Phase 1 - Purpose Definition

Oncovirus Knowledge Graph

1 Introduction

Oncoviruses, which are viruses that can induce cancer development, are responsible for a significant portion of human cancers. Examples include Human Papillomavirus (HPV), Hepatitis B and C viruses (HBV, HCV), and Epstein–Barr Virus (EBV). Understanding the molecular interactions between viral and host proteins, as well as their influence on cellular pathways and phenotypes, is essential to reveal oncogenic mechanisms and develop targeted therapies or preventive vaccines.

However, the knowledge surrounding oncoviruses is fragmented across multiple data sources (UniProt, KEGG, Reactome, literature). The relationships between viruses, proteins, pathways, phenotypes, and cancers are not easily integrated or queried in a unified way.

A Knowledge Graph (KG) provides a powerful framework to integrate this heterogeneous information into a connected structure that enables semantic reasoning, hypothesis generation, and advanced querying. This project aims to build such a KG following the ITELOS methodology, starting with Phase I to define the KG’s purpose, domain scope, intended scenarios, personas, and competency questions, followed by a visualization of the ER diagram.

2 Purpose Definition

2.1 Informal Purpose

The purpose of this knowledge graph is to establish a unified and semantically rich representation of the biological mechanisms through which viruses contribute to cancer development. It seeks to integrate dispersed information about oncoviruses, their encoded proteins, and the host cellular components they interact with, in order to elucidate the molecular and systemic connections that underlie viral oncogenesis.

By connecting entities such as viruses, host and viral proteins, cellular pathways, phenotypes, cancers, and vaccines, the graph will enable researchers to navigate complex relationships that are otherwise scattered across literature and databases. The goal is to create a framework that not only supports data integration but also facilitates reasoning, discovery, and hypothesis generation.

This knowledge graph will serve as a bridge between molecular biology, computational modeling, and clinical understanding, allowing scientists to trace how a virus infects a host, manipulates its molecular machinery, disrupts cellular pathways, and ultimately drives tumorigenesis. Furthermore, it will provide a foundation for vaccine and therapeutic research by highlighting targets and pathways implicated in virus-induced cancers.

2.2 Domain of Interest

The Domain of Interest encompasses the molecular, cellular, and clinical dimensions of virus-induced cancers, focusing on the interactions between oncogenic viruses, host cellular machinery, and cancer-related pathways. The knowledge graph integrates biological, molecular, and clinical information to model how viruses contribute to tumorigenesis across different organisms and tissue types.

2.2.1 Space

The spatial scope of this study primarily involves human biological systems, where the focus lies on host–virus interactions at the molecular and cellular levels. Specifically, the graph captures protein interactions within cellular compartments and within tissues or organs that serve as the primary sites of infection and cancer development (e.g., liver for HBV/HCV, cervix for HPV, lymphatic tissue for EBV).

The viruses represented include major oncoviruses of clinical relevance, such as Human Papillomavirus (HPV), Hepatitis B Virus (HBV), Hepatitis C Virus (HCV), Epstein–Barr Virus (EBV), and Human T-lymphotropic Virus (HTLV-1), reflecting a global distribution across diverse populations and cancer types.

2.2.2 Time

The temporal scope extends from the initial viral infection to the development and progression of virus-induced cancers. This includes stages such as viral entry, genomic integration, latency, immune evasion, cellular transformation, and tumor formation.

While the knowledge graph is not bound to specific chronological datasets, it integrates findings from studies conducted between 2000 and 2025, encompassing both historical discoveries and contemporary research in virology, oncology, and bioinformatics.

2.3 Scenarios

These represent practical use cases demonstrating how the KG can be used.

1. **Identify Virus–Cancer Links:**
   A researcher queries which viruses are causally associated with specific cancer types (e.g., HPV → cervical cancer).

2. **Map Virus–Host Interactions:**
   A bioinformatician explores which human (host) proteins interact with viral proteins to identify potential therapeutic targets.

3. **Analyze Disrupted Pathways:**
   A molecular biologist traces which signaling pathways are affected by viral proteins to understand oncogenic mechanisms.

4. **Explore Vaccine Coverage:**
   A public health analyst checks which viruses have vaccines and which remain without preventive measures.

5. **Phenotype-based Discovery:**
   A virologist studies virus phenotypes (e.g., tropism, virulence) and correlates them with cancer incidence or severity.

6. **Functional Context of Proteins:**
   A computational scientist investigates how host and viral protein functions/localizations relate to pathway disruptions.

7. **Comparative Oncovirus Analysis:**
   An academic compares multiple oncoviruses to identify shared pathways, target proteins, or phenotypic traits.

2.4 Personas

Personas represent typical users of the knowledge graph and their goals.

1. **Dr. Elisa Marino – Virologist**
   Focus: Understanding viral protein functions and phenotypes.
   Uses KG to connect viral mutations with oncogenic potential.

2. **Dr. James Liu – Computational Biologist**
   Focus: Integrating datasets and analyzing protein–protein interaction networks.
   Uses KG to run queries linking virus-host interactions with pathway disruptions.

3. **Dr. Maria Rossi – Oncologist**
   Focus: Translating molecular data into clinical understanding.
   Uses KG to explore which viral infections are associated with specific cancers.

4. **Dr. Daniel Ortega – Vaccine Researcher**
   Focus: Designing or improving vaccines targeting oncoviruses.
   Uses KG to identify viral proteins that are ideal immunogenic candidates.

5. **Dr. Anna Becker – Molecular Pathologist**
   Focus: Examining cellular pathways impacted by viral infections in tumor samples.
   Uses KG to cross-reference altered pathways and cancer data.

6. **Prof. Samuel Green – Bioinformatics Educator**
   Focus: Teaching integrative omics and knowledge graph technologies.
   Uses KG as a pedagogical example of biological knowledge modeling.

7. **Dr. Rania Al-Hassan – Data Scientist**
   Focus: Developing machine learning models on biological graphs.
   Uses KG embeddings to predict novel virus–cancer associations.

3 Competency Questions (CQs)

These questions define what the knowledge graph must be able to answer.

1. **Virus–Cancer Relationships**
   Which viruses are known to cause specific types of cancer?

2. **Protein Interactions**
   Which host proteins interact with specific viral proteins, and what is their biological function?

3. **Pathway Involvement**
   Which pathways involve the host proteins that interact with viral proteins?

4. **Vaccine Coverage**
   Are there existing vaccines targeting a given oncovirus?

5. **Viral Phenotypes**
   What phenotypic traits (e.g., virulence, tropism) are associated with a specific virus?

6. **Functional Context**
   What are the annotated functions and localizations of host and viral proteins involved in oncogenic interactions?

7. **Knowledge Discovery**
   Can we infer new virus–cancer associations through shared host pathways or phenotypic similarities?

4 Concepts Identification – ETypes & Attributes

Entity Name	Category	Description / Role in KG	Attributes
Virus	**Core**	Represents oncogenic viruses responsible for specific cancers.	id, virusCode, organismName, species, host
Cancer	**Core**	Represents cancers associated with viral infections.	id, type, primarySite
Vaccine	**Core**	Represents preventive vaccines targeting oncoviruses.	id, name, description, source

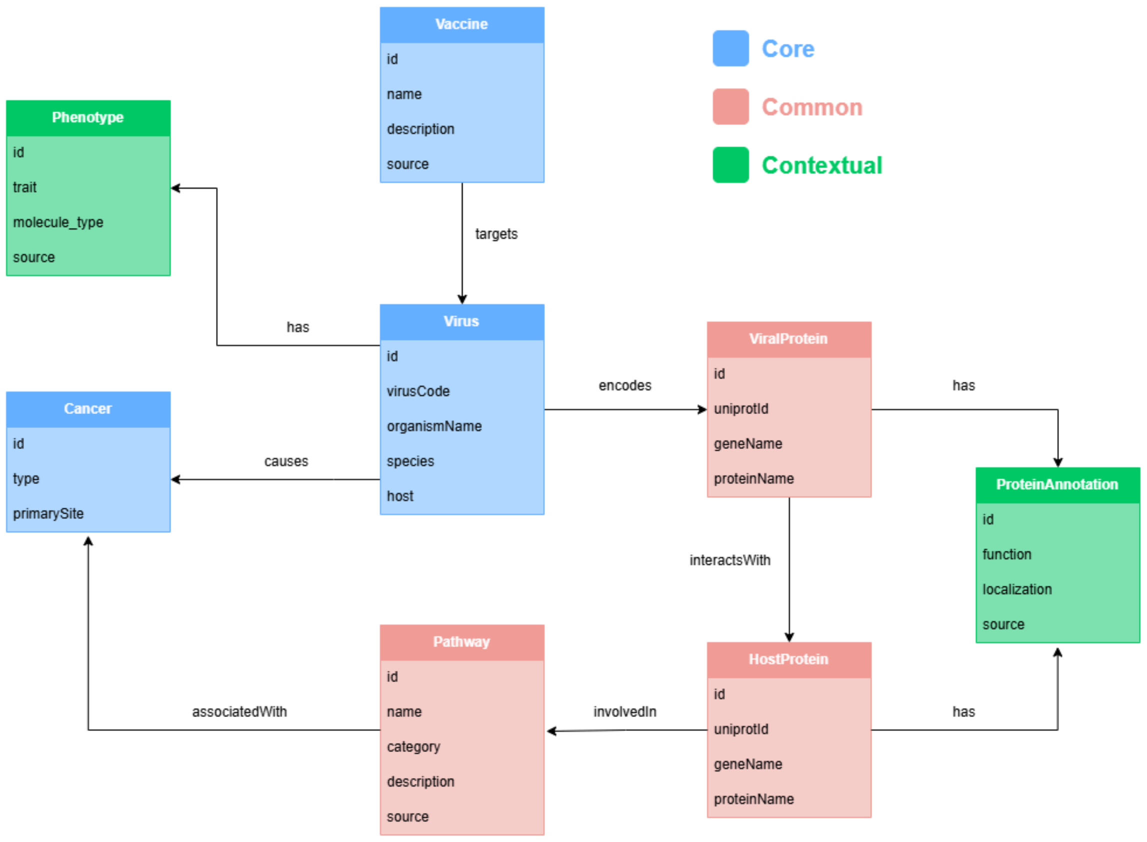
ViralProtein	**Common**	Encoded proteins from oncoviruses that interact with host proteins.	id, uniprotId, geneName, proteinName
HostProtein	**Common**	Human proteins targeted or affected by viral proteins.	id, uniprotId, geneName, proteinName
Pathway	**Common**	Biological pathway or process involving host proteins.	id, name, category, description, source
ProteinAnnotation	**Contextual**	Stores metadata describing protein-specific properties (function, localization, evidence).	id, function, localization, source
Phenotype	**Contextual**	Represents observable virus-related traits (e.g., virulence, tropism, infectivity).	id, trait, type, source

5 Relationships (Edges)

- **Vaccine targets → Virus**
- **Virus causes → Cancer**
- **Virus has → Phenotype**
- **Virus encodes → ViralProtein**
- **ViralProtein interactsWith → HostProtein**
- **ViralProtein has → ProteinAnnotation**
- **HostProtein has → ProteinAnnotation**
- **HostProtein involvedIn → Pathway**
- **Pathway associatedWith → Cancer**

6 ER Diagram

Based on the defined Competency Questions, we created the following Entity–Relationship (ER) diagram. We took into consideration all relevant biological and biomedical scenarios, prioritizing the oncovirology and cancer research areas of interest. To achieve this, we integrated data derived from public biomedical sources and viral oncology literature, focusing on the relationships between viruses, cancers, vaccines, and molecular interactions. We carefully inspected each data source and corresponding schema, noting variables and features of potential interest. During the ER diagram development, we aimed to group and diversify features, Entity Types (ETypes), and properties in alignment with the formalized purpose of modeling virus–host–cancer interactions.



We define the following Entity Types (ETypes) and categorize them as Common, Core, or Contextual, following the same principles used in our domain analysis:

• Common ETypes:

– **ViralProtein:** represents proteins encoded by oncoviruses that can directly or indirectly interact with host proteins. This EType is fundamental for capturing the molecular mechanisms through which viruses modulate host cellular pathways.
– **HostProtein:** represents human proteins targeted or affected by viral proteins. This EType provides the necessary foundation to describe virus–host interactions at the molecular level.
– **Pathway:** represents biological pathways or molecular processes involving host proteins. It enables grouping of host proteins into functional contexts, facilitating higher-level reasoning about disease mechanisms.

- **Core ETypes:**

– **Virus:** represents oncogenic viruses responsible for specific cancers. This EType serves as a primary node in the knowledge graph, connecting to multiple other entities (e.g., cancers, vaccines, and viral proteins).
– **Cancer:** represents types of cancers associated with viral infections. It allows linking molecular mechanisms to clinical outcomes and supports competency questions about oncogenesis and disease prevalence.
– **Vaccine:** represents preventive vaccines developed to target or neutralize specific oncoviruses. This EType helps address competency questions related to prevention and immunization strategies.

- **Contextual ETypes:**

– **ProteinAnnotation:** stores metadata describing protein-specific properties such as function, subcellular localization, and evidence type. This EType enriches the molecular entities (viral and host proteins) with additional biological meaning.
– **Phenotype:** represents observable virus-related traits such as virulence, cell tropism, or infectivity. This EType is crucial for contextualizing viral behavior and linking it to pathogenic potential.

Relationships between ETypes

We identify a set of biologically meaningful relationships capturing causation, interaction, and participation patterns within the system. Some are straightforward hierarchical or associative links (e.g., “Vaccine targets Virus”), while others represent complex molecular associations (e.g., “ViralProtein interactsWith HostProtein”). These edges are central to addressing the competency questions concerning oncogenic mechanisms, molecular targets, and preventive interventions.