Pathway Analysis: Comprehensive Guide

From Genes to Biological Insights

Pathway analysis transforms lists of differentially expressed genes into interpretable biological knowledge. Instead of examining hundreds of individual genes, we identify enriched biological pathways, processes, and functions.

Analysis Methods

Over-Representation Analysis (ORA)

How it works:

- Tests if DE genes are over-represented in a pathway
- Uses Fisher's exact test on 2×2 contingency table
- Simple and interpretable results

Limitations:

- Requires arbitrary threshold (e.g., $p < 0.05$, $|FC| > 2$)
- Loses information about genes near threshold
- Ignores magnitude of gene expression changes

Gene Set Enrichment Analysis (GSEA)

How it works:

- Uses all genes ranked by expression change
- Calculates enrichment score (ES) using Kolmogorov-Smirnov test
- No arbitrary threshold needed

Advantages:

- More sensitive than ORA
- Considers all genes and their ranks

- Identifies coordinated changes in pathways

Feature	ORA	GSEA
Input	List of significant genes	All genes with expression values
Threshold	Required (e.g., $p < 0.05$, $\|FC\| > 2$)	Not required
Statistical Test	Fisher's exact test / Hypergeometric test	Kolmogorov-Smirnov-like statistic
Sensitivity	Lower (loses borderline genes)	Higher (uses all genes)
Computational Cost	Fast	Slower (permutation testing)
Best For	Quick exploratory analysis	Comprehensive pathway analysis

Pathway Databases

Gene Ontology (GO)

Gene Ontology is a standardized vocabulary system that describes gene functions across all species. It provides a hierarchical, structured representation of biological knowledge organized into three independent ontologies: Biological Process (BP), Molecular Function (MF), and Cellular Component (CC). GO is the most widely used annotation system in bioinformatics and is continuously updated by the research community.

Biological Process (BP)

Series of molecular events with a defined beginning and end. Examples: "cell cycle", "immune response", "metabolic process". Represents biological objectives accomplished by ordered assemblies of molecular functions.

Molecular Function (MF)

Activities at the molecular level. Examples: "kinase activity", "DNA binding", "receptor activity". Describes tasks performed by individual gene products without specifying where or when they occur.

Cellular Component (CC)

Locations where gene products are active. Examples: "nucleus", "mitochondrion", "plasma membrane". Describes subcellular structures and macromolecular complexes.

Hierarchical Structure

Terms organized in directed acyclic graph (DAG) from general to specific. Child terms inherit relationships from parents, enabling analysis at different granularities.

Species-Independent

Applies to all organisms from bacteria to humans. Same GO term can annotate orthologous genes across species, facilitating comparative analyses.

Evidence Codes

Each annotation has evidence code (IEA: inferred, IDA: direct assay, etc.) indicating confidence level and source of annotation.

💡 Example Use Case: Cell Cycle Analysis

Scenario: You identified 150 upregulated genes in cancer cells vs. normal cells.

GO Analysis Results:

- **BP:** "Cell cycle" (GO:0007049, p = 1.2e-15) - 45 genes
- **BP:** "DNA replication" (GO:0006260, p = 3.4e-12) - 28 genes
- **MF:** "DNA-binding transcription factor activity" (GO:0003700, p = 2.1e-8) - 22 genes
- **CC:** "Nucleus" (GO:0005634, p = 5.6e-10) - 78 genes

Interpretation: Cancer cells show dysregulation of cell cycle control and DNA replication, with many transcription factors activated in the nucleus - consistent with uncontrolled proliferation.

🎯 When to use GO: Best for broad functional categorization, understanding where and how gene products function, and comparing results across different studies and organisms.

KEGG (Kyoto Encyclopedia of Genes and Genomes)

KEGG is a comprehensive database that integrates genomic, chemical, and systemic functional information. It focuses on molecular interaction networks including metabolic pathways, signaling cascades, and disease pathways. KEGG is particularly valuable for understanding how genes work together in biological systems and for visualizing pathway diagrams showing enzymes, substrates, and products.

Visual Pathway Maps

Detailed, hand-drawn pathway diagrams showing molecular interactions, substrates, and products with spatial organization. KEGG Mapper allows overlaying expression data onto these maps.

Metabolic Pathways

Comprehensive coverage of metabolic processes including glycolysis, TCA cycle, amino acid metabolism, and lipid metabolism with chemical structures and enzyme reactions.

Signaling Pathways

Well-curated signal transduction pathways (MAPK, PI3K-Akt, Wnt, etc.) showing protein-protein interactions, phosphorylation cascades, and transcriptional regulation.

Disease Pathways

Comprehensive disease pathway collection showing genetic and environmental factors in cancer, neurodegenerative diseases, metabolic disorders, and infectious diseases.

Drug Information

Includes drug-target relationships and pharmacological pathways. Useful for drug discovery, drug repurposing studies, and understanding mechanism of action.

Multi-Omics Integration

Links genes, proteins, compounds, reactions, and diseases. Enables integration of genomics, transcriptomics, metabolomics, and disease data in single framework.

💡 Example Use Case: Drug Response Study

Scenario: Analyzing gene expression changes in cells treated with a kinase inhibitor.

KEGG Analysis Results:

- **hsa04010** - MAPK signaling pathway (p = 2.3e-18) - 45 genes affected
- **hsa04151** - PI3K-Akt signaling pathway (p = 1.5e-14) - 38 genes affected
- **hsa04115** - p53 signaling pathway (p = 4.2e-12) - 22 genes affected
- **hsa05200** - Pathways in cancer (p = 8.7e-10) - 52 genes affected

Visualization: KEGG Mapper colored the affected genes on pathway diagrams, showing ERK and downstream targets downregulated (blue), apoptosis genes upregulated (red), and cell cycle checkpoint genes activated.

Interpretation: The drug effectively inhibits MAPK signaling, leading to cell cycle arrest and apoptosis activation - confirming mechanism of action.

```
# Example: Mapping expression data to KEGG pathway library(pathview) # Gene expression data (log2 fold changes)
gene.data <- setNames(deg_results$log2FoldChange, deg_results$entrez_id) # Visualize on MAPK pathway
pathview(gene.data = gene.data, pathway.id = "04010", # MAPK pathway species = "hsa", # Homo sapiens out.suffix =
"MAPK_drug_treatment")
```

🎯 When to use KEGG: Best for understanding molecular mechanisms, visualizing expression changes on pathway diagrams, drug target analysis, and integrating metabolomics with genomics data.

Reactome

Reactome is a peer-reviewed, expert-curated database of biological pathways with a focus on human biology. It represents pathways as a series of molecular events (reactions) organized hierarchically. Reactome provides detailed pathway diagrams with emphasis on molecular interactions, post-translational modifications (PTMs), protein complexes, and subcellular localization. The database is particularly strong in signal transduction and immune system pathways.

Reaction-Based Model

Pathways decomposed into individual molecular reactions. Each reaction explicitly shows inputs, outputs, catalysts, and regulators with stoichiometry and compartmentalization.

Peer-Reviewed Content

All pathways reviewed by experts in respective fields. Higher confidence in accuracy compared to automated annotations. Each pathway linked to supporting literature.

Detailed Annotations

Includes specific protein modifications (phosphorylation sites), complex formations, subcellular locations, and disease variants with precise molecular detail.

Hierarchical Organization

Top-level pathways subdivided into smaller units. Enables analysis at different granularity levels from broad processes to specific molecular events.

Cross-References

Extensive links to external databases (UniProt, ChEBI, Ensembl, GO). Enables comprehensive data integration across multiple biological resources.

Pathway Browser

Interactive visualization tool with zoom capabilities, overlay of expression data, and exploration of related pathways with intuitive interface.

💡 Example Use Case: Immune Response Analysis

Scenario: RNA-seq data from T cells stimulated with antigen compared to unstimulated controls.

Reactome Analysis Results:

- **R-HSA-202403** - TCR signaling (p = 1.5e-22) - 58 entities

- **R-HSA-202424** - Downstream TCR signaling (p = 3.2e-20) - 42 entities
- **R-HSA-1280215** - Cytokine signaling in immune system (p = 8.1e-18) - 67 entities
- **R-HSA-983705** - Interleukin-2 family signaling (p = 2.4e-15) - 28 entities

Detailed Insights:

- ZAP70 phosphorylation events increased (specific sites: Y319, Y493)
- LAT-SLP76 complex formation upregulated
- Nuclear translocation of NFAT transcription factors detected
- Reactions in cytoplasm and nucleus compartments identified

Interpretation: Complete activation of TCR signaling cascade with proper complex formation and nuclear translocation - indicating functional T cell activation.

```
# Example: Reactome pathway analysis library(ReactomePA) library(org.Hs.eg.db) # Convert gene symbols to Entrez IDs gene.entrez <- bitr(de.genes, fromType="SYMBOL", toType="ENTREZID", OrgDb=org.Hs.eg.db) # Enrichment analysis reactome.result <- enrichPathway(gene = gene.entrez$ENTREZID, pvalueCutoff = 0.05, pAdjustMethod = "BH", readable = TRUE) # Visualize as network library(enrichplot) cnetplot(reactome.result, categorySize="pvalue")
```

🎯 When to use Reactome: Best for detailed mechanistic understanding, analyzing protein modifications and complexes, studying immune system pathways, and when you need expert-curated, peer-reviewed pathway information.

MSigDB (Molecular Signatures Database)

MSigDB is a comprehensive collection of annotated gene sets developed at the Broad Institute. It aggregates gene sets from multiple sources including pathway databases, published literature, and computational analyses. MSigDB is organized into collections (C1-C8 and Hallmark), each serving different analysis purposes. It's the standard resource for Gene Set Enrichment Analysis (GSEA) and is particularly valuable for its curated "Hallmark" gene sets representing well-defined biological states and processes.

Hallmark Gene Sets (H)

50 refined sets representing well-defined biological states. Reduced redundancy and noise. Most commonly used for initial GSEA analysis. Examples: HALLMARK_HYPOXIA, HALLMARK_APOPTOSIS.

Curated Gene Sets (C2)

~6,300 gene sets from pathway databases (KEGG, Reactome, BioCarta), chemical and genetic perturbations. Most comprehensive collection for exploring diverse hypotheses.

Ontology Gene Sets (C5)

Gene Ontology terms organized by BP, CC, and MF. Provides GO analysis within MSigDB framework. Useful for functional annotation with GSEA method.

Oncogenic Signatures (C6)

Gene sets representing signatures of cellular pathways disrupted in cancer. Includes signatures for oncogenes (MYC, RAS) and tumor suppressors (TP53, RB1).

Immunologic Signatures (C7)

Gene sets representing cell types and states of immune system. Derived from microarray and RNA-seq studies of immune cells. Useful for immune profiling.

Cell Type Signatures (C8)

Gene sets from single-cell RNA-seq studies. Cell type markers for various tissues and cell types. Essential for deconvolution and cell type identification.

💡 Example Use Case: Cancer Subtype Classification

Scenario: Gene expression data from breast cancer samples - understanding biological differences between subtypes.

MSigDB GSEA Results (Hallmark):

- **Basal-like subtype enriched:**
  - HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION (NES=2.45, FDR=0.001)
  - HALLMARK_HYPOXIA (NES=2.12, FDR=0.002)

- HALLMARK_GLYCOLYSIS (NES=1.98, FDR=0.003)

- **Luminal subtype enriched:**
  - HALLMARK_ESTROGEN_RESPONSE_EARLY (NES=2.87, FDR<0.001)
  - HALLMARK_ESTROGEN_RESPONSE_LATE (NES=2.34, FDR=0.001)

Using C6 (Oncogenic Signatures):

- MYC_TARGETS highly enriched in aggressive tumors
- TP53_DN signatures in basal-like subtype

Interpretation: Clear molecular distinction - basal-like shows aggressive phenotype with EMT and metabolic reprogramming, while luminal shows estrogen dependence.

```
# Example: MSigDB analysis with fgsea library(fgsea) library(msigdbr) # Get Hallmark gene sets hallmark_sets <- msigdbr(species = "Homo sapiens", category = "H") hallmark_list <- split(hallmark_sets$gene_symbol, hallmark_sets$gs_name) # Prepare ranked gene list gene_ranks <- setNames(deg_results$stat, deg_results$gene) gene_ranks <- sort(gene_ranks, decreasing = TRUE) # Run GSEA fgsea_results <- fgsea(pathways = hallmark_list, stats = gene_ranks, minSize = 15, maxSize = 500, nPermSimple = 10000) # Filter significant pathways sig_pathways <- fgsea_results[padj < 0.05]
```

🎯 When to use MSigDB: Best for GSEA analysis, exploring diverse biological hypotheses, cancer research, when you want both broad (Hallmark) and comprehensive (C2) options, and for standardized analysis workflows.

Choosing the Right Database

Database	Best For	Strengths	Considerations

Gene Ontology	Broad functional annotation, cross-species comparisons	• Universal applicability • Most comprehensive • Hierarchical structure • Widely adopted	• Can be redundant • Varying evidence quality • Less detailed mechanisms
KEGG	Metabolic pathways, visual interpretation, drug studies	• Beautiful pathway maps • Metabolic focus • Drug information • Easy visualization	• Reference pathways • License restrictions • Update frequency
Reactome	Detailed mechanisms, immune pathways, PTM analysis	• Peer-reviewed • Reaction-level detail • Excellent for signal transduction • Well-annotated complexes	• Mainly human-focused • Smaller than others • Can be very detailed
MSigDB	GSEA, cancer research, hypothesis generation	• Hallmark sets (refined) • Comprehensive collections • Regular updates • GSEA-optimized	• Overlapping sets in C2 • Varying quality • Can be overwhelming

Best Practices for Pathway Analysis

1. Multiple Testing Correction

Always apply FDR or Bonferroni correction. Testing thousands of pathways inflates false positives. Use adjusted p-value < 0.05 as cutoff.

2. Consider Gene Set Size

Filter pathways with 15-500 genes. Very small sets are unstable; very large sets are too general and lack specificity.

3. Validate Results

Cross-validate findings across multiple databases. Examine individual genes in significant pathways for biological sense.

4. Use Background Appropriately

5. Report Methodology

6. Interpret Biologically

For ORA, use all genes measured in experiment as background, not entire genome. Affects statistical significance dramatically.

Clearly state: database version, analysis method (ORA/GSEA), p-value cutoffs, correction method, and software version.

Statistical significance ≠ biological importance. Consider effect sizes, biological context, and literature support.

⚠️ Common Pitfalls to Avoid

- **Cherry-picking results:** Don't select only pathways that fit your hypothesis
- **Ignoring redundancy:** Multiple related pathways may represent same biology
- **Over-interpreting p-values:** p=0.049 vs p=0.051 is not meaningful difference
- **Wrong background:** Using wrong gene universe inflates false positives
- **Outdated databases:** Use current versions - biology knowledge evolves
- **Ignoring directionality:** Check if genes are up or down-regulated in pathway

Summary & Key Takeaways

🎯 Quick Reference Guide

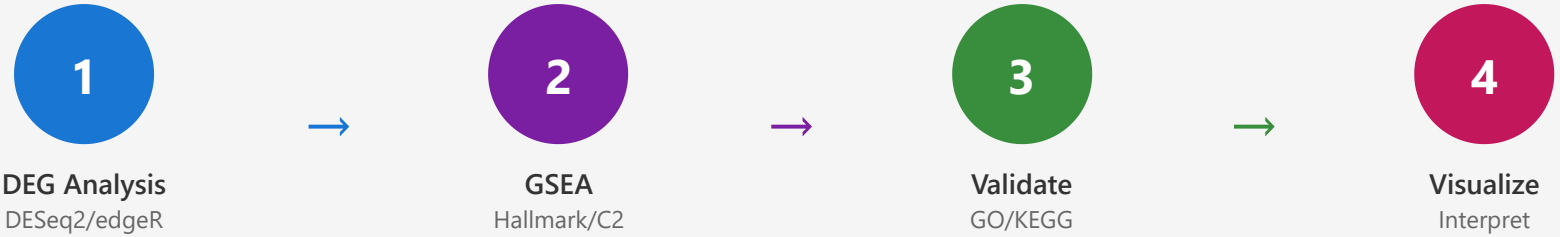
Analysis Methods

- **ORA:** Fast, simple, requires thresholds
- **GSEA:** More powerful, uses all genes
- **Recommendation:** Use GSEA when possible
- **Why:** GSEA detects subtle coordinated changes

Database Selection

- **Starting point:** MSigDB Hallmark or GO BP
- **Mechanism:** KEGG or Reactome
- **Validation:** Use multiple databases
- **Species:** Non-human → prioritize GO

📊 Typical Workflow



✅ Analysis Checklist

- ☑ Multiple testing correction applied
- ☑ Appropriate background used
- ☑ Gene set size filtered (15-500)
- ☑ Multiple databases consulted
- ☑ Individual genes examined
- ☑ Biological context considered
- ☑ Methods clearly documented
- ☑ Results validated

📚 Additional Resources

GO Consortium
geneontology.org
~44,000 terms

KEGG
genome.jp/kegg
~500 pathways

Reactome
reactome.org
~2,500 pathways

MSigDB
gsea-msigdb.org
~25,000 gene sets

🧬 Remember

Pathway analysis transforms gene lists into biological insights.

Choose your tools wisely, validate thoroughly, and always think biologically!

The goal is understanding, not just statistics.