million (TPM) / counts per million (CPM) threshold in a certain fraction of samples.

- **Example**: Genes must have a CPM ≥ 1 or 2 in at least X samples, where X could be the size of the smallest experimental group

- Genes with extremely low expression are more prone to noise and can inflate multiple testing corrections.

Gene Filtering Strategies

Sample Proportion

- Retain only those genes that are "detected" (above a given expression threshold) in a specific percentage (e.g., 50%, 75%, etc.) of the samples, or in all samples of at least one condition.

- If a gene is detected in only one or two samples, it may not be informative across biological replicates or conditions.

- **Example**: A gene must have a minimum of 1 read per million in at least 50% of all samples

Gene Filtering Strategies

Variance or Coefficient of Variation (CV) Filtering

- Filter out genes whose expression values show very little variation across all samples (e.g., genes that are nearly constant).

- Genes that do not vary much are unlikely to be differentially expressed between conditions.

- Compute the standard deviation or CV of normalized expression across samples > remove genes below a certain variance threshold (e.g., "remove the bottom 20% of genes by variance").

Gene Filtering Strategies

Filtering by Group-Specific Expression

- Ensure that a gene is expressed in at least one experimental group to a certain minimum level.

- Require that the gene has a minimum of X counts in at least Y replicates of at least one condition (e.g., "At least 5 counts in ≥2 samples in at least one group").