Translational Research

Comprehensive multiomics integration: bridging proteomics and genomics

Khatereh Motamedchaboki, Dave Abramowitz, Eric M. Wilson, Mark R Ressler, and Samad Jahandideh

Thermo Fisher Scientific, San Jose, CA, USA • Contact: khatereh.motamed@thermofisher.com

Introduction

Translational researchers face significant challenges in accessing, analyzing, and integrating diverse omics data types, highlighting the need for a robust data integration solution. Existing workflows for proteogenomics data integration are often complex and not user-friendly, especially for non-experts. This complexity hinders a comprehensive understanding of the relationship between genetic variations and protein changes, which is crucial for advancing personalized medicine and leveraging proteogenomics for population and precision health studies.

In this study, we present preliminary data on an advanced data integration workflow that seamlessly integrates proteomics and protein post-translational modifications (PTMs) data with genomic sequence analysis and RNA-seq data. This workflow enables researchers to investigate specific genes or gene sets and retrieve transcriptomics and proteomics data, thereby facilitating a more comprehensive understanding of health and disease states and progressions.

Materials and methods

Training and Testing Dataset: We have used CPTAC[1] proteogenomics dataset as an invaluable resource for our data analysis and generative AI training, enabling us to conduct advanced proteogenomics in cancer research. Clinical Proteomic Tumor Analysis Consortium (CPTAC) dataset, is a comprehensive dataset to accelerate our understanding of the molecular basis of cancer. CPTAC genomic data including; whole genome sequencing (WGS), whole exome sequencing (WXS), and RNA sequences with the GRCh38 reference genome using GDC DNA-Seq Analysis Pipelines and mRNA Analysis Pipelines, respectively was accessed from the Genomic Data Commons (GDC) for over 1500 cancer samples, encompassing a wide range of cancer types, including Endometrial, Renal, Lung Adenocarcinoma and Squamous Cell Carcinoma, Breast, Colon, Ovarian, Brain, Head and Neck, and Pancreatic cancers. Additionally, CPTAC proteomic and phosphoproteomics data were through the CPTAC Proteomics Data Commons (PDC). Clinical data (CNV) provided with the datasets were used for labeling.

We employ Generative AI to generate synthetic datasets from the CPTAC dataset. The characteristics of the generated data, including the distribution of demographic and clinical variables, were then examined. We used the Classification and Regression Trees algorithm which can handle complexity while maintaining interpretability and accessibility. One of its key advantages is its versatility, as it effectively handles both categorical and numerical features across classification and regression tasks. This also accommodates missing values and outliers, reducing the need for extensive data preprocessing. This models the relationships and distributions within the original dataset and uses the generated models to create new, artificial data points that maintain the statistical properties and correlations of the original data without directly copying any individual records. This approach effectively anonymizes the data while retaining its utility for analysis.

Rigorous testing is done to ensure the reliability, and base line performance of our model versus published models like Multi-Omics Factor Analysis (MOFA)[2]. Sample classification performance were evaluated with biological insight gained from our integrated proteogenomics data analysis.

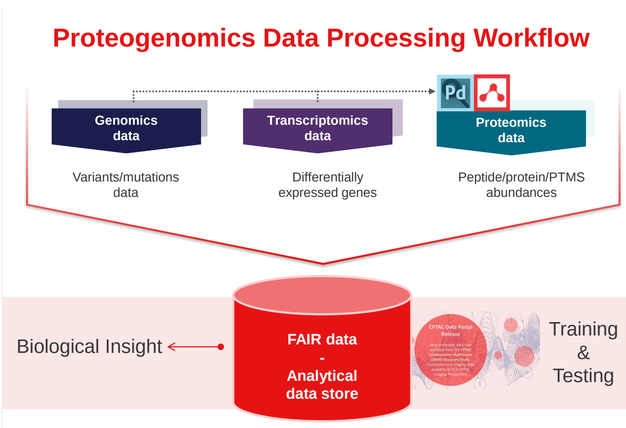


Figure 1. CPTAC Proteogenomics Dataset. The CPTAC, genomics, transcriptomics, proteomics and phosphoproteomics data were utilized to develop an integrated proteogenomics data analysis pipeline.

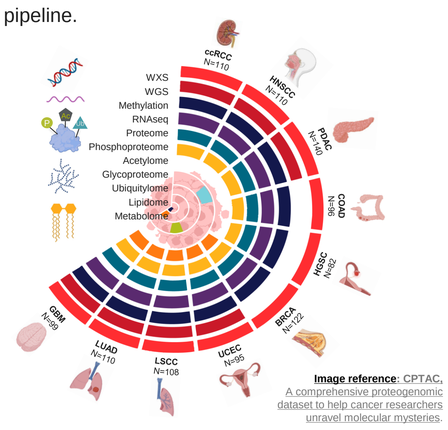
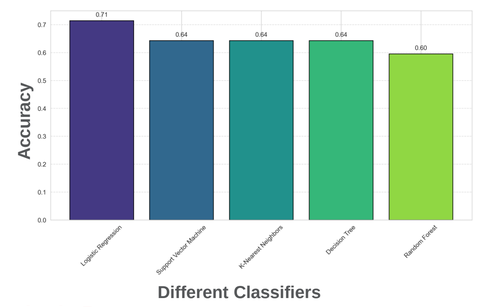


Figure 2. Classification Performance. CPTAC Lung squamous cell carcinoma (LSCC) dataset was analyzed with MOFA and with our integration model trained on synthetic dataset with enhanced classification performance.

MOFA



Synthetic Data

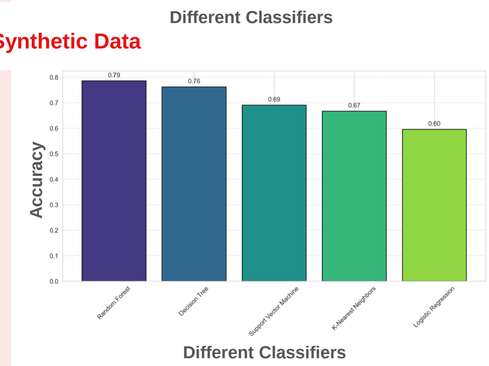
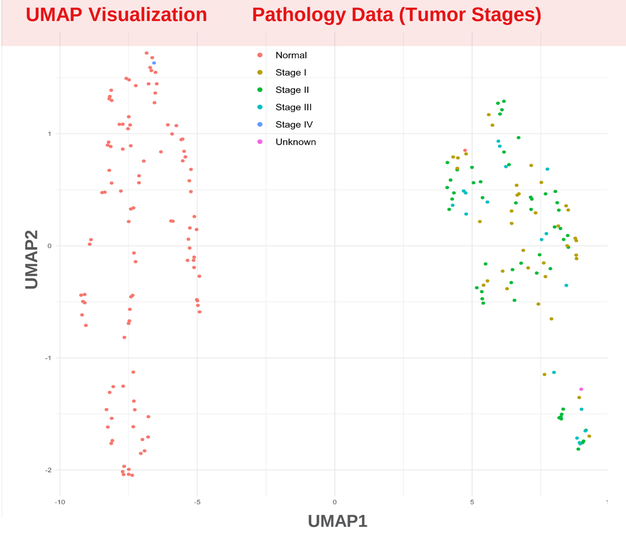


Figure 3. Clustering and Subtyping Performance. CPTAC LSCC dataset was analyzed with our model and the UMAP data visualization of integrated data highlight clear separation between healthy and cancer with improved detection of subgroups.

UMAP Visualization **Pathology Data (Tumor Stages)**



Results

Here we show the performance of our integration model with CPTAC, Ovarian Cancer and Lung squamous cell carcinoma proteogenomic data providing an enhanced integration workflow. We report the disease-related proteogenomics profile, with valuable insights into the molecular mechanisms underlying cancer progression.

Figure 4. Integrated Molecular Profiling: An overview of diverse data modalities, allowing selection of proteogenomic factors explaining different biological insight. Factor 1 in CPTAC Ovarian Cancer highlight shows strong variance in genomics mutations and proteomics versus factor 2 with strong variance in transcriptomic and phosphoproteomics profile.

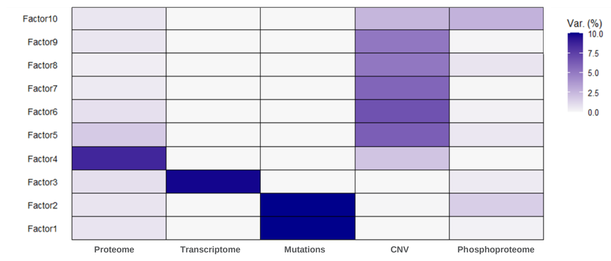


Figure 5. Insight into Protein Regulations. Proteins (A) and Pathways (B) with up and down regulation pattern in factor 1 integrated proteogenomics analysis of CPTAC ovarian cancer highlighting know biomarkers (Red) and new insights (Blue).

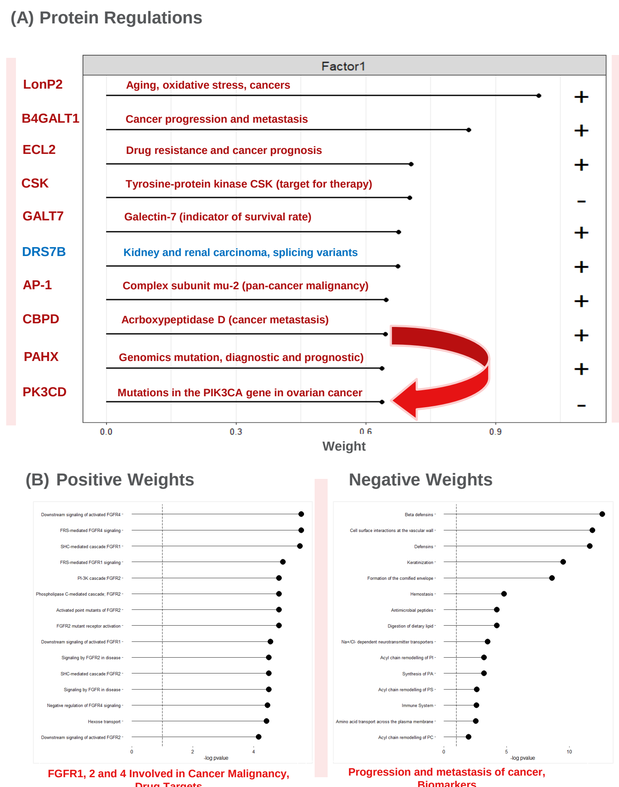


Figure 6. Insight into Disease Progression. Integrated proteogenomics profiling of CPTAC ovarian cancer dataset, identifies samples with known mutations and proteins biomarkers in fast (Red) and slow progressing tumors (Grey).

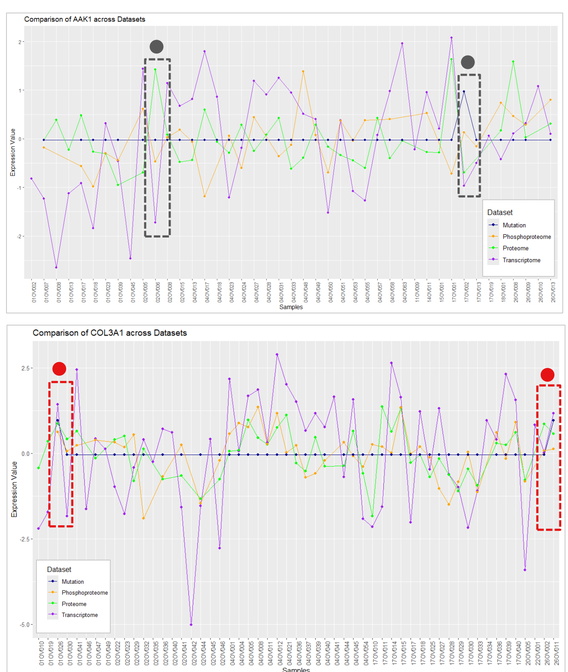


Figure 7. Integrated Molecular Profiling: An overview of diverse data modalities with Factor 1 impact in CPTAC Lung squamous cell carcinoma (LSCC) showing strong variance in transcriptomics and proteomics and Factor 2, phosphoproteomics (A) LSCC cancer stage classification with Factor 1 (B).

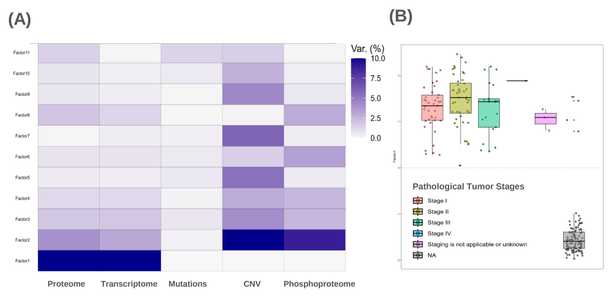


Figure 8. Proteogenomics Sample Classification: CPTAC Lung squamous cell carcinoma (LSCC) classification with Factor 1 data modalities enabling clear classification of samples with cancer from healthy and early and late stages of LSCC. Sample classification based on proteomics data is shown on left (A) and further separation of late and early stage LSCC were achieved with additional transcriptomics data (B).

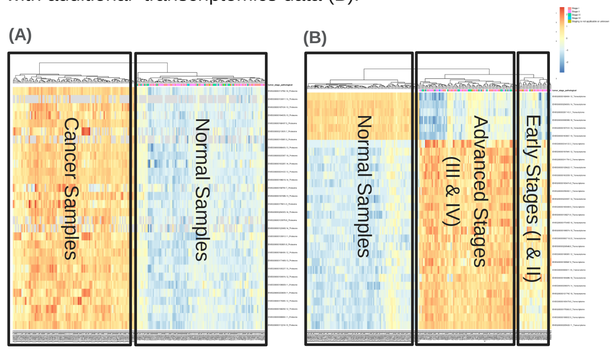


Figure 9. Improved Performance Across Disease Progression. Enhanced classification of CPTAC Lung squamous cell carcinoma (LSCC) compared to baseline performance with MOFA.

	MOFA (LR)				Synthetic Data Approach (RF)			
	Normal	Stage I	Stage II	Stage III	Normal	Stage I	Stage II	Stage III
N (total - real data)	103	41	44	22				
N (test set - 20% of real data)	21	8	9	4	21	8	9	4
True Positives (TP)	20	4	5	1	21	3	5	3
False Positives (FP)	0	7	5	0	0	4	5	0
False Negatives (FN)	1	4	4	3	0	4	4	1
True Negatives (TN)	21	27	28	38	21	31	28	38
Sensitivity	95%	50%	56%	25%	100%	43%	56%	75%
Specificity	100%	79%	85%	100%	100%	89%	85%	100%
Accuracy	98%	74%	79%	93%	100%	81%	79%	98%

Conclusions

- Successful proteogenomic data integration across various CPTAC cancer studies.
- Highlighting potential targets and pathways implicated with the highest weights in healthy and diseased states.
- Providing subtyping, classification and biological insight on complex biological system and diseases,
- Closing the gap between genomics and proteomics.

References

1. Yize Li et al, Proteogenomic data and resources for pan-cancer analysis. Cancer Cell (2023).
2. Ricard Argelaguet et al., Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. Mol Syst Biol (2018)

Trademarks/licensing

Science at a scan
Scan the QR code on the right with your mobile device to download this and many more scientific posters.

Learn more at **thermofisher.com/translationalresearch**