A Visual Data Mining System for Subcellular Location-specific Knowledge Mining from Multi-Omics Data in Cancer

Minwoo Pak[1], Dabin Jeong[2], Sungjoon Park[1], Jeonghyeon Gu[3] and Sun Kim[124*]

1 Department of Computer Science and Engineering Seoul National University, Seoul, Republic of Korea.
2 Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul, Republic of Korea.
3 Interdisciplinary Program in Artificial Intelligence, Seoul National University, Seoul, Republic of Korea.
4 Institute of Engineering Research Seoul National University, Seoul, Republic of Korea.
* Correspondence: sunkim.bioinfo@snu.ac.kr

Motivation

Cancer dynamics is highly complex that data from a single layer of omics does not give the complete information to study a phenotypic characteristics of a cancer type.
While transcriptomic and proteomic abundance carry crucial information of cell states, studies have found that the two omics often do not agree with each other.
To better capture the biological context of interest, it is imperative to comparatively analyze them in the context of protein localization and biological pathway.
Although there are many data analysis tools for gene expression or protein quantification data, none of them allows researchers to compare different abundance data in the context of protein localization.

Methods: ALPACA

We present ALPACA (A Location-wise Proteome/transcriptome Abundance Comparative Analyzer) a visual data mining system that comparatively analyzes transcriptomic and proteomic abundance data of different cancers in location-wise and biological pathway-specific way.
Our system compartmentalizes the whole transcriptome and proteome abundance and visually presents the discrepancies of different cancers using subcellular locations and biological pathways as a filtering and sorting mechanisms.
Such filtering and sorting adds biological context of interest to the data analysis to aid the identification of potential biomarkers. ALPACA enables researchers explore the vast search space of numerous cancer types and pathway combinations with little effort and time.

Data Sets

Protein Localization Data

COMPARTMENTS, UniProt, Human Protein Atlas, LocTree3 → 34,172 unified protein-location labels

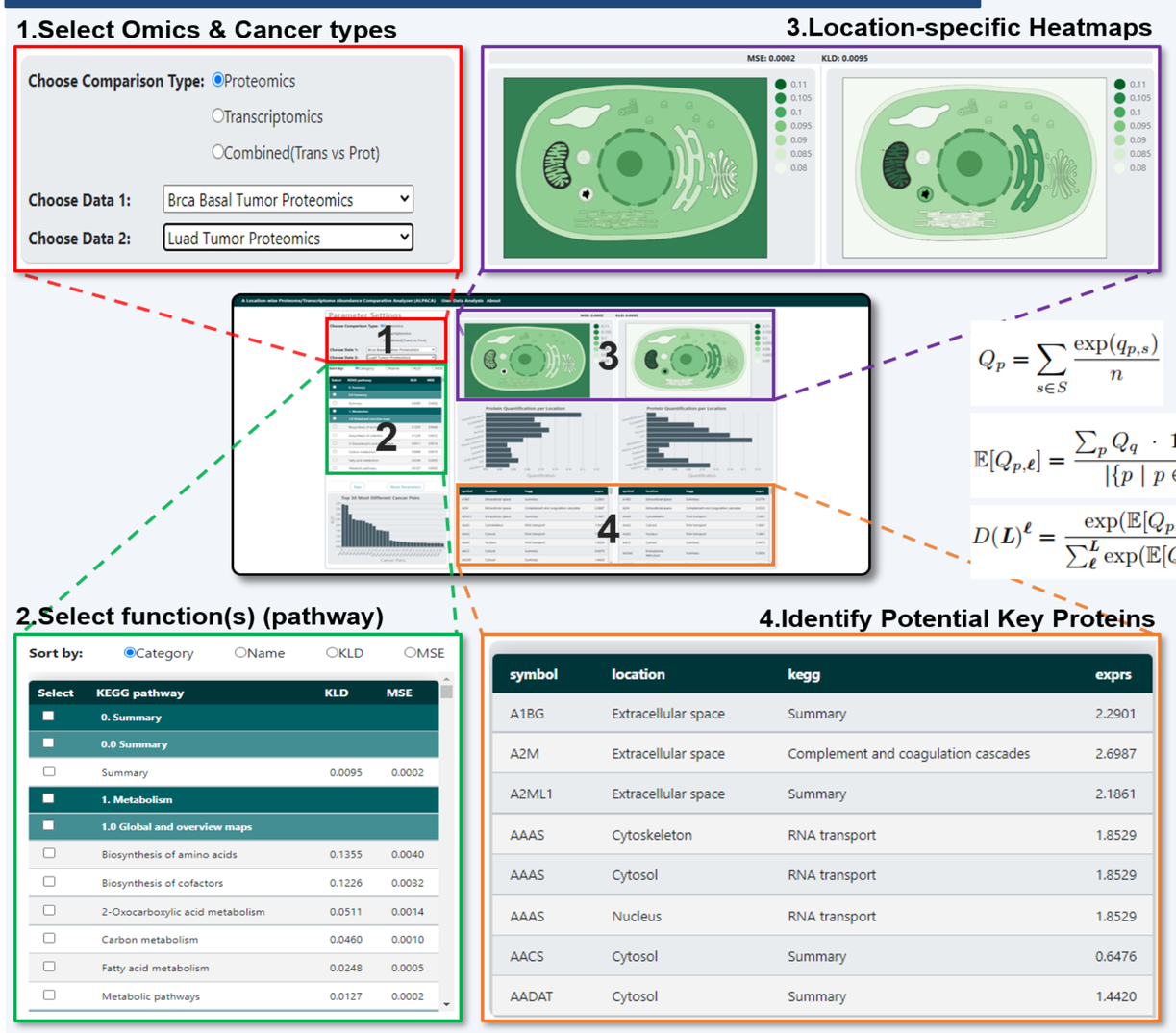
Proteomic / transcriptomic Abundance Data

CPTAC API → Transcriptome / Proteome for 11 Cancer Types/Subtypes

Pathway Data

KEGG pathway database → All Pathway Categories, All Constituent Genes for each Pathway

ALPACA Data Mining Process



$$Q_p = \sum_{s \in S} \frac{\exp(q_{p,s})}{n}$$

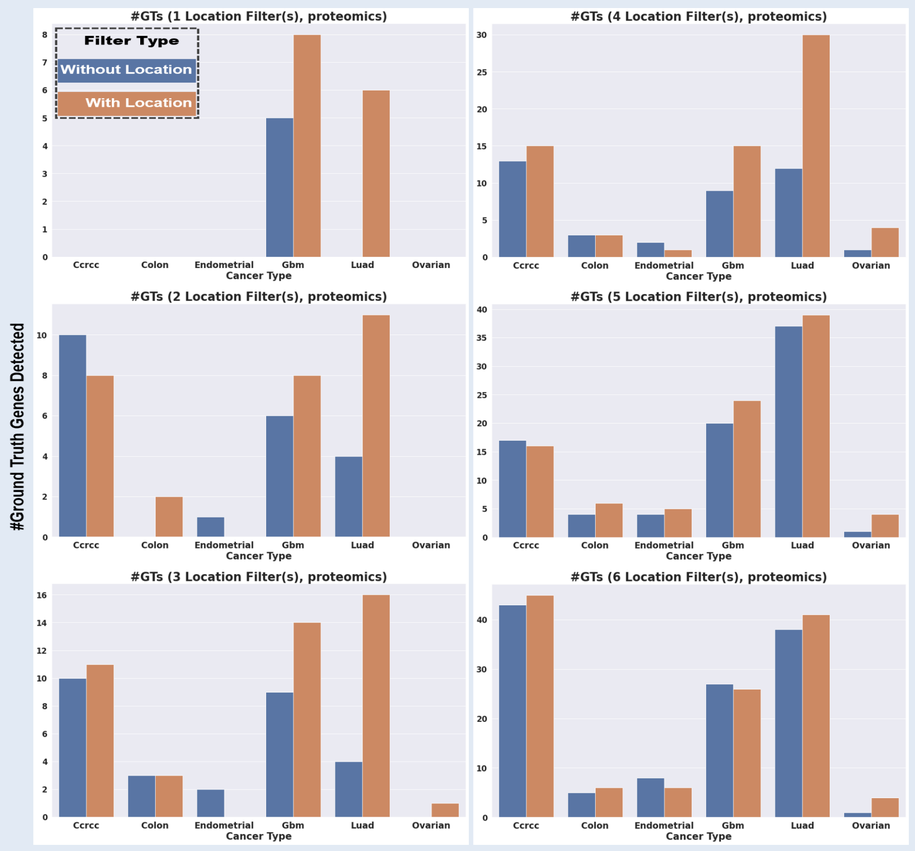
$$\mathbb{E}[Q_{p,\ell}] = \frac{\sum_p Q_q \cdot \mathbb{1}(p \in \ell)}{|\{p \mid p \in \ell\}|}$$

$$D(\boldsymbol{L})^{\ell} = \frac{\exp(\mathbb{E}[Q_{p,\ell}])}{\sum_{\ell}^{L} \exp(\mathbb{E}[Q_{p,\ell}])}$$

Results

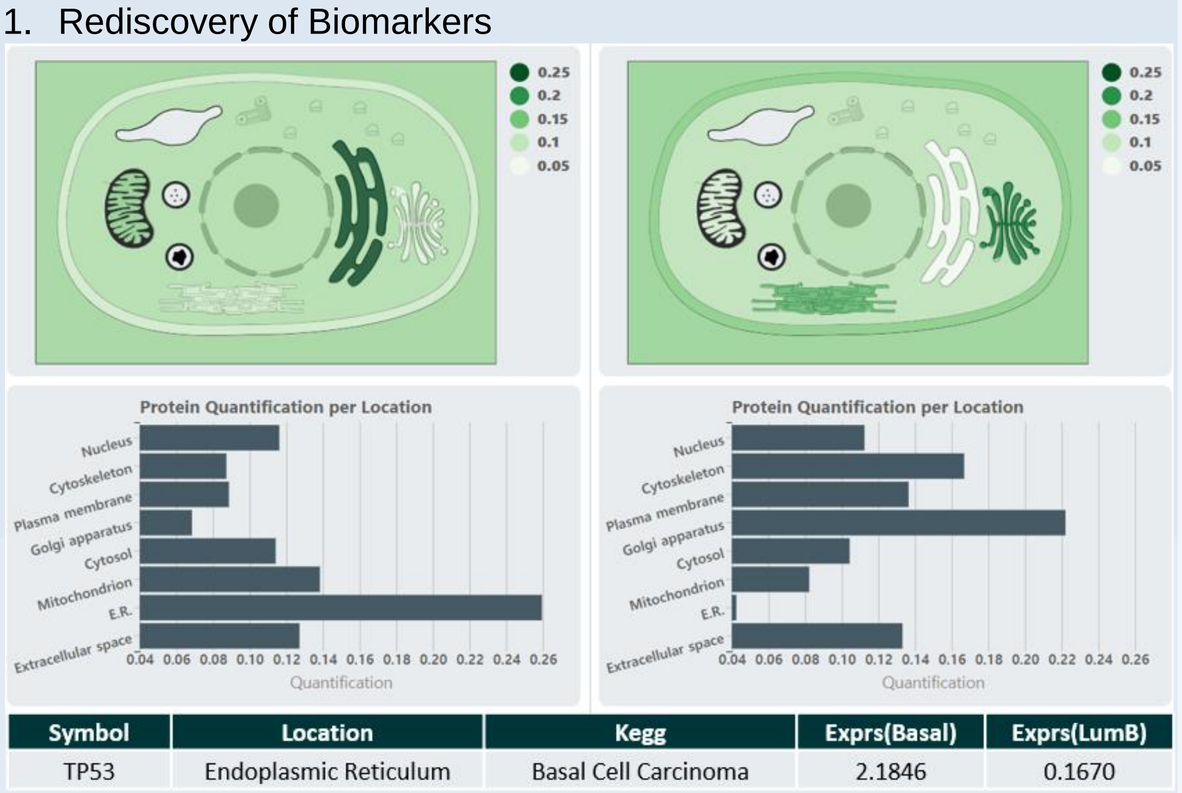
Power of Location Information

We demonstrate the utility of our system through cases studies of various scenarios. Guided by the quantitative difference metrics that ALPACA provides, we were able to efficiently search through different combinations of cancers and pathways to narrow down on the potential key proteins that can help explain the manifestation of a phenotypic difference between two cancers.



Data Mining Case Study

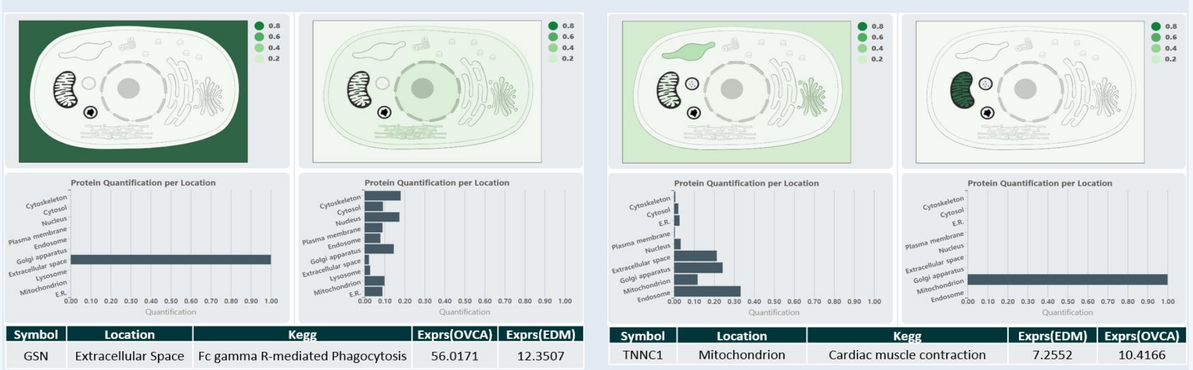
1. Rediscovery of Biomarkers



Symbol	Location	Kegg	Exprs(Basal)	Exprs(LumB)
TP53	Endoplasmic Reticulum	Basal Cell Carcinoma	2.1846	0.1670

Largest discrepancy in E.R.: Overexpression of TP53 → TP53 implies its mutations. Known as a sign of poor prognosis. Supports the worst prognosis of Basal Subtype.

1. Data Mining Guided by Difference Metrics



Symbol	Location	Kegg	Exprs(OVCA)	Exprs(EDM)
GSN	Extracellular Space	Fc gamma R-mediated Phagocytosis	56.0171	12.3507

Symbol	Location	Kegg	Exprs(EDM)	Exprs(OVCA)
TNNC1	Mitochondrion	Cardiac muscle contraction	7.2552	10.4166

Pathways with the largest location enrichment discrepancies according to the difference metrics reveal known biomarkers that may explain the phenotypic characteristics of Gynaecological cancers.
GSN: known for poor prognosis in OVCA. ALPACA shows that GSN is indeed expressed the highest in OVCA compared to other Gynaecological cancers.