References:

Integrating Large-Scale Knowledge Graphs to Enhance Transcriptomics Analysis

Wanjun Gu, Gianmarco Bellucci, Braian Peetoom, Maura McDonagh and Sergio Baranzini

Department of Neurology, Weill Institute for Neurosciences, University of California, San Francisco, San Francisco, CA

University of California San Francisco

Introduction

Unlocking Insights from Omics Data: High-throughput omics technologies generate vast transcriptomic datasets but extracting holistic and comprehensive biological insights remains a challenge.

Overcoming Noise and Threshold Limitations: A method is needed to filter through noise and uncover biologically meaningful signals, even beyond conventional significance thresholds.

Cell-Type-Specific Analysis Matters: Understanding transcriptomic changes at the cell-type level is crucial for accurate interpretation, yet many existing tools lack this capability.

Introducing tKOI: The transcriptomics Knowledge-graph-driven Omics Integration (tKOI) pipeline leverages large-scale biological knowledge graphs to enhance functional, network-based transcriptomic analysis and enable cell-type-aware discoveries.

Method

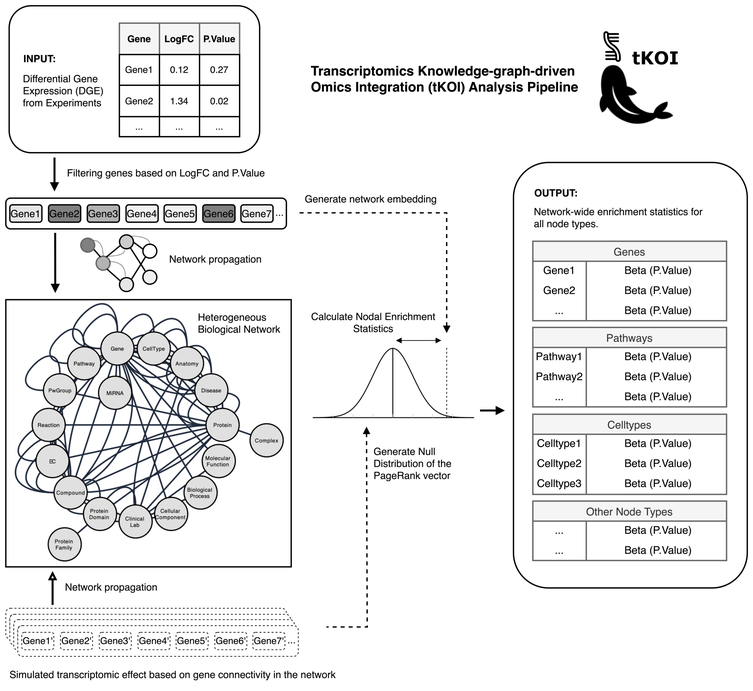

Figure 1 and Table 1. Summary of the tKOI analysis pipeline

Step	Description
Input	Filtered RNA-seq DGE results (LogFC, p-values)
Knowledge Graph	Pruned subset of SPOKE, containing 18 biological node types (genes, pathways, cell types, diseases, etc.)
Network Propagation	SoftMax transformation prioritizes highly upregulated genes Personalized PageRank (PPR) spreads transcriptomic signals across the network
Statistical Inference	Null distribution generated via topologically matched randomization
Output	Ranked enrichment scores for genes, pathways, cell types, enabling cell-type-aware and network-informed analysis

Pathway Analysis

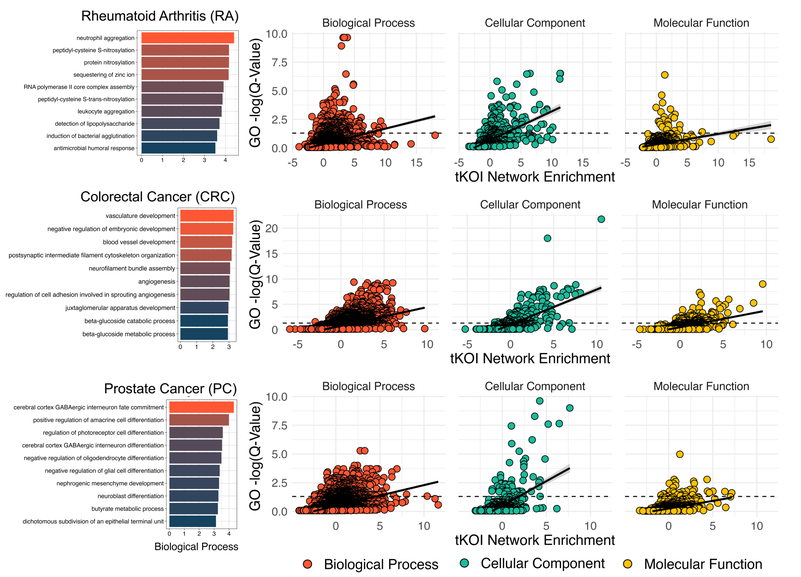

Figure 2: Validation of tKOI GO Enrichment Analysis Across Disease Datasets. This figure compares Gene Ontology (GO) enrichment results obtained from tKOI with those from clusterProfiler, a widely used GO enrichment tool, across three disease datasets: Rheumatoid Arthritis (**Top**), Colorectal Cancer (**Middle**), and Prostate Cancer (**Bottom**). The bar plots on the left display the top 10 most enriched biological processes identified by tKOI for each dataset. The scatter plots on the right illustrate the consistency between tKOI enrichment scores and those from clusterProfiler for biological processes (**Red**), cellular components (**Green**), and molecular functions (**Yellow**). A positive correlation between the two methods indicates that tKOI effectively captures biologically relevant pathways.

Gene Exploration

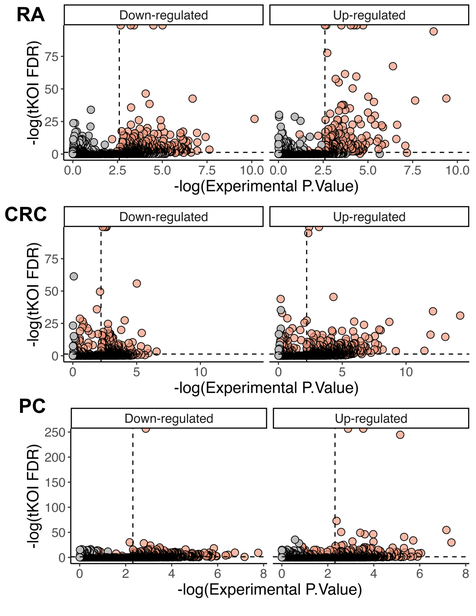

Figure 3. Retainment and Discovery of Significant Genes in tKOI Analysis

Significance between experimental differential expression analysis and tKOI enrichment results were compared across the three tested datasets. Most genes that are significant in the experiment also show strong significance in tKOI, demonstrating **retainment** of key experimental findings. However, tKOI also identifies **discovery** genes, genes that are only significant in the network-based analysis but not in the experiment. These genes, despite not meeting experimental significance thresholds, may hold high biological relevance in the network context. By leveraging knowledge graph propagation, tKOI enhances gene discovery beyond conventional differential expression analysis.

Simulations

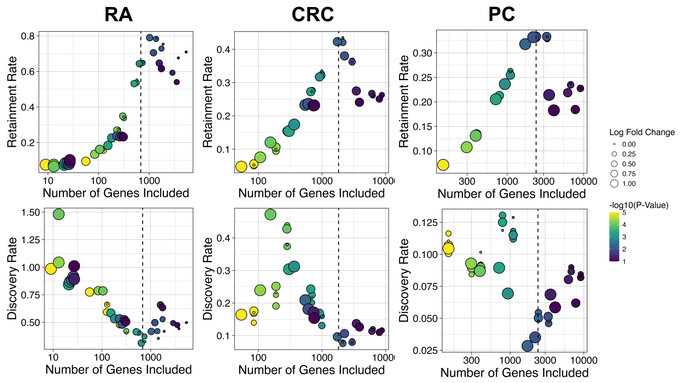

Figure 3: Simulation-Based Optimization of Significance Thresholds for tKOI Analysis

Simulated differentially expressed gene datasets were generated by varying log fold change (LFC) thresholds from 0 to 1 and p-value cutoffs from 0.1 to 0.00001. The top row shows the retainment rate, representing the proportion of experimentally significant genes that remain significant in tKOI. The bottom row depicts the discovery rate, capturing genes that gain significance in tKOI despite being borderline or insignificant in the experiment. The color scale reflects the statistical significance of the genes, while point sizes indicate LFC magnitude. This approach enables the identification of optimal significance thresholds that balance sensitivity and discovery potential in tKOI analysis.

Cell Type Specificity

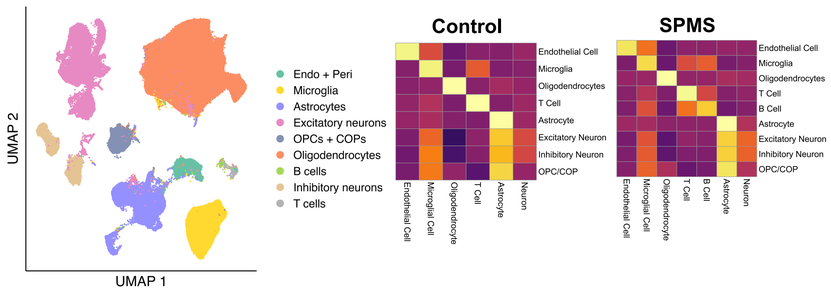

Figure 4. Cell-Type-Specific Analysis of Secondary Progressive Multiple Sclerosis (SPMS) Using tKOI

This figure demonstrates the cell-type specificity of tKOI by applying it to a Multiple Sclerosis dataset. (Left) A UMAP plot of single-cell RNA sequencing data shows the clustering of different cell types, including oligodendrocytes, microglia, astrocytes, neurons, and immune cells. (Right) Heatmaps display the cell-cell interaction scores inferred from tKOI for control (B) and secondary progressive multiple sclerosis (SPMS) (C) conditions. Marker genes identified within each cell type cluster exhibit similar cell-type network enrichment profiles, demonstrating that tKOI effectively captures cell-type-specific pathway activity and interactions in disease contexts.

Conclusion

In summary, tKOI leverages large-scale biological knowledge graphs for enhanced transcriptomic analysis.

Comparable to Leading Tools: tKOI produces Gene Ontology (GO) enrichment results on par with ClusterProfiler.

Beyond GO & KEGG: tKOI uncovers novel biological insights beyond traditional pathway annotations.

Discovery of Hidden Signals: Identifies disease-associated genes missed by conventional statistical thresholds but biologically relevant.

Cell-Type Specificity: Captures cell-type-enriched signals, enabling more precise transcriptomic interpretation.