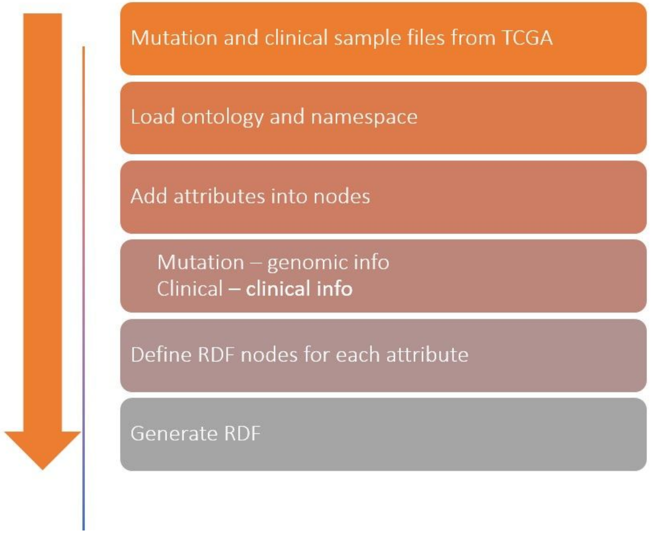
TCGA data to Knowledge Graphs

Team members:
Chiao-Feng Lin (Team lead), Rachit Kumar, Soham Shirolkar, Aniket Naik

Workflow



Mutation and clinical sample files from TCGA

Load ontology and namespace

Add attributes into nodes

Mutation – genomic info
Clinical – **clinical info**

Define RDF nodes for each attribute

Generate RDF

Dataset used:

Colon Cancer (CPTAC-2 Prospective, Cell 2019)

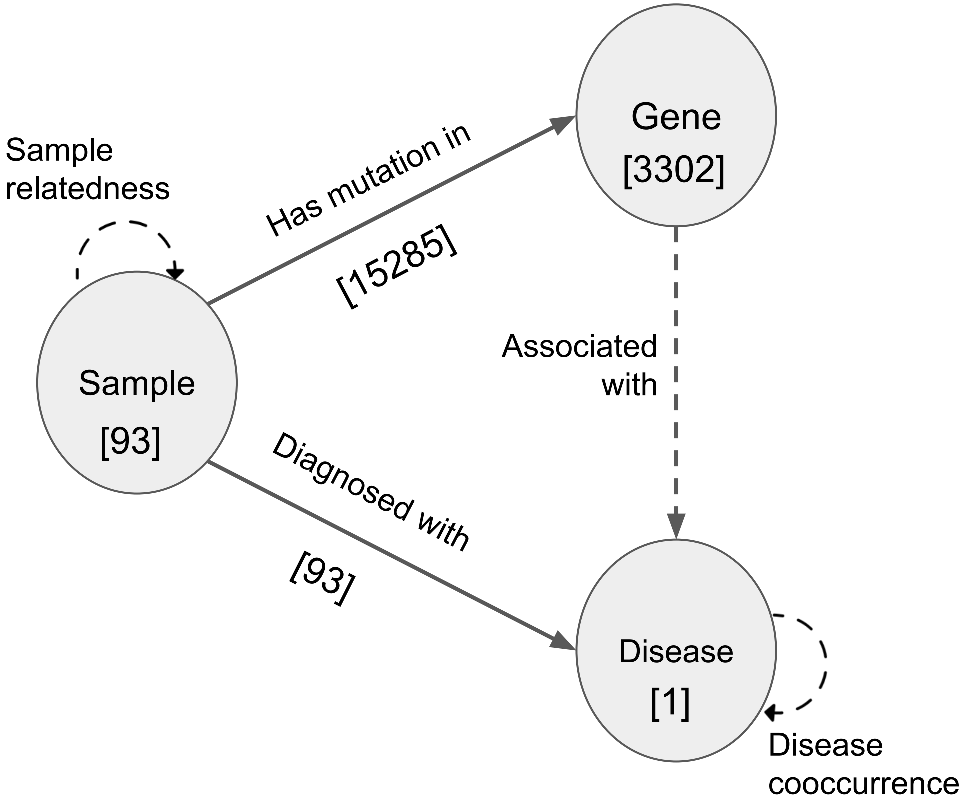
Framework to work with a single cohort, but graph schema will be the same from cohort to cohort

RDF: Resource Data Framework (standard graph format, stores information as concepts and triplets)

Methodology

- Assumption: VCF already converted to MAF
- Import required libraries - rdflib(Graph, namespace, literal), pandas, requests
- Load file and filter out required variables - mutations having ClinVAR annotations (mutations and clinical data)
- Map genes to HGNC IDs and diseases to MONDO IDs
- Construct RDF triplets (Samples -> Genes, Samples -> Diseases)
- Store and output as RDF graph

Results (Schema and Graph Details)



Solid lines represent implemented edges

Dashed lines represent edges to be implemented

(Framework allows for extensions across cohorts easily, just concatenate the MAF files)

Future Steps

- Further enrich the graph with Gene-Disease associations (DisGeNet)

- Visualize the entire graph (GraphDB)

- Include more node attributes
    - (individual patient clinical information - Age, Sex, etc.)

- Provide explicit framework for combining cohorts (concatenating cohort files)

- Wrap explicit VCF -> MAF conversion (rather than relying on user/dataset)