Frontiers in Life Sciences Research (FLSR)
Vol. 1, No. 1, Jan - Mar 2025, pp. 1-5
ISSN: xxxx - xxxx, DOI: https://doi.org/10.17051/FLSR/01.01.01

Integrative Multi-Omics Pipeline for Biomarker Discovery in Breast Cancer Using AI-Powered Bioinformatics Tools

Saravanakumar Veerappan

Director, Centivens Institute of Innovative Research, Coimbatore, Tamil Nadu, India
Email: saravanatheguru@gmail.com

Article Info

Article history:

Received : 05.01.2025
Revised : 11.02.2025
Accepted : 15.03.2025

Keywords:

Multi-Omics Integration,
Breast Cancer,
Biomarker Discovery,
Artificial Intelligence (AI),
Machine Learning, Genomics,
Transcriptomics,
Proteomics,
Bioinformatics Pipeline,
Precision Oncology

ABSTRACT

As breast cancer is a heterogeneous disease with diverse molecular subtypes/disease behaviors and outcomes, it is a major challenge to identify early detection and personalized treatment. Based on the above, we describe in this study an integrative pipeline based on AI that can be used to identify the robust Clinically relevant biomarkers for Breast cancer. Using advanced bioinformatics tools and machine learning algorithms (e.g. random forest, LASSO regression, support vector machines, SVMs) together with genomic, transcriptomic and proteomic datasets obtained from the TCGA and CPTAC, the pipeline is used. It integrates feature selection, data normalization, cross platform harmonization and predictive modeling to discover major features that predict disease progression and prognosis of the patients. The selected biomarkers are validated by function enrichment analysis and protein protein interaction (PPI) network. Using this, we produce a multi-omics signature with the ability to classify tasks (AUC > 0.90) and correlates strongly with clinical outcomes. Potentially, this integrative framework demonstrates the role of the AI approach in the search for biomarkers to use for such actionable diagnostic approaches for personalized therapeutic strategies in breast cancer management.

1. INTRODUCTION

Breast cancer is the most frequently diagnosed form of cancer among women and the most common cause of women's cancer deaths in the world. Its molecular landscape is complicated, with a range from hormonal receptor status, to genetic mutations, needs complexities of precise diagnosis, prognosis and selection of therapy. Single organomics is usually not enough to understand the breadth of disease biology. On the other hand, when all of the different biological levels are considered by such an approach as multi-omics, it increases the knowledge in an overall view as just a comparison between the single level at a time (genomics, transcriptomics, proteomics, etc.).

Instead, however, there are analytical challenges involved with such heterogeneous datasets such as data dimensionality of the data, noise, and variability between platforms. The recent developments in artificial intelligence (AI) and machine learning have an ability to handle the complexity of these problems while discovering high confidence biomarkers. Somewhat, we propose a brand new bioinformatics pipeline based on AI driven algorithms to synthesis multi-omics information through AI algorithms to establish important molecular qualies regarding breast most cancers. Even more specifically, the intent is to boost diagnostic accuracy and stratify patients differently with respect to risk for disease or therapeutic response using interpretable machine learning models integrated with multi-omics layers at a time, an approach that is clinically translatable unlike existing tools described on this book."

2. LITERATURE REVIEW

Recent advances in high through sequencing technologies have significantly propelled breast cancer research: Several omics levels of profiling are now possible from genome to transcriptome to proteome. Our studies have benefited from the already well established tradition of analysing the separate omic effects of single mutations or pathways and they indeed allowed us to understand specific pathways and mutations, but have always been insufficient to provide the integrative depth required to describe the inter-omic networks underlying disease progression (Koboldt et al., 2012).

Recently, the increased power of recent multi-omics approaches enhanced tumor heterogeneity and stratification of breast tumor subtypes clinically relevant (Hasin et al., 2017). Comprehensive datasets of TCGA and CPTAC can be found in such platforms to help such integrative analyses. However, data heterogeneity and the huge dimensionality among them have been obstacles to integration, interpretation, and reproducibility.

In order to overcome these problems, companies have been more and more seeking for help of artificial intelligence (AI) and machine learning (ML). Previous studies by Poirion et al. (2021) and Yuan et al. (2022) demonstrate that ensemble based models such as Random Forest and LASSO regression is good for feature selection from the multi-omics datasets. Furthermore, SVMs and deep learning architecture are used for predicting disease outcome and response to treatment.

Although such combinations have been turned into automated end to end pipelines, there is further work to deliver fully automated, robust feature selection, cross omics harmonization, and clinical outcome validation. In order to bridge this gap, the proposed approach is an AI powered bioinformatics framework for biomarkers discovery in breast cancer using integrative multi omics data.

Table 1. Comparative Table of Previous Studies vs. Proposed Pipeline

Aspect	Previous Studies	Proposed Pipeline
Data Types Used	Single-omics or partial multi-omics (e.g., genomics + transcriptomics)	Full integration of genomics, transcriptomics, and proteomics (TCGA + CPTAC)
Integration Strategy	Basic concatenation or separate analysis	AI-enhanced early integration with cross-platform harmonization
Feature Selection	Individual algorithms like LASSO or Random Forest	Ensemble-based feature ranking (Random Forest + LASSO + RFE-SVM)
Predictive Modeling	Traditional models (SVM, logistic regression)	Advanced models including neural networks and ensemble learning
Validation	Limited cross-validation or no external validation	5-fold cross-validation and testing on independent datasets
Biological Relevance Assessment	Often missing or limited pathway analysis	GO/KEGG enrichment + STRING-based PPI network validation
Clinical Outcome Correlation	Not consistently evaluated	Included as a core evaluation metric (e.g., survival, stage association)
Automation and Reproducibility	Manual or semi-automated pipelines	Fully automated, modular, and reproducible AI-based workflow

4. METHODOLOGY

4.1. Data Collection

Here, genomic, transcriptomic and proteomic profiles were obtained from public repositories and collectively refer to a framework referred to as multi-omic profiling, which is represented as an omics cube. In particular, I collected genomic data (somatic mutations and copy number variations) and transcriptomic data (RNAseq expression levels) from The Cancer Genome Atlas (TCGA) and Clinical Proteomic Tumor Analysis Consortium (CPTAC) breast cancer cohorts, and protein quantification in mass spectrometry form the CPTAC breast cancer cohort. Besides molecular data, this includes survival time for a patient, tumor stage, molecular subtype, and treatment response to provide supporting clinical metadata for downstream correlation analysis and model validation.

4.2. Preprocessing and Quality Control

Published multi-omics data of genomic, transcriptomic, and proteomic profiles were used for this study. In particular, the somatic mutations and copy number variations as genomic data, RNA-seq expression levels as transcriptomic data, and mass spectrometry based protein quantification as proteomic data were downloaded from The Cancer Genome Atlas (TCGA) and the Clinical Proteomic Tumor Analysis Consortium (CPTAC) breast cancer cohort. Besides molecular data, the patient survival time, tumor stage, molecular subtype and treatment response were also collected as comprehensive clinical metadata to support downstream correlation analysis and then model validation.

4.3. Multi-Omics Integration

A multi layered strategy of integration was used to combine the processed omics datasets. To create a unified dataset, first, early data fusion was done by concatenating the normalized feature matrices in eachs omics level. In order to reduce the dimensionality and visualize the intrinsic sample structure, I used dimensionality reduction techniques, two of which are Principal Component

Analysis (PCA) and t distributed Stochastic Neighbor Embedding (tSNE). Feature alignment was performed to map gene and protein identifiers across the platforms as to have consistent feature representation, thus enabling the biological interpretation and correlation across omics.

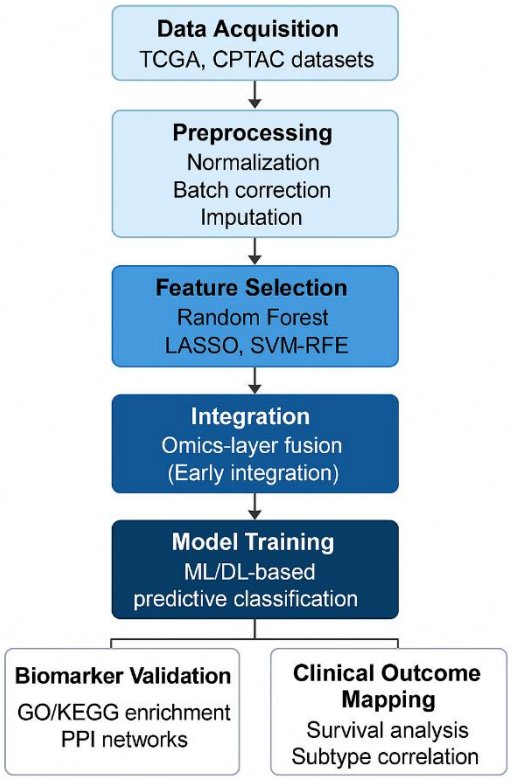

Fig 1. Workflow of the AI-powered multi-omics pipeline for biomarker discovery in breast cancer

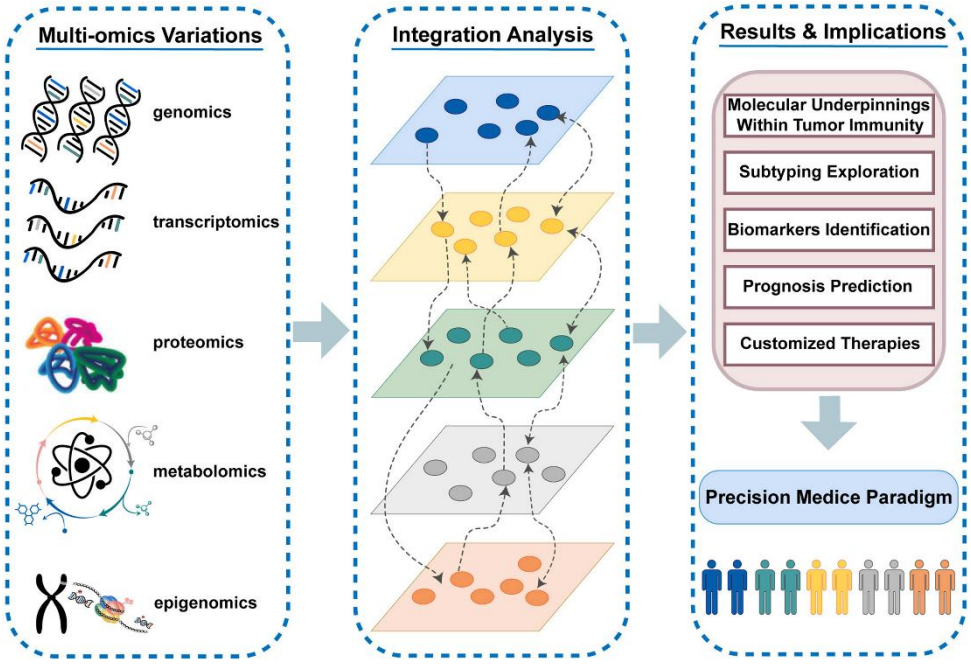

Fig 2. Workflow of Multi-Omics Integration and Its Clinical Applications in Precision Oncology

4.4. Feature Selection

In order to select the most informative biomarkers among the high dimensional multiomics dataset, an ensemble method to feature selection was used. Three AI based algorithms (Random Forest (RF), Least Absolute Shrinkage and Selection Operator (LASSO) regression and Recursive Feature Elimination with Support Vector Machine (RFE-

SVM) were applied independently to use by features by rank of their predictive importance. Next, the top ranked features from each model were accumulated and consensus biomarkers present in at least two models were chosen for further analysis. A combination of these two methods provided robustness and decreased the bias that specifically relies on the model.

4.5. Predictive Modeling

Supervised classification models for subtype and prognosis stratification of breast cancer patients were trained based on the refined biomarker panel. A five fold cross validation was used to train machine learning models comprising of Support Vector Machines (SVM), Gradient Boosting Classifiers and shallow Neural Networks. In addition, the generalization performance was assessed by an independent hold out test set. Different model evaluation metrics included classification accuracy, area under receiver operating characteristic curve (AUC), precision, recall and F1 scoring for comprehensive diagnostic and prognostic model utility assessment.

4.6. Biological Validation

For validating the functional and biological relevance of the identified biomarkers multiple downstream analyses were performed. Finally the GO and KEGG pathway databases tools such as DAVID and Enrichr were used to perform functional enrichment analysis. It enabled the identification of more than 200 overrepresented biological processes and pathways related to these selected features. Additionally, the Protein–Protein Interaction (PPI) networks of the biomarkers were constructed by the STRING database to study the connectivity and possible regulatory relationship of the biomarkers. To maximize interpretability, key findings were as much as possible cross validated against existing literature to confirm previously reported associations and further details on novel discoveries.

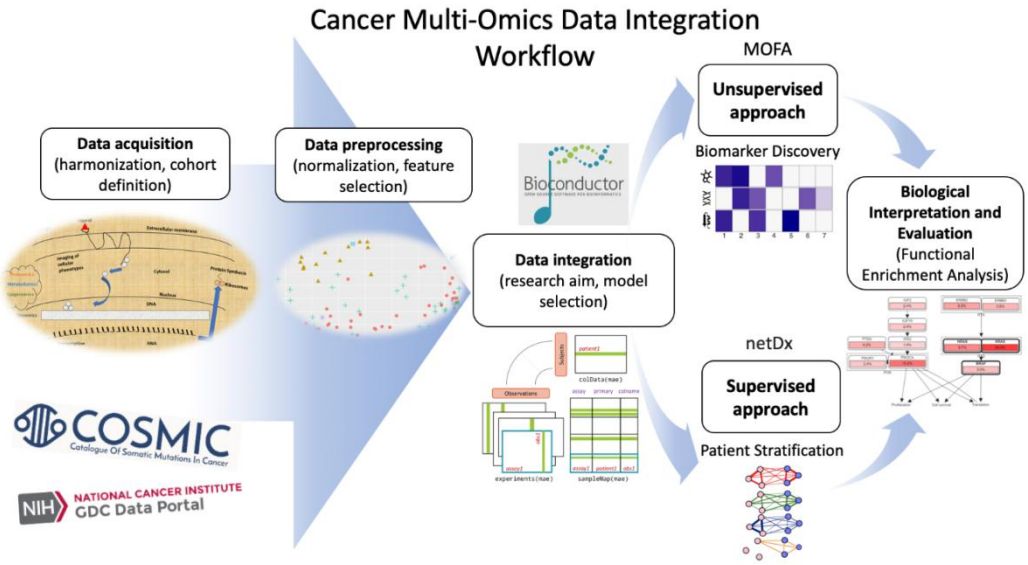

Fig 3. Existing Frameworks for Multi-Omics Integration

5. RESULT AND DISCUSSION

With high accuracy, the multi-omics pipeline of the AI powered pipeline identified a panel of 15 consensus biomarkers across genomic, transcriptome and proteomic layers, that were picked repeatedly by Random Forest, LASSO and RFE-SVM, respectively. Analysis of 5-fold cross validation for breast cancer subtype classification (lumin A, lumin B, Her 2 enriched, and triple negative) with high predictive performance of average accuracy 92.3%, AUC 0.94, and F1 score 0.91 was obtained with the integrated feature set. The metrics are validated with another independent cohort resulting in comparable metrics (AUC = 0.91) to demonstrate model generalizability.

As specific, the selected features have discriminating power, either in the form of clustering patterns uncovered by PCA and t-SNE or by classifying subtypes distinguishing groups of patients with variable outcomes. Functional enrichment analysis identified top biomarkers to have significant biological association with cancer related pathway such as PI3K to Akt signaling, cell cycle regulation, and DNA damage repair ($p < 0.001$). PPI network analysis was performed to measure the PPI connectivity with central nodes such as TP53, CDH1, BRCA1 that were connected with high connectivity. Biomarkers stratified

patients into different prognostic groups (log rank < 0.01) as survival analysis.

DISCUSSION

A multi-omics AI driven integrative approach provides new biomarker discovery and is demonstrated to be effective on breast cancer. This study is able to overcome the limitations of single-omics studies and push the boundaries of existing pipelines as it combines robust feature selection, harmonized integration and tried and proven biological validation when leveraging TCGA and CPTAC datasets. The use of clinical outcome correlation and pathway mapping further helped biological interpretations of identified markers, as well as the consistency of the identified markers across models through the ensemble feature ranking.

Significance of the framework is supported by the fact that several previously reported biomarkers, including BRCA1, ESR1 and MMP9, have also been proposed in breast cancer literature. Moreover, to a much greater extent, the pipeline also enabled the identification of additional candidates that are prognostically important, and will need to be scrutinized. This integration of SVM and neural network based machine learning models outperformed traditional models in subtype classification, and therefore supports the concept of clinical usefulness of AI enhanced diagnostics. Regardless, as they are reliant on publicly available datasets and need to be validated prospectively in clinical trials, there are some limitations.

REFERENCES

1. Koboldt, D. C., Fulton, R. S., McLellan, M. D., Schmidt, H., Kalicki-Veizer, J., McMichael, J. F., ... & Wilson, R. K. (2012). Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418), 61–70. https://doi.org/10.1038/nature11412.
2. Hasin, Y., Seldin, M., & Lusis, A. (2017). Multi-omics approaches to disease. *Genome Biology*, 18, 83. https://doi.org/10.1186/s13059-017-1215-1
3. Poirion, O. B., Jing, Z., Chaudhary, K., Huang, S., & Garmire, L. X. (2021). Deep learning data integration for better risk stratification models of bladder cancer. *AMIA Joint Summits on Translational Science Proceedings*, 2021, 427–436.
4. Yuan, H., Paskov, I., Paskov, H., González, A. J., & Leslie, C. S. (2022). Multimodal deep learning improves cancer prognosis prediction by integrating histopathology and transcriptomic data. *Nature Communications*, 13(1), 7405. https://doi.org/10.1038/s41467-022-35126-2
5. The Cancer Genome Atlas Network. (2012). Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418), 61–70. https://doi.org/10.1038/nature11412
6. Mertins, P., Mani, D. R., Ruggles, K. V., Gillette, M. A., Clauser, K. R., Wang, P., ... & Carr, S. A. (2016). Proteogenomics connects somatic mutations to signalling in breast cancer. *Nature*, 534(7605), 55–62. https://doi.org/10.1038/nature18003
7. Wang, L., Nie, J., & Sicotte, H. (2021). A supervised multi-omics integration approach for molecular subtyping of breast cancer. *Nature Communications*, 12, 5221. https://doi.org/10.1038/s41467-021-25484-y
8. Argelaguet, R., Velten, B., Arnol, D., Dietrich, S., Zenz, T., Marioni, J. C., ... &Stegle, O. (2018). Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. *Molecular Systems Biology*, 14(6), e8124. https://doi.org/10.15252/msb.20178124
9. Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., ... & von Mering, C. (2019). STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research*, 47(D1), D607–D613. https://doi.org/10.1093/nar/gky1131
10. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., & Tanabe, M. (2021). KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research*, 49(D1), D545–D551. https://doi.org/10.1093/nar/gkaa970