Cutting through the complexity of genomic data : A general method to identify candidate genes

Narmada Sambaturu[1], Sridhar Hannenhalli[1,2] and Nagasuma Chandra[1]

[1]Indian Institute of Science, Bengaluru, India

[2]University of Maryland, College Park, USA

INTRODUCTION

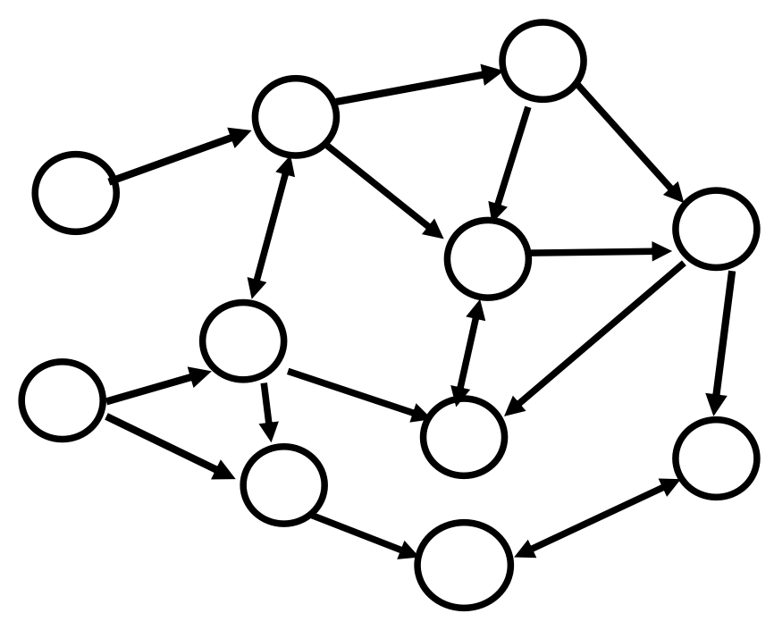
- Genomic and transcriptomic data from biological and clinical samples can capture information relevant to the tissue or disease under study.
- Most studies involve the selection of candidate genes for probing their function, mechanism or other application.
- However, noise caused by inherent biological heterogeneity confounds this selection.
- Protein-protein interaction networks give a bird's eye view of the paths along which information can flow in a system.
- Although such networks are typically built for an organism as a whole and not all interactions are relevant for a particular cell type or disease condition, integrating these two data can help extract condition-specific information.

OBJECTIVES

- We showcase a method to integrate protein-protein interaction networks (PPIs) with transcriptomic data to obtain a condition-specific sub-network, or *top-network*.
- This top-network can then be mined to identify candidate genes to address the question of interest.

METHODS

Input 1. Protein-protein interaction network

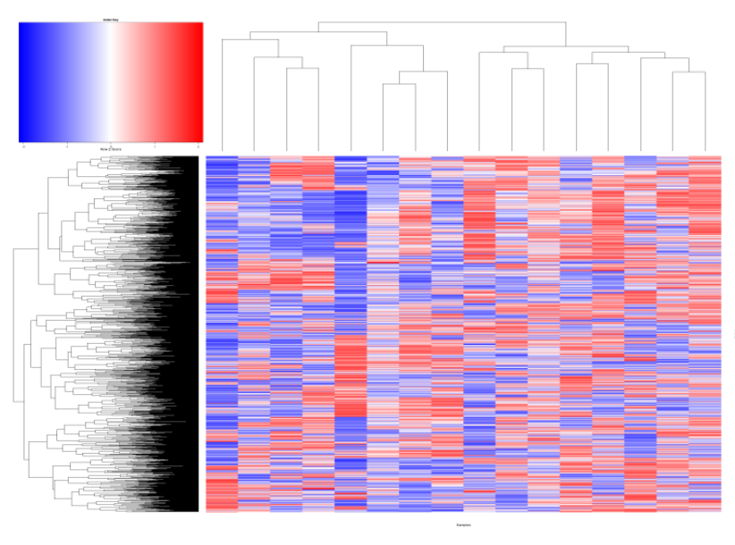


Human interaction network*

- Sources: In-house curated network (STRING v 10, SignaLink v 2.0, Cancer Cell Map, BioGRID, Multinet)
- Genes (nodes): 17,062
- Interactions (edges): 2,08,760

*Network published in NPJ Systems Biology

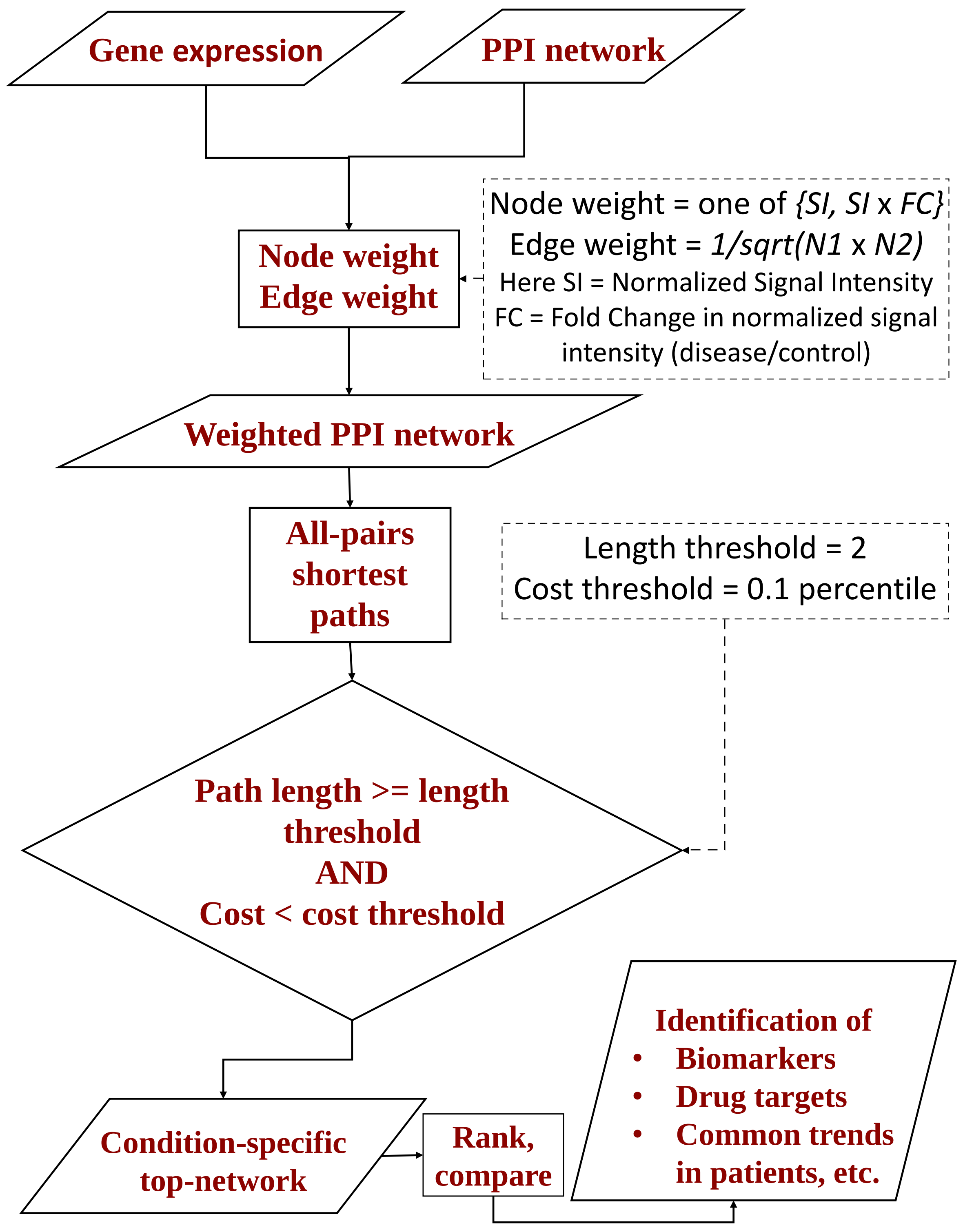
Input 2. Gene expression data



Transcriptome data

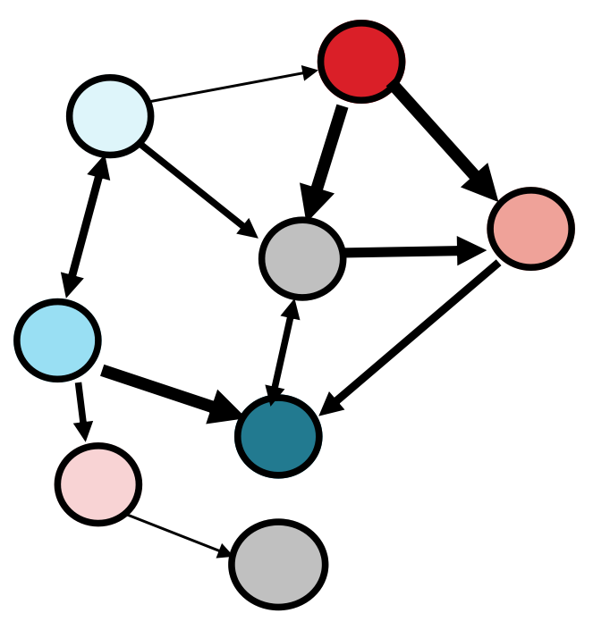
- Active tuberculosis vs healthy controls, GSE19491
- Normal liver tissue, GTEx v6p RNA-SeQCv1.1.8
- Up-regulation of PARK2 in glioma cell line (U251), GSE61973

Algorithm



Gene expression
PPI network
Node weight
Edge weight
Node weight = one of *{SI, SI x FC}*
Edge weight = *1/sqrt(N1 x N2)*
Here SI = Normalized Signal Intensity
FC = Fold Change in normalized signal intensity (disease/control)
Weighted PPI network
All-pairs shortest paths
Length threshold = 2
Cost threshold = 0.1 percentile
Path length >= length threshold AND Cost < cost threshold
Condition-specific top-network
Rank, compare
Identification of
Biomarkers
Drug targets
Common trends in patients, etc.

Output. Condition-specific top-network



Condition-specific top-network

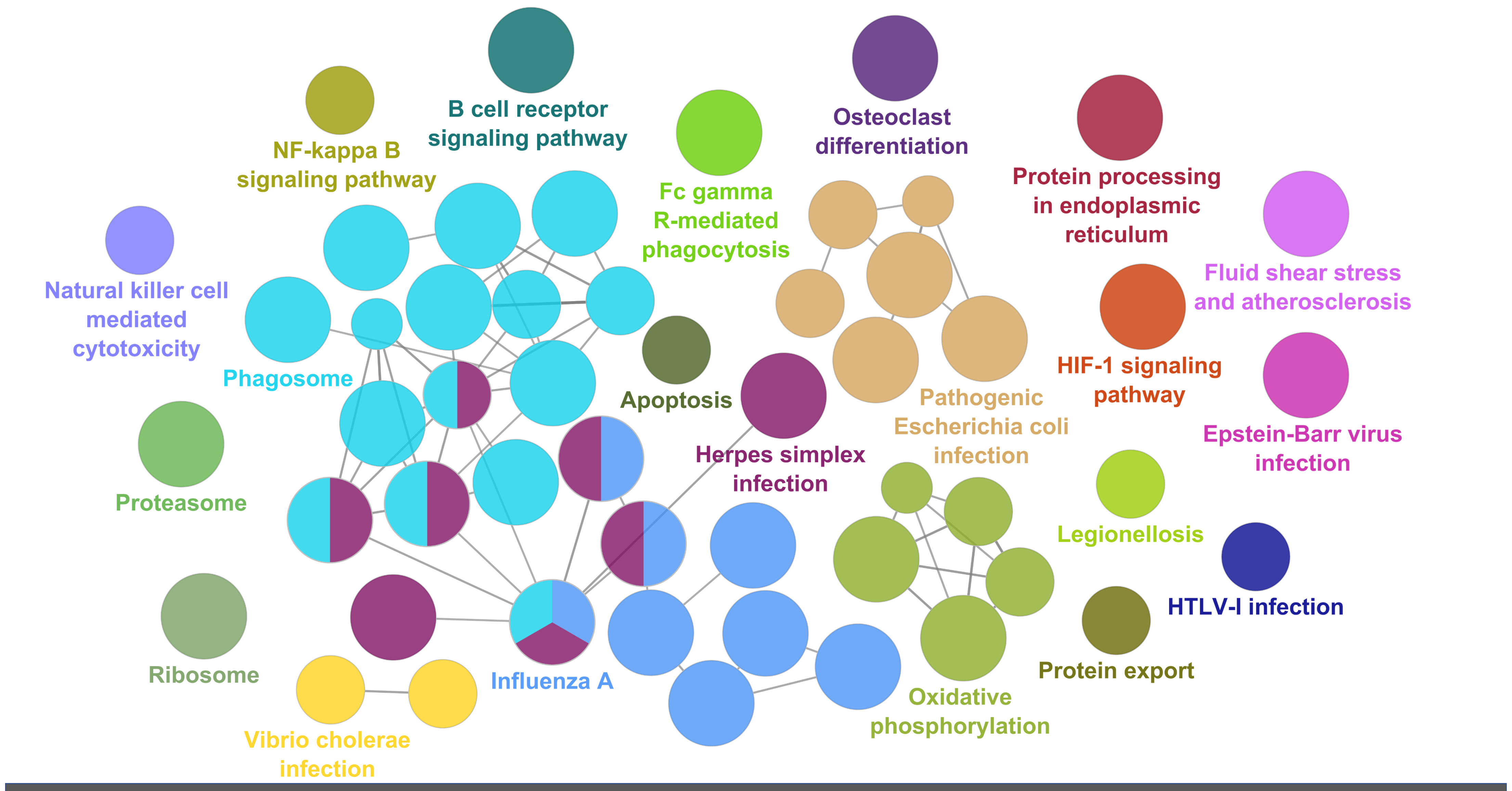
- A sub-network with condition-specific nodes and interactions
- Candidate genes can be obtained by ranking nodes of this network using any appropriate centrality measure

RESULTS - Top-networks identified by our algorithm are enriched in condition-specific signals.

1. Disease - tuberculosis

Data: In-house curated PPI + Whole blood of patients with active tuberculosis vs healthy controls (GSE19491)

Inference: Top-network is characteristic of host responses in tuberculosis infection, such as NF-kappa B signaling pathway, Natural killer cell mediated cytotoxicity, among others.

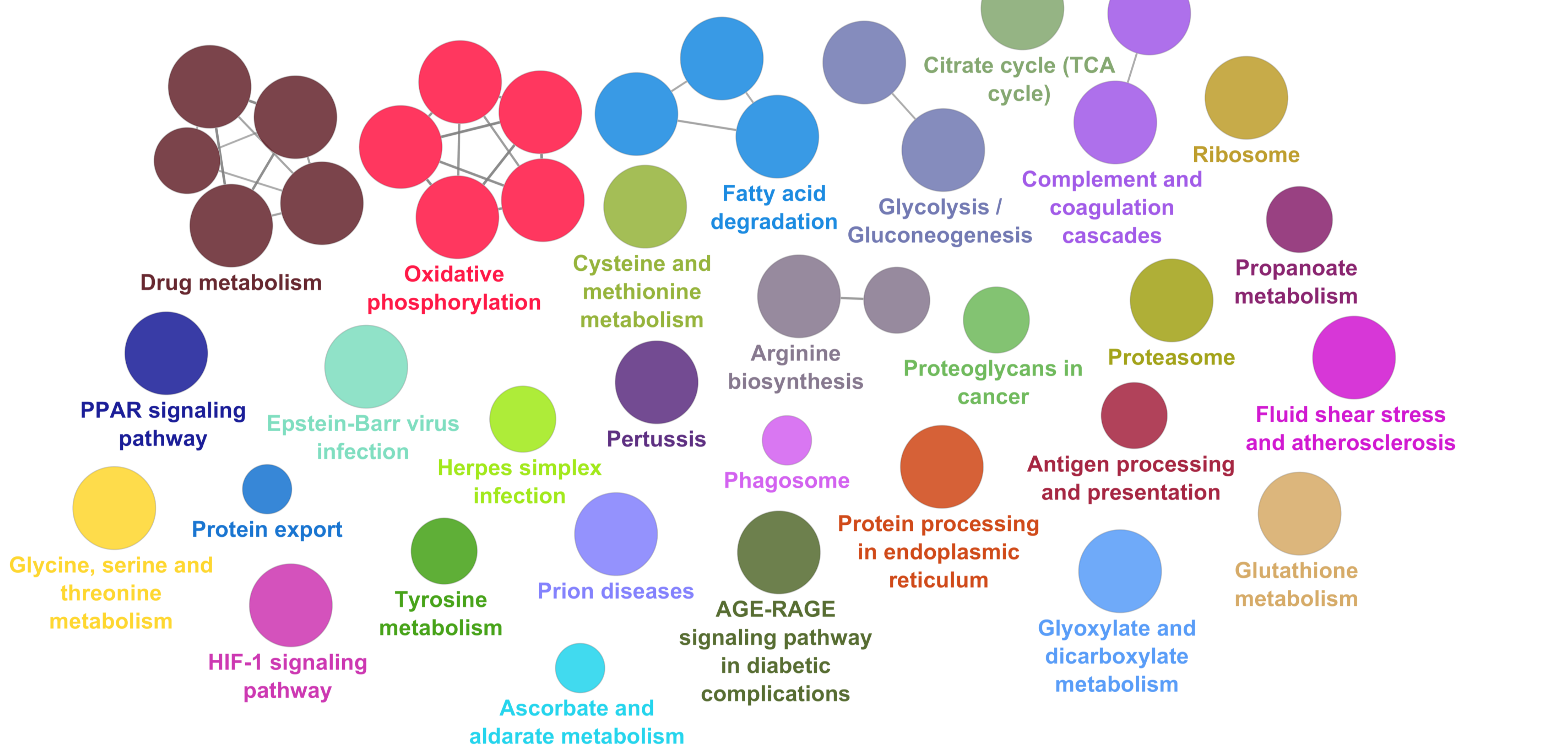


NF-kappa B signaling pathway
B cell receptor signaling pathway
Fc gamma R-mediated phagocytosis
Osteoclast differentiation
Protein processing in endoplasmic reticulum
Fluid shear stress and atherosclerosis
Natural killer cell mediated cytotoxicity
Phagosome
Apoptosis
HIF-1 signaling pathway
Pathogenic Escherichia coli infection
Epstein-Barr virus infection
Proteasome
Herpes simplex infection
Legionellosis
HTLV-I infection
Ribosome
Influenza A
Protein export
Oxidative phosphorylation
Vibrio cholerae infection

2. Normal tissue - liver

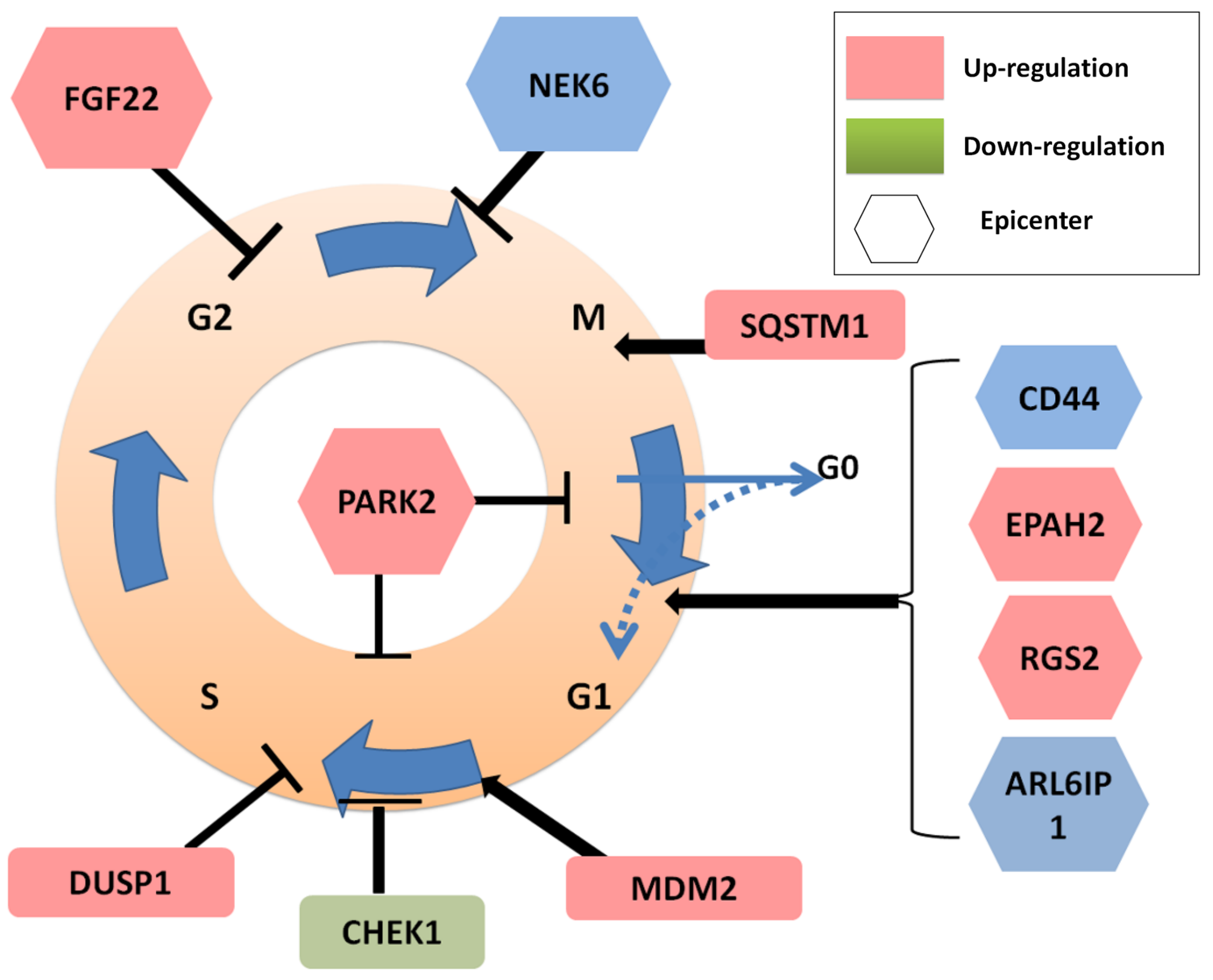
Data: In-house curated PPI + Liver tissue (GTEx v6p RNA-SeQCv1.1.8)

Inference: The top-network highlights *top-activity paths* in the liver tissue, such as Drug metabolism, Cysteine and methionine metabolism, among others.



Citrate cycle (TCA cycle)
Ribosome
Fatty acid degradation
Glycolysis / Gluconeogenesis
Complement and coagulation cascades
Drug metabolism
Oxidative phosphorylation
Cysteine and methionine metabolism
Propanoate metabolism
Arginine biosynthesis
Proteoglycans in cancer
Proteasome
PPAR signaling pathway
Epstein-Barr virus infection
Pertussis
Fluid shear stress and atherosclerosis
Herpes simplex infection
Phagosome
Antigen processing and presentation
Protein export
Protein processing in endoplasmic reticulum
Glycine, serine and threonine metabolism
Tyrosine metabolism
Prion diseases
AGE-RAGE signaling pathway in diabetic complications
Glutathione metabolism
Glyoxylate and dicarboxylate metabolism
HIF-1 signaling pathway
Ascorbate and aldarate metabolism

3. Targeted up-regulation of PARK2



Data: In-house curated PPI + Targeted up-regulation of PARK2 gene in human glioma cell line (U251) (GSE61973)

Inference: Ranking genes in the top-network based on centrality identified PARK2 as the most influential gene. Other highly ranked genes were found to enable or counter the activity of PARK2.

Selected applications

- **Blood-based biomarker** for **prognosis of treatment** in **tuberculosis** patients (Chandrani et al, Gordon Research Conference, 2017)
- **Gene signature** to discriminate **primary from metastatic melanoma** (Metri et al, Scientific Reports, 2017, *In Press*)
- **Blood-based biomarker** for **pulmonary tuberculosis** (Sambarey et al, EBioMedicine, 2017)
- Identification of **'common-core' in tuberculosis** (Sambarey et al, NPJ Systems Biology and Applications, 2017)
- Identification of **'epicenters' of perturbation** (Sambaturu et al, BMC genomics, 2016)

Conclusions

- The algorithm can answer different biological questions depending on the weighting scheme and filtration methods used.
- The algorithm also serves as a general framework for incorporation of other omics data.