AWS Open Data Hackathon

October 1-3, 2025

Group 1

Project Outline and Goal

Essence: Build a knowledge graph connecting variants, drugs, and clinical evidence to identify therapeutic opportunities

Details: Use gene-centric somatic variants data from MTP and TCGA and link with drug targets from OncoDB, evidence-based variant interpretation from CIViC, biological context and functional relationships from MSigDB, and variant/allele metadata from HGNC reference. Attempt to also include protein interaction data from StringDB

Introduction to Technical Implementation

- Use Neo4j in Docker as the graph database
- Nodes: genes, variants, cancer_types, histological data, drug targets, pathways
- Edges:gene_has_variant_src, ASSOCIATED_WITH_PATHWAY, TARGETED_BY_DRUG
- Cypher queries for flexible retrieval and downstream analytics

Technical Methods

- Data Processing

- Download datasets (TSV, JSON)
- Standardize schemas for each dataset as collected from different sources, using gene_name as primary key

Technical Methods

- Knowledge Graph Construction

- Define node types:
- Define relationships:
- Define properties
- Convert datasets into triples compatible with Neo4j
- Create ETL pipelines that automatically ingest data into Neo4j running on docker

Technical Methods

- **Analysis and Querying**
  - Use Cypher
    - Use case: filter for specific cancer_type,
      - Extract drug_targets, pathway level insights

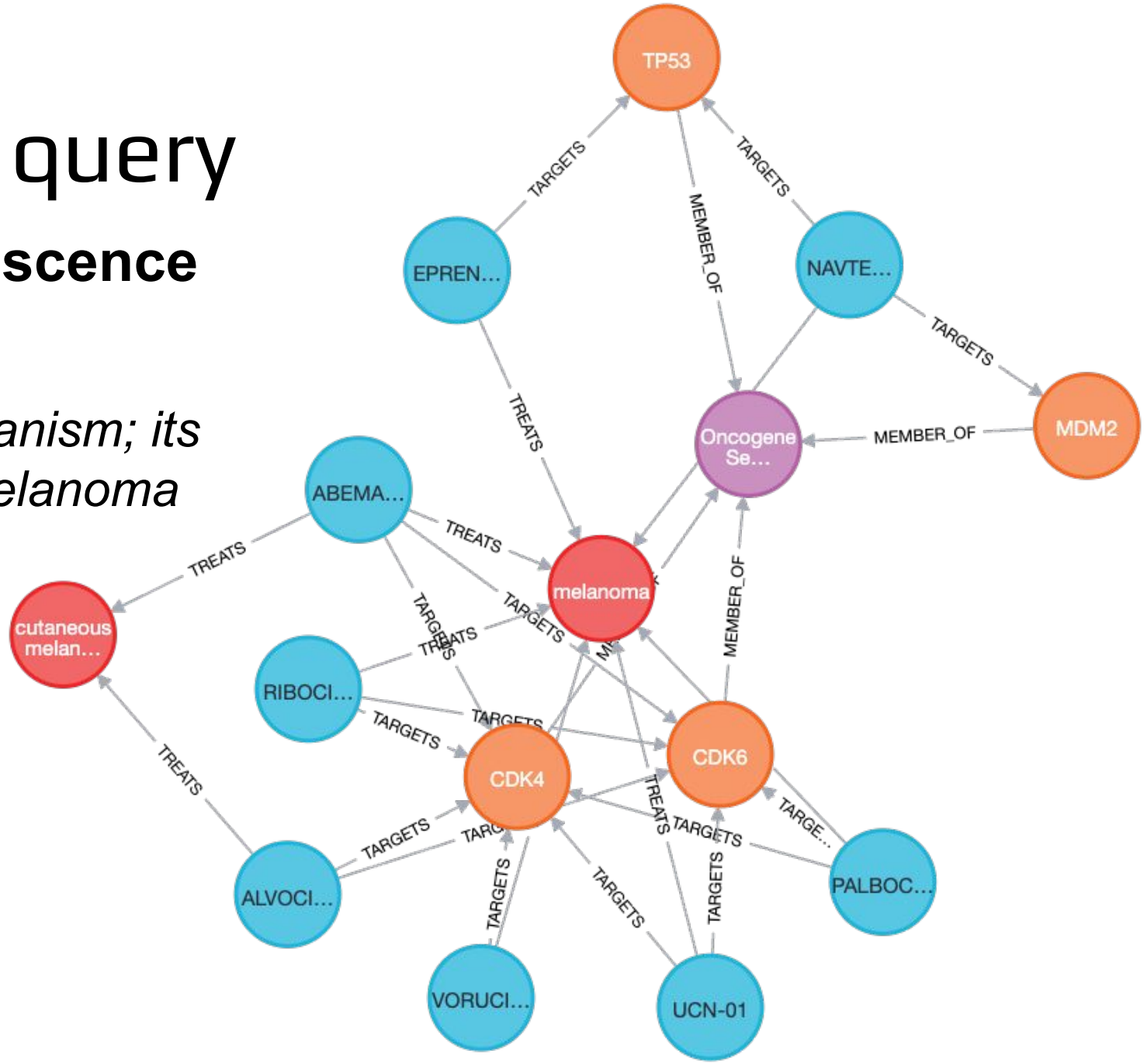
- **Visualization/Outputs**
  - Graphs connecting genes-variants-drug targets-clinical interpretation-pathway level insight

Pathway-centric query

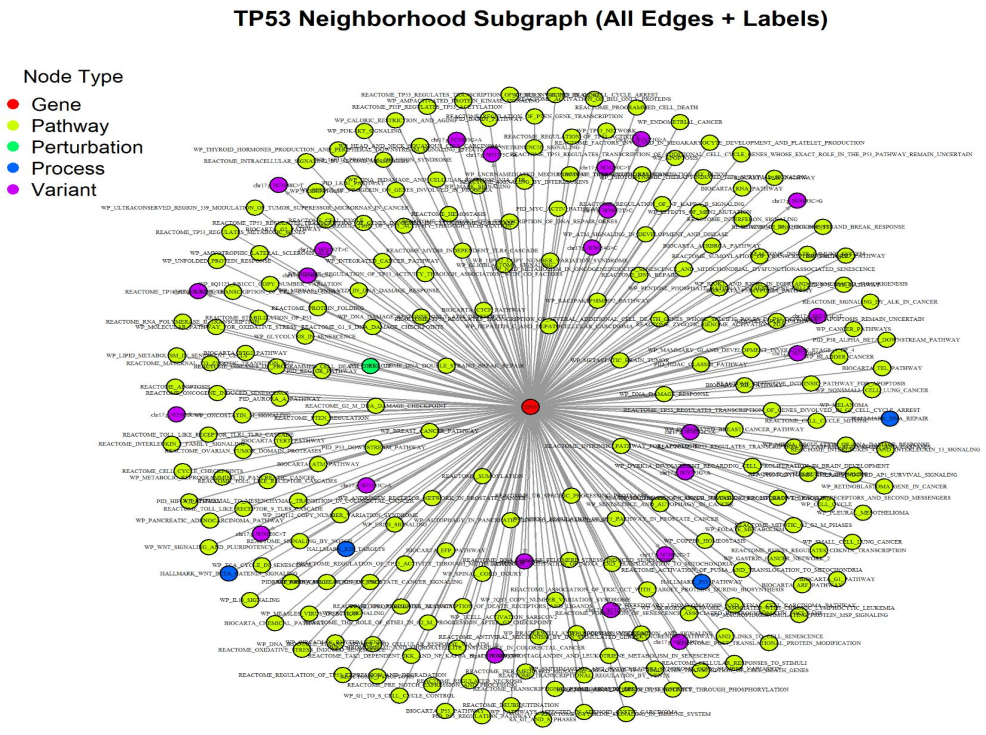
Oncogene Induced Senescence in **Melanoma**

- is a tumor-suppressive mechanism; its dysregulation contributes to melanoma progression.

Genes **highlight targets** to restore or bypass senescence; linking them to **drugs** reveals **repurposing**, combinations, and new interventions.



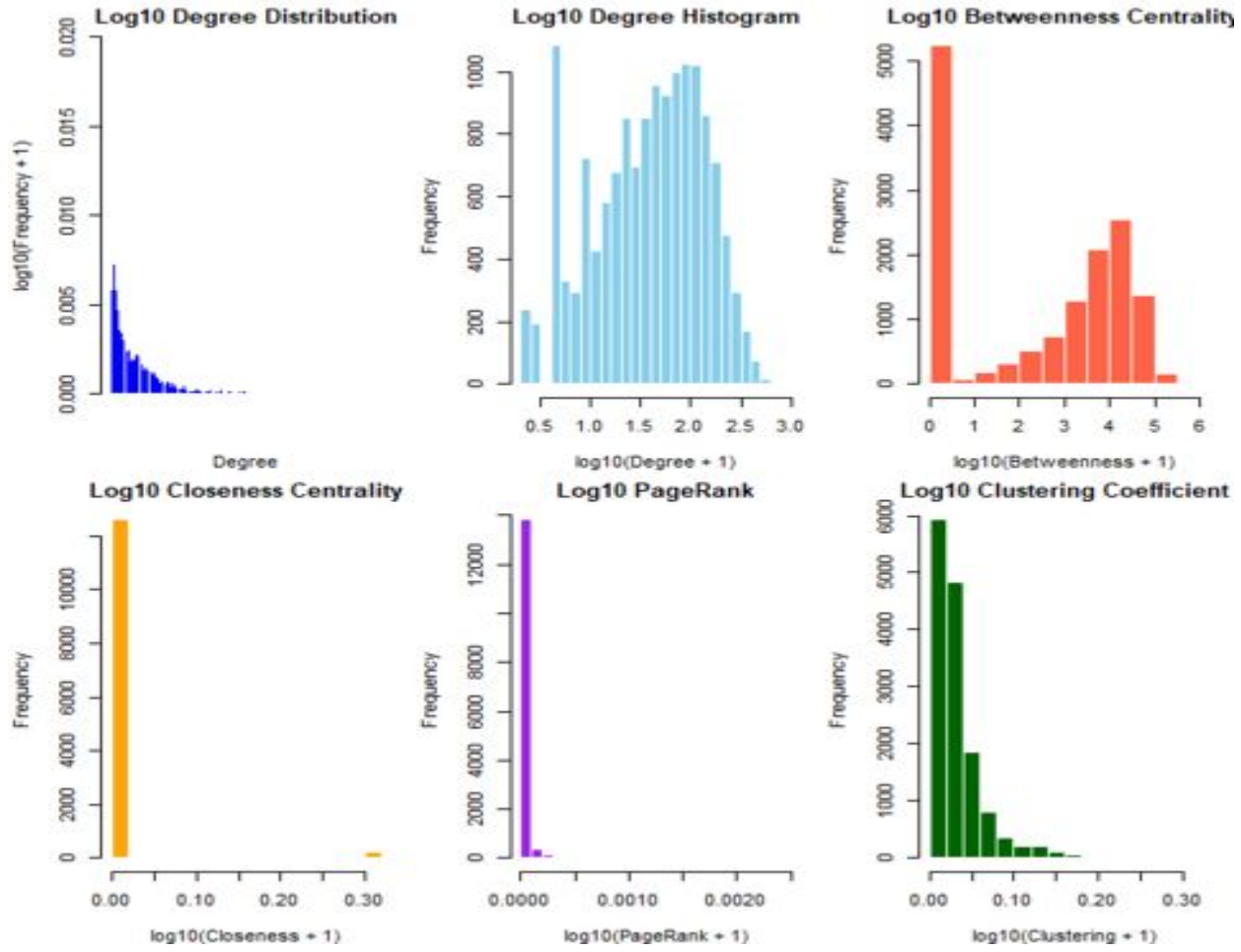
Creating Induced Subgraphs



TP53 Subgraph
(genes+variants+pathways+processes+perturbations)

- Induced subgraphs can be created given any set of nodes.
- Subgraph includes neighboring nodes and their inner connections.
- Can be used for downstream ML model training and inference.

Statistics on the GeNETwork KG subgraph



Induced subgraph of KG (CIViC, subsets of MSigSB, and GTEx co-expression

Degree Distribution: connections per node

Degree Histogram: how often nodes with different numbers of connections appear

Betweenness Centrality: key bridges connecting different parts of the network

Closeness Centrality: shows how close each node is to all other nodes in the network

PageRank: shows the importance of nodes each one connects to

Clustering Coefficient: shows how tightly knit a node's neighbors are

Future Directions

- ClinicalTrials.gov integration
- Add new data layers such as CNAs, gene fusions, drug sensitivity screens
- Graph neural networks could predict novel drug-gene associations
- Expand beyond cancer to include off-label therapeutic indications
- Integrate gene-based GWAS to find new targets

Group Members

US folks:

- Chantera Lazard
- Sangeeta Shukla
- Taha Ahooyi
- Deanne Taylor

UK folks:

- Christine Withers
- Vivien Ho
- Polina Rusina
- Ben Wingfield
- Seeta Ramaraju Pericherla
- Kart Subramanian