Integrating molecular signature databases to holistically reveal gene signaling events

Ivor Ho[1,2], Shan He[1], Vakul Mohanty[1], Ken Chen[1]

[1]Department of Bioinformatics and Computational Biology, University of Texas MD Anderson Cancer Center, Houston, Texas
[2]Department of Biological Science, Florida State University, Tallahassee, Florida

THE UNIVERSITY OF TEXAS
MD Anderson Cancer Center
Making Cancer History®

Introduction

Curated molecular interaction databases capture important aspects of biological systems such as signaling pathways, ligand-receptor interactions and co-functional gene modules. Information encoded in these databases can facilitate deeper interrogation of large omics datasets. Currently, there are limited efforts to integrate multiple interaction databases, capturing different aspects of biological systems, and utilizing them for various applications. One application of these integrated graphs is to predict effects of gene perturbations.

Methods

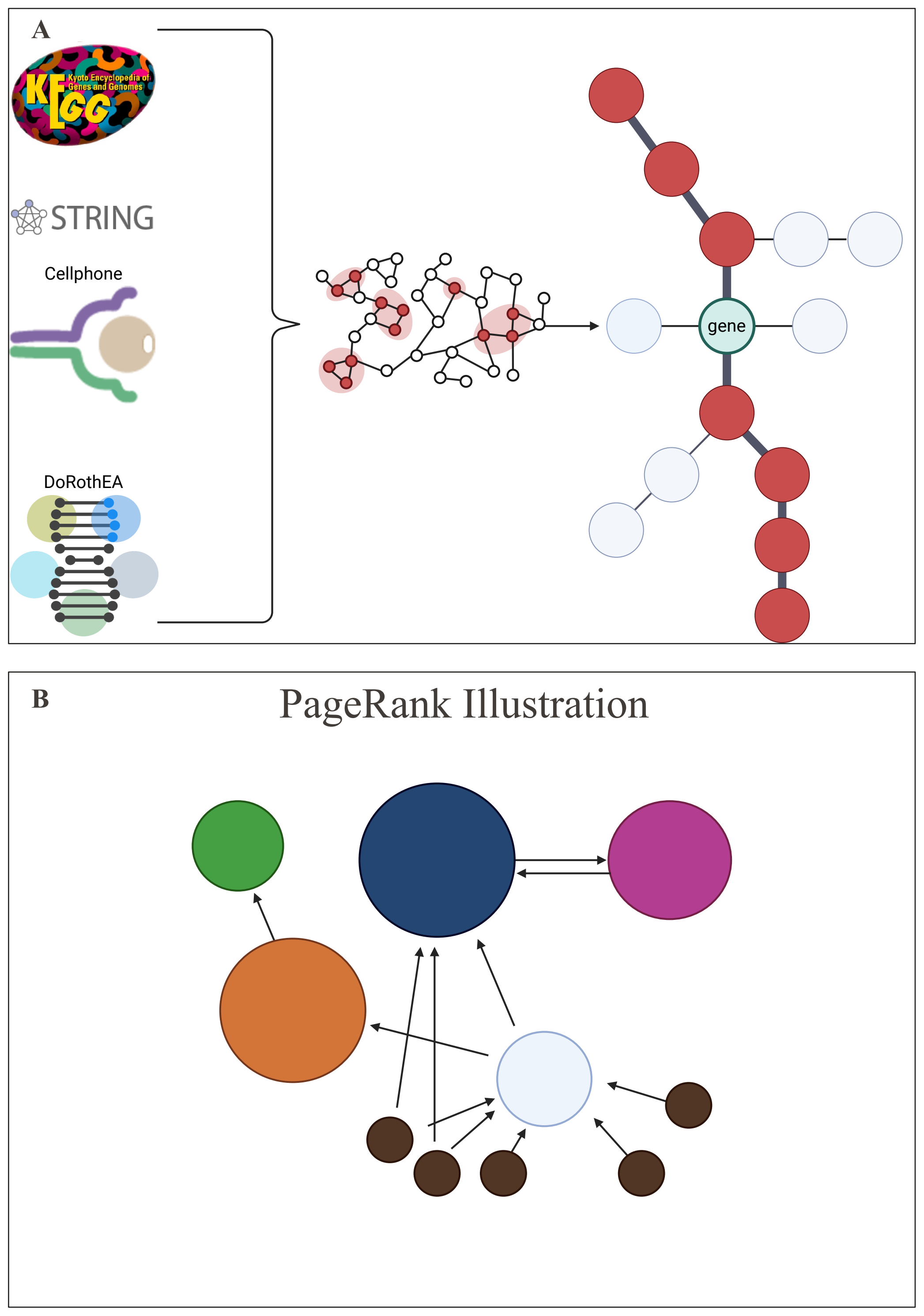

Figure 1. Network construction and analysis (A) Curated interaction databases are merged into a directed knowledge graph. From this graph, each perturbation gene's neighborhood is extracted for downstream analysis. (B) The full graph is analyzed with a personalized PageRank algorithm that quantifies the influence of each gene within the graph.

Graph Statistics

Source	Nodes	Total Edges	Activating	Inhibiting
Cellphone	713	1553	1553	0
DoRothEA	20295	454504	410583	43921
KEGG	3429	21185	17291	3894
STRING	16703	1019970	0	0
Full Graph	41140	1497212	429427	47815

Densest subgraphs

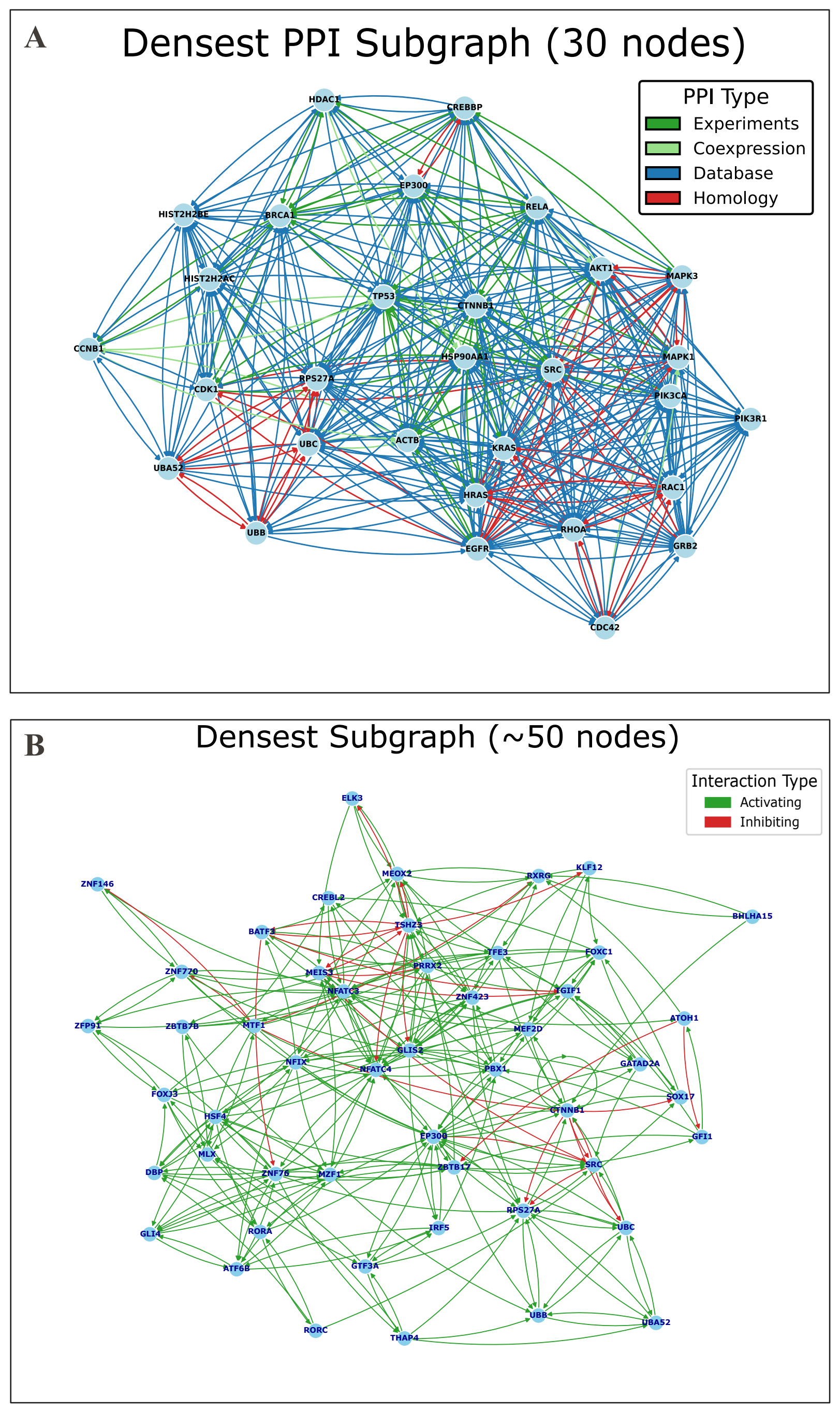

Figure 2. Densest network subgraphs in knowledge graph. (A) The densest subgraph from the graph-derived PPI network, with edges colored by evidence type. (B) Densest subgraph from the knowledge graph. Each gene shares many interactions with others, creating a highly coordinated hub.

Gene subgraphs

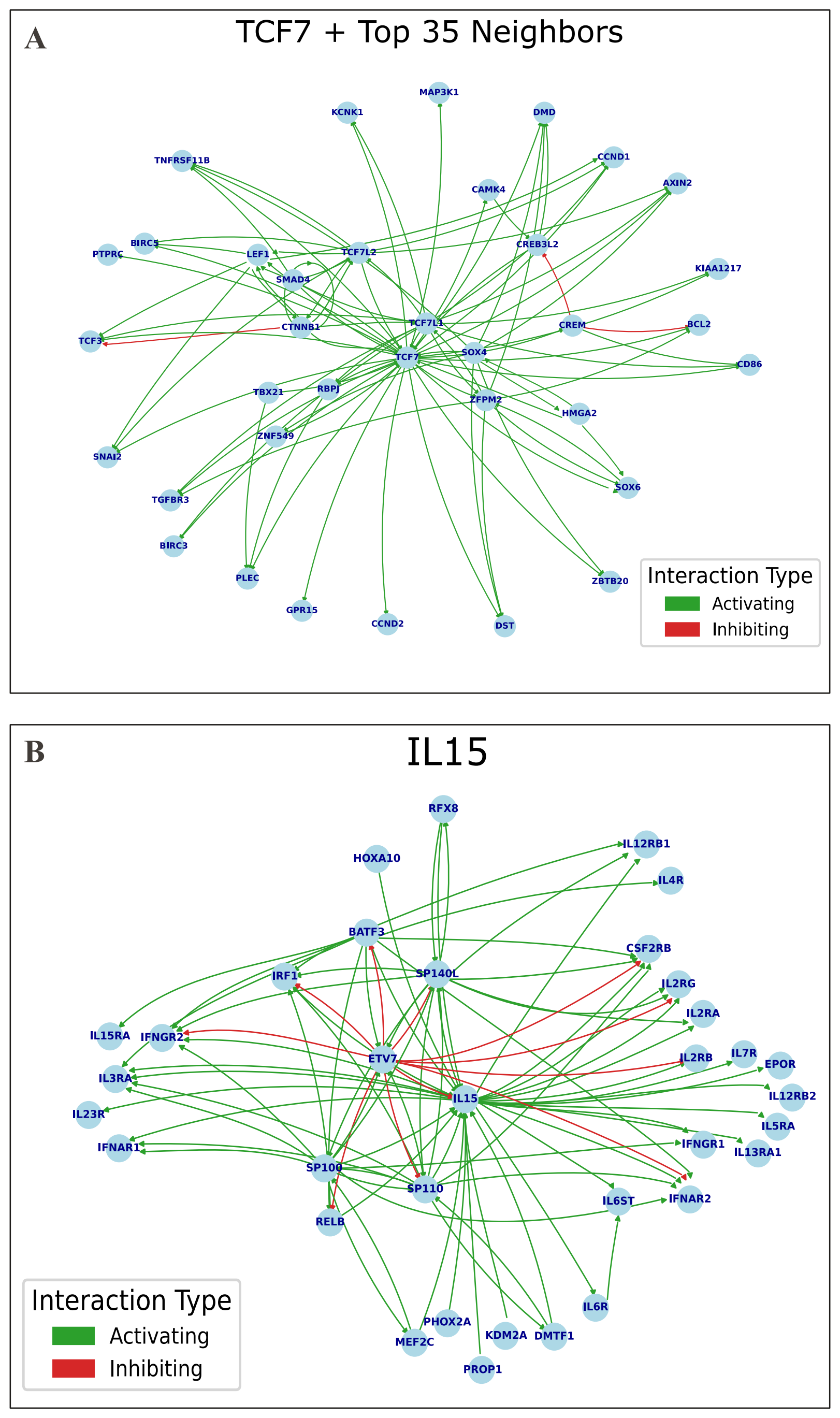

Figure 4. Interaction subnetworks for selected perturbation genes. (A) Directed neighborhood of TCF7 with top 35 neighbors, showing all direct upstream and downstream genes with activating and inhibitory edges. (B) Directed neighborhood of IL15, with same format.

Perturbation Prediction

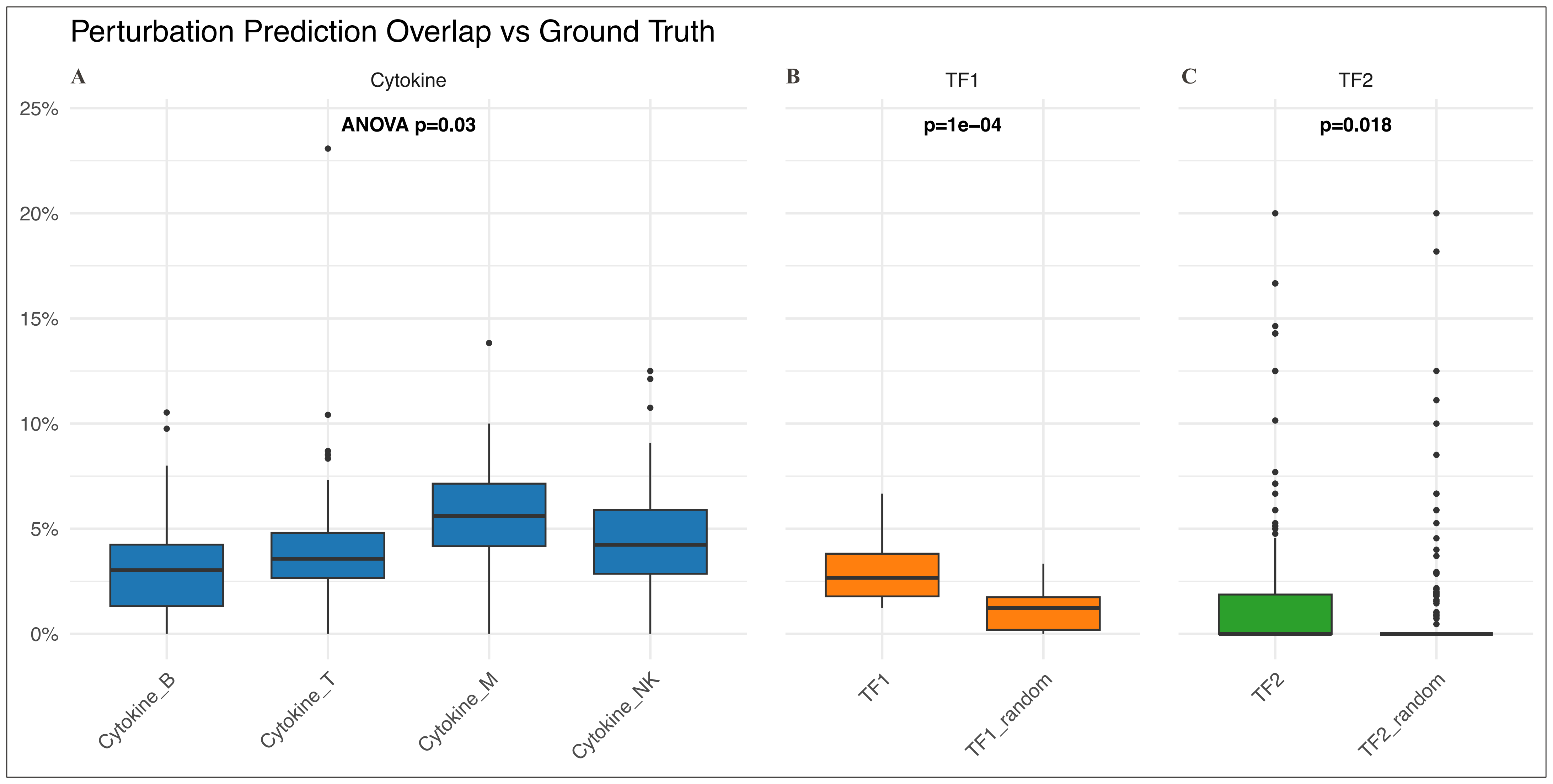

Figure 3. Overlap between PageRank predicted genes and experimentally determined DEGs for gene perturbations (A) Distribution of overlap between the top 1% PageRank-predicted genes and differentially expressed genes (DEGs) across B-cells, T-cells, macrophages, and NK cells. (one-way ANOVA, p = 0.03) (B) Comparison of prediction overlap from a CRISPR knock-out experiment (TF1) versus randomized graph (two-sample t-test, p = 1e-04). (C) Overlap from a similar knock-out experiment (TF2) versus randomized graph. (two-sample t-test, p = 0.018)

Downstream Analysis

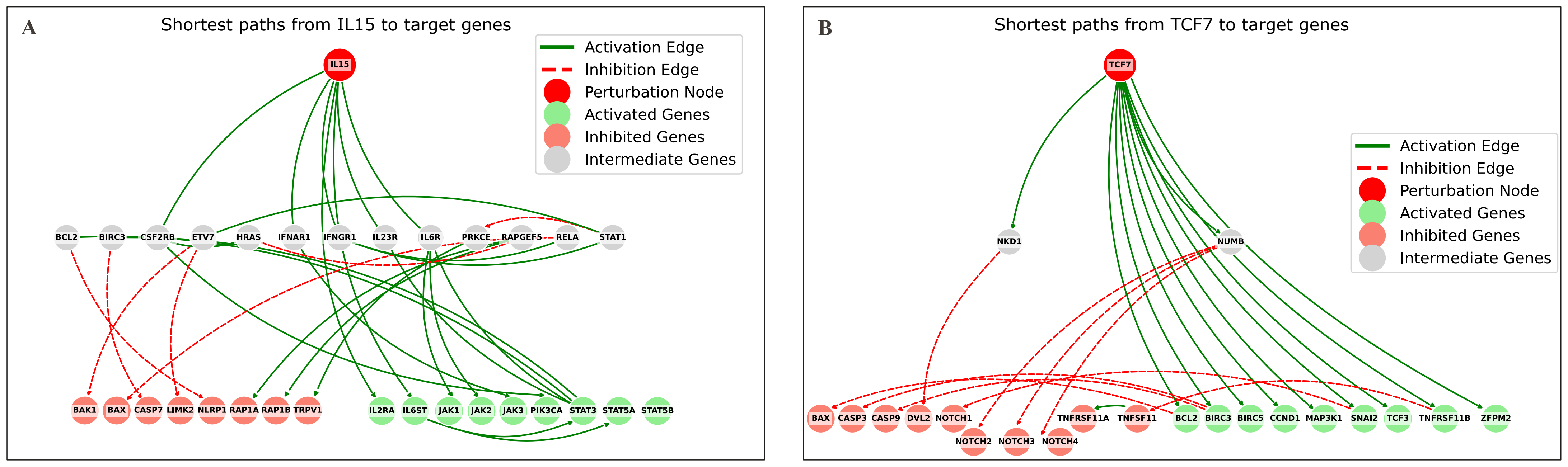

Figure 5. Shortest path subnetworks from perturbed genes in knowledge graph. (A) Paths from IL15 gene, showing activation (solid green) and inhibition (dashed red) edges traveling through intermediate genes (grey). (B) Paths from TCF7 gene, with same color conventions. These visualizations highlight how each perturbation propagates through directed paths in the network.

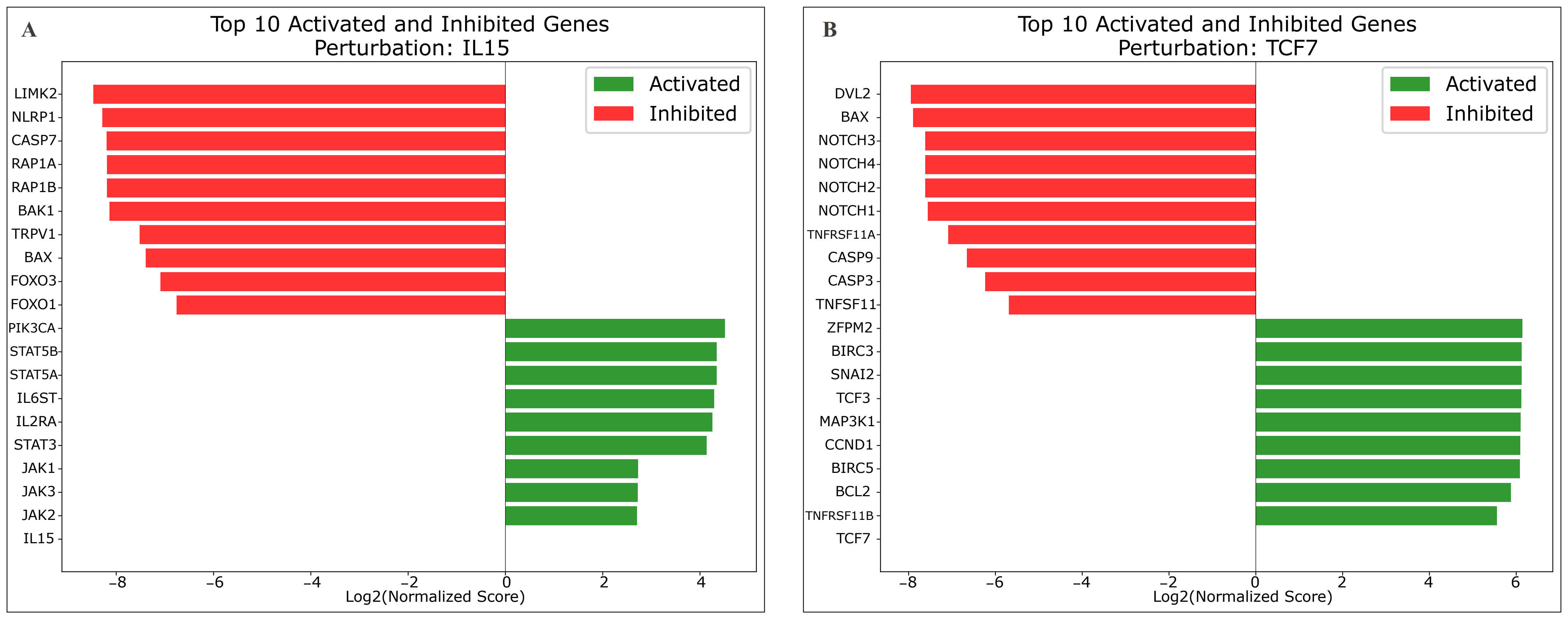

Figure 6. Top downstream effectors from directed PageRank. (A) Horizontal bar plot of the ten highest-scoring activated and inhibited genes following IL15 perturbation. (B) Bar plot for TCF7 perturbation, using the same Log2 normalized score scale. Each gene's score reflects its proximity and path strength from the perturbed gene through the directed network.

Conclusions

Transforming curated molecular interaction databases into a knowledge graph allows advanced network analysis that makes it possible to accurately identify downstream genes and study their relationships and pathways. By mapping molecular interactions onto a graph, data can be converted into an interactive and rich network that highlights important genes, reveals signaling pathways, and drive experimental hypotheses.

Acknowledgements

IH is supported by the CPRIT Training Program RP210028

IH is an undergraduate MD Anderson CATALYST trainee