Omer_collab_tumour_data_2025

July 11th

Cell status	Annotation coarse	Annotation granular	top_sig_enriched_pathways	sig_upreglated_top10
Malignant	AC-gliosis-like	AC-gliosis-like 1	['HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION', 'GOBP_AMEBOIDAL_TYPE_CELL_MIGRATION', 'GOBP_TISSUE_MIGRATION', 'GOBP_REGULATION_OF_EPITHELIAL_CELL_MIGRATION', 'GOBP_ENDOTHELIAL_CELL_MIGRATION']	['POSTN', 'JPH1', 'EYA4', 'RTN1', 'LAMA2', 'AC092957.1', 'TNC', 'IGFBP7', 'COL23A1', 'NAMPT']
Malignant	AC-gliosis-like	AC-gliosis-like 2	['GOBP_WOUND_HEALING', 'HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION', 'GOBP_MICROTUBULE_BASED_MOVEMENT', 'GOBP_AMEBOIDAL_TYPE_CELL_MIGRATION', 'GOBP_EPITHELIAL_CELL_PROLIFERATION']	['CFAP54', 'IGFBP7', 'VOPP1', 'AQP4', 'LANCL2', 'AC074351.1', 'DTHD1', 'AC012405.1', 'RTN1', 'MTRNR2L12']

Problem: GO/pathway enrichment results are to terms that are very general terms and refer to processes which may not be relevant to this ***context***.

Potential LLM/agent-based solutions:

1. Use LLM queries containing gene lists plus context to mine the literature, returning references →
    1. Use agentic AI to extract Gene:GO links → Validate against GO DBs; test enrichment scores for GO terms against longer gene lists/trancriptomic data
    2. Second round query for genes necessary for processes/pathways returned → bioinformatic tests of transcriptomic data

2. Find GO/reactome/pathway-DB annotations for genes on list → Use LLMs to rank relevance to context → test enrichment scores against bioinformatic data

LLMs - background

- LLM searches can mine latent knowledge from the literature, returning supporting references Adding relevant context to these searches increases the likelihood that the latent knowledge returned is relevant to the question being asked.
- LLM results can include hallucinations, but these are reduced by deep-search (multiple rounds of searching and synthesis) and can be ameliorated by LLM driven checks against returned refs.

cell_status	annotation_granular	sig_upreglated_top10
Malignant	AC-gliosis-like 1	['POSTN', 'JPH1', 'EYA4', 'RTN1', 'LAMA2', 'AC092957.1', 'TNC', 'IGFBP7', 'COL23A1', 'NAMPT']

Working pipeline

Partially working / in progress

TODO

Contextual deep search

The following is a list of genes enriched in Malignant cells from a (human) patient with glioblastoma annotated as Astrocyte-gliosis-like. What might this list say about the state, properties and interactions of this cell type? Do not limit your answers to the cancer literature, include normal cell-cell interactions, development and cross-cutting metabolic states. ...

1. Extracellular Matrix Remodeling and Invasion

Primary genes involved: POSTN, TNC, LAMA2, COL23A1

Process description: These cells exhibit extensive ECM remodelling capabilities, facilitating invasion and metastasis[1][2]. **POSTN (periostin)** acts as a matricellular protein that promotes cancer cell survival, EMT, invasion, and metastasis through integrin receptor interactions[1]. In glioblastoma specifically, POSTN promotes stem cell self-renewal via the αVβ3/PI3K/AKT/β-catenin/FOSL1 pathway while recruiting immunosuppressive microglia[3]. **TNC (tenascin-C)** is an ECM glycoprotein that promotes cell migration and invasion, while **LAMA2** encodes laminin alpha-2, a critical basement membrane component[4]....

Assertion validation against refs /weeding (Paper QA based RAG tests)

extracellular matrix remodeling ==> GO:0030198 extracellular matrix organization

LAMA2 : basement membrane ==> GO:0005604 basement membrane

Agent validates via GO DB queries

(knowledge graph mapping agent)

Extract gene: process/cell component relationships & map to GO, reactome...

Second round query for gene necessary for processes

ECM remodeling facilitating invasion and metastasis

Required genes for this function:
- **Integrin signaling pathway**: ITGA1, ITGA8, ITGAV, ITGB1, ITGB3, FERMT2[5]
- **FAK signaling**: PTK2 (FAK), SRC, BCAR1
- **ECM remodeling**: MMP2, MMP9, TIMP1, TIMP2, COL1A1, COL4A1
- **PI3K/AKT pathway**: PIK3CA, PIK3R1, AKT1, AKT2, MTOR, PTEN[6][7]

Grounding agent - Extracted biological process terms + GO mappings from output contextual query for:

cell_status	annotation_granular	sig_upreglated_top10
Malignant	AC-gliosis-like 1	['POSTN', 'JPH1', 'EYA4', 'RTN1', 'LAMA2', 'AC092957.1', 'TNC', 'IGFBP7', 'COL23A1', 'NAMPT']

- extracellular matrix remodeling ==> GO:0030198 extracellular matrix organization
- metabolic stress adaptation ==> None*
- calcium dysregulation ==> GO:0055074 calcium ion homeostasis
- enhanced invasive potential ==> None*
- epithelial-mesenchymal transition ==> GO:0001837 epithelial to mesenchymal transition
- calcium homeostasis dysregulation ==> GO:0055074 calcium ion homeostasis
- ER stress ==> GO:0034976 response to endoplasmic reticulum stress
- cell cycle control ==> None*
- DNA repair ==> GO:0006281 DNA repair
- cellular senescence ==> GO:0090398 cellular senescence
- stress response ==> GO:0006950 response to stress
- immune modulation ==> None*
- angiogenesis ==> GO:0001525 angiogenesis
- vascular remodeling ==> GO:0001974 blood vessel remodeling
- cancer-associated fibroblast interactions ==> None*

Coming soon knowledge graph agent extract GO:gene relationships

*Where there is no relevant GO term, USE LLMs to populate with genes on-the-fly.

Suggested potential integration into GO?

Find GO/reactome/pathway-DB annotations for genes on list → Use LLMs to rank relevance to context

Glioblastoma-Relevant Processes and Components: Analysis for Astrocyte-like Malignant Cells

Based on comprehensive research into glioblastoma biology and molecular pathways, I have evaluated each process and cellular component in your list for their relevance to astrocyte-like malignant cells from glioblastoma. The analysis considers three key criteria: **relevance to glioblastoma**, **specificity**, and **biological/medical significance**.

Process/Component	Relevance to Glioblastoma	Specificity	Biological/Medical Relevance	Overall Score
insulin receptor signaling pathway	High	Moderate	High	**High**
adipose tissue development	Low	Low	Low	**Low**
positive regulation of gene expression	High	Low	Moderate	**Moderate**
positive regulation of canonical NF-kappaB signal transduction	High	High	High	**High**
positive regulation of ERK1 and ERK2 cascade	High	High	High	**High**
inflammatory response	High	Moderate	High	**High**
signal transduction	High	Low	Moderate	**Moderate**
cell-cell signaling	High	Moderate	High	**High**

Review draft pipeline outputs

- Issues – deepsearch via API returns limited refs compared to chat
  - Could try using bot to query web interface?

https://github.com/Cellular-Semantics/omer_gene_tumour_gene_list/

Questions:

Are results sufficiently interesting/usable?

Tools

- Reference-returning LLM queries (Rapidly evolving)
  - Perplexity Deepsearch/Pro – Direct LLM query => broad queries can work well.
  - PaperQA: low hallucination queries of academic literature. Works best with focused questions.
- Assertion checker (working beta – statistical validation in progress):
  - Decomposes referenced text into atomic assertions; downloads references; Uses paperQA to test assertions against references. Reformats results to table (.md or .csv)
- Agent driven grounding/annotation – alternative to NLP. Combines LLM latent knowledge of synonymy and stemming with agent driven queries of ontology search APIs.
  - Cell type annotator for published h5ad datasets (paper provides context). (working prototype)
- Aurelian (collaborative development with LNBL): a collection of agents for working with ontologies, ontology annotations & knowledge graphs
- Agent driven ontology editing
  - Edits ontologies from GitHub ticket requests, generating Pull Requests for review.
  - Rapidly evolving. Coming soon – new term definition suggestion; assertion checker.

New test case

- Astroycte tumor before progression to aggressive
  - Gene cassettes from co-expression + factorization across tumors => longer gene lists to test (up to ~200 genes)
  - Try explicitly asking for processes in common (pseudo go enrichment)